UNIVERSITE CLAUDE BERNARD - LYON 1
FACULTE DE PHARMACIE
INSTITUT DES SCIENCES PHARMACEUTIQUES ET BIOLOGIQUES

THESE N°32

THESE

POUR LE DIPLOME D'ETAT DE DOCTEUR EN PHARMACIE

Présentée et soutenue publiquement le 08 mars 2022 par

M. ROBERT Jean-Philippe
Né le 26 janvier 1995
A Chatenay-Malabry (92)

M. ROCHE Valentin
Né le 19 janvier 1996
A Bron (69)

*****

ELABORATION D'UNE METHODE DE DETECTION PRECOCE D'EVENEMENTS INDESIRABLES DECLARES PAR LES PATIENTS SUR LES RESEAUX SOCIAUX : CAS DU LEVOTHYROX® SUR LE SITE DOCTISSIMO®

*****

JURY

PRESIDENT : M. DUSSART Claude, Professeur des Universités - Praticien Hospitalier

DIRECTRICE : Mme SALAM Hanan, Professeur Assistant, Docteur en intelligence artificielle

TUTEUR PEDAGOGIQUE : M. ARMOIRY Xavier, Professeur des Universités – Praticien Hospitalier

MEMBRES : Mme BARDEL DANJEAN Claire, Maitre de Conférences Universitaires – Praticien Hospitalier

M. BOUREILLE Antoine, Pharmacien d'officine, Docteur en pharmacie

UNIVERSITE CLAUDE BERNARD LYON 1

• Président de l'Université	Frédéric FLEURY
• Présidence du Conseil Académique	Hamda BEN HADID
• Vice-Président du Conseil d'Administration	Didier REVEL
• Vice-Président de la Commission Recherche	Petru MIRONESCU
• Vice-Président de la Formation et de la Vie Universitaire	Céline BROCHIER

Composantes de l'Université Claude Bernard Lyon

SANTE

UFR de Médecine Lyon Est	Directeur : Gilles RODE
UFR de Médecine Lyon Sud Charles Mérieux	Directrice : Carole BURILLON
Institut des Sciences Pharmaceutiques et Biologiques.	Directeur : Claude DUSSART
UFR d'Odontologie	Directrice : Dominique SEUX
Institut des Sciences et Techniques de Réadaptation	Directeur : Xavier PERROT (ISTR)

SCIENCES ET TECHNOLOGIES

UFR Fédération Sciences (Chimie, Mathématique, Physique)	Directeur : M. Bruno ANDRIOLETTI
UFR Biosciences	Directrice : Mme Kathrin GIESELER
Département composante Informatique	Directeur : M. Behzad SHARIAT
Département composante Génie Electrique et des procédés (GEP)	Directrice Mme Rosaria FERRIGNO
Département composante Mécanique	Directeur : M. Marc BUFFAT
UFR Sciences et Techniques des Activités Physiques et Sportives (STAPS)	Directeur : M. Yannick VANPOULLE
Polytech Lyon	Directeur : M. Emmanuel PERRIN
I.U.T. LYON 1	Directeur : M. Christophe VITON
Institut des Sciences Financières et d'Assurance (ISFA)	Directeur : M. Nicolas LEBOISNE
Observatoire de Lyon	Directrice : Mme Isabelle DANIEL

UNIVERSITE CLAUDE BERNARD LYON 1
ISPB -Faculté de Pharmacie Lyon

LISTE DES DEPARTEMENTS PEDAGOGIQUES

DEPARTEMENT PEDAGOGIQUE DE SCIENCES PHYSICO-CHIMIQUE ET PHARMACIE GALENIQUE

- **CHIMIE GENERALE, PHYSIQUE ET MINERALE**

Monsieur Raphaël TERREUX (PR)
Madame Julie-Anne CHEMELLE (MCU)

- **CHIMIE ANALYTIQUE**

Madame Anne DENUZIERE (MCU)
Monsieur Lars-Petter JORDHEIM (MCU-HDR)
Madame Christelle MACHON (MCU-PH)
Monsieur Waël ZEINYEH (MCU)

- **PHARMACIE GALENIQUE -COSMETOLOGIE**

Madame Marie-Alexandrine BOLZINGER (PR)
Madame Stéphanie BRIANCON (PR)
Monsieur Fabrice PIROT (PU-PH)
Monsieur Eyad AL MOUAZEN (MCU)
Madame Sandrine BOURGEOIS (MCU)
Madame Danielle CAMPIOL ARRUDA (MCU)
Madame Ghania HAMDI-DEGOBERT (MCU-HDR)
Monsieur Plamen KIRILOV (MCU)
Madame Giovanna LOLLO (MCU)
Madame Jacqueline RESENDE DE AZEVEDO (MCU)
Monsieur Damien SALMON (MCU-PH)
Madame Eloïse THOMAS (MCU)

- **BIOPHYSIQUE**

Monsieur Cyril PAILLER-MATTEI (PR)
Madame Laurence HEINRICH (MCU)
Monsieur David KRYZA (MCU-PH-HDR)
Madame Sophie LANCELOT (MCU-PH)
Madame Elise LEVIGOUREUX (MCU-PH)

DEPARTEMENT PEDAGOGIQUE PHARMACEUTIQUE DE SANTE PUBLIQUE

- **DROIT DE LA SANTE**

Madame Valérie SIRANYAN (PR)
Madame Maud CINTRAT (MCU)

- **ECONOMIE DE LA SANTE**

Madame Nora FERDJAOUI MOUMJID (MCU-HDR)
Monsieur Hans-Martin SPÄTH (MCU-HDR)

- **INFORMATION ET DOCUMENTATION**

Monsieur Pascal BADOR (MCU-HDR)

- **INGENIERIE APPLIQUEE A LA SANTE ET DISPOSITIFS MEDICAUX**

Monsieur Xavier ARMOIRY (PU-PH)
Madame Claire GAILLARD (MCU)

- **QUALITOLOGIE – MANAGEMENT DE LA QUALITE**

Madame Alexandra CLAYER-MONTEMBAULT (MCU)
Monsieur Vincent GROS (MCU-enseignant contractuel temps partiel)
Madame Audrey JANOLY-DUMENIL (MCU-PH)
Madame Pascale PREYNAT (MCU-enseignant contractuel temps partiel)

- **MATHEMATIQUES – STATISTIQUES**

Madame Claire BARDEL-DANJEAN (MCU-PH-HDR)
Madame Marie-Aimée DRONNE (MCU)
Madame Marie-Paule GUSTIN (MCU-HDR)

- **SANTE PUBLIQUE**

Monsieur Claude DUSSART (PU-PH)
Madame Chloé HERLEDAN (AHU)

DEPARTEMENT PEDAGOGIQUE SCIENCES DU MEDICAMENT

- **CHIMIE ORGANIQUE**

Monsieur Pascal NEBOIS (PR)
Madame Amanda GARRIDO (MCU)
Madame Christelle MARMINON (MCU)
Madame Sylvie RADIX (MCU-HDR)
Monsieur Luc ROCHEBLAVE (MCU-HDR)

- **CHIMIE THERAPEUTIQUE**

Monsieur Marc LEBORGNE (PR)
Monsieur Thierry LOMBERGET (PR)
Monsieur Laurent ETTOUATI (MCU-HDR)
Monsieur François HALLE (MCU)
Madame Marie-Emmanuelle MILLION (MCU)

- **BOTANIQUE ET PHARMACOGNOSIE**

Madame Marie-Geneviève DIJOUX-FRANCA (PR)
Madame Anne-Emmanuelle HAY DE BETTIGNIES (MCU)
Madame Isabelle KERZAON (MCU)
Monsieur Serge MICHALET (MCU)

- **PHARMACIE CLINIQUE, PHARMACOCINETIQUE ET EVALUATION DU MEDICAMENT**

Madame Christelle CHAUDRAY-MOUCHOUX (PU-PH)
Madame Catherine RIOUFOL (PU-PH)
Madame Magali BOLON-LARGER (MCU-PH)
Monsieur Teddy NOVAIS (MCU-PH)

Madame Céline PRUNET-SPANO (MCU)
Madame Florence RANCHON (MCU-PH)
Madame Delphine HOEGY (AHU)

DEPARTEMENT PEDAGOGIQUE DE PHARMACOLOGIE, PHYSIOLOGIE ET TOXICOLOGIE

- **TOXICOLOGIE**

Monsieur Jérôme GUITTON (PU-PH)
Madame Léa PAYEN (PU-PH)
Monsieur Bruno FOUILLET (MCU)

- **PHYSIOLOGIE**

Monsieur Christian BARRES (PR)
Madame Kiao Ling LIU (MCU)
Monsieur Ming LO (MCU-HDR)

- **PHARMACOLOGIE**

Monsieur Sylvain GOUTELLE (PU-PH)
Monsieur Michel TOD (PU-PH)
Monsieur Luc ZIMMER (PU-PH)
Monsieur Roger BESANCON (MCU)
Monsieur Laurent BOURGUIGNON (MCU-PH)
Madame Evelyne CHANUT (MCU)
Monsieur Nicola KUCZEWSKI (MCU)
Madame Dominique MARCEL CHATELAIN (MCU-HDR)

- **COMMUNICATION**

Monsieur Ronald GUILLOUX (MCU)

- **ENSEIGNANTS CONTRACTUELS TEMPS PARTIEL**

Madame Aline INIGO PILLET (MCU-enseignant contractuel temps partiel)
Madame Pauline LOUBERT (MCU-enseignant contractuel temps partiel)
Madame Levgeniia CHICHEROVA (ATER)

DEPARTEMENT PEDAGOGIQUE DES SCIENCES BIOMEDICALES A

- **IMMUNOLOGIE**

Monsieur Guillaume MONNERET (PU-PH)
Madame Morgane GOSSEZ (MCU-PH)
Monsieur Sébastien VIEL (MCU-PH)
Monsieur David GONCLAVES (AHU)

- **HEMATOLOGIE ET CYTOLOGIE**

Madame Christine VINCIGUERRA (PU-PH)
Madame Sarah HUET (MCU-PH)
Monsieur Yohann JOURDY (MCU-PH)
Madame Amy DERICQUEBOURG (AHU)

- **MICROBIOLOGIE ET MYCOLOGIE FONDAMENTALE ET APPLIQUEE AUX**

BIOTECHNOLOGIES INDUSTRIELLES

Monsieur Frédéric LAURENT (PU-PH)
Madame Florence MORFIN (PU–PH)
Madame Veronica RODRIGUEZ-NAVA (PR)
Monsieur Didier BLAHA (MCU-HDR)
Madame Ghislaine DESCOURS (MCU-PH)
Madame Anne DOLEANS JORDHEIM (MCU-PH-HDR)
Madame Emilie FROBERT (MCU-PH)
Monsieur Jérôme JOSSE (MCU)

DEPARTEMENT PEDAGOGIQUE DES SCIENCES BIOMEDICALES B

- **BIOCHIMIE – BIOLOGIE MOLECULAIRE - BIOTECHNOLOGIE**

Madame Pascale COHEN (PR)
Madame Caroline MOYRET-LALLE (PR)
Madame Emilie BLOND (MCU-PH)
Monsieur Karim CHIKH (MCU-PH)
Madame Carole FERRARO-PEYRET (MCU-PH-HDR)
Monsieur Anthony FOURIER (MCU-PH)
Monsieur Boyan GRIGOROV (MCU)
Monsieur Alexandre JANIN (MCU-PH)
Monsieur Hubert LINCET (MCU-HDR)
Monsieur Olivier MEURETTE (MCU-HDR)
Madame Angélique MULARONI (MCU)
Madame Stéphanie SENTIS (MCU)
Monsieur Jordan TEOLI (AHU)

- **BIOLOGIE CELLULAIRE**

Madame Bénédicte COUPAT-GOUTALAND (MCU)
Monsieur Michel PELANDAKIS (MCU-HDR)

INSTITUT DE PHARMACIE INDUSTRIELLE DE LYON

Madame Marie-Alexandrine BOLZINGER (PR)
Monsieur Philippe LAWTON (PR)
Madame Sandrine BOURGEOIS (MCU)
Madame Marie-Emmanuelle MILLION (MCU)
Madame Alexandra MONTEMBAULT (MCU)
Madame Angélique MULARONI (MCU)
Madame Marie-Françoise KLUCKER (MCU-enseignant contractuel temps partiel)
Madame Valérie VOIRON (MCU-enseignant contractuel temps partiel)

PR : Professeur des Universités
PU-PH : Professeur des Universités-Praticien Hospitalier
MCU : Maître de Conférences des Universités
MCU-PH : Maître de Conférences des Universités-Praticien Hospitalier
HDR : Habilitation à Diriger des Recherches
AHU : Assistant Hospitalier Universitaire
ATER : Attaché temporaire d'enseignement et de recherche

L'Apothicaire

(Musée des Hospices Civils de Lyon)

Serment des Pharmaciens
Au moment d'être reçu
Docteur en Pharmacie,

En présence des Maitres de la Faculté, je fais le serment :

- D'honorer ceux qui m'ont instruit(e) dans les préceptes de mon art et de leur témoigner ma reconnaissance en restant fidèle aux principes qui m'ont été enseignés et d'actualiser mes connaissances
- D'exercer, dans l'intérêt de la santé publique, ma profession avec conscience et de respecter non seulement la législation en vigueur, mais aussi les règles de Déontologie, de l'honneur, de la probité et du désintéressement
- De ne jamais oublier ma responsabilité et mes devoirs envers la personne humaine et sa dignité
- En aucun cas, je ne consentirai à utiliser mes connaissances et mon état pour corrompre les moeurs et favoriser des actes criminels.
- De ne dévoiler à personne les secrets qui m'auraient été confiés ou dont j'aurais eu connaissance dans l'exercice de ma profession
- De faire preuve de loyauté et de solidarité envers mes collègues pharmaciens
- De coopérer avec les autres professionnels de santé.

Que les Hommes m'accordent leur estime si je suis fidèle à mes promesses. Que je sois couvert(e) d'opprobre et méprisé(e) de mes confrères si j'y manque.

Remerciements

À Monsieur Claude DUSSART,

Nous vous remercions de nous faire l'honneur de présider ce jury de thèse. Merci pour votre disponibilité. Veuillez trouver ici le témoignage de notre profond respect et de notre reconnaissance.

À Madame Hanan SALAM,

Merci infiniment pour votre soutien, votre écoute et votre disponibilité tout au long de ce travail. Vos enseignements de data sciences sont à l'origine de ce sujet et ont permis de mener à bien ce projet. Nous en sommes sûrs, l'utilisation de ces connaissances et compétences sera un précieux atout pour nos exercices futurs ! Ce fut un réel plaisir de travailler avec vous.

À Monsieur Xavier ARMOIRY,

Nous vous remercions de nous avoir accompagnés et d'avoir pris le temps nécessaire pour nous permettre d'améliorer et peaufiner ce travail. Soyez assuré de notre gratitude.

À Madame Bardel-Danjan,

Nous vous remercions de nous faire l'honneur de juger ce travail. Veuillez recevoir nos remerciements les plus sincères.

À Monsieur Antoine BOUREILLE,

Merci pour toutes ces années d'études passées ensemble à l'ISPB et cette inoubliable année de folies à l'AAEPL. Merci d'avoir accepté de juger ce travail et d'être présent aujourd'hui.

À Camille,

Nous te remercions chaleureusement pour ton temps et tes relectures. Tes avis ont été précieux pour achever ce travail.

Remerciements personnels de Jean-Philippe Robert

Je tiens en tout premier lieu à renouveler, à titre personnel, mes remerciements pour le jury de cette thèse. Ce travail que vous jugez m'est important en termes de signification. C'est la fin d'une longue formation qui me permettra d'exercer les préceptes de mon art en toute conscience et indépendance. J'ai choisi ce métier, j'aime la science et l'Homme. Merci d'avoir contribué à devenir le professionnel de santé que je suis aujourd'hui. C'est un véritable honneur que de pouvoir présenter ce travail intellectuellement complexe qui est surtout, pour moi, l'occasion de prononcer le serment de Galien. Ce serment m'est cher tant la puissance de signification de chacun de ses mots résonne en moi. La complexité de ce travail est pour moi un jeu, toutefois, ma place est dorénavant au côté de mes patients à l'officine. Ce dernier point, je l'ai compris bien tard et j'espère que la profession continuera à faire le maximum pour défendre et promouvoir cette voie.

Je dédie tout particulièrement ce travail à mes parents. Un couple extraordinaire qui a su m'élever dans un cadre prospère, auréolé de valeurs familiales, morales et éthiques fortes qui font de moi ce que je suis aujourd'hui. La vie n'est simple pour personne, elle ne l'a pas été pour moi, malgré la chance inouïe que j'ai eu de pouvoir grandir en France. Je vous remercie infiniment de m'avoir tenu, soutenu et guidé quand bien même mes meilleurs défauts auraient pu me conduire loin de là où je suis aujourd'hui. Ces traits qui m'ont parfois causé beaucoup de torts sont aujourd'hui ma force : mon intuition, ma curiosité, ma créativité, mon empathie, mon amour, mon excentricité, mon humour et ma spontanéité. Maman, Papa, merci pour tout. Katia, Christophe et vos familles respectives ; je vous remercie également à ce titre d'avoir eu votre contribution dans la construction de l'Homme et non l'homme que je suis.

Merci à toi Valentin d'avoir été, depuis notre première année dans le supérieur, un ami et un collaborateur de génie. Tu as des qualités exceptionnelles, qui m'ont permis d'apprendre beaucoup de choses sur moi. Nous avons tous deux des ambitions personnelles et professionnelles partagées ; j'espère que nous continuerons à les entretenir. Merci à mes amis qui ont également contribués à ce que je suis : Nicolas, Jonathan, Louis, Cédric et Florent. Merci à mes camarades de la faculté devenus confrères, à mon ancien bureau de l'ANEPF, à tous les associatifs et étudiants que j'ai eu l'occasion de rencontrer et que je revois encore aujourd'hui. Léa, pour des raisons qui sont miennes je tiens à te remercier de m'avoir fait grandir.

Remerciements personnels de Valentin Roche

À mes parents,

Merci d'avoir tout mis en œuvre pour me permettre aujourd'hui d'exercer un métier qui me passionne et dans lequel je suis épanoui. Rien n'aurait été pareil si vous ne m'aviez pas transmis toutes ces valeurs et cette éducation. Merci de continuer à me soutenir et à croire en moi tous les jours.

À ma sœur,

Merci d'avoir toujours été autant attentionnée avec moi. Le premier confinement nous a permis de grandement nous rapprocher et j'en suis très heureux. Ta persévérance m'impressionne et je suis persuadé que tu réussiras ce dont tu rêves depuis si longtemps.

T'es la meilleure kiki !!

À mes grands-parents, tantes, oncles, cousins et cousines,

Merci pour tous ces moments vécus ensemble. Le Pharmacien que je suis aujourd'hui, c'est aussi grâce à vous.

PS : Merci Adri d'autant me distraire avec toutes ces péripéties (sentimentales ?).

À Jean-Philippe,

Depuis notre rencontre en PACES, au-delà d'être un ami, tu as été un vrai partenaire. Ces journées de travail riches en nouvelles idées et ces folles soirées sont et seront toujours un plaisir à partager. Je te souhaite que cette expérience dans le Sud t'apporte ce que tu recherches.

À Yohan,

Merci pour tous tes conseils et le temps passé à nous aider pendant le projet.

Aux Ceupeds,

Merci pour cette année de folie passée à l'amicale, toutes ces soirées et évènements organisés ! Elle restera ma meilleure expérience étudiante grâce à vous.

Aux teams Lyon Sud et Allez les bleus,

Je ne sais même pas quoi dire tellement vous comptez pour moi. Une seule chose suffira : Tié la famille !

Table des matières

Table des figures

Table des tableaux

Table des annexes

Abréviations

ACC : Autorisation d'Accès Compassionnel

AFMT : Association Française des Malades de la Thyroïde

AMA : Association Médicale Américaine

AMM : Autorisation de Mise sur le Marché

ANSM : Agence Nationale de Sécurité du médicament

API : Application Programming Interface

ASC : Aire Sous la Courbe

ATP : Adénosine Tri-Phosphate

AUC : Area Under Curve

BERT : Bidirectional Encoder Representations from Transformers

BCPNN : Bayesian Confidence Propagation Neural Network

BNPV : Base Nationale de PharmacoVigilance

CIP : Code Identifiant de Présentation

CMUH : Comité des Médicaments à Usage Humain

CRPV : Centre Régional de Pharmacovigilance

CSS : Cascading Style Sheets

CSV : Comma-Separated Values

CTPV : Comité Technique de PharmacoVigilance

DMP : Dossier Médical Partagé

ECG : Electrocardiogramme

EIM : Effets Indésirables des Médicament

EMA : European Medicines Agency

ENS : Espace Numérique de Santé

EPPH : Effet de Premier Passage Hépatique

FDA : Food and Drug Agency

HAS : Haute Autorité de Santé

HDH : Health Data Hub

HTML : HyperText Markup Language

IA : Intelligence Artificielle

IBE : Individual BioEquivalence

IER : Individual Exposition Ratio

IRM : Imagerie par Résonance Magnétique

LDA : Latent Dirichlet Allocation

LSA : Latent Semantic Analysis

MGPS : Multi-Item Gamma Poisson Shrinker

NLP : Natural Language Processing (Traitement Naturel du Langage en français)

NMF : Non-Negative Matrix Factoring

OMS : Organisation mondiale de la Santé

pH : Potentiel Hydrogène

PDF : Portable Document Format

PRR : Proportional Ratio Reporting

RCP : Résumé des Caractéristiques du Produit

RegEx : Regular Expression

RNN : Réseaux Neuronaux Récurrents

SNDS : Système National des Données de Santé

T3 : Liothyronine (3,5,3'-triiodothyronine ou 3,3',5'-triiodothyronine)

T4 : Thyroxine (3,5,3',5'-tétraiodothyronine)

TPO : Thyroperoxydase

TRH : Hormone thyréotrope

TSH : Thyréostimuline

UE : Union Européenne

UMLS : Unified Medical Language System

URL : Uniform Ressource Locator

Web : World Wide Wed

Introduction

L'analyse des données de vie réelle est devenue incontournable et représente un enjeu majeur pour améliorer la qualité des soins et la régulation du système de santé. C'est une réalité qui accompagne la transformation numérique de tout le secteur de la santé. L'intérêt de ces analyses est multiple et intervient à plusieurs étapes de la vie d'un médicament, au niveau des évaluations médico-économiques et de la phase IV dite « post-AMM » (Autorisation de Mise sur le Marché) de pharmacovigilance.

L'exploitation de ces données est rendue possible via le développement de nouveaux outils permettant de recueillir, d'analyser et de croiser une importante quantité de données. Ils permettent d'appliquer des traitements statistiques et algorithmiques (Big Data). Ces données sont désormais exploitables car, en plus des outils adéquats, il existe des enregistrements de celles-ci dans de nombreux systèmes informatisés, en particulier sur les réseaux sociaux qui constituent une source unique d'informations.

Ces données de vie réelle ne sont pas issues des cadres expérimentaux classiques (essais randomisés contrôlés, en double aveugle contre référence ou placebo). De ce fait, elles sont générées par l'environnement quotidien du patient. Les signaux sont ainsi détectés et parfois rapportés par les équipes de soins primaires (point d'entrée principal dans le système de pharmacovigilance). Toutefois, il est fréquent que le patient n'exprime pas aux équipes de soins ses ressentis (effet dit de blouse blanche). Ce phénomène s'atténue une fois qu'il a regagné sa sphère de confort. Ces mêmes données sont ainsi beaucoup plus fréquemment et spontanément échangées avec la famille, les amis et les communautés d'intérêts par différents canaux (à l'oral, par messages ou via les réseaux sociaux). Parallèlement, il y a presque 40 millions de français (1) actifs sur les réseaux sociaux. Ce canal favorise les échanges patients et les retours d'expériences. C'est sans surprise que les réseaux sociaux généralistes (Facebook®, Twitter®, Youtube®, Instagram®, etc.) ou spécialisés (Doctissimo®, sites d'associations de patients, etc.) recueillent aujourd'hui des quantités importantes d'informations dont l'étude est devenue cruciale pour la recherche biomédicale et les systèmes de santé. Le Healthcare Data Institute, « think tank » ou groupe de penseurs dédié au Big Data dans le domaine de la santé a lancé une réflexion sur les données des patients générées au cours de l'année 2017. L'objet était d'analyser les enjeux et les perspectives d'usage des données générées sur les réseaux sociaux en santé.

L’objectif de ce travail est d’élaborer une méthode de détection précoce d’évènements indésirables déclarés par les patients sur les réseaux sociaux. L’étude se base sur l’affaire du Levothyrox® à partir des données du sous-forum « endocrinologie » du site Doctissimo®. Cette affaire a fortement été relayée dans les médias en août 2017 à la suite du changement, en mars de la même année, de la formulation galénique de la spécialité (modification du type et de la teneur des excipients). De nombreux patients ont rencontré des effets indésirables liés à une fluctuation du fonctionnement thyroïdien alors qu’ils étaient déjà traités et stabilisés par le Levothyrox® (3 millions de patients atteints de pathologies thyroïdiennes). L’idée directrice est de comprendre dans quelle mesure le résultat du traitement des données échangées sur le forum Doctissimo® peut être utilisé en pharmacovigilance et si ces données aurait pu permettre une meilleure réactivité de la part des laboratoires Merck Santé et des pouvoirs publics.

Pour répondre à cet objectif, il s’agit dans un premier temps d’extraire les données du forum puis de les nettoyer. Dans un second temps, il s’agit de les analyser à l’aide de différents algorithmes utilisant les techniques actuelles de *data science* et de *machine learning* (algorithmes auto-apprenant, dits « d’intelligence artificielle »), afin de détecter si les problèmes relayés médiatiquement en août 2017 auraient pu être identifiés plus tôt.

1. L’affaire Levothyrox® en France

1.1 Physio-pathologie thyroïdienne et thérapeutique

1.1.1 Physiologie de la glande thyroïde

L’absence de glande thyroïde entraîne une diminution des activités physiques et psychiques, une diminution de la résistance au froid et, chez l'enfant, un retard de croissance caractéristique. A l’inverse, un excès d'activité thyroïdienne conduit à un épuisement de l'organisme avec des signes généraux tels que de la nervosité, des tremblements, des troubles cardiaques ou une production excessive de chaleur. Cette glande synthétise des hormones indispensables à la santé, intervenant sur de nombreuses fonctions physiologiques : croissance osseuse, développement mental, stimulation de la consommation d'oxygène des tissus, transformations des graisses et des sucres, etc.

Les hormones thyroïdiennes sont des molécules iodées, dérivées de la tyrosine (la thyronine est un assemblage de deux molécules de tyrosine), dont la biosynthèse est permise par l’apport exogène d’iode d’origine alimentaire :

- La thyroxine, ou 3,5,3',5'-tétraiodothyronine ou T4, active.
- La 3,5,3'-triiodothyronine, ou T3, active.
- La 3,3',5'-triiodothyronine, ou T3 inverse/reverse (T3r), inactive.
- On retrouve également les précurseurs de ces hormones (3-monoiodotyrosine, ou MIT, et la 3,5-diiodotyrosine, ou DIT).

La sécrétion de ces hormones est régulée par une hormone hypophysaire, la TSH (*Thyroid-Stimulating Hormone*), elle-même régulée par la TRH (*Thyreo-Realising Hormon*) hypothalamique.

Les formes libres sont les seules actives d'un point de vue physiologique ; capables de franchir les parois vasculaires et les membranes cellulaires. La demi-vie biologique est de sept jours pour la T4 et d'un jour pour la T3. La diffusion des deux hormones dans tous les tissus se fait librement et ne dépend que de la concentration des formes libres. Pour pallier d’éventuels déficits de production en hormones thyroïdiennes, la thyroïde a la capacité de constituer une réserve hormonale d’environ 100 jours.

La T4 agit en intracellulaire après transformation en T3. Lorsque la concentration de T4 diminue, l'axe hypothalamo-hypophysaire est stimulé et la sécrétion de TSH augmente. À l'inverse, lorsque la concentration de T4 augmente, l'axe hypothalamo-hypophysaire est freiné et la sécrétion de TSH diminue. La synthèse des hormones thyroïdiennes est donc autorégulée.

Les hormones thyroïdiennes ont de très nombreuses actions physiologiques. Elles agissent sur pratiquement tous les tissus, très souvent en synergie avec la noradrénaline et le glucagon. Elles n'interviennent pas dans la croissance fœtale pendant les dix premières semaines du développement. Ce n'est qu'à partir de la onzième semaine, qu'elles sont essentielles pour la différenciation et la maturation des tissus fœtaux, en particulier le squelette et surtout le cerveau. Elles jouent un rôle dans la myélinisation des axones et la prolifération axonale et dendritique. Les hormones thyroïdiennes de la mère ne peuvent pas combler un éventuel déficit, car elles ne passent pas la barrière placentaire.

À la naissance, la croissance dépend du fonctionnement normal de la thyroïde. Un déficit en hormones thyroïdiennes entraîne une diminution de croissance de la plupart des organes, en particulier de la croissance linéaire des os longs, conduisant à un nanisme disharmonieux et des altérations profondes du système nerveux central, aboutissant rapidement au crétinisme. Ces effets sont irréversibles, s'ils se manifestent au cours du développement fœtal et au début de la vie post-natale, sauf si l'hypothyroïdie est immédiatement corrigée. Ceci souligne l'importance du diagnostic systématique, à la naissance, d'un éventuel déficit, compte tenu de la gravité des lésions et du fait que le traitement sera d'autant plus efficace qu'il sera plus précoce.

1.1.2 L'hyperthyroïdie

Biologiquement, l'hyperthyroïdie se caractérise par une augmentation des hormones thyroïdiennes circulantes (T4 libre et/ou T3 libre). De ce fait, la TSH est diminuée via le rétrocontrôle négatif hypophysaire. L'augmentation du métabolisme peut s'observer par des modifications biologiques : hypocholestérolémie, neutropénie, hyperglycémie. Le diagnostic se fait par un dosage de la TSH en première intention, puis le dosage de la T4 libre et de la T3 libre permet d'affiner le diagnostic de l'hyperthyroïdie. Pour compléter ces examens biologiques, la scintigraphie à l'iode, l'échographie thyroïdienne ou les bilans immunologiques permettent de remonter à l'origine de l'hyperthyroïdie. D'installation rapide et parfois brutale, l'hyperactivité sécrétoire de la thyroïde constitue une maladie sérieuse au vu du risque de complications graves, au niveau cardiaque. La prise en charge repose sur des hormones thyroïdiennes substitutives et sur les antithyroïdiens de synthèse. Un suivi biologique et médical régulier est impératif. Il existe cinq éléments étiologiques principaux pour expliquer la survenue d'une hyperthyroïdie, de la plus à la moins fréquente :

- La maladie de Basedow (hyperthyroïdie auto-immune).
- Les nodules hypersécrétants.
- Les hyperthyroïdies iatrogènes (interférons, lithium, iode, hormones thyroïdiennes…).
- L'adénome hypophysaire à TSH.
- Les tumeurs thyroïdiennes ou les métastases thyroïdiennes.

Cliniquement, on peut retrouver :

- Des signes cardiovasculaires parfois graves (tachycardie sinusale, palpitations, dyspnée).
- Des signes neuropsychiatriques (tremblements, nervosité, agitation, asthénie, insomnie, trouble de l'humeur…).
- Des signes neuromusculaires tel que le signe du tabouret : difficulté à se relever de la position accroupie ou assise, le malade prend appui avec ses mains sur les genoux et « grimpe » le long de ses cuisses.
- Des troubles sexuels (aménorrhée, impuissance, gynécomastie).
- Des signes cutanés (phanères fins et cassants, peau moite, prurit).

Le traitement débute généralement par une bi-thérapie à base d'hormones thyroïdiennes (lévothyroxine par exemple) et d'anti-thyroïdiens de synthèse (carbimazole, thiamazole,

benzylthiouracile…). Des traitements symptomatiques aspécifiques existent, pour limiter les signes cliniques (bêtabloquants, benzodiazépines, ralentisseurs du péristaltisme intestinal et antispasmodiques). Selon la gravité des symptômes, il est également possible de détruire la thyroïde par radiothérapie (iode radioactif pendant 6 mois) ou par thyroïdectomie chirurgicale.

1.1.3 L’hypothyroïdie

L'hypothyroïdie est moins fréquente avec une prévalence de 1,9% pour les hommes et 3,3% chez les femmes au sein de la population française (2). C'est un syndrome caractérisé par une carence en hormones thyroïdiennes. A l'inverse de l'hyperthyroïdie, elle s'installe de façon lente et progressive. C’est une pathologie dans laquelle la production d’hormones thyroïdiennes est diminuée voire absente, ce qui perturbe le fonctionnement de la glande thyroïde provoquant des altérations du fonctionnement de l’organisme (asthénie, troubles digestifs, cardiovasculaires, nerveux, etc.). Le traitement repose alors presque exclusivement sur la supplémentation en hormones thyroïdiennes. Ces hypothyroïdies peuvent avoir plusieurs étiologies :

- Acquises centrales ou périphériques (iatrogénie, cancers, etc.).
- Auto-immunes (thyroïdite d’Hashimoto).
- Congénitales centrales, périphériques ou transitoires.

Seul le bilan sanguin permet d’affiner le diagnostic tant les symptômes sont peu spécifiques. Ce bilan se base sur le dosage de la TSH (et de la TRH dans certains cas), des hormones T4 libres et T3 libres. On couple également ce bilan par des tests immunologiques (anti-TPO, antithyroglobuline) et radiographiques (échographie et scintigraphie comme pour l’hyperthyroïdie). La valeur de la TSH permet d’identifier l’origine de la pathologie :

- TSH faible : atteinte de l’axe hypothalamo-hypophysaire, le rétrocontrôle n’existe plus, nécessite un dosage de la TRH.
- TSH normale : exclut une atteinte directe de la thyroïde (ex : thyroïdite d’Hashimoto, présence d'anticorps anti-TPO).
- TSH élevée : origine primaire (atteinte thyroïdienne), inhibition du rétrocontrôle négatif par manque d’hormones thyroïdiennes.

En parallèle, l’hypothyroïdie entraîne également d’autres perturbations biologiques (hypercholestérolémie, anémie macrocytaire, hyponatrémie, augmentation des transaminases au niveau musculaire). Sur le plan clinique, on observe alors des conséquences sur l’ensemble de l’organisme, à l’instar de l’hyperthyroïdie :

- Signes nerveux : asthénie chronique, irritabilité, hypersensibilité, dépression, amnésie, ralentissement cognitif, hypothermie, etc.

- Signes uro-digestifs : calculs, constipation, ballonnements, etc.
- Signes cardiovasculaires : hypertension, signes d'angor, bradycardie, dyspnée d'effort.
- Signes cutanéo-muqueux et musculo-squelettiques : peau sèche et froide, phanères secs et cassants, épaississement des muqueuses, myasthénie, raideurs, crampes, arthralgies, etc.

Le principe de la prise en charge est la normalisation de la TSH en compensant le déficit en hormones thyroïdiennes. La supplémentation se fait à l'aide de 3,5,3',5'tétraiodo-L-thyronine encore appelée lévothyroxine (substances active contenue dans le Levothyrox®), elle se fait par palier successifs jusqu'à atteindre la dose nécessaire à l'euthyroïdie stabilisée (de 25 μg à 200 μg par jour).

1.1.4 La place de la lévothyroxine sodique dans l'arsenal thérapeutique

La supplémentation en hormones thyroïdiennes concerne uniquement deux molécules endogènes : la thyroxine (T4), précurseur de la liothyronine (triiodothyronine ou T3) qui sont actives sous forme libre. Ces hormones ont des effets physiologiques sur tout l'organisme et leur équilibre est très fragile. De fortes variations peuvent découler de la rupture de cet équilibre dû à une très faible part de la fraction libre de ces hormones thyroïdiennes dans la circulation sanguine.

Il n'existe aucun traitement permettant de restaurer in situ la synthèse endogène normale de ces médiateurs, en cas de défaillance centrale ou périphérique des fonctions thyroïdiennes. Seule la substitution par des hormones de synthèse, dont le mécanisme d'action est strictement identique, permet de restaurer une concentration plasmatique et une physiologie normale. Il s'agit de mettre en place un traitement visant l'euthyroïdie, pris à vie par le patient. Il existe plusieurs stratégies thérapeutiques dont la bi-thérapie (T4-T3) en cas de non-réponse. Les médicaments contiennent uniquement les formes lévogyres de ces hormones, seuls énantiomères réellement actifs (L-thyroxine plutôt que la D-thyroxine, dextrogyre et peu active). Ils sont administrés par voie orale, la biodisponibilité est très bonne par cette voie (80 à 95%) lorsque la prise se fait à jeun. La voie parentérale (intravasculaire IV ou intramusculaire IM) est réservée aux hypothyroïdies secondaires (post-partum, chirurgie ou tumeur) et notamment en cas de coma myxœdémateux (rare urgence médicale de ces pathologies ; c'est une complication de l'hypothyroïdie) ; c'est un coma calme, hypothermique avec bradycardie. Il peut être déclenché par une infection, un traumatisme, une intervention chirurgicale ou par l'arrêt du traitement. Le pronostic, sombre, est dominé par plusieurs manifestations respiratoires : bradypnée, encombrement bronchique, épanchements pleuraux. Ces hormones de substitution (en particulier la lévothyroxine) sont indiquées dans toutes circonstances où l'on souhaite freiner l'action de la TSH, principalement dans des cas d'hypothyroïdie ; l'objectif est de restaurer un rétrocontrôle négatif normal pour atteindre l'objectif thérapeutique visant l'euthyroïdie. La liothyronine (T3) et le tiratricol sont utilisés en deuxième intention lorsqu'il existe des résistances périphériques aux hormones thyroïdiennes ; lorsque le patient ne répond pas au traitement par lévothyroxine seule. C'est le cas pour certains cancers TSH-dépendants, certains goitres simples et certains nodules, ainsi que dans

le traitement substitutif des hypothyroïdies dans le cas où un effet rapide ou transitoire est souhaité (effet freinateur insuffisant de la TSH).

Sans entrer de manière exhaustive dans les différentes interactions possibles avec ce traitement substitutif, il est à noter que de nombreux facteurs influencent l'efficacité de la supplémentation en T4 et/ou T3. Il s'agit d'être prudent en cas de co-administration d'anti-vitamines K (risque hémorragique), des antidiabétiques oraux (effets antagonistes sur la glycémie), de médicaments ayant des propriétés d'induction enzymatique ou ayant une forte liaison aux protéines plasmatiques (compétition de fixation : possible augmentation de la fraction libre et donc de l'effet des hormones thyroïdiennes). De plus, les cations métalliques (aluminium, fer, etc.), les résines échangeuses d'ion (cholestyramine), les oestro-mimétiques (contraception, soja...) et certains médicaments (sertraline, chloroquine, proguanil, etc.) peuvent en diminuer l'activité. De façon spécifique, l'amiodarone inhibe la transformation périphérique de T4 en T3 du fait de sa teneur élevée en iode. Elle peut alors déclencher une hyper- ou une hypothyroïdie. D'autres composés ont la capacité d'inhiber la biotransformation périphérique, on retrouve le propylthiouracile, les glucocorticoïdes, les bêtabloquants ou les médicaments iodés (produits de contrastes, antiseptiques, etc.).

Les hormones thyroïdiennes passent difficilement la barrière placentaire, ce qui autorise le traitement sans suivi particulier au cours de la grossesse, hormis des dosages trimestriels pour s'assurer de l'efficacité. A contrario, une proportion significative est éliminée dans le lait maternel : le traitement reste possible mais sous surveillance, le tiratricol est à proscrire. L'administration doit se faire si possible à jeun ou à distance des repas. Il est à noter que ces traitements substitutifs sont sensibles à la chaleur et à la lumière. La conservation doit se faire dans l'emballage d'origine et dans un endroit sec et frais (8°C à 15°C), ce qui n'est que rarement pris en compte ; que cela soit chez le patient, à l'hôpital, durant le transport, le stockage chez le grossiste ou le pharmacien. L'ensemble de ces facteurs jouent un rôle majeur dans le maintien de l'euthyroïdie et expliquent de nombreux cas d'échec de traitement par survenue d'effets secondaires ou par impossibilité d'adapter correctement les posologies. Les disparités inter-individuelles se retrouvent principalement lors de l'instauration du traitement ou des modifications des doses (point important à considérer pour expliquer une partie du problème en ce qui concerne l'affaire du Levothyrox®). Le tableau suivant reprend les principaux symptômes relatifs à un sur- ou à un sous-dosage :

Tableau 1 : Signes principaux d'un dosage trop faible ou trop élevé en hormones thyroïdiennes

Effets d'un sous-dosage	***Effets d'un surdosage***
Asthénie, tremblements, intolérance au froid	Sueurs, thermophobie, fébricule
Faciès lunaire, dépilation (sourcils), infiltration cutanée (myxoedèmes)	/
Douleurs et raideurs musculaires	Myasthénie, hypotrophie musculaire
Prise de poids sans augmentation de l'appétit	Amaigrissement sans perte d'appétit, augmentation du catabolisme cellulaire
Angine de poitrine, bradycardie, essoufflement	Tachycardie, dyspnée d'effort, arythmie, hypertension artérielle, insuffisance cardiaque
Constipation	Diarrhées
Nervosité, dépression, majoration des tableaux de démence chez le sujet âgé	Nervosité, irritabilité, insomnies, tremblements fins des extrémités
Voix rauque (épaississement muqueux)	/

La lévothyroxine fait partie d'un groupe générique depuis 2010. Compte-tenu de la spécificité du mode d'action et la variabilité inter-individuelle, l'intervalle de bioéquivalence entre le Levothyrox® et ses spécialités génériques a été resserré à [90% ; 111%] pour l'aire sous la courbe des concentrations plasmatiques mesurées entre 0 et 48 heures après la prise. Sa substitution nécessite de remplir les mêmes conditions que pour les autres molécules tombées dans le domaine public. Néanmoins, l'adaptation posologique très fine et la marge thérapeutique étroite nécessitent la prise de précautions particulières. Chez de nombreux patients, une faible variation de lévothyroxine peut perturber l'équilibre thérapeutique. Il semble ainsi de mise d'initier un traitement avec une seule spécialité puis de réaliser l'ensemble des adaptations nécessaires (en général, pendant 3 à 4 semaines) avec cette même spécialité. Il est déconseillé de substituer le traitement ainsi mis en place par un équivalent d'une autre marque et en cas de rupture, les surveillances cliniques et biologiques doivent être renforcées.

1.2 Changement de formulation du Levothyrox®

1.2.1 Analyse du dossier

Avec 3 millions de personnes atteintes en France, le Levothyrox® fait partie des trois médicaments les plus prescrits. À la suite de la décision de l'Agence Nationale de Sécurité du Médicament (ANSM) en 2011, la formule du Levothyrox® a été modifiée par le laboratoire allemand Merck® dans le but d'améliorer la stabilité et la conservation pour la forme solide (comprimé sécable). Une lettre d'information de Merck Santé (Annexe 2) à destination des professionnels de santé a été diffusée dès le 27 février 2017, celle-ci rappelle l'importance d'avoir un suivi régulier du fait de la marge thérapeutique étroite de la lévothyroxine. Le principe actif reste le même, il s'agit d'une modification de la formule au niveau des excipients. Le lactose est remplacé par le mannitol et par l'ajout d'acide citrique (rôle de conservateur) (3). Cette nouvelle formulation, censée être plus stable et répondant à un besoin de maintenir les concentrations sanguines dans l'objectif thérapeutique, sera commercialisée et distribuée dès mars 2017. Une note présentée sous la forme Question/Réponse (4) a été rendue publique au même moment par l'ANSM, dans le but de rassurer professionnels et patients : « Levothyrox® : changement de formule et de couleurs des boîtes et blisters » (le détail en Annexe 1). Aucun élément susceptible d'indiquer la potentielle survenue d'effets indésirables ne transparaît. Le changement de formule est présenté comme étant parfaitement anodin sur la conduite normale du traitement, comme on peut le voir dans le passage et l'illustration présentés ci-dessous :

« 4. Quels sont les risques liés au changement de formule ?

Aucun changement du profil de tolérance n'est attendu, le principe actif restant de la lévothyroxine sodique de même source. La bioéquivalence entre l'ancienne et la nouvelle formule a été démontrée. Seuls les excipients ont été modifiés. La bioéquivalence entre l'ancienne et la nouvelle formule a été démontrée par des études de biodisponibilité. Il a ainsi été mis en évidence que les nouveaux excipients ne modifient ni la quantité de substance active qui passe dans le sang, ni la vitesse à laquelle elle atteint l'organe cible. Cette bioéquivalence est la garantie d'une efficacité et d'une sécurité identique à celle de l'ancienne formule.

Par mesure de précaution, si vous pensez avoir des symptômes traduisant un déséquilibre thyroïdien (question 5 : « Quels sont les symptômes qui doivent m'alerter sur un déséquilibre

thyroïdien ? ») nous vous recommandons de contacter votre médecin pour contrôler votre TSH. »

Figure 1 : Infographie de Merck Santé diffusée au début de la distribution de la nouvelle formule du Levothyrox®

Pourtant, ce changement de formule a provoqué une augmentation inédite de la fréquence de signalement d'effets indésirables : ils étaient plus de 17 000 patients recensés entre juin 2017 et fin janvier 2018 d'après un rapport de pharmacovigilance remis à l'ANSM. A ce même moment, au début de l'année 2018, on peut lire dans un rapport de l'ANSM, que seulement vingt-trois cas validés d'hyperthyroïdie iatrogène entre 2009 et 2011 avaient justifié le retrait du marché de l'ancienne formule de Levothyrox®, selon la publication de l'ANSM.

Afin d'augmenter les capacités de traitement dans un délai restreint, des ressources supplémentaires ont été allouées aux Centres Régionaux de PharmacoVigilance (CRPV). Par précaution et en tenant compte du domaine thérapeutique concerné (quand bien même la bioéquivalence entre l'ancienne et la nouvelle formule soit démontrée, tout comme l'absence d'impureté ou de problème dans la fabrication), il a été recommandé à de nombreux patients de réaliser un dosage de la TSH plusieurs semaines après le début de la prise de la nouvelle formule du Levothyrox®.

Une enquête de pharmacovigilance a également été conduite sur la période allant de fin mars 2017 au 15 septembre 2017, lors de la commercialisation de la nouvelle formule, pour étudier les événements signalés. Les premiers résultats ont été présentés durant un Comité Technique de PharmacoVigilance (CTPV) à la date du 10 octobre 2017. Parmi les cas rapportés, 5 062 ont été identifiés comme prioritaire et enregistrés dans la Base Nationale de PharmacoVigilance (BNPV). Ils correspondent aux signalements entraînant des répercussions sur la vie familiale, professionnelle ou sociale, ainsi que les cas les plus documentés. Les troubles mentionnés le plus fréquemment sont l'asthénie, l'insomnie, les céphalées, les vertiges, les douleurs musculo-squelettiques (dont des arthralgies), la dégradation et/ou la chute des phanères dont les cheveux (effets secondaires déjà connus avec l'ancienne formule). La survenue de déséquilibres thyroïdiens pour certains patients lors du changement entre l'ancienne et la nouvelle formule du Levothyrox® a alors été confirmée (5).

Cette enquête préliminaire a permis de conclure que le profil clinique des effets indésirables rapportés avec la nouvelle formule était similaire au profil de tolérance connu avec l'ancienne formule. Il a été précisé que tout changement de spécialité ou de formule peut modifier l'équilibre hormonal et nécessiter un ajustement posologique, ce qui peut prendre un certain délai (compte-tenu des propriétés de ce traitement de la marge thérapeutique étroite pour arriver à l'euthyroïdie).

Une nouvelle enquête a été déclenchée sur la période du 15 septembre au 30 novembre 2017. Les résultats ont été soumis au CTPV du 30 janvier 2018, en présence de représentants d'associations de patients et de professionnels de santé. 12 248 nouveaux cas ont été enregistrés dans la BNPV, puis analysés sur cette période. Ces nouveaux signalements ont été très majoritairement déclarés par les patients eux-mêmes (90 %). Le pourcentage de patients signalant des effets indésirables est estimé à 0,75 % des patients traités avec le médicament mis en cause (contre 0,00007% avec l'ancienne formule). Les effets indésirables rapportés sont en tout point identiques à ceux identifiés depuis que l'on utilise ce traitement substitutif. Cependant ils ont été rapportés à une fréquence inédite et inattendue. Sur l'ensemble de ces signalements, une attention particulière a été portée sur 339 cas sélectionnés pour leurs critères de gravité (décès, hospitalisation, tératogenèse, mise en jeu du pronostic vital, handicap, etc.). 19 cas de décès ont ainsi été analysés : aucun lien n'a été établi avec la nouvelle formule. Un cas de suicide a par ailleurs conduit à une analyse approfondie sur 79 cas similaires (idées suicidaires). De la même façon, les données ne sont

pas suffisamment complètes pour permettre d'établir un lien de causalité sur la survenue de ces troubles psychiatriques et l'utilisation de la nouvelle formule.

Parmi les cas déclarés, environ 4 000 comportent des informations concernant les bilans sanguins et thyroïdiens. Seulement 1 745 sont suffisamment documentés et ont pu permettre une analyse détaillée confirmant la survenue possible (dans environ un tiers des cas) de déséquilibres thyroïdiens dus au changement de formule. Le profil d'effets indésirables est similaire chez tous les patients en hypothyroïdie ou en hyperthyroïdie. Pour le reste, les déclarations ont été faites alors que les dosages de TSH sont dans les normes attendues, ce qui révèle un potentiel effet nocebo dans le sens où il est tout à fait anormal de constater ce type d'effets indésirables avec des valeurs normales au niveau du bilan biologique. Finalement, l'analyse de l'ensemble des cas des différentes enquêtes diligentées ne permet pas la mise en évidence de nouveaux effets indésirables avec la nouvelle formule ni ne fournit des éléments explicatifs pour corroborer telle ou telle hypothèse.

Une étude pharmaco-épidémiologique a ensuite été demandée par l'ANSM, pour l'ensemble des patients traités. Le premier volet de cette étude a pour objectif de décrire les caractéristiques et l'état de santé des patients qui ont changé de formule entre les mois de mars 2017 et de juin 2017. Il en ressort que la population utilisant le Levothyrox® se caractérise comme tel (6) :

- 85% de femmes.
- Moyenne de 64 ans d'âge.
- Changement de formule réalisé en moyenne, en mai 2017.
- Pas de modification statistiquement significative du dosage moyen prescrit.
- Augmentation de la fréquence des dosages de la TSH en laboratoire d'analyse médical.

A ce moment, concernant l'offre thérapeutique, hormis le Levothyrox® présenté dans sa nouvelle formule, plusieurs médicaments à base de lévothyroxine sodique ont une AMM et sont disponibles en France (4) :

- La L-THYROXINE SERB, solution buvable en gouttes du laboratoire SERB.
- La spécialité générique THYROFIX, comprimé (quatre dosages) du laboratoire UNI-PHARMA.

- La spécialité TCAPS sous forme de capsule molle (douze dosages) des laboratoires GENEVRIER.
- La L-THYROXINE HENNING, comprimé du laboratoire SANOFI (à disposition depuis mi-octobre 2017 par le biais d'importations, s'est vu délivrer le 25 janvier 2018 des AMM en France pour différents dosages).
- Des stocks de produits strictement identiques à l'ancienne formulation du Levothyrox® (EUTHYROX) ont également été mis à disposition dès octobre 2017 par le biais d'importations. La prescription de l'Euthyrox® est destinée en dernier recours aux patients qui rencontrent des effets indésirables durables. À la demande des pouvoirs publics, Merck Santé a poursuivi les importations, jusqu'à nos jours où on peut encore retrouver ce médicament en pharmacie.

Au vu des données de l'Assurance Maladie, le nombre de patients traités par l'une des alternatives précitées a été estimé à environ 500 000 en 2017 (plus de 15% des patients).

Parallèlement, des voix se font entendre et ouvrent le débat médiatique avec de nombreux témoignages, des éléments et analyses différentes : professionnels de santé, associations de patients, patients eux-mêmes, s'approprient cette affaire et tentent d'enrichir les investigations (6). Figure emblématique de cette mobilisation, Sylvie Robache, patiente traitée par le Levothyrox® après un cancer de la thyroïde en 2016, a lancé une pétition en ligne fin juin 2017 afin de dénoncer la nouvelle formule du médicament, « trop de patients ne supportent pas le nouveau Levothyrox®, ils ressentent d'importants effets secondaires ». En cause, le changement de formule fin mars 2017. Au travers de cette mobilisation, elle souhaite s'adresser directement au laboratoire Merck® ainsi qu'à l'ANSM et demande que « les laboratoires concernés reviennent à l'ancienne formule ». Cette pétition a été signée par plus de 155 000 personnes fin août 2017 et compte aujourd'hui pas moins de 345 000 signatures (11% de la population traitée par lévothyroxine sodique) (7).

1.2.2 Les lacunes d'une industrie et des autorités mis en perspective de cette affaire

Le 14 septembre 2017, Mme Laurence Cohen (sénatrice communiste du Val-de-Marne) a interrogé Agnès Buzyn, ministre de la Santé, sur les effets indésirables de la nouvelle formule du Levothyrox®. Prescrit à près de 3 millions de français, la nouvelle formule, disponible depuis quelques mois, semble poser de nombreux effets secondaires (asthénie, malaises, troubles musculaires, troubles intestinaux, etc.).

Malgré que des vérifications supplémentaires soient nécessaires pour traiter l'ensemble des données recueillis par la pharmacovigilance (1 500 cas à ce moment), elle souligne que cette nouvelle formule pose question, sur le point de vue de la bioéquivalence. Elle reporte également que des patients en viennent à arrêter le traitement ou à se reporter sur une autre spécialité à base de lévothyroxine (ce qui risque d'entraîner une rupture de stock pour celles-ci). L'inquiétude et la mobilisation des patients ne cesse de grandir au fil des jours.

En attendant le traitement par l'ANSM des données recueillies, elle questionne sur la manière dont la ministre de la santé entend intervenir auprès de l'ANSM et du laboratoire concerné pour demander le retrait immédiat de la nouvelle formule au profit de l'ancienne, ou a minima, « laisser le choix aux patients entre ces deux versions, et ce, par souci de principe de précaution et de santé publique » (8).

Pour réponse à la question de Mme Cohen, Agnès Buzyn a répondu : « [...] Sans minimiser ni nier les symptômes ressentis par certains patients, ils sont invités à se tourner vers leur médecin traitant ou leur endocrinologue pour trouver ensemble le dosage adéquat de la nouvelle formule du Levothyrox®. Il faut garder à l'esprit que le seul danger pour ces patients est qu'ils arrêtent de prendre leur traitement. Le risque sanitaire pour les patients de la nouvelle formule est inchangé. L'ANSM a vérifié la conformité de la nouvelle formule et n'a relevé aucune impureté dans le Levothyrox®. Une enquête de pharmacovigilance supplémentaire est en cours et donnera ses résultats en octobre 2017. L'ANSM sera parfaitement transparente sur toutes les mesures de suivi et invitera les associations de patients pour leur présenter les résultats. En outre, la ministre des solidarités et de la santé reconnaît que cette spécialité bénéficie, en France, d'un quasi-monopole, qu'il convient d'ouvrir à d'autres médicaments. [...] » (11).

Le 25 juin 2017, la Cour d'appel de Lyon juge en faveur d'une condamnation de Merck Santé au titre de préjudice moral. Chaque plaignant devant être indemnisé d'un montant de 1 000 euros. L'initiateur de cette action collective, l'avocat Christophe Leguevaques, non content de cette décision, décide de s'attaquer à l'ANSM (la réunion d'information s'est tenue le 14 septembre 2017). Il considère l'ANSM responsable au même titre que le laboratoire. L'agence est accusée de manquements et d'insuffisances puisqu'elle n'aurait pas vérifié en détails les travaux produits par Merck Santé. Il reproche également la collusion entre l'agence et le laboratoire, au détriment du ressenti et de l'écoute des patients. Dernier argument mis en perspective : des changements de formulation de ce médicament ont déjà été opéré dans d'autres pays dans le passé et ont causés, à chaque fois, un problème de santé. Il considère donc « que l'ANSM a failli à sa mission de protection de la santé publique et, à ce titre, elle doit rendre des comptes et doit être déclarée responsable par le tribunal administratif » (9). En mai 2021, un rapport d'expertise demandé par la justice montrait effectivement que les formules n'étaient pas interchangeables (10).

En novembre 2017, l'Association Française des Malades de la Thyroïde (AFMT) a lancé une pétition, exigeant de la ministre de la santé de l'époque, Agnès Buzyn, « des réponses rapides et pérennes à la crise du Levothyrox® » (11). L'AFMT pointe du doigt le coût humain et financier qui sera imputable à cette crise sanitaire. Elle dénonce aussi le quasi-monopole du Levothyrox® de Merck Santé, malgré de nombreuses alternatives. D'un côté, l'Ordre des Médecins demande aux médecins de ne pas laisser le choix aux patients. De l'autre côté, en pharmacie, le stock des produits apparentés est très faible et il en va de même du côté de l'approvisionnement des grossistes. La culpabilisation des malades est dénoncée et malgré la reconnaissance du problème dès le mois de septembre, rien ne s'est passé en deux mois. Il est alors expressément demandé, au travers de cette pétition, « des réponses urgences et des actes conséquents sur ces faits » (11).

Demandé par le tribunal de Marseille, le rapport d'expertise mentionné précédemment alimente le doute sur la composition de la nouvelle formule. Le journal Le Monde s'est procuré le document et indique que « le changement de formule du médicament a été insuffisamment évalué » et que l'ANSM aurait été au courant du potentiel problème, dès le mois de septembre 2017 (12).

Le rapport d'expertise demandé s'inscrit dans le cadre d'une information judiciaire ouverte début mars 2018 pour « blessures involontaires et mise en danger d'autrui ». Il a ensuite été

élargie un an plus tard au chef d' « homicide involontaire ». Les experts listent une série de failles réglementaires et de manquements du laboratoire et des autorités sanitaires, ce qui explique certains des effets indésirables liés au changement de formule, rapporte Le Monde. Les auteurs du rapport estiment plus que probable l'existence de différences entre les deux formules du médicament. Ils évoquent la vitesse de dissolution du principe actif et le changement d'excipient : remplacement du lactose par un mélange de mannitol et d'acide citrique.

Dans les documents saisis par la justice, une note a été retrouvée, dans laquelle le référent en pharmacocinétique de l'ANSM reconnaît la non-substituabilité du groupe générique de la lévothyroxine. Les auteurs du rapport considèrent donc que « le fait que le médicament soit substituable a été imposé ». L'AFMT, à l'origine de la plainte devant le tribunal de grande instance de Marseille, s'est réjouie des conclusions de ce rapport : « Les résultats de l'expertise pénale prouvent que ce n'était pas un effet nocebo » (13).

1.3 L'équivalence thérapeutique

1.3.1 Notions de pharmacocinétique : biodisponibilité et bioéquivalence

Pour bien comprendre les enjeux liés à l'affaire du Levothyrox®, il est important d'introduire les notions de pharmacocinétiques fondamentales utilisées dans les études qui permettent aux industriels et aux autorités de santé de mettre à disposition de la population les médicaments. En effet, dans ce cas précis, il s'agit d'un changement de formulation pour limiter les variations constatées de teneur en principe actif dans l'ancienne formule du Levothyrox®. La dose de principe actif étant strictement identique entre les anciennes et nouvelles formules, les études réalisées visaient à démontrer la bioéquivalence stricte pour s'assurer de l'innocuité et de l'absence de modifications de l'état clinique (à l'instar des procédures permettant d'autoriser un médicament générique).

La pharmacocinétique est définie comme étant le devenir biologique d'une substance active dans l'organisme dans lequel elle est administrée. Deux aspects fondamentaux coexistent : l'aspect qualitatif (processus physiologiques impliqués dans le devenir des médicaments) et l'aspect quantitatif (relation existante entre la dose administrée et la concentration sanguine en résultant, obtenue via des paramètres pharmacocinétiques calculés et caractéristiques de chaque substance pour un patient donné). L'aspect qualitatif de la pharmacocinétique distingue les phases d'entrée (absorption, distribution) et de sortie permettant ainsi l'élimination (métabolisation, excrétion). L'un des paramètres pharmacocinétique majeur est la biodisponibilité qui se définit comme étant la fraction active et donc inchangée d'un médicament qui atteint la circulation sanguine (circulation systémique).

Lors de l'usage d'un médicament, il est souhaité que le principe actif puisse être absorbé et acheminé jusqu'à son site d'action pour exercer correctement ses propriétés pharmacologiques. En effet, pour avoir un effet thérapeutique, il ne suffit pas que le médicament entre dans l'organisme. Il est nécessaire que la dose soit disponible au niveau de la zone cible à traiter. De plus, la substance active doit atteindre la zone ciblée en un temps spécifique et y rester pendant une période définie.

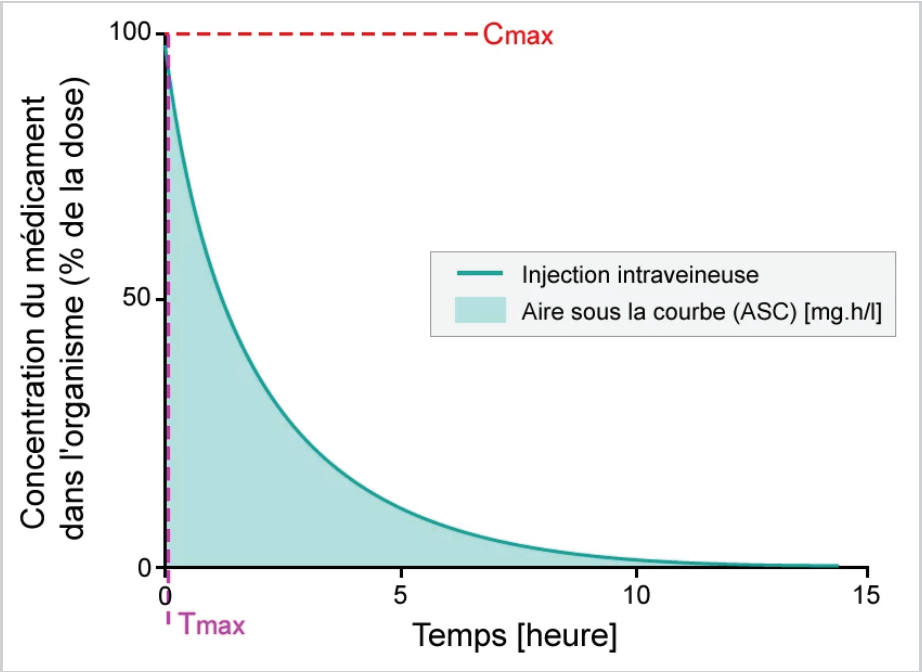

Figure 2 : Pourcentage de principe actif dans le corps ou biodisponibilité après injection directe dans la circulation sanguine, étudié sur une période de 15 heures

Lors d'une injection d'une substance active, celle-ci atteint sa cible via la circulation sanguine, après un trajet complexe et de multiples étapes de métabolisation (notamment hépatique via le passage dans la veine porte). Immédiatement après l'injection, la biodisponibilité est de 100% (administration sanguine directe). C'est ce qui est remarqué sur l'axe des ordonnées de la figure ci-dessus. Une fraction de la dose administrée est ensuite métabolisée et excrétée (rôle majeur du foie et des reins). C'est ce qui explique la diminution progressive de la concentration sanguine en principe actif au fil du temps. La diminution de la concentration n'est pas la même entre tous les médicaments, la notion de demi-vie est alors capitale. Certains médicaments sont lentement métabolisés et agirons, de fait, très longtemps, alors que d'autres devront être administrés très régulièrement pour maintenir l'effet thérapeutique. C'est l'étude de l'aire sous la courbe (ASC ou AUC) qui permet d'évaluer le profil de biodisponibilité d'une substance active et de le comparer à d'autres médicaments. L'ASC représente l'exposition totale à une substance active reçue par l'organisme. Le temps nécessaire à l'obtention de la plus forte concentration de substance active dans le sang est défini comme étant le Tmax (présent sur l'axe des abscisses dans la figure ci-dessus). La concentration sanguine maximale observée à Tmax est définie comme étant le Cmax.

Seule l'administration dans la circulation générale permet d'obtenir une biodisponibilité de 100%. La voie per os (ingestion orale) a donc forcément une biodisponibilité inférieure à 100% (voir figure ci-dessous, où l'on observe une biodisponibilité maximale de 55%).

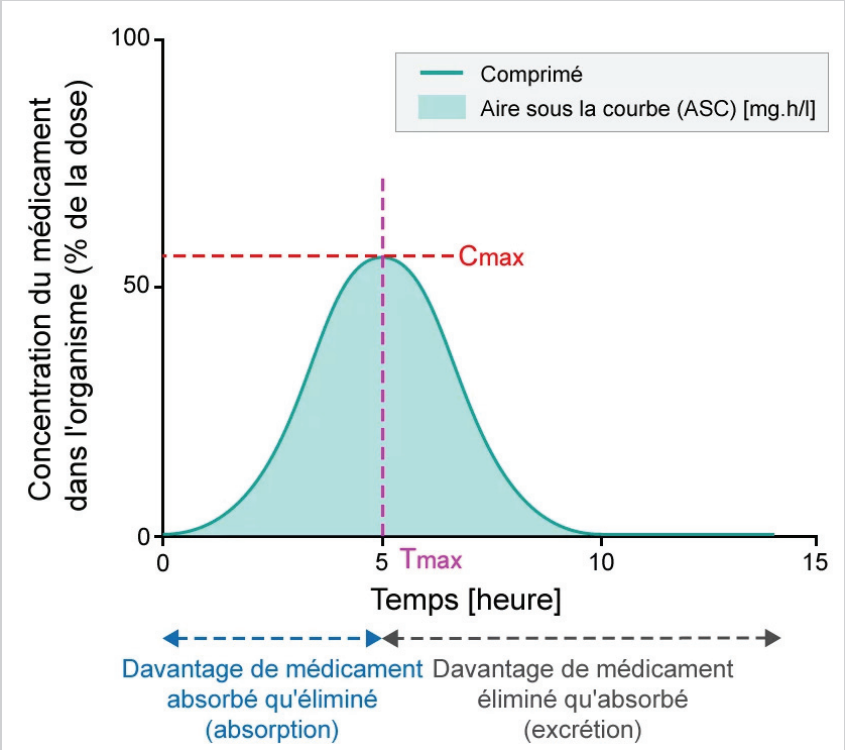

Figure 3 : Pourcentage de principe actif après la prise d'un comprimé, étudié sur une période de 15 heures

La biodisponibilité plus faible de la voie orale par rapport à la voie intraveineuse s'explique par de nombreux mécanismes qui se retrouvent à toutes les étapes de la vie du médicament après l'ingestion. Les enzymes salivaires (amylases principalement) peuvent dégrader un certain nombre de substances, ensuite interviennent les estérases gastriques. Le passage dans l'intestin réduit la quantité de substance active disponible via trois mécanismes :

- Le premier est l'absorption intestinale. Elle est permise par la présence de microvillosités qui composent la muqueuse de l'intestin. Elles permettent le passage des nutriments et des médicaments dans la circulation sanguine. Cette capacité d'absorption dépend du médicament, du bol alimentaire en présence et de l'état des parois (muqueuse intestinale).
- Le deuxième mécanisme est l'intervention enzymatique. Elle permet l'assimilation normale des nutriments : amylases, lipases, estérases, trypsine, chymotrypsine, carboxypeptidase, ribonucléases et désoxyribonucléases. Elles ont la capacité

d'influer sur de très nombreuses substances actives, cela va de la biotransformation à la dégradation complète.

- Le troisième mécanisme est l'effet de premier passage hépatique (EPPH). Tout ce qui est absorbé via l'intestin passe dans la circulation sanguine et transite directement dans le système porte pour subir une première biotransformation avant d'être pompé par le cœur dans la veine cave et d'atteindre la circulation artérielle.

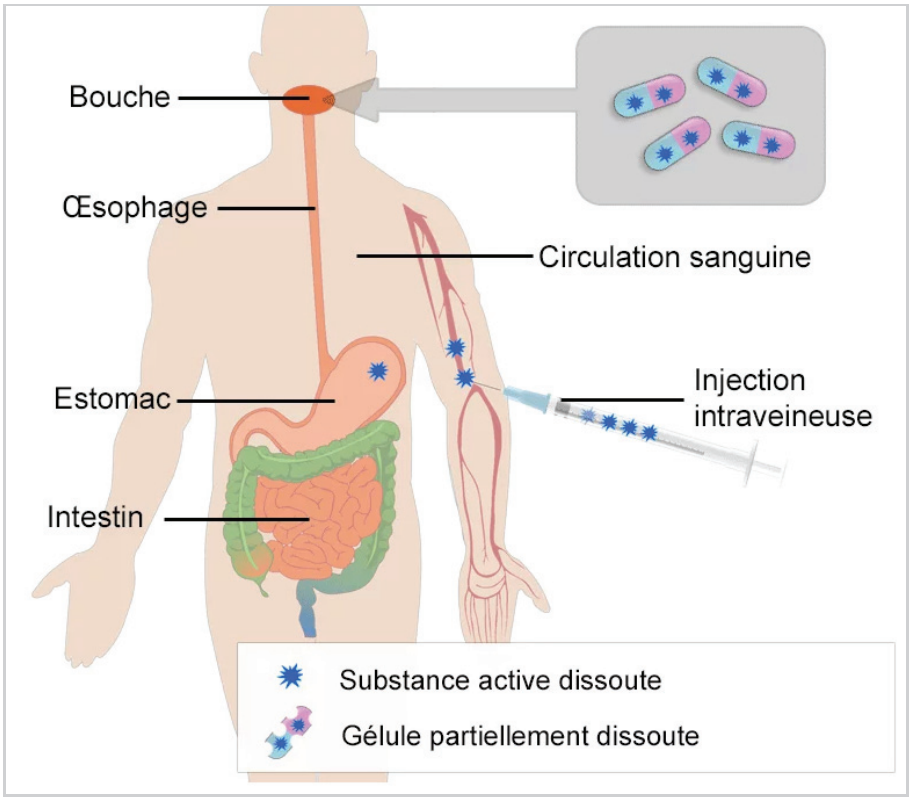

Figure 4 : Mise en parallèle de deux voix d'administration : per-os et parentérale (injection intraveineuse)

Concernant les médicaments à marge thérapeutique étroite (très faible dosage, dose thérapeutique proche de la dose toxique, etc…), cela explique les précautions et le suivi biologique strict mis en place ; c'est le cas de la lévothyroxine sodique, principe actif du Levothyrox®.

La biodisponibilité est donc affectée par de nombreux facteurs tels que la formulation du médicament, la voie d'administration, les aspects interindividuels et génétiques, la coprésence de substances issues de l'alimentation ou d'autres médicament, la présence de pathologie gastro-intestinales (principalement des inflammations), etc… Deux principes actifs sont dits bioéquivalents lorsque, administrés à la même concentration, ils engendrent les mêmes effets

pharmacologiques et thérapeutiques. Par extension, il s'agit de la relation entre deux formulations différentes du même médicament, dans la même présentation et avec un profil de biodisponibilité équivalent. La biodisponibilité peut donc permettre la comparaison de médicaments différents contenant la même substance active, fabriqués par des laboratoires pharmaceutiques différents.

Cependant, les études de bioéquivalence ne sont normalement menées qu'auprès de volontaires en bonne santé afin de réduire la variabilité non attribuable aux différences entre les produits. On pourrait donc s'interroger si le médicament générique produisait des effets différents dans la population de patients ciblée, compte tenu de facteurs comme les comorbidités, les autres thérapeutiques utilisées, de facteurs physiologiques comme les variabilités métaboliques, le pH gastrique, la flore bactérienne… Les études de bioéquivalence portent en général sur des doses uniques d'un médicament. Il est en théorie possible que les excipients que contient une préparation générique modifient l'absorption et le métabolisme à l'état d'équilibre, mais pas après une dose unique. Ces différences sont toutefois très peu probables et sont normalement mises en évidence au cours de l'étude de bioéquivalence. Toute différence pouvant exister est négligeable comparativement à la variabilité des conditions dans les voies gastro-intestinales et de son effet sur l'absorption (14).

1.3.2 Bioéquivalent synonyme d'interchangeable ?

Il est important de reconnaître que l'essai de bioéquivalence, mené sur des volontaires sains conformément aux directives de l'European Medicines Agency (EMA) et de la Food and Drug Agency (FDA) (15)(16)(17), ne garantit pas que chaque patient de la population cible, qui passe d'une ancienne formulation de référence à une nouvelle formulation, sera exposé de manière similaire à la lévothyroxine. En 2010, un changement de ligne directrice caractérisant la bioéquivalence a été admis. Le nouvel objectif est de s'assurer que l'impact des variations de formulation sont détectées via les paramètres pharmacocinétiques tels que l'AUC et la Cmax. Avant 2010, les observations cliniques entraient aussi en considération dans le dossier de bioéquivalence. En s'affranchissant des observations cliniques, cette ligne directrice laisse place à des variabilités interindividuelles plus importantes, surtout pour des médicaments à marge thérapeutique étroite et qui se conserve mal à des températures supérieures à 20°C comme le Levothyrox® (18)(19). Cette nouvelle position de l'union européenne (UE) est juridiquement plus défendable que les orientations précédentes, mais elle considère implicitement que les sujets sains homogènes impliqués dans ce type d'essais sont moins représentatifs d'une population de patients cible hétérogène.

La possibilité de changer de formulation pour soutenir la substitution d'un produit par un autre n'est pas traité par les lignes directrices de l'EMA ou de la FDA. La politique de substitution est une question nationale et non réglementée par l'UE (15). La démonstration de la bioéquivalence via les études ad-hoc est utilisée dans l'approbation préalable à la commercialisation de nouvelles formulations génériques. Cependant le nouveau Levothyrox® n'est pas une nouvelle formulation générique proposée comme alternative possible à l'ancien Levothyrox® (aujourd'hui Euthyrox®). Il s'agit d'une nouvelle formulation destinée à le remplacer. Le nombre de patients pour lesquels cette évolution a été imposée en France entre mars et juin 2017 est estimé à 2 188 432 (20). Par conséquent, la question clé qui aurait dû être abordée avant la commercialisation de l'Euthyrox® est : un patient déjà traité par le Levothyrox® peut-il passer en toute sécurité et efficacement de cette formulation (qui ne sera plus disponible) à la nouvelle formulation, pour un dosage donné ? Une étude démontrant la bioéquivalence ne répond pas à cette question, c'est-à-dire que la démonstration de ce type d'étude, entre l'Euthyrox® et le Levothyrox® ne garantit pas leur interchangeabilité.

Le concept sous-jacent de l'interchangeabilité est que chaque patient a sa propre fenêtre thérapeutique optimale, c'est-à-dire une gamme de concentrations plasmatiques offrant une

efficacité et une sécurité optimales. Si un changement de formulation est effectué, celle-ci doit garantir un profil d'exposition au médicament similaire et donc situé dans la fenêtre thérapeutique du patient en question. C'est ce qui peut garantir le même niveau de sécurité et d'efficacité (21). De façon à mieux apprécier l'interchangeabilité de deux formulations, le concept de bioéquivalence individuelle (IBE) a été introduit il y a plus de 25 ans (22), de façon à dépasser les limites des essais de bioéquivalences classiques. Une étude IBE compare l'exposition obtenue avec chaque formulation, pour chaque sujet, de manière individuelle. Cela garantit que chaque individu réponde de manière similaire aux deux formulations. L'étude IBE nécessite d'estimer l'écart de biodisponibilité entre les deux formulations en établissant non seulement des moyennes au sein d'une population mais aussi en utilisant la variance intra-individuelle et la variance d'interaction individuelle par formulation. Ce terme d'interaction est la mesure dans laquelle les différences individuelles entre les deux formulations sont similaires d'un sujet à l'autre. La bioéquivalence individuelle a été à la fois largement discutée et contestée puis, finalement, non adoptée par les autorités réglementaires. Cependant, conclure que le concept d'IBE n'est pas cliniquement pertinent n'est pas acceptable. Les études IBE nécessitent des protocoles bien plus compliqués et coûteux que les études classiques de bioéquivalence et elles sont de plus associées à plusieurs problèmes réglementaires (conception différente de l'étude, inclusion de sujets malades nécessaires alors que la bioéquivalence est normalement définie sur sujets sains, etc…) (17). Il est probable que la raison de la non-adoption se situe ici, même si le consensus actuel considère qu'il n'y a aucune preuve d'échec des études de bioéquivalence classiques pour les génériques mis sur le marché.

Toutefois, un avis du groupe de travail sur la bioéquivalence des populations individuelles de la FDA (21) a été donné : l'interaction individuelle par formulation est le terme de variance le plus pertinent à explorer pour assurer l'interchangeabilité. Dans cet avis, il est proposé que la bioéquivalence entre les deux formulations de Levothyrox® ainsi que l'interaction individuelle par formulation sont indispensables pour évaluer si l'interchangeabilité est réellement établie. Pour le Levothyrox®, plus de 50 % des sujets inclus dans un grand essai de bioéquivalence ayant un protocole conforme aux exigences de l'EMA étaient en fait en dehors de la plage de bioéquivalence définie à priori.

En raison des préoccupations du public et des médias, et de la volonté des autorités de régulation françaises d'assurer une transparence totale vis-à-vis de cette crise majeure de santé publique, le dossier de bioéquivalence, y compris ses données brutes, ont été rendus

publiques sur le site internet de l'ANSM. Les profils concentration-temps de T4 de 204 individus sains, pour les anciennes et les nouvelles formulations, ont été récupérés. Des échantillons de sang ont été prélevés avant l'administration (valeur initiale) et régulièrement jusqu'à soixante-douze heures après l'administration. D'après les recommandations de 2010 provenant de l'EMA : « *If the substance being studied is endogenous, the calculation of pharmacokinetic parameters should be performed using baseline correction so that the calculated pharmacokinetic parameters refer to the additional concentrations provided by the treatment* ». La T4 étant une hormone endogène, la correction doit alors s'appliquer. A l'issu de l'analyse comparant les ratio d'AUC ancienne-nouvelle formulation pour la T4 non ajustée et ajustée, on se rend compte que moins de 50 % des sujets (32,8 %) étaient situés dans l'intervalle de bioéquivalence défini a priori de 0,9 à 1,11 pour la T4 ajustée, contre 83,3% sans ajustements (23).

Tableau 2 : Nombre de cas (sur 204 sujets) étudiés dans chaque classe de ratio d'exposition individuelle (IER)

Class intervals	*Unadjusted T4 AUCnew/AUCold ratio*	*Adjusted T4 AUCnew/AUCold ratio*
< 0,8	2 1%	40 20%
0,8-0,9	17 8%	32 16%
0,9-1,11 (marge thérapeutique étroite)	170 83%	67 33%
1,11-1,25	15 7%	26 13%
> 1,25	0 0%	39 19%

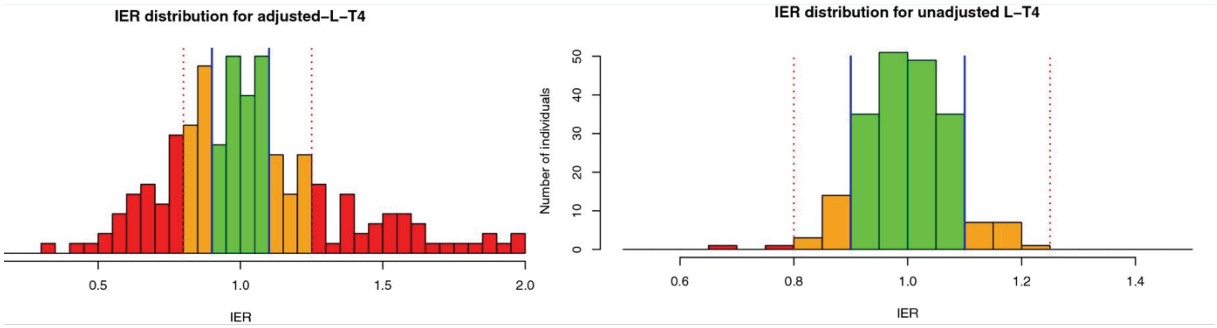

Figure 5 : Distribution du ratio d'exposition individuelle (IER) (AUCnew/AUCold) obtenue avec les concentrations plasmatiques de T4 ajustées au taux basal et de T4 non ajustées

Cependant, il est proposé que l'IBE en se concentrant sur la variabilité intra-individuelle, mérite d'être pris en considération. Il n'est donc pas possible d'affirmer que les deux formulations ne sont pas interchangeables. Néanmoins, le tracé de l'IER observé met en évidence un « signal d'avertissement » majeur à considérer pour deux raisons. Premièrement, moins de 50 % des sujets se situent dans l'intervalle de bioéquivalence défini a priori entre 0,90 et 1,11 lorsque, conformément à la directive européenne, l'AUC ajustée au taux basal est prise en compte. Deuxièmement, il y a un résultat apparemment plus favorable, lorsque l'AUC non ajustée est prise en compte. Bien qu'une telle analyse de données ne soit pas recommandée par les lignes directrices de l'UE, elle constitue une considération importante lors de l'examen de la pertinence d'une étude de bioéquivalence individuelle. En effet, pour les sujets sains de cet essai, ayant une fonction thyroïdienne normale, la T4 administrée a probablement déclenché une rétroaction négative sur la sécrétion endogène de T4. La conséquence étant un effet tampon sur la concentration plasmatique de T4, entraînant ainsi une dispersion de l'IER plus faible que lorsque l'AUC ajustée est utilisée. Il est alors possible d'émettre l'hypothèse que de tels ajustements physiologiques rapides seront moins efficaces voire absents dans la population clinique ciblée. Ces patients ayant soit une fonction thyroïdienne réduite soit une absence totale de fonction thyroïdienne (thyroïdectomie). Dans ce cas, ce sont les effets indésirables qui ont déclenché l'ajustement posologique requis pour assurer un état euthyroïdien individuel. Par conséquent, la pertinence d'utiliser des sujets euthyroïdiens sains pour évaluer la bioéquivalence pour les formulations de Levothyrox® est discutable.

La lévothyroxine est classée dans le système de classification biopharmaceutique comme une substance de classe III. C'est-à-dire une substance ayant une solubilité élevée mais une

faible perméabilité (24). La nouvelle formulation du Levothyrox® a introduit du mannitol, un excipient osmotique considéré comme critique (25), en particulier pour les médicaments de classe III (26). Les composés à faible perméabilité sont souvent soumis à une absorption dépendante du site d'action, et leur biodisponibilité dépend du temps de transit du tractus gastro-intestinal, qui peut être influencé par le mannitol. Par exemple, la biodisponibilité de l'antagoniste des récepteurs H2, la cimétidine, dans un comprimé à croquer contenant 2,264 grammes de mannitol, a été réduite de 29 % à cause d'une réduction du temps de transit dans l'intestin grêle de 20 % (27). Pour la nouvelle formulation du Levothyrox®, la quantité de mannitol est d'environ 70 mg pour un comprimé de 100 mg et un patient peut prendre deux comprimés (dans le cas d'un dosage particulier). Une petite quantité de mannitol, d'environ 140 mg, pourrait affecter le temps de transit de l'intestin grêle et donc être associée à une diminution de la biodisponibilité de la lévothyroxine. De plus, selon Chen et al. (26), la relation dose-réponse quantitative du mannitol sur l'absorption de la cimétidine ne peut pas être extrapolée à d'autres substances car, outre un effet osmotique, un excipient osmotiquement actif peut influencer soit le mécanisme d'absorption, soit le site d'absorption. Pour le sorbitol, un isomère du mannitol, il a été rapporté que de très petites quantités (7 mg ou 50 mg) peuvent affecter l'absorption du médicament et cet effet semble dépendre du sujet (28).

2.Les données de vie réelle

2.1 Introduction au cycle de vie du médicament

2.1.1 Préparation du médicament

La préparation d'un médicament est initiée pour donner suite à l'expression d'un besoin médical, à l'identification d'un marché ou à l'avancée de la recherche publique ou privée. Elle peut être décomposée en trois phases :

- La recherche
- Les études pré-cliniques
- Les essais cliniques

2.1.1.1 La recherche

Il s'agit dans un premier temps d'identifier toutes les molécules intéressantes pour la cible thérapeutique, en moyenne, 10 000 sont identifiées. C'est l'étape de screening, elle se fait grâce à des ordinateurs qui testent des molécules de différentes bases de données appelées « chimiothèques » afin d'identifier certaines propriétés physicochimiques, pharmacocinétiques et pharmacologiques. Ces techniques de criblage existent depuis les années 1990 mais ce n'est que depuis quelques années, avec l'augmentation de la puissance de calcul des ordinateurs que ces outils sont réellement efficaces. Les 10 000 molécules identifiées sont ensuite évaluées lors de tests biologiques (mesure d'absorbance, de luminescence, de fluorescence et utilisation de techniques d'imageries) pour obtenir une centaine de molécules retenues (29).

Par la suite, des essais sont menés sur ces molécules pour sélectionner les dix candidats médicaments potentiels qui feront l'objet d'un dépôt de brevet. Le brevet permet d'assurer une exclusivité temporaire aux inventeurs. Il rémunère le travail de recherche et l'investissement qui sera engagé par la suite (études cliniques, processus de recherche et développement et de mise sur le marché) (30).

2.1.1.2 Les études pré-cliniques

L'enjeu des études pré-cliniques est d'identifier, parmi les 10 molécules sélectionnées, celles qui peuvent être testées sur l'homme lors d'essais cliniques. Les tests menés sont d'abord in vitro, puis in vivo.

Les principaux essais réalisés sur ces molécules sont :

- La pharmacologie expérimentale : des essais d'efficacité sont réalisés sur des systèmes moléculaires inertes, sur des cellules et cultures et, enfin, sur des modèles animaux.
- La toxicologie : ces études évaluent les risques d'effets secondaires des futurs médicaments.
- La pharmacocinétique et le métabolisme du médicament : ces études portent sur des propriétés pharmaceutiques de la molécule telles que l'absorption, le métabolisme, la distribution et l'élimination. Mais elles ont aussi pour but de prouver les propriétés pharmacologiques.

Ces tests ont pour but de déterminer :

- La dose maximale tolérée chez l'animal. Un facteur multiplicateur sera utilisé pour déterminer cette dose chez l'homme (31).
- La dose sans effets observables.
- La dose sans effets toxiques observables. L'application d'un facteur multiplicateur sur cette dose permet de déterminer la première dose maximale sécuritaire à utiliser chez l'humain (31).

Les molécules pour lesquelles les résultats de ces études sont positifs, passent en phase d'essai clinique chez l'homme. Les résultats des études pré-cliniques doivent respecter les bonnes pratiques de laboratoire (32).

2.1.1.3 Les essais cliniques

Parmi ces dix molécules sélectionnées lors des études pré-cliniques, une seule est testée lors des essais cliniques (30). La molécule lauréate présente les meilleures propriétés pour répondre au besoin thérapeutique tout en étant bien tolérée chez l'Homme. Les essais cliniques se déroulent en quatre phases, sont strictement encadrés par la loi et doivent se dérouler selon les bonnes pratiques cliniques (31). Ces bonnes pratiques répondent à des critères éthiques afin de garantir la sécurité des patients et l'obtention de leur adhésion à l'essai. Ils doivent être informés, donner leur consentement éclairé et être avertis des risques

éventuels. Afin de protéger les participants, Bioethics International publie un indicateur sur le niveau d'éthique des essais cliniques des compagnies pharmaceutiques. Aussi, l'ANSM réalise régulièrement des inspections pour s'assurer du bon respect de ces pratiques.

Pour chaque indication, quatre phases d'essais cliniques sont menées :

- Phase 1 : Appelée phase de « tolérance » ou « d'innocuité » :
    - Population éligible : Nombre restreint de volontaires sains (20 à 80 participants) (31).
    - L'objectif étant d'évaluer la tolérance, l'absence d'effets indésirables et son activité pharmacologique (cinétique et métabolisme chez l'homme). Il s'agira d'administrer de façon croissante des doses de la molécule étudiée.
- Phase 2 : Appelée phase « pilote » :
    - Population éligible : Nombre restreint de volontaires malades (inférieur à 500 participants) (31).
    - Il s'agit de déterminer la dose optimale pour laquelle l'effet thérapeutique est le plus haut et les effets indésirables les plus bas. Une balance entre ces deux effets est recherchée.
    - Cette phase est divisée en deux :
    - Phase 2a : Estimation de l'efficacité de la molécule sur une faible quantité de malades (100 à 200 participants).
    - Phase 2b : Détermination de la dose thérapeutique de la molécule sur une quantité plus importante de malades (100 à 300 participants).
- Phase 3 : Appelée phase « pivot » :
    - Population éligible : Grand nombre de volontaires malades (plusieurs milliers de participants) (31).
    - C'est durant cette phase que l'on étudie le rapport entre la tolérance et l'efficacité. La molécule sera testée face à un placebo ou un médicament de référence. Généralement, ni le médecin, ni le malade ne savent quel traitement reçoit chacun des malades, l'étude est donc menée en double aveugle afin d'éviter tout jugement faussé.
    - En fonction du résultat de cette phase, le laboratoire pourra, si les résultats sont satisfaisants, faire une demande d'autorisation de mise sur le marché (AMM) de la molécule qui peut aboutir à la commercialisation d'un nouveau produit (31).

- Phase 4 **:** Appelée phase de « production et commercialisation ». C'est à ce moment qu'intervient la pharmacovigilance ; uniquement si le traitement arrive sur le marché et donc s'il obtient une AMM. Elle s'étendra pendant toute la durée de vie du médicament pour suivre le produit dans les conditions réelles d'utilisation, évaluer sa tolérance à grande échelle et détecter tous types d'effets secondaires inattendus, désirables (extension d'utilisation) ou non. Pour garantir la sécurité du produit, les autorités compétentes (ANSM pour la France) réalisent régulièrement des contrôles en laboratoire et des inspections des sites de production et de recherche. En cas de mise en danger de la santé des patients, l'AMM peut être retirée à tout moment. L'objectif de la phase 4 est aussi d'améliorer les connaissances sur le médicament, de mettre au point de nouvelles formes galéniques ainsi que de découvrir de nouvelles indications thérapeutiques (31).

2.1.2 La détection du signal en pharmacovigilance

Aujourd'hui, une grande quantité d'information circule au sein des bases de données servant la pharmacovigilance et le suivi post-AMM des produits de santé. C'est notamment grâce aux notifications, aux études pharmaco-épidémiologiques ou à la littérature que cela est rendu possible. Ces sources de données constituent un flux complexe dont le traitement nécessite beaucoup de ressources et de temps. Pour améliorer la vitesse d'obtention et la qualité des résultats de ces traitements de données, l'un des défis actuels consiste à perfectionner les techniques dites de « détection du signal ». En 2000, l'Organisation Mondiale de la Santé (OMS) définissait le signal comme étant une « information notifiée concernant une possible relation de cause à effet entre la survenue d'un événement et la prise d'un médicament, la relation étant inconnue jusqu'alors ou bien incomplètement documentée ». Il s'agit d'être en mesure de détecter automatiquement des signaux d'alerte dans la grande masse d'information de départ. Ce sont ces signaux qui peuvent déclencher ou non une enquête approfondie des autorités compétentes (surtout si le signal est fort). La fouille dans le signal fait ressortir des signaux forts et faibles mais également du bruit de fond qui doit être atténué au maximum. C'est le rôle des outils mathématiques, statistiques et informatiques, développés ces dernières années et utilisés dans le cadre de notre travail. Le but étant alors d'éviter que le bruit soit pris pour un signal (faux positif) ou qu'un signal se fonde dans le bruit (faux négatif). C'est la raison pour laquelle de nouveaux outils informatiques et statistiques ont été mis en avant dans ce domaine, on peut d'ailleurs citer l'avènement de l'usage de l'intelligence artificielle.

Au niveau national, c'est l'ANSM qui pilote la pharmacovigilance. Elle est assistée par un réseau régional de 31 Centres Régionaux de Pharmacovigilance (CRPV). Tous les médicaments utilisés chez l'Homme sont concernés, qu'ils soient obtenus sur ordonnance ou non, y compris les médicaments homéopathiques, ceux à base de plantes et les préparations magistrales ou hospitalières font l'objet d'un suivi par la pharmacovigilance. Ainsi, la pharmacovigilance consiste à recueillir toutes les informations sur les effets secondaires présumés ou non dans le cadre de l'utilisation conforme ou non à l'AMM. Les effets secondaires sont une grande famille d'événements observés à la suite de la prise de médicaments. Ils surviennent en plus de l'effet principal du traitement et regroupent :

- Les effets indésirables : Ils sont imprévisibles et apparaissent aux doses normales chez certains patients. Leur fréquence d'apparition augmente lorsque plusieurs médicaments sont administrés.
- Les effets latéraux : Ils ne sont pas des effets thérapeutiques recherchés mais sont inévitables et surviennent aux doses d'administration normales chez tous les individus.
- Les effets toxiques : Ils sont inévitables mais surviennent uniquement lors d'un surdosage chez tous les individus.

Il arrive parfois que les effets latéraux prennent finalement le dessus sur l'effet pharmacologique initial. L'exemple du sildénafil (Viagra®) peut être évoqué, cette molécule était initialement prévue pour traiter l'hypertension pulmonaire et l'angor. Lors des essais cliniques de phase I, les chercheurs se sont rendu compte que les résultats sur les indications initiales étaient défavorables. Cependant, en maintenant haute la concentration en monoxyde d'azote dans l'organisme, il permettait également de favoriser l'érection masculine. Le laboratoire Pfizer® décida alors de repositionner le sildénafil sur cette indication, cette molécule fut la première à permettre de traiter l'impuissance et rencontra un fort succès commercial (33).

Concernant les effets indésirables, trois types existent (34) :

- Dose-dépendant : Il s'agit d'être vigilant surtout lorsque la marge thérapeutique est étroite, lorsque la fonction rénale et/ou hépatique sont altérées ou lors d'interactions médicamenteuses.
- D'origine allergique : Le médicament se comporte comme un antigène ou un allergène. L'organisme qui est sensibilisé provoquera une réaction allergique lors d'une exposition ultérieure.
- Idiosyncrasique : Ils correspondent aux effets indésirables rares qui apparaissent de façon inhabituelle. Ils sont observés en condition normale d'utilisation du médicament et sont souvent dus à des anomalies liées à la singularité de chaque patient (déficit enzymatique, variation génétique) qui perturbent le métabolisme du médicament.

La détection du signal consiste en l'identification d'une relation encore inconnue entre la prise d'un médicament et la survenue d'un événement qui lui serait imputé (35). Cette détection se fait aujourd'hui via :

- Le signalement spontané des effets secondaires par les professionnels de santé, les industriels, les patients et associations agréées de patients. Les médecins, chirurgiens-dentistes, sage-femmes et pharmaciens ont l'obligation de déclarer immédiatement tout effet indésirable suspecté d'être dû à un médicament, dont ils ont connaissance, au CRPV dont ils dépendent (36).
- L'observation, le recueil, l'exploitation et l'évaluation de toute information concernant le risque d'effets indésirables à un niveau individuel ou dans un contexte populationnel.
- La réalisation d'études ou de travaux concernant la sécurité et l'emploi des médicaments (publications scientifiques).
- Les études épidémiologiques (elles sont surtout utilisées pour valider les hypothèses).

De nouveaux outils de détection du signal sont actuellement en développement et présentent un fort potentiel en ce qui concerne les signaux faibles, inobservables avec les techniques habituelles. Ces outils font appel à des lois statistiques utilisées via des algorithmes informatiques afin d'avoir la puissance de calcul nécessaire pour analyser une importante quantité de données.

Quelques outils statistiques ainsi utilisés :

- *Bayesian Confidence Propagation Neural Network* (BCPNN) est un réseau de neurones faisant appel au théorème de Bayes. Dans le cadre de l'application de cette méthode à la détection d'effets indésirables liés à des médicaments, l'algorithme va chercher toutes les associations possibles entre un médicament et ses effets indésirables de manière impartiale afin de détecter les problèmes significatifs les plus précocement possible.
- *Multi-Item Gamma Poisson Shrinker* (MGPS) est un algorithme analysant la fréquence d'apparition d'une liste de mots considérés comme intéressants au cours d'une période par rapport à une période de référence. Pour diminuer les biais de l'algorithme, il est possible de découper la période à analyser en plusieurs sous-périodes qui pourraient avoir des fréquences différentes. Ces strates permettent de lisser les fréquences d'apparition extrêmes et d'éviter de conclure qu'un ensemble d'éléments est inhabituellement fréquent dans la période d'observation alors qu'il l'est en réalité seulement dans une seule strate. Ce découpage permet de limiter les faux positifs.
- *Proportional Ratio Reporting* (PRR) est le rapport entre la fréquence à laquelle un évènement indésirable spécifique est signalé pour le médicament d'intérêt (par rapport

à tous les événements indésirables signalés pour le médicament) et la fréquence à laquelle le même événement indésirable est signalé pour tous les médicaments d'un groupe de comparaison (par rapport à tous les événements indésirables pour les médicaments du groupe de comparaison).

- Régression logistique est une méthode d'analyse statistique utilisée pour prédire des valeurs de données sur la base d'observations antérieures d'un ensemble de données. La fonction utilisée pour cette méthode est la fonction sigmoïde. La régression logistique prend un nombre limité de valeurs, elle est catégorielle. On parle de régression linéaire binaire si elle en prend deux ou multi-linéaire si elle en prend plus de deux.

Ces quatre méthodes peuvent être à la base d'algorithmes de *machine learning* supervisés ou non supervisés (37)(38)(39).

Le *machine learning* supervisé est une méthode d'apprentissage automatique basée sur l'entraînement de l'algorithme sur une base de données qui a été manuellement labellisée par l'Homme. La machine connaît donc les réponses qu'on attend d'elle et doit déceler les différences entre les différents labels pour être capable de labelliser la base de données qui ne l'est pas. Pour améliorer sa précision, ses premières réponses sont corrigées à la main pour qu'elle s'améliore. Pour une application destinée à détecter le genre d'une personne (homme ou femme), l'algorithme est entraîné sur une base de données où le genre lui est indiqué (label). Il cherche à détecter les similarités entre les photos de même label et les différences entre celles de labels différents (homme ou femme). Une base de données de photos non labellisées est ensuite donnée à l'algorithme qui doit indiquer s'il s'agit d'un homme ou d'une femme.

Le *machine learning* non-supervisé est différent par le fait que les réponses ne sont pas présentes dans un jeu de données. La machine doit proposer ses propres réponses en se basant sur les caractéristiques communes des données du jeu qui lui sont fournies. Par exemple, si des photos d'animaux sont proposées à l'algorithme, il pourrait classer les photos en fonction des différents animaux présents sur les photos (chien, éléphant, chat, girafe). Ce type de *machine learning* est caractérisé de non-supervisé car l'algorithme n'a aucune référence dans la base de données à analyser, il est laissé à son propre mécanisme pour en faire ressortir les structures intéressantes.

Ces algorithmes étant très récents, ils sont encore très peu utilisés en pharmacovigilance et font aujourd'hui l'objet d'études scientifiques qui ont pour but de mettre au point une approche reproductible qui permettrait de développer des outils de veille.

En effet, aujourd'hui, ce sont les professionnels de santé qui fait remonter les problèmes à la suite des observations cliniques qu'ils font. Ici, l'objectif est d'utiliser les capacités de calculs des ordinateurs via des algorithmes informatiques pour étudier une masse de données que l'Homme est incapable d'étudier lui-même. Cette nouvelle piste d'étude ne se substitue donc en rien à la pharmacovigilance actuelle mais permet d'étudier une source de données (celle présente sur internet) qui est inutilisée aujourd'hui alors qu'elle représente la plus importante masse d'information disponible et actualisée en continu.

2.2 Des données essentielles pour la qualité des soins et l'efficience de notre système de santé

2.2.1 Qu'est-ce que la donnée de vie réelle

Avec l'omniprésence des outils numériques dans le quotidien, la quasi-totalité des données sont numérisées et stockées sur des serveurs. Que ce soit l'historique de navigation internet, les déplacements ou encore les données de santé. Tout est en ligne avec différents niveaux de sécurité pour y accéder. En parallèle, l'accroissement des capacités de calcul des ordinateurs permet d'analyser une quantité toujours plus importante de données, rendant possible d'appréhender de plus en plus précisément l'usage, l'efficacité et la tolérance des médicaments. La donnée de vie réelle correspond à toutes les données qui ne sont pas collectées dans un cadre expérimental (essai clinique), mais qui sont générées dans la pratique courante, notamment à l'occasion des soins réalisés en routine (40)(41).

Avant l'avènement des outils informatiques, les données de santé qui étaient à disposition étaient de la donnée structurée issue principalement des essais cliniques. La donnée de vie réelle est quant à elle complètement déstructurée et ne provient pas d'un cadre expérimental. Elle peut provenir des notes qu'un médecin a pris lors d'une consultation, de données provenant d'objets connectés, ou encore des réseaux sociaux et plus largement du web. L'intérêt de ces données de vie réelle réside dans le fait qu'elles ne sont pas observables lors des essais cliniques. Elles décrivent l'impact qu'ont les médicaments sur la qualité de vie des patients ou le comportement de ces derniers vis-à-vis de leurs traitements. L'inconvénient est qu'elles n'ont pas de comparateur car ce sont des études monobras non contrôlées. Le travail qui est fait lors de ces études est d'analyser le chemin des patients sans avoir le contrôle sur les facteurs d'exposition. Aucun plan expérimental n'est établi donc aucun lien de causalité ne peut être formellement conclu.

Tout l'enjeu de l'analyse de ces données de vie réelle réside donc dans la capacité à les structurer pour les rendre exploitables. Les études en vie réelle et les essais cliniques sont donc complémentaires. La vie réelle permet :

- De vérifier à quel point les conditions des essais cliniques sont observées et/ou applicables dans la vraie vie.

- De détecter via des populations numériques plus importantes, des effets qui ne seraient pas repérables dans les études cliniques.
- D'augmenter la durée de suivi des patients et d'observer des effets à long terme.
- De suivre des patients qui sont rarement inclus dans les essais cliniques (grande fragilité, multiples comorbidités).
- De détecter de nouvelles tendances : plaintes soudaines à la suite d'une modification mineure (changement de packaging, de formule…), ruptures de médicaments, mésusage/abus, etc.

2.2.2 L’utilisation de la donnée de vie réelle en France

En 2016, la France a lancé un important projet concernant la donnée de vie réelle avec la création du SNDS.

Cette base de données regroupe :

- Les données de l’Assurance Maladie.
- Les données des hôpitaux.
- Les causes médicales de décès.
- Les données relatives au handicap.
- Un échantillon de données en provenance des organismes d’Assurance Maladie complémentaire.

Cependant, cette base présente des limites dans le sens où elle ne propose pas de données cliniques (comptes rendus d’examens cliniques et biologiques, motifs de consultation, imageries, etc…). Pour élargir le champ d’action du SNDS à ces données, l'État a légiféré en 2019 en faveur de la loi Organisation et transformation du Système de Santé. Cette loi a ainsi permis de créer en 2019 le Health Data Hub (HDH) qui lance des appels à projets de recherche, pour travailler sur les données du SNDS.

Depuis sa création en 2019, le HDH a permis à plus de 1600 projets de voir le jour. Il y a bientôt un an (juin 2021), a été signé un partenariat entre l’ANSM et le HDH afin de faciliter l’accès aux données disponibles sur les médicaments, améliorer la transparence et les connaissances des traitements. Le HDH permet un accès facilité à la majorité des données de santé provenant des établissements de santé. Cette donnée est exprimée selon des termes normés et scientifiques, est objective et a été formulée par des professionnels de santé. Bien que la France possède l’une des plus importantes bases de données au monde, le web rassemble une quantité presque infinie d’informations. Cette information est en revanche très subjective, dépendante du ressenti de chaque patient, basée sur des témoignages et donc non normée. Cette donnée présente beaucoup plus de biais et de bruit de fond mais en termes de suivi au jour le jour des traitements et de détection précoce des signaux faibles pour signaler des effets indésirables, elle présente un potentiel très prometteur.

Le dernier exemple français qui pourrait être évoqué est le Dossier Médical Partagé (DMP). Ce projet d’initiative public a vu le jour en 2011 et les premiers dossiers ont été créés en 2014.

Un DMP est un carnet de santé numérique qui rassemble toutes les informations médicales de chaque patient. Le but étant que tout le monde puisse, où qu'il soit, avoir accès à ses propres informations médicales. Les professionnels de santé peuvent aussi avoir accès au DMP des patients. Les données accessibles sont cependant différentes selon la profession du professionnel de santé, des droits d'accès différents sont accordés. Lors de sa création le DMP était principalement destiné aux médecins généralistes pour leur fournir la vision la plus globale possible de l'historique médical de chaque patient afin de dispenser les soins les plus adaptés et d'éviter les redondances d'examens et de prescriptions. Aujourd'hui, le projet n'a toujours pas atteint les objectifs attendus du fait de sa complexité d'utilisation. L'ensemble de l'information est stocké sous forme de nombreux fichiers PDF très chronophages à étudier. Depuis le 1er juin 2021, un Espace Numérique de Santé (ENS) est créé pour tous les français comme indiqué dans le projet de loi « Ma santé 2022 ». Sur cet espace, il est possible de consulter son DMP (qui sera revu pour rendre son usage facilité), ses données générées par des objets connectés, ses données de remboursements et d'échanger avec des professionnels de santé.

En termes d'innovation autour de la donnée de vie réelle de santé, les deux sources majeures qui présentent un fort potentiel sont :

- La donnée provenant du Web et en particulier des réseaux sociaux.
- La donnée provenant des dispositifs connectés.

Les milliards d'utilisateurs des réseaux sociaux en font des bases de données à fort potentiel pour la recherche. Aujourd'hui, Twitter® est un réseau social fortement utilisé car les messages des utilisateurs sont limités en nombre de caractères. Les tweets constituent une version condensée du message et nécessitent donc moins de nettoyage que sur les autres réseaux sociaux, il est ainsi plus facile d'obtenir des résultats significatifs.

Les dispositifs connectés représentent la deuxième source majeure de données de vie réelle. Sont retrouvés, les dispositifs médicaux connectés (tensiomètres, glucomètres, oxymètres, balances, etc…), les accessoires de prêt à porter connectés (montres, bijoux, vêtements...) et bien sûr les smartphones avec leurs nombreuses applications de santé. Tous ces objets collectent de la donnée en continu et permettent, si la donnée est utilisée, de mettre en place une médecine personnalisée préventive et prédictive. L'avantage principal de ces outils connectés est qu'ils permettent en continue de collecter de la donnée et de suivre les patients en impactant au minimum leur quotidien.

Pour faciliter l'accès et la sécurité de cette donnée (bien qu'elle soit pour une grande partie en libre accès sur le web), des systèmes de *blockchains* sont en développement. Pour rappel, la *blockchain* est une technique de stockage et de transmission d'information sans organe de contrôle. Pour que l'information soit transférée, elle doit être validée par tous les acteurs de la chaîne comme elle se situe de partout à la fois. Ainsi, avec une technologie de *blockchain*, la donnée de vie réelle peut être accessible par tous les acteurs de façon sécurisée. Sans cette technologie, l'information peut être stockée et envoyée n'importe où. Pour que ce ne soit pas le cas, de nombreuses contraintes et barrières de sécurité sont mises en place, en contrepartie, l'accès à la donnée est rendu plus difficile. La *blockchain* permet donc de s'affranchir de ces contraintes.

3. Détection précoce d'événements au travers de l'analyse des réseaux sociaux : socle expérimental basé sur l'affaire du Levothyrox® et l'analyse de la donnée sur les forums Doctissimo®

Ce travail de thèse a fait l'objet de la rédaction d'un article intitulé « *AI-based Approach for Safety Signals Detection from Social Networks : Application to the Levothyrox® Scandal in 2017 on Doctissimo® Forum* » (Journal : *Artificial Intelligence in Medicine*).

3.1 Intérêt de ce travail et pertinence des méthodes employées

3.1.1 Limites des essais cliniques et des outils actuels de pharmacovigilance

Lors de la mise sur le marché d'un médicament, les autorités sanitaires disposent de données suffisamment robustes pour conclure à une balance bénéfice-risque d'un produit dans une indication donnée. Les profils de tolérance et d'efficacité sont également définis dans le cadre d'un usage sur l'être humain. Cependant, elles n'empêchent pas la survenue de scandales à postériori. C'est le cas de la récente affaire du benfluorex, principe actif contenu dans la spécialité Mediator®, aujourd'hui retirée du marché pour mésusage. Il était initialement prescrit contre le diabète avant d'être prescrit pour les patients souhaitant perdre du poids qui est une indication hors-AMM. Il a conduit au décès d'environ 1300 personnes en France (45). Dans cet exemple, ni les essais cliniques ni le suivi post-AMM ne permettent de déceler le problème. C'est à la suite de ce scandale que le système de pharmacovigilance a été modifié, en 2011, du fait d'un constat de sous-notification de la part des professionnels de santé.

La compréhension des limites des outils est la clé pour générer une innovation favorable pour le système de santé et les patients. Dès 1976, les premiers centres régionaux de pharmacovigilance ont vus le jour après avoir constaté les limites des essais cliniques concernant la connaissance et le suivi des médicaments. Les principales limites sont (43) :

- Les cohortes comprenant un faible nombre de sujets : de 20 sujets en phase I jusqu'à plusieurs milliers de sujets en phase III. A titre de comparaison, le Levothyrox® est utilisé par plus de 3 millions de personnes en France.
- La durée des essais planifiée en amont et la non-prise en considération des spécificités de certaines interventions (médicaments, acte de chirurgie…).
- La phase de sécurité (toxicité) : l'évaluation est faite sur des sujets jeunes et sains uniquement.
- La phase d'efficacité : les sujets sont sélectionnés selon des règles strictes et ne permettant pas d'évaluer l'impact d'une thérapeutique sur les co-morbidités ou les co-administrations d'autres traitements.

Le faible nombre de sujets impliqués au cours des essais cliniques vis-à-vis de la population cible du traitement une fois l'AMM obtenue, constitue la principale limite à l'évaluation de la tolérance et de la sécurité du produit évalué. Le rôle de la phase IV de ces essais vise à suivre le produit tout au long de sa commercialisation, de façon à collecter des éléments non identifiables (statistiquement rares) durant les essais. C'est ainsi que certains produits voient leurs balances bénéfices-risques réévaluées et parfois, ils sont purement et simplement retirés du marché. La détection de certains événements indésirables rares et graves se fait donc à postériori. D'autres facteurs conduisent au même résultat, tels que les différents biais liés aux protocoles des essais cliniques. Ceux-ci diffèrent selon les produits (maladies rares/orphelines, oncologie, dispositifs médicaux, etc.), les réglementations ou les pays (procédure commune d'autorisation de mise sur le marché au niveau européen). Ces biais peuvent être pris en compte puis limités.

Le système de pharmacovigilance actuel est basé sur la notification et la déclaration de rapports de pharmacovigilance. Les professionnels de santé (médecins, pharmaciens, dentistes, sages-femmes) se doivent de notifier tout effet indésirable qu'ils pourraient imputer à un médicament, et sans délai. Toutefois, le scandale du Mediator® a mis en évidence des carences importantes de ce système de surveillance post-AMM (phase IV). Il a conduit à une amélioration des processus de déclaration, afin d'inciter les professionnels de santé à déclarer davantage (44). Malgré l'amélioration continue des pratiques et des organisations, les lacunes d'antan n'ont pas disparues :

- Le faible nombre de transmission des déclarations de pharmacovigilance malgré une amélioration du système de déclaration en 2011 (45).

- La méconnaissance des règles de déclaration de pharmacovigilance. Règles qui sont encore davantage méconnues pour les autres types de vigilances (matériovigilance, réactovigilance, etc.).
- Toutes les informations recueillies ne peuvent être prises en compte. Celles-ci sont collectées selon un format particulier et défini à l'avance pour permettre leur exploitation. De ce fait, il arrive que certains éléments significatifs puissent ne pas être pris en compte (paramètre biologique, environnement, co-morbidité...).
- La donnée de vie réelle collectée par les professionnels de santé est rarement inscrite dans le dossier patient ou dans les déclarations de pharmacovigilance.

De ce fait, la quantité de données collectée est limitée. Certaines interprétations statistiques sont difficiles et parfois, un produit passe entre les mailles du filet et peut se retrouver au cœur d'un scandale sanitaire : thalidomide, poches pour transfusion sanguine, prothèses PIP®, Mediator®, Levothyrox®.

Néanmoins, ces outils introduits progressivement au cours du XXème siècle restent l'un des piliers des sciences médicales modernes, de la médecine fondée sur les faits (*Evidence-Based Medicine*). Ils sont le garant d'un niveau minimal de sécurité et d'efficacité pour une intervention et une indication donnée. S'ils sont indispensables, certains événements portant à l'intérêt des populations peuvent encore survenir, du fait de certaines limites détaillées précédemment. L'idée directrice de ce projet est d'apporter une approche complémentaire visant à réduire les biais et les contraintes inhérentes aux systèmes d'information actuels. La donnée de vie réelle est aujourd'hui un champ crucial, au centre des réflexions et des enjeux de la pharmacovigilance et plus largement de la santé. C'est ce type de données qui peut être exploitée à l'aide des nouvelles technologies, notamment dans le domaine des *data sciences* et de l'intelligence artificielle. Il est alors possible de récolter une masse très importante de données de vie réelle anonymisées, qui, en termes de valeur, permet d'outrepasser l'effet dit de la « blouse blanche » et du faible nombre d'observations collectées à l'hôpital ou dans les cabinets médicaux. Ces données permettent de dégager, au sens statistique, de nouvelles tendances peu ou pas détectables via l'emploi des outils traditionnels. L'utilisation de ces algorithmes, en complément des outils actuels de pharmacovigilance permettrait de renforcer la sécurité et le suivi des produits de santé.

3.1.2 Utilisation de l'intelligence artificielle en santé

L'intelligence artificielle occupe une place grandissante dans le secteur de la santé. Comme évoqué précédemment, elle complète les actions humaines et permet d'optimiser les soins et le suivi des patients. L'association médicale américaine (AMA) définit l'IA comme une "intelligence augmentée" qui améliore l'intelligence humaine plutôt que de la remplacer. Aussi, une récente enquête de l'AMA montre que les médecins sont très réceptifs aux outils de santé numériques. Ils considèrent qu'ils peuvent être intégrés en douceur dans leur pratique actuelle et améliorent les soins en renforçant la relation patients-médecins (46).

L'intégration de l'IA dans les flux de travail des professionnels de santé permet ainsi l'utilisation du *big data* pour optimiser l'utilisation des informations des patients. L'objectif étant de renforcer la prise de décision basée sur les preuves et d'améliorer la qualité, la sécurité et l'efficacité des soins coordonnés ou non.

Aujourd'hui, les dossiers de santé électroniques (comme le DMP en France) sont de formidables bases de données pour intégrer des algorithmes d'IA. Actuellement, ils sont principalement des systèmes numériques d'enregistrement et de stockage d'informations aux capacités limitées. L'intelligence artificielle peut transformer ces dossiers en des outils intelligents. Plusieurs champs d'applications sont aujourd'hui déjà envisageables pour (47) :

- Fournir une aide à la décision clinique au moment des soins et améliorer la précision du diagnostic et des recommandations de traitement. Les algorithmes d'IA sont aujourd'hui capables :
    - D'analyser l'imagerie médicale (scanner, IRM, ECG, radiographies...).
    - D'étudier les données génomiques et comportementales (symptômes, antécédents familiaux...) pour permettre une médecine plus personnalisée et hiérarchiser les populations en fonction des facteurs de risques.
- Optimiser l'utilisation des ressources à disposition :
    - Intégrer des données provenant de dispositifs mobiles qui peuvent être connectés.
    - Utiliser le traitement naturel du langage pour analyser les données de santé narratives (notes du médecin, rapports, cliniques) et fournir des résumés critiques des informations clés.

- Améliorer la recherche clinique sur les données de vie réelles. Cela est permis en collectant, et combinant les différentes sources de données de vie réelle avec les résultats des essais cliniques pour optimiser la recherche clinique et les soins.

Ces dernières années, le *machine learning* a prouvé sa fiabilité dans le diagnostic et la détection des maladies. En effet, de nombreux algorithmes ont été approuvés par la FDA via la soumission de déclaration de pré-commercialisation « 510(k) » et l'approbation de pré-commercialisation (PMA : *premarket approval*). Une autre soumission préalable existe, c'est la soumission de novo. Chacun de ces types de soumission donne lieu à une décision de la FDA qui autorise (510(k)), approuve (PMA) ou accorde (de novo) des droits de commercialisation au demandeur retenu. La plupart des algorithmes approuvés l'ont été dans le domaine de la radiologie, de la cardiologie, de l'oncologie et de la dermatologie. Ils sont principalement basés sur de la reconnaissance d'images (48).

Dans le secteur privé, de nombreuses entreprises développent des solutions de santé embarquant des algorithmes d'intelligence artificielle notamment :

- Alphabet® qui a développé AlphaFold® pour prédire la structure tridimensionnelle des protéines à partir de la séquence d'acides aminés. L'algorithme couvre la majorité des protéines de la base de données Uniprot® (49).
- IBM® annonce en 2013 que la première application commerciale du logiciel Watson® serait destinée à aider à la prise de décision des différentes recommandations de traitements du cancer du poumon au *Memorial Sloan Kettering Cancer Center* de New York (50).
- Nuance® propose des outils de *Natural Language Processing* (NLP) qui peuvent être intégrés aux dossiers de santé électroniques pour faciliter la documentation clinique et la saisie des données.
- Amazon® dispose d'un outil de NLP (Comprehend Medical®) pour analyser du contenu clinique non structuré.

Certains pays tels que le Royaume-Uni, le Canada, les États-Unis et la Chine ont d'ores et déjà implémenté, dans leurs institutions, des outils d'intelligence artificielle provenant d'entreprises privées pour améliorer la qualité des soins. Parmi ces quatre pays pionniers, c'est la Chine qui est en train de devenir la première puissance mondiale en termes d'intelligence artificielle. Ceci est permis grâce à l'état chinois qui met en place une stratégie nationale globale visant à soutenir l'IA via des financements publics, à son réseau d'hôpitaux

publiques et à une gouvernance des données moins restrictive (51). En 2018, le gouvernement chinois a publié la directive « *Internet Plus Healthcare* » qui indique que les technologies d'IA doivent être utilisées pour offrir des services médicaux et de santé publique, faciliter les pratiques des médecins de famille, faciliter l'approvisionnement en médicaments et le paiement des factures médicales, et fournir une formation médicale. Aujourd'hui, dans beaucoup d'hôpitaux chinois, les patients sont accueillis et orientés par des robots et de nombreuses entreprises privées s'associent à des hôpitaux pour proposer des services expérimentaux de diagnostic basés sur l'IA (52) :

- Tencent® avec son algorithme AIMIS® est présent dans une centaine d'hôpitaux pour détecter des cancers de l'œsophage, du poumon, du col de l'utérus, du sein et la rétinopathie diabétique grâce à l'analyse de l'imagerie médicale.
- Baidu® a lancé en 2016 Melody the Medical Assistant®, un *chatbot* basé sur l'IA conçu pour converser avec les patients et collecter des données sur leur état afin de faire gagner du temps aux médecins.
- Alibaba®, a créé ET Medical Brain® pour aider les médecins dans les domaines de l'imagerie médicale, du développement de médicaments et de la gestion de la santé.

Bien que l'IA ait encore besoin de temps pour aboutir sur des technologies matures, elle est de plus en plus présente dans le secteur de la santé et se positionne comme un assistant des professionnels de santé. C'est parce que c'est une technologie prometteuse qu'autant de pays misent sur elle et investissent du temps et de l'argent. Les outils les plus aboutis actuellement sont ceux basés sur la reconnaissance d'images. Les champs d'application se développent avec les *chatbots*, le *drug design* ou encore la gestion administrative des patients. En se positionnant sur le créneau de la pharmacovigilance, ce travail de thèse s'inscrit exactement dans la continuité de l'extension des champs d'applications des programmes d'IA. Comme présenté précédemment, les outils actuels de pharmacovigilance montrent leurs limites. C'est ainsi, et dans la même logique de complémentarité des outils existant et des professionnels de santé que l'algorithme proposé se positionne pour améliorer et optimiser le suivi post-AMM des médicaments.

En plus de faciliter la surveillance et la prise en charge des patients, une optimisation du système de pharmacovigilance permettrait de réduire fortement le coût imputé au système d'assurance maladie. En effet, la dernière version Européenne de la législation de pharmacovigilance estimait, grâce à une optimisation des pratiques existantes, sauver 5000

vies et économiser 2,5 milliards d'euros par an en Europe (53). En France, la iatrogénie est à l'origine de 130 000 hospitalisations (6,5% des admissions, 9% des séjours hospitaliers et 15% des entrées en réanimation) et de plus de 12 000 décès annuels en France. En plus d'avoir pour origine des erreurs humaines dans plus de 50% des cas, ces événements se traduisent par un coût moyen par patient de 13 000 euros (54). Nombreux sont les cas purement associés à un problème de bon usage du produit de santé (erreur de prescription, âge, polymédication, mésusage) (55). Il reste que certains événements iatrogènes particulièrement graves sont liés à des effets indésirables rares et ou graves non connus et non associés à la thérapeutique concernée. Dans ces cas, le rôle de la pharmacovigilance est absolument fondamental et conduit chaque année à des réévaluations de certains médicaments ; parfois allant jusqu'au retrait pur et simple de l'AMM.

3.1.3 Revue de la littérature

Dans le cadre de ce travail, la revue de la littérature est incontournable, d'autant plus que le sujet traité est innovant (pas de commercialisation de ce type de solution à l'heure actuelle). Seule certitude au début de ce travail : la donnée générée par les utilisateurs des réseaux sociaux est précieuse et insuffisamment étudiée de nos jours. Celle-ci permet d'obtenir des informations sur la vie réelle des patients qui est une source d'information de plus en plus utilisée par les autorités et le secteur privé. Lors de la revue des travaux déjà réalisés sur le sujet, sept articles suffisamment détaillés et pertinents ont permis de concevoir les premières itérations algorithmiques.

3.1.3.1 Analyse du sentiment des commentaires des patients

Plusieurs travaux portent sur l'analyse du sentiment des commentaires des patients sur les forums médicaux. En général, la polarité (positive ou négative) est détectée pour tenter de comprendre la satisfaction des patients sur divers aspects de leurs soins de santé. Par exemple, une équipe a étudié comment les gens expriment leur opinion sur les médecins et les médicaments en explorant l'analyse des sentiments basée sur le lexique et l'apprentissage supervisé (56). Des modèles permettant de détecter la polarité des sentiments (positif/négatif) dans les commentaires sont entraînés séparément sur les médicaments et sur les médecins. Il s'avère que les avis sur les médicaments sont plus difficiles à classer que ceux sur les médecins. L'utilisation du langage informel est la principale difficulté pour analyser les critiques sur les médecins. Pour les médicaments, en plus du langage informel, des terminologies spécifiques (effets indésirables, noms de médicaments) sont utilisées et apportent une plus grande diversité lexicale, donc plus de complexité, pour caractériser les critiques.

Dans le même esprit, une autre approche (57) tente de catégoriser la polarité des commentaires des patients concernant leurs soins de santé à l'hôpital (recommandation de l'hôpital, propreté de l'hôpital, qualité des traitements prodigués aux patients). Dans ce travail, sont étudiés l'évolution dans le temps de la polarité des sentiments des commentaires des patients comme indicateur possible pour la détection d'évènements indésirables liés à l'utilisation de produits de santé.

3.1.3.2 Recherche ciblée d'informations médicales

Certains travaux abordent la recherche d'informations ciblées en utilisant des messages cliniques ou des commentaires de patients. Par exemple, une approche pour le filtrage de messages cliniques pertinents a été proposée (58) et le NLP est utilisé pour identifier des phrases et des mots-clés dans les messages postés sur une liste de diffusion d'e-mail. Les phrases et mots-clés sélectionnés sont ensuite utilisés pour identifier, filtrer et stocker les messages cliniquement pertinents pour une analyse ultérieure. Les techniques de pré-traitement comprennent la suppression des *stopwords*, la conversion des majuscules en minuscules, la suppression des mots de moins de 3 caractères, la sélection des termes d'une longueur supérieure à cinq et d'une fréquence supérieure à 7, l'analyse des 300 bi-grammes et tri-grammes les plus fréquents. Les commentaires sont ensuite triés selon le nombre de n-grammes et mots-clés qu'ils contiennent ainsi que le calcul de leur occurrence par an.

Un système de filtrage semi-automatique a également été proposé (59) et permet d'identifier des termes pertinents dans des jeux de données en vue de leur inclusion dans une liste collaborative et en accès libre de vocabulaire de santé grand public. Le système se compose de trois parties principales : un robot d'exploration du Web et un analyseur de code HTML, un système de filtrage de termes candidats qui utilise des techniques de NLP telles que des méthodes de reconnaissance de termes, et une interface de vérification humaine. Dans ce travail est effectué une analyse de la fréquence des mots et des bi-grammes dans les commentaires des patients. Ensuite est étudié si l'évolution dans le temps de cette fréquence peut être indicative pour la détection d'évènements indésirables liés à l'utilisation de produits de santé.

3.1.3.3 Identification et extraction des effets indésirables

L'identification et l'extraction des effets indésirables des médicaments à partir des données des médias sociaux ont également attiré l'attention de la communauté des chercheurs en pharmacovigilance. Parmi les différentes approches, une étude utilise une technique de classification supervisée sur les forums de santé pour analyser l'observance et les comportements déviants des patients avec leurs traitements (60). L'approche proposée utilise la tokenisation, le marquage POS et la lemmatisation comme techniques de prétraitement. Les méthodes de classification NaiveBayes, Random Forest et Simple Logistic sont ensuite utilisées. Une analyse manuelle du contenu des messages a révélé que la mauvaise observance est dans 28% des cas liée à une diminution de la consommation de médicaments,

dans 27% liée à une surconsommation et dans 6% des cas liée à une consommation anarchique. De nombreux travaux portent sur l'identification et l'extraction des effets indésirables des médicaments (EIM) à partir des données des médias sociaux. Plusieurs chercheurs français ont publié un article (61) présentant un panorama des différentes approches existantes : les approches traditionnelles et les approches d'apprentissage profond.

3.1.3.4 Approches traditionnelles

Parmi les travaux existants dans ce domaine, un protocole standardisé a été proposé (62). Celui-ci concerne l'évaluation d'un logiciel basé sur le NLP qui extrait des EIM à partir de messages de forums de santé. L'extraction de l'information sur les EIM est réalisée en extrayant la relation entre le médicament et les événements indésirables, puis testée par rapport à un modèle réalisé manuellement par deux personnes spécialisées en terminologie médicale. L'une des approches utilise une machine d'apprentissage automatique pour extraire les EIM (63). Les caractéristiques des EIM sont obtenues en concaténant les chaînes de caractères (obtenus à l'aide d'un moteur d'apprentissage récurrent) et les enchaînements de mots (obtenus à partir d'un modèle pré-entraîné). Un F-score de 87,5 % a été obtenu avec cette méthode. De même, une méthode de détection des EIM dans les médias sociaux basée sur un algorithme SVM et des modèles lexicaux de type « *skip-gram* » a été proposé dans le cadre d'un autre travail (64).

Des méthodes de filtrage des termes relatifs aux troubles ne correspondant pas à des événements indésirables ont également été retrouvées dans la littérature. Une technique (65) exploite une approche basée sur la distance (en nombre de mots) entre le terme du médicament et le terme du trouble/symptôme. L'analyse est réalisée sur un corpus de paires médicament-trouble/symptôme provenant de cinq forums français via un modèle gaussien et un algorithme de modélisation par maximisation de l'espérance. Les résultats montrent que la distance entre les termes peut être utilisée pour identifier les faux positifs, améliorant ainsi la détection des EIM dans les médias sociaux.

3.1.3.5 Approches d'apprentissage profond

Récemment, l'apprentissage profond a attiré les chercheurs pour proposer des approches de détection et d'extraction d'EIM à partir des médias sociaux (66). Parmi les architectures utilisées, sont retrouvées des approches basées sur les réseaux neuronaux récurrents (RNN) (67). Cette approche RNN labellise les mots avec les labels relatifs aux EIM qui leur

correspondent. Les mots sont ensuite traduits en vecteurs via les techniques de *word-embedding*. Ces vecteurs sont ensuite utilisés comme caractéristique d'entrée du RNN. Un réseau de neurones de type Bi-LSTM (68) est proposé pour la détection et l'identification des effets secondaires des médicaments non signalés par les professionnels de santé. L'approche utilise des encapsulations de phrases BERT (*Bidirectional Encoder Representations from Transformers*) qui surpassent les architectures d'apprentissage profond standard.

Malgré le succès de l'apprentissage profond dans l'identification et la détection des EIM, il n'a pas été étudié pour la détection des signaux de sécurité potentiels. Aussi, aucune des approches existantes n'a exploré l'utilisation des images des nuages de mots comme donnée d'entrées d'un réseau neuronal convolutif profond.

3.1.3.6 Traitement de substitution des hormones thyroïdiennes

Très peu de travaux abordent l'analyse de données concernant le traitement hormonal substitutif de la thyroïde. Des travaux utilisent le NLP pour identifier les thèmes préoccupant des patients qui utilisent des médicaments pour le traitement hormonal substitutif de la thyroïde (69). Le travail est réalisé sur un ensemble de données collectées sur WebMD® aux États-Unis. Ils utilisent des analyses de régressions multiples pour examiner la prédictibilité des préoccupations des patients vis-à-vis de leur traitement. Leur étude révèle six thèmes distincts de problèmes de médication liés à la prise de lévothyroxine. Aussi, elle montre que la satisfaction du traitement à la lévothyroxine est très dépendante de la survenue ou non d'EIM. L'approche de pré-traitement consiste à supprimer les *stopwords*, à utiliser la tokenisation, à réaliser un *stemming* et à analyser la fréquence des mots. C'est l'algorithme *Latent Dirichlet Allocation* (LDA) qui est utilisé pour détecter les sujets de préoccupation des patients.

Par rapport aux approches examinées ci-dessus, cette thèse s'inscrit dans la catégorie de la détection des signaux de sécurité en pharmacovigilance. Sont analysés les commentaires des patients pour extraire les indicateurs possibles d'un signal de sécurité. Une architecture de réseau neuronal convolutifs profond est proposée. Elle prend en donnée d'entrée les images des nuages de mots extraites des commentaires des patients. Sont explorées différentes techniques de prétraitement via des algorithmes de NLP et leur effet sur la performance du modèle d'apprentissage profond. L'approche proposée est holistique dans le sens où est examiné le contenu global des messages des patients et leur évolution dans le temps.

Ces articles donnent une bonne idée des étapes nécessaires au nettoyage et à l'analyse d'un jeu de données :

- Utilisation de « *stopwords* » que l'on ne veut pas retrouver dans les données, utilisation de dictionnaire pour les discriminer (ex : le, bonjour, Danone, etc.).
- Utilisation de n-grammes permettant de réaliser des analyses sur des groupes de mots pertinents plutôt que des mots seuls (perte de sens).
- Utilisation des règles « RegEx » (*Regular Expression*) pour supprimer les données indésirables (passage en minuscule, suppression des accents, suppression des caractères spéciaux, phrases ayant moins de 3 mots, suppressions des chiffres, suppression des dates…).
- Utilisation de codifications internationales (la classification ATC dans le cas de ces articles).
- Utilisation de techniques de NLP.

Sans calquer les expérimentations déjà réalisées, ces articles permettent de réduire le temps de réflexion pour construire la méthode qui est réalisée dans ce travail. Certaines technologies sont reprises, de même que la chronologie de certaines démarches (nettoyage puis analyse de fréquence, par exemple). Au-delà des ressources bibliographiques, de nombreux outils de data science et d'IA sont déjà employés dans d'autres secteurs d'activité. De ce fait, il est nécessaire de réaliser le panorama actuel des technologies disponibles pour sélectionner celles qui répondent le mieux aux besoins de ce travail.

3.1.4 Techniques actuelles

3.1.4.1 Traitement automatique du langage naturel (Natural Language Processing – NLP)

Le traitement automatique du langage dit « naturel » est une technologie permettant à des machines d'analyser le langage humain grâce à l'IA. Il existe de nombreuses techniques autour du NLP, telles que :

- L'analyse de la fréquence des mots et de leur historique, plus généralement des n-grammes qui sont des associations de n termes du discours.
- L'analyse de sentiments.
- La traduction automatique.
- La détection de thèmes (nouvelles tendances Twitter®, nouveaux sujets abordés dans les médias, améliorer un produit à partir de retours utilisateurs dans des commentaires, etc…).
- La classification de messages (exemple : *spams*).
- La reconnaissance vocale.
- Les assistants personnels tels que Apple Siri®, Microsoft Cortana®, Amazon Alexa®, etc.
- Les *chatbots*.
- La génération automatique de texte.

Les phrases seront transformées en des structures de données utilisables par les algorithmes de Traitement du Langage Naturel (*Natural Language Processing* - NLP) : liste de mots (ou tokens), graphiques, matrices, etc… Avant cette transformation, les textes font habituellement l'objet de différents traitements dont le choix dépendra essentiellement des algorithmes utilisés et des buts poursuivis. Pour limiter les temps de traitement, compte-tenu du volume important des données manipulées, il faudra s'assurer que ne sont conservés que les mots qui peuvent contribuer de manière significative à l'obtention du résultat. Les tâches les plus couramment exécutée sont :

- Passage en minuscules, sauf si la distinction entre majuscule et minuscule est porteuse de sens.
- Suppression des accents, en français notamment.
- Tokenisation : découpe le texte en phrases et les phrases en mots.

- Suppression de la ponctuation, des espaces inutiles, des caractères spéciaux, des mots dont la longueur est inférieure à une certaine limite.
- Suppression des *stopwords* (mots fréquents de la langue qui ne sont pas porteur de sens dans l'analyse envisagée : et, avec, déjà, je, suis, es, est, etc.).
- Lemmatisation.
- *Stemming*.

La « tokenisation » cherche à transformer un texte en une série de « tokens » individuels pour appliquer les transformations souhaitées dessus. Chaque token représente un mot. La complexité de la tâche réside dans la gestion des mots tels que : « J'ai froid ». Il faut que le modèle de tokenisation sépare le « J' » comme étant un premier mot.

3.1.4.2 La reconnaissance des Entités Nommées (Named Entities Recognition – NER)

C'est une sous-tâche de l'activité d'extraction de l'information des corpus documentaires. Elle consiste à trouver des objets textuels (un mot, un groupe de mots, etc.) qu'il est possible d'identifier dans des classes telles que des noms de personnes, des noms d'organisations, des lieux géographiques, des quantités, des distances, des valeurs, des dates, etc.

Exemple : « Henri a acheté 3000 actions de la société SANOFI en 2013 » Le résultat de la détection des entités nommées sous forme du texte étiqueté avec des balises XML respectant le standard d'étiquetage ENAMEX est donné ci-dessous :

<ENAMEX TYPE="PERSON">Henri</ENAMEX> a acheté <NUMEX TYPE="QUANTITY">3000</NUMEX> actions de la société<ENAMEX TYPE="ORGANIZATION">SANOFI</ENAMEX> en <TIMEX TYPE="DATE">2013</TIMEX>.

C'est une phase d'analyse clé permettant de représenter un texte sous forme d'un graphe représentant les entités du corpus et les relations entre ces entités. Les techniques utilisées sont basées sur des méthodes d'apprentissage. On soumet un ensemble de texte annoté où chaque mot ou groupe de mots se sera vu attribuer une étiquette (tag) au standard ENAMEX ou équivalent à un système d'apprentissage comme un réseau neuronal. Une fois la phase d'apprentissage terminée, le réseau neuronal est capable d'analyser un texte et d'attribuer à chaque mot ou groupe de mots l'étiquette vue comme la plus probable au regard du modèle élaboré lors de la phase d'apprentissage.

En python, on peut faire appel à deux librairies, NLTK ou Spacy. On trouve des modèles prêts à l'emploi résultant d'une phase d'apprentissage utilisant des textes de la langue courante. Les modèles sont assez facilement disponibles en anglais. Cependant, on trouve peu de modèles prêts à être utilisés en français et quand ils sont disponibles, les résultats ne sont pas toujours convaincants. La performance des modèles de reconnaissance d'entités nommées est largement dépendante de la phase d'apprentissage. Il est essentiel de vérifier qu'elle a été conduite sur des textes annotés représentatifs du domaine étudié. D'autant plus que le secteur médical est richement doté en terminologies et codifications qui ont été spécifiquement développées (SNOMED CT, CIM, VIDAL, ATC…). Après apprentissage, il est essentiel que le modèle ne laisse pas passer des entités identifiées dans ces bases de données ainsi que les relations entre ces entités, notamment : Pathologies / Symptômes / Principes actifs clés de la pharmacopée / Principes actifs utilisables par pathologie / Nom commercial des médicaments et composition. Développer un modèle par apprentissage en utilisant NLTK ou Spacy est un travail considérable mais qui mériterait d'être exploré.

3.1.4.3 Annotation sémantique (Named Entity Linking - NEL)

Comme vu précédemment, les méthodes permettant d'identifier les entités nommées comportent très généralement deux phases :

- Une phase d'apprentissage au cours de laquelle on utilise un ensemble de textes annotés pour construire un modèle.
- Une phase de prévision dans laquelle on soumet un texte au modèle pour y détecter les entités nommées.

Ce procédé fonctionne assez bien pour des textes courants rédigés avec un vocabulaire courant. Lorsque les textes à analyser utilisent un vocabulaire riche, hors du vocabulaire courant, les résultats de l'analyse seront décevants, sauf à augmenter très considérablement la richesse des textes annotés utilisés pendant la phase d'apprentissage. L'apprentissage de la langue et de ses structures les plus générales peut se faire avec un vocabulaire relativement limité. Dans un deuxième temps, cette compétence de base pourra être étendue par l'utilisation d'un dictionnaire qui permet d'élargir la capacité à décrypter un texte. Dans le champ de l'Entity Linking, les entités identifiées sont reliées à une liste de concepts rassemblés dans une base de connaissance. Cette base de connaissance réuni des termes médicaux avec leurs propriétés (nom, type, identifiant unique…). Un exemple de base de connaissance médicale est l'Unified Medical Language System (UMLS). UMLS est une compilation de

termes médicaux à côté d'autres vocabulaires. Les entités identifiées sont reliées à la base de connaissance en utilisant des techniques comme le TF-IDF et le *text matching*. Malheureusement, UMLS est une base de connaissance développée en langue anglaise. Un problème majeur est la faiblesse de l'offre en matière de base de connaissance médicale en français directement utilisable dans des outils de traitement du langage naturel comme Spacy.

Un effort important reste à faire pour développer des bases de connaissance médicales en français et facilement utilisables. La base de connaissance réunit tous les concepts des terminologies médicales avec, pour chacun d'eux, les attributs clés comme noms, synonymes, définitions, et informations contextuelles qui permettront de lever les possibles ambiguïtés. Les bases de connaissances doivent réunir l'ensemble des concepts médicaux susceptibles d'être reliés à un terme médical présent dans un texte lorsqu'on met en œuvre un traitement de ce type. Les termes médicaux des textes soumis à un traitement de type *Medical Entity linking* seront reliés à un concept de la base de connaissance représentant une maladie, un symptôme, un médicament, une classe thérapeutique, une posologie, un traitement, etc. A chaque entité de nature médicale pourront être associées des métadonnées ou des annotations récupérées dans la liste des attributs du concept dans la base de connaissance. Ces métadonnées ou annotations seront utilisées dans les analyses de traitement du langage naturel selon leur pertinence vis-à-vis de l'objectif poursuivi.

3.1.4.4 Reconnaissance et suppression de l'ambiguïté d'entités nommées (Named Entity Recognition and Disambiguation - NERD)

Une autre étape clé de l'analyse consisterait à éliminer l'ambiguïté dans le texte et lier une identité au bon concept en utilisant le contexte. Par exemple, le mot « froid » peut avoir une signification différente en fonction du contexte : « Conserver le vaccin au froid » ou « J'ai attrapé froid ».

3.1.4.5 L'identification de thèmes (Topic modeling)

La deuxième fonctionnalité principale du NLP est l'extraction de thèmes, plus communément appelée *topic modeling*. Le *topic modeling* peut s'appliquer à toute forme de texte : *mails, tickets, feedbacks*, etc., pour avoir une vision globale des préoccupations des individus. Les principaux modèles de *topic modeling* sont non-supervisés. C'est-à-dire qu'ils n'apprennent pas à lier des messages à un thème donné, ils découvrent eux-mêmes les thèmes. Il est également possible d'associer des messages à des thèmes connus via des modèles de *machine learning*. Des textes connus et clairement associés à ces thèmes sont

utilisés pour entrer dans la phase d'apprentissage. Mais avant d'être analysés, les messages doivent passer par la même préparation que pour l'analyse de sentiments. Ensuite, il existe, là encore, plusieurs méthodologies possibles :

- *Latent Dirichlet Allocation* (LDA) : modèle probabiliste et algorithmique parcourant les messages pour former des groupes de mots qui co-occurrent souvent et ainsi découvrir des thèmes.
- *Latent Semantic Analysis* (LSA) : modèle d'algèbre linéaire décomposant le lien « terme-document » en un lien « terme-thème » + « thème-document ».
- *Non-negative Matrix Factoring* (NMF) : modèle d'algèbre linéaire réalisant le même travail que la LSA pour découvrir des variables latentes, les thèmes. Les deux modèles se différencient par leur méthode de décomposition mais la LSA est plus fréquemment utilisée notamment par son caractère unique et son interprétation un peu plus aisée (grâce à l'importance des thèmes).

3.1.4.6 La Classification dans des thèmes connus

Il est également possible d'associer des messages à des thèmes connus. On utilise des modèles de *machine learning*. Des textes connus et clairement associés à ces thèmes sont utilisés pour entrer en phase d'apprentissage. Plusieurs méthodes sont disponibles, notamment :

- L'algorithme des k plus proches voisins (kNN).
- Les SVM.
- La régression logistique.
- Les réseaux de neurones.

Un des inconvénients de ces approches est la nécessité de fixer le nombre de thèmes à priori. Il existe des critères statistiques pour donner une indication sur le nombre de thèmes optimal mais il reste nécessaire, pour choisir sa méthodologie ou son nombre de thèmes, de s'assurer de la pertinence de l'interprétation des thèmes. Les modèles fournissent le numéro du thème auquel le message est associé, voire l'importance relative de chaque thème pour le message. Quoiqu'il en soit, le thème ne sera qu'une liste de mots auquel il faut associer un nom.

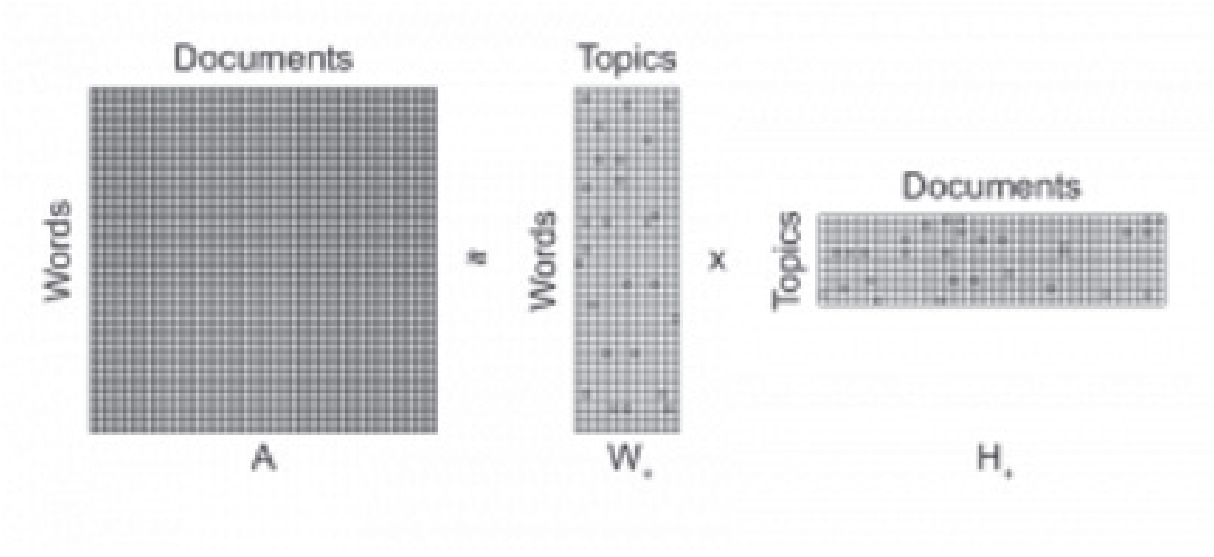

Figure 6 : Illustrant du topic modeling

3.1.4.7 La détection d'évènements (Events Detection)

Cette méthode est dans la continuité du « *Topic modeling* » et de la classification selon des thèmes connus. Un travail de recherche important a été initié dans ces dernières années visant à rechercher des événements dans des collections de *tweets* (70). De nombreuses réflexions permettent d'enrichir l'approche initiale principalement basée sur une analyse de fréquence des mots et la recherche d'anomalies statistiques faisant penser à un choc exogène. La tâche consistant à analyser des documents textuels pour détecter des évènements est appelée détection d'évènements. C'est un problème de découverte, c'est-à-dire la recherche de structures sémantiques nouvelles dans une collection de textes. Il s'agit donc de trouver de nouveaux événements ou de reconnaître des événements déjà identifiés. Quelques remarques fondamentales :

- La plupart des textes n'ont aucune relation à un quelconque évènement.
- Le nombre de textes faisant référence à un événement donné peut être très variable. L'importance d'un évènement ne doit pas être uniquement basé sur le nombre de textes qui y font référence. En effet, cela risque de masquer de nombreux évènements importants qui apparaissent à bas bruit au travers des textes.
- Le dernier challenge consiste à choisir une définition du terme « Événement » adapté au contexte particulier des média sociaux.
- Dans le contexte des média sociaux, et particulièrement de Tweeter, une proposition de définition de la notion d'événement a été proposée par Dou et al.,

2012 (71)(72) : « *An occurrence causing change in the volume of text data that discusses the associated topic at a specific time. This occurrence is characterized by topic and time, and often associated with entities such as people and location* ».

Deux catégories principales d'événements ont été mises en évidence dans une étude récente (73) :

- « *Close domain »* principalement dédié à la recherche d'évènements bien spécifiques qui peuvent être définis à priori comme un tremblement de terre ou une épidémie de grippe.
- « *Open domain* » dont l'objectif est de détecter de nouveaux évènements, sans connaissance à priori des évènements recherchés (74).

Le texte, après traitement initial, est converti en un graphe constitué de nœuds et de relations. A chaque relation est attaché un poids qui est le nombre de co-occurrences. Un poids élevé montre une association forte entre deux termes dans le test. Une fois l'ensemble des entités et des relations mises en évidence on supprime les relations dont la fréquence d'occurrence est faible. On fait ainsi apparaître des sous-ensembles indépendants d'entités et de relations qui n'entretiennent plus de relation avec les îlots voisins. Ces sous-ensembles indépendants correspondent en général à des évènements spécifiques.

3.2 Matériel et méthodes

Dans cette section, le cadre de l'étude sera précisé et les analyses menées tout au long de l'expérimentation seront présentées. L'étude peut être scindée en six sous-étapes, seule les cinq premières seront abordées dans cette partie :

1. Extraction de la donnée
2. Nettoyage de la donnée
3. Analyse de la fréquence des mots, calcul de corrélation entre deux termes et extrapolation à des bi-grammes
4. *Machine learning* (analyse de similarité, analyse de sentiments)
5. Analyses des réseaux de neurones convolutifs entraînés sur des nuages de mots
6. Extraction des effets indésirables rapportés et analyse de leur occurrence

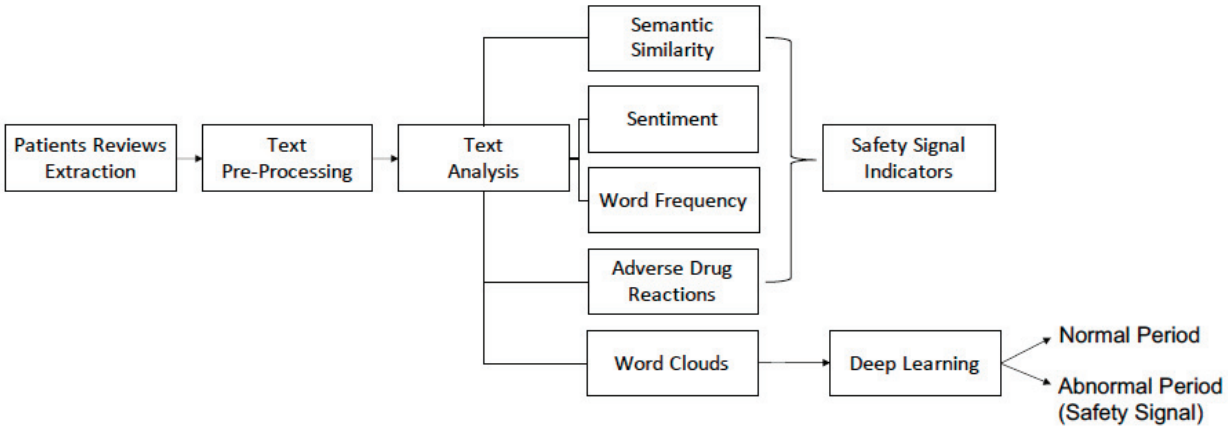

Figure 7 : Diagramme du déroulement de l'étude conduite

Afin de limiter les sources d'erreurs pour obtenir la méthodologie la plus précise possible, il est choisi de se focaliser sur un unique médicament ainsi qu'une unique source de données. C'est le forum Doctissimo® et le médicament Levothyrox® qui sont retenus.

Doctissimo® est choisi car il est le forum de santé le plus utilisé en France par les patients consommateurs de médicaments. En effet, il se classe en première position avec 61% des utilisateurs de réseaux sociaux de santé, largement devant le forum « santé médecine le journal de femmes » qui est second avec 5% des utilisateurs. D'autres sources d'informations existent comme Twitter® qui est le site internet le plus utilisé au monde par les patients consommateurs de médicaments. Twitter® rassemble 52% de ces patients contre 27% pour

tous les forums de discussion confondus. Cependant, l'accès aux données de Twitter® est payant, c'est la raison pour laquelle Twitter n'est pas retenu (75). Le Levothyrox® est quant à lui choisi car c'est un médicament pour lequel des données récentes et facilement accessibles sont disponibles. En effet, l'affaire du Levothyrox® de 2017 a fortement été realyée dans les médias, et notamment sur les forums de discussions.

Toute la partie algorithmique décrite dans cette thèse est réalisée en langage Python. Ce langage est choisi car il est très flexible et polyvalent. Il peut à la fois servir à développer des logiciels, des applications mais est aussi le langage le plus populaire dans le *big data*, l'application de calculs mathématiques ou de *machine learning*. Le *big data* est défini comme un ensemble très volumineux, véloce et diversifié de données qu'un outil classique de gestion de base de données ne peut manipuler.

Le *machine learning* est, quant à lui, une technologie permettant à l'ordinateur d'apprendre une certaine tâche sans qu'il ne soit programmé à cet effet. L'ordinateur est donc entraîné via des jeux de données d'une taille telle, qu'ils sont qualifiés comme appartenant au *big data*. Aussi, l'utilisation du langage Python est favorisée du fait de l'importante quantité de ressources et d'exemples d'applications des algorithmes en ligne et en libre accès. Ils sont donc une vaste source d'inspiration au cours de ce travail contrairement aux autres langages qui sont moins intuitifs à manipuler et donc moins utilisés.

Concernant l'environnement d'exécution des algorithmes, c'est Google Collaboratory® qui est sélectionné. De nombreux logiciels existent mais celui-ci permet une flexibilité de travail grâce à la possibilité de travailler à plusieurs en même temps sur les mêmes fichiers et de les partager facilement à d'autres utilisateurs. La façon dont le travail collaboratif est permis sur Google Collaboratory® est la même que les autres logiciels de la suite Google® type Google Doc®, Google Slide®, Google Sheet®, etc.

3.2.1 Extraction de la donnée

L'extraction de la donnée correspond à la première étape de l'étude. L'objectif est de récupérer toute la donnée nécessaire à l'étude. Pour ce faire, un algorithme de *scraping* est mis au point. Le *scraping* est une technique d'extraction du contenu de sites web via un programme informatique dans le but de l'utiliser dans un contexte différent. Ce processus de *web scraping* est similaire à un copier/coller de l'information, depuis Doctissimo®, dans une base de données telle que Excel®.

L'extraction est réalisée sur le forum « thyroïde et problèmes endocriniens » avec le mot clé « levothyrox ». Il est décidé de prendre seulement en considération les informations de ce forum avec un mot clé défini pour limiter la quantité de données extraites. En effet, lors de l'extraction, Doctissimo® bloque la tâche de *scraping* lorsque l'ordinateur extrait plus de 8 000 fils de discussions. Le site internet détecte que l'activité menée n'est pas réalisée par un humain mais par une machine. Pour contourner cette restriction, il est possible de créer un outil de gestion de bases de données mais sa mise en place nécessite beaucoup de temps. Aussi, comme la quantité de donnée serait plus importante, la durée d'exécution des algorithmes des étapes suivantes serait fortement rallongée et les ordinateurs utilisés ne sont pas capables de gérer une masse de donnée si importante. Grâce à cette focalisation sur les résultats d'un mot clé sur un forum particulier, la base de données totalise 110 260 commentaires sur 20 ans.

L'algorithme suis les étapes suivantes pour extraire la donnée :

1. L'algorithme se rend sur le site internet du forum Doctissimo®.
2. Il sélectionne le forum « thyroïde et problèmes endocriniens », entre le mot clé « levothyrox » et lance la recherche.

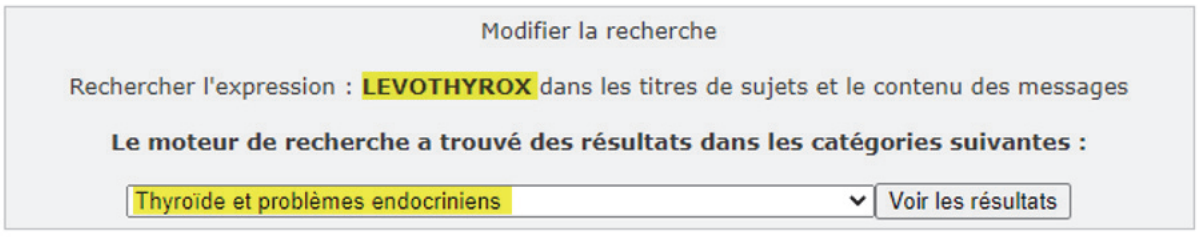

Figure 8 : Visualisation de la recherche effectuée automatiquement par l'algorithme

Les résultats de la recherche sont classés par sujets pour que l'utilisateur puisse retrouver plus rapidement le fil de discussion qu'il recherche. Au sein de chaque sujet sont retrouvés les commentaires que les internautes échangent.

3. L'algorithme va récupérer toutes les *Uniform Ressource Locator* (URL) correspondant à chaque sujet. Au total, 7650 sont enregistrés (juste en dessous de la limite à 8 000).
4. Pour les 7 650 URL de sujets extraits, la machine consulte chaque URL afin d'accéder au code *HyperText Markup Language* (HTML) de la page et copier/coller dans un fichier Excel :
    a. La date du commentaire.
    b. Le pseudo de la personne qui a écrit le commentaire.
    c. Le texte du commentaire.
    d. Le lien URL du commentaire.

La création d'une page web nécessite deux types de codes : le code HTML et le code *Cascading Style Sheets* (CSS). Le code HTML renferme l'information et la structure de la page web (titres, sous-titres, paragraphes). Le code CSS permet la mise en forme de la page web (couleurs du texte, insertion d'images, etc.). Pendant cette phase de *scraping*, un pré-nettoyage des commentaires est effectué avant l'écriture dans la base de données Excel® pour permettre un découpage efficace et optimisé :

- Suppression des apostrophes.
- Remplacement des virgules par des espaces.
- Remplacement des sauts de lignes par des espaces.
- Remplacement des émoticônes par des espaces.
- Suppression des liens insérés.
- Remplacement des images par des espaces.
- Suppression de balises particulières : br, span, table, strong, div, etc…
- Suppression des accents.

Les précédentes étapes décrites permettent l'obtention d'une base de données sous la forme d'un fichier Excel® au format *Comma-Separated Values* (CSV) dans lequel une ligne correspond à un commentaire et dont les caractéristiques de chaque commentaire (date, pseudo, texte et URL) sont séparés par des virgules. Ce format CSV consiste à représenter des données tabulaires sous forme de texte avec pour séparateur entre les données de chaque colonne, une virgule.

date,pseudo,texte,url

21/03/2021,freesia53,bonjour j'ai une thyroïdite d'hashimoto,https://forum.doctissimo.fr/…

17/03/2021,patrickB,bonjour à tous j'ai une hyperdthyroïdie,https://forum.doctissimo.fr/…

13/03/2021,Susaaanne,bonjour pourquoi quel est votre medecin?,https://forum.doctissimo.fr/…

Figure 9 : Visualisation de la donnée extraite en format csv délimité par des virgules

Tableau 3 : Visualisation de la donnée extraite sous forme de tableau

date	*pseudo*	*texte*	*url*
21/03/2021	freesia53	bonjour j'ai une thyroïdite d'hashimoto	https://forum.doctissimo.fr/…
17/03/2021	patrickB	bonjour à tous j'ai une hyperdthyroïdie	https://forum.doctissimo.fr/…
13/03/2021	Susaaanne	bonjour pourquoi quel est votre medecin?	https://forum.doctissimo.fr/…

Ce format de base de données est facile d'utilisation et utilisable pour des bases de données peu volumineuses. En effet, Excel® limite la taille d'affichage des fichiers CSV à 1 048 576 lignes mais il est possible d'ouvrir ceux qui sont plus volumineux avec d'autres logiciels. Avec les 110 260 commentaires extraits de Doctissimo®, Excel® est amplement suffisant pour manipuler la base de données.

3.2.2 Nettoyage de la donnée

Une fois la base de données obtenue, il s'agit de la nettoyer pour supprimer tous les motifs indésirables. Dans un premier temps, il est nécessaire d'importer la base de données dans Google Collaboratory® depuis Google Drive® où elle est stockée sous le nom « dataset_doctissimo_22_03_2020 ». Il est décidé de stocker la donnée sur Google Drive® afin d'apporter de l'agilité au projet grâce à la possibilité de travailler de façon collaborative en temps réel sur les mêmes fichiers et de les partager avec toutes les parties prenantes. Pour la travailler et qu'elle soit clairement lisible, la librairie pandas sera utilisée. Pandas est l'abréviation anglaise « données de panel » (*panel data*), elle permet d'utiliser des tableaux de données appelés *dataframes*, qui sont des outils de manipulation de la donnée (stockée sur la mémoire vive). Dans ce travail, les *dataframes* contiennent les informations de la base de données Excel®.

	date	user	text	url
0	21/03/2020	freesia53	b"Bonjour. Je suis suivie pour une thyroidite ...	https://forum.doctissimo.fr/sante/thyroide-pro...
1	13/03/2020	NotYourMajesty	b"Bonjour a tous ! Ayant une hypotyroidie je ...	https://forum.doctissimo.fr/sante/thyroide-pro...
2	13/03/2020	petitbouchon	b"Bonjour Pourquoi votre medecin vous prescrit...	https://forum.doctissimo.fr/sante/thyroide-pro...
3	13/03/2020	Susanne in F	b"L'equivalent de 25 T3 serait 100 T4 . #034;...	https://forum.doctissimo.fr/sante/thyroide-pro...
4	30/09/2007	ale14za	b"Bonjour ma fille agee d'un mois et demi a ...	https://forum.doctissimo.fr/sante/thyroide-pro...

Figure 10 : Dataframe pandas de la base de données après extraction

Avant de rentrer dans le vif du sujet, il s'agit, pour faciliter la compréhension, de définir certains termes généraux correspondants à différents outils de développement informatiques utilisés :

- Librairie / bibliothèque : Ensemble de fonctions, regroupées dans un ensemble appelé librairie ou bibliothèque afin de les utiliser sans avoir à les réécrire. Chaque librairie est spécialisée dans un certain domaine (calculs, graphisme, formatage de texte, génération de documents, etc.). Après importation des librairies, il est possible d'utiliser toutes leurs fonctions.
- Module : Fichier contenant du code qui peut être de toute sorte (librairie, fonction, variable, etc.), qu'il sera possible d'importer dans un autre script de code. L'intérêt de ces modules est de rendre plus compréhensible le code final en évitant d'avoir à

réécrire le contenu du module à chaque fois qu'il est nécessaire de l'utiliser. En langage python, le fichier comporte l'extension « .py ».

- Fonction : Ensemble d'instructions permettant de réaliser une tâche précise qui peut être utilisée autant de fois que souhaité en l'exécutant. En développement informatique, exécuter une fonction se dit « appeler » la fonction car elle porte un nom. La syntaxe ce ces fonctions est toujours la même et respecte celle présentée sur la figure ci-dessous :

```
def nom_fonction(liste de paramètres):
    bloc d'instructions
```

Figure 11 : Syntaxe d'une fonction

A titre d'exemple, une fonction utilisée dans cette étude est celle qui permet de supprimer les caractères correspondant à de la ponctuation dans les commentaires.

- Variable : Entité qui contient une valeur pendant une période limitée. La valeur qu'elle contient est stockée durant le temps d'exécution du script, cette valeur peut changer au cours de l'exécution.

Par exemple, une variable utilisée lors de cette étude s'appelle « extraction_sorted ». Elle correspond, lors de sa création, à la base de données entière dans laquelle les commentaires sont triés par date. A la seconde étape où cette variable est utilisée, une fonction permettant de convertir tous les caractères des commentaires en minuscules est appelée sur cette variable. D'autres fonctions sont par la suite appelées comme la suppression de la ponctuation ou la suppression de mots indésirables. A la fin de l'exécution du script, la variable a subi une succession de modifications, la valeur qu'elle contient varie donc au fil de l'exécution.

Différentes librairies Python sont importées pour réaliser les tâches de nettoyage :

- Pandas : Permet d'utiliser des tableaux de données appelés *dataframes*, qui sont des outils de manipulation de la donnée (stockée sur la mémoire vive).
- Matplotlib : Destinée à tracer et visualiser des données sous forme de graphiques.
- Spacy : Destinée à faire du Traitement Naturel du Langage. Le NLP étant une technique d'intelligence artificielle permettant aux ordinateurs de lire, déchiffrer, comprendre et donner un sens au langage humain. C'est ce type d'algorithmes qui

sont utilisés par les applications de reconnaissances vocales type Amazon Alexa® ou Apple Siri®.

- NLTK : Destinée au NLP comme Spacy.

Différents modules ont été importés pour les tâches de nettoyage :

- re : Module des expressions régulières. En fonction de l'expression régulière écrite, le module reconnaît tous les motifs (lettres, chiffres, caractères spéciaux) correspondant à l'expression régulière afin d'appliquer des modifications sur celles-ci. Par exemple, l'expression régulière « ^a-zA-Z0-9 » va rechercher tous les caractères autres que les lettres majuscules, minuscules et les chiffres. Le symbole « ^ » signifie « tout sauf ».
- string : Permet de définir des données comme étant des chaînes de caractères afin d'appliquer des fonctions utilisables sur des chaînes de caractères.
- csv : Permet d'importer et d'exporter des fichiers CSV, ce sont ces derniers qui sont utilisés pour la gestion de la base de données.
- datetime : Permet de manipuler les dates et le temps. Il accepte différents types de dates et permet d'accéder au jour, au mois, à l'année souhaitée, mais également à l'heure, la minute, la seconde ou la microseconde si l'information est contenue dans la date.

Concernant l'étendue de la base de données, elle va de 2000 à 2020. Il s'agit, dans un premier temps, de classer la donnée par ordre chronologique et de connaître l'évolution de la fréquence des commentaires postés sur le forum au cours du temps. Le graphique de la figure suivante est ainsi tracé (à l'aide de la librairie matplotlib).

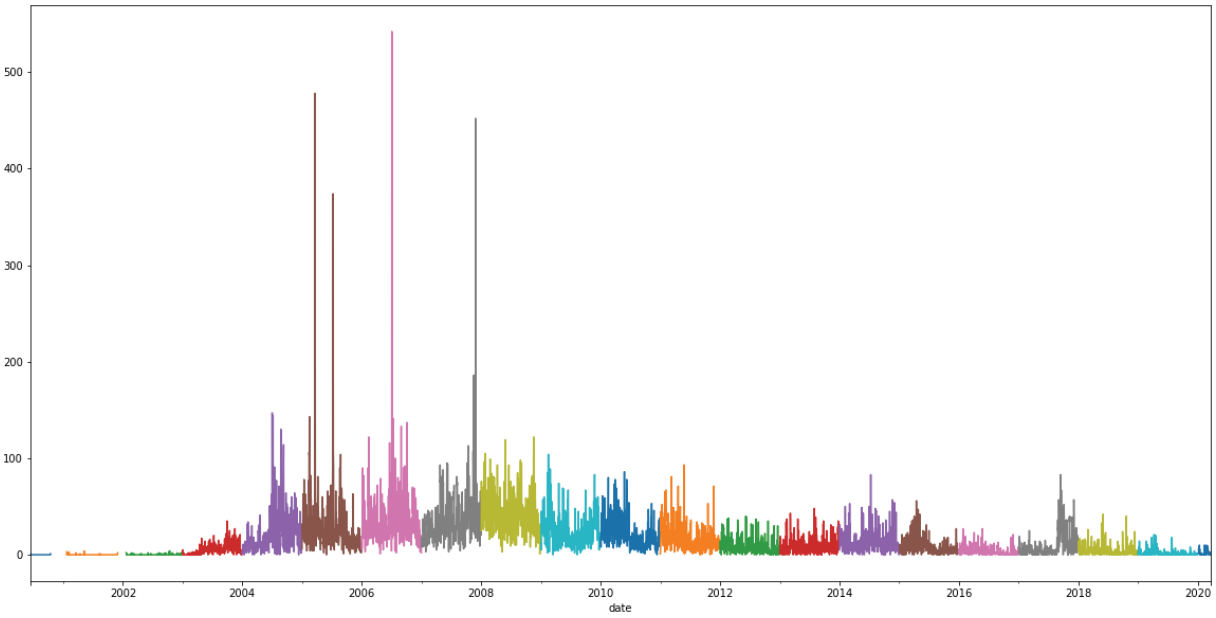

Figure 12 : Nombre de commentaires postés en fonction du temps sur la période 2000 à 2020

La figure 12 met en évidence une fréquentation du forum plus importante entre 2004 et 2012 qu'après 2012. Il est difficile d'expliquer les causes de cette diminution de fréquentation, hormis l'hypothèse qu'une partie des utilisateurs de Doctissimo® auraient migrés vers d'autres canaux de communication tels que les réseaux sociaux généralistes. Entre 2012 et 2016 est observé une occurrence de commentaires plus importante qu'entre 2016 et 2020 mais moins importante qu'avant 2012. L'affaire du Levothyrox® ayant eu lieu en 2017, il est décidé de restreindre la base de données de travail à la période 2016-2020 pendant laquelle l'occurrence des commentaires est stable.

Un fois la donnée restreinte à la période 2016-2020, des étapes de nettoyage sont effectuées. L'objectif étant de standardiser au maximum chaque mot afin que ceux de la même famille se ressemblent au maximum. Cette démarche augmente la performance et la précision des algorithmes de NLP qui seront par la suite utilisés. En effet, l'algorithme sera plus efficace si les mots « mange », « manger », « mangez », « mangé », « mangée », « mangées », sont tous regroupés sous la même orthographe.

Ainsi, six fonctions principales sont créées pour nettoyer les commentaires. Les fonctions sont appliquées dans l'ordre suivant :

1. « doctissimo_sort_range » : Permet de trier les commentaires dans la base de données par ordre chronologique descendant.

2. « dataframe_preprocessing » : Permet de réaliser les grandes étapes de nettoyage du corps du texte :

   a. Suppression des accents.

   b. Suppression des caractères spéciaux, c'est-à-dire tout caractère autre que les lettres de A à Z et les chiffres de 0 à 9.

   c. Conversion de toutes les lettres en minuscules.

   d. Suppression des motifs indésirables. Ceux qui ont été identifiés dans la base de données sont « #034 » et « #039 ».

   e. Suppression des gifs (images animées). Comme la base de données est extraite depuis le code HTML, les gifs sont identifiables par leur lien finissant par le motif « .gif ».

   f. Suppression des liens grâce à leur identification via le motif « http » ou « https ».

   g. Suppression des chiffres ou nombres contenus dans des mots.

   h. Suppression des chiffres et nombre hormis les dates.

   i. Suppression des alinéas et sauts de lignes.

   j. Suppression des caractères isolés.

   k. Suppression des espaces multiples.

Au cours de ces étapes, les modules regex et string sont utilisés pour passer en revue toute la base de données à chaque étape de la fonction et appliquer les modifications de texte en conséquence. La figure suivante permet d'illustrer le fonctionnement de la fonction « dataframe_preprocessing » et l'utilisation des modules regex et string.

```
def dataframe_preprocessing(df):
        df['text'] = df['text'].str.normalize('NFKD').str.encode('ascii',errors='ignore').str.decode('utf-8') # Remove accent
        df['text'] = [re.sub(r'[^a-zA-Z0-9 ]', ' ', str(x)) for x in df['text']] # Remove special characters and punctuation
        df['text'] = df['text'].str.lower() # Convert ['text'] string to lowercase
        df['text'] = df['text'].str.replace('#034', '', regex=True) # Remove #034 pattern
        df['text'] = df['text'].str.replace('#039', '', regex=True) # Remove #039 pattern
        df['text'] = df['text'].str.replace(".*gif", "", regex=True) # Remove all gifs
        df['text'] = df['text'].str.replace("http.* ", "", regex=True) # Remove http links
        df['text'] = df['text'].str.replace("https.* ", "", regex=True) # Remove https links
        df['text'] = [re.sub(r'[A-Za-z]+\d+|\d+[A-Za-z]+','', str(x)) for x in df['text']] # Delete numbers between alphabetic chars
        df['text'] = [re.sub(r'\b(?!(\D\S*|[12][0-9]{3})\b)\S+\b','', str(x)) for x in df['text']] # Numbers except dates
        df['text'] = df['text'].str.replace('\n',' ', regex=True).replace('\t',' ', regex=True) # Remove line breaks and tabulations
        df['text'] = [re.sub(r'(^| ).( |$)', ' ', str(x)) for x in df['text']] # Remove single characters
        df['text'] = [re.sub(r'\s+', ' ', str(x)) for x in df['text']] # Delete multiple spaces
        return df
```

Figure 13 : Fonction « dataframe_preprocessing »

3. « doctissimo_words_improvment » : Permet de corriger l'orthographe des mots les plus fréquents de la base de données. Pour se faire, l'affichage de nuages de mots par mois, semaine et jour a été réalisé pour se rendre compte des différents

orthographes utilisées. Le nuage de mots présente les mots les plus fréquents du corpus et les représente dans différentes tailles selon leur fréquence d'apparition. Ainsi, plus le mot est gros et plus il est utilisé dans la partie de la base de sonnées correspondante (mois, semaine, jour).

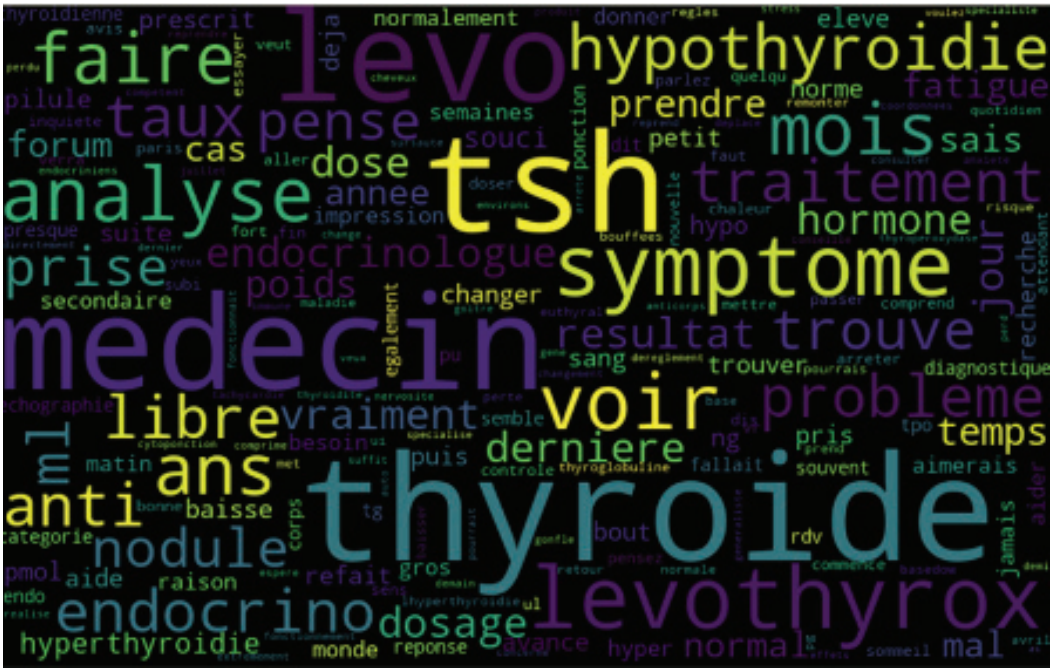

Figure 14 : Nuage de mots de septembre 2016, après suppression des stopwords et avant exécution de la fonction « doctissimo_words_improvement »

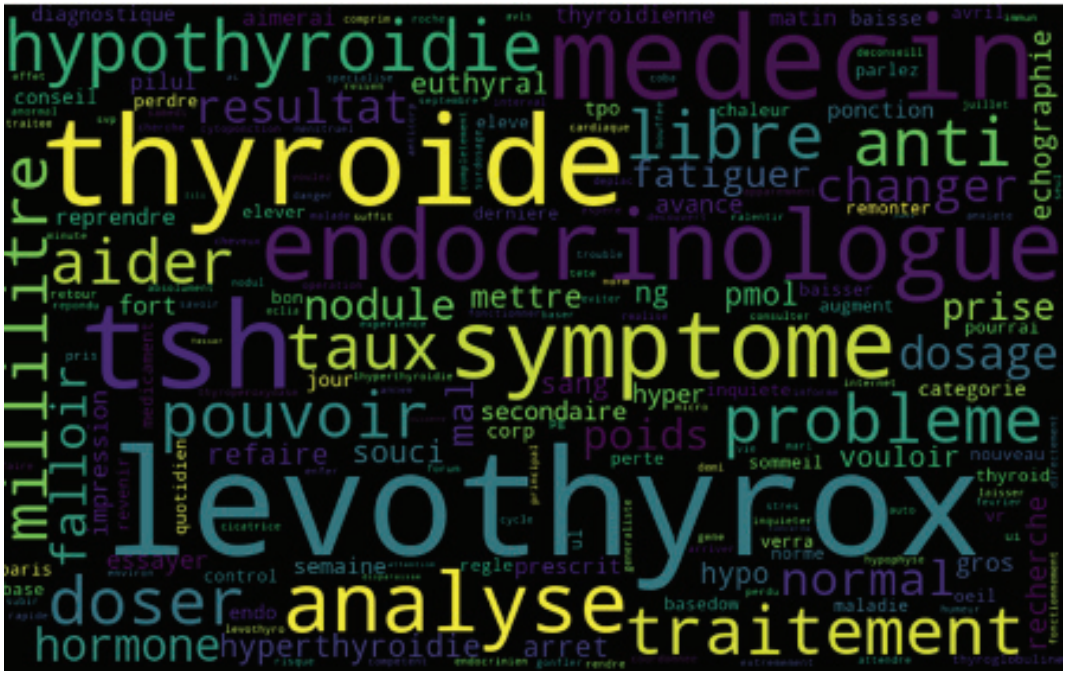

Figure 15 : Nuage de mots de septembre 2016 après exécution de toutes les fonctions jusqu'à « dataframe_duplicata_less3words »

Les nuages de mots représentés dans les figures précédentes présentent les résultats respectivement avant l'exécution de la fonction « doctissimo_words_improvment » et après l'exécution de la fonction « dataframe_duplicata_less3words », donc après nettoyage complet de la base de données. Il est possible d'observer que de nombreuses orthographes différentes d'un même mot sont présentes dans la première figure, ainsi que de nombreux mots indésirables qui ne présentent pas d'intérêt pour l'analyse. Les méthodes utilisées pour supprimer ces mots sont détaillées par la suite. La figure ci-dessous permet d'illustrer le fonctionnement de la fonction « doctissimo_words_improvment ».

```python
def doctissimo_words_improvment(df):
        df['text'] = df['text'].str.split() # Split string
        i=0
        for line in df['text']:
                new_line = []
                for word in line:
                        for imprv in words_improvment:
                                if word in imprv and word != imprv[0]:
                                        word = imprv[0]
                                        break
                        new_line.append(word)
                df.at[i, 'text'] = new_line
                i += 1
        df['text'] = df['text'].apply(' '.join) # Join string
        return df
```

Figure 16 : Fonction « doctissimo_words_improvement »

Cette fonction permet, grâce à la liste « words_improvement », présentée en partie dans la figure 17, de transformer les différentes orthographes fréquemment observées d'un mot en une orthographe commune. L'objectif est de redonner au mot le poids qu'il devrait avoir dans le corpus. Pour chaque ligne de la liste « words_improvement », les mots en deuxième position et plus (levo, levothyro et levotyrox pour la première ligne) prennent l'orthographe du premier mot de cette même ligne (levothyrox pour la première ligne).

```python
words_improvment = [['levothyrox', 'levo', 'levothyro', 'levotyrox'],
                    ['euthyrox', 'leuthyrox', 'eutyrox', 'leutyrox'],
                    ['lthyroxine', 'lthyroxin', 'ltyroxine', 'ltyroxin'],
                    ['hypothyroidie', 'lhypothyroidie', 'hypotyroidie', 'lhypotyroidie'],
                    ['comprime', 'comprim'],
                    ['cytomel', 'cynomel'],
                    ['controle', 'control'],
                    ['changer', 'change', 'chang', 'changement'],
                    ['allemagne', 'allemand'],
                    ['generaliste', 'generalist'],
                    ['arret', 'arreter', 'arrete'],
                    ['excipient', 'excipients'],
                    ['laboratoire', 'laboratoir'],
                    ['poids', 'poid'],
                    ['hormone', 'hormon', 'dhormone', 'dhormon', 'lhormone', 'lhormon'],
                    ['correcte', 'correct'],
                    ['courche', 'coucher', 'chouchee'],
                    ['neomercazole', 'neomercazol'],
                    ['enceinte', 'enceint'],
                    ['ancien', 'ancienne', 'lancien', 'lancienne'],
                    ['francais', 'fraincaise', 'francai'],
```

Figure 17 : Apperçu d'une partie de la liste « words_improvement »

4. « dataframe_stopwords_wtd » : Permet d'épurer le texte des mots indésirables. Trois étapes principales fractionnent cette fonction :

a. La création d'une liste de mots indésirables à exclure de la base de données. Ces mots sont qualifiés d'indésirables car ils sont très récurrents dans les commentaires et sont neutres, ils n'apportent aucune plus-value en termes de détection du signal. Ces termes ont, comme pour la liste « words_improvement », été identifiés manuellement dans le texte. Ces mots sont regroupés au sein d'un fichier Excel® intitulé exclusion.csv. La figure ci-dessous présente une partie des mots présents dans le fichier exclusion.csv afin de se rendre compte des mots qui sont supprimés.

actuellement	bonsoir	corp	demain
aller	bout	cote	demande
annee	bref	coucou	demander
arrive	ca	coup	dernier
arriver	cas	courage	derniere
attendre	cause	courir	detre
aussi	cest	croi	devoir
avi	ceter	crois	di
avoir	chose	daccord	dis
ba	chri	daller	dit
besoin	comme	dapre	dire
bientot	commence	dautre	doc
bisou	comprend	davance	donc
bisous	comprendre	davoir	donne
bon	compte	deja	donner
bonjour	connai	dejer	droit

Figure 18 : Liste non exhaustive des mots présents dans le fichier « exclusion.csv »

b. La suppression des *stopwords* ou « mots vides » en français. Ils sont les mots tellement fréquents qu'ils apportent beaucoup de bruit de fond dans les analyses. Un mot est qualifié de vide lorsqu'il n'est pas discriminant et ne permet pas de distinguer les commentaires les uns par rapport aux autres. Ils ne présentent donc aucun intérêt de les conserver. Des listes de *stopwords* existent en accès libre sur internet, c'est l'une d'entre elles qui a été utilisée. Les *stopwords* les plus fréquents en français sont « le », « la », « les », « de », « du », « ce », etc.

5. La suppression de *stopwords* supplémentaires. En effet, ceux qui ont été supprimés ont permis de supprimer du bruit de fond mais après analyse de la base de données et des nuages de mots, de nombreux *stopwords* persistent. Une liste de *stopwords* additionnelle est créée en croisant plusieurs listes en accès libre sur internet et en analysant la base de données et les nuages de mots. Cette liste s'appelle « additional_stopwords », elle permet un nettoyage bien plus précis. La figure suivante présente un aperçu des *stopwords* supplémentaires qui sont supprimés.

```
additional_stopwords = ['a', 'abord', 'afin', 'ah', 'ai', 'ainsi', 'allaient', 'allo', 'allô', 'allons', 'alors', 'apres', 'après', 'assez', 'attendu', 'aucun',
                        'b', 'bonjour', 'bonsoir', 'bah', 'beaucoup', 'bien', 'bigre', 'bon', 'boum', 'br', 'bravo', 'brr', 'brrr',
                        'ca', 'ça', 'car', 'ceci', 'cela', 'celle', 'celle-ci', 'celle-la', 'celle-là', 'celles', 'celles-ci', 'celles-la', 'celles-là', 'celui'
                        'da', 'debout', 'debut', 'début', 'dedans', 'dehors', 'dela', 'delà', 'depuis', 'derriere', 'derrière', 'dés', 'dès', 'desormais', 'déso
                        'e', 'effet', 'eh', 'elle-meme', 'elle-même', 'elles', 'elles-memes', 'elles-mêmes', 'encore', 'entre', 'envers', 'environ', 'ès', 'essa
                        'f', 'facon', 'façon', 'fais', 'faisaient', 'faisant', 'fait', 'faites', 'feront', 'fi', 'flac', 'floc', 'fois', 'font', 'force', 'fumes
                        'g', 'gens',
                        'h', 'ha', 'haut', 'he', 'hé', 'hein', 'helas', 'hélas', 'hem', 'hep', 'hi', 'ho', 'hola', 'holà', 'hop' ,'hormis', 'hors' ,'hou', 'houp
                        'i', 'ici', 'importe',
                        'jusqu', 'jusqua', 'jusque', 'juste',
                        'k',
                        'là', 'laquelle', 'las', 'lequel', 'lès', 'lesquelles', 'lesquels', 'leurs', 'longtemps', 'lorsque', 'lui-meme', 'lui-même',
                        'maint', 'maintenant', 'malgre', 'malgré', 'meme', 'memes', 'mêmes' ,'merci', 'mien', 'mienne', 'miennes', 'miens', 'mille', 'mince', 'm
                        'na', 'nai', 'nas', 'neanmoins', 'néanmoins', 'neuf', 'neuvieme', 'neuvième', 'ni', 'nombreuses', 'nombreux', 'nommes', 'nommés', 'non',
                        'o', 'onsoir', 'onjour', 'ô', 'oh', 'ohe', 'ohé', 'ole', 'olé', 'olle', 'ollé', 'onze', 'onzieme', 'onzième', 'ore', 'où', 'ouf', 'ouias
                        'p', 'paf', 'pan', 'parce', 'parmi', 'parmis', 'parole', 'partant', 'particulier', 'particuliere', 'particulière', 'particulierement', '
                        'q', 'quand', 'quant', 'quant-a-soi', 'quant-à-soi', 'quant-a-soit', 'quanta', 'quarante', 'quatorze', 'quatre', 'quatre-vingt', 'quatri
                        'r', 'revoici', 'revoila', 'revoilà', 'rien',
                        'sacrebleu', 'sans', 'sapristi', 'sauf', 'seize', 'selon', 'sept', 'septieme', 'seulement', 'si', 'sien', 'sienne', 'siennes', 'siens',
                        'tac', 'tandis', 'tant', 'té', 'tel', 'telle', 'tellement', 'telles', 'tels', 'tenant', 'tic', 'tien', 'tienne', 'tiennes', 'tiens', 'to
                        'u', 'unes', 'uns',
                        'v', 'va', 'vais', 'valeur', 'valeurs', 'vas', 've', 'vé', 'vers', 'via', 'vif', 'vifs', 'vingt', 'vivat', 'vive', 'vives', 'vlan', 'voi
                        'w',
                        'x',
                        'z', 'zut']
```

Figure 19 : Liste « additional_stopwords »

6. « dataframe_lemmatization » : Permet d'affiner l'homogénéisation des mots du corpus par rapport à celle qui est faite lors de l'exécution de la fonction « doctissimo_words_improvment ». Cette fonction permet d'appliquer une lemmatisation du corpus. La lemmatisation consiste à réduire les mots à leur lemme commun pour diminuer les variations d'orthographes entre les mots qui ont un sens similaire. En guise d'exemple, la lemmatisation transforme les mots « petit », « petite », « petits », « petites » en leur lemme commun qui est « petit ». Lors de l'exécution de cette fonction, c'est grâce à la librairie Spacy que la lemmatisation est effectuée. La figure ci-après illustre la fonction permettant d'appliquer la lemmatisation.

```
def dataframe_lemmatization(df):
        # Lemmatization with nlp_fr
        df['text'] = df['text'].apply(lambda x: [y.lemma_ for y in nlp_fr(x)]).apply(' '.join)
        return df
```

Figure 20 : Fonction « dataframe_lemmatization »

Une autre technique peut être utilisée et est assez proche de la lemmatisation. C'est le *stemming* ou racinisation en français. Elle consiste à transformer les mots en leur radical commun. A la différence du lemme de la lemmatisation qui correspond à un terme issu de l'usage usuel, le radical correspond souvent à aucun mot. Par exemple les mots "chercher", "chercha", "cherche", etc. seront transformés en "cherch" via le *stemming* et en "chercher" via la lemmatisation.

7. « dataframe_duplicates_less3words » : Permet de réaliser 3 tâches de nettoyage :
    a. Supprimer les commentaires qui ont moins de trois mots. Une phrase ayant au minimum trois mots (un sujet, un verbe et un complément), les commentaires qui ne sont pas des phrases sont supprimés pour s'affranchir des commentaires parasites.
    b. Suppression des commentaires dupliqués. Il arrive que des utilisateurs publient plusieurs fois le même message.
    c. Suppression des lignes de la base de données où au moins une cellule est vide. Ces commentaires sont identifiables par le motif NaN (Not a Number) ou NaT (Not a Timstamp) signifiant respectivement « pas un nombre ou non numérique » et « pas une date ». La figure suivante permet d'illustrer le fonctionnement de la fonction « dataframe_duplicates_less3words ».

```
def dataframe_duplicates_less3words(df):
        words_count = df['text'].str.count(' ') + 1 # 'text' characters counter
        df['words_count'] = words_count # Add words_count column on the dataframe
        before_deleting = df['text'].count()
        print('\n******************\nNumber of rows BEFORE deleting the rows that contain less than 3 words : ' + str(before_deleting) + ' rows')
        df.drop(df[df['words_count'] < 3].index, inplace = True) # Remove rows that contain less than 3 words
        after_deleting = df['text'].count()
        print('Number of rows AFTER deleting the rows that contain less than 3 words : ' + str(after_deleting) + ' rows')
        diff_deleting = before_deleting - after_deleting
        print('Difference : ' + str(diff_deleting) + ' rows')
        df.drop_duplicates() # Drop duplicates rows
        df.dropna() # Drop the rows even with single NaN or single missing values.
        after_del_duplicates = df['text'].count()
        print('Number of rows after deleting duplicates : ' + str(after_del_duplicates) + ' rows')
        diff_duplicates = after_deleting - after_del_duplicates
        total_delete = diff_deleting + diff_duplicates
        print('Difference : ' + str(diff_duplicates) + ' rows')
        print('Total number of deleted rows : ' + str(total_delete) + '\n******************')
        return df
```

Figure 21 : Fonction « dataframe_duplicates_less3words »

Une fois ces fonctions créées, il s'agit de les appliquer successivement sur la base de données pour effectuer le nettoyage. La base de données est ensuite exportée sur Google Drive® dans un nouveau fichier CSV portant le nom « dataset_doctissimo_updated.csv » où toute la donnée nettoyée de l'étude est stockée. L'étape suivante consiste à réaliser les premières analyses textuelles des commentaires pour en faire ressortir les motifs intéressants.

3.2.3 Analyse de la fréquence des mots, calcul de corrélation entre deux termes et extrapolation à des bi-grammes

La première voie explorée est l'analyse de la fréquence des mots dans les commentaires. Grâce à la structuration de la base de données lors du *scraping* et l'utilisation de la bibliothèque pandas, il est facile d'accéder à la fréquence des mots dans les messages. Elle peut être établie sur une période (jour, semaine, mois, année) pour analyser leur évolution.

Différentes librairies Python sont importées pour réaliser l'analyse de fréquences. Comme pour l'étape de nettoyage, pandas et matplotlib sont nécessaires. La bibliothèque numpy est aussi utilisées, elle permet de manipuler des matrices, des tableaux multidimensionnels et d'appliquer des fonctions mathématiques sur celles-ci. Numpy permettra notamment de calculer les corrélations.

Différents modules ont été importés pour les tâches de nettoyage :

- « itertools » : Permet de mettre à disposition une liste d'outils afin de réaliser des itérations, c'est-à-dire répéter une action. En programmation, cette répétition d'actions va permettre de parcourir les éléments d'un objet. Ici, l'objet est la base de données et les éléments sont les mots. Au sein du module itertools, différents itérateurs sont mis à disposition, celui qui est utilisé est l'itérateur chain() qui permet de concaténer (relier) des listes ou chaînes de caractères. Ici, l'outil est utilisé pour relier tous les commentaires ensemble afin que tous les mots des commentaires soient dans un seul et même bloc pour analyser leur fréquence.
- « collections » : Python possède différents types de structures pour organiser les données dont les listes, les dictionnaires et les tuples. Les listes sont des structures de données modifiables contenant une série de valeurs pouvant être les caractères ou des chiffres. Les dictionnaires sont des structures de données non ordonnées d'objets pouvant être des caractères ou des chiffres. Pour retrouver une donnée du dictionnaire, elle est appelée grâce à leur clé. Les tuples ont les mêmes caractéristiques que les listes mais ils ne sont pas modifiables. Le module collections permet de manipuler ces types de structures. Par son intermédiaire est utilisé l'attribut « counter() » qui permet de dénombrer l'occurrence des éléments

d'une structure de données. Il est ainsi mis à profit pour étudier la fréquence des mots de la base de données.

- « wordcloud » : Permet de représenter visuellement les mots les plus fréquents d'un corpus. Sur la représentation, plus le mot est gros et plus il apparaît fréquemment dans le texte. La couleur et la position des mots n'ont pas de signification, ils sont attribués de façon à afficher le plus clairement le plus grand nombre de mots possible.

Le processus de rédaction du code permettant l'analyse de la fréquence des mots est similaire à celui pour le nettoyage. Dans un premier temps est importée la donnée, puis des fonctions spécifiques aux tâches souhaitant être réalisées sont définies, enfin il s'agit d'exécuter les fonctions sur la donnée importée. Avant de décrire les analyses réalisées et d'expliquer leur pertinence, le concept de n-grammes est à définir pour faciliter la compréhension.

Les n-grammes sont des séquences contiguës de N éléments dans une phrase. N pouvant être n'importe quel entier positif. Bien souvent, N n'excède pas 3 car il est rare d'observer fréquemment plus de 3 mots adjacents dans différentes phrases. Dans ce travail, les bi-grammes sont utilisés pour connaître les associations de mots les plus fréquentes selon les différentes périodes de la base d'étude. Ainsi, les analyses de fréquences réalisées sont les suivantes :

- Analyse de la fréquence des mots sur toute la période étudiée (2016 à 2020) et par période (années, mois, semaines, jour).
- Analyse des bi-grammes sur toute la période étudiée (2016 à 2020) et par période (années, mois, semaines, jour).
- Analyse de la corrélation entre les mots les plus fréquents.
- Analyse de la corrélation entre les bi-grammes les plus fréquents.

Concernant les nuages de mots « wordcloud », il s'agit de l'une des techniques de visualisation de données les plus populaires pour représenter de la donnée en format chaîne de caractères. Ils permettent de visualiser les mots les plus fréquents dans un échantillon en utilisant la taille des caractères comme indicateur de fréquence ou d'importance. Ici les images générées ont une taille de 800 x 500 pixels avec un nombre maximal de mots de 200. La taille maximale d'un caractère est paramétrée à 110 pixels et le minimum à 4 pixels. Une échelle

relative (RS = 0,5) est utilisée. La taille de caractère (FS : font size) est calculée en fonction de la fréquence comme présenté ci-dessous :

$$FS = \left(RS * \left(\frac{frequency}{last_{freq}} \right) + (1 - RS) \right) * FS$$

3.2.4 Machine learning (analyse de similarité, analyse de sentiment)

L'analyse de sentiment permet d'associer une polarité à un message. Il peut représenter un sentiment, une satisfaction, une opinion, etc. C'est une technique très utilisée en marketing pour :

- Apprécier l'avis des internautes sur un produit, une marque, et l'évolution de cet avis dans le temps.
- Repérer les influenceurs qui génèrent des messages avec une polarité fortement positive.
- Identifier les problèmes associés aux sentiments négatifs.

Pour apprécier la polarité d'un texte plusieurs modèles peuvent être utilisés :

- Le dictionnaire : Chaque mot qu'il contient est associé à un score de sentiment. Pour chaque message un score global de sentiment est obtenu à partir des scores des mots qui le composent. L'avantage de cette méthode est qu'elle est simple à comprendre, à expliquer et à implémenter. Cependant, elle ne tient pas compte du contexte dans lequel le mot est employé et ne gère pas du tout les sarcasmes, l'ironie, etc.
- L'apprentissage supervisé : Il consiste à entraîner un modèle de *machine learning* à différencier les messages positifs des négatifs à partir de données labellisées. Exemples de modèles : SVM, régression logistique, XGBoost, etc. L'utilisation de classification naïve bayésienne est aussi une approche supervisée possible. Elle calcule pour un message la probabilité de chaque classe de sentiment (positive, négative ou neutre) sachant les mots qui le composent. Ces méthodes sont souvent plus performantes que l'approche par dictionnaire mais sont moins facilement explicables à des interlocuteurs métier. Elles nécessitent en outre un corpus de messages dont on connaît déjà le sentiment.
- Le Word Embedding : Il consiste à utiliser un réseau de neurones qui attribue à chaque mot du texte, un vecteur en fonction des mots qui l'entourent. Le principe étant que les mots qui apparaissant dans des contextes similaires possèdent des vecteurs correspondants qui sont relativement proches. Ces vecteurs peuvent être pré-calculés et utilisés en entrée des modèles supervisés ou être découverts à la suite de l'entraînement d'un réseau de neurones adapté à la problématique (*machine learning*

non supervisé). Cette méthode est très performante car elle prend en compte le contexte dans lequel le mot est utilisé. Cependant, comme tout réseau neuronal, il est difficile à expliquer et à interpréter.

- Les interfaces de programmation ou API (*Application Programming Interface*) : Il est également possible d'utiliser des API qui permettent de renforcer les services et les usages d'un programme. Il en existe de très performantes, telles que l'API Natural Language de Google® ou l'API Cognitive Services de Microsoft Azure®.

Toutes ces méthodes ont un objectif similaire : donner à chaque message une polarité positive ou négative. Dans ce travail, c'est l'algorithme Fasttext qui est utilisé pour l'analyse de similarité par vectorisation. C'est un algorithme de *word embedding*. Il est l'étape préliminaire à l'analyse sentimentale via les programmes de *machine learning* (sklearn est utilisé dans ce projet après entrainement sur un jeu de donnée Twitter®). Il est assez naturel de penser à l'analyse de sentiments pour associer à un message un sentiment positif ou négatif. Le terme de polarité à la place de sentiment est aussi fréquemment retrouvé. La méthode utilisée ici s'appuie sur une technique d'apprentissage automatique (*machine learning*). L'entraînement de l'algorithme sklearn est réalisé sur une base de données de tweets déjà labélisés positifs ou négatifs. La polarité de chaque message du forum est ensuite prédite en exécutant l'algorithme entrainé. Pour que le modèle de *word embedding* soit le plus efficient possible, son entraînement doit se faire sur une base de données dans la même langue et sur le sujet le plus similaire à la base de données à labéliser. L'idéal ici serait que les tweets soient en français et évoquent des sujets médicaux. Cependant, aucune base de données ne répondant à ces deux critères n'a été trouvée. Les tweets sont à la base en anglais puis ont été traduits en français mais ils n'évoquent pas uniquement des sujets médicaux.

3.2.5 Réseau neuronal convolutif entraîné sur les nuages de mots « Word Clouds Convolutional neural network » WC-CNN

Le concept des nuages de mots a déjà été introduit au cours de la deuxième étape. Si le nettoyage a en partie été basé sur cet élément, c'est pour pouvoir conduire des analyses sur des nuages de mots nettoyés. Ainsi, ils ne contiennent pas de mots parasites ce qui améliore la performance des algorithmes utilisés. Il est décidé de conduire des analyses via un réseau neuronal convolutif sur les nuages de mots correspondant à différentes périodes (jour, semaine, mois).

Mais, qu'est-ce qu'un réseau neuronal convolutif ? C'est un algorithme qui recherche les points communs entre les différents nuages de mots. Ces points communs sont les caractéristiques (*features* en anglais) communes aux différents nuages de mots. A la différence des outils d'apprentissage traditionnels, les réseaux de neurones définissent eux-mêmes les caractéristiques pertinentes à étudier au sein des éléments qui lui sont donnés (nuages de mots dans le cas de cette étude). Ces réseaux de neurones peuvent être utilisés dans des cas d'apprentissages supervisés ou non-supervisés, selon que la donnée soit labellisée ou non.

Comment le réseau neuronal fonctionne-t-il ? Il se décompose en plusieurs couches/convolutions permettant d'analyser successivement les différents aspects d'une donnée d'entrée. Ici, ce sont des images correspondant aux nuages de mots qui sont analysées. Les couches de convolutions permettent de traiter les images, chaque couche permet d'effectuer une action (nettoyage, suppression du bruit, faire ressortir les contours, etc.). Le mécanisme de la couche de convolution consiste en la transformation numérique que chaque pixel d'une image selon une caractéristique définie.

Comme présenté sur la figure suivante, chaque pixel en noir et blanc est transformé en un chiffre correspondant à son niveau de gris. Ensuite, la fenêtre "image pièce" de taille 3*3 pixels se déplace sur l'image et récupère les valeurs de chaque pixel. Cette fenêtre est multipliée par un *kernel* de même taille. C'est une multiplication terme à terme qui est effectuée où chaque pixel est multiplié par le chiffre correspondant du *kernel*. Les caractéristiques (chiffres 1 et 0) du *kernel*, permettent d'analyser les propriétés de l'image. Sur la figure ci-dessous, les chiffres

1 représentent une croix qui permettent d'analyser les lignes obliques de l'image. L'opération est détaillée sur la figure ci-dessous.

Figure 22 : Illustration d'une couche de convolution

Le résultat est mis de côté puis l'image pièce est déplacée d'un cran et l'opération est répétée jusqu'à la fin de l'image pour obtenir l'image de convolution, comme illustré sur la figure ci-dessous.

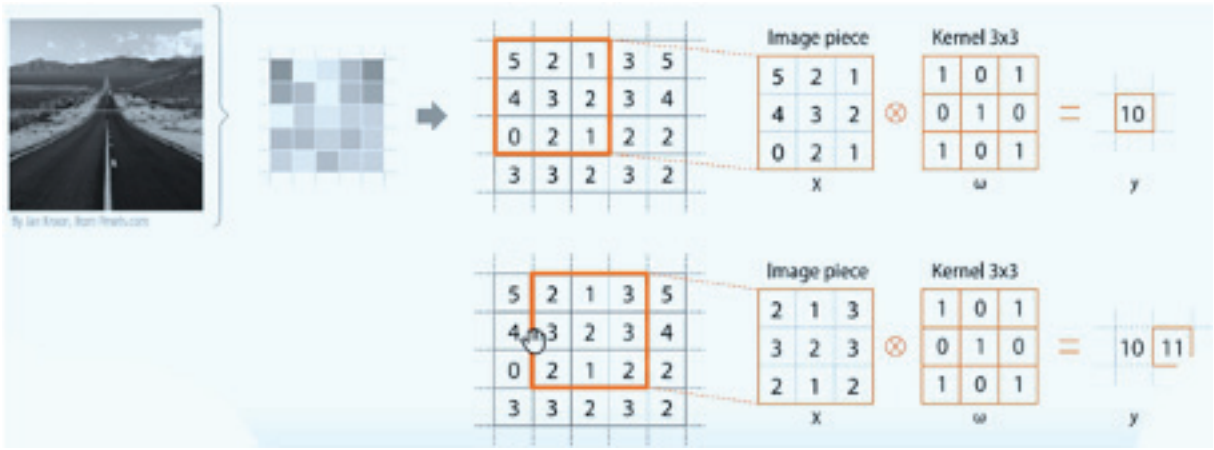

Figure 23 : Illustration du décalage d'une image pièce au sein d'une couche de convolution

Pour une image en couleur, le principe est le même sauf que le *kernel* est constitué de trois couches correspondant au code couleur RVB (rouge, vert bleu) de chaque pixel (illustration sur la figure ci-dessous).

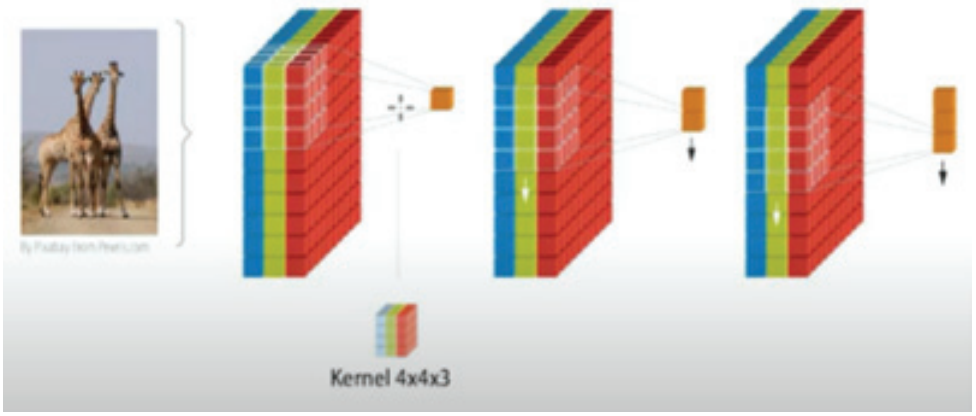

Figure 24 : Illustration du décalage d'une image pièce au sein d'une convolution pour une image en couleur

Dans chaque couche de convolution sont utilisés plusieurs *kernels* différents qui sont appelés des "filtres". Le nombre de filtres utilisé est une puissance de deux, ils sont dans ce travail compris entre 2^5 à 2^8 soit respectivement 32 à 256 filtres dans chaque couche de convolution pour des *kernels* de taille 2 par 2 à 6 par 6 pixels. Le nombre et la taille des *kernels* sont déterminés ainsi car c'est ce qui est le plus fréquemment observé dans la littérature. Plusieurs tests sont réalisés en faisant varier les paramètres : nombre de couches de convolutions (une à cinq), nombre de filtres, taille des *kernels*, traitements réalisés après chaque couche de convolution.

Les traitements appliqués après les couches de convolutions sont :

- Le *pooling* (mise en commun) (76) : Il diminue le risque de sur-apprentissage du réseau de neurones en sous-échantillonnant l'image pour réduire la quantité de paramètres et de calculs dans le réseau. Il est très fréquent de faire une étape de *pooling* entre deux couches de convolutions successives. La figure suivante illustre la translation réalisée par le *pooling*.

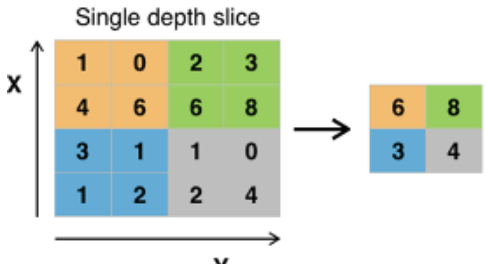

Figure 25 : Illustration du pooling (mise en commun)

- Le *flattening* (aplanissement) (77) : Il réduit de deux à une dimension l'image de convolution comme présenté sur la figure ci-dessous. Le but étant de vectoriser les couches de convolution pour les entrer ensuite dans un nouveau réseau de neurones.

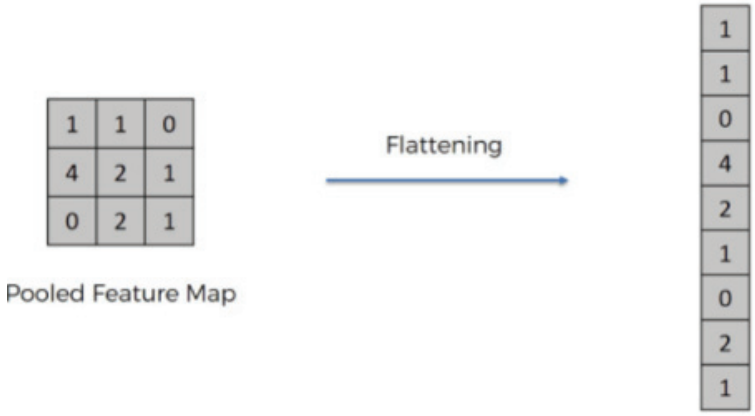

Figure 26 : Illustration du flattening (aplatissement)

- Le dropout (décrochage ou abandon) (78) : Il permet, comme le *pooling*, de réduire le sur-apprentissage en supprimant temporairement une partie des neurones du réseau.

L'architecture CNN proposée dans le cadre de ce travail peut se représenter comme indiqué sur la figure suivante :

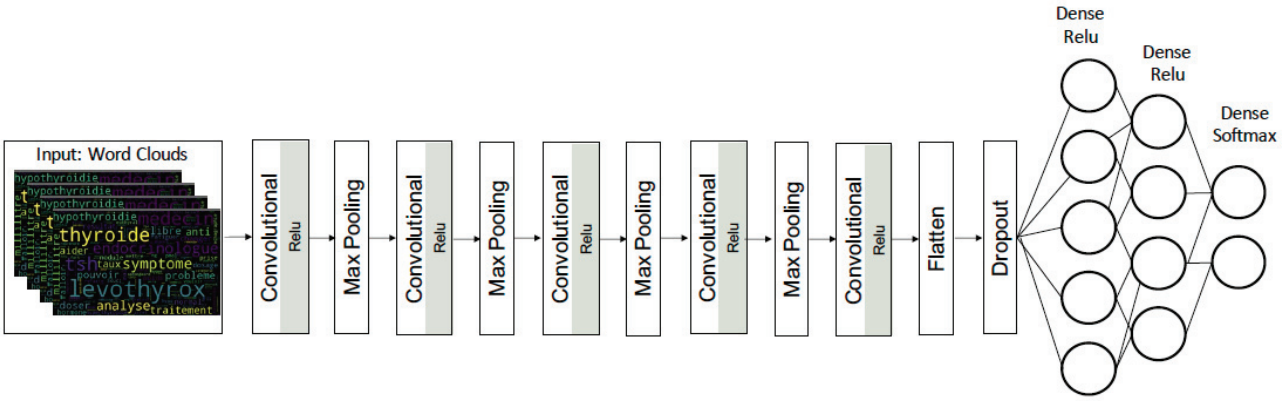

Figure 27 : Architecture CNN proposée

Le réseau de neurone proposé peut être nommé WC-CNN. C'est une architecture qui apprend et déduit les caractéristiques différenciantes des nuages de mots extraits des commentaires des patients selon différentes résolutions temporelles (jour, semaine, mois). La figure suivante permet d'illustrer la fonction du réseau neuronal qui présente les meilleurs résultats. Il est observé l'application de cinq couches de convolutions, la première appliquant 128 filtres (*kernels*) de taille 6 par 6 pixels. Sont retrouvés aussi l'utilisation des trois traitements présentés précédemment, le *pooling*, le *flattening* et le *dropout*.

```
def cnn_model(size, num_cnn_layers):
    NUM_FILTERS = 32
    KERNEL = (3, 3)
    #MIN_NEURONS = 20
    MAX_NEURONS = 120

    model = Sequential()
    model.add(Conv2D(128, (6,6), input_shape=(100,100,3), activation='relu'))
    model.add(MaxPooling2D(pool_size=(2,2)))
    model.add(Conv2D(128, (6,6), activation='relu'))
    model.add(MaxPooling2D(pool_size=(2,2)))
    model.add(Conv2D(128, (6,6), activation='relu'))
    model.add(MaxPooling2D(pool_size=(2,2)))
    model.add(Conv2D(128, (3,3), activation='relu'))
    model.add(MaxPooling2D(pool_size=(2,2)))
    model.add(Conv2D(128, (2,2), activation='relu'))
    model.add(Flatten())
    model.add(Dropout(0.5))
    model.add(Dense(256, activation='relu'))
    model.add(Dense(50, activation='relu'))
    model.add(Dense(2, activation='softmax'))

    model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])

    #print(model.summary())

    return model
```

Figure 28 : Illustration de la configuration paramétrique du réseau neuronal utilisé

Afin de tester la performance du réseau neuronal, les nuages de mots ont été labellisés normaux ou anormaux en fonction de trois périodes d'anormalité :

- Juillet à décembre 2017 (l'information du scandale est relayée dans les médias)
- Mai 2017 à février 2018
- Mars 2017 à avril 2018

Les images de convolution des nuages de mots sont présentées aléatoirement au réseau neuronal qui doit identifier si l'image est normale ou anormale. Les pourcentages de prédiction corrects et incorrects permettent d'établir la performance du réseau neuronal. Pour ces trois périodes d'anormalité, quatre nettoyages différents des nuages de mots sont appliqués afin de tester s'il y a des variations de performance de prédiction dépendantes du nettoyage. Les nuages de mots sont réalisés selon trois fréquences différentes : quotidienne, hebdomadaire et mensuelle, dans le but de connaître la donnée la plus pertinente à utiliser pour obtenir les meilleurs résultats de prédiction. L'intérêt de définir trois périodes d'anormalité est d'identifier si une détection précoce du problème lié au changement de formule du Levothyrox® est possible.

Cette détection précoce peut être confirmée si :

- Les performances du réseau neuronal sont bonnes sur la période juillet à décembre 2017. Cela signifie qu'il existe bien une différence entre les nuages de mots labellisés normaux et anormaux.
- Les résultats observés sur la période mai 2017 à février 2018 et/ou mars 2017 à avril 2018 sont aussi bons que sur la période juillet à décembre 2017. Cela confirme que les nuages de mots de deux ou trois périodes d'anormalité sont similaires (comme les trois périodes d'anormalité renferment la période juillet à décembre 2017).

Si les deux conditions précédentes sont remplies, il est possible de conclure qu'une différence avec la période de normalité est identifiable et qu'une détection précoce est possible.

3.2.6 Extraction des effets indésirables rapportés et analyses de leur occurrence

Après avoir conduit les analyses sur la fréquence des mots et de *machine learning*, il est décidé de faire une analyse d'occurrence et de bi-grammes, mais cette fois-ci sur les effets indésirables uniquement.

Il est alors créé une liste des effets indésirables et de leurs synonymes. Tous les mots de la base de données sont ensuite supprimés sauf ceux de la liste. Chaque commentaire ne contient plus que les effets indésirables.

Les analyses menées sont les mêmes et suivent la même méthodologie que celles présentée dans la partie 3.2.3 :

- Analyse de fréquences
- Analyse des bi-grammes

L'intérêt est d'étudier l'évolution dans le temps de la fréquence à laquelle sont cités les effets indésirables, savoir lesquels sont les plus fréquents ainsi que de connaître les associations les plus courantes.

3.3 Résultats

Dans cette partie sont présentés les résultats de l'analyse des données exposées dans la partie « matériel et méthodes ».

3.3.1 Analyse de la fréquence des mots et n-grammes

Les deux premières étapes de notre démarche consistent à collecter de la donnée puis à la formater. L'objectif de ce formatage, ou nettoyage, est de récupérer la donnée d'intérêt (sélection de la plage 2016-2020 et exclusion du reste) et d'uniformiser le texte récupéré. Ce texte est stocké dans la colonne « *text* » de notre fichier CSV source. Le fichier source sert ensuite de base aux manipulations de données avec la bibliothèque pandas. La qualité du nettoyage est primordiale pour obtenir des résultats pertinents. Il est important de garder à l'esprit que des transformations ont été appliquées ayant pour conséquence de transformer et supprimer des mots dans notre jeu de données. Cela est nécessaire pour le traitement statistique et pour les algorithmes d'IA mais explique aussi la présence de mots mal orthographiés dans les résultats présenté ci-après.

La troisième étape s'attache à l'analyse de la fréquence des mots, à l'évaluation de la pertinence des n-grammes puis l'étude de la corrélation entre les termes du corpus. L'extraction de la donnée s'étend jusqu'au 22 mars 2020 (date à laquelle l'extraction est réalisée). Pour l'interprétation des résultats, l'année 2020 n'est pas prise en compte comme seulement une partie de l'année est extraite.

Top word occurence history

	2016-12-31	2017-12-31	2018-12-31	2019-12-31	2020-12-31
an	214.0	441.0	239.0	155.0	19.0
cortisol	16.0	30.0	40.0	21.0	44.0
dosage	217.0	433.0	236.0	80.0	34.0
doser	271.0	402.0	175.0	82.0	14.0
falloir	350.0	770.0	401.0	122.0	41.0
fatiguer	203.0	330.0	185.0	95.0	47.0
formule	2.0	548.0	119.0	23.0	2.0
hormone	225.0	255.0	170.0	75.0	37.0
levothyrox	530.0	1711.0	449.0	224.0	33.0
mal	159.0	455.0	197.0	80.0	21.0
medecin	548.0	1203.0	655.0	392.0	83.0
norme	333.0	416.0	319.0	133.0	14.0
resultat	303.0	331.0	270.0	106.0	45.0
sang	278.0	391.0	203.0	79.0	15.0
savoir	180.0	450.0	217.0	82.0	31.0
symptome	223.0	379.0	217.0	151.0	26.0
thyroid	303.0	463.0	314.0	138.0	38.0
traitement	328.0	502.0	297.0	198.0	43.0
tsh	720.0	1064.0	772.0	306.0	122.0

Figure 29 : Capture d'écran des résultats obtenus lors de l'exécution de la fonction « top word occurrence history » (annexe 5)

La figure précédente présente les mots les plus fréquemment observés sur quatre années et leur fréquence d'apparition dans chacune d'elle. Parmi ces résultats, il est observé en 2017 :

- 1,74 fois plus d'occurrence qu'en 2016 et 3,49 fois plus que les années postérieures à 2017 pour le mot « dosage » ou « doser ».
- 1,63 fois plus d'occurrence qu'en 2016 et 2,63 fois plus que les années postérieures à 2017 pour le mot « fatiguer ».
- 274 fois plus d'occurrence qu'en 2016 et 14,22 fois plus que les années postérieures à 2017 pour le mot « formule ».
- 3,23 fois plus d'occurrence qu'en 2016 et 5,72 fois plus que les années postérieures à 2017 pour le mot « levothyrox ».
- 2,86 fois plus d'occurrence qu'en 2016 et 4,00 fois plus que les années postérieures à 2017 pour le mot « mal ».
- 2,20 fois plus d'occurrence qu'en 2016 et 2,45 fois plus que les années postérieures à 2017 pour le mot « medecin ».
- 1,70 fois plus d'occurrence qu'en 2016 et 2,12 fois plus que les années postérieures à 2017 pour le mot « symptome ».
- 1,53 fois plus d'occurrence qu'en 2016 et 2,11 fois plus que les années postérieures à 2017 pour le mot « traitement ».

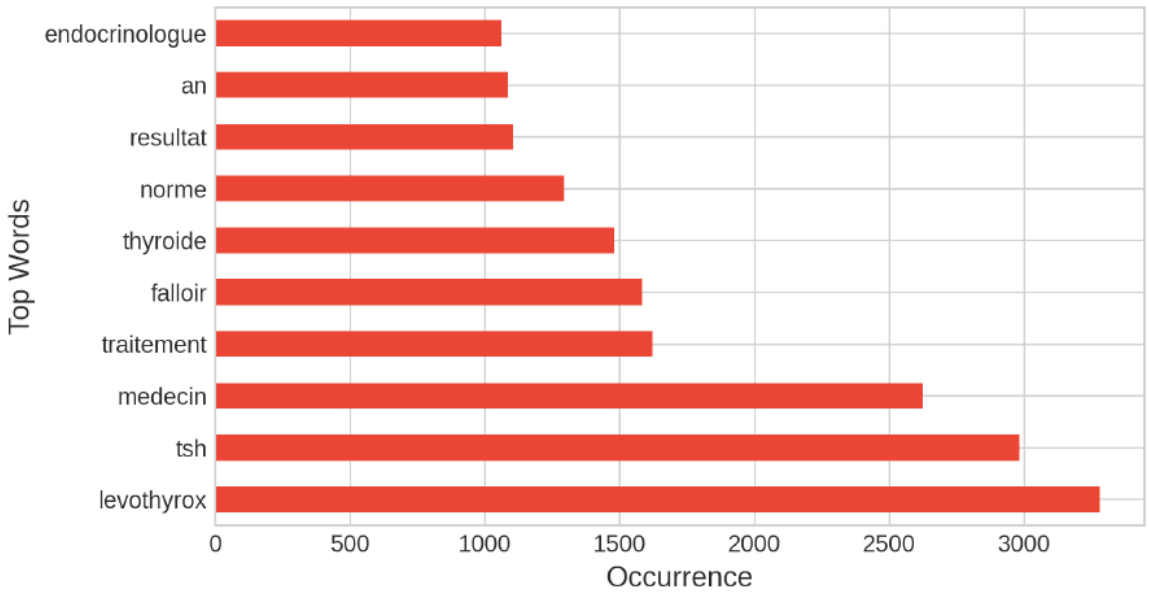

Figure 30 : « Top word occurrence » (2016-2020)

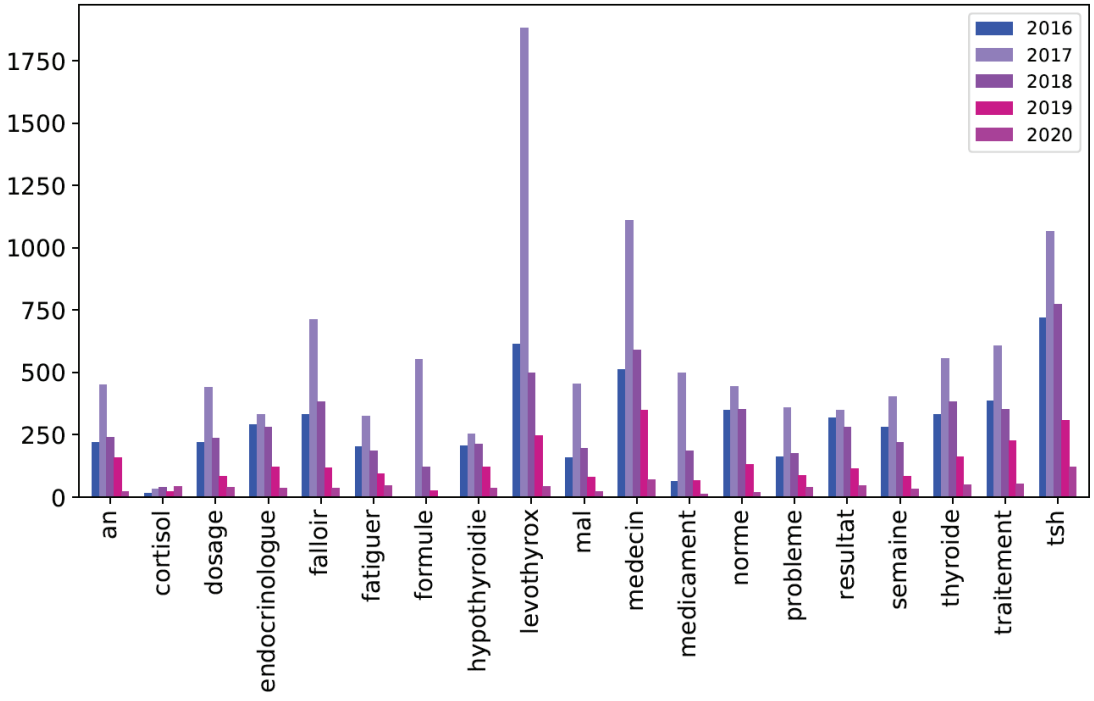

Figure 31 : « Top word occurrence » par année (période 2016-2020)

L'observation de ces résultats indique bien une forte activité des utilisateurs, en lien avec le Levothyrox® en 2017. De nombreux mots présents ici le sont du fait de la surreprésentation de leurs occurrences durant l'année 2017 ; à commencer par l'exemple frappant de « formule » qui n'était utilisé qu'à deux reprises en 2016 et 548 fois l'année d'après. Il s'est révélé qu'avec ces résultats, il était effectivement possible d'approfondir ce travail en utilisant les effets indésirables et les termes significatifs s'y rapportant (ex : « dosage », « doser », « fatiguer », « mal », « symptome », « traitement »). La recherche de corrélation entre les mots ne donne pas de bons résultats, ce qui conduit à estimer l'utilisation de n-grammes. Dans ce travail, ce sont les bi-grammes, donc 2 mots qui sont étudiés. Ci-dessous, la figure présentant les premiers résultats obtenus, à savoir l'occurrence des 10 bi-grammes les plus fréquents pour la période choisie.

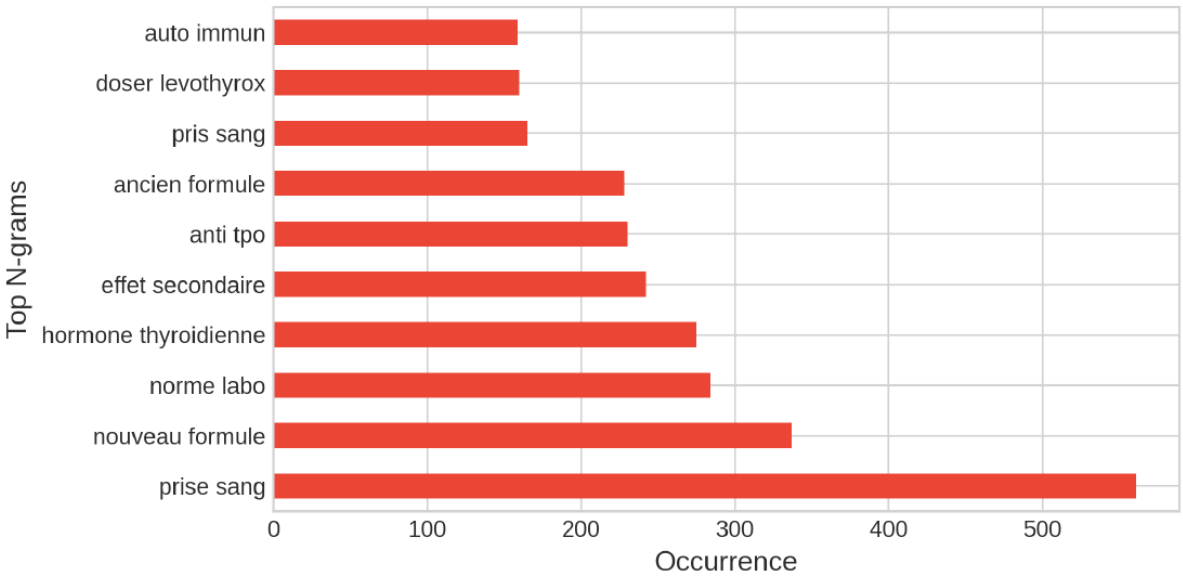

Figure 32 : « Top bi-gram occurrence » (2016-2020)

Les résultats obtenus sont très prometteurs, il est décidé de systématiser l'utilisation de ces bi-grammes dans le reste de l'expérimentation. Par la suite, le raisonnement utilisé précédemment a été appliqué *stricto sensu*, pour dégager :

- Un « top n-gram history » qui rassemble l'ensemble des 10 bi-grammes les plus fréquents de chaque année
    - L'analyse faite au travers des bi-grammes se révèle plus pertinente. Les termes affichés sont cohérents et correspondent davantage à ce qui est attendu.
    - Certains bi-grammes sont particulièrement révélateurs : « ancien formule », « effet secondaire », « formule levothyrox », « nouveau formule ».

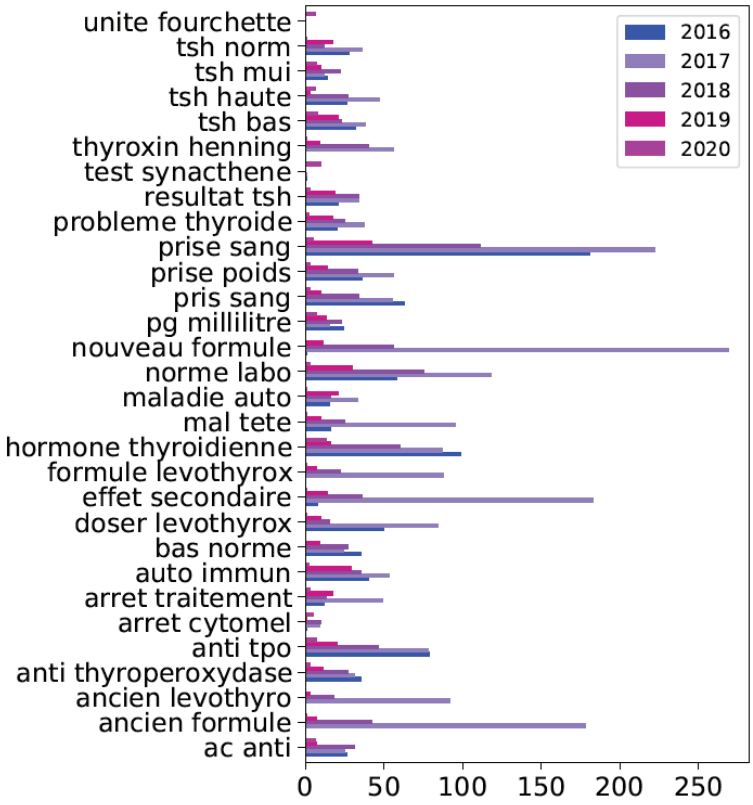

Figure 33 : « Top n-gram » par année (2016-2020)

- Un « top n-gram history 2 » qui présente les 10 bi-grammes les plus fréquents de chaque année

- Encore plus probant, des bi-grammes inexistants en 2016 deviennent certains des bi-grammes les plus observés de l'année du changement de formule.

- Un « top n-gram occurrence correlations 2016-2020 » qui affiche les corrélations significatives (> 0,95) entre deux bi-grammes. Il est observé plusieurs fois le bi-gramme « ancien formule » qui est l'un des bi-grammes les plus fréquents dans le classement « top-n-gram history » et en 2017 dans le classement « top n-gram history 2 ».
    - Les six corrélations les plus fortes entre deux bi-grammes sont tous dans le classement « top n-gram history » et concernent le changement de formule du Levothyrox®
    - C'est ici la preuve indéniable de ce qui pouvait se poser en hypothèse : il s'est passé quelque chose de nouveau en 2017, au point de dominer les tendances sur une période de quatre ans. De plus, en s'intéressant à un découpage mensuel, il est observé sur la figure 35 que ce phénomène a débuté avec les annonces médiatiques, en juillet 2017.

Le tableau 4 présente les 10 bi-grammes les plus fréquents pour les années comprises dans la période étudiée (2016-2020). De nombreux bi-grammes apparaissant en 2017 n'existaient pas avant le changement de formule. Certains sont directement liés à ce changement : « nouveau formule », « ancien formule », « ancien levothyrox », « formule levothyrox », etc. Les meilleures corrélations entre deux bi-grammes identifiés parmi les meilleurs de chaque année sont présentés dans le tableau 5.

Tableau 4 : Les 10 bi-grammes les plus fréquents pour les années comprises entre 2016 et 2020 inclus

2016	***2017***	***2018***	***2019***	***2020***
prise sang	nouveau formule	prise sang	prise sang	hormone thyroidienne
hormone thyroidienne	prise sang	norme labo	norme labo	test synacthene
anti tpo	effet secondaire	hormone thyroidienne	auto immun	tsh bas
pris sag	ancien formule	nouveau formule	maladie auto	anti tpo
norme labo	norme labo	anti tpo	tsh bas	pg millilitre
doser levothyrox	mal tete	ancien formule	anti tpo	tsh mui
auto immun	ancien levothyrox	thyroxin henning	resultat tsh	ac anti

prise poids	formule levothyrox	effet secondaire	arret traitement	tsh haute
anti thyroperoxydase	hormone thyroidienne	auto immun	probleme thyroide	unite fourchette
bas norme	doser levothyrox	pris sang	tsh norme	arret cytomel

Tableau 5 : Les meilleures corrélations entre deux bi-grammes parmi les meilleurs de chaque année

Bi-gramme 1	***Bi-gramme 2***	***Corrélation***
Ancien levothyrox	Nouveau formule	0,9999287212
Ancien formule	Nouveau formule	0,9995480603
Ancien formule	Formule levothyrox	0,9993103979
Ancien formule	Ancien levothyrox	0,9992348913
Ancien levothyrox	Effet secondaire	0,9989443185
Effet secondaire	Nouveau formule	0,9988471039
Formule levothyrox	Nouveau formule	0,9988416757
Ancien levothyrox	Formule levothyrox	0,9982914213
Ancien formule	Effet secondaire	0,9971094253
Effet secondaire	Formule levothyrox	0,9966794925

La figure de la page suivante illustre les résultats bruts, obtenus à suite de l'exécution de l'algorithme.

Top n_gram history

	2016-12-31	2017-12-31	2018-12-31	2019-12-31	2020-12-31
ac anti	26.0	25.0	31.0	7.0	6.0
ancien formule	0.0	178.0	42.0	7.0	1.0
anti thyroglobuline	33.0	27.0	25.0	17.0	4.0
anti thyroperoxydase	34.0	31.0	27.0	11.0	3.0
anti tpo	79.0	78.0	46.0	20.0	7.0
anticorps anti	52.0	55.0	29.0	9.0	4.0
arret cytomel	1.0	9.0	10.0	1.0	5.0
auto immune	37.0	51.0	36.0	27.0	2.0
carence iode	0.0	1.0	1.0	0.0	6.0
doser levothyrox	51.0	85.0	17.0	10.0	1.0
effet secondaire	8.0	183.0	36.0	14.0	1.0
formule levothyrox	0.0	88.0	21.0	7.0	1.0
hormone thyroidienne	91.0	84.0	56.0	15.0	12.0
maladie auto	15.0	33.0	16.0	21.0	1.0
medecin traitant	41.0	52.0	17.0	19.0	3.0
norme labo	50.0	114.0	73.0	32.0	3.0
nouveau formule	1.0	269.0	56.0	11.0	0.0
perte cheveux	11.0	29.0	18.0	19.0	2.0
pg millilitre	24.0	15.0	23.0	13.0	7.0
pris sang	69.0	67.0	43.0	14.0	6.0
prise sang	176.0	218.0	104.0	41.0	3.0
test synacthene	1.0	1.0	0.0	0.0	7.0
thyroxin henning	0.0	73.0	52.0	11.0	1.0
tsh bas	32.0	38.0	23.0	21.0	8.0
tsh mui	14.0	12.0	22.0	10.0	7.0
tsh norme	33.0	59.0	31.0	21.0	1.0

Top n_gram history 2

	2016-12-31	2017-12-31	2018-12-31	2019-12-31	2020-12-31
0	prise sang	nouveau formule	prise sang	prise sang	hormone thyroidienne
1	hormone thyroidienne	prise sang	norme labo	norme labo	tsh bas
2	anti tpo	effet secondaire	hormone thyroidienne	auto immune	anti tpo
3	pris sang	ancien formule	nouveau formule	maladie auto	pg millilitre
4	anticorps anti	norme labo	thyroxin henning	tsh bas	test synacthene
5	doser levothyrox	formule levothyrox	anti tpo	tsh norme	tsh mui
6	norme labo	doser levothyrox	pris sang	anti tpo	ac anti
7	medecin traitant	hormone thyroidienne	ancien formule	medecin traitant	carence iode
8	auto immune	anti tpo	auto immune	perte cheveux	pris sang
9	anti thyroperoxydase	thyroxin henning	effet secondaire	anti thyroglobuline	arret cytomel

Top n_gram occurence correlations 2016-2020

	Item 1	Item 2	correlation
89	ancien formule	nouveau formule	0.999548
84	ancien formule	formule levothyrox	0.999266
224	formule levothyrox	nouveau formule	0.999147
210	effet secondaire	nouveau formule	0.998847
205	effet secondaire	formule levothyrox	0.997412
83	ancien formule	effet secondaire	0.997109
108	anti tpo	pris sang	0.997097
240	hormone thyroidienne	pris sang	0.996754
94	anti tpo	anticorps anti	0.996132
128	anticorps anti	pris sang	0.994146
101	anti tpo	hormone thyroidienne	0.990027
129	anticorps anti	prise sang	0.989693
121	anticorps anti	hormone thyroidienne	0.989083
109	anti tpo	prise sang	0.983944
304	pris sang	prise sang	0.976534
250	maladie auto	perte cheveux	0.976121
87	anti thyroperoxydase	pris sang	0.974985
267	medecin traitant	tsh bas	0.973781
193	doser levothyrox	medecin traitant	0.969854
49	anti thyroglobuline	anti thyroperoxydase	0.969646
279	norme labo	tsh norme	0.969632
171	auto immune	tsh norme	0.968517
80	anti thyroperoxydase	hormone thyroidienne	0.967119
184	carence iode	test synacthene	0.965775
72	anti thyroperoxydase	anti tpo	0.965400
312	prise sang	tsh bas	0.964150
169	auto immune	tsh bas	0.961597
241	hormone thyroidienne	prise sang	0.961181
264	medecin traitant	prise sang	0.958783
323	tsh bas	tsh norme	0.956814
199	doser levothyrox	prise sang	0.953508

Figure 34 : Résultats obtenus lors de l'exécution des fonctions liées aux bi-grammes (annexe 5)

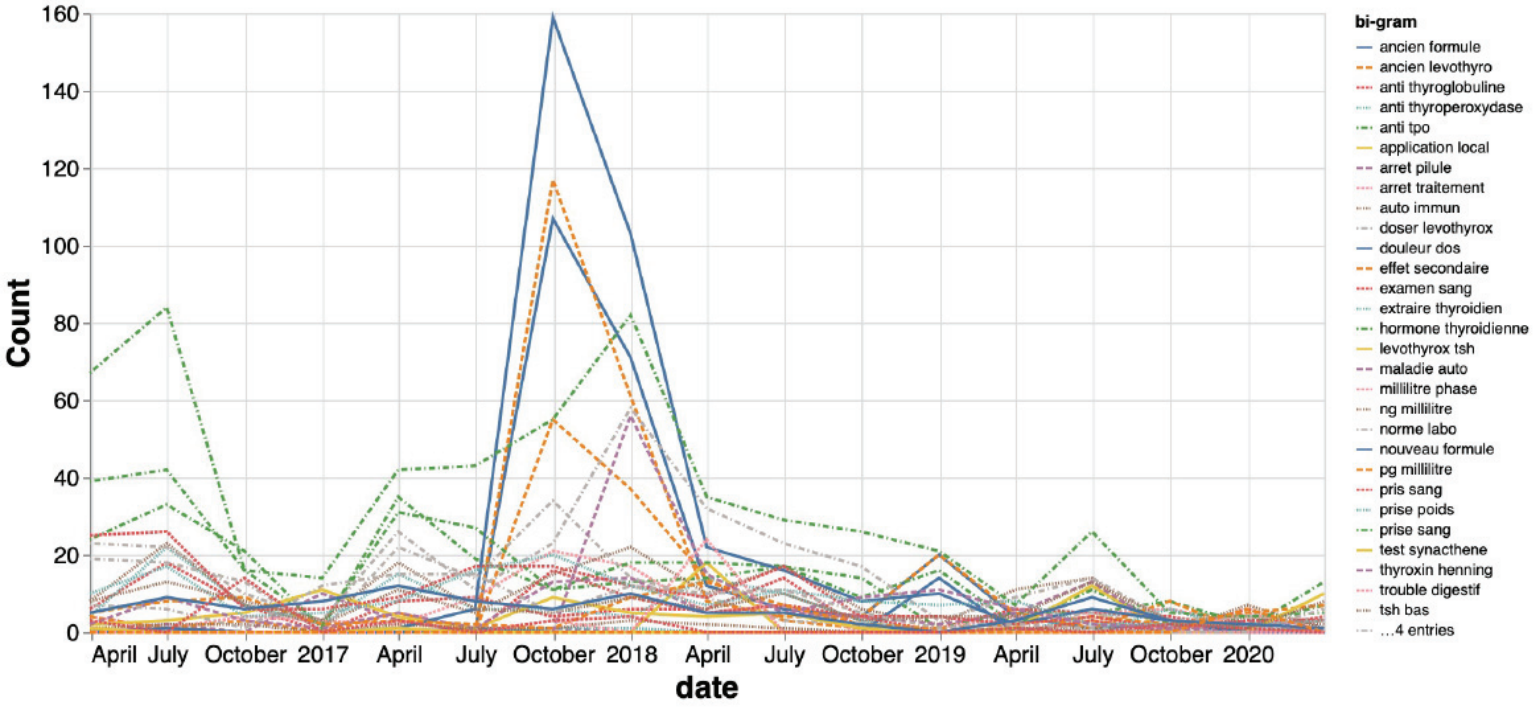

Figure 35 : Fréquence d'apparition des "top n-gram" pour la période 2016-2020

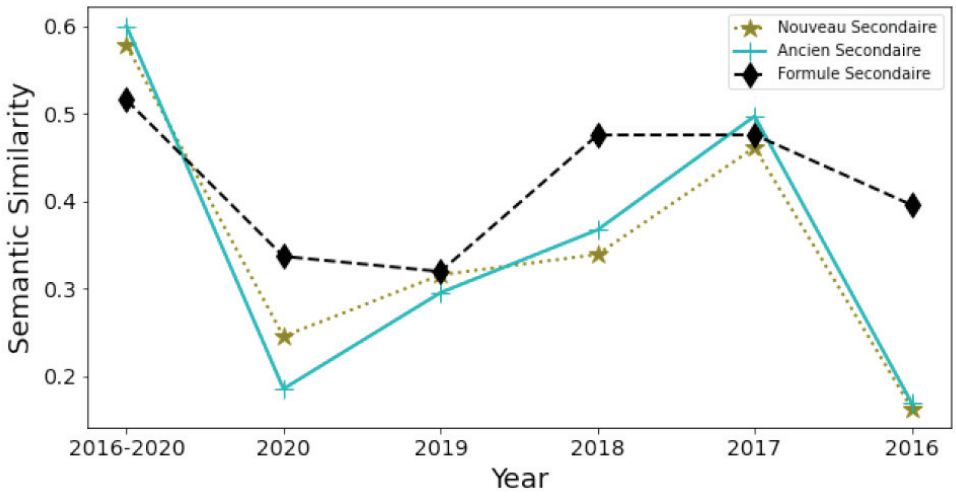

Figure 36 : Analyse de similarité sémantique appliquée aux bi-grammes

A ce stade, il est impossible de parler de détection précoce. Il s'agit d'un moyen de confirmer expérimentalement une tendance prévisible, de la nommer et de la chiffrer, au travers de la donnée générée par les utilisateurs d'un forum. Il s'agit désormais d'utiliser des outils conçus spécifiquement pour le traitement et l'analyse massive de donnée, afin de déterminer s'il est possible de détecter précocement ce qui allait se passer avant les annonces médiatiques.

3.3.2 Analyse de contenu textuel via un algorithme de NLP (Traitement du Langage Naturel)

Plusieurs outils d'IA et de NLP (traitement du langage naturel) sont utilisés de façon itérative : Fasttext, sklearn et Spacy. Comme expliqué précédemment, le modèle Fasttext est un algorithme de *word embedding*. Il permet d'attribuer, à chaque mot, un vecteur à plusieurs dimensions (60 dimensions dans ce travail), comme illustré sur la figure suivante pour le mot « levothyrox ».

```
Keyed vectors for levotyrox
[-0.1515551  -0.0431994   0.02470955  0.23442402  0.07182983 -0.51890326
  0.10021465  0.36545977 -0.16588293 -0.18264422  0.26579508 -0.08827683
 -0.22856167  0.08466751  0.12038728  0.13065295  0.11172011 -0.15287635
  0.05534374 -0.695484   -0.17109105 -0.02951328  0.35382134 -0.12830362
 -0.42894897 -0.26008755 -0.4690509   0.27304554 -0.37958375  0.38361332
 -0.3518691  -0.36105013 -0.2549195   0.07133891 -0.04836788  0.08016964
  0.02666572  0.18020687 -0.05995554 -0.11599924  0.24733661 -0.0512655
  0.16236569  0.18382019  0.27828327 -0.35402256 -0.21740484  0.282825
  0.23846106 -0.00398343 -0.29370084  0.5027071   0.07110989  0.27753288
 -0.22058323  0.29546317  0.14892468 -0.0786326   0.15001339  0.02198214]
```

Figure 37 : Représentation du vecteur à 60 dimensions du mots « levothyrox » par l'algorithme de machine learning Fasttext

3.3.2.1 Analyse des similarités sémantiques

Il est possible de visualiser les similarités entre les mots en utilisant la similarité cosinus sur modèle Fasttext. Il permet de réduire les dimensions des vecteurs de chaque mot de 60 à 2 afin de les visualiser sur un graphique et de se rendre compte des termes que l'algorithme considère comme proches. Cette transformation est permise via la librairie sklearn, permettant à tout être humain de visualiser simplement les similarités proposées par l'algorithme.

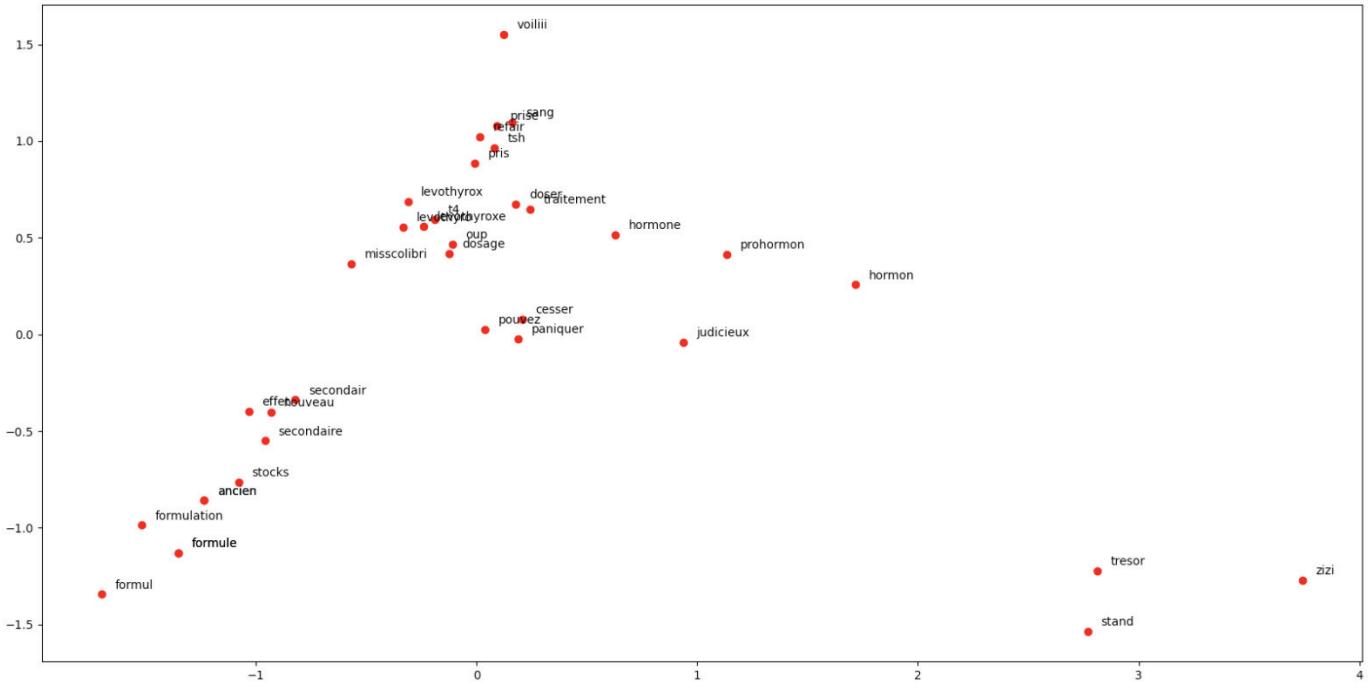

Figure 38 : Visualisation du résultat de sklearn sur notre jeu de donnée, en utilisant des vecteurs de mots bidimensionnels (utilisation de la bibliothèque PCA de sklearn)

3.3.2.2 Analyse de l'évolution des sentiments

L'étape suivant l'utilisation des outils de *machine learning* consiste à analyser les évolutions des sentiments perçus. Pour cela, la librairie « Spacy » est utilisée de manière à associer automatiquement un sentiment, négatif (0) ou positif (1) à toutes les entrées de notre table de données.

```
Success rate:  74.66246206132749 %

All sentiments
                                                text         prediction
0  suivie thyroidite hasimoto deconvenuer levothy...  __label__negative
1  medecin prescrire hypothyroidi secondaire tsh ...  __label__negative
2  equivalent pifometre savoir exactement dosage ...  __label__negative
3  hypothyroidie traitemer ans quotidiennement mi...  __label__negative
4  date savoir fille atteinte syndrome interrupti...  __label__negative
```

Figure 39 : Résultat après exécution de l'algorithme sklearn (association sentimentale après entrainement sur un jeu de test)

L'objectif est de tenter de détecter avant juillet 2017, par une autre méthode, la survenue d'un événement ou à minima, d'augmenter la fenêtre de détection d'un problème. Les résultats les plus éloquents sont présentés sur la figure suivante mais ne permettent pas de conclure à une détection précoce. En effet, le pic de commentaires négatifs se situe après le mois d'août 2017. Rien n'est observé plus tôt hormis le pic au mois de mars 2017, avant le changement

de formule. La figure ci-après reprend l'évolution des commentaires positifs et négatifs par an, par mois (2016-2020) et par mois durant l'année 2017. Au cours de la période juillet 2017 à janvier 2018, il est observé un nombre de commentaires négatifs bien supérieur aux commentaires positifs alors qu'en temps normal les fréquences sont très proches.

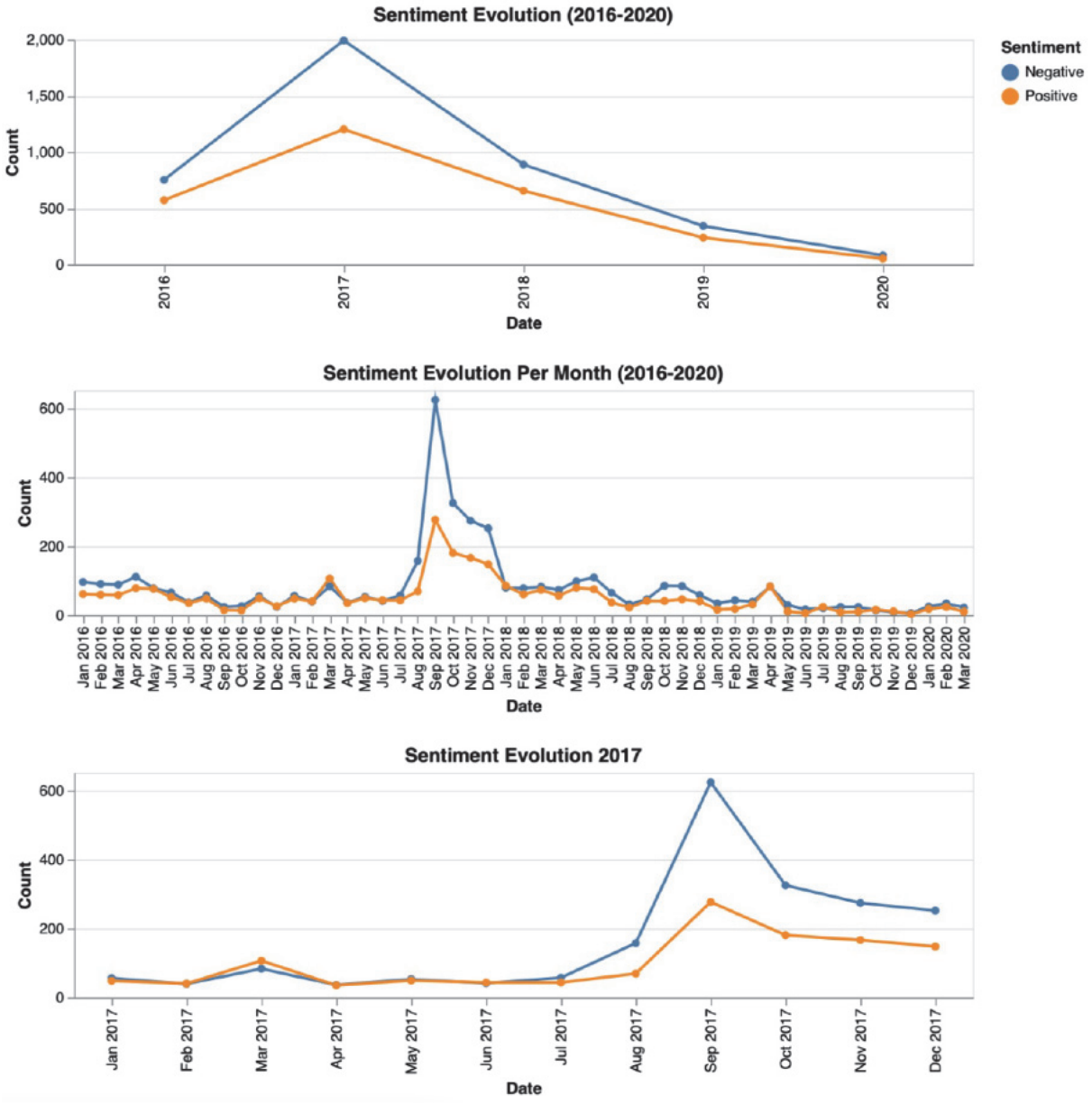

Figure 40 : Histogramme représentant l'évolution sentimentale des commentaires de patients, par année, de 2016 à 2020

3.3.2.3 Evaluation de la performance du CNN (Réseau de Neurones Convolutifs)

L'utilisation de réseaux de neurones a permis de mettre en évidence des résultats très intéressants concernant la détection précoce des événements indésirables.

Le tableau suivant présente les résultats du réseau neuronal utilisé après optimisation des paramètres.

Période anormale : Juillet à Décembre 2017				
Nettoyage effectué	**Jour**	**Semaine**	**Mois**	**Toutes périodes confondues**
Etapes 1 2 3 4 5	0,557	0,519	0,250	0,572
Etapes 1 2 3 4	0,554	0,611	0,500	0,527
Etapes 1 2 3 5	0,548	0,593	0,667	0,493
Etapes 1 2 5	0,533	0,463	0,333	0,527
Période anormale : Mai 2017 à Février 2018				
Nettoyage effectué	**Jour**	**Semaine**	**Mois**	**Toutes périodes confondues**
Etapes 1 2 3 4 5	0,531	0,625	0,750	0,566
Etapes 1 2 3 4	0,547	0,523	0,550	0,537
Etapes 1 2 3 5	0,494	0,455	0,400	0,519
Etapes 1 2 5	0,492	0,580	0,650	0,519
Période anormale : Mars 2017 à Avri 2018				
Nettoyage effectué	**Jour**	**Semaine**	**Mois**	**Toutes périodes confondues**
Etapes 1 2 3 4 5	0,566	0,476	0,500	0,574
Etapes 1 2 3 4	0,547	0,444	0,571	0,553
Etapes 1 2 3 5	0,531	0,532	0,464	0,541
Etapes 1 2 5	0,584	0,589	0,429	0,555

Figure 41 : Résultats du réseau de neurones présentant la meilleure performance

Comme expliqué dans la partie explicative du réseau neuronal, différents nettoyages des nuages de mots sont réalisés pour savoir si des variations de performance de prédiction sont observées.

Quatre nettoyages différents sont réalisés, les étapes de nettoyage effectuées sont les suivantes :

1. Suppression des commentaires de moins de trois mots comme une phrase contient au moins un sujet, un verbe et un complément.
2. Suppression des « *stopwords* ».

3. Lemmatisation des commentaires.
4. Améliorations de la lemmatisation par la création de plusieurs listes permettant de corriger l'orthographe des mots observés dans les nuages de mots.
5. Création d'une liste appelée « words_to_delete », destinée à nettoyer les nuages de mots des mots parasites non liés au contexte médical.

Les résultats du réseau neuronal sont présentés dans trois tableaux correspondants chacun aux trois périodes d'anormalité. Dans chaque tableau :

- La première colonne permet d'indiquer les étapes de nettoyage effectuées.
- Les colonnes 2, 3 et 4 présentent les résultats obtenus selon la fréquence à laquelle les nuages de mots ont été réalisés.
- La colonne 4 présente les résultats obtenus lorsque le réseau de neurones est entraîné et testé sur les nuages de mots des trois périodes confondues (quotidienne, hebdomadaire et mensuelle).

Entre les différents tests réalisés, il est observé qu'en augmentant le nombre de couches de convolutions, donc en utilisant un réseau neuronal d'apprentissage plus profond, les résultats sont meilleurs. Un résultat est considéré comme acceptable lorsque la prédiction est supérieure à 0,56 et considéré comme bon lorsqu'elle est supérieure à 0,6. En effet, comme seulement deux labels sont ici présents, une prédiction aléatoire correspond à un résultat de 0,5, s'il y en avait eu trois, une prédiction aléatoire correspondrait à un résultat de 0,33. N'étant pas expert des réseaux neuronaux et les sources bibliographiques les utilisant étant assez limitées, il est décidé de s'appuyer sur l'expérience du docteur *Hanan Salam*, directrice de ce sujet thèse pour fixer le seuil à partir duquel un résultat peut être considéré comme significatif.

Pour le réseau de neurones entier, il est observé 19 résultats supérieurs à 0,56 (cases vertes de la figure précédente) sur 48 résultats totaux. C'est un résultat qui peut paraître faible mais en réalité il est très bon. En comparaison avec les réseaux de neurones moins profond (deux couches de convolution), entre zéro et trois résultats supérieurs à 0,56 sont observés.

Pour la période d'anormalité juillet à décembre 2017, la prédiction la plus élevée est obtenue via les nuages de mots mensuels (précision de 0,667), suivie par les nuages de mots hebdomadaires avec une précision de 0,611. Les mêmes conclusions sont tirées sur la période d'anormalité mai 2017 à février 2018 avec trois très bons résultats de 0,75 et 0,65 pour les nuages de mots mensuels et de 0,625 pour les nuages de mots hebdomadaires. Pour la

période mars 2017 à avril 2018, les résultats sont un peu moins bons mais restent acceptables. Lors des tests avec les autres réseaux de neurones, la période mars 2017 à avril 2018 présente fréquemment des résultats un peu en dessous des deux autres. Lorsque tous les nuages de mots sont confondus, les résultats sont moins bons et presque identiques à ceux observés avec les nuages de mots quotidiens (toujours autour de 0,5). Cette observation est expliquée par le nombre très important de nuages de mots quotidiens par rapport aux hebdomadaires et mensuels qui viennent diminuer fortement la performance de prédiction.

Le nettoyage de la donnée a peu d'impact sur les performances de prédiction des réseaux de neurones. En effet, des résultats supérieurs à 0,6 sont observés avec tous les différents types de nettoyages. Ces observations sont également retrouvées lors des tests effectués en faisant varier les paramètres des réseaux de neurones. Ainsi, il est observé que la performance de prédiction augmente avec l'augmentation du nombre de couches de convolution et est meilleure sur les nuages de mots hebdomadaire et mensuels. Ce n'est pas observable sur ce réseau de neurones mais lors des différents tests effectués, il est identifié que les nuages de mots mensuels présentent de meilleurs résultats que les hebdomadaires.

Aussi, les résultats des périodes d'anormalité juillet à décembre 2017 et mai 2017 à février 2018 étant tout aussi bons, il est possible de conclure qu'une détection précoce de signaux indicateurs d'évènements anormaux est identifiable à partir du mois de mai 2017. Il est remarqué que ces résultats sont moins bons sur la période d'anormalité de mars 2017 à avril 2018. Ce moins bon résultat est explicable car la nouvelle formule est arrivée sur le marché au mois de mars et était encore peu consommée par les patients. Aussi, la fin de la période d'anormalité s'étendant jusqu'en avril 2018, les patients échangeaient moins sur le sujet après décembre 2017. Les nuages de mots sont moins proches de ceux observés entre juillet et décembre 2017 qui est la période au cours de laquelle une explosion de la fréquence des commentaires est observée sur Doctissimo®.

3.3.3 Analyse des effets indésirables

Concernant la dernière partie de ce travail expérimental, elle s'attache à étudier l'évolution de l'occurrence des effets indésirables. Un traitement supplémentaire de la donnée permet de garder uniquement ce que l'algorithme identifie comme étant un effet indésirable. Seuls subsistent la date et le texte se rapportant à l'un des effets indésirables répertoriés. Le résultat de ce traitement est présenté ci-après :

```
*****************
Dataframe side effects
*****************

date
2020-03-21                                              [insomnie]
2020-03-13                                                      []
2020-03-13                                                      []
2020-03-13                                                      []
2020-03-11                                                      []
2020-03-09                                                      []
2020-03-09                                              [deprimer]
2020-03-07                                                      []
2020-03-06                                                      []
2020-03-05                                                      []
2020-03-04                                      [fatiguer, vertige]
2020-03-04                                                      []
2020-03-04                                                      []
2020-03-04                                                      []
2020-03-04                                                      []
2020-03-03                                                      []
2020-03-03                                                      []
2020-03-03    [fatiguer, depressionnair, pertecheveux, pertu...
2020-03-02                                              [fatiguer]
2020-03-02                                                      []
Name: side_effects, dtype: object
```

Figure 42 : Version finale du jeu de donnée utilisé après suppression des mots qui ne sont pas des effets indésirables

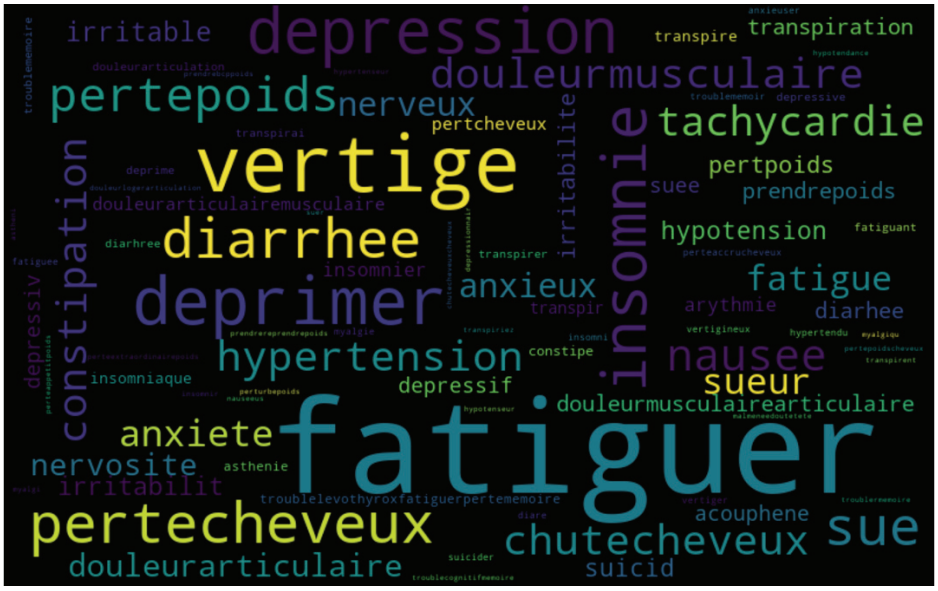

Figure 43 : Visualisation des principaux effets indésirables mentionnés

L'histogramme présenté dans la figure suivante détaille l'occurrence des effets indésirables par commentaire posté. Deux périodes sont étudiées : 2016 à 2020 et durant l'année 2017.

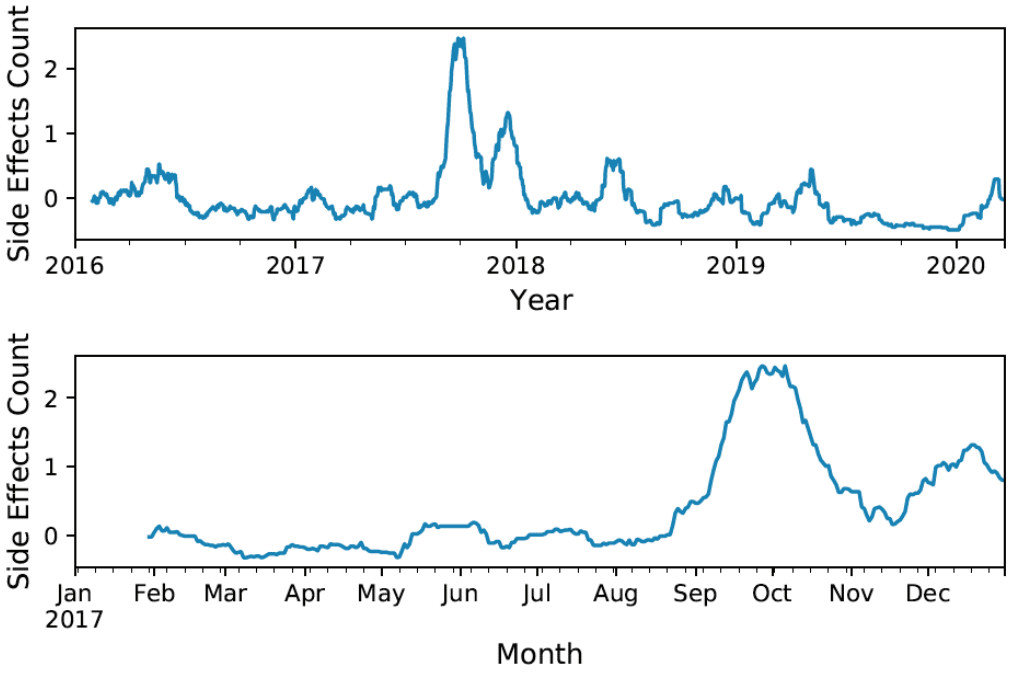

Figure 44 : Occurrence des effets indésirables entre 2016 et 2020 puis durant l'année 2017

Autre résultat de l'expérimentation, l'historique quotidien sur 4 ans ou sur l'année 2017 de l'occurrences des effets indésirables relevés dans les messages. Une fonction de normalisation est appliquée sur ces résultats (comparaison sur les figures suivantes des courbes avant et après normalisation). La fonction normalisée est une transformation linéaire de la fonction de départ pour obtenir une moyenne nulle et un écart type égal à 1 sur l'intervalle. Si l'hypothèse, que la courbe de l'occurrence des effets indésirables suit une loi normale (Gauss) est posée, la probabilité d'être entre 0 et 1 d'écart type est de 34 %. La probabilité d'être entre -1 et +1 d'écart type est de 68 %. La probabilité que la courbe passe au-dessus de 1 est inférieure à 16 %. La probabilité d'être au-dessus de 2 est de 4,55 %. Ici, le pic de la courbe normalisée culmine à 2,5 sur la période 2016-2020. L'évènement observé ne devrait donc normalement pas se produire (très faible probabilité de l'ordre du pourcentage). C'est un événement notable, statistiquement significatif et de longue durée, qui est observé fin 2017. Ces résultats sont présentés sur les deux pages qui suivent.

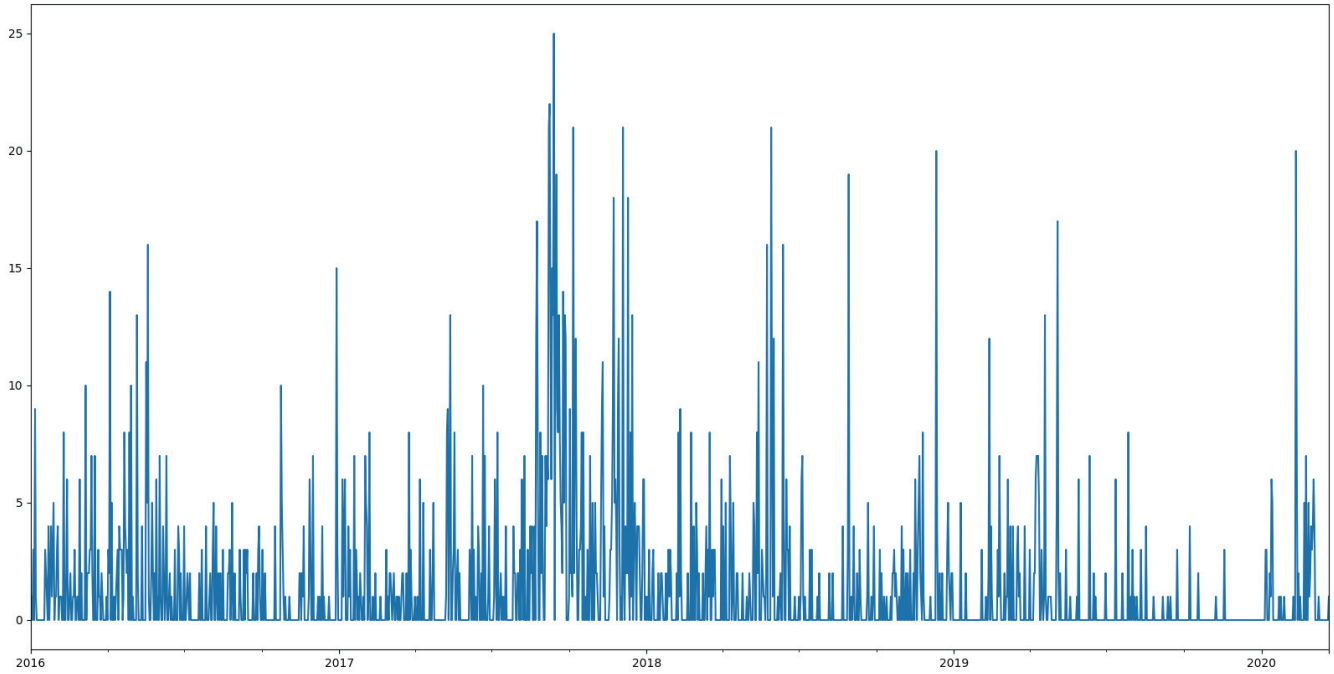

Figure 45 : Occurrence quotidienne des effets indésirables relevés dans les messages (période 2016-2020)

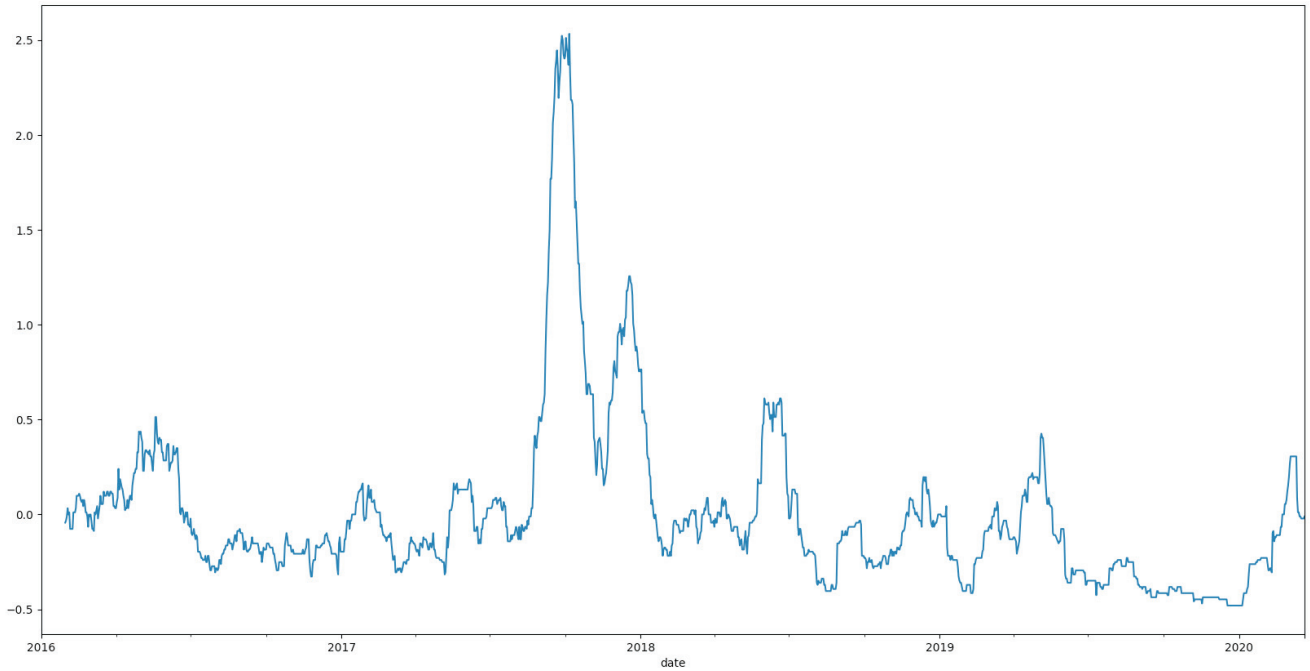

Figure 46 : Normalisation appliquée à l'occurrence quotidienne des effets indésirables relevés dans les messages (période 2016-2020)

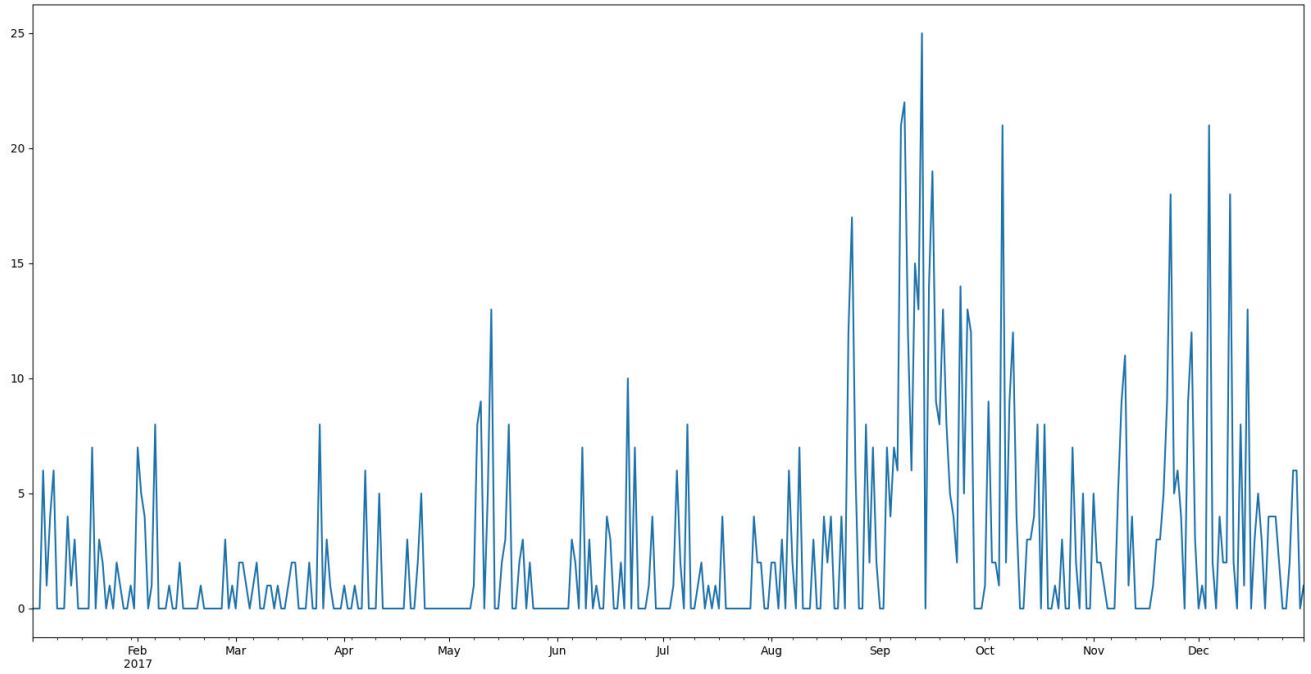

Figure 47 : Occurrence quotidienne des effets indésirables relevés dans les messages en 2017

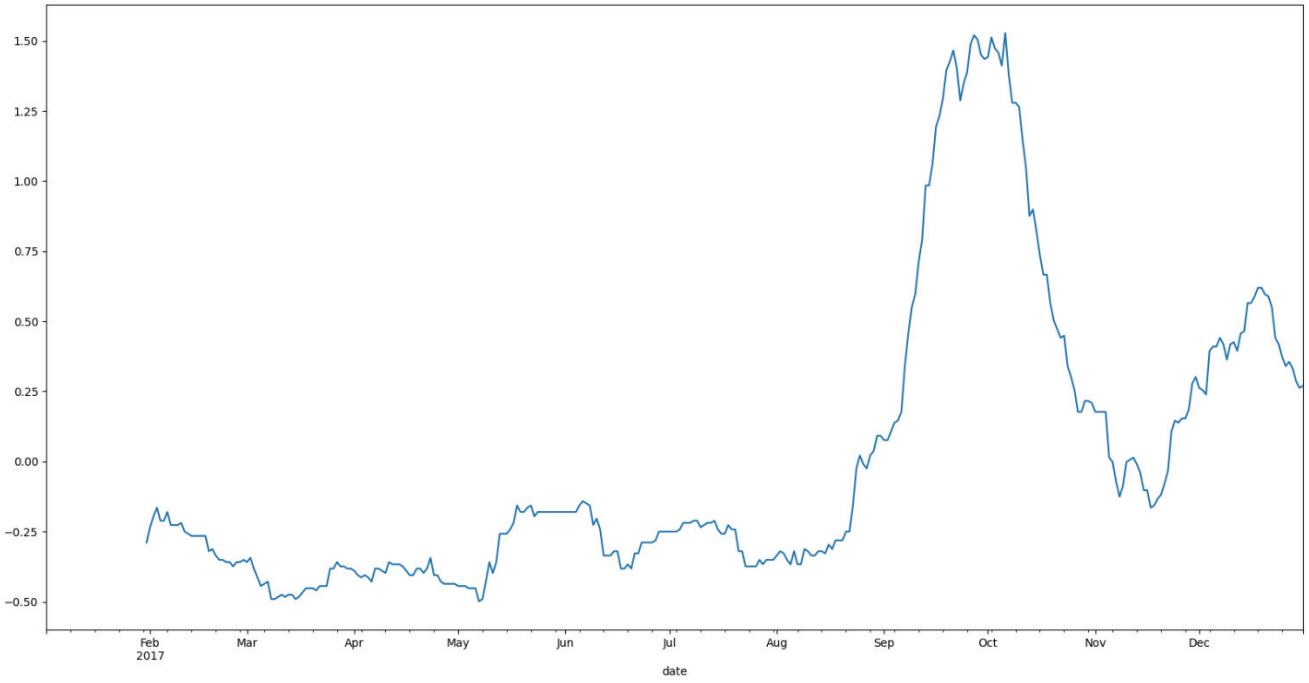

Figure 48 : Normalisation appliquée à l'occurrence quotidienne des effets indésirables relevés dans les messages en 2017

Un réajustement des graphiques de la figure 44 est permis suite à l'obtention d'un corpus contenant les effets secondaires les plus fréquents/probables.

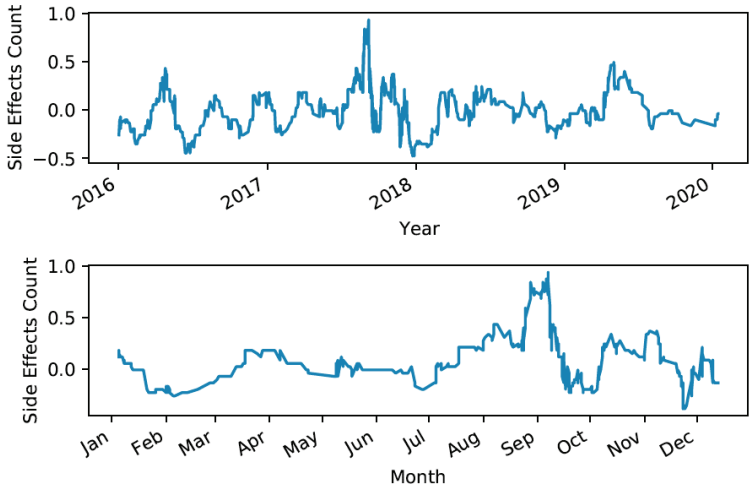

Figure 49 : Visualisation des effets secondaires les plus fréquents issus du corpus analysé

Dernier résultat, la figure suivante présente les meilleures combinaisons d'effets secondaires pour la période 2016-2020 (ensemble des top 10).

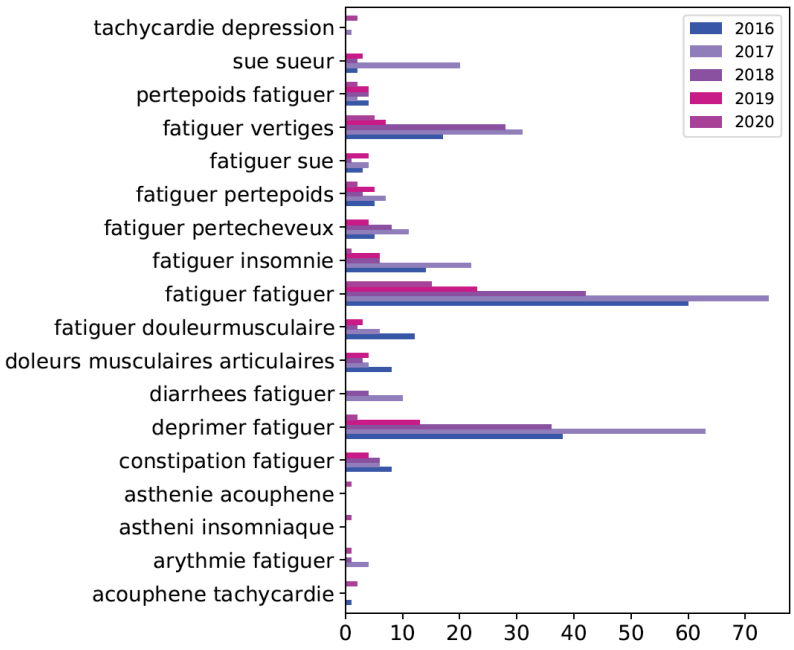

Figure 50 : « Top n-gram » des effets secondaires

3.4 Discussion

Le travail conduit ne permet pas d'explorer toutes les possibilités offertes par le traitement du langage naturel (NLP). La démarche s'est construite de manière itérative, en l'absence de référence. Elle est basée sur l'utilisation de connaissances médicales, en traitement des données et par l'intermédiaire du langage python dont les performances en font un outil de choix à disposition de la communauté scientifique :

- C'est un langage interprété de haut niveau, donc facilement compréhensible, par opposition aux langages de bas niveau tel que « l'assembleur » ou le langage binaire. Celui-ci est très performant car il appelle des bibliothèques objets créées à partir de programmes en C compilés (dans le but d'améliorer la vitesse d'exécution).
- C'est un langage parfaitement adapté à la communauté scientifique puisqu'il dispose de nombreuses et excellentes bibliothèques open source destinées au calcul scientifique, à l'analyse statistique et offrant des solutions performantes dans le domaine de l'IA et de l'apprentissage automatique (pandas, scipy, numpy, la suite Anaconda, etc…).
- Il existe une communauté mondiale de développeurs qui partagent leurs expériences et difficultés.

Assez naturellement, le problème est abordé en explorant les pistes les plus immédiatement accessibles. Des conclusions très intéressantes sont tirées mais, comme lors d'une randonnée en montagne, le premier col cache longtemps le suivant. Le temps et les ressources sont comptés, il faut accepter de s'arrêter ici mais non sans esquisser la suite du voyage.

3.4.1 Points forts et limites de l'expérimentation

L'émergence d'une problématique autour du Levothyrox® est visible dès la fin de l'été 2017 et cette expérience le confirme par la seule analyse des commentaires d'utilisateurs sur une seule plateforme. En s'appuyant sur l'analyse statistique de séries temporelles représentant la fréquence des mots ou de n-grammes, il s'avère que de nombreux mots courants du vocabulaire peu ou pas pertinents viennent polluer le traitement statistique. Il devient dès lors indispensable d'utiliser des terminologies adaptées et codifiées, dans le but de créer de manière efficiente un référentiel médical extrêmement précis. C'est ce qui peut permettre d'élaborer des dictionnaires de mots courants visant à éliminer le bruit de fond et à repérer des effets indésirables de manière reproductible pour toutes les classes de médicaments et toutes les données médicales. Il convient alors d'utiliser des classifications telles que le CIM-10, la classification ATC. Il s'agit alors de rassembler sous une entité bien désignée tous les mots se rapportant à un symptôme précis, en incluant les fautes d'orthographe et les différentes façons de décrire ce symptôme.

L'utilisation d'outils statistiques de base ne permet pas de mettre en évidence une détection précoce d'évènements anormaux. En effet, l'application d'une fonction de normalisation à l'occurrence quotidienne des effets secondaires rapportés dans les messages est la preuve flagrante de l'existence d'un événement anormal et significatif durant la période où le « scandale » a éclaté dans la presse (été 2017). Cependant, elle ne permet pas de mettre en évidence la survenue d'événements anormaux avant juillet 2017.

C'est par l'utilisation d'outils informatiques plus poussés qu'une détection précoce d'évènements indésirables est possible. Le réseau neuronal convolutif, qui est un outil de *deep learning*, permet effectivement de détecter à partir du mois de mai 2017 les premières anormalités.

Ce sont ces outils qui permettront d'améliorer la détection des signaux faibles et donc anticiper les problèmes des médicaments.

Les résultats obtenus sont très encourageants et il existe un réel potentiel pour les améliorer. En effet, dans ce travail, des biais limitent la capacité d'obtention de meilleurs résultats et le déploiement et la généralisation de ces méthodes. Un seul médicament a été étudié, au travers des messages des utilisateurs d'un seul et unique forum de discussion. Le médicament en question a fait l'objet d'une affaire, non du fait de l'action néfaste du principe

actif (par sa nature toxique ou son mécanisme d'action) mais du fait d'un changement de formulation qui a impliqué des ratés de communication. L'effet iatrogène, réel pour certains patients mais rapidement résolu par ajustement posologique, n'est pas la seule cause. Il pourrait être intéressant et très pertinent d'appliquer cette méthodologie sur un nombre minimal de sources d'information (réseaux sociaux) et de spécialités pour comparer les résultats obtenus.

3.4.2 Perspectives, extrapolation à d'autres activités

L'élaboration d'une méthode exploitant les nouvelles possibilités de l'informatique et destinée à introduire une nouvelle discipline appliquée à la santé (pharmacovigilance) est un travail complexe et exigeant. En effet, de nombreuses contraintes sont présentes pour maintenir une bonne qualité des soins : validation par les pairs, robustesse de la conception, preuves scientifiques, démonstration du fonctionnement en situation réelle, etc.

De ce fait, ce travail de recherche constitue un premier socle faisant office de preuve de concept. Pour aller plus loin avec ces nouveaux outils et obtenir des résultats robustes, il s'agit de continuer la recherche et le développement avec une équipe pluridisciplinaire pour aboutir à une méthode reproductible sur l'ensemble de la pharmacopée. Tout l'enjeu de la réussite d'un tel projet réside dans l'obtention d'une quantité suffisante de données sur lesquelles il s'agit d'appliquer un nettoyage précis pour obtenir de la donnée de qualité. A partir de là, il est possible d'envisager tout type d'analyses sur toutes les classes de médicaments confondues, même les traitements peu répandus ou très innovants tels que les CAR-T Cells.

L'objectif initial était la mise en œuvre d'une méthode optimisant la détection du signal en pharmacovigilance. L'exemple du Levothyrox® a été utilisé dans ce cadre mais il est évident que la méthode peut se généraliser à d'autres produits que le médicament. L'évaluation du dispositif médical est également un sujet critique puisque ces produits de santé sont très répandus et ont un cycle de vie globalement plus court que pour le médicament. La surveillance est d'autant plus complexe que les processus de matériovigilance sont moins connus et maîtrisés. L'exemple d'Essure®, peut-être mis en parallèle avec le cœur de ce travail puisqu'internet a permis à de nombreuses victimes de mettre en cause ce produit avant toute réaction des autorités sanitaires. Essure® est un dispositif médical de classe III visant à la stérilisation définitive de la femme en âge de procréer (hystérectomie), développé puis commercialisé en France dès 2012 par Conceptus® puis racheté par Bayer® en 2013. Cet implant, conçu à partir de polymères et d'alliages métalliques et destiné à être introduit dans les trompes de Fallope par hystéroscopie, s'est vu impliqué dans 42 signalements à l'ANSM durant l'année de commercialisation. En 2015, ce dispositif médical est placé sous surveillance renforcée par le ministère de la Santé puis est suspendu le 3 août 2017 (79). Pendant ce temps, l'association R.E.S.I.S.T a recensé 2884 victimes françaises ayant ressenties différents symptômes allant de l'asthénie à des perforations en passant les ménorragies (80). Pour certaines, l'errance thérapeutique fût réelle et de longue durée (81).

L'agroalimentaire pourrait aussi bénéficier de ce type d'outils, du fait des nombreux risques et incidents imputables aux produits alimentaires (allant de l'intoxication alimentaire à la présence d'objets étrangers dans les produits).

Concernant les pistes d'évolutions du projet, plusieurs évidences existent à ce stade.

Twitter® est le site internet le plus utilisé au monde, aussi, les tweets nécessitent moins de nettoyage du fait qu'ils soient limités en nombre de caractères, il pourrait être intéressant de réaliser une étude similaire sur cette source d'informations.

L'entrainement des algorithmes de *machine learning* sur une base de données en français et ayant pour thèmes principaux des sujets médicaux permettrait d'augmenter la performance des algorithmes.

Comme présenté dans la partie « points forts et limites de l'expérimentation », il est possible d'améliorer le processus de détection d'événements anormaux en associant des techniques de traitement du langage naturel avec les bases de connaissances structurées. Deux librairies python peuvent être envisagées pour mettre l'accent sur la détection des entités nommées et leurs relations, ainsi que les événements clés des commentaires échangés. Il s'agit de NLTK et de Spacy. Une autre technique qui n'a pas été utilisée mais qu'il serait intéressant d'envisager est l'application de techniques du *topic modelling* pour suivre l'évolution et la corrélation des sujets principaux évoqués par les patients. Comme présenté dans la partie « L'identification de thèmes (Topic modeling) », l'utilisation des algorithmes LDA, LSA et NMF permettent la mise en place de ces analyses.

Enfin, après avoir mis en place toutes ces techniques, l'utilisation du *clustering* permettrait de combiner toutes les caractéristiques de chaque commentaire pour en faire ressortir des différences et similarités. Le *clustering* est une technique d'apprentissage non-supervisé qui permet de rechercher les points communs et différences des entités analysées (commentaires ou périodes dans ce projet). Ici, les différentes caractéristiques pourraient être :

- Les mots identifiés comme pertinents présents dans chaque commentaire ou période (effets indésirables, médicaments, etc…).
- Le sujet principal de chaque commentaire ou période.
- L'association des sujets principaux de chaque commentaire ou période en fonction de la plus forte corrélation obtenue.
- Le sentiment positif ou négatif de chaque commentaire ou période.

- Les bi-grammes les plus fréquents présents dans chaque commentaire ou période.
- Etc.

Le *clustering* cherche à identifier ce qui n'est pas visible dans la donnée et regroupe les commentaires ou périodes les plus proches. Cette technique permet d'identifier les périodes au cours desquelles les échanges des patients sont similaires. Pour mettre au point une méthode visant à détecter des événements indésirables, l'affaire du Levothyrox® n'est peut-être pas le meilleur exemple. Il serait intéressant de reproduire le raisonnement sur d'autres spécialités pharmaceutiques considérées comme ayant un bénéfice/risque faible et/ou une forte iatrogénie et pour lesquelles de la donnée est accessible sur les réseaux sociaux.

4. Conclusions

THESE SOUTENUE PAR : M. ROBERT Jean-Philippe et M. ROCHE Valentin

La nouvelle formule du Levothyrox® a été commercialisée en France en mars 2017. Une augmentation de la fréquence des effets indésirables dû à la prise du médicament a été identifiée et relayée dans les médias à partir du mois de juillet 2017. En réponse à cet incident, le laboratoire Merck® a réintroduit en France l'ancienne formule du Levothyrox® en octobre 2017. L'objectif de ce projet était d'élaborer une méthode pour détecter plus précocement, via des algorithmes informatiques, les effets indésirables du médicament mentionnés sur le sous-forum endocrinologie du site Doctissimo®. La construction du prototype a été réalisée selon un mode séquentiel et empirique en l'absence de méthode de référence. Ce travail s'intègre dans une démarche d'optimisation de la pharmacovigilance en envisageant l'utilisation des data sciences pour collecter les données de vie réelle des patients.

La première phase du projet consistait à extraire puis nettoyer les données. La deuxième phase qui analysait les meilleurs bi-grammes sur la période 2016 à 2020 révélais un seul et unique pic de fréquence entre août 2017 et janvier 2018. Le bi-gramme le plus fréquent (« Ancien Formule »), a un taux d'apparition 160 fois supérieur en août 2017 qu'en mars 2017 et 6 fois supérieur qu'en janvier 2018. « Effets secondaires » apparaît 8 fois en 2016, 14 fois en 2019 et 183 fois en 2017. Le même profil est observé pour d'autres bi-grammes (« Nouvelle Formule », « Ancien Formule », « Doser Levothyrox »). La troisième phase du traitement algorithmique visait à extraire et à analyser les effets indésirables. Mi-septembre 2017, il est constaté dans les commentaires une fréquence d'apparition moyenne de 25 par jour contre 6 par jour entre 2016 et 2020. L'application d'une fonction de normalisation à la courbe d'occurrence des effets indésirables en fonction du temps concluait qu'un événement inhabituel et significatif s'est bel et bien produit en 2017. La quatrième phase a permis, via l'utilisation de réseaux de neurones convolutifs, d'identifier des similarités entre les termes utilisés durant la période de mai 2017 à février 2018. Aussi, l'algorithme a confirmé que les termes cités pendant la période d'anormalité sont différents de ceux de la période de normalité. Ainsi, il a été conclu qu'une détection de signaux précoces, indicateurs d'évènements anormaux était possible à partir du mois de mai 2017.

Ce travail, précurseur dans le domaine de la pharmacovigilance présente des résultats très encourageants. Ils mettent en évidence une capacité réelle d'innovation concernant l'utilisation des data sciences et de l'intelligence artificielle à des fins de traitement statistiques des données de vie réelle des patients. Il permet d'envisager, après ajustement des biais et mise en place des pistes d'améliorations, l'extrapolation du modèle à d'autres sources de données et d'autres scripts d'analyses (avec ou sans l'utilisation de l'intelligence artificielle). Toutefois, le chemin est encore long pour espérer rencontrer ces outils dans la pratique quotidienne. La formation d'une équipe pluridisciplinaire sera un prérequis pour construire un outil applicable à tous les médicaments et pathologies sur tous les réseaux sociaux.

Le Président de la thèse,
Nom : Pr. C. Dussart

Vu et permis d'imprimer, Lyon, le 14 FEV. 2022
Vu, le Directeur de l'Institut des Sciences Pharmaceutiques et Biologiques, Faculté de Pharmacie

Signature :

Pour le Président de l'Université Claude Bernard Lyon 1,

Professeur C. DUSSART

Bibliographie

1. We Are Social. Digital report 2021 : Les dernières données de notre état des lieux du digital dans le monde. [En ligne]. 2021. [cité le 17 février 2022]. Disponible : https://wearesocial.com/fr/blog/2021/01/digital-report-2021-les-dernieres-donnees-de-notre-etat-des-lieux-du-digital-dans-le-monde/

2. Estaquio C, Castetbon K, Valeix P. Maladies thyroïdiennes dans la cohorte SU.VI.MAX. Estimation de leur incidence et des facteurs de risques associés. 2009. 59 p. Disponible : https://www.santepubliquefrance.fr/docs/maladies-thyroidiennes-dans-la-cohorte-su.vi.max.-estimation-de-leur-incidence-et-des-facteurs-de-risque-associes-1994-2002

3. Prescrire. Lévothyrox : des milliers de signalements en lien avec un changement de formulation. Prescrire. 2017;37(n°408):756

4. ANSM. Lévothyrox : changement de formule et de couleurs des boîtes et blisters. [En ligne]. 2017. [cité le 17 février 2022]. Disponible : https://ansm.sante.fr/actualites/levothyrox-levothyroxine-changement-de-formule-et-de-couleur-des-boites

5. Sénat. Changement de formule du Lévothyrox pour les personnes souffrant de troubles thyroïdiens. [En ligne]. 2018. [cité le 17 février 2022]. Disponible : https://www.senat.fr/questions/base/2018/qSEQ180404650.html

6. Casassus B. Risks of reformulation: French patients complain after Merck modifies levothyroxine pills. BMJ. 16 février 2018;360:714.

7. Mes opinions. Pétition contre le nouveau Lévothyrox dangereux pour les patients ! [En ligne]. 2017. [cité le 17 février 2022]. Disponible : https://www.mesopinions.com/petition/sante/contre-nouveau-levothyrox-dangereux-patients/31185

8. Sénat. Effets indésirables graves de la nouvelle formule du Levothyrox. [En ligne]. 2017. [cité le 17 février 2022]. Disponible : https://www.senat.fr/questions/base/2017/qSEQ170901196.html

9. Ladepeche. Lévothyrox : une nouvelle action collective lancée contre l'Agence nationale de sécurité du médicament. [En ligne] 2017. [cité le 17 février 2022]. Disponible :

https://www.ladepeche.fr/amp/2021/09/06/levothyrox-une-nouvelle-action-collective-lancee-contre-lagence-nationale-de-securite-du-medicament-9772637.php

10. Ladepeche. Les manquements du laboratoire Merck sur le Levothyrox pointés par un rapport d'expertise. [En ligne] 2021. [cité le 17 février 2022]. Disponible : https://www.ladepeche.fr/2021/05/25/levothyrox-les-manquements-du-laboratoire-pointes-par-un-rapport-dexpertise-9565597.php

11. Mes opinions. #Levothyrox : Madame Buzyn laissez la justice faire son travail, ne l'entravez pas ! [En ligne]. 2017. [cité le 17 février 2022] Disponible : https://www.mesopinions.com/petition/sante/levothyrox-madame-buzyn-laissez-justice-faire/35862

12. Le monde. Levothyrox : les experts judiciaires éreintent Merck et les autorités sanitaires. [En ligne]. 2021. [cité le 17 février 2022] Disponible : https://www.lemonde.fr/planete/article/2021/05/28/levothyrox-les-experts-judiciaires-ereintent-merck-et-les-autorites-sanitaires_6081795_3244.html

13. Le Quotidien du Médecin. Lévothyrox : un rapport d'expertise judiciaire pointerait des manquements du laboratoire Merck et de l'ANSM. [En ligne]. 2021. [cité le 17 février 2022] Disponible : https://www.lequotidiendumedecin.fr/actus-medicales/medicament/levothyrox-un-rapport-dexpertise-judiciaire-pointerait-des-manquements-du-laboratoire-merck-et-de

14. ACTMS. Biodisponibilité et bioéquivalence [En ligne]. Date non communiquée. [cité le 17 février 2022]. Disponible : https://www.cadth.ca/sites/default/files/pdf/What_Are_Bioavailability_and_Bioequivalence_f.pdf

15. EMA. Committee for Medicinal Products for Human Use. Guideline on the investigation of bioequivalence. [En ligne]. 2010. [cité le 17 février 2022] Disponible : http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/01/WC500070039.pdf

16. FDA. Statistical Approaches to Establishing Bioequivalence. [En ligne]. 2020. [cité le 17 février 2022]. Disponible : https://www.fda.gov/regulatory-information/search-fda-guidance-documents/statistical-approaches-establishing-bioequivalence

17. Toutain PL, Gandia P, Bousquet-Mélou A. Problèmes et difficultés rencontrés lors de

la planification et de l'analyse d'un essai de bioéquivalence. La Lettre du Pharmacologue. 2014;28(n°2):37

18. EMA. Overview of comments received on draft guideline on the investigation of bioequivalence [En ligne]. 2010. [cité le 17 février 2022] Disponible : https://www.ema.europa.eu/en/documents/other/overview-comments-received-draft-guideline-investigation-bioequivalence-cpmp/ewp/qwp/1401/98-rev-1_en.pdf

19. Morais J, Lobato M. The new European Medicines Agency guideline on the investigation of bioequivalence. Basic Clin Pharmacol Toxicol. 2010;106(n°3):221-5.

20. ANSM. L'ANSM publie les résultats des enquêtes nationales de pharmacovigilance sur les spécialités à base de lévothyroxine [En ligne]. 2018. [cité le 17 février 2022] Disponible : https://ansm.sante.fr/actualites/lansm-publie-les-resultats-des-enquetes-nationales-de-pharmacovigilance-sur-les-specialites-a-base-de-levothyroxine-communique

21. Chen M, Patnaik R, Hauck W, Schuirmann D, Hyslop T, Williams R. An individual bioequivalence criterion: regulatory considerations. Statistics in Medicine. 2000;19(n°20):2821-42.

22. Anderson S, Hauck W. Consideration of individual bioequivalence. J Pharmacokinet Biopharm. 1990;18:259-73.

23. Concordet D, Gandia P, Montastruc J-L, Bousquet-Mélou A, Lees P, Ferran A, et al. Levothyrox® New and Old Formulations: Are they Switchable for Millions of Patients?. Clin Pharmacokinet. 2019;58(n°7):827-33.

24. Lindenberg M, Kopp S, Dressman J. Classification of orally administered drugs on the World Health Organization Model List of Essential Medicines according to the Biopharmaceutics Classification System. Eur J Pharm Biopharm. 2004;58:265-78.

25. Blume H, Schug B. The biopharmaceutics classification system (BCS): class III drugs—better candidates for BA/BE waiver?. Eur J Pharm Sci. 1999;9:117-21.

26. Chen M, Sadrieh N, Yu L. Impact of osmotically active excipients on bioavailability and bioequivalence of BCS class III drugs. AAPS J. 2013;15:1043-50.

27. Adkin D, Davis S, Sparrow R, Huckle P, Wilding I. The effect of mannitol on the oral bioavailability of cimetidine. J Pharm Sci. 1995;84:1405-9.

28. Garcia-Arieta A. Interactions between active pharmaceutical ingredients and excipients affecting bioavailability: impact on bioequivalence. Eur J Pharm Sci. 2014;65:89-97.

29. CultureSciencesChimie. Concevoir des candidats médicaments sur Internet. [En ligne]. 2020. [cité le 17 février 2022]. Disponible : https://culturesciences.chimie.ens.fr/thematiques/chimie-organique/chimie-pharmaceutique/concevoir-des-candidats-medicaments-sur-0

30. Leem. Recherche et développement. [En ligne]. 2021. [cité le 17 février 2022]. Disponible : https://www.leem.org/recherche-et-developpement

31. Wikipedia. Essai clinique. [En ligne]. 2021. [cité le 17 février 2022]. Disponible : https://fr.wikipedia.org/w/index.php?title=Essai_clinique&oldid=186870385

32. ANSM. Bonnes pratiques de laboratoire. [En ligne]. 2021. [cité le 17 février 2022]. Disponible : https://ansm.sante.fr/documents/reference/bonnes-pratiques-de-laboratoire

33. Wikipedia. Sildénafil. [En ligne]. 2021. [cité le 17 février 2022]. Disponible : https://fr.wikipedia.org/w/index.php?title=Sild%C3%A9nafil&oldid=183953996

34. Le manuel MSD. Effets indésirables des médicaments. [En ligne]. 2021. [cité le 17 février 2022]. Disponible : https://www.msdmanuals.com/fr/professional/pharmacologie-clinique/effets-ind%C3%A9sirables-des-m%C3%A9dicaments/effets-ind%C3%A9sirables-des-m%C3%A9dicaments

35. WHO. Identification, évaluation d'un signalement et déclenchement d'une alerte. [En ligne]. 2005. [cité le 17 février 2022]. Disponible : https://www.who.int/medicines/areas/quality_safety/safety_efficacy/trainingcourses/5evaluation_alerte.pdf

36. Ministère des Solidarités et de la Santé. La déclaration des effets indésirables. [En ligne]. 2018. [cité le 17 février 2022]. Disponible : https://solidarites-sante.gouv.fr/soins-et-maladies/medicaments/la-surveillance-des-medicaments/article/la-declaration-des-effets-indesirables

37. Microsoft experiences . Apprentissage supervisé et non supervisé. [En ligne]. 2020. [cité le 17 février 2022]. Disponible : https://experiences.microsoft.fr/articles/intelligence-artificielle/apprentissage-supervise-et-non-supervise-quelles-differences/

38. DataScientest. Apprentissage non supervisé. [En ligne]. 2021. [cité le 17 février 2022]. Disponible : https://datascientest.com/apprentissage-non-supervise

39. Lateral. A fastText-based hybrid recommender. [En ligne]. 2016. [cité le 17 février 2022]. Disponible : https://www.lateral.io/resources-blog/fasttext-based-hybrid-recommender

40. Haute autorité de santé. Construction et dialogue des savoirs, vers de meilleures décisions individuelles et collectives en santé. [En ligne]. 2019. [cité le 17 février 2022]. Disponible : https://www.has-sante.fr/upload/docs/application/pdf/2019-11/colloque-has-d-polton.pdf

41. Bégaud B, Polton D, Von Lennep F. Les données de vie réelle, un enjeu majeur pour la qualité des soins et la régulation du système de santé L'exemple du médicament. Ministère de Solidarités et de la Santé ; 2017. 105. Disponible : https://solidarites-sante.gouv.fr/IMG/pdf/rapport_donnees_de_vie_reelle_medicaments_mai_2017vf.pdf

42. Fournier A, Zureik M. Estimate of deaths due to valvular insufficiency attributable to the use of benfluorex in France. Pharmacoepidemiol Drug Saf. 2012;21(n°4):343-51.

43. COFER. Item 169 : Évaluation thérapeutique et niveau de preuve. [En ligne]. 2011. [cité le 17 février 2022]. Disponible : http://campus.cerimes.fr/rhumatologie/enseignement/rhumato24/site/html/cours.pdf

44. BENSADON A-C, Marie E, Morelle A. Rapport sur la pharmacovigilance et gouvernance de la chaîne du médicament. Inspection Générale des Affaires Sociales ; 2011. 103 p. Disponible : https://www.igas.gouv.fr/IMG/pdf/Synthese_RM2011-103P_pharmacovigilance.pdf

45. Humbert X, Chrétien B, Sassier M, Coquerel A, Alexandre J, Fedrizzi S. Évaluation d'un nouvel outil en pharmacovigilance : la déclaration simplifiée en ligne pour les médecins généralistes. Sante Publique (Bucur). 2018;30(n°2):225-32.

46. American Medical Association. AMA passes first policy recommendations on augmented intelligence. [En ligne]. 2018. [cité le 17 février 2022]. Disponible : https://www.ama-assn.org/press-center/press-releases/ama-passes-first-policy-recommendations-augmented-intelligence

47. Chen M, Decary M. Artificial intelligence in healthcare: An essential guide for health leaders. Healthc Manage Forum. 2020;33(n°1):10-8.

48. FDA. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. FDA. [En ligne]. 2021. [cité le 17 février 2022]. Disponible : https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

49. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(n°7873):583-9.

50. Forbes. IBM's Watson Gets Its First Piece Of Business In Healthcare. [En ligne]. 2021. [cité le 17 février 2022]. Disponible : https://www.forbes.com/sites/bruceupbin/2013/02/08/ibms-watson-gets-its-first-piece-of-business-in-healthcare/?sh=74a9392d5402

51. Lee K-F. AI Superpowers: China, Silicon Valley, and the New World Order. Houghton Mifflin Harcourt; 2018. 275 p

52. Yan NT, Lu C, Wang F. China Focus: Tech giants tap into AI healthcare market. Xinhua. 2018

53. EMA. Better vigilance for health protection and innovation : Overviw of the new EU pharmacovigilance legislation. [En ligne]. 2015. [cité le 17 février 2022]. Disponible : https://www.ema.europa.eu/en/documents/leaflet/better-vigilance-health-protection-innovation_en.pdf

54. Arcizet J, Leroy B, Renzullo C, Mondoloni P, Donier L, Penaud J-F, et al. Iatrogénie médicamenteuse responsable d'hospitalisation en réanimation : étude descriptive dans un centre hospitalier. J Pharm Clin. 2018;37(n°2):111-20.

55. AMELI. La iatrogénie médicamenteuse. [En ligne]. 2022. [cité le 17 février 2022]. Disponible : https://www.ameli.fr/assure/sante/medicaments/medicaments-et-situation-de-vie/iatrogenie-medicamenteuse

56. Jiménez-Zafra SM, Martín-Valdivia MT, Molina-González MD, Ureña-López LA. How do we talk about doctors and drugs? Sentiment analysis in forums expressing opinions for medical domain. Artif Intell Med. 2019;93:50-7.

57. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online. J Med Internet Res. 2013;15(n°11):239.

58. Bekhuis T, Kreinacke M, Spallek H, Song M, O'Donnell JA. Using Natural Language Processing to Enable In-depth Analysis of Clinical Messages Posted to an Internet Mailing List: A Feasibility Study. J Med Internet Res. 2011;13(n°4):e1799.

59. Doing-Harris KM, Zeng-Treitler Q. Computer-Assisted Update of a Consumer Health Vocabulary Through Mining of Social Network Data. J Med Internet Res. 2011;13(n°2):37.

60. Bigeard E, Grabar N. Detection and analysis of medical misbehavior in online forums. HAL. 2020. 7 p. Disponible : https://hal.archives-ouvertes.fr/hal-02430533/document

61. Lardon J, Abdellaoui R, Bellet F, Asfari H, Souvignet J, Texier N, et al. Adverse Drug Reaction Identification and Extraction in Social Media: A Scoping Review. J Med Internet Res. 2015;17(n°7):171.

62. Arnoux-Guenegou A, Girardeau Y, Chen X, Deldossi M, Aboukhamis R, Faviez C, et al. The Adverse Drug Reactions From Patient Reports in Social Media Project: Protocol for an Evaluation Against a Gold Standard. JMIR Res Protoc. 2019;8(n°5):11448.

63. El-allaly E, Sarrouti M, En-Nahnahi N, Ouatik El Alaoui S. An adverse drug effect mentions extraction method based on weighted online recurrent extreme learning machine. Comput Methods Programs Biomed. 2019;176:33-41.

64. Bollegala D, Maskell S, Sloane R, Hajne J, Pirmohamed M. Causality Patterns for Detecting Adverse Drug Reactions From Social Media: Text Mining Approach. JMIR Public Health Surveill. 2018;4(n°2):51.

65. Abdellaoui R, Schück S, Texier N, Burgun A. Filtering Entities to Optimize Identification of Adverse Drug Reaction From Social Media: How Can the Number of Words Between Entities in the Messages Help? JMIR Public Health Surveill. 22 juin 2017;3(n°2):6577.

66. Lee CY, Chen Y-PP. Prediction of drug adverse events using deep learning in pharmaceutical discovery. Brief Bioinform. 2021;22(n°2):1884-901.

67. Cocos A, Fiks AG, Masino AJ. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. J Am Med Inform Assoc JAMIA. 2017;24(n°4):813-21.

68. Fan B, Fan W, Smith C, Garner H "Skip". Adverse drug event detection and extraction from open data: A deep learning approach. Inf Process Manag. 2020;57(n°1):102131.

69. Park SH, Hong SH. Identification of Primary Medication Concerns Regarding Thyroid Hormone Replacement Therapy From Online Patient Medication Reviews: Text Mining of Social Network Data. J Med Internet Res. 2018;20(n°10):11085.

70. Amosse E. Event detection and analysis on short text messages [Thèse de doctorat]. Université Côte d'Azur ; 2017

71. Dou W, Wang X, Ribarsky W, Zhou M. Event Detection in Social Media Data. University of North Carolina, IBM Almaden Research Center. 2012. 4 p. Disponible : https://webpages.charlotte.edu/wdou1/publications/2012/Dou-EventDetectionTasks-2012.pdf

72. Dou W, Wang X, Skau D, Ribarsky W, Zhou M. LeadLine: Interactive visual analysis of text data through event identification and exploration.IEEE. 2012. Disponible : https://ieeexplore.ieee.org/abstract/document/6400485

73. Farzindar A, Khreich W. A Survey of Techniques for Event Detection in Twitter. Comput Intelligence. 2015;31:132–32

74. Ritter A, Clark S, Mausam, Etzioni O. Named Entity Recognition in Tweets: An Experimental Study. Association for Computational Linguistics. 2011. p 1524-34. Disponible : https://aclanthology.org/D11-1141.pdf

75. Digimind. La voix des patients et des laboratoires sur le web et les médias sociaux [En ligne]. 2019. [cité le 17 février 2022]. Disponible : https://cdn2.hubspot.net/hubfs/636866/DIGIMIND-Industrie_Pharmaceutique-Voix-des-patients-et-laboratoires-web-et-medias-sociaux.pdf

76. Wikipedia. Réseau neuronal convolutif. [En ligne]. 2021. [cité le 17 février 2022]. Disponible : https://fr.wikipedia.org/wiki/Réseau_neuronal_convolutif

77. SuperDataScience Team. Convolutional Neural Networks (CNN): Step 3 - Flattening. [En ligne]. 2018. [cité le 17 février 2022]. Disponible : https://www.superdatascience.com/blogs/convolutional-neural-networks-cnn-step-3-flattening

78. Wikipedia. Abandon (réseaux neuronaux). [En ligne]. 2022. [cité le 17 février 2022]. Disponible : https://fr.wikipedia.org/wiki/Abandon_(réseaux_neuronaux)

79. Dupont E. Stérilisation tubaire par ESSURE® : prévalence des effets indésirables, à propos d'une cohorte de 642 patientes opérées au CHU Amiens-Picardie. [Mémoire]. Amiens, France : Université de Picardie Jules Verne ; 2018

80. Lindheim SR, Madeira JL, Bagavath B, Petrozza JC. Social media and Essure hysteroscopic sterilization: a perfect storm. Fertil Steril. 2019;111(n°6):1105-6.

81. Jegaden M, Pourcelot A-G, Fernandez H, Capmas P. Surgical removal of essure® micro inserts by vaginal hysterectomy or laparoscopic salpingectomy with cornuectomy: Case series and follow up survey about device-attributed symptoms resolution. J Gynecol Obstet Hum Reprod. 2020;49(n°8):101781.

Agence nationale de sécurité du médicament et des produits de santé

Questions/Réponses

Levothyrox : changement de formule et de couleurs des boîtes et blisters

---

1. A quoi sert la lévothyroxine ?

La lévothyroxine est une hormone de substitution thyroïdienne utilisée dans les hypothyroïdies (insuffisance de sécrétion de la glande thyroïde ou absence de celle-ci) ou dans les situations où il est nécessaire de freiner la sécrétion d'une hormone stimulant la thyroïde, appelée TSH (*Thyroid stimulating hormone*).

Pour plus d'informations concernant les propriétés de la lévothyroxine, vous pouvez consulter la base de données publique des médicaments à cette adresse : http://base-donnees-publique.medicaments.gouv.fr/

2. Mon médecin m'a dit que la formule de Levothyrox® allait changer, quels sont les changements ?

Les changements effectués sont :

- une optimisation de la formule visant à garantir une teneur en substance active (la lévothyroxine) plus constante pendant toute la durée de conservation du produit,
- la suppression d'un excipient à effet notoire : le lactose.
- Le format, les couleurs des boites et des blisters ont changé : **pensez à bien vérifier le dosage indiqué sur la boite et celui de votre ordonnance.**

3. Le pharmacien m'a délivré une boite différente de celle que je prends d'habitude, que dois-je faire? Est-ce que je dois aller voir mon médecin ?

Même si la boite et le blister de votre médicament a pu changer de couleur (dans un souci d'harmonisation avec les autres pays d'Europe), votre pharmacien vous a délivré le même dosage en Levothyrox®. Cela ne change rien pour vous. Il vous suffit de :

- Bien vérifier le nom et le dosage du médicament qui vous a été délivré.
- Prendre les nouveaux comprimés de Levothyrox® exactement de la même façon que vous preniez l'ancienne formule.

En effet, la dose de lévothyroxine que vous prenez est ajustée en fonction de vos besoins, votre suivi thyroïdien n'est donc pas modifié.

Cependant, nous vous recommandons de contacter votre médecin pour contrôler votre TSH dans les 6 à 8 semaines après le début de la prise de la nouvelle formule si :

- votre équilibre thérapeutique a été particulièrement difficile à atteindre
- vous avez un cancer de la thyroïde
- vous avez une maladie cardiovasculaire (insuffisance cardiaque ou coronarienne et/ou des troubles du rythme)
- le patient ou la patiente est un enfant
- le patient ou la patiente est une personne âgée

Si vous êtes enceinte, nous vous recommandons de contacter votre médecin pour contrôler votre TSH dans les 4 semaines après le début de la prise de la nouvelle formule

4. Quels sont les risques liés au changement de formule ?

Aucun changement du profil de tolérance n'est attendu, le principe actif restant de la lévothyroxine sodique de même source.

La bioéquivalence entre l'ancienne et la nouvelle formule a été démontrée.
Seuls les excipients ont été modifiés. La bioéquivalence entre l'ancienne et la nouvelle formule a été démontrée par des études de biodisponibilité. Il a ainsi été mis en évidence que les nouveaux excipients ne modifient ni la quantité de substance active qui passe dans le sang, ni la vitesse à laquelle elle atteint l'organe cible. Cette bioéquivalence est la garantie d'une efficacité et d'une sécurité identique à celle de celle l'ancienne formule.

Par mesure de précaution, si vous pensez avoir des symptômes traduisant un déséquilibre thyroïdien (cf question 5 «Quels sont les symptômes qui doivent m'alerter sur un déséquilibre thyroïdien ?») nous vous recommandons de contacter votre médecin pour contrôler votre TSH.

5. Quels sont les symptômes qui doivent m'alerter sur un déséquilibre thyroïdien ?

Les symptômes cliniques d'un déséquilibre thyroïdien ne sont pas très spécifiques et restent variables d'un patient à l'autre.

Hypothyroïdie : une fatigue inhabituelle, une constipation, une sensation de ralentissement général sont les symptômes les plus fréquents liés à un taux insuffisant d'hormone thyroïdienne.

Hyperthyroïdie : des sueurs, une tachycardie, des palpitations, une excitation sont des symptômes évoquant un taux trop élevé d'hormones thyroïdiennes.

La probabilité de survenue de ces symptômes lors d'une substitution dose pour dose de Levothyrox® est faible et leur absence ne suffit pas à prédire que l'équilibre thérapeutique soit bon. D'où la nécessité de recourir à des dosages hormonaux (TSH) lors de la surveillance de ce traitement.
Si votre état clinique est stable et que vous ne présentez pas les caractéristiques citées ci-dessus (enfant, personne âgée, femme enceinte, équilibre thérapeutique difficile à atteindre, cancer de la thyroïde ou une maladie cardiovasculaire), un dosage, une à deux fois par an, est suffisant.

Comme pour tout médicament, en cas d'évènement indésirable ou pour toute question relative à la prise du médicament, vous ne devez pas hésiter à consulter votre médecin. Vous pouvez également déclarer les effets indésirables directement via le système national de déclaration auprès de l'Agence nationale de sécurité du médicament et des produits de santé (ANSM) et du réseau des Centres Régionaux de Pharmacovigilance, via le site internet de l'ANSM, rubrique Déclarer un effet indésirable: http://ansm.sante.fr

6. Que faire si je ressens ces symptômes ?

Si vous constatez un des symptômes décrits ci-dessus (voir question 5) ou tout autre symptôme inhabituel, cela peut être la conséquence d'un déséquilibre thyroïdien ou d'une autre pathologie. Aussi, l'ANSM vous recommande de prendre contact, dans les meilleurs délais, avec votre médecin.

7. Est-ce que les modalités de prise de cette nouvelle formule sont différentes ?

Non, les modalités de prise de votre médicament sont inchangées. Dans tous les cas, il est essentiel de toujours respecter la posologie, les modalités de prise et de suivi indiquées par votre médecin.

8. A qui puis-je m'adresser pour obtenir des informations complémentaires ?

N'hésitez pas à interroger votre pharmacien, votre médecin généraliste, votre endocrinologue, qui pourront vous apporter des informations complémentaires.

9. A quelle date cette nouvelle formule sera disponible ?

Elle sera disponible à partir de fin mars 2017, progressivement pour l'ensemble des dosages de la gamme.

10. Que dois-je faire de mes anciennes boites de Levothyrox® ?

Vous pouvez utiliser toutes les boites de l'ancienne formule si vous les avez conservées correctement. Dès lors que vous avez commencé à utiliser la nouvelle formule, il est recommandé de rester sur cette formule.

11. Et si jamais je mélange les deux formules, est-ce qu'il y a un risque pour ma santé ?

Non, cela ne présente pas de risque pour votre santé. Cependant, par mesure de précaution, si vous pensez avoir des symptômes traduisant un déséquilibre thyroïdien (cf. question 5 « Quels sont les symptômes qui doivent m'alerter sur un déséquilibre thyroïdien ?») nous vous recommandons de contacter votre médecin pour contrôler votre TSH.

12. Que faire si je change de pharmacie et que je bénéficie d'une délivrance de formule différente ?

Si le pharmacien vous délivre une boite de nouvelle formule pour un dosage donné pour la première fois : ceci est tout à fait normal, les modalités de prise de votre médicament sont inchangées. Dès lors que vous avez commencé à utiliser la nouvelle formule, il est recommandé de rester sur cette formule.

Si le pharmacien vous délivre une boite de l'ancienne formule alors que vous êtes déjà passé à la nouvelle formule pour un dosage donné, signalez-le au pharmacien lors de la délivrance, il n'est pas recommandé d'utiliser l'ancienne formule après un passage à la nouvelle formule.

13. Que faire si je prends des dosages différents avec une boîte de l'ancienne formule et une boîte de la nouvelle formule ?

Si vous n'avez pas d'autre choix, il est possible de « panacher » les boîtes (exemple : boite de Levothyrox 25 µg de l'ancienne formule et boite de Levothyrox 100 µg de la nouvelle formule), cependant il conviendra de vérifier la TSH en cas de symptômes alertant sur un déséquilibre thyroïdien.

Annexe 1 : Note de l'ANSM à la parution de la nouvelle formule du Levothyrox®

INFORMATIONS SÉCURITÉ PATIENTS

MERCK

INFORMATION TRANSMISE SOUS L'AUTORITE DE L'ANSM

Lettre aux professionnels de santé

27 Février 2017

LEVOTHYROX® (levothyroxine) comprimés sécables nouvelle formule : suivi des patients à risque pendant la période de transition

Information destinée aux médecins généralistes, endocrinologues, pédiatres, chirurgiens ORL, gynécologues obstétriciens, cardiologues, gériatres, pharmaciens officinaux et hospitaliers.

Madame, Monsieur, Cher confrère,

En accord avec l'Agence nationale de sécurité du médicament et des produits de santé (ANSM), le laboratoire Merck souhaite porter à votre connaissance les informations suivantes.

Résumé

- Une nouvelle formule de Levothyrox® comprimés sécables est mise à disposition à partir de fin mars 2017
- Elle se caractérise par une amélioration de la stabilité en substance active durant toute la durée de conservation du produit et par la suppression d'un excipient à effet notoire, le lactose.
- **Les modalités de prise et de suivi sont inchangées hormis pour les patients à risque pour qui un suivi spécifique et un contrôle de l'équilibre thérapeutique est recommandé.**
- Il est rappelé que le Levothyrox® est un produit à marge thérapeutique étroite.

Pour les médecins prescripteurs :

- Pour les patients à risque : confirmer le maintien de l'équilibre thérapeutique par une évaluation clinique et biologique.

Pour les pharmaciens :

- Les codes CIP et UCD sont modifiés
- La présentation des boites et les couleurs sont modifiées selon les dosages
- **La mise à disposition des nouvelles boites se fera au fur et à mesure de l'écoulement des stocks** des anciennes boites, dosage par dosage. A réception des nouvelles boites, les pharmacies sont invitées à les mettre à disposition auprès des patients, uniquement après écoulement des stocks des anciennes boites.
- Il est recommandé de limiter la coexistence des anciennes et nouvelles boites
- **Il est nécessaire d'informer les patients** du changement de couleur des boites et des blisters de la plupart des dosages et de l'importance **de terminer leur stock de l'ancienne formule AVANT de passer à la nouvelle formule, pour ne plus changer ensuite.**

Informations complémentaires de sécurité et recommandations

Levothyrox® est prescrit dans le traitement des hypothyroïdies (insuffisance de sécrétion de la glande thyroïde) et des circonstances associées ou non à une hypothyroïdie où il est nécessaire de freiner la sécrétion de TSH (hormone stimulant la glande thyroïde).

La lévothyroxine sodique est une hormone thyroïdienne de synthèse à marge thérapeutique étroite. Lors de la phase de transition, il est recommandé de surveiller l'équilibre thérapeutique chez certains patients à risque dans les catégories suivantes: les patients qui reçoivent un traitement pour le cancer de la thyroïde mais qui présentent également une maladie cardiovasculaire (insuffisance cardiaque ou coronarienne et/ou des troubles du rythme), les femmes enceintes, les enfants et

Contact expéditeur : informations@securite-patients.info

les personnes âgées ; et dans certaines situations pour lesquelles l'équilibre thérapeutique a été particulièrement difficile à atteindre.

- Chez ces patients, le maintien de l'équilibre thérapeutique doit être confirmé **par une évaluation clinique et biologique** (contrôle de la TSH réalisé entre 6 et 8 semaines après la transition sauf chez les femmes enceintes où un dosage toutes les 4 semaines est recommandé).
- Le dosage de TSH permet à lui seul de confirmer le maintien de l'euthyroïdie et s'inscrit dans la surveillance habituelle de l'hormonothérapie substitutive conformément au RCP.
- Un dosage de la T4l reste justifié dans certaines conditions particulières notamment en cas d'insuffisance antéhypophysaire.

Les caractéristiques et les codes CIP et UCD des présentations sont modifiés en conséquence, comme indiqué dans le tableau ci-dessous :

	CIP	UCD	Ancienne couleur boite et blister *	Nouvelle couleur boite et blister *
Levothyrox 25 µg	3400930065556 (30 cps) 3400930065570 (90 cps)	3400894232513	Vert foncé	Vert foncé
Levothyrox 50 µg	3400930065662 (30 cps) 3400930065686 (90 cps)	3400894232681	Orange	Gris
Levothyrox 75 µg	3400930065785 (30 cps)	3400894232742	Violet	Violet
Levothyrox 100 µg	34009300 65891 (30 cps)	3400894232162	Rose	Bleu
Levothyrox 125 µg	3400930066010 (30 cps)	3400894233343	Jaune	Bleu clair
Levothyrox 150 µg	3400930066188 (30 cps)	3400894232223	Bleu foncé	Rouge
Levothyrox 175 µg	3400930066249 (30 cps)	3400894232391	Vert-bleu	Orange
Levothyrox 200 µg	3400930023341 (30 cps)	3400894232452	Rouge	Rouge foncé

*Voir Q&A sur le site de l'ANSM

Déclaration des effets indésirables

L'ANSM rappelle que les professionnels de santé doivent déclarer immédiatement tout effet indésirable suspecté d'être dû à un médicament dont ils ont connaissance au centre régional de pharmacovigilance dont ils dépendent géographiquement.
Par ailleurs, tout signalement de risque d'erreur médicamenteuse, d'erreur potentielle ou d'erreur avérée sans effet indésirable, inhérent aux médicaments peut être transmis directement au Guichet Erreurs Médicamenteuses.
Pour plus d'information, consulter la rubrique « Déclarer un effet indésirable » sur le site Internet de l'ANSM : http://ansm.sante.fr

Information médicale

Pour toute information, vous pouvez prendre contact avec le laboratoire Merck Santé s.a.s - Information médicale et Pharmacovigilance, au numéro suivant : 0800 888 024 (Service & Appels gratuits)

Nous vous remercions de prendre en compte cette information.
Valérie LETO-ESPIRAT
Pharmacien Responsable Merck Serono s.a.s.

Les informations complémentaires sont accessibles sur le site de l'ANSM à l'aide du lien suivant : http://ansm.sante.fr

Contact expéditeur : informations@securite-patients.info

Annexe 2 : Lettre d'information de Merck Santé destinée aux professionnels de santé

```
import re
import sys
import csv
import threading
import timeit
import time
import pandas as pd
import urllib.request
import unicodedata
import re
import pandas as pd
from google.colab import files

class GetAllPages_topic_Thread(threading.Thread):

    def __init__(self, turl, page, nb_page, sujet):
        threading.Thread.__init__(self)
        self.url = turl
        self.page = page
        self.nb_page = nb_page
        self.sujet = sujet

    def run(self):
        global all_msg
        if debug:
          print("[Ouverture URL de \"" + self.sujet + "\" page " + str(page) + "] :
" + str(self.url))

        html = get_html(url, 5)
        if html == -1:
          print(f"[Error get messages] --> abort page {url}")
          return

        only_messages = re.search('<div id="topic" >(.*?)<div
class="bottom_action_topic_menu">', html, re.MULTILINE | re.DOTALL).group(1)
        messages_page = re.findall('class="md-topic_post(.*?)/table>',
only_messages, re.MULTILINE | re.DOTALL)

        for message in messages_page:
          if re.match('.*data-id_user.*', message, re.DOTALL):
            user = re.search('data-id_user.+?>(.+?)<', message).group(1)
          elif re.match('.*itemprop="name"', message, re.DOTALL):
              user = re.search('itemprop="name".*?>(.+?)<', message).group(1)
          elif re.match('.*Profil supprimé.*', message, re.DOTALL):
              user = "Profil supprimé"
          else:
              user = "[ERROR_Encodage_user_unknown]"
          date = re.search('Posté le ([0-9/]+)', message).group(1)
          if re.match('.*itemprop="citation".*', message, re.DOTALL):
```

```
            message = re.sub('itemprop="citation".+?</span></span>', '', message,
flags=re.DOTALL)
          text = re.search('itemprop="text" hidden>(.*?)</span><div>', message,
re.MULTILINE | re.DOTALL).group(1)
          text = clean_message(text)
          all_msg = all_msg.append({'date':date, 'user':user, 'text':text,
'url':self.url}, ignore_index=True)

        if debug:
          print("[" + str(len(messages_page)) + " new msg sur \"" + self.sujet +
"\" de la page " + str(self.page) + " sur " + self.nb_page + "]")


def clean_message(msg):
    msg = re.sub('&#039;', '\'', msg)#apostrop"he
    msg = re.sub(',', ' ', msg)  # Pour un decoupage correct sur excel
    msg = re.sub('[>\r\n]+', ' ', msg) #Saut de ligne
    msg = re.sub(':\w+:', ' ', msg) #les smiley :happy:
    msg = re.sub('http\://.+?\.html', '', msg) #les liens copi
    msg = re.sub('<img.*?/>', ' ', msg)  #suppr les images
    msg = re.sub('<br.*?>', ' ', msg) #suppr les balise br
    msg = re.sub('<a (.*?)</a>', ' ', msg) #suppr les liens externe
    msg = re.sub('</?span.*?>', ' ', msg)
    msg = re.sub('</?table.*?>', ' ', msg)
    msg = re.sub('</?[a-z][a-z]?>', ' ', msg) #</i> <lu> et bien d'autre
    msg = re.sub('&[a-z#0-9]{1,4};', ' ', msg) #&#034;   &euro; &gt; &lt;
    msg = re.sub('\[#[0-9]+ size=[0-9]+\]', ' ', msg)
    msg = re.sub('</?strong>', ' ', msg)
    msg = re.sub('</?div>?', ' ', msg)
    msg = unicodedata.normalize('NFD', msg).encode('ascii', 'ignore')  # suppr les
accents
    return(msg)

    def get_nbr_page(html):
    list_pages = re.search('pagination_main_visible(.+?)/div', html).group(1)
    if re.match(r".*href.*", list_pages):
        return(re.findall("\">([0-9]+)<", html)[-1])
    else:
        return("1")

def get_html(url:str, max_attempt:int):
  attempt = 1
  while (attempt <= max_attempt):
    try:
      with urllib.request.urlopen(url) as response:
        html = response.read().decode('utf-8')
        return html
    except OSError as e:
      print(f"[Error {e.code}] {e.reason} : {url}")
      if e.code == 503:
```

```
        time.sleep(60)
      time.sleep(1)
      attempt += 1
  return -1

# Executiuon start here
search = "levothyrox"
rubrique = "18*sante"
debug = True
save_tmp_out = True

print("Recherche de <"+search+"> dans la rubrique <"+rubrique+">")
if debug:

print('http://forum.doctissimo.fr/search_result.php?post_cat_list='+rubrique+'&sear
ch='+search+'&resSearch=250')
with
urllib.request.urlopen('http://forum.doctissimo.fr/search_result.php?post_cat_list=
'+rubrique+'&search='+search+'&resSearch=250') as response:
  html = response.read().decode('utf-8')
if re.match(r".*aucune réponse n'a été trouvée.*", html, re.MULTILINE|re.DOTALL):
    print("La recherche de <"+search+"> dans la rubrique <"+rubrique+"> donne aucun
résultat")
    sys.exit()
nb_page_topic = get_nbr_page(html)

print(nb_page_topic+" page(s) de 250 topics sur le sujet <"+search+"> dans la
rubrique <"+rubrique+">")

all_topics_url = []
page = 1
if debug:
    nb_page_topic = "1"
while page <= int(nb_page_topic):
    print("telechargement de page " + str(page) + " sur " + nb_page_topic)
    if debug:

print('http://forum.doctissimo.fr/search_result.php?post_cat_list='+rubrique+'&sear
ch='+search+'&resSearch=250&page='+str(page))
    with
urllib.request.urlopen('http://forum.doctissimo.fr/search_result.php?post_cat_list=
'+rubrique+'&search='+search+'&resSearch=250&page='+str(page)) as response:
        html = response.read().decode('utf-8')
    topics = re.findall(r"</?t.*?sujet ligne_booleen(.+?)</tr>", html, re.MULTILINE
| re.DOTALL)
    for topic in topics:
        if debug:
            print(re.search(r"href=\"(.+?)\"", topic).group(1))
        all_topics_url.append(re.search(r"href=\"(.+?)\"", topic).group(1))
    page += 1
```

```
print("nb total de topic = " + str(len(all_topics_url)))

start = timeit.default_timer()
all_msg = pd.DataFrame(columns=['date','user', 'text', 'url'])
threadList = []
nb_topic = 0

for url in all_topics_url:
    if debug:
        print(url)
        time.sleep(2)
    html = get_html(url, 5)

    if html == -1:
      print(f"[Error get nb page] --> abort topic {url}")
      continue

    sujet_topic = re.search("forum.doctissimo.fr/sante/.+/(.*?)sujet_",
url).group(1)
    nb_page_topic = get_nbr_page(html)

    if debug:
        print("topic \""+sujet_topic+"\" avec "+str(nb_page_topic)+" page(s)")
    page = 1
    while page <= int(nb_page_topic):
        clean_url = re.search(r"(.*)_", url).group(1)
        newthread = GetAllPages_topic_Thread(clean_url + "_" + str(page) + ".htm",
page, nb_page_topic, sujet_topic)
        newthread.start()
        time.sleep(0.1)
        threadList.append(newthread)
        page += 1

    if (nb_topic % 100 == 0):
      print(str(nb_topic) + " topics extrait sur " + str(len(all_topics_url)) + ".
Messages récoltés : " + str(len(all_msg)))

    if len(all_msg) > 1000 and save_tmp_out == True:
        all_msg.to_csv("tmp_out.csv", sep=',', encoding='utf-8', index=False)
        print("Fichier temporaire save --> tmp_out.csv")
        save_tmp_out = False
    nb_topic += 1

print("Attente des threads")
for curThread in threadList :
    curThread.join()
all_msg.to_csv("out.csv", sep=',', encoding='utf-8', index=False)
stop = timeit.default_timer()
m, s = divmod(stop - start, 60)
h, m = divmod(m, 60)
```

```
print(str(len(all_msg)) + " messages total récoltés en %dh %02dmin et %02ds" % (h,
m, s))

url = "http://forum.doctissimo.fr/sante/arthrose-os/maigrir-sujet_149370_1.htm" #
profil supprimé + citation normol
url = "http://forum.doctissimo.fr/sante/thyroide-problemes-
endocrinologiques/endocrinologue-belgique-sujet_160644_1.htm" # encodage user avec
space authorisé (Susanne in F)
url = "http://forum.doctissimo.fr/sante/thyroide-problemes-
endocrinologiques/supportez-thyroxine-sanofi-sujet_171008_1.htm" #avec hidden dans
code devant user name
url = "http://forum.doctissimo.fr/sante/thyroide-problemes-
endocrinologiques/demande-renseignements-sujet_171692_1.htm" #encodage citaiton
different
url = "http://forum.doctissimo.fr/sante/regles-problemes-gynecologiques/retart-
regles-sujet_222033_1.htm" # encodage citation différent
url = "http://forum.doctissimo.fr/sante/thyroide-problemes-
endocrinologiques/probleme-couple-tyroide-sujet_152716_2.htm" #message de fay41ft
le 24/04/2006 "..." cité absent sur la page web mais présent dans le tableau ??
url = "https://forum.doctissimo.fr/sante/thyroide-problemes-
endocrinologiques/interruption-pituitaire-hypophysaire-sujet_156698_1.htm"

df = pd.DataFrame(columns=['date','user', 'text', 'url'])
try:
    with urllib.request.urlopen(url) as rep:
        html = rep.read().decode('utf-8')
except (http.client.IncompleteRead) as e:
    html = e.partial.decode('utf-8')

only_messages = re.search('<div id="topic" >(.*?)<div
class="bottom_action_topic_menu">', html, re.MULTILINE | re.DOTALL).group(1)
messages_page = re.findall('class="md-topic_post(.*?)/table>', only_messages,
re.MULTILINE | re.DOTALL)
for message in messages_page:
    if re.match('.*data-id_user.*', message, re.DOTALL):
        user = re.search('data-id_user.+?>(.+?)<', message).group(1)
    elif re.match('.*itemprop="name"', message, re.DOTALL):
        user = re.search('itemprop="name".*?>(.+?)<', message).group(1) #parfois
hidden est rajouté dans le code source donc .+? après name
    elif re.match('.*Profil supprimé.*', message, re.DOTALL):
        user = "Profil supprimé"
    else:
        user = "[ERROR_Encodage_user_unknown]"
    date = re.search('Posté le ([0-9/]+)', message).group(1)
    if re.match('.*itemprop="citation".*', message, re.DOTALL):
        message = re.sub('itemprop\=\"citation\".+?</span><span', '', message,
flags=re.DOTALL)
    text = re.search('itemprop="text" hidden>(.*?)</span>[<div>|<span
itemprop="author"]', message, re.MULTILINE | re.DOTALL).group(1)
    text = clean_message(text)
```

```
    df = df.append({'date':date, 'user':user, 'text':text, 'url':url},
ignore_index=True)
print(str(len(messages_page)))
```

Annexe 3 : Code source python du script de scraping (récupération automatisée de la donnée)

```
# -*- coding: utf-8 -*-
import time
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import re
import string
import csv
import nltk
import ssl
import sys
import Spacy
import fr_core_news_sm
from nltk.corpus import stopwords
try:
    _create_unverified_https_context = ssl._create_unverified_context
except AttributeError:
    pass
else:
    ssl._create_default_https_context = _create_unverified_https_context
nltk.download('stopwords')
stop = stopwords.words('french')
nlp_fr = fr_core_news_sm.load()

# Functions & Dictionaries
def doctissimo_sort_range(df):
        df['date'] = pd.to_datetime(df['date'], format='%d/%m/%Y')
        df = df.sort_values(by=['date'], ascending=False)
        df['year'] = df['date'].dt.year
        df = df.set_index(df['date'])
        df = df[:'2015-12-31']
        df['index'] = range(0, len(df))
        df = df.set_index(df['index'])
        del df['index']
        df['text'] = df['text'].astype(str)
        return df

def doctissimo_words_improvment(df):
        df['text'] = df['text'].str.split() # Split string
        i=0
        for line in df['text']:
                new_line = []
                for word in line:
```

```
                        for imprv in words_improvment:
                                if word in imprv and word != imprv[0]:
                                        word = imprv[0]
                                        break
                        new_line.append(word)
                df.at[i, 'text'] = new_line
                i += 1
        df['text'] = df['text'].apply(' '.join) # Join string
        return df

def dataframe_preprocessing(df):
        df['text'] =
df['text'].str.normalize('NFKD').str.encode('ascii',errors='ignore').str.decode('ut
f-8') # Remove accent
        df['text'] = [re.sub(r'[^a-zA-Z0-9 ]', ' ', str(x)) for x in df['text']] #
Remove special characters and punctuation
        df['text'] = df['text'].str.lower() # Convert ['text'] string to lowercase
        df['text'] = df['text'].str.replace('#034', '', regex=True) # Remove #034
pattern
        df['text'] = df['text'].str.replace('#039', '', regex=True) # Remove #039
pattern
        df['text'] = df['text'].str.replace(".*gif", "", regex=True) # Remove all
gifs
        df['text'] = df['text'].str.replace("http.* ", "", regex=True) # Remove
http links
        df['text'] = df['text'].str.replace("https.* ", "", regex=True) # Remove
https links
        df['text'] = [re.sub(r'[A-Za-z]+\d+|\d+[A-Za-z]+','', str(x)) for x in
df['text']] # Delete numbers between alphabetic chars
        df['text'] = [re.sub(r'\b(?!(\D\S*|[12][0-9]{3})\b)\S+\b','', str(x)) for x
in df['text']] # Numbers except dates
        df['text'] = df['text'].str.replace('\n',' ', regex=True).replace('\t',' ',
regex=True) # Remove line breaks and tabulations
        df['text'] = [re.sub(r'(^| ).( |$)', ' ', str(x)) for x in df['text']] #
Remove single characters
        df['text'] = [re.sub(r'\s+', ' ', str(x)) for x in df['text']] # Delete
multiple spaces
        return df

def dataframe_stopwords_wtd(df):
        # Words to delete (csv file)
        words_to_delete = []
        words_count = 0
        with open(dir + 'exclusions.csv', newline='') as csvfile:
            reader = csv.reader(csvfile, delimiter=' ', quotechar='|')
            for row in reader:
                words_to_delete.append('; '.join(row))
        words_count = len(words_to_delete)
        words = df['text'].str.split()
```

```
        words = words.apply(lambda x: [item for item in x if item not in
words_to_delete]) # Words in "text" of dataframe) without excluded words
        words = words.apply(lambda x: [item for item in x if item not in stop]) #
From library "FR"
        words = words.apply(lambda x: [item for item in x if item not in
additional_stopwords]) # From dictionary
        df['text'] = words.apply(' '.join)
        return df

def dataframe_lemmatization(df):
        # Lemmatization with nlp_fr
        df['text'] = df['text'].apply(lambda x: [y.lemma_ for y in
nlp_fr(x)]).apply(' '.join)
        return df

def dataframe_duplicata_less3words(df):
        words_count = df['text'].str.count(' ') + 1 # 'text' characters counter
        df['words_count'] = words_count # Add words_count column on the dataframe
        before_deleting = df['text'].count()
        print('\n******************\nNumber of rows BEFORE deleting the rows which
contains less than 3 words : ' + str(before_deleting) + ' rows')
        df.drop(df[df['words_count'] < 3].index, inplace = True) # Remove rows
which contains less than 3 words
        after_deleting = df['text'].count()
        print('Number of rows AFTER deleting the rows which contains less than 3
words : ' + str(after_deleting) + ' rows')
        diff_deleting = before_deleting - after_deleting
        print('Difference : ' + str(diff_deleting) + ' rows')
        df.drop_duplicates() # Remove duplicates rows
        df.dropna() # Drop the rows even with single NaN or single missing values.
        after_del_duplicates = df['text'].count()
        print('Number of rows after deleting duplicata : ' +
str(after_del_duplicates) + ' rows')
        diff_duplicates = after_deleting - after_del_duplicates
        total_delete = diff_deleting + diff_duplicates
        print('Difference : ' + str(diff_duplicates) + ' rows')
        print('Total number of deleted rows : ' + str(total_delete) +
'\n******************')
        return df

def lemmatizer(text):
    sent = []
    doc = nlp_fr(text)
    for word in doc:
        sent.append(word.lemma_)
    return " ".join(sent)

# StopWords dictionary
additional_stopwords = ['a', 'abord', 'afin', 'ah', 'ai', 'ainsi', 'allaient',
'allo', 'allô', 'allons', 'alors', 'apres', 'après', 'assez', 'attendu', 'aucun',
```

```
'aucune', 'aucuns', 'aujourd', 'aujourdhui', 'auquel', 'auquelle', 'auquelles',
'auquels', 'aussi', 'autre', 'autres', 'auxquelles', 'auxquels', 'avant', 'avoir',
                        'b', 'bonjour', 'bonsoir', 'bah', 'beaucoup', 'bien',
'bigre', 'bon', 'boum', 'br', 'bravo', 'brr', 'brrr',
                        'ca', 'ça', 'car', 'ceci', 'cela', 'celle', 'celle-ci',
'celle-la', 'celle-là', 'celles', 'celles-ci', 'celles-la', 'celles-là', 'celui',
'celui-ci', 'celui-la', 'celui-là', 'cent', 'cependant', 'certain', 'certaine',
'certaines', 'certains', 'certes', 'cet', 'cette', 'ceux', 'ceux', 'ceux-ci',
'ceux-là', 'ceux-là', 'chacun', 'chaque', 'cher', 'chere', 'chère', 'cheres',
'chères', 'chers', 'chez', 'chiche', 'chut', 'ci', 'cinq', 'cinquantaine',
'cinquante', 'cinquantieme', 'cinquantième', 'cinquieme', 'cinquième', 'clac',
'clic', 'combien', 'comme', 'comment', 'compris', 'concernant', 'contre', 'couic',
'crac',
                        'da', 'debout', 'debut', 'début', 'dedans', 'dehors',
'dela', 'delà', 'depuis', 'derriere', 'derrière', 'dés', 'dès', 'desormais',
'désormais', 'desquelles', 'desquels','dessous', 'dessus', 'deux', 'deuxieme',
'deuxième', 'deuxiemement', 'deuxièmement', 'devant', 'devers', 'devra', 'devrait',
'different', 'différent', 'differente', 'différente', 'differentes', 'différentes',
'differents', 'différents', 'dire', 'divers', 'diverse', 'diverses', 'dix', 'dix-
huit', 'dix-neuf', 'dix-sept', 'dixieme', 'dixième', 'doit', 'doivent', 'donc',
'dont', 'douze', 'douzieme', 'douzième', 'dring', 'droite', 'duquel', 'durant',
                        'e', 'effet', 'eh', 'elle-meme', 'elle-même', 'elles',
'elles-memes', 'elles-mêmes', 'encore', 'entre', 'envers', 'environ', 'ès',
'essai', 'etaient', 'etais', 'etait', 'etant', 'etante', 'etantes', 'etants',
'etat', 'état', 'etats', 'états', 'etc', 'ete', 'etee', 'etees', 'etes', 'etiez',
'etions', 'étions', 'etre', 'être', 'euh', 'eumes', 'eux-memes', 'eux-mêmes',
'excepte', 'excepté',
                        'f', 'facon', 'façon', 'fais', 'faisaient', 'faisant',
'fait', 'faites', 'feront', 'fi', 'flac', 'floc', 'fois', 'font', 'force', 'fumes',
'futes',
                        'g', 'gens',
                        'h', 'ha', 'haut', 'he', 'hé', 'hein', 'helas', 'hélas',
'hem', 'hep', 'hi', 'ho', 'hola', 'holà', 'hop' ,'hormis', 'hors' ,'hou', 'houp',
'hue', 'hui', 'huit', 'huitieme', 'huitième', 'hum', 'hurrah',
                        'i', 'ici', 'importe',
                        'jusqu', 'jusqua', 'jusque', 'juste',
                        'k',
                        'là', 'laquelle', 'las', 'lequel', 'lès', 'lesquelles',
'lesquels', 'leurs', 'longtemps', 'lorsque', 'lui-meme', 'lui-même',
                        'maint', 'maintenant', 'malgre', 'malgré', 'meme', 'memes',
'mêmes' ,'merci', 'mien', 'mienne', 'miennes', 'miens', 'mille', 'mince', 'mine',
'moi-meme', 'moi-même', 'moins', 'mot', 'moyennant',
                        'na', 'nai', 'nas', 'neanmoins', 'néanmoins', 'neuf',
'neuvieme', 'neuvième', 'ni', 'nombreuses', 'nombreux', 'nommes', 'nommés', 'non',
'nôtre', 'notres', 'nôtres', 'nous-meme', 'nous-memes', 'nous-memes', 'nous-mêmes',
'nouveau', 'nouveaux', 'nul',
                        'o', 'onsoir', 'onjour', 'ô', 'oh', 'ohe', 'ohé', 'ole',
'olé', 'olle', 'ollé', 'onze', 'onzieme', 'onzième', 'ore', 'où', 'ouf', 'ouias',
'oust', 'ouste', 'outre',
```

```
                      'p', 'paf', 'pan', 'parce', 'parmi', 'parmis', 'parole',
'partant', 'particulier', 'particuliere', 'particulière', 'particulierement',
'particulièrement', 'passe', 'passé', 'pendant', 'personne', 'personnes', 'peu',
'peut', 'peuvent', 'peux', 'pff', 'pfff', 'pffff', 'pfft', 'pfut', 'piece',
'pièce', 'pif', 'plein', 'pleins', 'plouf', 'plupart', 'plus', 'plusieurs',
'plutot', 'plutôt', 'pouah', 'pourquoi', 'premier', 'premiere', 'première',
'premierement', 'premièrement', 'pres','près', 'proche', 'psitt', 'puisque',
                      'q', 'quand', 'quant', 'quant-a-soi', 'quant-à-soi',
'quant-a-soit', 'quanta', 'quarante', 'quatorze', 'quatre', 'quatre-vingt',
'quatrieme', 'quatrième', 'quatriemement', 'quatrièmement', 'quel', 'quelconque',
'quell', 'quelle', 'quelle', 'quelles', 'quelles', 'quelque', 'quelques',
'quelquun', 'quels', 'quest', 'quiconque', 'quil', 'quils', 'quinze', 'quoi',
'quoique',
                      'r', 'revoici', 'revoila', 'revoilà', 'rien',
                      'sacrebleu', 'sans', 'sapristi', 'sauf', 'seize', 'selon',
'sept', 'septieme', 'seulement', 'si', 'sien', 'sienne', 'siennes', 'siens',
'sinon', 'six', 'sixieme', 'sixième', 'soi', 'soi-meme', 'soi-même', 'soient',
'sois', 'soixante', 'sous', 'suivant', 'sujet', 'surtout',
                      'tac', 'tandis', 'tant', 'té', 'tel', 'telle', 'tellement',
'telles', 'tels', 'tenant', 'tic', 'tien', 'tienne', 'tiennes', 'tiens', 'toc',
'toi-meme', 'toi-même', 'touchant', 'toujours', 'tous', 'tout', 'toute', 'toutes',
'treize', 'trente', 'tres', 'très', 'trois', 'troisieme', 'troisième',
'troisiemement', 'troisièmement', 'trop', 'tsoin', 'tsouin',
                      'u', 'unes', 'uns',
                      'v', 'va', 'vais', 'valeur', 'valeurs', 'vas', 've', 'vé',
'vers', 'via', 'vif', 'vifs', 'vingt', 'vivat', 'vive', 'vives', 'vlan', 'voici',
'voie', 'voient', 'voila', 'voilà', 'vont', 'vôtre', 'votres', 'vôtres', 'vous-
memes', 'vous-mêmes', 'vu',
                      'w',
                      'x',
                      'z', 'zut']

# Words_improvment dictionary
words_improvment = [['levothyrox', 'levo', 'levothyro', 'levotyrox'],
                    ['euthyrox', 'leuthyrox', 'eutyrox', 'leutyrox'],
                    ['lthyroxine', 'lthyroxin', 'ltyroxine', 'ltyroxin'],
                    ['hypothyroidie', 'lhypothyroidie', 'hypotyroidie',
'lhypotyroidie'],
                    ['comprime', 'comprim'],
                    ['cytomel', 'cynomel'],
                    ['controle', 'control'],
                    ['changer', 'change', 'chang', 'changement'],
                    ['allemagne', 'allemand'],
                    ['generaliste', 'generalist'],
                    ['arret', 'arreter', 'arrete'],
                    ['excipient', 'excipients'],
                    ['laboratoire', 'laboratoir'],
                    ['poids', 'poid'],
                    ['hormone', 'hormon', 'dhormone', 'dhormon', 'lhormone',
'lhormon'],
```

```
                    ['correcte', 'correct'],
                    ['courche', 'coucher', 'chouchee'],
                    ['neomercazole', 'neomercazol'],
                    ['enceinte', 'enceint'],
                    ['ancien', 'ancienne', 'lancien', 'lancienne'],
                    ['francais', 'fraincaise', 'francai'],
                    ['manque', 'manqu'],
                    ['medicament', 'medicaments', 'medoc', 'medocs'],
                    ['stress', 'stres'],
                    ['analyse', 'danalyse'],
                    ['hasimoto', 'dhashimoto', 'hasimoto'],
                    ['thyroidite', 'thyroidit', 'tyroidite', 'tyroidit'],
                    ['angoisse', 'dangoisse'],
                    ['hypo', 'lhypo', 'dhypo'],
                    ['hyper', 'lhyper', 'dhyper'],
                    ['euthyral', 'leuthyral', 'deuthyral', 'eutyral', 'leutyral',
'deutyral'],
                    ['autre', 'lautre'],
                    ['pmol', 'pmoil'],
                    ['soucis', 'souci'],
                    ['augmente', 'augment', 'augmenter', 'daugmenter', 'daugmente',
'daugment'],
                    ['sentais', 'sentai'],
                    ['pensezvous', 'pensezvou'],
                    ['echographie','echo', 'lecho', 'lechographie'],
                    ['aller', 'alle', 'allee'],
                    ['adenomegalie', 'dadenomegalie'],
                    ['jaurai', 'jaurais'],
                    ['elever', 'eleve', 'elevee'],
                    ['cheveux', 'cheveu'],
                    ['devrai', 'devrais'],
                    ['menopause', 'menopaus'],
                    ['nodule', 'nodul'],
                    ['reactive', 'reactiv'],
                    ['periode', 'period'],
                    ['epuiser', 'epuisee', 'epuise'],
                    ['arriver', 'arrive', 'narrive', 'narriv'],
                    ['memoire', 'memoir'],
                    ['parcours', 'parcour'],
                    ['message', 'messages'],
                    ['specialiste', 'specialist'],
                    ['apriori', 'priori'],
                    ['delais', 'delai'],
                    ['gonfler', 'gonfle', 'gonflee'],
                    ['ovaire', 'lorvair', 'lovair'],
                    ['precis', 'preci'],
                    ['prend', 'prends'],
                    ['fatiguer', 'fatigue', 'fatiguee'],
                    ['deprimer', 'deprime', 'deprimee', 'deprim'],
                    ['penser', 'pense', 'pensee', 'pensez'],
```

```
                    ['secretaire', 'secretair'],
                    ['quelque', 'quelques', 'quelqu'],
                    ['cause', 'caus'],
                    ['lobe', 'lob'],
                    ['t3', 't3l', 'ft3'],
                    ['t4', 't4l', 'ft4'],
                    ['norme', 'norm'],
                    ['doser', 'dosee', 'dose'],
                    ['endocrinologue', 'lendocrinologue', 'endocrino',
'lendocrino'],
                    ['continu', 'continue'],
                    ['interval', 'intervalle'],
                    ['thyroidien', 'tyroidien'],
                    ['soir', 'soiree'],
                    ['conseil', 'conseille'],
                    ['anticorps', 'anticorp'],
                    ['ablation', 'lablation'],
                    ['aider', 'aide', 'maide', 'maider'],
                    ['ordre', 'ordr', 'lordre', 'lordr'],
                    ['operation', 'loperation'],
                    ['remercier', 'remercie'],
                    ['marseille', 'marseill']]

# Execution start here
dir = "FOLDER PATH OF THESE SCRIPTS" # ie: /Users/jp/Documents-non-icloud/thèse-
levothyrox/algo/2_datasets_formating_cleaning.py
df_doctissimo = pd.read_csv(dir + 'dataset_doctissimo_22_03_2020.csv',
encoding='utf8')
print('\n*****************\nFile <dataset_doctissimo_22_03_2020.csv> has been
loaded')
df_french_tweets = pd.read_csv(dir + 'french_tweets.csv', encoding='utf8')
print('File <french_tweets.csv> has been loaded\n*****************\n')

# Doctissimo : start
start = time.time()
df = df_doctissimo.copy()
del df_doctissimo
# Doctissimo : sort by date and select range : 2016-2020
df = doctissimo_sort_range(df)
print('*****************\nSorting and selecting range : 2016-2020...
check\n*****************')
# Doctissimo dataframe preprocessing
df = dataframe_preprocessing(df)
print('\n*****************\nData cleaning and formating...
check\n*****************\n')
print(df.iloc[3,2])
# Doctissimo : words improvment
df = doctissimo_words_improvment(df)
print('\n*****************\nWords improvment... check\n*****************\n')
print(df.iloc[3,2])
```

```
# Doctissimo stop words removing
df = dataframe_stopwords_wtd(df)
print('\n*****************\nRemoving stop words and WTD...
check\n*****************\n')
print(df.iloc[3,2])
# Doctissimo lemmatization (COMMENT TO ENHANCE TIME OF EXECUTION)
df = dataframe_lemmatization(df)
print('\n*****************\nLemmatization... check\n*****************\n')
print(df.iloc[3,2])
# Doctissimo final cleaning step
df = dataframe_duplicata_less3words(df)
print('\n*****************\nRemoving duplicates and rows which contains less than 3
words... check\n*****************\n')
df.to_csv(dir + 'dataset_doctissimo_updated.csv', index=False, sep=',',
header=True, encoding='utf8')
print('\nFile <dataset_doctissimo_updated.csv> has been exported')
end = time.time()
print('Elapsed time - Doctissimo:', end - start, 's')
print()
print(df)

# French_tweets : start
start = time.time()
df = df_french_tweets.astype(str).copy()
del df_french_tweets
# French_tweets dataframe preprocessing
df = dataframe_preprocessing(df)
# French_tweets : labeling format
df['sentiment'] = df['label']
df['sentiment'] = df['sentiment'].str.replace("0", "negative", regex=True)
df['sentiment'] = df['sentiment'].str.replace("1", "positive", regex=True)
col = ['sentiment', 'text']
df = df[col]
df['sentiment']=['__label__'+ s for s in df['sentiment']]
# French_tweets stop words removing
df = dataframe_stopwords_wtd(df)
# French_tweets lemmatization (COMMENT TO ENHANCE TIME OF EXECUTION)
df = dataframe_lemmatization(df)
print('*****************\nAll formating steps... check\n*****************')
# French_tweets final cleaning step
df = dataframe_duplicata_less3words(df)
del df['words_count'] # Remove "words_count" column
df.to_csv(dir + 'french_tweets_updated.csv', index=False, sep=',', header=True,
encoding='utf8')
print('\nFile <french_tweets_updated.csv> has been exported')
end = time.time()
print('Elapsed time - French_tweets: ', end - start, 's')
print()
print(df)
```

Annexe 4 : Code source python du script de nettoyage de la donnée

```
# -*- coding: utf-8 -*-
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import csv
from common import word_occurence, n_gram, words_cloud

def word_occurence_history(df, period):
    df2 = None
    df3 =df.groupby(pd.Grouper(freq=period))
    for period, group in df3:
        df1 = word_occurence(group)
        df1.rename(columns={'occurence': period}, inplace=True)
        if df2 is None :
            df2 = df1
        else:
            df2 = pd.concat([df2, df1], axis=1, join='outer', sort=True)
    df2 = df2.fillna(0)
    return df2

def top_word_occurence_history(df, period, limit):
    df2 = None
    df = word_occurence_history(df, period)
    # Convert to percent of column total
    #df = df.div(df.sum(axis=0), axis=1).multiply(100)
    periods = list(df.columns)
    for per in periods :
        df1 = df.nlargest(limit, [per])                    # Period top rows
        if df2 is None :
            df2 = df1
        else:
            df2 = pd.concat([df1, df2], join='inner')       # Concatenate period
top 10
    df2 = df2.groupby(level=0).last()                      # Clean duplicate rows
    return df2

def n_gram_history(df, n_gram_size, period):
    # n-gram : yearly analysis
    df2 = None
    df3 = df.groupby(pd.Grouper(freq=period))
    for period, group in df3:
        df1 = n_gram(group, n_gram_size)
        df1.rename(columns={'occurence': period}, inplace=True)
        if df2 is None :
            df2 = df1
        else:
            df2 = pd.concat([df2, df1], axis=1, join='outer', sort=True)
    df2 = df2.fillna(0)
```

```
    return df2

def top_n_gram_history(df, n_gram_size, period, limit):
    # n-gram : yearly analysis
    df2 = None
    df = n_gram_history(df, n_gram_size, period)
    # To convert to percent of column total, uncomment next line
    #df = df.div(df.sum(axis=0), axis=1).multiply(100)
    for period in list(df.columns):
        df1 = df.nlargest(limit, [period])                    # Period top rows
        if df2 is None:
            df2 = df1
        else:
            df2 = pd.concat([df1, df2], join='inner')         # Concatenate
period top 10
    df2 = df2.groupby(level=0).last()                         # Clean duplicate
rows
    return df2

def top_n_gram_history_2(df, n_gram_size, period, limit):
    # n-gram : yearly analysis
    df = n_gram_history(df, n_gram_size, period)
    df2 = pd.DataFrame()
    periods = list(df.columns)
    for period in periods:
        top_index_list = list(df.nlargest(limit, [period]).index)
        df2[period] = top_index_list
    return df2

def getCorrelations(df):
    # df : historical dataframe
    df_cor_data = []
    df = df.transpose()

    columns_list = list(df.columns)
    for i in range(0, len(columns_list)):
        col_1 = columns_list[i]
        for j in range (i+1, len(columns_list)):
            col_2 = columns_list[j]
            df1 = pd.merge(df[col_1], df[col_2], left_index=True, right_index=True,
how='inner', suffixes=('_left', '_right'))
            cor = df.corr()[col_1][col_2]
            df_cor_data.append({'Item 1':col_1, 'Item 2':col_2, 'correlation':cor})
            #cor =
np.corrcoef(security_1.data.loc[start:end]['Log_Returns(AdjClose)'],
security_2.data.loc[start:end]['Log_Returns(AdjClose)'])[1, 0]
    df_cor = pd.DataFrame(df_cor_data)
    df_cor = df_cor.sort_values(by='correlation', ascending=False)
    return df_cor
```

```
def getTopCorrelations(df):
    df = getCorrelations(df)
    df = df.loc[(df['correlation']>0.95)]
    df = df.sort_values(by=['Item 1', 'Item 2'])
    return df

# Added 2021/09/21
def getTopCorrelations_2(df):
    # df : historical dataframe
    df = df.transpose()
    dict = {}
    for year in range(2016, 2020):
        df0 = df.loc[df.index.year==year]
        df_cor_data = []
        columns_list = list(df0.columns)
        for i in range(0, len(columns_list)):
            col_1 = columns_list[i]
            for j in range (i+1, len(columns_list)):
                col_2 = columns_list[j]
                df1 = pd.merge(df0[col_1], df0[col_2], left_index=True,
right_index=True, how='inner', suffixes=('_left', '_right'))
                cor = df0.corr()[col_1][col_2]
                df_cor_data.append({'Item 1':col_1, 'Item 2':col_2,
'correlation':cor})
                #cor =
np.corrcoef(security_1.data.loc[start:end]['Log_Returns(AdjClose)'],
security_2.data.loc[start:end]['Log_Returns(AdjClose)'])[1, 0]
        df_cor = pd.DataFrame(df_cor_data)
        df_cor = df_cor.sort_values(by='correlation', ascending=False)
        df_cor = df_cor.loc[(df_cor['correlation']>0.95)]
        df_cor = df_cor.sort_values(by=['Item 1', 'Item 2'])
        dict[year] = df_cor
    return dict

# Execution starts here
dir = "FOLDER PATH OF THESE SCRIPTS" # ie: /Users/jp/Documents-non-icloud/thèse-
levothyrox/algo/3_word_frequency_analysis.py
top_size = 15
n_gram_size = 2
limit = 10
period = 'M'        # 'Y', 'M', 'Q'
show = True
df = pd.read_csv(dir + 'data/df_doctissimo.csv')

# Cast date to datetime
df['date'] = pd.to_datetime(df['date'])

# Cast text column to string
df['text'] = df['text'].astype(str)
```

```
# Set date column as index
df = df.set_index('date')

# Words cloud after cleaning
words_cloud(df, True)

# Top word occurence
print()
print('Top word occurence')
print(word_occurence(df).nlargest(10, ['occurence']))
print()

# Top word occurence history
print()
print('Top_word_occurence_history')
#print(word_occurence_history(df, 'Y').nlargest(10, [2016, 2017, 2018, 2019,
2020]))
print(top_word_occurence_history(df, 'Y', 10))
print()

# Top n-gram occurence
print()
print('Top N-gram occurrence')
print(n_gram(df, n_gram_size).nlargest(10, ['occurence']))
print()

# Top_n_gram_history
print()
print('Top_n_gram_history')
#print(n_gram_history(df, n_gram_size, 'Y').nlargest(10, [2016, 2017, 2018, 2019,
2020]))
print(top_n_gram_history(df, n_gram_size, 'Y', 10))
print()

# Top_n_gram_history_2
print()
print('Top_n_gram_history_2')
print(top_n_gram_history_2(df, n_gram_size, 'Y', 10))
print()

print()
print('Top word occurence correlations')
print(getCorrelations(top_word_occurence_history(df, 'Y', 10)))
print()

print()
print('Top n_gram occurence correlations')
print(getCorrelations(top_n_gram_history(df, n_gram_size, 'Y', 10)))
print()
```

```
print()
print('Top n_gram occurence correlations')
print(getTopCorrelations(top_n_gram_history(df, n_gram_size, 'Y', 10)))
print()

#df1 = top_word_occurence_history(df, 'Y', 10).transpose()
#df1 = top_word_occurence_history(df, 'M', 5).transpose()
df1 = top_n_gram_history(df, 2, 'Q', 5).transpose()
#df1 = top_n_gram_history(df, 2, 'M', 10).transpose()

df1.plot()
plt.show()

dict = getTopCorrelations_2(top_word_occurence_history(df, 'M', 10))
for year, df in dict.items():
    print()
    print(year)
    print(df)
```

Annexe 5 : Code source python du script d'analyse de la fréquence des mots

```
# -*- coding: utf-8 -*-
from fasttext import train_supervised
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn
from gensim.models.fasttext import FastText
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split
import string
import nltk
from nltk.corpus import stopwords
from io import StringIO
import csv
import time

# Execution start here
dir = "FOLDER PATH OF THESE SCRIPTS" # ie: /Users/jp/Documents-non-icloud/thèse-
levothyrox/algo/data/
vector_size = 60
window = 40
min_count = 3
sample = 1e-2
test_size = 0.2                              # test/(test+train)
# Limiting datasets size during test phase
max_tweets = None                           # Or None / 10000
max_texts = None                              # Or None / 100
```

```
df = pd.read_csv(dir + 'dataset_doctissimo_updated.csv')
# Limit size for initial testing
if not max_texts is None:
    df = df.sample(n=max_texts)

# Fasttext (Word2Vec) machine learning model
word_tokenized_corpus = df['text'].str.split()
start = time.time()
ft_model = FastText(word_tokenized_corpus,
                    vector_size=vector_size,
                    window=window,
                    min_count=min_count,
                    sample=sample,
                    sg=1,
                    epochs=10)

end = time.time()
print()
print('Elapsed time', end - start, 's')
print()
print('Keyed vectors for levotyrox')
print(ft_model.wv['levothyrox'])
semantically_similar_words = {words: [item[0] for item in
ft_model.wv.most_similar([words], topn=5)] for words in ['levothyrox', 'formul',
'secondair', 'sang', 't4', 'hormon']}
print()
print("Semantically similar words to ['levothyrox','formul', 'secondair', 'sang',
't4', 'hormon']" )
for k, v in semantically_similar_words.items():
    print(k + ':' + str(v))
print()
print('Similarity levothyrox / secondair', ft_model.wv.similarity(w1='levothyrox',
w2='secondair'))
print()
print('Similarity levothyrox / formul', ft_model.wv.similarity(w1='levothyrox',
w2='formul'))
all_similar_words = sum([[k] + v for k, v in semantically_similar_words.items()],
[])
print()
print('Similar words')
print(all_similar_words)

word_vectors = ft_model.wv[all_similar_words]

pca = PCA(n_components=2)

p_comps = pca.fit_transform(word_vectors)
word_names = all_similar_words

plt.figure(figsize=(20, 10))
```

```
plt.scatter(p_comps[:, 0], p_comps[:, 1], c='red')

for word_names, x, y in zip(word_names, p_comps[:, 0], p_comps[:, 1]):
    plt.annotate(word_names, xy=(x+0.06, y+0.03), xytext=(0, 0), textcoords='offset
points')

plt.show()

# French_tweets
french_tweets = pd.read_csv(dir + 'french_tweets_updated.csv')

# Limit size for initial testing
if not max_tweets is None:
    french_tweets = french_tweets.sample(n=max_tweets)

# Split train/test samples : option 2
train, test = train_test_split(french_tweets, test_size=test_size)

# Save train to csv
train.to_csv(dir + 'train.csv', index=False, sep=' ', header=False,
quoting=csv.QUOTE_NONE, quotechar="", escapechar=" ")

model = train_supervised(dir + 'train.csv', epoch=10)

predictions = []
for line in test['text']:
    pred_label = model.predict(line, k=-1, threshold=0.5)[0][0]
    predictions.append(pred_label)

# you add the list to the dataframe, then save the datframe to new csv
test['prediction'] = predictions

s_positive = len(test[test['sentiment']=='__label__positive'].index)
s_negative = len(test[test['sentiment']=='__label__negative'].index)
p_positive = len(test[test['prediction']=='__label__positive'].index)
p_negative = len(test[test['prediction']=='__label__negative'].index)
success = len(test[test['prediction']==test['sentiment']].index)
total = len(test.index)
print(s_positive, s_negative, p_positive, p_negative)
print('Success rate: ', success/total*100, '%')


# Drop all columns except the 'text' column
df = df[['text']]

predictions=[]
for line in df['text']:
    pred_label = model.predict(line, k=-1, threshold=0.5)[0][0]
    predictions.append(pred_label)
```

```
df['prediction'] = predictions
print()
print('All sentiments')
print(df.head())
print()
print('Positive sentiment')
print(df[df['prediction']=='__label__positive'].head())
print()
print('Negative sentiment')
print(df[df['prediction']=='__label__negative'].head())
n_positive = len(df[df['prediction']=='__label__positive'].index)
n_negative = len(df[df['prediction']=='__label__negative'].index)
print('Count positive', n_positive)
print('Count negative', n_negative)

df.to_csv(dir + 'sentiment_prediction.csv')
```

Annexe 6 : Code source python du premier algorithme d'intelligence artificielle utilisé (machine learning : "word2vec")

```
# -*- coding: utf-8 -*-
import numpy as np
import pandas as pd
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
import unicodedata as unicodedata
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import GaussianNB
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
import re
import string
import time

class DenseTransformer():
    def fit(self, X, y=None, **fit_params):
        return self
    def transform(self, X, y=None, **fit_params):
        return X.todense()

# Stemmers remove morphological affixes from words, leaving only the word stem.
ps = PorterStemmer()

# Execution start here
dir = "FOLDER PATH OF THESE SCRIPTS" # ie: /Users/jp/Documents-non-icloud/thèse-
levothyrox/algo/
n = 10000   # tweets subset size (testing)
n2 = 100    # df_doctissimo subset size (testing)
```

```
t0 = time.process_time()

tweets = pd.read_csv('data/french_tweets_updated.csv')

Ntotal = len(tweets)
print()
print('Tweets file size', Ntotal)

t1 = time.process_time()
elapsed_time10 =t1 - t0

# Limit the size for the test phase
tweets = tweets.sample(n=n)

#print(tweets)
#List to hold cleaned tweets and labels
X = [word for word in tweets['text']]
y = list(tweets['sentiment'].values)
#print(X)

t2 = time.process_time()
elapsed_time21 =t2 - t1

print()
print('Time to load', n, 'rows', elapsed_time10, 's')
print('Time to clean', n, 'rows', elapsed_time21, 's')
print('Time to clean full dataset', Ntotal, 'rows', Ntotal*elapsed_time21/n/60,
'mn')

# First you split to train/split and then you train all the steps of your model.
X_train, X_test, y_train, y_test = train_test_split(X, y)

# Use Pipeline as your classifier, this way you don't need to keep calling a
transform and fit all the time.
classifier = Pipeline([('cv', CountVectorizer(max_features=300)), ('to_dense',
DenseTransformer()), ('n_b', GaussianNB())])

# Here you train all steps of your Pipeline in one go.
classifier.fit(X_train, y_train)

t3 = time.process_time()
elapsed_time32 =t3 - t2
print('Time to fit model', elapsed_time32, 'size', n)

y_pred  = classifier.predict(X_test)

combined = np.vstack((y_test, y_pred)).T
comb = pd.DataFrame(data=combined, columns = ['test', 'predicted'])
x1 = len(comb[comb['test'] == comb['predicted']])
```

```
print()
print('Sample size', len(comb))
print('Prediction accuracy on sample', 100*x1/len(comb), '%')

# Load blog messages
df = pd.read_csv(dir + 'data/df_doctissimo.csv')

# Cast date to datetime
df['date'] = pd.to_datetime(df['date'])

# Cast text column to string
df['text'] = df['text'].astype(str)

# Set date column as index
df = df.set_index('date')

# Extract subset for testing purpose only
df = df.head(n2)

# Predict sentiment
to_predict = [word for word in df['text']]
predicted = classifier.predict(to_predict)

df = df.assign(sentiment=predicted)

print()
print(df[['text', 'sentiment']])

# Save result
df.to_csv(dir+'data/'+'with_polarity.csv')
```

Annexe 7 : Code source python du deuxième algorithme d'IA utilisé (sklearn)

```
# -*- coding: utf-8 -*-
import time
import datetime as dt
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import string
import csv
import os

# Execution start here
dir = "FOLDER PATH OF THESE SCRIPTS" # ie: /Users/jp/Documents-non-icloud/thèse-
levothyrox/algo/data/
df = pd.read_csv(dir + 'dataset_doctissimo_updated.csv', encoding='utf8')
print('\n*****************\nFile <dataset_doctissimo_updated.csv> has been loaded')
label = pd.read_csv(dir + 'sentiment_prediction.csv', encoding='utf8')
```

```
print('File <sentiment_prediction.csv> has been loaded\n****************\n')

# Define frequencies
frequencies = {}
frequencies['Y'] = {'label': 'yearly', 'format': '%Y'}
frequencies['M'] = {'label': 'monthly', 'format': '%Y_%m'}
frequencies['W'] = {'label': 'weekly', 'format': '%Y_%U'}
frequencies['D'] = {'label': 'daily', 'format': '%Y_%m_%d'}

def word_occurence(df):
    words = df['text'].str.split()
    full_list = list(itertools.chain(*words))
    counts = Counter(full_list)
    index = []
    values = []
    for key, item in counts.items():
        index.append(key)
        values.append(item)
    return pd.DataFrame(data={'occurence':values}, columns=['occurence'],
index=index)

def n_gram(df, n_gram_size):
    # An n-gram is a contiguous sequence of n items from a given sample of text
    tokens = ' '.join([text for text in df['text']])
    tokens = tokens.split()
    ngrams = zip(*[tokens[i:] for i in range(n_gram_size)])
    list = [' '.join(ngram) for ngram in ngrams]
    counts = Counter(list)
    index = []
    values = []
    for key, item in counts.items():
        index.append(key)
        values.append(item)
    return pd.DataFrame(data={'occurence':values}, columns=['occurence'],
index=index)

def words_cloud(period, df, show=False):
    #fig = None
    wc = None
    all_text = ' '.join([text for text in df['text']])
    all_text = all_text.strip()
    try:
        wc = WordCloud(width=800, height=500,random_state=21, max_font_size=110,
collocations=False).generate(all_text)
    except ValueError:
        print('Value Error: ', period.strftime('%Y_%m_%d'), 'text: [',
all_text,']')
        return None
    fig = plt.figure(figsize=(20, 12))
    plt.imshow(wc, interpolation='bilinear')
```

```
    plt.axis('off')
    if show:
        plt.show()
    plt.close(fig) # Close the window displaying WC (too much memory used)
    return fig

def sentiment_number(row):
   if row['sentiment'] == '__label__positive':
      return 1
   if row['sentiment'] == '__label__negative':
      return 0

def save_csv(df):
    for key, frequency in frequencies.items():
        df_grouped = df.groupby(pd.Grouper(freq=key))
        index = 0
        f = open(dir+'clustering/' + frequency['label'] + '.csv', 'w')
        for period, group in df_grouped:
            if len(group) > 0:
                df1 = word_occurence(group).nlargest(10, ['occurence'])
                lst_1 = []
                if not df1.empty:
                    lst_1 = [text for text in df1.index]
                if len(lst_1) < 10:
                    lst_1.extend(['']*(10-len(lst_1)))
                fragment_1 = ','.join(lst_1)
                df2 = n_gram(group, 2).nlargest(10, ['occurence'])
                lst_2 = []
                if not df2.empty:
                    lst_2 = [text for text in df2.index]
                if len(lst_2) < 10:
                    lst_2.extend(['']*(10-len(lst_2)))
                fragment_2 = ','.join(lst_2)
                xs_p = 100*len(group[group['sentiment'] ==
'__label__positive'])/len(group)
                xs_n = 100*len(group[group['sentiment'] ==
'__label__negative'])/len(group)
                line = str(index) + ',' +
period.strftime(frequencies[key]['format']) + ',' + fragment_1 + ',' + fragment_2 +
',' + str(xs_p) +',' + str(xs_n) + '\n'
                f.write(line)
            index+=1
        f.close()

# Save WC in folders : yearly - monthly - weekly - daily and sort by normal_0 or
abnormal_1 tag // 30 min of execution time
def save_word_clouds(df, start, end):
    # Define intervals
    intervals = []
    # Normal
```

```
    intervals.append({'index':0, 'name':'normal', 'mask':(df.index < start) |
(df.index > end)})
    # Abnormal
    intervals.append({'index':1, 'name':'abnormal', 'mask':(df.index >= start) &
(df.index <= end)})
    for item in intervals:
        for key, frequency in frequencies.items():
            df_grouped = df[item['mask']].groupby(pd.Grouper(freq=key))
            for period, group in df_grouped:
                if len(group.index) > 0:
                    file_path = dir+'cnn/'+ frequency['label'] + '/' + item['name']
+ '/'  + 'world_cloud_' + frequency['label'] + '_' +
period.strftime(frequency['format']) + '.png'
                    #print(period, file_path)
                    fig = words_cloud(period, group)
                    if not fig is None:
                        fig.savefig(file_path)

def clean_folders():
    folders = []
    folders.append(dir+'clustering')
    folders.append(dir+'cnn/yearly/normal')
    folders.append(dir+'cnn/monthly/normal')
    folders.append(dir+'cnn/weekly/normal')
    folders.append(dir+'cnn/daily/normal')
    folders.append(dir+'cnn/yearly/abnormal')
    folders.append(dir+'cnn/monthly/abnormal')
    folders.append(dir+'cnn/weekly/abnormal')
    folders.append(dir+'cnn/daily/abnormal')
    for folder in folders:
        for filename in os.listdir(folder):
            file_path = os.path.join(folder, filename)
            try:
                if os.path.isfile(file_path) or os.path.islink(file_path):
                    os.unlink(file_path)
                elif os.path.isdir(file_path):
                    shutil.rmtree(file_path)
            except Exception as e:
              print('Failed to delete %s. Reason: %s' % (file_path, e))

# Execution starts here
clean_folders()
print('All files in clustering & CNN folders are
deleted\n*******************************')

# DF formating & CSV export for clustering
print('DF label file (sentiment_prediction.csv)\n*******************************')
print(label)
print('\nDF without labeling\n*******************************')
print(df)
```

```
df['sentiment'] = label['prediction'] # Add a column
df['date'] = pd.to_datetime(df['date']) # Cast date to datetime
df['text'] = df['text'].astype(str) # Cast text column to string
df['text'] = df['text'].str.replace('[{}]'.format(string.punctuation), '') # Clean
text

df = df.set_index('date') # Set date column as index
print('\nDF with labeling\n******************************')
print(df)

# Add a numeric column reflecting sentiment value
df['sentiment_number'] = df.apply (lambda row: sentiment_number(row), axis=1)
print('\nDF with labeling and sentiment number\n******************************')
print(df)

# Save labeled dataframe to csv
df.to_csv(dir + 'dataset_doctissimo_updated_labeled.csv', index=False, sep=',',
header=True, encoding='utf8')
# Generate and save top_10 words, top_10 bi-grams, +/- sentiments in csv files
save_csv(df)
print('\n******************************')
print('CSV files are saved in dir+clustering/ : yearly.csv, monthly.csv,
weekly.csv, daily.csv')

# Charts & plots
# Distribution of messages by sentiment
# All dates
fig, ax = plt.subplots(figsize=(20, 10))
df['sentiment_number'].value_counts().plot(ax=ax, kind='bar')
print('Distribution of messages by sentiment (0: negative / 1: positive):\n-> all
dates\n******************************')
plt.show()
# Selected year
fig, ax = plt.subplots(figsize=(20, 10))
df[df.index.year == 2016]['sentiment_number'].value_counts().plot(ax=ax,
kind='bar')
print('\nDistribution of messages by sentiment (0: negative / 1: positive):\n->
2016\n******************************')
plt.show()
# Selected month
fig, ax = plt.subplots(figsize=(20, 10))
df[(df.index.year == 2016) & (df.index.month ==
2)]['sentiment_number'].value_counts().plot(ax=ax, kind='bar')
print('\nDistribution of messages by sentiment (0: negative / 1: positive):\n->
2016-02\n******************************')
plt.show()
# Selected day
fig, ax = plt.subplots(figsize=(20, 10))
```

```
df[df.index == '2016-03-18']['sentiment_number'].value_counts().plot(ax=ax,
kind='bar')
print('\nDistribution of messages by sentiment (0: negative / 1: positive):\n->
2016-03-18\n*******************************')
plt.show()
# Selected range
fig, ax = plt.subplots(figsize=(20, 10))
df[(df.index >= '2016-03-18') & (df.index < '2016-03-
25')]['sentiment_number'].value_counts().plot(ax=ax, kind='bar')
print('\nDistribution of messages by sentiment (0: negative / 1: positive):\n->
2016-03-18 to 2016-03-25\n*******************************')
plt.show()
# Historical line chart per selected sentiment
freq = pd.offsets.Day(30)
fig, ax = plt.subplots(figsize=(20, 10))
ax = df[(df.index.year == 2016) & (df['sentiment'] ==
'__label__positive')]['sentiment_number'].resample(freq).sum().plot.line(ax=ax)
print('\nHistorical line chart per selected sentiment (__label__positive) in
2016\n*******************************')
plt.show()
# Historical line chart of comments per user
fig, ax = plt.subplots(figsize=(20, 10))
df[(df.index.year == 2016) & (df.index.month ==
2)]['user'].value_counts().plot(ax=ax, kind='bar')
print('\nComments per user in 2016-02\n*******************************')
plt.show()

# WC saving for CNN algorithm
start = '2017-07-01' # Start of date range qualified as abnormal
end = '2017-12-31' # End of date range qualified as abnormal
print('Start of date range qualified as abnormal : ' + start + '\nEnd of date range
qualified as abnormal : ' + end + '\n*******************************\n')
print('Generating .png files...')
save_word_clouds(df, start, end) # Generate and save wordclouds as .png image files
print('\n*******************************\n.png files have been saved in
dir+cnn/\n*******************************\n')
```

Annexe 8 : Code source python du script d'analyse sentimentale et de préparation de la donnée en vue d'analyses complémentaires (clustering & CNN)

```
# -*- coding: utf-8 -*-
import time
import datetime as dt
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import string
import nltk
import csv
import re
```

```
import sys
import seaborn as sns
from wordcloud import WordCloud
from common import n_gram, n_gram_history, top_n_gram_history
#nltk.download('punkt')

def get_side_effects(text, side_effects):
    se_list = []
    for regex in list(side_effects):
        if re.search(regex, str(text)):
            result = re.search(regex, str(text)).group(1)
            result = result.replace(' ', '')
            result = result.replace('\'', '')
            result = result.replace('[', '')
            result = result.replace(']', '')
            result = result.replace(',', ' ')
            se_list.append(result)
    return se_list

# Execution start here
dir = "FOLDER PATH OF THESE SCRIPTS" # ie: /Users/jp/Documents-non-icloud/thèse-
levothyrox/algo/data/
side_effects_min_freq = 10
#pd.options.mode.chained_assignment = None

levo_side_effects_extended = ['fatigu\S*', 'astheni\S*',
              'insomni\S*',
              'ma\S* d\S* tete', 'ma\S* \S* d\S* tete' 'cephal\S*',
              'vertig\S*',
              'depressi\S*', 'deprim\S*', 'suicid\S*',
              'douleur\S* musculair\S*', 'douleur\S* \S* musculair\S*',
'myalgi\S*',
              'douleur\S* articulair\S*', 'douleur\S* \S* articulair\S*',
'douleur\S* a\S* articulation\S*', 'douleur\S* \S* a\S* articulation\S*',
'douleur\S* d\S* articulation\S*', 'douleur\S* \S* d\S* articulation\S*',
'douleur\S* articulation\S*', 'douleur\S* \S* articulation\S*', 'arthralgi\S*',
              'chut\S* d\S* cheveu\S*', 'chut\S* \S* d\S* cheveu\S*', 'chut\S*
cheveu\S*', 'chut\S* \S* cheveu\S*', 'pert\S* cheveu\S*', 'pert\S* \S* cheveu\S*',
'pert\S* d\S* cheveu\S*', 'pert\S* \S* d\S* cheveu\S*',
              'pri\S* d\S* poid\S*', 'pri\S* \S* d\S* poid\S*', 'prendre poid\S*',
'prendre \S* poid\S*', 'prendre d\S* poid\S*', 'prendre \S* d\S* poid\S*',
              'pert\S* poid\S*', 'pert\S* \S* poid\S*', 'pert\S* d\S* poid\S*',
'pert\S* \S* d\S* poid\S*',
              'troubl\S* memoir\S*', 'troubl\S* \S* memoir\S*', 'troubl\S* \S* \S*
memoir\S*', 'troubl\S* \S* \S* \S* memoir\S*',
              'anxie\S*',
              'nervosit\S*', 'nerveu\S*', 'irritabilit\S*', 'irritabl\S*',
              'nausee', 'nauseeu\S*',
              'diar\S*',
              'constip\S*',
```

```
                'sue', 'suee', 'suees', 'suer', 'sueur\S*', 'transpi\S*',
                'acouphen\S*',
                'tachycardi\S*','arythmi\S*', 'hyperten\S*', 'hypoten\S*', 'hyper
ten\S*', 'hypo ten\S*']

# Read data
df = pd.read_csv(dir + 'dataset_doctissimo_updated_labeled.csv', encoding='utf8')
print('\n*****************\nFile <dataset_doctissimo_updated_labeled.csv> has been
loaded\n*****************\n')
# Drop useless columns
df = df.drop(columns = ['user', 'url', 'year', 'words_count', 'sentiment'])
# Cast date to datetime
df['date'] = pd.to_datetime(df['date'])
# Set date column as index
df = df.set_index('date')
# Side effects listed in text
df['side_effect_count'] =
df['text'].str.count(r'\b|\b'.join(levo_side_effects_extended))

# Downsample to sample size = 1 day
ONEDAY = pd.offsets.Day(1)
df_daily = df.resample(ONEDAY)["side_effect_count"].sum()

# Daily occurences of side effects reported in messages, all dates
fig, ax = plt.subplots(figsize=(20, 10))
ax = df_daily.plot.line(ax=ax)
#plt.show()
plt.savefig(dir + 'SE-2016-2020.png')
print('\n*****************\nDaily occurence in 2016-2020...
Check\n*****************\n')

# Daily occurences of side effects reported in messages, 2017
fig, ax = plt.subplots(figsize=(20, 10))
ax = df_daily[df_daily.index.year==2017].plot.line(ax=ax)
#plt.show()
plt.savefig(dir + 'SE-2017.png')
print('\n*****************\nDaily occurence in 2017... Check\n*****************\n')

# Normalize
fig, ax = plt.subplots(figsize=(20, 10))
df_normalized = (df_daily - df_daily.mean())/df_daily.std()
ax1 = df_normalized.rolling(window=30).mean().plot.line(ax=ax)
#plt.show()
plt.savefig(dir + 'SE-2017-normalized.png')
print('\n*****************\nNormalize... Check\n*****************\n')

df['side_effects'] = df['text'].apply(get_side_effects,
side_effects=levo_side_effects_extended_1)
print('\n*****************\nDataframe side effects\n*****************\n')
print(df['side_effects'].head(20))
```

```
# Build most common side effects list
a = ' '.join(np.concatenate(df['side_effects']))
words = nltk.word_tokenize(a)
word_dist = nltk.FreqDist(words)
most_common_side_effects = pd.DataFrame(list(filter(lambda x: x[1] >=
side_effects_min_freq, word_dist.items())), columns = ['side_effect', 'frequency'])
mcse_list = most_common_side_effects['side_effect'].to_list()

# Restrict side effects to most_common_side_effects and store in column
df['most_common_side_effects'] = df['side_effects'].apply(lambda l :[x for x in l
if np.isin(x, mcse_list)])

# New dataframe
vector = pd.DataFrame(columns = mcse_list)
for side_effect in mcse_list:
    vector[side_effect] =
df.loc[df['most_common_side_effects'].astype(bool)]['most_common_side_effects'].app
ly(lambda l: int(side_effect in l))
# For testing purpose
#print(vector.astype(bool).sum(axis=0))
print('\n*****************\nVector\n*****************\n')
print(vector.head(20))

# Correlation matrix
correlations = vector.corr()
print('\n*****************\nCorrelation matrix\n*****************')
print(correlations.head(20))

cor_1 =
correlations.where(np.triu(np.ones(correlations.shape)).astype(bool)).stack().sort_
values(ascending=False).reset_index()
cor_1 = cor_1.loc[cor_1['level_0'] != cor_1['level_1']]
cor_1.columns = ['side_effect_1','side_effect_2','Correlation']
print(cor_1.head(20))

# plot the correlation matrix
plt.figure(figsize=(20,20))
sns.heatmap(correlations, cmap='RdBu', vmin=-1, vmax=1, square = True,
cbar_kws={'label':'correlation'})
#plt.show()
plt.savefig(dir + 'heatmap.png')

all_text = ' '.join([text for text in
df['side_effects'].apply(str).str.replace('[{}]'.format(string.punctuation), '',
regex=True)])
print('\n*****************\nWC Generating...\n*****************')
print('Number of words in all_text:', len(all_text))
wordcloud = WordCloud(width=800, height=500,random_state=21, max_font_size=110,
collocations=False).generate(all_text)
```

```
plt.figure(figsize=(20, 12))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off');
#plt.show()
plt.savefig(dir + 'SE-WC.png')

df1 = df.copy()
# Create a 'text' column for use in N_gram and top_N-gram functions
# This has to be improved
# Option 0 : pass a series
# Option 1 : pass the column name in the finctions parameters
# Option 2 : limite de dataframe to a single column
df1['text'] = df1['side_effects'].apply(lambda l: ' '.join(l))
df2 = n_gram_history(df1, 2, 'Y')
print('\n*****************\nn_gram_history\n*****************')
print(df2.head(20))
print('\n*****************\ntop_n_gram_history\n*****************')
df3 = top_n_gram_history(df1, 2, 'Y', 10)
print(df3.head(20))

# Save file
#df.to_csv(dir + 'dataset_doctissimo_side_effect_1.csv', index=False, sep=',',
header=True, encoding='utf8')
```

Annexe 9 : Code source python du script visant à l'étude statistique des résultats et des données obtenues

```
import pandas as pd
import numpy as np
import os
import glob
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix
from keras.models import Sequential, load_model
from keras.layers import Dense, Conv2D, Activation, Dropout, Flatten, MaxPooling2D
from keras.callbacks import ModelCheckpoint, EarlyStopping
from sklearn.model_selection import StratifiedKFold
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score

#one hot encoding function
def one_hot_encoder(df_name, df_column_name, suffix=''):
    temp = pd.get_dummies(df_name[df_column_name]) #get dummies is used to create
dummy columns
    df_name = df_name.join(temp, lsuffix=suffix) #join the newly created dummy
columns to original dataframe
    df_name = df_name.drop(df_column_name, axis=1) #drop the old column used to
create dummy columnss
    return df_name
```

```
#function to draw confusion matrix
def draw_confusion_matrix(true,preds):
    conf_matx = confusion_matrix(true, preds)
    sns.heatmap(conf_matx, annot=True,annot_kws={"size": 12},fmt='g', cbar=False,
cmap="viridis")
    plt.show()
    #return conf_matx

def cnn_model(size, num_cnn_layers):
    NUM_FILTERS = 32
    KERNEL = (3, 3)
    #MIN_NEURONS = 20
    MAX_NEURONS = 120
    model = Sequential()
    model.add(Conv2D(128, (6,6), input_shape=(100,100,3), activation='relu'))
    model.add(MaxPooling2D(pool_size=(2,2)))
    model.add(Conv2D(128, (6,6), activation='relu'))
    model.add(MaxPooling2D(pool_size=(2,2)))
    model.add(Conv2D(128, (6,6), activation='relu'))
    model.add(MaxPooling2D(pool_size=(2,2)))
    model.add(Conv2D(128, (3,3), activation='relu'))
    model.add(MaxPooling2D(pool_size=(2,2)))
    model.add(Conv2D(128, (2,2), activation='relu'))
    model.add(Flatten())
    model.add(Dropout(0.5))
    model.add(Dense(256, activation='relu'))
    model.add(Dense(50, activation='relu'))
    model.add(Dense(2, activation='softmax'))
    model.compile(loss='categorical_crossentropy', optimizer='adam',
metrics=['accuracy'])
    #print(model.summary())
    return model

def fit_and_evaluate(tr_x, ts_x, tr_y, ts_y, tro_y, tso_y, EPOCHS=100,
BATCH_SIZE=50):
    model = None
    model = cnn_model(IMAGE_SIZE, 2)
    results = model.fit(tr_x, tr_y, epochs=EPOCHS, validation_split = 0.2,
batch_size=BATCH_SIZE,  verbose=1)
    print("Val Score: ", model.evaluate(ts_x, ts_y))
    cc=model.predict_classes(ts_x)
    del model
    return cc
#callbacks=[early_stopping, model_checkpoint],

path='./Data'
fdal = open('./Results/Consolidated.txt','w')
fname =glob.glob(path+'/'+'*.csv')
for fn in fname:
```

```
    print(fn)
    train_images = pd.read_csv(fn)
    train_images_x = train_images.iloc[:,1:]
    train_images_array = train_images_x.values
    train_x = train_images_array.reshape(train_images_array.shape[0], 100, 100, 3)
    train_x_scaled = train_x/255
    IMAGE_SIZE = (100, 100, 3)
    train_images_y = train_images[['0']]
    #do one hot encoding with the earlier created function
    train_images_y_encoded = one_hot_encoder(train_images_y, '0', 'lab')
    #get the labels as an array
    train_images_y_encoded = train_images_y_encoded.values
    train_images_y=train_images.iloc[:,0].values
    # Defining the CNN Architecture
    model = cnn_model(IMAGE_SIZE, 2)
    model.summary()
    pat = 5
    early_stopping = EarlyStopping(monitor='val_loss', patience=pat, verbose=1)
    model_checkpoint = ModelCheckpoint('fas_mnist_1.h5', verbose=1,
save_best_only=True)
    n_folds=5
    epochs=100
    batch_size=50
    model_history = []
    logo = StratifiedKFold(n_splits=n_folds)
    subject_index = 1
    preds=[]
    trues=[]
    for train_index, test_index in logo.split(train_x_scaled, train_images_y,
train_images_y_encoded):
        train_x, test_x = train_x_scaled[train_index], train_x_scaled[test_index]
        train_y,test_y= train_images_y_encoded[train_index],
train_images_y_encoded[test_index]
        train_y_org,test_y_org=train_images_y[train_index],
train_images_y[test_index]
        model_checkpoint =
ModelCheckpoint('.\Models\\valen_'+str(subject_index)+'.h5', verbose=1,
save_best_only=True)
        ft=fit_and_evaluate(train_x, test_x, train_y,
test_y,train_y_org,test_y_org, epochs, batch_size)
        print("=======*12, end="\n\n\n")
        subject_index = subject_index + 1
        preds=np.concatenate((preds,ft),axis=0)
        trues=np.concatenate((trues,test_y_org))
    cr=classification_report(trues,preds)
    cnf=confusion_matrix(trues, preds)
    print(cr)
    print(cnf)
    print(fn)
    op=fn.split('/')[2].split('.')[0]
```

```
    fd = open('./Results/'+op+'.txt','w')
    np.savetxt(fd, cnf, delimiter=",")
    fd.close()
    fdal.write(op)
    fdal.write('.txt\n')
    fdal.write(str(accuracy_score(trues,preds)))
    fdal.write('\n')
fdal.close()
```

Annexe 10 : Code source python du script d'intelligence artificielle de « Convolutional Neural Network »

Accessible en téléchargement via le lien suivant (104 Mo) :

https://drive.google.com/uc?export=download&id=110CnvJORmRhraNVE3clNnl8YNF9litqN

ROBERT Jean-Philippe – ROCHE Valentin

Élaboration d'une méthode de détection précoce d'évènements indésirables déclarés par les patients sur les réseaux sociaux : cas du Levothyrox® sur le site Doctissimo®

Th. D. Pharm., Lyon 1, 2022, 194 p.

RESUME

Basée sur l'affaire du Levothyrox® de 2017, l'objectif de ce projet est d'élaborer une méthode pour détecter plus précocement, via des algorithmes informatiques, les effets indésirables du médicament sur le sous-forum endocrinologie du site Doctissimo®. Cette approche algorithmique permet d'étendre les capacités de la pharmacovigilance. En effet, les réseaux sociaux constituent une source de données inépuisable pour la détection de signaux faibles. Dans le cadre de ce travail, la construction du prototype s'est faite selon un mode séquentiel et empirique en l'absence de méthode de référence. Le langage informatique python, différentes librairies (pandas, numpy, nltk, Spacy, etc.), ainsi que plusieurs algorithmes open source d'intelligence artificielle (Fasttext, sklearn, gensim, CNN) ont été utilisés.

Ce projet, précurseur dans le domaine de la pharmacovigilance, présente des résultats très encourageants. Ils mettent en évidence une capacité réelle d'innovation concernant l'utilisation des data sciences et de l'intelligence artificielle à des fins de traitement statistiques des données de vie réelle des patients. Il permet d'envisager, après ajustement des biais et mise en place des pistes d'améliorations, l'extrapolation du modèle à d'autres sources de données et d'autres scripts d'analyses (avec ou sans l'utilisation de l'intelligence artificielle). Toutefois, le chemin est encore long pour espérer rencontrer ces outils dans la pratique quotidienne. La formation d'une équipe pluridisciplinaire sera un prérequis pour construire un outil applicable à tous les médicaments et pathologies sur tous les réseaux sociaux.

MOTS CLES

Levothyrox
Machine learning
Pharmacovigilance

JURY

M. DUSSART Claude, PU-PH
MME. SALAM Hanan, Docteur en intelligence artificielle
M. ARMOIRY Xavier, PU-PH
MME. BARDEL-DANJEAN Claire, MCU-PH
M. BOUREILLE Antoine, Docteur en pharmacie

DATE DE SOUTENANCE

Mardi 8 mars 2022

CONTACT

M. Claude DUSSART – claude.dussart@univ-lyon1.fr