



HAL
open science

Détection automatisée de signaux de pharmacovigilance dans les médias sociaux

Pierre Karapetiantz

► **To cite this version:**

Pierre Karapetiantz. Détection automatisée de signaux de pharmacovigilance dans les médias sociaux. Informatique et langage [cs.CL]. Sorbonne Université, 2023. Français. NNT : 2023SORUS077 . tel-04137249

HAL Id: tel-04137249

<https://theses.hal.science/tel-04137249v1>

Submitted on 22 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



THESE DE DOCTORAT EN
SCIENCES DE L'INFORMATION BIOMEDICALE

ECOLE DOCTORALE 393
PIERRE LOUIS DE SANTE PUBLIQUE :
EPIDEMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMEDICALE

Détection automatisée de signaux de pharmacovigilance dans les médias sociaux

Auteur :

Pierre KARAPETIANTZ

Dirigée par :

Marie-Christine JAULENT

Cédric BOUSQUET

Thèse réalisée dans le Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances pour la e-Santé,
LIMICS-UMRS1142
Campus des Cordeliers
15 rue de l'école de médecine, 75006

Soutenue le 24 mars 2023, Devant le jury composé de :

Marie-Christine JAULENT, Directrice de Recherche INSERM (Directrice de thèse)

Cédric BOUSQUET, Praticien hospitalier, CHU de Saint-Etienne (Encadrant)

Pascal STACCINI, PU-PH, Université Côte d'Azur (Rapporteur)

Béatrice TROMBERT PAVIOT, PU PH, CHU de Saint-Etienne (Rapporteur)

Mehdi BENKEBIL, Chef de pôle Plateforme de réception et d'orientation des signalements, ANSM (Examineur)

Bernd AMANN, professeur en informatique et membre du laboratoire LIP6 (Examineur, Président du jury)

Thierry DART, Directeur de pôle, Agence du numérique en santé (Examineur)

Sandrine KATSAHIAN, PU-PH, Université de Paris 5 René Descartes (Examineur)

Remerciements

J'adresse mes plus sincères remerciements à ma directrice de thèse, Marie-Christine Jaulent, ainsi qu'à mon encadrant, Cédric Bousquet, pour leurs conseils et pour m'avoir soutenu tout au long de ce travail de thèse.

Je souhaite également remercier les membres de l'U1142 pour leur accueil et leurs conseils.

J'exprime ma gratitude à Madame Béatrice Trombert Paviot et Monsieur Pascal Staccini qui ont bien voulu être rapporteurs, ainsi qu'à Madame Sandrine Katsahian et Messieurs Bernd Amann, Mehdi Benkebil et Thierry Dart qui ont accepté d'être examinateurs.

Je remercie également mes parents, mon frère, ma nièce, ainsi que mes amis pour leur soutien.

Résumé

Introduction

La pharmacovigilance est définie par l'Organisation Mondiale de la Santé (OMS) comme « la science et les activités relatives à la détection, la compréhension et la prévention des effets indésirables ou de tout autre problème lié aux médicaments. » Les systèmes de pharmacovigilance s'appuient sur de grandes bases de données regroupant des notifications individuelles d'événements indésirables susceptibles d'être liés à la prise de médicaments. L'importante sous-déclaration des effets indésirables dus aux médicaments dans les bases nationales de pharmacovigilance (estimée en France entre 78% et 99% entre 1997 et 2002) ainsi que les différents scandales pharmaceutiques de ces dernières années ont fait se tourner la pharmacovigilance vers de nouvelles sources de données. Ainsi, il est reconnu que la déclaration spontanée des effets indésirables par des professionnels de santé n'est pas la source unique permettant d'alimenter les bases de données de pharmacovigilance, et une activité importante des agences réglementaires consiste à surveiller ces différentes sources de données afin de pouvoir prendre des décisions éclairées sur les mesures nécessaires à la réduction du risque médicamenteux.

Matériel et méthode

Dans cette thèse, nous avons choisi de nous intéresser aux médias sociaux, mais d'autres sources émergent également telles que les dossiers patients électroniques dans les hôpitaux ou les bases médico-administratives. Si l'analyse des médias sociaux ne peut être considérée comme un système de détection de signaux prioritaire, l'exploration de cette source de données est importante, les données y étant différentes de celles retrouvées dans les bases traditionnelles. Notre hypothèse est que des analyses statistiques peuvent être adaptées aux données issues des médias sociaux pour une analyse régulière et opérationnelle.

Pour se faire, j'ai appliqué des méthodes de détection du signal et réalisé une analyse des thèmes abordés par les utilisateurs de réseaux sociaux. J'ai conçu et utilisé une interface utilisateur qui a été implémentée grâce à l'aide d'un informaticien du laboratoire et qui complète une chaîne de traitement permettant d'appliquer ces analyses manière quotidienne.

Résultats

Finalement, cette chaîne de traitement composée de 4 modules indépendants et modifiables permet d'étudier de manière efficace les données issues de forums français de discussion en santé pour une utilisation en pharmacovigilance.

Discussion

J'ai pu montrer au cours de mon travail de thèse qu'utiliser cette chaîne de traitement, au travers d'une interface, permet une étude quantitative simple, sans besoin de coder, et dont les résultats sont cohérents avec la littérature, et permettant même d'obtenir de nouvelles informations à propos des médicaments.

Summary

Pharmacovigilance, defined by the World Health Organization as “the science and activities relating to the detection, assessment, understanding, and prevention of adverse effects or any other drug-related problem”, depends mainly on spontaneous reporting. Adverse Drug Reaction (ADR) reporting was first limited to healthcare professionals but has progressively opened to patients using dedicated platforms in many countries in the last decades. Several studies have demonstrated the value of the information given by patients on ADRs. However, under-reporting and incomplete data are still major limitations for pharmacovigilance (in France, under-reporting was estimated between 78% and 99% between 1997 and 2002). Moreover, several « scandals » made pharmacovigilance to explore different data sources in order to make decisions to reduce drug-related risks.

In this thesis, we chose to focus on social media, but other data sources are to explore such as electronic health record in hospital or medico-administrative databases. If the analysis of social media cannot be considered as a prior signal detection system, the exploration of this data source is important, as the data there being different from those found in traditional databases.

Our hypothesis is that statistical analyzes can be adapted to data from social media for regular and operational analysis. To do so, we applied signal detection methods and an analysis of the topics addressed by social network users. Then we implemented and used a user interface, completing a processing chain to apply these analyzes on a daily basis. We have therefore developed a pipeline made up of 4 independent and modifiable modules to effectively study data from French health discussion forums for use in pharmacovigilance. We have shown that using this processing chain, through an interface, allows a simple quantitative study, without the need to code, and whose results are consistent with the literature, and even allowing to obtain new information about medication.

Table des matières

Remerciements	3
Résumé	4
Summary	6
Abréviations	11
Avant-Propos	13
Chapitre 1	
Introduction	15
1. La pharmacovigilance.....	15
2. Les notifications	16
3. Implication et limites de la pharmacovigilance	18
4. Déclarations spontanées en France	19
5. Les médias sociaux.....	21
6. Hypothèse, objectif et plan de thèse	22
Chapitre 2	
Etat de l'art	23
1. Terminologies en pharmacovigilance	23
1.1 MedDRA.....	23
1.2 Classification ATC.....	24
1.3 UMLS.....	25
1.4 Langage patient : de possibles ressources ?.....	25
2 Identification des effets indésirables liés aux médicaments dans les messages.....	29
3 Algorithmes de détection du signal en pharmacovigilance	30
3.1 Méthodes basées sur la disproportionnalité.....	30
3.2 Autres méthodes	33
3.3 Conclusion sur les méthodes de détection du signal	36
4 La pharmacovigilance dans les médias sociaux	37
4.1 Recommandations.....	37
4.2 Ethique.....	37
4.3 Réglementations.....	37
4.4 Avantages et inconvénients de la pharmacovigilance dans les médias sociaux	38
4.5 Analyse des thèmes abordés dans les médias sociaux.....	42
4.6 Chaînes de traitement : évaluation de l'existant.....	43
Chapitre 3	
Matériel	50
1. Choix des forums.....	50
2. Module Scraper.....	50
3. Module annotation	55
Chapitre 4	
Méthode	60
1. Création et évaluation du module d'analyse (module 3)	60
1.1. Cas d'usages pour la detection du signal.....	62
1.2. Cas d'usage pour les Topic Models.....	65
2. Création du module 4 : l'interface utilisateur.....	67

2.1. Conception.....	67
2.2. Choix technologique pour l'implémentation.....	71
Chapitre 5	
Résultats	72
1. Module d'analyse : Détection du signal.....	72
1.1. Cas d'usage du Levothyrox	72
1.2. Cas d'usage des Vaccins HPV	77
2. Module d'analyse : Topic Models	78
3. Module 4 : L'interface.....	79
3.1. Implémentation du module de l'interface	79
3.2. Echanges avec les utilisateurs	82
Chapitre 6	
Discussion.....	85
1. Module d'analyse : Détection du signal.....	85
1.1. Cas d'usage du Levothyrox	85
1.2. Cas d'usage des vaccins HPV	85
2. Module d'analyse : Topic Models	86
3. Discussion : chaîne de traitement.....	86
4. Discussion générale.....	88
Chapitre 7	
Conclusion	90
Bibliographie	91
Annexe 1 : Description des projets	97
Annexe 2 : Captures d'écran de la BNPV	100

Liste des figures

<i>Figure 1 : BNPV - Informations sur le cas</i>	18
<i>Figure 2 : Portail de déclaration - Questionnaire</i>	20
<i>Figure 3 : Portail de déclaration - Description de l'EI</i>	20
<i>Figure 4 : Exemple de classification MedDRA</i>	24
<i>Figure 5 : Exemple de réseau neuro-bayésien</i>	32
<i>Figure 6 : Structure du Vigi4Med Scraper adapté pour PHARES</i>	54
<i>Figure 7 : Structure de la base de données PHARES</i>	54
<i>Figure 8 : Message numéro 29605 du forum alloDocteurs</i>	55
<i>Figure 9 : Exemple de message annoté</i>	56
<i>Figure 10 : Capture d'écran de la maquette</i>	68
<i>Figure 11 : Diagramme UML - Cas d'utilisation</i>	68
<i>Figure 12 : Diagramme UML – Classes</i>	69
<i>Figure 13 : Sources et évolution temporelle des messages concernant le levothyrox</i>	73
<i>Figure 14 : Distribution des SOCs des EIs retrouvés avec le levothyrox</i>	73
<i>Figure 15 : Sources et évolution temporelle des messages concernant le couple Levothyrox-fatigue</i>	75
<i>Figure 16 : Evolution temporelle du PRR pour le couple Levothyrox-fatigue</i>	75
<i>Figure 17 : Résultats du CISL pour l'EI fatigue</i>	76
<i>Figure 18 : Points de rupture identifiés par le CPA pour l'évolution temporelle du nombre de mentions du couple Levothyrox-fatigue</i>	76
<i>Figure 19 : Evolution du PRR pour les 5 EIs sélectionnés.</i>	77
<i>Figure 20 : Sélection du nombre de thèmes via la perplexité</i>	78

Liste des tableaux

<i>Tableau 1: Classification ATC du méthylphénidate</i>	<i>25</i>
<i>Tableau 2 : Table de contingence pour le couple Médicament i - E_j</i>	<i>31</i>
<i>Tableau 3 : Evaluation des chaînes de traitement selon nos critères.....</i>	<i>Erreur ! Signet non défini.</i>
<i>Tableau 4 : Volumétrie de l'extraction</i>	<i>53</i>
<i>Tableau 5 : Extrait de la base de données.....</i>	<i>58</i>
<i>Tableau 6 : Volumétrie des annotations</i>	<i>59</i>
<i>Tableau 7 : Annotations de Levothyrox par forum.....</i>	<i>72</i>
<i>Tableau 8 : 20 PTs les plus retrouvés avec le levothyrox.....</i>	<i>74</i>
<i>Tableau 9 : Principaux thèmes abordés dans les topics revus manuellement</i>	<i>79</i>

Abréviations

ADR-Prism	Adverse Drug Reactions from Patient Reports in Social Media
ANSM	Agence Nationale de Sécurité du Médicament et des produits de santé
API	Interface de programmation d'application
ATC	Système de classification anatomique, thérapeutique et chimique
BCPNN	Bayesian confidence propagation neural network
BERT	Bidirectional Encoder Representations from Transformers
BNPV	Base Nationale de Pharmacovigilance
CHV	Consumer Health Vocabulary
CISL	Class Imbalanced Subsampling Lasso
CISMef	Fondation Health On the Net
CPA	Change Point Analysis
CRPV	Centre Régionale de Pharmacovigilance
CSV	Comma-Separated Values
CUI	Concept Unique Identifiant
DOM	Document Object Model
DTM	Matrice terme-document
EI	Effet indésirable
EMA	Agence européenne des médicaments
FAERS	FDA Adverse Event Reporting System
FDA	Agence américaine des produits alimentaires et médicamenteux
HAS	Haute Autorité de Santé
HeTOP	Health Terminology/Ontology Portal
HITEx	Health Information Text Extraction
HLGT	Groupe de termes de haut niveau
HLT	Terme de haut niveau
HPV	Papillomavirus humain
ICH	Conseil international d'harmonisation des exigences techniques pour l'enregistrement des médicaments à usage humain
LIMICS	Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé
LIMSI	Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur
LLT	Terme de plus bas niveau
MedDRA	Dictionnaire Médical pour les Activités Règlementées
MeSH	Medical Subject Headings
MGPS	Multi Gamma Poisson Shrinker
OMS	Organisation Mondiale de la Santé
PHARES	PHarmacovigilance dans les Réseaux Sociaux
PREDOSE	PREscription Drug abuse Online Surveillance and Epidemiology

PRR	Proportional Reporting Ratio
PT	Terme préférentiel
RCP	Résumé des Caractéristiques du Produit
RGPD	Règlement sur la Protection des Données
ROC	Receiver operating characteristic, pour « caractéristique de fonctionnement du récepteur »
ROMEDI	Référentiel ouvert du médicament
ROR	Reporting odds ratio
SHA1	Secure Hash Algorithm 1
SOC	System Organ Class, Classification par discipline médicale
TAL	Traitement automatique du langage
TDAH	Trouble déficitaire de l'attention/hyperactivité
UMC	Observatoire d'Uppsala
UML	Langage de Modélisation Unifié
UMLS	Unified Medical Language System
VEM	Variational Expectation-Maximization
Vigi4Med	Vigilance dans les forums sur les médicaments
WEB-RADR	Recognizing Adverse Drug Reactions

Avant-Propos

Avant ce travail de thèse, à partir de janvier 2016, j'ai participé à trois projets évaluant l'analyse des médias sociaux pour la pharmacovigilance présentés en Annexe 1.

Ingénieur d'étude dans l'équipe 22 de l'Inserm U1138

Après mon master de recherche Santé Publique parcours biostatistique obtenu en 2015, j'ai travaillé en tant qu'ingénieur d'étude au sein de l'équipe 22 de l'unité Inserm U1138 : « Sciences de l'information au service de la médecine personnalisée » de Janvier 2016 à Mai 2017. Au cours de cette expérience professionnelle, j'ai travaillé sur deux projets décrits en annexe traitant de l'utilisation des médias sociaux pour la pharmacovigilance :

- le projet ADR-PRISM (Adverse Drug Reactions from Patient Reports in Social Media), et
- le projet Vigi4Med (Vigilance dans les forums sur les médicaments)

Pour le projet ADR-PRISM, mon travail a porté sur l'application de méthodes de détection du signal, notamment pour la publication d'une étude quantitative et qualitative sur le Méthylphénidate dans laquelle une méthode permettant d'analyser les thèmes sous-jacents des messages citant cette molécule a également été menée [1].

Avec notamment Xiaoyi Chen et Armelle Guenegou-Arnoux de l'équipe 22 de l'unité U1138, nous avons mis au point un protocole d'évaluation d'un logiciel permettant d'extraire les effets indésirables liés aux médicaments à partir des messages extraits de forums de discussion, en précisant la méthode de calcul de l'effectif statistiquement nécessaire à cette évaluation [2]. Puis nous avons évalué l'outil développé lors du projet ADR-PRISM en suivant ce protocole [3].

Pour le projet Vigi4Med, j'ai appliqué des méthodes de détection du signal afin de sélectionner les messages qui allaient ensuite être revus par des pharmacovigilants via une interface créée par Damien Leprovost, du LIMICS.

Ingénieur d'étude au LIMICS, U1142

A la fin de mon contrat avec l'équipe 22 de l'unité U1138, j'ai été recruté en juin 2017 par le LIMICS (unité U1142) pour continuer à travailler sur le projet Vigi4Med, puis également sur le projet en découlant : le projet PHARES (PHarmacovigilance dans les REseaux Sociaux).

Tout d'abord, j'ai utilisé les données saisies par des pharmacovigilants du Centre Régional de Pharmacovigilance (CRPV) de l'Hôpital Européen Georges Pompidou. Les pharmacovigilants sont chargés de recueillir des dossiers d'effets indésirables médicamenteux et de saisir les observations dans la base nationale de pharmacovigilance, de répondre aux questions sur le médicament et d'informer les professionnels de santé. J'ai comparé les données saisies pour Vigi4Med à celles des cas de pharmacovigilance de la base nationale de l'ANSM pour six médicaments. Cette comparaison concernait principalement l'informativité, c'est-à-dire si l'information était disponible dans les messages et dans les cas de pharmacovigilance de la base nationale [4]. Puis, nous avons réalisé une publication similaire en se concentrant sur le Baclofène [5]. J'ai obtenu des résultats satisfaisants en appliquant des méthodes de détection du signal aux données de Vigi4Med pour retrouver les effets indésirables connus du Baclofène ainsi que des effets inattendus [6]. Dans une étude sur le Protelos [7], j'ai présenté et démontré la complémentarité de deux outils permettant l'un de mener une étude qualitative (l'outil créé par Damien Leprovost) et l'autre une étude quantitative (outil créé par Juliette Faille, alors étudiante à Supelec, durant son stage de 2 mois que j'ai participé à encadrer).

Master d'Informatique Biomédical

Souhaitant réaliser une thèse, je m'étais inscrit au concours pour avoir une bourse de l'Ecole Doctorale 393. Mais, à la suite d'une hospitalisation m'ayant empêché de me présenter à l'oral, je me suis inscrit en master de santé publique en Informatique Biomédicale. J'ai réalisé mon stage de master de 5 mois (février 2019 à juin 2019) en travaillant sur le projet PHARES, dont l'objectif était de réaliser une suite logicielle allant de l'extraction des messages à leur analyse via une interface utilisateur. J'ai participé à la conception de cette suite logicielle (en précisant les spécifications et en rédigeant le guide utilisateur et le cahier des charges de l'interface). L'extraction des messages a été réalisée par Bissan Audeh du LIMICS, leur annotation par le LIMSI, j'ai travaillé sur le module d'analyse, et l'implémentation de l'interface a été réalisée par Akram Redjdal du LIMICS.

Thèse

Au cours de cette thèse, j'ai participé à l'application et l'évaluation de différentes méthodes d'analyse ainsi que de la chaîne de traitement avec la publication de 2 articles publiés sur l'application et l'évaluation de méthodes d'analyses et un troisième article en cours de soumission qui traite de la chaîne de traitement, de son utilisation et de son évaluation. Ce dernier article compare nos besoins aux chaînes de traitement existantes, puis décrit notre chaîne de traitement et montre son application avec le cas d'usage du Levothyrox présenté dans cette thèse.

Chapitre 1

Introduction

1. La pharmacovigilance

La pharmacovigilance est définie par l'Organisation Mondiale de la Santé (OMS) comme « la science et les activités relatives à la détection, la compréhension et la prévention des effets indésirables ou de tout autre problème lié aux médicaments » [1].

L'activité de pharmacovigilance se distingue selon qu'elle intervient avant ou après la mise sur le marché du médicament. En phase clinique, il s'agit de tester la toxicité et les éventuels effets indésirables qui doivent tenir compte de la balance bénéfique/risque et de la disponibilité d'autres molécules également efficaces qui présenteraient moins d'effets indésirables pour envisager la mise sur le marché du médicament.

Dans ce travail de thèse nous nous intéressons exclusivement à l'activité de pharmacovigilance qui débute une fois le médicament commercialisé, et qui a pour but de détecter les effets indésirables plus rares ainsi que les complications liées à l'administration sur une plus longue période. En effet, les conditions d'exposition en vie réelle sont très différentes de celles des essais cliniques dans lesquels la population est très homogène, avec des caractéristiques spécifiques et des populations de taille limitée. La pharmacovigilance post-commercialisation a donc pour but de détecter le plus précocement possible les effets indésirables non identifiés lors des essais cliniques.

Les systèmes de pharmacovigilance s'appuient sur de grandes bases de données regroupant des notifications individuelles d'événements indésirables soupçonnés d'être liés à la prise de médicaments. Il existe de nombreuses agences nationales et supranationales disposant de tels systèmes de notifications spontanées : l'Agence américaine des produits alimentaires et médicamenteux (FDA), l'Agence européenne des médicaments (EMA), l'Observatoire d'Uppsala (Uppsala Monitoring Center, UMC), centre collaborateur en pharmacovigilance pour l'OMS. En France, la base nationale de pharmacovigilance (BNPV) est gérée par l'Agence Nationale de Sécurité du Médicament et des Produits de Santé (ANSM). Elle fût construite en 1979 et est constituée des notifications reçues par les 32 centres régionaux de pharmacovigilance répartis sur l'ensemble du territoire afin de favoriser les échanges avec les professionnels de santé (CRPV). Les CRPV portent à la connaissance de l'ANSM les signaux potentiels qu'ils auraient observés. Un signal est défini par l'OMS comme « une relation entre un médicament et un effet, cette relation étant à ce jour inconnue ou non documentée ». La détection du signal est donc la détection d'un nouveau « couple » médicament-Effet

Indésirable (EI), mais aussi de nouveaux facteurs de risque ou interactions. Cette détection peut se faire manuellement par revue des cas. C'est un processus complexe qui comporte l'identification d'association(s) inconnue(s), l'évaluation causale et la quantification de l'association. Afin de détecter les problèmes liés aux médicaments et prioriser leur évaluation par des experts (les bases de pharmacovigilance ayant des tailles variables allant de 400 000 à plus de 4 millions de cas), des méthodes statistiques de détection de signal ont été proposées entre 1998 et 2001 et sont aujourd'hui appliquées en routine sur les bases de pharmacovigilance [2]. Historiquement, ces méthodes sont basées sur des mesures de disproportionnalité entre le nombre de cas observé et celui attendu. Cependant ces méthodes ont leurs limites. Elles ne permettent pas de détecter des augmentations soudaines du nombre de cas, ce qui explique le développement récent de méthodes d'analyses temporelles qui mettent en évidence des changements au niveau des tendances. Elles ne prennent pas en compte la possibilité d'interactions médicamenteuses, d'où le développement de méthodes basées sur la régression logistique. Certains types de méthodes pourraient être associés, comme celles basées sur la disproportionnalité et les méthodes d'analyses temporelles, mais l'intérêt d'une telle approche reste insuffisamment évalué [3]. Ces méthodes sont bien décrites mathématiquement mais aucune n'est considérée comme méthode de référence. Ainsi, l'OMS via sa plateforme Vigilyze appliquait le Bayesian confidence propagation neural network (BCPNN) [4], remplacé en 2014 par VigiRank [5], la FDA applique le Multi Gamma Poisson Shrinker (MGPS) [6,7], l'EMA applique le Proportional Reporting Ratio (PRR) [8] et le Centre de Pharmacovigilance des Pays-Bas, le Lareb, applique la régression logistique [9].

2. Les notifications

Cette introduction générale de la pharmacovigilance souligne le rôle essentiel des notifications. Ces notifications sont historiquement rapportées par des professionnels de santé (médecins, pharmaciens, laboratoires pharmaceutiques, etc.) et contiennent des informations sur 1) le patient, 2) le médicament (nom commercial, principe actif, posologie, date de traitement, etc...) et 3) le(s) effet(s) indésirable(s) (EI(s)) constaté(s) (durée avant apparition, évolution, gravité, etc...). Une méthode d'imputabilité commune est appliquée, afin de contrôler la qualité des informations recueillies. Cette méthode a été élaborée en 1978 par Dangoumau et al. [10] et a été actualisée en 1985 par Bégau et al. [11]. L'imputabilité intrinsèque est distinguée de l'imputabilité extrinsèque. L'imputabilité intrinsèque cherche à établir la relation de causalité entre chaque médicament et la survenue d'un EI.

La relation intrinsèque s'évalue selon un score d'imputabilité chronologique et un score d'imputabilité sémiologique. L'imputabilité chronologique dépend de la compatibilité chronologique entre la prise médicamenteuse et le délai de survenue de l'EI, l'évolution à l'arrêt du médicament (déchallenge) et

la récurrence en cas de nouvelle administration du médicament (rechallenge). L'imputabilité sémiologique dépend de l'explication pharmacodynamique (mécanisme d'action), des facteurs favorisants, des diagnostics différentiels possibles et des examens complémentaires de laboratoire prouvant la cause médicamenteuse (positif, non fait, négatif).

Le score d'imputabilité extrinsèque repose sur la recherche de cas similaires dans la littérature selon différents critères :

- Effet notoire, décrit : référencé dans les ouvrages de référence (dictionnaire des médicaments, Vidal, Martindale, Meyler's Side Effects of Drugs).
- Effet non notoire dans les documents usuels : publié à une ou deux reprises avec une sémiologie différente ou un médicament voisin.
- Effet non décrit : non décrit dans la littérature.
- Effet non décrit : non décrit après recherche exhaustive dans la littérature.

La base française est constituée de six blocs de données (figure 1 et figures en Annexe 2) :

1. Les caractéristiques administratives du cas : date de déclaration, type d'effet, date de dernière information et traçabilité de celui qui saisit et celui qui valide.
2. Le déclarant : type (professionnel de santé, ou patient et si professionnel de santé quel type), type d'exercice (hôpital, libéral, institutionnel, clinique ...) et localisation géographique.
3. Le patient : initiales, sexe âge (date de naissance) poids taille, antécédents.
4. Les médicaments :
 - a. Le(s) médicament(s) suspecté(s) : nom, voie d'administration, date de début et de fin, posologie, indication.
 - b. L'imputabilité de(s) médicament(s) suspecté(s).
5. Les effet(s) indésirable(s) : date de survenue, description de l'effet, gravité et évolution
6. Champ commentaire (libre).

Figure 1 : BNPV - Informations sur le cas

3. Implication et limites de la pharmacovigilance

La pharmacovigilance s'implique également dans la gestion de l'iatrogénie médicamenteuse, c'est-à-dire les conséquences néfastes pour la santé d'interventions médicales (tels que les erreurs de diagnostic, une prévention ou prescription inadaptée ou la complication d'un acte thérapeutique) ou de recours aux soins ou de l'utilisation d'un produit de santé, prenant en compte l'ensemble des risques médicamenteux, aussi bien dans le cadre d'une utilisation normale que d'un mésusage ou d'un abus.

L'exercice de la pharmacovigilance peut nécessiter la réalisation d'études et de travaux concernant la sécurité d'emploi des médicaments et produits de santé, ainsi que la recherche et l'analyse des données contenues dans le dossier préclinique d'expérimentation animale ou dans le dossier des essais cliniques d'un médicament ou produit, ou d'informations concernant sa fabrication et sa conservation. La pharmacovigilance tient également compte des produits pouvant modifier leurs risques et bénéfiques (interactions). Un comité technique de pharmacovigilance centralise et évalue l'ensemble des informations sur le risque médicamenteux afin d'émettre un avis à l'ANSM sur les mesures à prendre pour informer sur ce risque, le prévenir ou en diminuer la gravité.

L'importante sous-déclaration des effets indésirables dus aux médicaments dans la base nationale de pharmacovigilance (estimée en France entre 78% et 99% entre 1997 et 2002 [12]) ainsi que les différents scandales pharmaceutiques de ces dernières années, comme la « crise sanitaire » du Médiator [13], ont fait se tourner la pharmacovigilance vers de nouvelles sources de données. Ainsi, Depuis plusieurs décennies, il est reconnu que la déclaration spontanée des effets indésirables par des professionnels de santé n'est pas la source unique permettant d'alimenter les bases de données de

pharmacovigilance et une activité importante des agences réglementaires consiste à surveiller ces différentes sources de données afin de pouvoir prendre des décisions éclairées sur les mesures nécessaires à la réduction du risque médicamenteux.

Dans mon travail de thèse, mon objet d'étude comme nouvelle source potentielle de données sera les réseaux sociaux mais je rappelle préalablement les autres sources ainsi que les sources potentielles qui font également aujourd'hui l'objet de recherches et d'études scientifiques

4. Déclarations spontanées en France

En France, ces déclarations peuvent être émises directement par le patient depuis 2011 – à la suite de la crise du Médiator – et plus récemment, un portail a été mis en place pour déclarer des effets indésirables plus simplement depuis un site internet (<https://signalement.social-sante.gouv.fr>). Il y est tout d'abord demandé si le déclarant est un particulier, un professionnel de santé ou bien un autre professionnel (entreprise, organisme exploitant fabricant, distributeur, importateur, mandataire, ...). Pour un particulier, il faut ensuite remplir un questionnaire (Figure 2) concernant ce qui a causé l'effet indésirable : s'il s'agit d'un produit à usage médical, un autre produit, un acte médical ou examen. L'étape suivante est la déclaration de l'effet indésirable : des informations personnelles du déclarant sont demandées, puis des informations sur la personne qui présente ou a présenté l'effet indésirable, suivies d'informations concernant le médicament. Il est demandé si d'autres médicaments ou produits (des compléments alimentaires par exemple) ont été consommés. Il est ensuite demandé - Figure 3 - de décrire le ou les effets indésirables, puis leurs conséquences. Enfin, il est demandé des informations sur un professionnel de santé pouvant communiquer des informations sur l'état de santé de la personne concernée aux pharmacovigilants recevant la déclaration, puis il est possible de joindre un fichier (par exemple un compte rendu d'hospitalisation, des photos, des résultats d'analyses). Avant la déclaration, un récapitulatif est présenté, puis le numéro de référence du signalement est envoyé par mail et il est indiqué quel CRPV prendra en charge ce signalement. Il est possible pour le déclarant de télécharger une copie de sa déclaration.

Figure 2 : Portail de déclaration - Questionnaire

Figure 3 : Portail de déclaration - Description de l'EI

Dans une étude comparant les déclarations en ligne et celles traditionnelles, Abadie et al. [14] ont montré une augmentation de 45% de déclaration par des professionnels de santé en ville après l'implémentation de cet outil. Les cas étaient transmis en moyenne 26 jours plus tôt qu'avec les déclarations traditionnelles. Les notifications en lignes provenaient principalement de professionnels de santé et traitaient le plus souvent de médicaments et effets indésirables neuropsychiatriques. Aucune différence significative sur la gravité des effets indésirables n'a été retrouvée. Avec ce portail, la déclaration des effets indésirables a donc été facilitée pour les professionnels de santé eux-mêmes.

D'autres sources émergent tels que les dossiers patients électroniques dans les hôpitaux [15–17] ou les bases médico-administratives [18]. L'utilisation des dossiers patients électroniques permettrait une analyse en temps réel de données recueillies en routine à l'hôpital, offrant une approche plus proactive de la pharmacovigilance. Il est nécessaire de développer des techniques de traitement automatique du langage (TAL) adaptées afin d'utiliser au mieux ces données. Les bases médico-administratives pourraient permettre une fenêtre d'exposition étendue, un recrutement quasi-

exhaustif (en particulier dans le cadre de pathologies rares), et l'accélération du recrutement de sujets et du recueil de données. En revanche, leur volumétrie pourrait conduire à une situation de surpuissance statistique, les délais d'accès sont longs, et l'investissement nécessaire à la compréhension de ces données est conséquent.

5. Les médias sociaux

Enfin, les médias sociaux semblent également prometteurs afin de pouvoir exprimer la voix du patient. Bien que certaines agences réglementaires ont commencé à intégrer différentes sources de données en plus des déclarations spontanées, ce n'est pas encore le cas pour les médias sociaux qui ont été explorés comme source potentielle [19–26]. Si l'analyse des médias sociaux ne peut être considérée comme un système de détection de signaux prioritaire, l'exploration de cette source de données est importante, les données y étant différentes de celles retrouvées dans les bases traditionnelles.

Dans un travail préalable à cette thèse, j'ai contribué à un travail de recherche visant à étudier en terme informationnel, l'intérêt de cette source d'information [26]. Nous avons comparé l'informativité de messages issus de forums de discussion relatifs à des effets indésirables liés à 6 médicaments aux cas correspondants dans la BNPV. Les cas de la BNPV étaient plus informatifs concernant les caractéristiques du patient (âge, sexe), du traitement (posologie, durée, délai d'arrivée de l'EI) et de l'évolution de l'effet indésirable. Cependant, l'indication du traitement était davantage renseignée dans les forums. Les cas de la BNPV étaient plus souvent graves que dans les forums (46% contre 4%), mais les forums contenaient davantage d'effets indésirables non attendus que dans la BNPV (24% contre 17%) et la distribution des System Organ Class (SOCs) de MedDRA (le plus haut niveau de la hiérarchie MedDRA - Dictionnaire Médical pour les Activités Réglementées - permettant de classer les effets indésirables) étaient différentes entre les deux sources de données.

Ainsi, le potentiel des données issues des réseaux sociaux, fréquentés et nourris par des millions d'utilisateurs chaque jour, dans le cadre de la pharmacovigilance a été confirmé mais pas l'application des méthodes statistiques de détection du signal. Cependant, la performance des méthodes statistiques est dépendante de la base de données sur laquelle elles sont appliquées [27]. Une adaptation pourrait être nécessaire pour optimiser leurs performances dans le cas des réseaux sociaux. De plus, un usage efficace des données issues des médias sociaux par les agences réglementaires nécessite une interface adaptée en terme d'usabilité permettant de réaliser des analyses et de visualiser les résultats facilement.

6. Hypothèse, objectif et plan de thèse

Notre hypothèse est que des analyses statistiques peuvent être adaptées aux données issues des médias sociaux pour une analyse régulière et opérationnelle ; l'objectif principal est de définir un cadre pour l'application des méthodes de détection du signal aux données issues des médias sociaux dans un environnement opérationnel. Pour répondre à cet objectif, deux objectifs secondaires sont visés :

1. Démontrer l'utilité de ces données pour d'autres types d'analyses comme celle des thèmes abordés par les utilisateurs de réseaux sociaux lorsqu'ils traitent d'un certain sujet.
2. Implémenter et utiliser une interface utilisateur, complétant une chaîne de traitement permettant d'appliquer les algorithmes de détection du signal, ainsi que l'étude des thèmes abordés par les utilisateurs de manière quotidienne.

Pour cela, nous allons tout d'abord voir l'état de l'art concernant les méthodes permettant de repérer et normaliser l'information présente dans les messages, les algorithmes de détection du signal et leur application à ce type de données, les études de thèmes abordés par les utilisateurs de médias sociaux pour la pharmacovigilance et les chaînes de traitement existantes pour l'utilisation en pharmacovigilance de ces données. Ensuite, nous verrons le matériel à disposition pour réaliser cette thèse, c'est-à-dire le choix des forums et les modules d'extraction des messages et de leur annotation, ainsi que les méthodes utilisées pour répondre aux objectifs : la création du module d'analyse avec les cas d'usages pour la détection du signal puis celui de l'analyse des thèmes abordés par les utilisateurs, la création du module de l'interface avec sa conception et les choix technologiques pour son implémentation. Enfin, je vous présenterai les résultats obtenus pour le module d'analyse (détection du signal et analyse des thèmes abordés) ainsi que pour l'interface, avant de voir les discussions de ces différents résultats ainsi qu'une discussion générale ainsi qu'une conclusion.

Chapitre 2

Etat de l'art

Les données sollicitées dans l'utilisation de messages issus de forums de discussion en santé pour la pharmacovigilance sont disponibles sous forme de champs de caractères libres. Afin de réaliser notre objectif, la première étape est donc d'y repérer l'information d'intérêt et de la normaliser. Afin d'y parvenir, des méthodes de traitement automatique du langage doivent être utilisées et la normalisation doit être faite selon une terminologie, celle utilisée en pharmacovigilance institutionnelle pour désigner les effets indésirables, ainsi qu'une terminologie pour désigner les médicaments. La deuxième étape est de prioriser l'information et générer des signaux potentiels, des méthodes de détection du signal peuvent être appliquées pour cela. Dans la première partie de cet état de l'art, nous présenterons les terminologies utilisées pour normaliser les effets indésirables et les médicaments ainsi que les ressources existantes pour normaliser le langage patient. Ensuite, nous verrons quelles méthodes de détection du signal sont utilisées en pharmacovigilance avant de décrire l'utilisation des données issues des médias sociaux pour la pharmacovigilance : les recommandations, les problèmes éthiques, les réglementations ainsi que les résultats d'études ayant évalué leur utilisation en pharmacovigilance avec un focus sur la détection du signal.

Afin de répondre à notre premier objectif secondaire, nous verrons différents résultats déjà obtenus sur ce type de données concernant l'étude des thèmes abordés par les utilisateurs de médias sociaux pour la pharmacovigilance.

Enfin, pour répondre à notre deuxième objectif secondaire, nous présenterons différentes chaînes de traitement existantes pour l'utilisation en pharmacovigilance de données issues de médias sociaux. Nous établirons plusieurs critères sur lesquels nous les évaluerons.

1. Terminologies en pharmacovigilance

1.1 MedDRA

MedDRA, Dictionnaire Médical pour les Activités Règlementaires¹ est un dictionnaire de terminologie médicale international des effets indésirables, cliniquement validé et utilisé par les autorités réglementaires et l'industrie biopharmaceutique et organisé selon une hiérarchie à cinq niveaux : le

¹ Medical Dictionary for Regulatory Activities, <https://www.meddra.org/>, créé par le Conseil international d'harmonisation des exigences techniques pour l'enregistrement des médicaments à usage humain (ICH, pour International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use),

niveau le plus élevé est la classification par discipline médicale (SOC), divisée en groupe de termes de haut niveau (HLGT), termes de haut niveau (HLT), termes préférentiels (PT) puis les termes de plus bas niveau (LLT). [28]

La figure 4 représente la classification MedDRA du PT « Thyroïdite auto-immune ». On y observe qu'un PT peut être lié à plusieurs SOC. Chaque PT doit être lié à au moins une SOC et un PT ne peut être lié à une SOC que par un seul chemin HLT=>HLGT=>SOC. De plus, chaque PT a une SOC primaire qui détermine sous quelle SOC le terme est comptabilisé dans les analyses de données cumulées (chemin préféré).

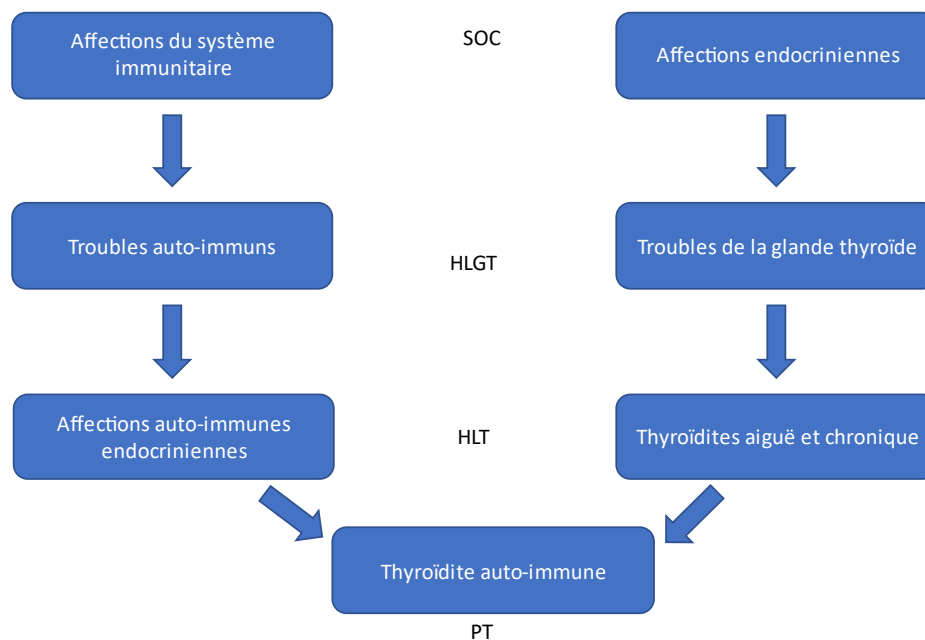


Figure 4 : Exemple de classification MedDRA

1.2 Classification ATC

Le Système de classification anatomique, thérapeutique et chimique (ATC), contrôlé par le Collaborating Centre for Drug Statistics Methodology de l'OMS (<https://www.who.int/tools/atc-ddd-toolkit/atc-classification>), permet de classer les médicaments.

La classification ATC est constituée de cinq niveaux correspondant aux organes (ou systèmes d'organes) cibles, et aux propriétés thérapeutiques, pharmacologiques et chimiques des différents médicaments. Le premier niveau (première lettre) définit le groupe anatomique parmi 14 différents. Le deuxième niveau (deux premiers chiffres) donne le sous-groupe pharmacologique ou thérapeutique principal. Les troisième et quatrième niveaux (deuxième et troisième lettres) correspondent à des sous-groupes chimiques, pharmacologiques ou thérapeutiques. Le cinquième et dernier niveau (deux derniers chiffres) indique la substance chimique. [29]

Nous avons choisi de l'utiliser car il s'agit de la classification utilisée par l'OMS dans son outil de veille de pharmacovigilance Vigilyze (<https://who-umc.org/covid-19/quantitative-vaccine-safety-surveillance-with-vigilyze/>). Dans cette classification, un médicament a un code ATC de 7 caractères correspondant à 5 niveaux de la classification. Le tableau 1 décrit la classification ATC du méthylphénidate.

Tableau 1: Classification ATC du méthylphénidate

Code ATC	Classe ATC
N	Système nerveux
N06	Psychoanaleptiques
N06B	Psychostimulants, agents utilisés en cas de trouble déficitaire de l'attention/hyperactivité (TDAH) et nootropiques
N06BA	Sympathicomimétiques centraux
N06BA04	Méthylphénidate

1.3 UMLS

Le méta-thésaurus UMLS (Unified Medical Language System, <https://www.nlm.nih.gov/research/umls/index.html>), créé en 1986, rassemble les concepts de nombreuses terminologies de santé. Il permet de relier différentes terminologies et leurs structures pour ainsi permettre des traductions entre elles. Chaque concept a un identifiant unique (Concept Unique Identifier, CUI), plusieurs mots différents peuvent avoir le même sens et sont donc codés avec le même identifiant.

1.4 Langage patient : de possibles ressources ?

L'objectif de détecter de l'information dans des champs de caractères libres rédigés par des patients nous a amené à rechercher des ressources potentielles pour « traduire » ce langage patient, c'est-à-dire de lier le langage patient aux ressources terminologiques présentées précédemment.

Huit études ont été identifiées [30–37], deux d'entre elles concernaient le français et les autres l'anglais.

MuEVO, un vocabulaire multi-expertise (patient/médecin) dédié au cancer du sein [30]

Les auteurs présentent une terminologie en français reliant le vocabulaire des professionnels de la santé à celui des patients dans le cadre du cancer du sein. Les auteurs présentent également une

méthode d'alignement de cette terminologie à des terminologies de référence en santé : Le Medical Subject Headings (MeSH, <https://www.hetop.eu/hetop/rep/fr/MESH/>), SNOMED (<https://www.snomed.org/>) et MedDRA.

Réconciliation du vocabulaire patient/médecin dans une ressource structurée [31]

Le but de cette étude qui fait suite à la précédente (menée par la même équipe de recherche) est d'utiliser des textes écrits par des patients pour construire un vocabulaire patient dans le cadre du cancer du sein. En combinant des méthodes basées sur la linguistique ainsi que des approches statistiques pour collecter différentes sortes d'expressions patients de messages issus du forum « cancerdusein.org » et de les relier au vocabulaire des professionnels de santé. Les auteurs ont construit un vocabulaire composé de 192 relations validées et formalisées dans une ontologie. Ces relations pouvaient par exemple être des « erreurs d'orthographe » comme le terme « hopital » pour « hôpital », des « abréviations » comme « psy » pour « psychologue », ou encore des associations comme « morphine » pour « douleur ».

Cas d'usage d'une terminologie patient dans un système de dossier médical électronique [32]

Les auteurs ont relié 340 expressions patients de diagnostics issus de dossier patients électroniques et d'un portail patient à des termes de l'UMLS en classant les paires en synonymes, presque synonymes ou termes associés. Par exemple, le terme patient « underactive thyroid » était relié à « hypothyroïdie ».

Terminologie médicale dans la communication entre patients sur internet [33]

Les auteurs ont étudié la terminologie médicale utilisée par les patients sur les forums de discussion. Ils ont codé, à partir de messages issus de forums, deux jeux de données provenant de deux forums patients dédiés aux problèmes de thyroïde.

Caractéristiques d'une terminologie consommateur pour la recherche d'informations en santé [34]

Afin de construire une terminologie d'expressions du langage patient, Zeng *et al.* ont pris le parti d'utiliser les requêtes rentrées par les utilisateurs dans les moteurs de recherche des sites et forums médicaux. Ils ont constitué un premier corpus en utilisant la partie « Find-a-Doctor » du site Web du *Brigham and Women's Hospital*, et un second corpus à partir des requêtes saisies sur le site MEDLINEplus le 1^{er} août 2001. Les auteurs ont étudié si les termes des ensembles correspondent exactement à un terme du metathesaurus UMLS. Toutes les expressions n'ayant pas de correspondance exacte dans UMLS furent mises en correspondance manuellement par un interne. Ce

dernier a examiné 668 termes sans correspondance UMLS. Les résultats de la mise en correspondance sont les suivants : 40% des termes examinés correspondaient exactement à des concepts UMLS, 48% correspondaient partiellement et 12% ne purent être mis en correspondance.

Concepts en santé de consommateurs ne pouvant être reliés à l'UMLS [35]

Ce travail académique avait deux objectifs : (1) identifier et caractériser les expressions en langage patient sans correspondance trouvée avec UMLS et (2) décrire la procédure de création de nouveaux concepts dans le processus de construction d'un lexique d'expressions en langage patient. Le matériel de base de cette étude consiste en deux ensembles : 12 millions de requêtes MedlinePlus (Set A) et des textes libres extraits manuellement de 25 forums de santé (Set B).

Pour chaque terme trouvé dans les deux ensembles de données et n'étant pas un synonyme d'un terme UMLS, trois relecteurs ont eu pour tâche de trouver le terme UMLS le plus proche et de qualifier la relation entre le terme en langage patient et le terme UMLS (plus précis, plus généraliste, vaguement lié, ou autre relation). Par la suite, les relecteurs durent déterminer, pour chaque terme en langage patient, son type sémantique (celui du terme UMLS en relation, s'il y en a un), le domaine médical, si le terme n'existe que dans le langage patient, sans équivalent dans le vocabulaire médical et si le terme peut être dérivé de termes existants par conjonction ou modification.

Sur les requêtes sur MedlinePlus, 31 des 44 nouveaux concepts ont pu être représentés comme des concepts plus spécifiques de concepts UMLS existants.

Parmi les 20 concepts tirés des textes libres des forums de santé, 12 avaient une « autre relation » avec des concepts UMLS existants.

Le nombre de concepts sans équivalence dans le vocabulaire médical était petit (6 dans les requêtes sur MedlinePlus et 11 sur les textes libres provenant de forums de santé), leur étude a été mutualisée. Sur les 17 concepts sans équivalence, 6 concepts peuvent être assignés à la santé sexuelle ou la gynécologie et 3 concepts à la beauté. Les autres domaines sont représentés par un unique concept : bien-être, oncologie, pathologie et génétique. Enfin, les 4 derniers concepts n'ont pas pu être assignés à un domaine.

PatientsLikeMe: Vocabulaire de consommateur en santé en tant que Folksonomie [36]

A l'image du Set B de Keselman et al. [35], les auteurs de cette étude se sont reposés sur les réseaux sociaux, et plus précisément PatientsLikeMe (<https://www.patientslikeme.com/>), pour trouver des termes en langage patient. C'est ainsi que 376 termes correspondant à un symptôme ont été soumis

par les membres de PatientsLikeMe. Par la suite, deux codeurs ont eu pour tâche d'analyser indépendamment ces termes et ont atteint 100% d'accord inter-évaluateur.

Parmi les 376 termes soumis par les patients, 90 termes (24%) correspondaient exactement à un terme UMLS et 71 concepts (19%) étaient des synonymes de termes UMLS existants. Parmi les 90 correspondances exactes, 93% venaient de SNOMED-CT, 88% de Read Codes et 86% de MedDRA.

209 termes (56%) n'ont pas trouvé de correspondance dans le metathesaurus UMLS pour différentes causes : fragment de terme UMLS (90 termes), deux concepts ou plus dans le terme (55), faute d'orthographe (11), ayant une composante temporelle (9), trop vague (6), argot (3), défaut de correspondance (3) et autre (70). Enfin, 6 concepts (3%) ont été notés comme ayant une relation « autre » : 1 comme doublon, 2 dont le sens a été impossible à déterminer et 3 considérés comme un défaut de correspondance.

Mise à jour d'un vocabulaire consommateur en santé via l'utilisation des medias sociaux [37]

Dans cette étude, les données du site PatientsLikeMe servent de base, comme dans la publication de Smith et Wicks [36]. A la différence de la publication présentée ci-dessus, les auteurs ne partent pas directement des données fournies par les utilisateurs, données qui sont seulement accessibles par les administrateurs du site PatientsLikeMe, mais des termes présents dans les pages publiques du site pour augmenter leur vocabulaire d'expressions patient (CHV pour Consumer Health Vocabulary). Une collecte automatique des pages Web, un filtrage des termes à l'aide d'outils de traitement de la langue et une interface de relecture manuelle ont été développés dans le cadre de cette étude.

La collecte automatique comprend trois sous-étapes. La première sous-étape est la collecte des pages en elle-même. La deuxième est le traitement du texte de la page pour trouver les composantes du langage, noms et entités nommées à l'aide de HITex [38], un outil de traitement automatique de la langue (TAL). La troisième sous-étape consiste à extraire tous les n-grammes (avec n de 1 à 7), même s'ils se chevauchent, c'est-à-dire qu'un mot peut être inclus dans plus d'un n-gramme. Les n-grammes contenant un symbole ou un nombre sont exclus, ainsi que ceux étant composés seulement de mots vides (par exemple, « le », « la », « et »).

La deuxième étape fait encore plus appel aux outils de TAL pour filtrer les n-grammes retournés par l'étape précédente. Dans un premier temps, les n-grammes sont comparés avec les termes du metathesaurus UMLS et aux termes déjà présents dans CHV. Les n-grammes déjà présents dans UMLS ou CHV sont exclus. Ensuite, les n-grammes sont recherchés sur une base de données de 70 000 enregistrements médicaux, ceux n'ayant pas été trouvés dans la base de données étant exclus. Enfin, un score est calculé pour chaque n-gramme : *C-value* pour les n-grammes en sortie de la première

étape et *termhood* pour les n-grammes n'étant pas trouvés dans la base de données des enregistrements médicaux.

La troisième et dernière étape consiste en la relecture manuelle des termes n'ayant pas été exclus. Les auteurs ont facilité cette étape grâce à un site Web interactif développé pour l'occasion. Le site Web a permis plus que la revue des n-grammes, à savoir la proposition de nouvelles adresses pour la collecte automatique, la revue des commentaires sur les n-grammes, le vote en faveur ou en défaveur de l'inclusion d'un n-gramme à CHV, la consultation des votes et des fréquences des n-grammes, et la consultation de trois exemples d'un n-gramme dans le contexte de la page dans laquelle il a été trouvé. La première étape a eu comme résultat 88 994 n-grammes dont ont été extraits 651 termes. Le filtrage par rapport à UMLS et CHV a permis d'écarter 1045 n-grammes, laissant 87 949 n-grammes (Les n-grammes déjà présents dans UMLS ou CHV sont exclus). Seulement 923 n-grammes (réduction de 99%) passèrent le filtre excluant les n-grammes non retrouvés dans la base de données des 70 000 enregistrements médicaux, ce qui représente 215 termes (réduction de 67%). En jouant sur les seuils pour les scores et en sélectionnant le meilleur, 662 et 170 termes candidats à la relecture ont été identifiés respectivement parmi ceux filtrés par rapport à UMLS et CHV et ceux présents dans la base de données des enregistrements médicaux.

Conclusion

Aucune terminologie satisfaisante en français n'existe à notre connaissance pour traduire le langage patient dans le cadre de la pharmacovigilance : lorsque des ressources existent en français, elles sont souvent spécifiques à un thème comme le cancer du sein alors que les effets indésirables liés aux médicaments sont très variés.

Une autre possibilité pour arriver à notre fin est le traitement automatique du langage. Le texte est annoté et l'information retrouvée sans se baser uniquement sur des ressources du langage patient. Le sous-chapitre suivant présentera notre solution pour l'annotation et la normalisation des couples potentiels.

[2 Identification des effets indésirables liés aux médicaments dans les messages](#)

Afin d'annoter les EIs dans le projet Vigi4Med, Morlane-Hondere et al. [39] détectent tout d'abord les entités correspondant à un médicament ou à un EI en utilisant des champs aléatoires conditionnels [40,41]. Cette méthode permet une prédiction prenant en compte le contexte (par exemple, les mots avant et après celui à annoter). Puis, une machine à vecteurs de support prédit la relation causale entre les deux entités identifiées afin de savoir si l'EI est bien dû au médicament. Finalement, cet algorithme d'apprentissage n'est pas basé sur une simple correspondance à un dictionnaire ou une terminologie

de langage patient : l'algorithme apprend à identifier un mot comme étant un effet indésirable, un médicament ou aucun des deux. Ensuite, alors, le verbatim identifié peut ou non être normalisé.

Des méthodes plus récentes de TAL sont maintenant utilisées[42–45], s'appuyant par exemple sur des réseaux de neurones convolutifs [43], et des représentations d'encodeur bidirectionnel à partir de Transformers (BERT) [45]. Hussain et al. ont proposé un réseau de neurones multitâches basé sur BERT avec optimisation des hyperparamètres permettant de classer une phrase et faire de la reconnaissance d'entités nommées [44]. Magge et al. ont proposé une chaîne de traitement composée de 3 réseaux de neurones BERT permettant la classification de phrases, l'extraction d'entités nommées, et finalement la normalisation en termes MedDRA [42].

3 Algorithmes de détection du signal en pharmacovigilance

Plusieurs algorithmes de détection automatisée de signaux ont été mis au point pour exploiter les bases de notifications spontanées afin de mettre en évidence des couples (médicament, EI) suspects. Les signaux générés sont ensuite étudiés de manière plus approfondie par des experts en pharmacovigilance pour être confirmés. Il est donc important de générer un nombre raisonnable de signaux avec le moins de faux positifs possible : c'est-à-dire avec le moins de couples médicament-EI considérés à tort comme signal par l'algorithme.

3.1 Méthodes basées sur la disproportionnalité

Les méthodes de détection utilisées en routine par les agences de santé publique sont appelées méthodes de disproportionnalité reposant sur trois étapes pour chaque couple étudié :

1. Déterminer le nombre de notifications impliquant ce couple.
2. Construire une mesure quantifiant la différence entre ce nombre et celui attendu sous l'hypothèse d'indépendance entre le médicament et l'EI.
3. Emettre un signal si cette différence est significative (parmi d'autres critères possibles).

Deux familles de méthodes de disproportionnalité se distinguent : les méthodes fréquentistes telles que le Reporting odds ratio (ROR) [2] et le Proportional Reporting Ratio (PRR) [8], et les méthodes bayésiennes comme le (Multi-Item) Gamma Poisson Shrinker (M)GPS [6,7] et le Bayesian Confidence Propagation Neural Network (BCPNN) [46,47].

a) PRR

Le PRR s'apparente à un risque relatif utilisé en épidémiologie pour lequel l'effectif des exposés (les patients consommateurs du médicament) n'est pas connu. Il est « remplacé » par le total des cas de pharmacovigilance contenus dans la base.

Il s'agit de comparer la survenue d'un effet indésirable donné pour un médicament donné à la survenue de cet effet indésirable pour tous les autres médicaments présents dans la base de données.

Il quantifie la force de l'association d'un médicament à un effet indésirable par rapport à l'association de cet effet indésirable pour tous les autres médicaments.

b) ROR

Le ROR s'apparente à l'odds ratio avec, de façon similaire au PRR, le total des cas de pharmacovigilance contenus dans la base en lieu et place du nombre d'exposés.

c) Formules PRR et ROR

Le PRR et le ROR se calculent directement depuis la table de contingence T de taille $I \times J$ de tous les médicaments et de tous les événements indésirables présents au moins une fois dans une notification spontanée. Lorsque plusieurs médicaments et/ou plusieurs événements indésirables sont présents dans une même notification, chacun des couples médicament-événement indésirable impliqués dans celle-ci est inclus dans la table de contingence.

Les statistiques d'intérêt des méthodes présentées dans la suite de ce document sont des mesures de disproportionnalité calculées pour chacun des couples (i, j) par agrégation de la table T en une table de contingence 2×2 (Tableau 2).

Tableau 2 : Table de contingence pour le couple Médicament i - El j

	El j	Tous les autres Els	Total
Médicament i	n_{ij}	$n_{i\bar{j}}$	$n_{i.}$
Tous les autres médicaments	$n_{\bar{i}j}$	$n_{\bar{i}\bar{j}}$	$n_{\bar{i}.}$
Total	$n_{.j}$	$n_{.\bar{j}}$	n

$$PRR_{ij} = \frac{n_{ij}/n_{i.}}{n_{\bar{i}j}/n_{\bar{i}.}}$$

$$ROR_{ij} = \frac{n_{ij} n_{\bar{i}\bar{j}}}{n_{\bar{i}j} n_{i\bar{j}}}$$

d) BCPNN

On s'intéresse à un réseau bayésien à propagation avant à deux couches dans lequel les nœuds en entrée sont les médicaments et les nœuds en sortie sont les événements indésirables. Chacun de ces nœuds représente une variable binaire et peut être ouvert ou éteint. La sortie d'un nœud est déterminée par les valeurs prises par tous les nœuds en entrée reliés à lui et les poids entre chacun de ces nœuds en entrée et le nœud de sortie (Figure 5).

Le poids entre deux nœuds, l'Information Component (IC), est une mesure de la dépendance entre l'entrée et la sortie. La règle de décision se fondera sur la mesure de cette valeur.

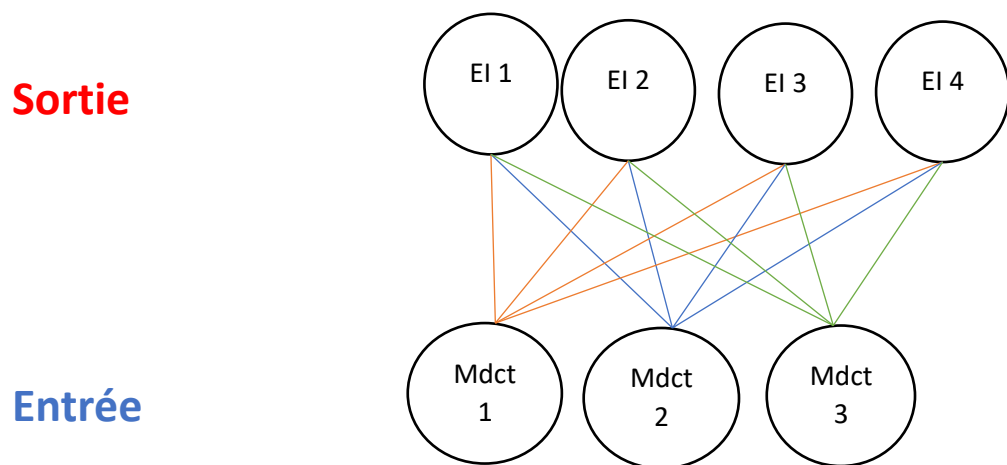


Figure 5 : Exemple de réseau neuro-bayésien

On note :

- $P(A) = P(EI = j)$, la probabilité que l'événement indésirable j soit présent sur la notification
- $P(D) = P(Mdct = i)$, la probabilité que le médicament i soit présent sur la notification
- $P(A, D)$, la probabilité que le médicament i et l'événement indésirable j soient tous les deux présents sur la notification
- $P(A|D)$, la probabilité que l'événement indésirable j soit présent sur la notification sachant que le médicament i l'est

Si $P(A|D) > P(A)$, cela signifie que la présence du médicament i sur une notification augmente la probabilité que l'événement indésirable j soit également présent sur la notification. Alors, le couple Médicament i – Événement indésirable j est plus fréquent qu'attendu dans la base de données.

D'après le théorème de Bayes [48], on peut écrire :

$$P(A|D) = \frac{P(A,D)}{P(D)} = P(A) \frac{P(A,D)}{P(A)P(D)}$$

$P(A)$ et $P(A|D)$ sont liés par le facteur $\left[\frac{P(A,D)}{P(A)P(D)} \right]$ dont le logarithme de base 2 est l'Information Component pour les événements A et D, mesure de la force de l'association entre ces deux événements.

$$IC_{ij} = \log_2 \left(\frac{P(A,D)}{P(A)P(D)} \right)$$

La règle de décision est la suivante : un signal est généré pour le couple (i, j) si le quantile à 2.5% de la distribution de IC_{ij} est plus grand que 0.

Pour cette approche Bayésienne, on suppose a priori une loi de probabilité pour $P(A)$, $P(D)$ et $P(A,D)$. En ajoutant de l'information à cette estimation a priori, on obtiendra une estimation a posteriori qui lors du prochain ajout d'information, sera considérée comme estimation a priori. L'estimation sera constamment améliorée par l'ajout d'information.

Utilisé par l'OMS jusqu'en 2014, le BCPNN a été remplacé par *vigiRank* [5], un modèle prenant en compte un ensemble d'informations de la notification dont la valeur de l'Information Component (la mesure de disproportionnalité issue du BCPNN).

e) (M)GPS EBGGM

Il s'agit d'une approche bayésienne conçue pour limiter le nombre de faux positifs au travers de multiples comparaisons. Elle implémente une valeur ajustée du ratio observé/attendu et pour la version utilisée par la FDA, peut être ajustée selon l'âge et le sexe. Nous en utiliserons une version simplifiée n'ajustant pas selon l'âge ni le sexe, non présents dans nos données.

3.2 Autres méthodes

Les méthodes de disproportionnalité, considérant la forme agrégée des données, souffrent du biais de masquage qui se produit lorsqu'un EI est fortement notifié avec un certain médicament : il est alors plus difficile de détecter un signal impliquant cet EI avec un autre médicament. De plus, ces méthodes ne tiennent pas compte de la co-prescription des médicaments ni de l'évolution temporelle du nombre de notifications concernant le couple d'intérêt. D'autres méthodes permettent d'en prendre compte et sont donc complémentaires aux méthodes de disproportionnalité.

a) Régression logistique : Class Imbalanced Subsampling Lasso

Cette méthode, décrite par Ismael Ahmed et al. [49] permet de prendre en considération l'association entre les médicaments. Elle implémente des odds ratio afin de sélectionner les médicaments les plus prédictifs d'un effet indésirable donné.

On considère un effet indésirable. Considérons N cas et N_m médicaments.

On peut alors construire :

- Un vecteur de 1 et de 0 en fonction de la survenue de l'effet indésirable dans le cas, et
- Une matrice de taille $N \times N_m$ comportant autant de lignes que de cas et constituée de 1 et de 0 selon la présence ou non de chaque médicament dans le cas.

On cherche à prédire la probabilité d'un événement en fonction de plusieurs variables. Ici, on modélise la probabilité qu'un effet indésirable se produise en fonction de la prise des différents médicaments.

$$\ln \frac{P(X|ADE = 1)}{P(X|ADE = 0)} = a + b_1 \times Mdc_1 + \dots + b_i \times Mdc_i + \dots + b_{N_m} \times Mdc_{N_m}$$

Sélection des médicaments les plus prédictifs de la présence de l'effet indésirable

- Si $b_i < 0$: le médicament i diminue le risque de survenue de l'EI
- Si $b_i > 0$: la prise du médicament i augmente le risque de survenue de l'EI

Le Least Absolute Shrinkage and Selection Operator (lasso) est une méthode de pénalisation pour la régression qui permet de forcer les coefficients des variables les moins importantes à 0 (utile en présence de nombreuses variables explicatives).

Principe du Class-imbalanced subsampling lasso : cette méthode prend en compte la large dimension des données et le déséquilibre entre la présence et l'absence de l'effet indésirable.

S_1 : ensemble des n_1 messages avec l'effet indésirable

S_0 : ensemble des n_0 messages sans l'effet indésirable

Dans notre cas : $n_0 \gg n_1$.

On effectue un tirage au sort avec remise de n_1 messages dans S_1 et de R messages dans S_0 . Avec $R = \max(4n_1, 4N_m)$ en analogie avec les études cas-témoins en épidémiologie.

On fixe E , le nombre maximum de variables explicatives dans le modèle.

On applique cette méthode sur B tirages au sort.

On s'intéresse finalement à la distribution empirique du nombre moyen de fois où un médicament est sélectionné par le lasso (coefficient supérieur à 0). Les médicaments retenus comme prédicteurs seront finalement ceux dont le α % quantile de cette distribution est supérieur à 0. Avec par exemple α égale à 5, 10 ou 15.

Le quantile α % d'une distribution est le nombre x tel que α % des valeurs de la distribution soient inférieures ou égales à x .

b) Change Point analysis

Il s'agit d'une méthode permettant de détecter des « points de ruptures » dans une évolution temporelle basée sur un changement brusque dans la moyenne (algorithme CUSUM) [50,51].

On s'intéresse à un couple médicament – effet indésirable. Soient X_1, \dots, X_n les nombres de fois où l'on retrouve ce couple dans la base à n temps différents (par exemple X_1 pourrait représenter le nombre d'occurrences du couple lors de la première année ou du premier mois).

On va alors calculer S_1, \dots, S_n de la manière suivante :

1. $\bar{X} = \frac{X_1 + \dots + X_n}{n}$
2. $S_0 = 0$
3. $S_i = S_{i-1} + (X_i - \bar{X})$ pour $i = 1, \dots, n$.

Et :

$$S_{diff} = S_{max} - S_{min}$$

$$S_{max} = \max_{i=0, \dots, n} S_i$$

$$S_{min} = \min_{i=0, \dots, n} S_i$$

On tire n échantillons de $n - 1$ valeurs de X que l'on ordonne au hasard (algorithme Jackknife : enlever une des observations par échantillon). Ces échantillons imiteraient le CUSUM si aucun changement n'était observé (comme n est grand).

On effectue les mêmes calculs pour chaque échantillon.

Une autre option –plus fiable lorsque n est petit - est d'effectuer un tirage de 1 000 échantillons de taille n avec remise. Le code R devra être adapté.

On appelle N le nombre total d'échantillons bootstrap et Y le nombre d'échantillons pour lesquels la valeur de S_{diff} est strictement inférieure à celle observée dans les données d'origine.

On peut alors calculer la confiance (en %) que l'on a dans le fait qu'un changement a bien eu lieu :

$$\text{Niveau de confiance} = \frac{100 \times Y}{N}$$

Après avoir identifié la présence d'un changement, il faut identifier à quelle période ce changement a eu lieu. On calcule :

$$|S_m| = \max_{i=0, \dots, n} |S_i|$$

Le point m est donc le dernier point avant le changement et le point $m + 1$ est le premier point après le changement.

Après ce changement détecté, les données sont divisées en 2 échantillons : du point 1 à m et du point $m + 1$ à n . L'algorithme peut alors être relancé jusqu'à ce que plus aucun changement ne soit détecté.

Finalement, on obtient en sortie de l'algorithme les dates auxquelles des augmentations ou diminutions significatives de la statistique d'intérêt sont détectées.

c) CPA appliqué sur les valeurs du PRR

Les méthodes vues précédemment peuvent être combinées afin de générer des signaux en prenant en compte différents critères. Par exemple, Trinh et al. [3] ont étudié la combinaison du PRR avec le CPA. Le CPA était alors appliqué sur la valeur mensuelle de la borne inférieure de l'intervalle de confiance du PRR. Un signal potentiel était généré si (1) le PRR était significativement supérieur à 1 avec au moins 5 cas et (2) le CPA détecte 2 augmentations successive de la statistique d'intérêt. En testant cette méthode dans cette étude préliminaire, elle semblait bénéfique afin de détecter plus rapidement des signaux et générer moins de faux positifs.

3.3 Conclusion sur les méthodes de détection du signal

Il n'existe pas de gold standard concernant les méthodes de détection du signal en pharmacovigilance. Nous pensons qu'il est donc utile de pouvoir appliquer des méthodes différentes et complémentaires et ainsi pouvoir comparer leurs résultats. C'est pourquoi toutes ces méthodes ont été sélectionnées pour être intégrées à ce travail.

4 La pharmacovigilance dans les médias sociaux

4.1 Recommandations

Le guide des bonnes pratiques en pharmacovigilance publié par l'EMA en 2017² contient des avis sur l'utilisation des données issues d'internet et des médias digitaux. L'EMA ne recommande pas aux détenteurs des droits des médicaments sur le marché de surveiller cette source de données mais recommande la surveillance des informations saisies par les usagers sur un site internet qu'ils sponsorisent ou dont ils sont responsables. Si un cas de pharmacovigilance potentiel est détecté, ils doivent transmettre l'information aux autorités compétentes.

Les recommandations du projet européen WEB-RADR (Recognizing Adverse Drug Reactions, 2014-2021) auquel l'OMS a participé sont de ne pas utiliser les médias sociaux comme une source de notifications spontanées [24].

4.2 Ethique

L'utilisation de ce type de données entraîne des questionnements d'ordre éthique. En effet, un commentaire mis manifestement à la disposition du public implique-t-il forcément un accord pour son utilisation dans une étude ? Il ne s'agit en tout cas pas d'un consentement éclairé [52].

Une autre problématique vient de l'utilisation par un tiers des données extraites pouvant mener à l'identification des auteurs : des compagnies d'assurances pourraient utiliser ces données afin de refuser certains contrats à des clients demandeurs. Les données permettant l'identification doivent être sécurisées et les données accessibles ne pas permettre l'identification [52].

Les investigateurs détectant un effet indésirable inattendu auraient intérêt à contacter l'auteur pour obtenir des informations complémentaires et assurer un suivi du cas. Mais cela n'est pas forcément possible. Le signal, en principe, devrait alors être évalué, ce qui peut être une perte de temps si l'information dans le message est erronée [52].

4.3 Réglementations

Dans l'Union-Européenne, le Règlement sur la Protection des Données (RGPD, <https://www.cnil.fr/fr/reglement-europeen-protection-donnees>), adopté par le Parlement Européen en avril 2016 et appliqué dans tous les états membres depuis le 25 mai 2018, renforce la protection des données individuelles en mettant en place le droit d'être informé sur l'utilisation des données personnelles. Comme il est impossible d'informer chaque utilisateur, le RGPD introduit deux conditions légales dans lesquelles le consentement éclairé n'est pas obligatoire et le texte peut être interprété

² [Guideline on good pharmacovigilance practices \(GVP\) - Module VI – Collection, management and submission of reports of suspected adverse reactions to medicinal products \(Rev 2\) \(europa.eu\)](https://www.cnil.fr/fr/reglement-europeen-protection-donnees)

comme autorisant l'utilisation des données extraites des médias sociaux pour la pharmacovigilance (Article 9) : « (e) le traitement porte sur des données à caractère personnel qui sont manifestement rendues publiques par la personne concernée; (...) (i) le traitement est nécessaire pour des motifs d'intérêt public dans le domaine de la santé publique (...) aux fins de garantir des normes élevées de qualité et de sécurité des soins de santé et des médicaments ou des dispositifs médicaux ». De plus, le RGPD contraint à « un traitement ultérieur ne permettant pas ou plus l'identification des personnes concernées » et ce à « chaque fois que ces finalités peuvent être atteintes ». Il est donc nécessaire de procéder à une dé-identification des données.

4.4 Avantages et inconvénients de la pharmacovigilance dans les médias sociaux

a) Inconvénients de la détection du signal

Tout d'abord, nous allons voir les résultats d'études qui ne sont pas en faveur de l'utilisation des données issues des médias sociaux pour la pharmacovigilance.

Dans une étude de Pierce et al. [53] un seul effet indésirable a été identifié dans les médias sociaux avant sa déclaration dans la base de pharmacovigilance de la FDA, la FDA Adverse Event Reporting System (FAERS) : la corrélation entre la crème solaire Banana Boat et des brûlures a été observée dans les médias sociaux le 2 juin 2012, soit 17 jours avant son premier rapport dans FAERS. De plus, dans cette même étude, sur 935 246 messages, seuls 6 décrivaient un cas de pharmacovigilance possible, probable ou certain.

Seulement 10 des 2 537 messages relatifs aux effets indésirables cardiovasculaires avec le rosiglitazone dans l'étude de Coloma et al. [54] correspondaient à des expériences personnelles.

Une étude visait à classer les médicaments entre ceux auxquels il fallait ajouter un avertissement directement sur la boîte de ceux pour lesquels cela n'était pas nécessaire [55] à l'aide d'un premier classifieur des messages dans lesquels on retrouve un ou des effets indésirables des autres, puis d'un second, triant les médicaments en différentes classes selon le nombre de messages retrouvés avec effet indésirable les concernant. Le second classifieur avait un F-score assez modeste (0.6). Mais avec cette méthode, deux médicaments ont été mal classés car le nombre de messages les concernant était faible.

Dans une étude de Duh et al. [56], les revues concernant l'atorvastatine n'ont pas permis de détecter un cas avant les notifications de FAERS.

En 2018, Caster et al. [24] ont montré que les algorithmes de détection de signaux sont peu performants sur les données issues de médias sociaux (messages provenant de Twitter, Facebook, et de 407 forums de patients) : l'aire sous la courbe ROC (de l'anglais *receiver operating characteristic*,

pour « caractéristique de fonctionnement du récepteur ») était entre 0,47 et 0,53 dans les médias sociaux vs. entre 0,64 et 0,69 pour VigiBase, la base de pharmacovigilance de l'OMS. De plus, sur Twitter et Facebook, seulement 16% et 6% des signaux de contrôles ont été détectés avant leur indexation dans les jeux de référence WEB-RADR et Harpaz [57], un jeu de référence temporel de couples médicaments-Els.

Bousquet et al. [25] ont apporté un commentaire à l'étude Web-RADR. Même s'ils approuvent le fait qu'aujourd'hui les médias sociaux ne doivent pas être utilisés en priorité dans l'analyse en pharmacovigilance, on doit garder à l'esprit qu'il s'agit de données non comparables aux bases traditionnelles aussi bien au niveau des médicaments que des profils des effets indésirables. En effet, les événements indésirables non graves y sont significativement plus présents (les médecins rapportent moins souvent les effets non graves). Or, des effets indésirables non graves peuvent avoir de graves conséquences comme la mauvaise observance du traitement dans des maladies comme la dépression, pouvant augmenter le risque de tentatives de suicide.

b) Avantages de la détection du signal

Cependant, d'autres études ont démontré le potentiel de l'utilisation de telles données en pharmacovigilance.

En effet, dans une revue de la littérature [53], 9 (64.2%) des 14 études ont montré que l'analyse des médias sociaux permettait la détection d'événement amenant des avertissements, un changement sur le résumé des caractéristiques du produit (RCP) ou un retrait du marché de 3 mois à 9 ans en avance par rapport aux bases traditionnelles. En effet, un message mentionnant une valvulopathie cardiaque avec le Mediator chez une femme a été posté sur un forum 7 mois avant le retrait du marché [58].

Chee et al. [19] ont analysé 4 médicaments (la cerivastatin, la trovafloxacin, l'hydromorphone et le rofécoxib) qui ont été retirés du marché. Des messages concernant des effets indésirables liés à ces médicaments ont été détectés sur Yahoo Groups avant leur retrait du marché. De plus, des messages à propos de la sibutramine - qui n'était pas sur la liste des médicaments surveillés au moment de l'étude mais qui a été retiré du marché car il augmentait le risque d'arrêt cardiaque et d'accident vasculaire cérébrale - ont été observés plus d'un an avant ce retrait du marché.

Dans une étude de Duh et al. [56], des mentions d'arrêts cardiaques en lien avec la sibutramine ont été trouvées 11 mois avant que l'effet ne soit déclaré sur FAERS.

Dans une étude d'Adjeroh et al. [54], pour 30 sur 46 (65.2%) médicaments pour lesquels il y a eu une alerte de la FDA par rapport à certains effets indésirables, ces effets indésirables étaient observables sur les médias sociaux. De plus, ces effets indésirables pouvaient être détectés de 3 à 35 mois avant l'alerte de la FDA.

Wu et al. [59] ont rapporté que des arythmies dues au propoxyphène/acétaminophène étaient observables en nombre depuis 2006 sur les médias sociaux, soit 4 ans avant le retrait du marché de ce médicament. Des effets indésirables de 3 autres médicaments (la simvastatine, le drospirénone et l'éthinylestradiol) ont été observés 4 à 6 ans avant une action de la FDA.

Une autre étude a montré que plusieurs EIs avaient été déclarés dans les médias sociaux plus tôt que dans FAERS : pensées suicidaires avec le prozac (1 an en avance), vision trouble avec le méthylphénidate (2 an en avance), la dépression avec le prozac (5 ans en avance), les pensées suicidaires avec le fluvoxamine (5 ans en avance), des problèmes rénaux avec le simvastatin (6 ans en avance) et des diarrhées avec le lansoprazole (13 ans en avance) [60].

Feldman et al. [61] ont rapporté que l'extraction non supervisée de couples statine-déficiences cognitive dans les forums de discussion identifiait ce couple 9 ans avant le changement dans le RCP. Dans la même étude, ils ont identifié une relation significative entre le bupropion et l'agitation, 7 ans avant une action de la FDA.

AZDrugMiner, développé par Liu et Chen [62], a pu détecter à une grande fréquence les couples rosiglitazone-infarctus du myocarde et problème cardiaque (respectivement 18 et 13% des effets indésirables) depuis les forums.

Les médias sociaux sont importants pour évaluer l'utilisation des médicaments comme décrit par Bhattacharya et al. [63]. En effet, les événements indésirables rapportés par les patients ont un impact sur leur qualité de vie et donc leur adhérence au traitement. Les utilisateurs de médias sociaux sont des jeunes adultes donc plus jeunes que les patients issus des bases traditionnelles et les médicaments les plus évoqués dans les médias sociaux ne sont pas nécessairement les plus vendus ou prescrits : les médicaments surveillés en pharmacovigilance ne sont pas forcément ceux que l'on retrouve le plus dans les médias sociaux [25].

Avant cette thèse, j'ai déjà pu travailler sur l'application des algorithmes de détection du signal dans les médias sociaux. En 2018, en travaillant sur le cas d'usage du baclofène [64], nous avons montré que la plupart des effets indésirables présents dans le RCP étaient retrouvés comme signaux par l'application des méthodes de détection du signal. Nous avons également montré qu'il était possible de détecter des effets indésirables inattendus.

Dans le cadre d'une étude sur le méthylphénidate [65] à laquelle j'ai participé, l'application du PRR permettait de retrouver 66.7% des signaux identifiés manuellement confirmant ainsi l'intérêt de l'application de telles méthodes. Parmi ces signaux, 38.5% étaient présents dans le RCP et 88.5% dans VigiBase.

A travers le cas d'usage d'une étude que j'avais menée sur le Protelos [66], les algorithmes de détection du signal avaient permis d'identifier comme étant des signaux 93.6% des 47 cas identifiés manuellement. Cependant, trois effets indésirables graves identifiés comme étant des signaux par les algorithmes étaient en fait des faux positifs : cela est dû au fait que la causalité est difficilement détectable par les algorithmes de traitement automatique du langage.

De plus, les forums de discussion semblent être à favoriser pour l'application de ce genre de méthodes par rapport aux données issues de Twitter et Facebook. En effet, en 2018, une étude française [67] a été menée pour évaluer les algorithmes de détection du signal dans les forums de discussion en les comparant aux Els présents dans VigiBase. La Sensibilité des signaux générés allait de 29% à 50.6% et la spécificité de 86.1% à 95.5%. De plus, 38% de ces signaux étaient détectés plus rapidement dans les forums que sur VigiBase.

Le problème est donc que d'après une étude importante [24], il semble que les médias sociaux ne contiennent pas suffisamment d'information pertinente (en quantité et qualité des messages) pour servir de matériel à des méthodes de détection du signal. Mais leur étude s'est principalement portée sur des données issues de Twitter et Facebook. Leur étude sur les forums a conduit à des conclusions similaires.

c) Limites et défis

Dans une scoping review [68], Lee et al. ont répertorié les défis/limites associé(e)s à l'étude des médias sociaux pour la pharmacovigilance :

- La couverture d'une telle étude :
 - o La population des utilisateurs n'est pas représentative de la population générale.
 - o Les médicaments consommés et les conditions médicales ne sont pas représentatifs.
 - o La période d'étude est limitée.
- La qualité des données :
 - o Problème lié au langage patient, à l'orthographe dans les médias sociaux.
 - o Identification des doublons.
 - o Manque d'information pour la causalité.
 - o Donnée incomplète, non vérifiable ou fausse
 - o La difficulté de détecter des signaux parmi le bruit
- Le traitement des données :
 - o Leur volume très important.
 - o L'analyse au niveau du mot n'atteint pas le niveau sémantique.

4.5 Analyse des thèmes abordés dans les médias sociaux

L'objectif secondaire de cette thèse est l'analyse des thèmes abordés par les utilisateurs de réseaux sociaux dans le cadre de la surveillance de la perception des vaccins. Pour cela, je vais appliquer une analyse de *topic models*.

Une analyse en *topic models* est une méthode d'apprentissage qui permet de mettre en lumière les thèmes sous-jacents à un corpus de texte.

C'est un modèle probabiliste qui permet de définir l'appartenance de textes à des topics, ou thèmes en faisant l'hypothèse que chaque texte contient une ou plusieurs thématiques et que ces thématiques peuvent être identifiées par les mots utilisés.

Avant d'appliquer cette analyse, un pré-traitement des données est nécessaire à la suite duquel une matrice terme-document peut être créée, indiquant la fréquence des termes dans la collection de documents. Ensuite, l'analyse en *topic models* est généralement basée sur l'allocation de Dirichlet latente [69].

Deux études utilisant les *topic models* sur les médias sociaux pour la pharmacovigilance ou la sécurité alimentaire ont déjà été publiées. En 2016, Sullivan et al. [70] ont utilisé les *topic models* sur des commentaires utilisateurs de compléments alimentaires sur Amazon.com, afin de pouvoir catégoriser ces compléments selon leur niveau de risque pour le consommateur : en biaisant l'analyse en *topic models* en garantissant que les mots identifiés comme des effets indésirables soient classés selon certains topics. Ainsi, le niveau de risque d'un complément dépendait de la distribution de ces topics. Cette méthode donnait des résultats similaires à une classification humaine des niveaux de risques des compléments pour 69,4% de ces compléments.

Dans une étude de Chen et al. [65] à laquelle j'ai participé, nous avons appliqué une analyse en *topic models* sur un corpus de messages issus de forums de discussion où le méthylphénidate avait été identifié. Parmi les 14 topics que nous avons identifiés, quatre ont été évalués comme présentant un intérêt : un topic sur les effets indésirables et la posologie, un autre sur l'utilisation du méthylphénidate par les adultes, et deux topics évocateurs du mésusage. Une classification hiérarchique descendante était alors appliquée aux topics d'intérêts afin d'étudier de manière plus précise leur constitution. Finalement, cette étude a permis de montrer que la plupart des topics identifiés dans les messages concernant le méthylphénidate étaient relatifs à son utilisation réglementaire, mais que cependant, les patients étaient soucieux des effets secondaires du traitement (3^{ème} topic le plus important). Et elle a également permis de retrouver des messages liés au mésusage de ce médicament.

Ces deux études justifient l'intérêt d'appliquer ce type de méthodes aux données issues des médias sociaux dans le cadre d'études de pharmacovigilance. Nous avons cherché à l'appliquer afin d'analyser l'hésitation vaccinale des vaccins contre la COVID-19.

4.6 Chaînes de traitement : évaluation de l'existant

Le second objectif secondaire de cette thèse concernait l'enrichissement d'une suite logicielle en partie développée préalablement dans le laboratoire par l'intégration des méthodes de détection du signal et la conception et le développement d'une interface utilisateur permettant d'appliquer les algorithmes et de visualiser leurs résultats.

Depuis les 20 dernières années, plusieurs chaînes de traitement permettant l'évaluation des messages issus des médias sociaux ont été décrits dans la littérature [21]. La détection d'effets indésirables requiert une analyse qualitative et quantitative des données. L'analyse qualitative des messages consiste à vérifier que le message contient des informations similaires à celles nécessaires à l'étude des cas de pharmacovigilance³. L'analyse quantitative considère les données de manière descriptive et en appliquant des analyses statistiques comme la détection du signal ou le change-point analysis (détection des points de ruptures dans une évolution temporelle). Dans ce travail de thèse, nous nous intéressons à l'analyse quantitative des données.

Dans cette section seront présentés quelques chaînes de traitement et outils qui ont été développés pour analyser les messages issus des médias sociaux dans le cadre de la pharmacovigilance, du mésusage ou de la toxicité de certains médicaments. Nous les évaluerons selon les critères suivants :

- Les critères liés à l'usage de ces chaînes de traitement :
 - o Permet-il l'analyse des effets indésirables liés aux médicaments ? (Nous nous intéressons aux couples médicaments-EIs)
 - o Est-il destiné à une utilisation en routine ? (Une chaîne de traitement devrait permettre une utilisation en routine)
 - o Est-il financé par le publique ou par le privé ? (Critère à caractère informatif).
- Les critères liés à l'extracteur :
 - o Est-il disponible à l'utilisation (open source) ? (Critère à caractère informatif)
 - o Permet-il l'extraction de messages issus de forums de discussion ? (Nous concentrons notre analyse sur ces forums).

³ Ces informations incluent : l'âge et le genre du patient, la chronologie, la gravité et le caractère attendu ou inattendu de l'effet indésirable, le dechallenge, le rechallenge et la durée avant le début de l'EI.

- Permet-il l'extraction de tous les messages d'une source de données ? (Nous recommandons l'extraction de tous les messages pour éviter des biais de sélection de messages).
- Applique-t-il la dé-identification ? (Pour des raisons éthiques et de réglementation nous favorisons la dé-identification).
- Les critères liés à l'annotateur :
 - Est-il adapté au français ? (Notre étude se limite au français).
- Ceux concernant les statistiques descriptives disponibles (ils permettent une analyse prenant en compte la temporalité) :
 - L'évolution temporelle du nombre de messages/annotations est-elle visualisable ?
 - Est-il possible de réaliser un change-point analysis ?
- Et, enfin, les critères liés aux méthodes de détection du signal (nous pensons que différentes méthodes de détection du signal devraient être applicables) :
 - Permet d'appliquer une méthode basée sur la régression logistique ?
 - L'évolution temporelle du PRR est-elle visualisable ?
 - Une méthode de détection du signal est-elle appliquée ?
- Une interface utilisateur est-elle implémentée ?

Les évaluations de ces chaînes de traitement selon ces critères sont résumées dans le Tableau 3.

PREDOSE

Le projet PREDOSE (PREscription Drug abuse Online Surveillance and Epidemiology, c'est-à-dire la surveillance et l'épidémiologie d'abus de médicaments prescrits) [71], financé par les Instituts Nationaux de Santé en 2011, est l'un des premiers projets implémentant une chaîne de traitement complète pour utiliser les forums de discussion pour analyser l'usage illégal d'opioïdes prescriptibles. La toxicovigilance, dont le but est de surveiller les effets toxiques d'un produit, d'une substance ou d'un mélange de produits sur l'Homme, a des caractéristiques communes avec la pharmacovigilance. Ainsi, cette approche est intéressante à considérer même si la pharmacovigilance se concentre sur les effets indésirables liés aux médicaments et non à la toxicité. Dans ce projet, l'information sémantique est extraite des médias sociaux en annotant les messages via l'ontologie d'abus de médicaments (<http://bioportal.bioontology.org/ontologies/DAO>) et en utilisant un extracteur de sentiments. Ce projet permet la consultation et la validation de cas ainsi que la surveillance de l'évolution temporelle du nombre de messages. Bien que ce projet ait clairement démontré la faisabilité d'exploiter les médias sociaux pour la surveillance de l'utilisation d'opioïdes, il n'était pas centré sur l'étude des effets indésirables et ne permettait pas l'application de méthodes de détection du signal. L'extraction

comportait certaines limites : l'absence de dé-identification et il était limité à trois forums spécifiques. De plus, les dernières démos de l'outil datent de 2013. PREDOSE est le prédécesseur du projet EDrugTrends [72,73] qui propose une plateforme qui facilite l'analyse des tendances sur Twitter et les forums de discussion en connaissances, attitudes et comportements liés à la consommation de cannabis et de cannabinoïdes de synthèse

INSIGHT EXPLORER

Insight Explorer [74] est un outil implémenté en 2012 par GlaxoSmithKline Inc durant le projet CRAWL (Contextualization of Real-World Drug Use through Social Listening) pour aider l'évaluation des messages issus de médias sociaux. Cet outil a ensuite été adapté pour le projet WEB-RADR (Recognizing Adverse Drug Reactions) pour analyser manuellement les messages et construire un gold standard. Cet outil n'est pas lié à un extracteur ni à une méthode d'annotations automatique spécifique et permet seulement l'annotation manuelle de messages. Cet outil permet le recueil d'un grand nombre d'informations liées à l'évaluation de la causalité et de la qualité. Les utilisateurs peuvent recueillir des données concernant le patient (tranche d'âge, genre, etc.) et le produit (nom, voie d'administration, durée d'utilisation, dosage, etc.). Mais Insight Explorer ne permet pas d'effectuer une analyse quantitative. A notre connaissance, cet outil n'a pas été utilisé depuis la fin du projet Web-RADR [75].

MEDWATCHER SOCIAL

MedWatcher Social [76] est une plateforme de surveillance des médias sociaux dont la première démo a été publiée en 2014. Elle a été développée par la FDA et le Brigham & Women's Hospital de Boston. L'objectif était d'analyser les données issues de médias sociaux, en particulier des forums de discussion relatifs à la santé, Twitter et Facebook, pour la détection d'effets indésirables liés aux médicaments. L'extraction quotidienne était réalisée par des machines d'acquisition collectant les données en utilisant des interfaces de programmation d'application (APIs) ou des extracteurs développés spécifiquement. Les médicaments et les symptômes étaient extraits automatiquement du texte par l'utilisation d'un dictionnaire auxquels des synonymes ont été ajoutés. Les symptômes étaient alors reliés à un terme MedDRA.

MedWatcher Social permet d'appliquer une méthode fréquentiste de détection du signal et d'analyser l'évolution temporelle du nombre de mention. Une approche statistique initialement utilisée pour classer les mails selon qu'il s'agisse de spam ou non [Robinson 2003] fût adaptée pour classer chaque message en deux catégories : contenant un effet indésirable ou non. Pour limiter le volume de données, l'extraction a été limitée aux messages contenant la mention d'un médicament (selon une liste de noms de médicaments). L'inconvénient étant que les erreurs de frappes, abréviations ou

surnoms ne sont pas pris en compte. La dernière publication dans laquelle MedWatcher Social fût utilisée date de 2019 [77].

YELESWARAPU ET AL.

Yeleswarapu et al. [78] ont décrit une chaîne de traitement semi-automatique développée et financée par Tata Consultancy Services (Inde) en 2014. L'objectif était d'extraire des paires médicament-effet indésirables de bases de données de pharmacovigilance et d'autres sources comme des abstracts de MEDLINE et des commentaires d'utilisateurs de sites internet en santé.

Les commentaires issus des sites web ont été extraits via Web-Harvest, un outil d'extraction open-source (<http://web-harvest.sourceforge.net>). Des méthodes de traitement automatique du langage ont ensuite été appliquées comme la reconnaissance d'entités nommées, la racinisation et la tokenization. Une méthode de détection du signal a également été implémentée. Cette chaîne de traitement n'était pas destinée à une utilisation en routine.

DOMINO

L'interface Domino [79] a été créée en 2018 par l'Université de Bordeaux et financée par l'ANSM. L'objectif était d'analyser le mésusage dans les forums de discussion en santé. Les messages étaient extraits avec des informations concernant la structure du forum (section, sous-section, fil de discussion...), la date de publication et des informations concernant les auteurs (date de naissance et genre). Un modèle sémantique a été utilisé pour stocker l'information extraite sous la forme d'un triplet RDF (Resource Description Framework, sujet-prédicat-objet). Les messages étaient alors annotés via la reconnaissance d'entité nommées en utilisant l'UMLS et ROMEDI ("référentiel ouvert du médicament") [79]. ROMEDI permet de relier une annotation d'un médicament au résumé des caractéristiques du produit (RCP). L'interface permet de visualiser des données agrégées ainsi que les messages annotés. Cet outil ne permet pas la surveillance des effets indésirables liés aux médicaments et n'inclut pas de méthodes de détection du signal. Il a premièrement été destiné aux antidépresseurs et d'autres catégories de médicaments sont actuellement en train d'y être intégrées.

NIKFARJAM ET AL.

Cette chaîne de traitement décrite par Nikfarjam et al. en 2019 [80] a été implémentée par l'Université de Stanford et Inspire (Etats-Unis), une entreprise avec un réseau social en santé de plus d'un million de patients et de professionnels de santé. L'objectif était d'étudier l'association d'anticancéreux avec des effets indésirables cutanés dans les messages de patients sur Inspire. Afin d'extraire les mentions d'événements indésirables, ils ont utilisé un réseau de neurones basé sur la reconnaissance d'entités nommées entraîné spécifiquement pour du contenu produit par des utilisateurs de médias sociaux.

Pour la normalisation, ils se sont concentrés sur quelques médicaments sélectionnés et sur seulement 8 effets indésirables reliés à l'identifiant conceptuel unique (CUI) de l'UMLS via un lexique. Cette chaîne de traitement a donc été créée pour cette utilisation spécifique et non pour être utilisée en routine. Une méthode de détection du signal a été implémentée (PRR). De plus, ils n'ont pas développé d'interface utilisateur pour faciliter les analyses, et les messages n'ont pas été extraits mais fournis par Inspire.

MAGGE ET AL.

Magge et al. [42] ont décrit en 2021 une chaîne de traitement implémentée par l'Université de Philadelphie (Etats-Unis), l'Université Fédérale de Kazan (Russie) et l'université de Leiden (Pays-Bas) et financée par des instituts nationaux de santé, la Librairie Nationale de Médecine, le fond SIDN et la Fondation Russe des Sciences. L'objectif du projet était l'extraction et la normalisation de mentions d'effets indésirables liés aux médicaments sur Twitter. Les tweets ont été collectés en utilisant l'API (interface de programmation d'application) de Twitter (<https://developer.twitter.com/en/docs/twitter-api>).

Leur chaîne de traitement était composée d'un classifieur d'effets indésirables pour identifier les tweets faisant l'objet d'un EI, d'une reconnaissance d'entités nommées, d'un normalisateur reliant les EIs à un code MedDRA. Ce normalisateur était limité aux EIs présents dans les données d'entraînement. Cette chaîne de traitement ne comportait pas de méthodes de détection du signal, s'intéressait uniquement à Twitter et n'était pas destinée à une utilisation en routine.

ENTREPRISES PRIVEES (exemples non exhaustifs)

Des entreprises privées comme Kappa Santé (<https://www.kappasante.com/>) à travers sa startup Kap Code et la plateforme DETEC'T proposent un outil d'analyse des médias sociaux, notamment pour la pharmacovigilance [81].

L'extracteur de Kap Code permet l'extraction des messages ainsi que de leurs métadonnées. Il permet la dé-identification des auteurs et de nouveaux forums ou sources de données peuvent facilement être ajoutées. Leur interface, destinée à une utilisation en routine, permet une analyse quantitative des messages, les requêtes concernent le médicament, la période, l'effet indésirable et le type de relation identifié.

L'objectif du projet ADR PRISM était lui aussi d'utiliser les données issues des médias sociaux pour la pharmacovigilance. Ce projet a été financé en 2013 par la *Direction Générale des entreprises et des collectivités territoriales* (Ile de France et Haute Normandie) et a été mené par des équipes de recherche publiques et des entreprises privées. Ils ont utilisé leur propre extracteur [82]. En plus du contenu du

message, les métadonnées liées au message étaient aussi extraites et stockées dans la base de données. L'annotation des messages a été réalisée par le Luxid Annotation server® via un web service permettant l'annotation automatique des messages. Les messages ont donc été annotés avec des concepts issus de la terminologie Vidal et du portail HeTOP (Health Terminology/Ontology Portal) développé par le CISMef (CHU de Rouen) qui permet l'accès à plus de 70 terminologies/ontologies en santé [83]. Des algorithmes de détection du signal ont été appliqués [65].

Expert System (<https://www.expert.ai/fr>) dans le cadre du projet ADR PRISM, a implémenté BIOPHARMA Navigator pour extraire les messages issus de forums de discussion ainsi que leurs métadonnées. L'interface développée par Expert System était destinée à l'annotation des messages. Une limite de cette interface était que l'utilisateur ne pouvait ni choisir les forums ni la période à analyser.

Récapitulatif

Parmi les dix chaînes de traitement identifiées, un même nombre était financé par le public et par le privé. Huit d'entre elles permettaient l'analyse des effets indésirables liés aux médicaments, mais seulement quatre étaient destinées à une utilisation en routine. Cinq des extracteurs étaient en open source, six extrayaient tous les messages ; six l'extraction de messages issus de forums, et seulement 3 réalisaient la dé-identification. Six permettaient la visualisation de l'évolution temporelle du nombre de messages mais permettait l'application du change point analysis était possible dans un seul d'entre elles. Les méthodes de détection du signal étaient disponibles pour 5 de ces chaînes de traitement, aucune ne permettant de visualiser l'évolution temporelle du PRR, ni d'appliquer une méthode basée sur la régression logistique. Finalement, une interface utilisateur avait été implémentée pour six d'entre elles.

Tableau 3 : Evaluation des chaînes de traitement selon nos critères

	Pipeline	PREDOSE	INSIGHT EXPLORER	MEDWATCHER SOCIAL	YELESWARAPU ET AL.	DOMINO	NICKFARJAM ET AL.	MAGGE ET AL.	ADR PRISM	DETECT'	EXPERT SYSTEM
	Etude des ADRs	X	✓	✓	✓	X	✓	✓	✓	✓	✓
	Routine	✓	X	✓	X	✓	X	X	X	✓	X
	Public/Privé	Public	Privé	Public	Privé	Public	Public et Privé	Public	Public et Privé	Privé	Privé
Extracteur	Tous les messages	✓	X	X	X	✓	X	✓	✓	✓	✓
	Dé-identification	X	X	X	X	X	X	X	✓	✓	✓
	Forums	✓	X	✓	X	✓	X	X	✓	✓	✓
	Open source	✓	✓	✓	X	✓	X	✓	X	X	X
Annotateur	Français	X	X	X	X	✓	X	X	✓	✓	✓
Statistiques	Evolution temporelle	✓	✓	✓	X	✓	X	✓	✓	✓	X
	Change point analysis	X	✓	X	X	X	X	X	X	✓	X
Détection du signal	Méthode de Détection Signal	X	✓	✓	✓	X	✓	X	✓	✓	✓
	Evolution temporelle PRR	X	✓	X	X	X	X	X	X	X	X
	Régression logistique	X	X	X	X	X	X	X	X	X	X
	Interface	✓	✓	✓	X	✓	X	X	✓	✓	✓

Chapitre 3

Matériel

Deux modules de la chaîne de traitement ont été développés préalablement à l'analyse des données. L'un permet d'extraire les messages issus de forums de discussion (le module Scrapper) et l'autre permet l'annotation de ces messages (le module annotation). Ce chapitre permettra de vous présenter ces deux modules, leurs fonctionnements ainsi que les volumes de données qu'ils ont permis d'obtenir.

1. Choix des forums

23 forums français de discussion ont été sélectionnés afin d'en extraire les messages. Ces forums sont tous publics et ne nécessitent aucune authentification. Les forums potentiellement intéressants ont été identifiés soit en réalisant une simple recherche sur Google ou bien via une recherche dans la liste de la Fondation Health On the Net (HON) des sites certifiés réalisée en collaboration avec la Haute Autorité de Santé (HAS) en français disponible via le lien suivant : https://www.has-sante.fr/jcms/c_334538/fr/la-certification-des-sites-internet-sante-de-novembre-2007-a-juillet-2013. Les sites devaient être hébergés en France et contenir un forum de discussion ou un endroit où partager des commentaires et contenait plus de 10 contributions de patients. Le Tableau 4 décrit le nombre de messages extraits et les dates de première et dernière extraction pour chaque forum au 28/02/2021.

2. Module Scraper

Le premier module permet d'extraire les messages issus de forums de discussion, il s'agit du Vigi4Med Scraper [84]. Ce logiciel a été créé dans le cadre du projet Vigi4Med décrit en avant-propos.

Il s'agit d'un logiciel en open-source composé de plusieurs éléments (Figure 6) permettant d'extraire des données structurées de forums de discussion. Il est hautement configurable : en utilisant un fichier de configuration (un par site internet), l'utilisateur peut librement choisir les données à extraire de n'importe quel forum. Ce fichier contient la liste des forums et les objets à extraire. Pour chaque URL dans cette liste, la fonction d'extraction génère une autre liste qui contient les URLs de tous les fils de discussion présents dans le forum afin d'extraire les messages de chaque fil de discussion avant d'extraire ceux du forum suivant. Avant d'effectuer la requête d'un URL, la fonction commence par vérifier les logs (c'est-à-dire les connexions) : si l'URL a déjà été extraite, elle sera ignorée ; sinon, l'URL est envoyée au proxy. Le proxy retrouvera la page HTML correspondante dans son cache ou récupérera

la page depuis le site web. Un Document Object Model (DOM) est généré à partir de la page HTML et les requêtes Xpath (les objets à extraire) dans le fichier de configuration sont exécutées sur le DOM pour obtenir des blocs de données structurés. Par exemple, si l'utilisateur définit les Xpaths des titres des messages, des auteurs et des dates de publication dans le fichier de configuration, l'algorithme produira des blocs de données structurés avec les valeurs de ces champs pour chaque message dans chaque page HTML extraite. Vigi4Med Scraper continuera d'extraire les fils de discussion et les messages jusqu'à la dernière page des fils de discussion, où le lien qui mène à la page suivante est absent.

L'extraction des messages mime en fait la structure du forum : où tous les messages appartiennent à des fils de discussion et chaque fil de discussion fait partie d'un forum ; un même site pouvant héberger plusieurs forums.

Afin de respecter le RGPD (<https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:32016R0679>) les données extraites sont dé-identifiées. Pour cela, nous définissons le pseudonyme de l'auteur, les informations du profil et les URLs des messages et des fils de discussion comme les identifiants à dé-identifier. Chaque identifiant est remplacé par une clé générée par la fonction de hachage cryptographique SHA1 (Secure Hash Algorithm 1, https://en.wikipedia.org/wiki/SHA-1#cite_note-1-3). Pour casser la dé-identification, il suffit de concaténer les données dé-identifiée avec leur clé. Il est d'usage que la dé-identification soit irréversible, cependant, nous avons fait l'hypothèse que, dans certains cas, il pourrait être nécessaire de retrouver les identifiants. Par exemple, en pharmacovigilance, la détection d'une exposition dangereuse à un médicament peut nécessiter de recommander au patient de contacter son médecin. Néanmoins, nous avons défini la dé-identification de façon à générer un graphe des clés dé-identifications séparé, qui sera conservé dans un endroit sécurisé dans l'utilisation en routine. De plus, ces clés ne seront pas utilisées par l'équipe qui analysera les données dé-identifiées. Afin de garantir la validité et la qualité de notre graphe RDF, la dé-identification distingue trois profils d'auteur : les auteurs connus (utilisateurs avec une page de profil), les auteurs inconnus (surnoms avec aucune page de profil) et les auteurs invalides (utilisateurs anonymes ou profils supprimés). Chaque auteur valide correspond à trois triplets dans le graphe RDF : le premier décrit son genre, le second correspond à sa page de profil et le troisième à son pseudonyme. Ainsi, pour dé-identifier les informations d'un utilisateur correspondant à un auteur connu, le script dé-identifiera le pseudonyme de l'auteur et la page de profil. Pour les auteurs inconnus, il dé-identifiera seulement le pseudonyme et un nœud vide est créé pour représenter la page de profil manquante. Les auteurs invalides n'apparaissent pas dans le graphe et leurs messages sont considérés comme des messages sans auteur. Cela évite de considérer les messages sans auteurs comme provenant du même « Inconnu ».

Pour éviter la surcharge du serveur hébergeant le forum pouvant être provoquée par les multiples connexions nécessaires à l'extraction, il est important d'éviter les doublons dans les requêtes et d'avoir un temps de pause entre deux connexions successives. Pour répondre au premier problème, un fichier log garde une trace de tous les URLs visités dans chaque site web. Avant de faire une requête d'un nouvel URL, le script vérifie si celui-ci a ou non déjà été visité. Afin de gérer les problèmes inattendus liés au fait de retrouver un URL déjà visité, un proxy avec une base de données cache a été choisi. Avant d'ouvrir une nouvelle connexion avec le serveur, la fonction vérifie si sa base de données contient déjà la page HTML désirée. Si tel est le cas, la page HTML est renvoyée à la fonction ; sinon, une connexion est établie avec le serveur et une copie de la page HTML reçue est sauvegardée dans le cache. De plus, le proxy assure un temps d'attente minimum de 0.330 secondes entre deux requêtes. Ce temps d'attente est augmenté à 2 secondes pendant les heures de bureau pour garantir les performances du réseau.

Le Vigi4Med Scraper a été adapté pour le projet PHARES. Cette adaptation était nécessaire car le projet PHARES requiert une extraction continue des forums pour permettre aux messages d'être ensuite annotés et disponibles pour l'analyse via l'interface de manière régulière. C'est pourquoi le script fût modifié afin de générer la donnée extraite en format tabulaire (CSV) et de la conserver dans une base de données. Ces nouveaux éléments ont été générés en plus du format RDF, conservé pour un alignement éventuel avec une ressource sémantique externe. L'avantage de générer des formats tabulaires était de faciliter l'entrée de données dans le module suivant (module d'annotation). Le format de la base de données (Figure 7), dont la structure est décrite dans la section consacrée au module d'annotation, a été utilisé afin d'optimiser l'interaction avec l'interface. Un fichier log a également été intégré avec la structure du scraper afin de garder une trace du dernier élément extrait afin que l'extraction journalière démarre où la dernière extraction s'était arrêtée.

Au total, ce travail de thèse a concerné 2 081 272 de messages extraits entre le 01/01/2017 et le 28/02/2021. Plus de la moitié de ces messages (52,10%) est issue de quatre forums de discussion : Vivre sans thyroïde, Doctissimo Santé, Doctissimo Nutrition et Info Bébé. Les dates des messages n'étaient pas disponibles pour le forum Bébé Maman (Tableau 4).

Tableau 4 : Volumétrie de l'extraction

Forum	Nombre de messages extraits	% Cumulé	Date de dernière extraction
Vivre sans thyroïde	524 629	25,21%	25/02/2021
Doctissimo Santé	248 691	37,16%	16/01/2021
Doctissimo Nutrition	183 730	45,98%	16/01/2021
Info Bébé	127 341	52,10%	08/03/2019
Atoute	118 415	57,79%	28/02/2021
Notre Famille	97 098	62,46%	26/10/2017
Magic Maman	96 713	67,10%	22/02/2021
Doctissimo Médicaments	95 531	71,69%	15/01/2021
Doctissimo Grossesse	93 449	76,18%	15/01/2021
Au Féminin	72 732	79,68%	09/01/2020
Maman Vie	69 167	83,00%	10/04/2019
Onmeda	61 428	85,95%	24/02/2021
Ados	58 181	88,75%	08/03/2019
Carenity	52 659	91,28%	29/08/2020
Famili	51 844	93,77%	17/11/2019
Baby France	43 806	95,87%	30/04/2018
Bébé Maman	38 450	97,72%	NA
Allo Docteurs	15 833	98,48%	09/02/2021
reboot	9 383	98,93%	25/02/2021
Futura	6 765	99,26%	22/02/2021
Sport Santé	6 350	99,56%	14/01/2020
Maladies Rares	4 827	99,80%	14/05/2020
Que Choisir	4 250	100,00%	11/02/2021
Total	2 081 272	-	28/02/2021

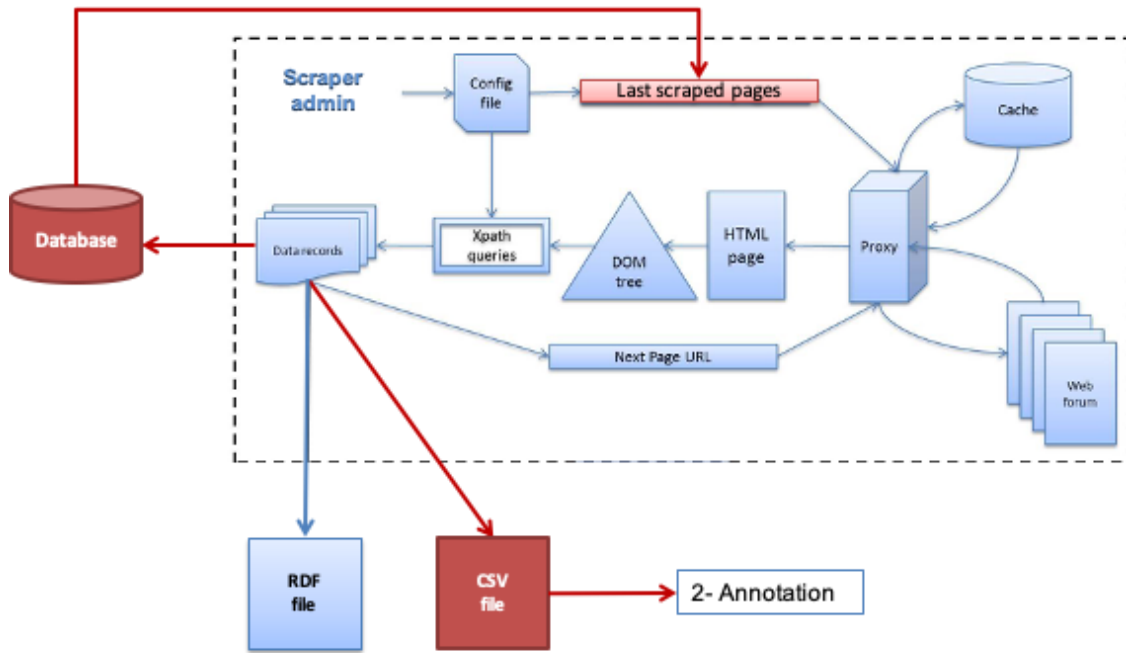


Figure 6 : Structure du Vigi4Med Scraper adapté pour PHARES



Figure 7 : Structure de la base de données PHARES

3. Module annotation

Notre objectif premier est de définir un cadre pour l'application des méthodes de détection du signal aux données issues des médias sociaux dans un cadre opérationnel en répondant également à deux objectifs secondaires : démontrer l'utilité de ces données pour d'autres types d'analyses comme celle des thèmes abordés par les utilisateurs de réseaux sociaux lorsqu'ils traitent d'un certain sujet, et l'implémentation et l'utilisation d'une interface utilisateur, complétant une chaîne de traitement permettant d'appliquer les algorithmes de détection du signal, ainsi que l'étude des thèmes abordés par les utilisateurs de manière quotidienne. Nous avons fait le choix d'utiliser la méthode de traitement automatique du langage présentée dans l'article de Morlane-Hondere et al. [39] car c'est celle qui fût utilisée pour les projets Vigi4Med et PHARES et que le module était donc déjà disponible. Ce module pourra facilement être modifié dans le futur pour tenir compte de l'évolution des performances des approches de TAL.

Les termes ont été annotés en utilisant une chaîne de traitement automatique du langage [39] qui détecte tout d'abord les entités correspondant à un médicament ou à un EI en utilisant des champs aléatoires conditionnels [40]. Cette méthode permet une prédiction prenant en compte le contexte (dans notre cas, les mots avant et après celui à annoter). Puis, une machine à vecteurs de support prédit la relation entre les deux entités identifiées afin de savoir s'il existe une relation causale exprimée dans la phrase entre l'EI et le médicament.

Par exemple, dans le message décrit figure 8, 4 effets indésirables ont été détectés : « maux de tête », « nodules cancéreux », « fatigue », et « perte de poids ». Chacun de ces EIs a été associé à chaque médicament mentionné dans le message (levothyrox® et calcium), formant 8 couples médicament-EI.

“Bonsoir on m’a oté la thyroïde début juillet. Je suis donc sous lévothyrox (le nouveau !!) Et je suis diabétique sous médocs. Mes questions du soir : j’ai de grosses insomnies, maux de tête non stop, fatigue, perte de poids (2kgs en 1 mois). Cela pourrait il venir du diabète qui en ce moment yoyotte pas mal ? Je fais une prise de sang ds 1 sem pr calcium et thyroïde. Je suis un peu perdue.... merci pour vos conseils. J’ai vu ma diabéto il y a 10 jrs mais elle ne m’a pas bcp éclairée ! Juste que la thyroïdectomie pouvait jouer sur le diabète. Je revois le chir qui m’a opérée le 24. Pour infos : on avait découvert des nodules cancéreux sur la thyroïde mais elle, elle n’avait rien. Voilà ! merci. J’ai un peu le moral ds les chaussettes ! (Les nuits sont fraîches alors je les remets) ???? “

Figure 8 : Message numéro 29605 du forum alloDocteurs

Ce module a été amélioré au cours du projet PHARES par le Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur (LIMSI) afin de normaliser les annotations selon la terminologie MedDRA et la classification ATC.

La figure 9 représente un message annoté où l'information à propos de l'EI est représentée en rouge et celle concernant le médicament, en bleu. Tout d'abord, les informations concernant le message sont collectées (identifiant du message, du forum, du fil de discussion, titre du fil de discussion, date du message et contenu de ce dernier). Dans cet exemple, 2 EIs ont été repérés dans le titre du fil de discussion (« infections », code MedDRA : 10021789, et « diabète de type 2 » sans code MedDRA identifié), et 4 EIs ont été détectés dans le message (« diabétique », « cicatrisation », « ampoule », « morsure de chien » en rouge). Deux médicaments ont été annotés dans le message (antibio, vaccin) et un a été associé à un code ATC (antibio).

```
<post id='1423'>
  <id_forum>carenity</id_forum>
  <id_fil>8570</id_fil>
  <titre_fil>
    <original>Les infections de la bouche causées par le diabète de type 2</original>
    <analyse>Les <patho id='1423_1' mdrfre='10021789'>infections</patho> de la
    bouche causées par le <patho id='1423_2'>diabète de type 2</patho>.</ analyse>
  </titre_fil>
  <date>2014-08-29 00:12:56</date>
  <texte>
    <original>bonsoir, décelée il y a 1 mois diabétique de type 2 j'ai rencontré des
    problèmes de cicatrisation lors d'ampoule et de morsure de chien malgres les soins
    apportés. antibio, vaccins , pansements tous les jours</original>
    <analyse>bonsoir, décelée il y a <duree id='1423_3'>1 mois</duree> <patho
    id='1423_4' mdrfre='10012641'>diabétique</patho> de type 2 j'ai rencontré des
    problèmes de <patho id='1423_5'>cicatrisation</patho> lors d'<patho id='1423_6'
    mdrfre='10005216'>ampoule</patho> et de <patho id='1423_7'
    mdrfre='10013589'>morsure de chien</patho> malgres les soins apportés.
    <trt id='1423_8' atc='S02CA05'>antibio</trt>, <trt id='1423_9'>vaccins</trt> ,
    pansements <freq id='1423_10'>tous les jours</freq></analyse>
  </texte>
</post>
```

Figure 9 : Exemple de message annoté

La sortie de ce module est constituée de fichiers CSV avec les variables suivantes :

- concernant le message :

- identifiant du message,
- nom du forum,
- date de publication,

- concernant l'EI :

- le verbatim détecté comme étant un EI,
- l'EI normalisé (selon la terminologie MedDRA),
- son code MedDRA,
- le concept unique identifier (CUI) de l'UMLS de l'EI,

- et concernant le médicament

- le verbatim identifié comme médicament,
- le médicament normalisé (selon la classification ATC),
- le principe actif
- et son code ATC.

Dans ces fichiers CSV, une ligne peut concerner une annotation de seulement un EI, seulement un médicament, ou d'un couple pour lequel un lien de causalité a été retrouvé. Le tableau 5 représente un extrait de la base de données pour l'exemple concernant le Levothyrox® vu précédemment.

Tableau 5 : Extrait de la base de données

Nom du Forum	Id message	Date	heure	Verbatim EI	EI normalisé	CUI	Verbatim médicament	Médicament normalisé	Principe actif	Code MedDRA	Code ATC
atoute	7354	10/08/2018	21:37:00	maux de tête	Céphalée	C0018681	lévothyrox	LEVOTHYROX	Levothyroxine sodique	----	H03AA01
atoute	7354	10/08/2018	21:37:00	maux de tête	Céphalée	C0018681	calcium	----	----	----	----
atoute	7354	10/08/2018	21:37:00	nodules cancreux	----	----	lévothyrox	LEVOTHYROX	levothyroxine sodique	----	H03AA01
atoute	7354	10/08/2018	21:37:00	nodules cancreux	----	----	calcium	----	----	----	----
atoute	7354	10/08/2018	21:37:00	fatigue	Fatigue	C0015672	lévothyrox	LEVOTHYROX	levothyroxine sodique	10016256	H03AA01
atoute	7354	10/08/2018	21:37:00	fatigue	Fatigue	C0015672	calcium	----	----	10016256	----
atoute	7354	10/08/2018	21:37:00	perte de poids	Poids diminué	C0043096	lévothyrox	LEVOTHYROX	levothyroxine sodique	10048061	H03AA01
atoute	7354	10/08/2018	21:37:00	perte de poids	Poids diminué	C0043096	calcium	----	----	10048061	----

Tableau 6 : Volumétrie des annotations

Forum	Total
Annotations couples	2 454 310
Annotations EI normalisés	1 988 451 (81,02%)
Annotations Médicaments normalisés	713 067 (29,05%)
Annotations Couples normalisés	580 299 (23,64%)

Au total (Tableau 6), 2 454 310 couples ont été annotés. Parmi eux, 1 988 451 annotations de verbatims en tant qu'effets indésirables ont été normalisées selon MedDRA (81,02%), et 713 067 verbatims identifiés comme étant des médicaments ont pu être normalisés selon la classification ATC (29,05%). Finalement, 580 299 couples annotés ont été normalisés (23,64%).

Chapitre 4

Méthode

Dans cette partie, nous allons tout d'abord décrire la méthode utilisée pour définir un cadre pour l'application des algorithmes de détection du signal aux données issues des médias sociaux de manière opérationnelle et pour une utilisation quotidienne. Nous verrons la méthode choisie pour l'objectif secondaire, qui est de démontrer l'utilité de ces données pour d'autres types d'analyses que la détection du signal, avec l'analyse des thèmes abordés par les utilisateurs de réseaux sociaux. Enfin, nous présenterons la conception, l'implémentation et l'utilisation d'une interface utilisateur permettant d'appliquer les algorithmes de détection du signal ainsi que, dans le futur, l'étude des thèmes abordés par les utilisateurs et de visualiser les résultats.

1. Création et évaluation du module d'analyse (module 3)

Cette partie concerne le module statistique que nous avons implémentée et qui fait partie de la chaîne de traitement. Nous avons déjà décrit les travaux préliminaires à ce travail de thèse dans le chapitre précédent.

Les algorithmes sélectionnés pour l'évaluation sont les suivants : PRR, ROR, BCPNN, CISL, CPA, et CPA combiné au PRR. Cette sélection prend en compte :

- des méthodes fréquentistes et bayésiennes utilisées en routine sur certaines bases de pharmacovigilance (PRR, ROR, et BCPNN),
- une méthode de régression logistique afin de prendre en compte la co-prescription (CISL), et
- le change-point analysis appliqué au nombre d'annotation pour détecter une augmentation significative de mentions du couple (CPA), ou appliqué au PRR, cette méthode, encore peu explorée, semblant obtenir des résultats satisfaisants.

Pour le PRR, le ROR et le BCPNN, le package PhVid [85] propose une implémentation et pourra être utilisé. Les données doivent être transformées sous la forme d'une base de données composée d'une ligne par couple d'annotation {médicament, EI} avec trois colonnes : le médicament, l'effet indésirable et le nombre d'occurrences de ce couple. Nous avons décidé de ne compter qu'une occurrence maximum par couple et par message : lorsqu'un couple est annoté plusieurs fois dans le même message, nous ne l'avons comptabilisé qu'une seule fois.

Pour le CISL (régression logistique, méthode décrite page 29), pour chaque effet indésirable, nous allons créer un vecteur présence/absence de l'EI dans le message ainsi qu'une matrice présence/absence des médicaments dans le message. En pharmacovigilance basée sur la déclaration spontanée, il est possible de savoir quels médicaments le patient a pris en même temps. Dans notre cas, nous faisons l'hypothèse que chaque médicament identifié dans le message a été pris par le patient. Le code en langage R de l'algorithme nous avait été fourni par Ismaël Ahmed, l'auteur de l'article introduisant cette méthode [49]. Il est désormais disponible dans le package adapt4pv [86]. Afin de générer un signal, on s'intéresse à la distribution empirique du nombre moyen de fois où un médicament est sélectionné par le lasso (lorsque le coefficient est supérieur à 0) : les médicaments retenus comme prédicteurs seront finalement ceux dont le $\alpha\%$ quantile de cette distribution est supérieur à 0. Avec par exemple α égale à 5, 10 ou 15.

Nous avons choisi d'étudier l'évolution temporelle mensuelle des logarithmes du PRR et du ROR avec leurs intervalles de confiance. Cette évolution temporelle nous permet de savoir à quelles périodes un signal a été détecté. Ainsi, nous pouvons générer un signal et le visualiser sur le graphe de cette évolution si la borne inférieure de l'intervalle de confiance de ces statistiques est supérieure à 0.

En effet, cela signifie que l'intervalle de confiance du PRR ou du ROR est supérieure à 1 et donc qu'il existe une disproportionnalité et que le couple d'intérêt est présent dans la base de manière significative.

Pour le CPA, nous allons créer un vecteur de la statistique étudiée (la fréquence d'annotation du médicament, de l'effet indésirable ou du couple) selon une certaine périodicité (mensuelle). Nous avons ensuite utilisé le code fourni dans [50,51].

Nous avons évalué ce module via les cas d'usage du Levothyrox et du vaccin contre les infections à papillomavirus humain, nous présentons dans ce chapitre les cas d'usage et les protocoles des études dont les résultats sont présentés dans le chapitre suivant

Nous avons évalué l'application de ces méthodes à travers deux cas d'usage décrits ci-après.

1.1. Cas d'usages pour la détection du signal

a) Analyse des mentions d'effets indésirables avec le Levothyrox

Contexte :

Afin de démontrer la performance et l'utilisation de la chaîne de traitement, nous avons choisi le Levothyrox comme cas d'usage. Ce médicament est prescrit en France depuis 1980 pour l'hyperthyroïdie et pour des circonstances dans lesquelles il est nécessaire de limiter la stimulation de l'hormone de la thyroïde. En 2017, une nouvelle formule du Levothyrox a été mise sur le marché. Elle est modifiée uniquement au niveau des excipients, le lactose étant remplacé par le manitol et l'acide citrique dans la nouvelle formule. Une pétition « Contre le nouveau Levothyrox dangereux pour les patients ! » a dépassé les 300 000 signatures sur mesopinions.com (<https://www.mesopinions.com/petition/sante/contre-nouveau-levothyrox-dangereux-patients/31185>). Une augmentation inattendue des notifications d'effets indésirables avec ce médicament a été observée, couverte par les médias (<https://www.leparisien.fr/societe/thyroide-les-malades-en-colere-contre-la-nouvelle-formule-du-levothyrox-23-08-2017-7207752.php>). En effet, plus de 32 000 EIs ont été déclarés par les patients en France en 2017, représentant 42% de toutes les déclarations collectées cette année et une augmentation inattendue du nombre de déclarations par les patients (5,8% en 2016) [87]. La plupart de ces déclarations concernaient la nouvelle formule du Levothyrox. En 2017, 1 664 notifications d'EIs ont été réalisées par des patients au CRPV de Nice. Après une revue de 1 544 notifications, 1 372 concernaient le Levothyrox et seulement 172 concernaient d'autres médicaments [87].

Objectif :

Cette étude nous permettra d'évaluer l'application des méthodes du module d'analyse ainsi que le module interface utilisateur.

Protocole de l'étude :

L'analyse a donc été réalisée au moyen de l'interface afin de vérifier dans quelle mesure on peut se passer de code en langage R et utiliser plutôt les composants de l'interface. La période d'étude était du 01/01/2017 au 28/02/2021 et les médicaments inclus dans l'étude étaient deux médicaments de la classe ATC "H03AA Hormones thyroïdiennes" : "Lévothyroxine sodique" et "Associations de Lévothyroxine et Liothyronine".

Les forums ont été explorés un à un pour vérifier qu'ils contenaient des informations à propos des médicaments sélectionnés (au moins 1 message) afin de réduire le temps de calcul : la transformation de la base de données afin d'appliquer le CISL nécessite un temps de calcul assez long. C'est pour cette

raison que nous avons décidé de limiter le nombre de forums sélectionnés, même si cela entraîne un biais dans les résultats obtenus.

Les messages ont été extraits, annotés et analysés via la chaîne de traitement. Les méthodes de détection du signal ont été appliquées pour 2 Els choisis du fait de leur apparition fréquente avec le Levothyrox dans nos données : “Fatigue” et “Faim”. Un signal peut être généré quand la borne inférieure de l’intervalle de confiance à 95 % du logarithme du PRR est supérieur à 0. Pour la régression logistique, nous avons appliqué le 10ème quantile.

b) [Analyse des mentions d’effets indésirables avec les vaccins contre les infections à papillomavirus humain](#)

Contexte :

L’hésitation vaccinale est un sujet important, surtout en France [88] où la peur de la vaccination peut mener à une protection plus faible de la population à des maladies infectieuses évitables. Les médias sociaux sont potentiellement intéressants pour étudier cette hésitation, et toute information peut être utile aux autorités de santé pour la diminuer et promouvoir le vaccin. Notre objectif était d’explorer les forums de discussions de santé en français pour détecter et investiguer des signaux de pharmacovigilance concernant les vaccins contre les infections à papillomavirus humain (HPV). Ces vaccins protègent contre un groupe de virus pouvant causer le cancer, spécialement de l’utérus et ont été initialement prescrits en priorité aux jeunes filles. Plusieurs études ont déjà analysé les vaccins HPV dans les médias sociaux. Suragh et al. [89] ont trouvé que 49% des 39 effets indésirables d’anxiété après vaccination sur Google et Facebook étaient liés aux vaccins HPV. Entre 2007 et 2015, presque 10% des sujets de discussion concernant les virus HPV sur Reddit concernaient les effets indésirables liés à ces vaccins [90]. De plus, comme le web 2.0 peut avoir une influence sur la décision de se faire vacciner ou non [91], il est préférable d’explorer les expériences personnelles et les opinions des utilisateurs de médias sociaux afin d’analyser la perception de ce vaccin et mesurer en quoi les médias sociaux peuvent influencer la population.

Objectif de l’étude :

Cette étude nous permettra d’évaluer l’application des méthodes du module d’analyse. Elle a fait l’objet d’une publication [92].

Protocole de l’étude :

Nous avons utilisé les trois premiers modules de la chaîne de traitement. Une étape préalable et indispensable était de changer le format des données avant de mener notre analyse. Les principales étapes du traitement des données étaient :

- La standardisation des dates ;
- Le mapping du CUI de l'UMLS au code MedDRA 21.1 ;
- Les termes correspondants au Système Organ Class « Circonstances sociales » et aux High Level Group Terms (HLGT) contenant la chaîne de caractère « procédure thérapeutique » ont été retirés de l'analyse ;
- Le verbatim utilisé par l'auteur du message identifié comme étant un traitement mais n'ayant pas été relié aux vaccins HPV s'ils contiennent les termes « Gardasil », « papillomavirus », « hpv », « Cervarix », ou les deux termes « vaccin » et « utérus » ;
- Seuls les couples codés en preferred terms MedDRA et 5^{ème} niveau de l'ATC ont été inclus dans l'étude.

Après traitement des données, nous avons appliqué différentes règles de décision pour des méthodes de détection du signal en pharmacovigilance [3,46,50]:

- 1) Bayesian Confidence Propagation Neural Network (BCPNN) : borne inférieure de l'intervalle de confiance à 95% de l'Information Component supérieur à 0 ;
- 2) Change-Point Analysis (CPA) : changement (en augmentation) avec 95% de confiance du nombre de couples ;
- 3) CPA combiné au Proportional Reporting Ration (PRR) pour 5 événements indésirables sélectionnés (« Syncope », « Réaction anaphylactique », « Arrêt cardiaque », « Suicide » et « Cécité ») : PRR supérieur à 1, couple détecté dans au moins 5 messages, et deux augmentations successives détectées par le CPA (avec 95% de confiance).

Pour chaque règle de décision, 50 messages ont été tirés au sort pour une analyse manuelle (150 au total). Ces messages ont été évalués selon les critères suivants :

- Mention du vaccin HPV (oui/non) ;
- Caractérisation de l'opinion vis-à-vis du vaccin HPV (positive/négative/neutre) ;
- Mention d'un événement indésirable (oui/non) ;
- Relation causale entre vaccin et événement indésirable (oui/non).

Les messages considérés comme cas de pharmacovigilance potentiels après analyse manuelle ont été analysés par un expert en pharmacovigilance. Toutes les analyses statistiques ont été réalisées sur R en utilisant le package PhVID et l'implémentation du CPA proposé par Xu et al.

1.2. Cas d'usage pour les Topic Models

Le module 3 concerne toutes les analyses que nous souhaitons mener sur les données, nous avons donc exploré l'analyse en Topic Models en vue d'une intégration de celle-ci à ce module.

Rappelons que l'hypothèse de l'analyse en Topic Models, comme présenté dans l'état de l'art, est qu'un corpus est un regroupement de textes, qui sont eux-mêmes une collection de thématiques, et qu'une thématique est un ensemble de mots.

Il existe de très nombreux algorithmes. Le plus fréquemment utilisé est l'*Allocation de Dirichlet Latente (LDA)*. Il s'agit d'un modèle probabiliste, un algorithme d'apprentissage automatique non supervisé dont les classes de sorties ne sont pas étiquetées. C'est-à-dire que l'algorithme identifie les thèmes, les mots et les messages appartenant à ce thème, mais il n'est pas en mesure de définir ce thème en termes de sens. C'est un processus de catégorisation des mots et des textes.

Il sera nécessaire d'indiquer le nombre de thèmes à identifier (on peut choisir ce nombre de thèmes à l'aide de méthodes telles que la perplexité, qui correspond à une expression de la log-vraisemblance) puis l'algorithme ira partitionner les documents, et attribuera à chaque mot et chaque document une probabilité d'appartenance à chaque thème.

Cas d'usage : Analyse des mentions de vaccins contre la Covid-19

Nous avons appliqué cette méthode aux vaccins contre la COVID-19 [93]. Depuis décembre 2019 (jusqu'en août 2021), d'après l'Organisation Mondiale de la Santé (OMS) [94], on a observé 199 466 211 cas de COVID-19 parmi lesquels 4 244 541 patients sont décédés. Des vaccins ont été développés et plus de 3 milliards de doses administrées [94]. L'OMS a défini l'hésitation vaccinale comme l'une des 10 menaces majeures pour la santé mondiale en 2019 [95]. Le Web 2.0 peut influencer la décision de se faire vacciner [91] et les médias sociaux sont potentiellement intéressants pour explorer l'hésitation vaccinale. L'objectif est d'identifier les sujets et tendances à propos des vaccins contre la COVID-19 dans les forums de santé français pour comprendre la perception des utilisateurs de médias sociaux de santé de ces vaccins avant le début de la campagne vaccinale.

Le traitement préliminaire des données était le suivant :

- Conversion du texte en minuscules
- Identification des vaccins contre la COVID-19
- Suppression des ponctuations et des mots vides

- Suppression de certains termes spécifiques aux forums
- Suppression des messages traitants de la grippe
- Suppression des *tokens* trop fréquents ou trop rares
- Suppression des mots utilisés pour créer le corpus
- Transformation des espaces multiples en espaces simples

La période d'étude est du 1^{er} février 2020 au 31 janvier 2021

Nous avons créé une matrice terme-document (DTM) à partir des messages après traitement préliminaire :

- Cette matrice témoigne de la fréquence des termes dans la collection de messages
- Les lignes correspondent aux messages et les colonnes aux termes

Nous avons décidé de garder unigrammes et bigrammes afin de prendre en compte les séquences de deux mots (comme « effets indésirable », « mal tête ») ainsi que les mots seuls.

Nous avons mené une analyse en Topic Model avec allocation de Dirichlet latent et l'algorithme Variational Expectation-Maximization (VEM) [69].

Nous avons appliqué une validation croisée à 10 blocs, testant les modèles pour un nombre de topics allant de 2 à 10 [96].

Les données ont été divisées en 10 jeux de données tests avec le reste des messages considérés comme données d'entraînement.

- Le nombre de topics pour le modèle final a été sélectionné en fonction de la perplexité minimum.
- Nous avons initialisé le paramètre alpha du modèle final avec la moyenne des alphas générés pour ce nombre de topics.

Pour les topics que nous avons trouvés potentiellement intéressants pour notre étude, nous avons ensuite tiré au sort 20 messages de moins de 500 mots pour une analyse manuelle.

2. Création du module 4 : l'interface utilisateur

2.1. Conception

L'interface a été développée de manière itérative :

1. Tout d'abord, les utilisateurs potentiels de l'interface ont été définis (l'ANSM, les CRPVs, les entreprises pharmaceutiques et les équipes de recherche), et leurs besoins recueillis puis comparés à l'existant (cf. partie état de l'art).
2. Ensuite, j'ai réalisé une maquette via Marvel app (plateforme internet pour réaliser des maquettes d'applications web, Figure 10) et modélisé les fonctionnalités de l'interface selon la méthode UML (Langage de Modélisation Unifié) [97,98] (diagrammes de cas d'utilisation, de classes et de séquence, Figures 11-14).
3. J'ai rédigé le cahier des charges et le guide utilisateur.
4. Choix technologique pour l'implémentation.

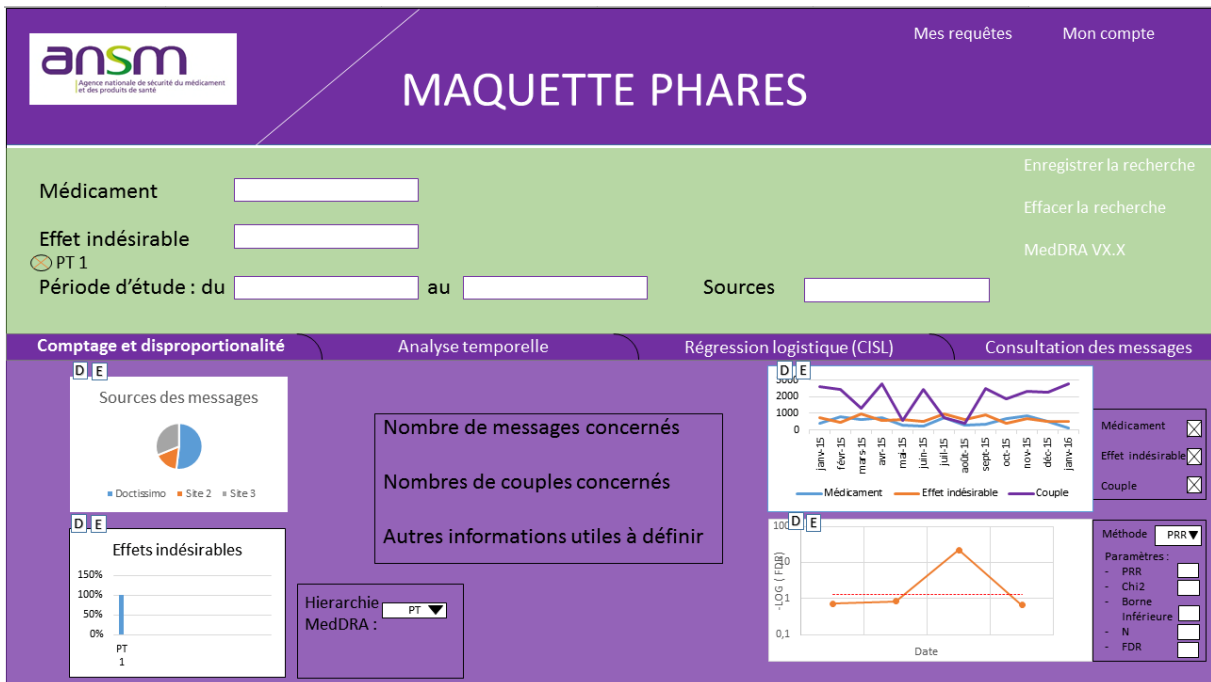


Figure 10 : Capture d'écran de la maquette

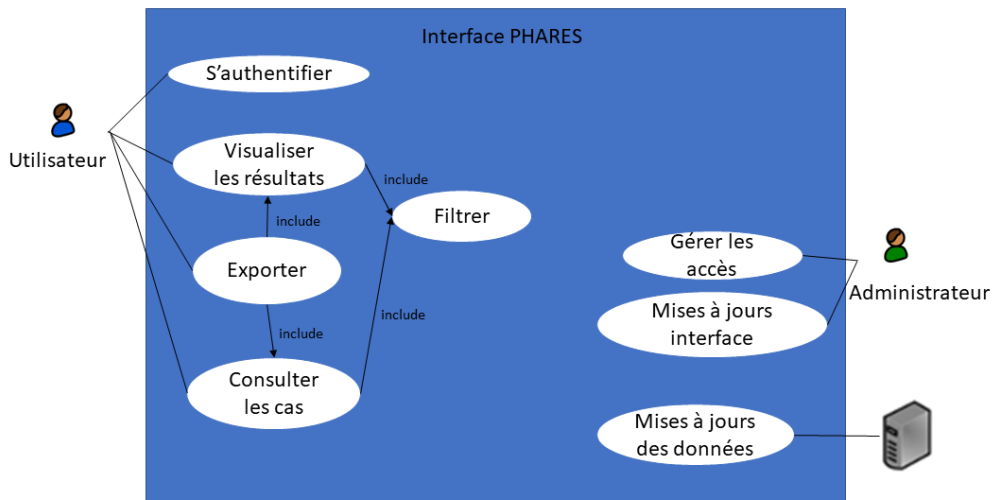


Figure 11 : Diagramme UML - Cas d'utilisation

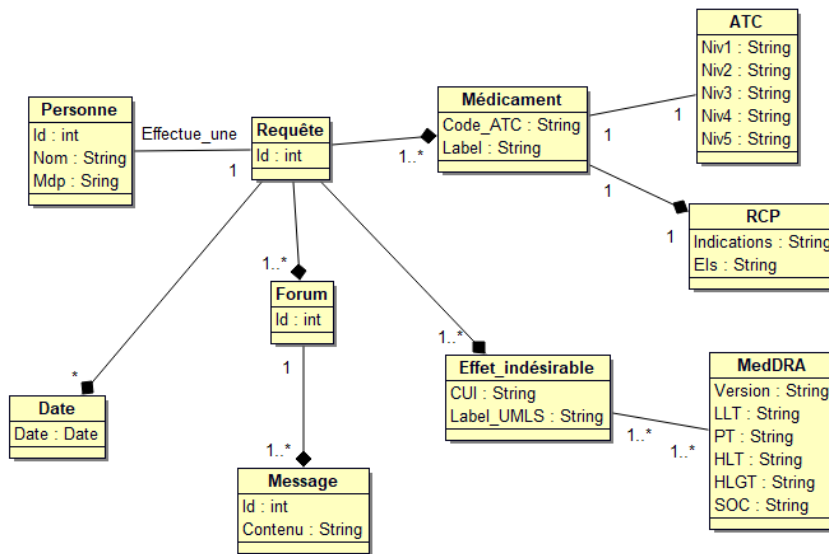


Figure 12 : Diagramme UML – Classes

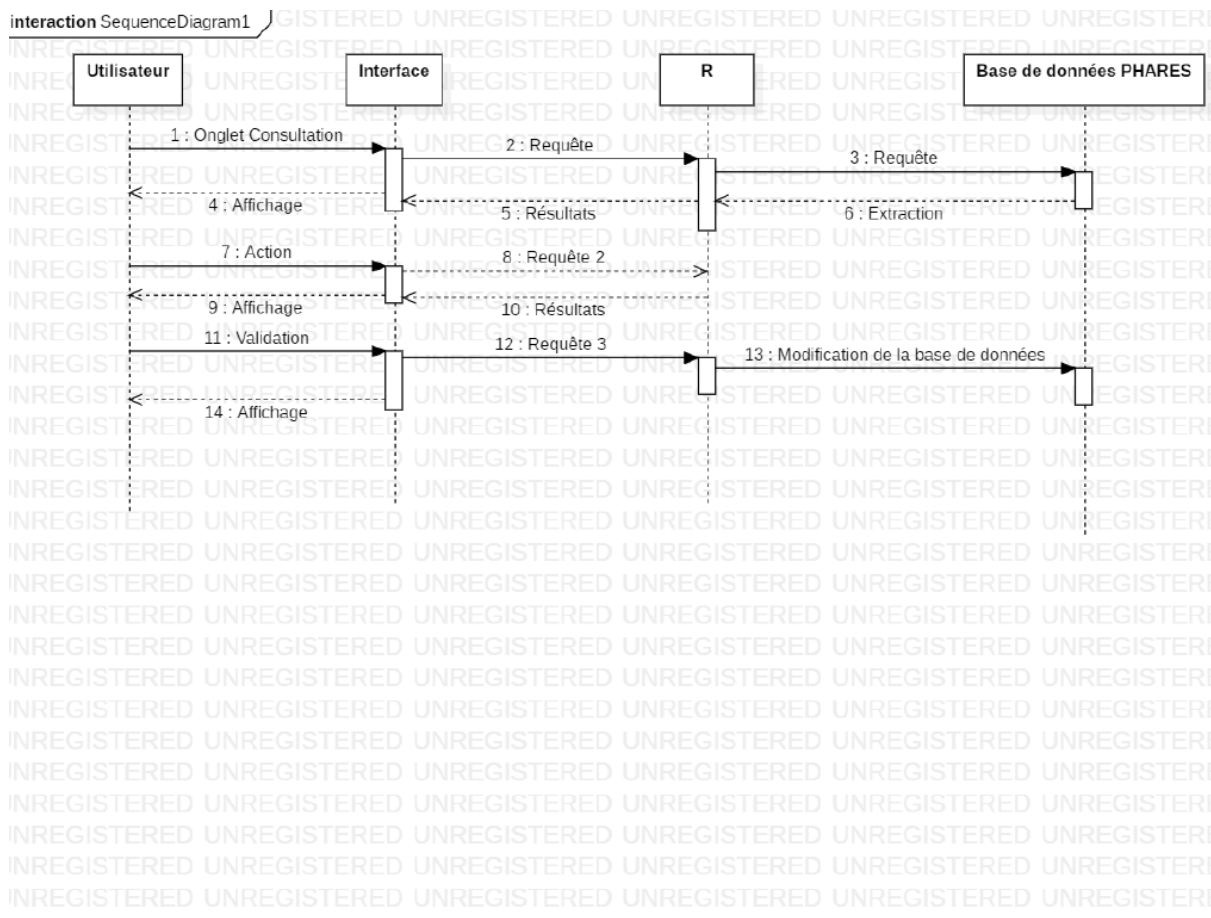


Figure 13 : Diagramme UML – Séquence 1

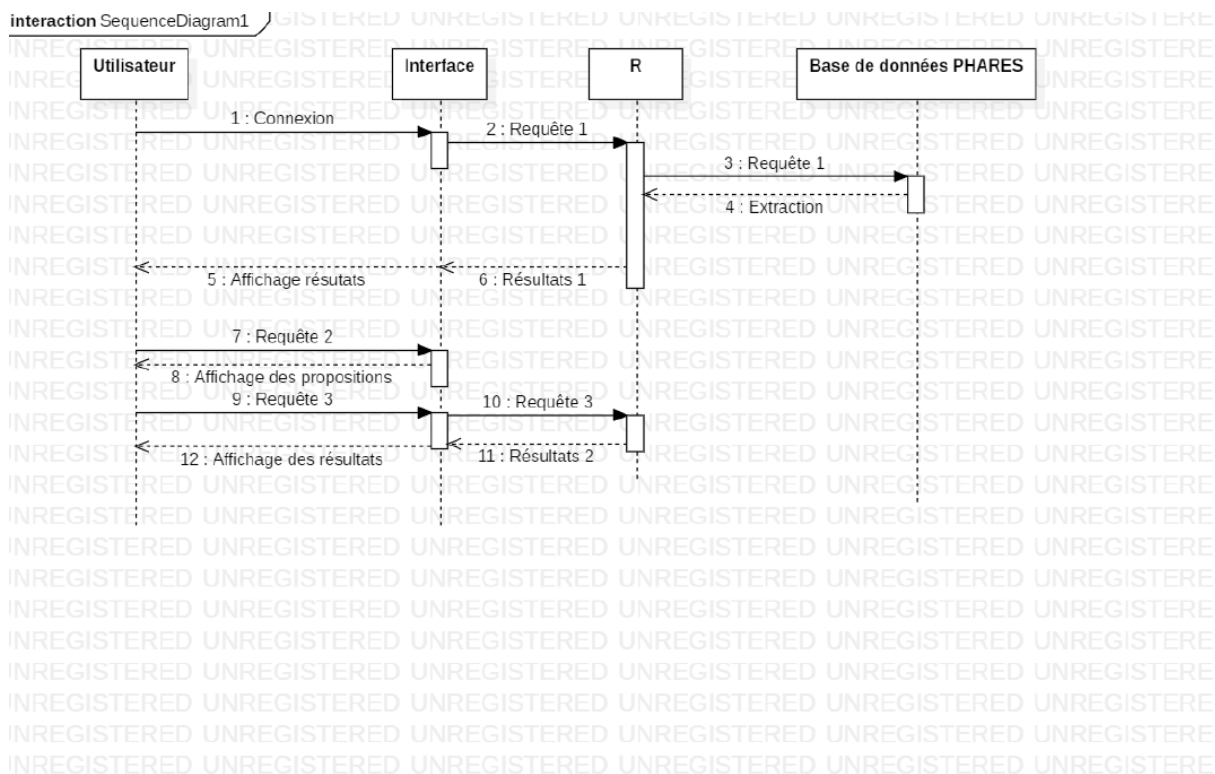


Figure 14 : Diagramme UML – Séquence 2

Le diagramme de cas d'utilisation (Figure 11) représente ce que l'utilisateur peut faire avec l'interface : s'y connecter, filtrer pour visualiser les résultats et consulter les cas ainsi que les exporter. La mise à jour des données est automatique, le scraper et l'annotateur fonctionnant en routine.

Le diagramme de classe (Figure 12) représente les classes (acteurs), leurs attributs et leurs relations. Par exemple, un utilisateur va effectuer une requête qui concernera la date, les médicaments, les effets indésirables et les messages (issus de forums).

Les diagrammes de séquence (Figures 13 et 14) sont des représentations graphiques des interactions entre les acteurs. La figure 13 représente la consultation via une requête exécutée via l'interface puis la modification de la base de données décrite Figure 13 via l'interface qui avait alors été envisagée (cela reste une perspective à moyen terme pour notre chaîne de traitement). La Figure 14 représente la connexion de l'utilisateur ainsi que l'application d'une requête via l'interface et l'affichage des résultats.

2.2. Choix technologique pour l'implémentation

Cette interface a été implémentée avec le langage et le logiciel R en utilisant le package Shiny (<https://cran.r-project.org/web/packages/shiny/shiny.pdf>) avec les contributions successives d'Akram Redjda et de la société de service Stat4Decision. En effet, le module statistique créé avant l'interface utilisait le langage R. Nous avons décidé d'utiliser ce même langage pour implémenter l'interface. L'administrateur gère les accès et les mises à jour de la chaîne de traitement.

Chapitre 5

Résultats

Dans ce chapitre, nous verrons les différents résultats obtenus au cours de cette thèse.

1. Module d'analyse : Détection du signal

Tout d'abord, je vais vous présenter les résultats obtenus pour répondre au premier objectif qui est de définir un cadre pour l'application des algorithmes de détection du signal aux données issues des médias sociaux de manière opérationnelle et pour une utilisation quotidienne. Pour cela, nous nous sommes intéressés à deux cas d'usages : l'un portant sur le Levothyrox, et l'autre sur les vaccins HPV.

1.1. Cas d'usage du Levothyrox

11 340 messages contenaient une annotation concernant le Levothyrox, pour un total de 50 127 annotations de ce médicament. Ces messages et annotations provenaient principalement du forum *Vivre sans thyroïde* et avaient été postés en milieu d'années 2017 (Figure 13, Tableau 7).

Les effets indésirables annotés avec le Levothyrox provenaient principalement des SOCs « Troubles généraux et anomalies au site d'administration » (29,6%), « Troubles du métabolisme et de la nutrition » (11,6%) et « Affections endocriniennes » (11,4%) (Figure 14). Les 20 PTs les plus souvent annotés avec le Levothyrox sont listés dans le Tableau 8.

Tableau 7 : Annotations de Levothyrox par forum

Forum	N	% cumulé
Vivre sans thyroïde	41211	82,21
Doctissimo Santé	4230	90,65
Doctissimo Grossesse	1476	93,60
Doctissimo Nutrition	1177	95,94
Carenity	863	97,67
Allo docteurs	502	98,67
Atoute	170	99,01
Doctissimo médicaments	166	99,34
Que choisir	85	99,51
Maladies rares	76	99,66
Au féminin	58	99,77
Sport santé	50	99,87
Onmeda	48	99,97
Famili	7	99,98
Futura	5	99,99
Maman vie	2	100,00
Magic maman	1	100,00

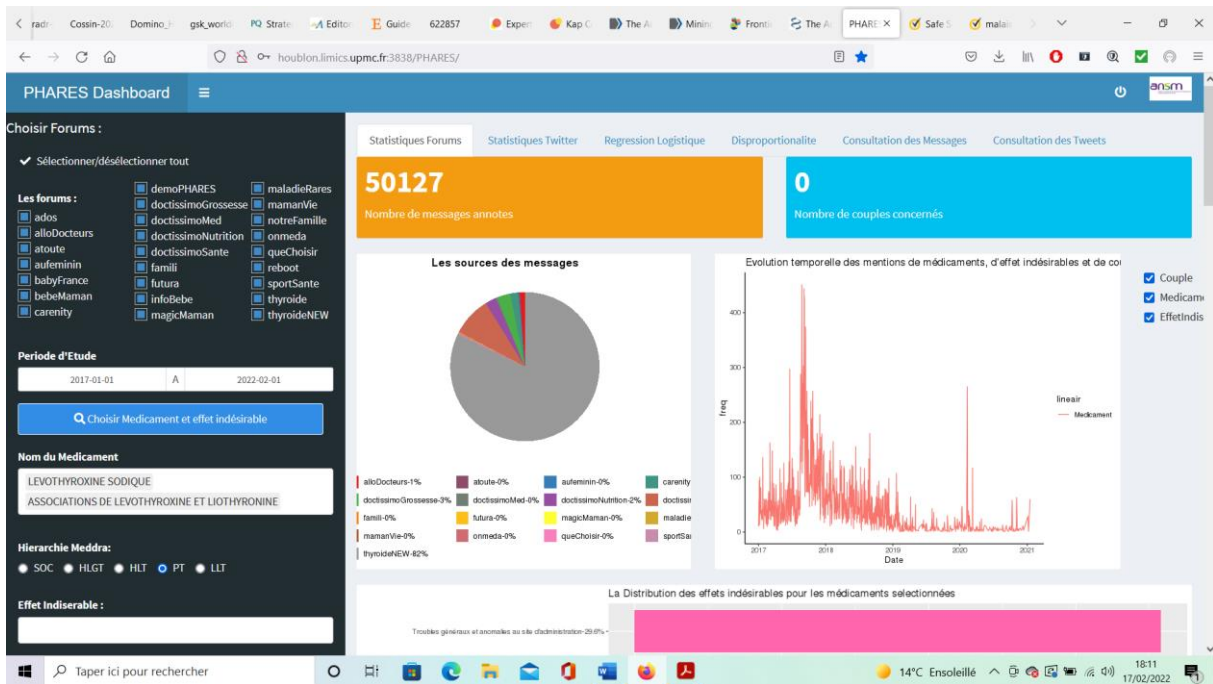


Figure 13 : Sources et évolution temporelle des messages concernant le levothyrox

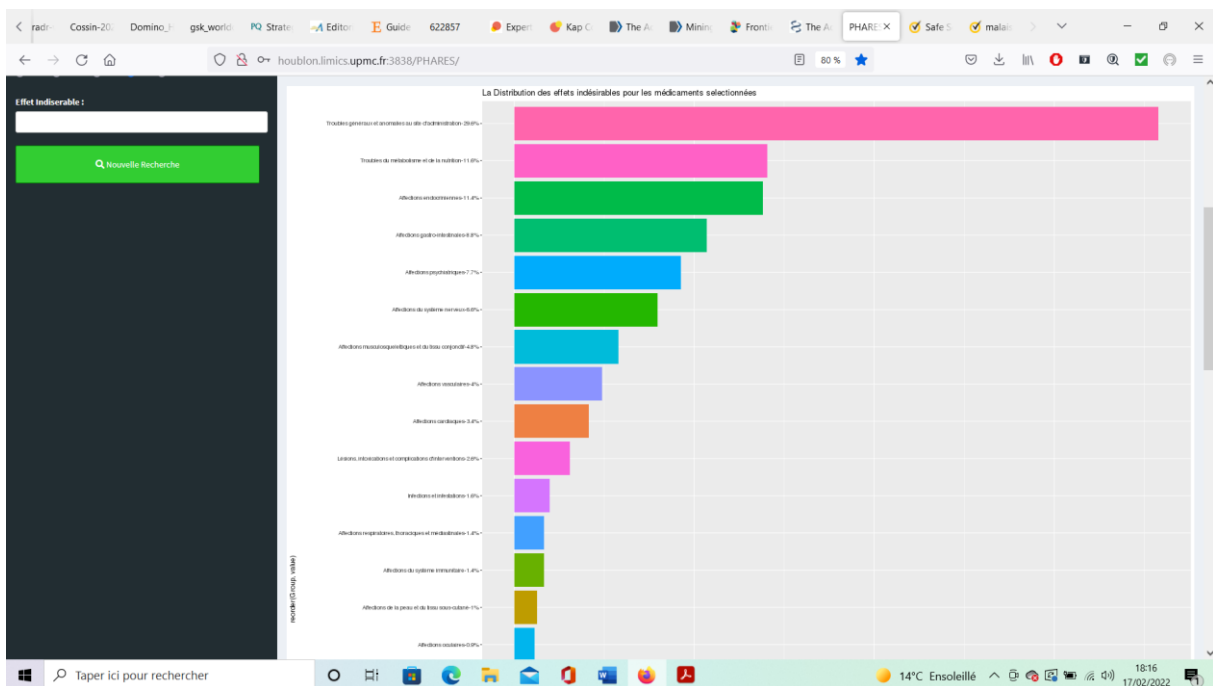


Figure 14 : Distribution des SOC des EIs retrouvés avec le levothyrox

Tableau 8 : 20 PTs les plus retrouvés avec le levothyrox

PT	n
Douleur	1882
Fatigue	1841
Malaise	1267
Hypothyroïdie	1110
Vertiges	912
Insomnie	627
Palpitations	571
Hyperthyroïdie	568
Tumeur maligne	560
Anxiété	498
Surdose	490
Tension nerveuse	484
Myalgie	409
Nausée	388
Stress	380
Diarrhée	354
Tachycardie	322
Spasmes musculaires	321
Convulsions	302
Arthralgie	276

Nous avons sélectionné le PT “Fatigue” pour la détection du signal. 85 976 messages étaient annotés avec le Levothyrox ou l’EI “Fatigue”, et le couple a été annoté 1 841 fois principalement en 2017 (Figure 15).

La Figure 16 montre l’évolution temporelle du logarithme du PRR pour le couple d’intérêt. Un signal est généré lorsque la borne inférieure de l’intervalle de confiance du logarithme du PRR est supérieure à 0.

11 médicaments ont été retrouvés comme signaux avec le PT « fatigue » en appliquant le CISL : Paclitaxel, Pegfilgrastim, Levothyrox®, Glatiramere Acetate, Escitalopram, Sulfate Ferreux, Association de Levothyrox® et Liothyronine, Secukinumab, Methotrexate et Bismuth Potassique, Tetracycline et Metronidazole (Figure 17).

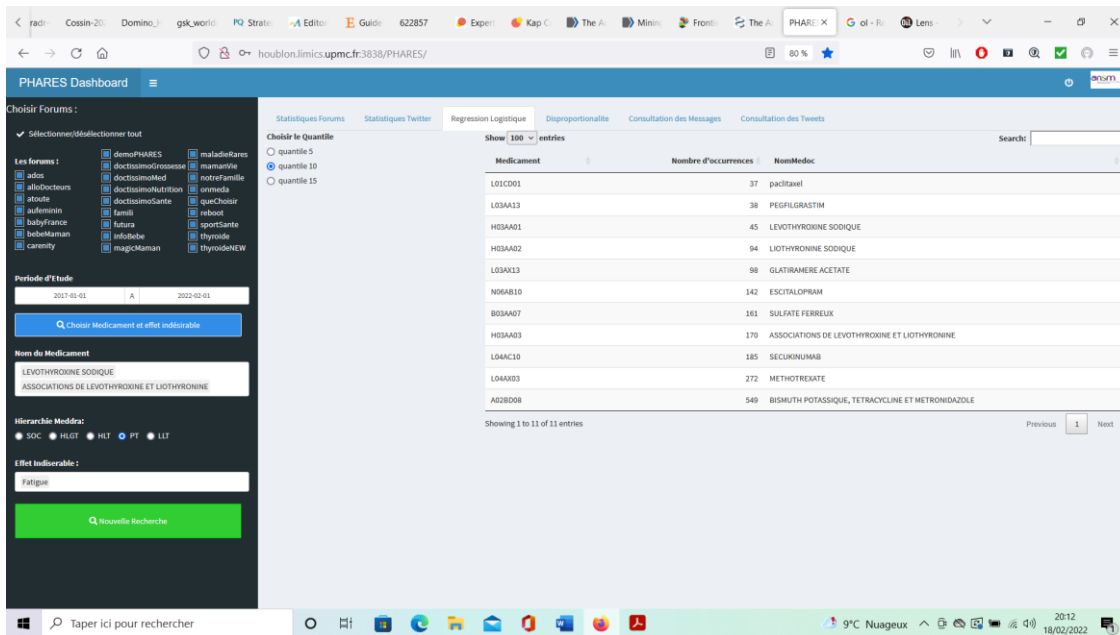


Figure 17 : Résultats du CISL pour l'EI fatigue

Le CPA a été appliqué sur l'évolution mensuelle du nombre de couples retrouvés avec le Levothyrox détecté dans les forums. Six points de rupture ont été détectés (Figure 18) et trois correspondaient à une augmentation du nombre de couples : en août 2017, et en septembre et décembre 2018.

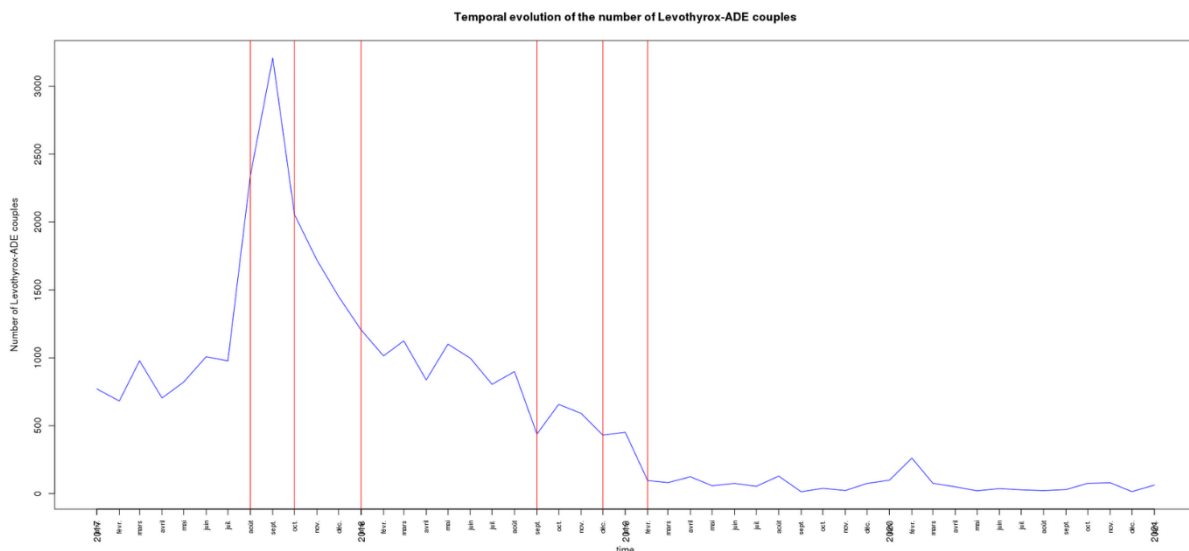


Figure 18 : Points de rupture identifiés par le CPA pour l'évolution temporelle du nombre de mentions du couple Levothyrox-fatigue

Ce cas d'usage montre que, dans le contexte du Levothyrox, les résultats obtenus par le module 3 (via le module 4) sont cohérents avec ceux retrouvés dans la littérature utilisant des sources de données plus traditionnelles en pharmacovigilance (voir le chapitre Discussion). Cela démontre le potentiel de l'utilisation de la chaîne de traitement pour la surveillance d'un médicament, non seulement de

manière rétrospective, mais aussi en temps réel en utilisant les données issues de médias sociaux. Des signaux peuvent potentiellement être générés via l'utilisation de cette chaîne de traitement.

1.2. Cas d'usage des Vaccins HPV

On a retrouvé 18 800 526 annotations (de médicaments, effets indésirables ou de couples) dans 3 482 010 messages postés entre le 14/06/1999 et le 13/04/2019 (770 annotations sans date) sur 21 forums de discussion. Après l'étape de préparation des données, la base était constituée de 165 663 couples {médicament, effet indésirable} différents pour un total de 2 523 564 couples concernant 1 561 médicaments et 3 978 effets indésirables annotés dans des messages postés entre décembre 1999 et avril 2019. Le vaccin HPV a été annoté dans 5 159 couples, avec 501 effets indésirables différents.

Quatre points de ruptures ont été identifiés avec un niveau de confiance de 95% en appliquant le CPA au nombre de couples retrouvés avec le vaccin : octobre 2007 (100%), août 2009 (98%), janvier 2012 (100%), et novembre 2015 (99.9%). Seulement un d'entre eux correspondait à une augmentation (octobre 2007). 124 signaux ont été générés en appliquant le BCPNN, 3 561 messages étaient concernés par ces signaux. Aucun point de rupture n'a été identifié dans l'évolution temporelle du PRR avec un niveau de confiance de 95% pour les 5 EIs sélectionnés (Figure 19).

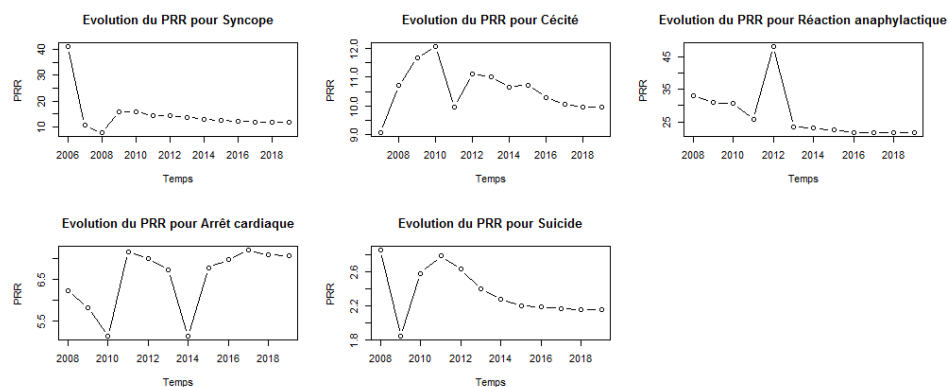


Figure 19 : Evolution du PRR pour les 5 EIs sélectionnés.

Sur les 150 messages tirés au sort, 138 ont été revus manuellement, 11 étaient des doublons et 1 correspondait à une image (non visible). La mention du vaccin HPV, un effet indésirable et d'une relation entre le vaccin et un EI étaient respectivement présents dans 126 (91,3%), 112 (81,2%) et 40 (29,0%) messages.

Parmi les messages où une relation a été identifiée, 4 ont été considérés comme des cas de pharmacovigilance potentiels. Les autres messages rapportaient des EIs observés dans des études à l'étranger. Deux messages correspondaient à des expériences personnelles d'EI potentiellement lié au vaccin : le premier rapportait une embolie pulmonaire à la suite du vaccin avec une temporalité

cohérente, et le second une expérience de plusieurs Els sans description complète. Parmi les 138 messages revus manuellement, 63 (46%) étaient neutres vis-à-vis du vaccin, 56 (40%) décrivaient une opinion négative et seulement 19 (14%) représentaient une opinion positive.

2. Module d'analyse : Topic Models

Sur les 66 176 messages postés entre le 1er février 2020 et le 31 janvier 2021 :

- 704 ont été identifiés comme concernant les vaccins contre le COVID-19
- 211 ont été exclus car ils concernaient également la grippe
- 8 ont été exclus après traitement (suppression de tous les termes des messages)
- Un total de 485 messages ont été inclus dans l'étude

La perplexité minimum en moyenne a été obtenue pour 10 topics (Figure 20). Nous avons donc choisi ce nombre de topics pour la suite de l'étude.

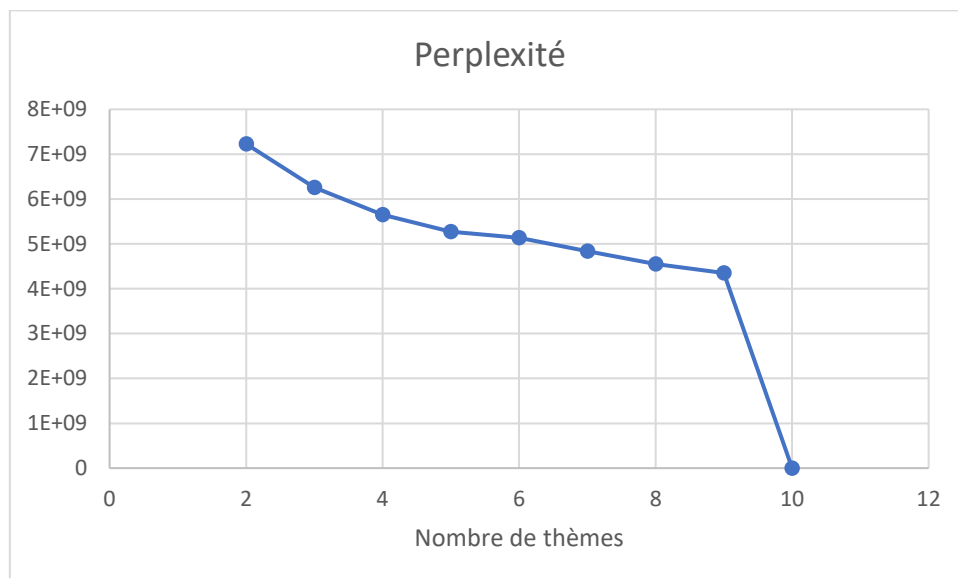


Figure 20 : Sélection du nombre de thèmes via la perplexité

Pour les 4 topics que nous avons trouvés potentiellement intéressants pour notre étude (topics 3, 4, 5 et 6), nous avons ensuite tiré au sort 20 messages de moins de 500 mots pour une analyse manuelle. Le tableau 9 résume les thèmes abordés dans ces 4 topics.

Tableau 9 : Principaux thèmes abordés dans les topics revus manuellement

Topic	Principaux thèmes
3	Connivence big-pharma-gouvernement Campagne de vaccination et date de début La vaccination des enfants Passeport sanitaire Chloroquine
4	La pandémie en Europe (son début) La recherche pour un vaccin Le refus, la méfiance et l'hésitation vaccinale Responsabilité citoyenne
5	Chloroquine Espoir d'un vaccin Remdesivir Critique des politiciens, les effets indésirables de la chloroquine Les problèmes économiques
6	Accouchement et COVID-19 Quelle durée avant la disponibilité d'un vaccin Les enfants et le COVID-19 L'immunité collective Complot du gouvernement

3. Module 4 : L'interface

Les utilisateurs potentiels ont été définis : l'ANSM, les CRPVs, les entreprises pharmaceutiques et les équipes de recherche.

3.1. Implémentation du module de l'interface

L'interface est un tableau de bord séparé en deux parties. La colonne de gauche en noir (Figure 21) permet de contrôler les différents paramètres avant de lancer une requête (liste des forums, période temporelle, choix des EIs et des médicaments). La partie de droite est dédiée aux résultats de la requête avec différents onglets correspondant aux différentes analyses qu'il est possible d'effectuer.

Avant d'appliquer une requête (Figure 21) :

- Le premier ("Statistiques forums") et cinquième ("Consultation des messages") onglets donnent le nombre total de messages annotés et le nombre total d'annotations parmi les messages téléchargés pour l'analyse via l'interface,

- Le deuxième onglet (“Statistiques twitter”) donne le nombre total de tweets annotés et le nombre total d’annotations parmi les tweets téléchargés pour l’analyse via l’interface,
- Le troisième (“Regression logistique”), quatrième (“Disproportionalité”) et septième (Change-point”) onglets ne sont pas disponibles sans lancer une requête car ils permettent la visualisation de statistiques calculées pour une sélection de médicament(s) et/ou d’EI(s), et le sixième onglet (“Consultation des tweets”) donne le nombre total de tweets annotés téléchargés pour l’analyse via l’interface.

L'interface permet de sélectionner différents paramètres pour lancer une requête. Ces paramètres sont : le choix des forums, de la période temporelle, du (des) médicament(s) selon la classification ATC et le (les) EI(s) selon la classification MedDRA.

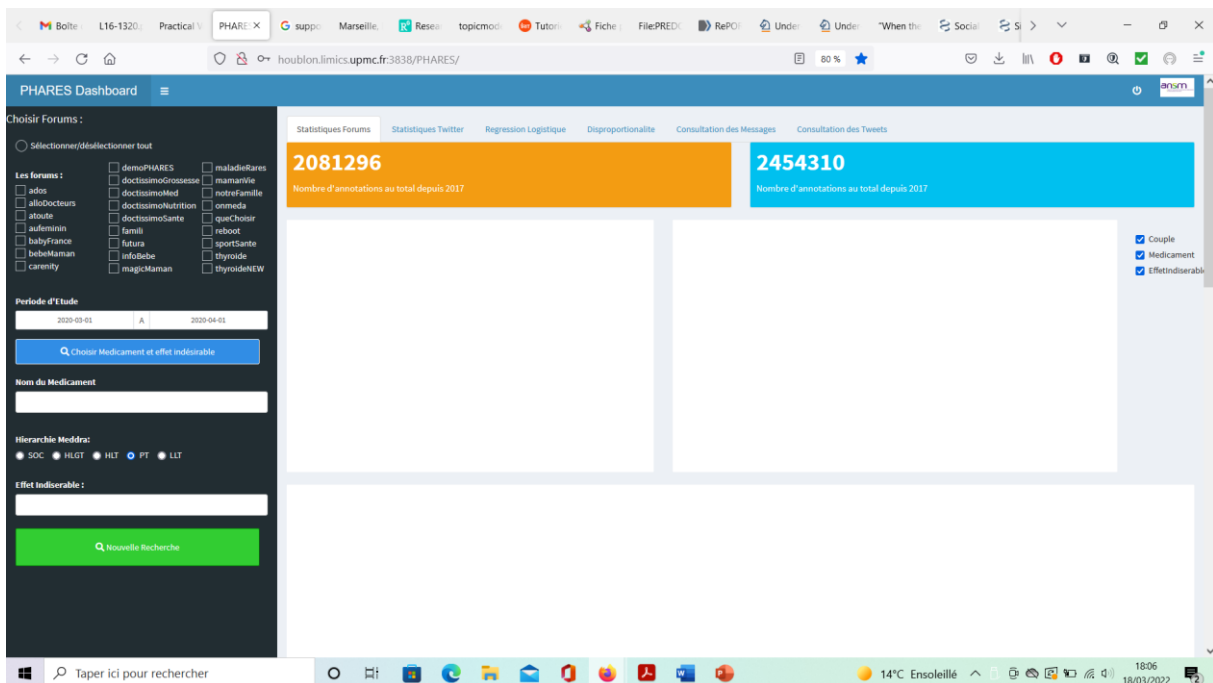


Figure 21 : Interface avant une requête

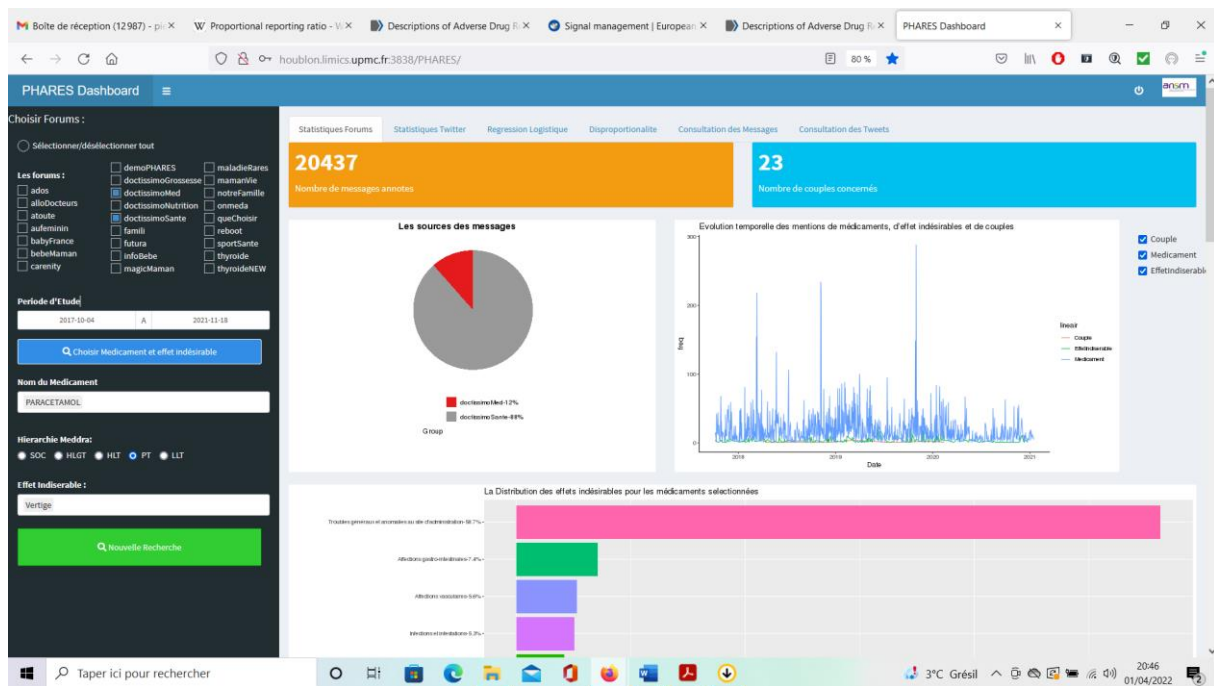


Figure 22 : Onglet "Statistiques forums" après une requête

Après avoir lancé une requête :

- Les onglets "Statistiques forums" et "Statistiques twitter" fournissent des statistiques descriptives, des diagrammes et graphique de l'évolution temporelle des annotations du médicament, de l'EI et/ou du couple sélectionné pour la requête (Figure 22).
- L'onglet "Régression logistique" permet de choisir un dernier paramètre spécifique (α), puis permet de visualiser les médicaments sélectionnés par la méthode.
- L'onglet "Disproportionnalité" permet de choisir la méthode entre le PRR et le ROR, puis l'évolution temporelle du logarithme de la statistique choisie et de son intervalle de confiance à 95% sont visualisables.
- L'onglet "Change-point" permet de choisir de réaliser l'analyse sur l'évolution temporelle du nombre d'annotations de couple(s), médicament(s) ou d'EI(s). La courbe correspondante est complétée par les différents points auxquels la valeur de la moyenne change de façon significative.
- Les onglets "Consultation des messages" et "Consultation des tweets" donnent le nombre total de messages/tweets annotés, le nombre total d'annotations du (des) médicament(s) et/ou EI(s) sélectionnés. Ils permettent aussi de visualiser et de télécharger une table avec, pour les forums : le forum, le message, le nom du topic, la date et heure de publication et les éléments annotés ; et pour les tweets : le tweet, la date et heure de publication et les éléments annotés.

Lorsqu'une requête est lancée, les données de la requête (les forums, la période, le(s) médicament(s) et le(s) EI(s)) sont envoyées au module statistique qui réalise les calculs et envoie les résultats au module de restitution.

Les onglets sont mis à jour dès que l'utilisateur modifie la requête. Si plusieurs médicaments ou EIs sont sélectionnés, l'analyse considérera l'union des médicaments comme un médicament ou l'union des EIs comme un EI.

Pour les algorithmes de disproportionnalité, des problèmes ont été notés lors de l'utilisation du package PhVid. Les algorithmes ont donc été implémentés manuellement.

3.2. Echanges avec les utilisateurs

Plusieurs allers-retours ont été réalisés avec les utilisateurs afin de faire remonter des bogues et problèmes.

Par exemple, les problèmes suivants ont été repérés et résolus :

1. Cohérence des labels des « nombre de messages annotés », « nombre de couples concernés » indiqués dans la page « statistique forum » et les labels « Nombre de messages annotés » et « Messages annotés, dont X annotations avec les médicaments sélectionnés »

Remplacer par :

- Nombre de messages avec le(s) médicaments
 - Nombre de messages avec le(s) effet(s) indésirables
 - Nombre de messages avec le(s) couple(s) sélectionnés
 - Nombre de couples avec le(s) médicament(s)
 - Nombre de couples avec le(s) effet(s) indésirable(s)
 -
2. Cohérence de la Régression logistique

Exemple avec le clomifène et l'infertilité : 554 couples infertilité-clomifène identifiés (Figure 23) mais seulement 4 messages concernés par le couple (Figure 24).

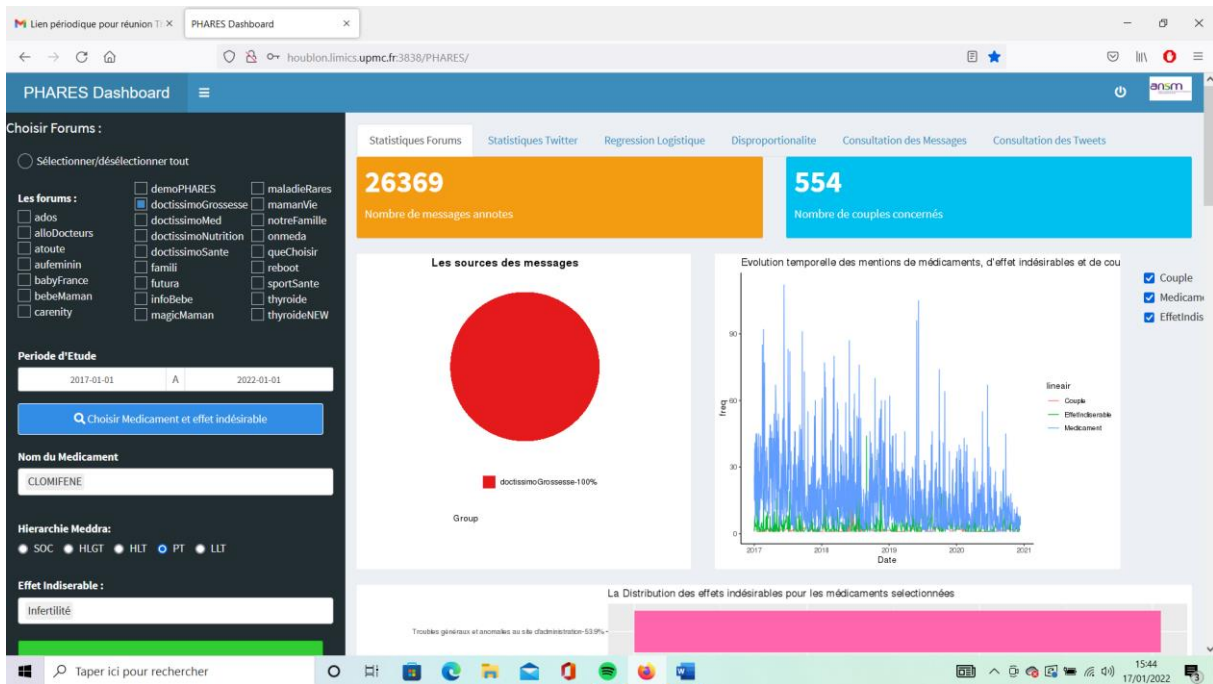


Figure 23 : Illustration du problème 1

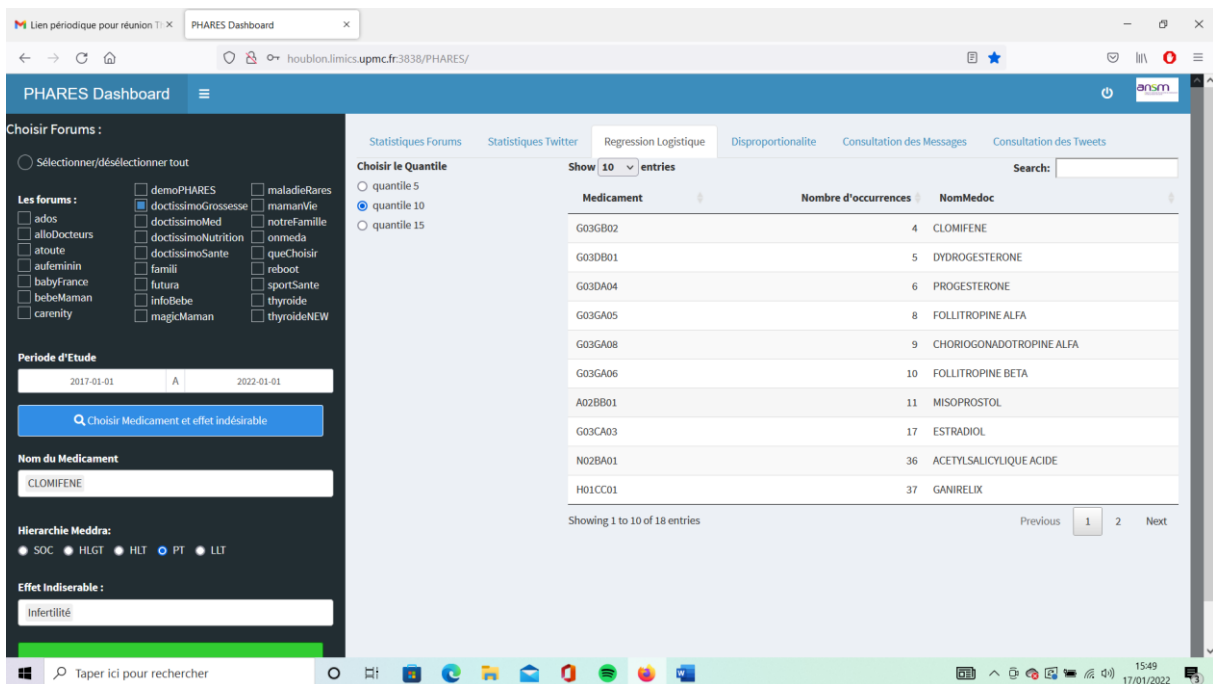


Figure 24 : Illustration du problème 2

- Il manque la distribution des PTs retrouvés avec le médicament d'intérêt ainsi que la distribution des médicaments retrouvés avec l'effet indésirable d'intérêt dans la page statistique forums.

4. Sur la page « consultation des messages », le téléchargement des messages résulte en un fichier Excel avec une seule case renseignée avec les informations suivantes « „NomForum", "Message", "ObjetMessage", "Date", "Heure", "range" ».

5. PRR, ROR

L'évolution temporelle du log de la statistique d'intérêt ou de la statistique d'intérêt présente encore des limites, par exemple sur une évolution du 12/2017 au 02/2018, l'évolution est tracée avec la valeur de la statistique aux dates : 12/2017, puis 12/2018, suivie du 01/2017, puis du 01/2018, etc... (Figure 25).

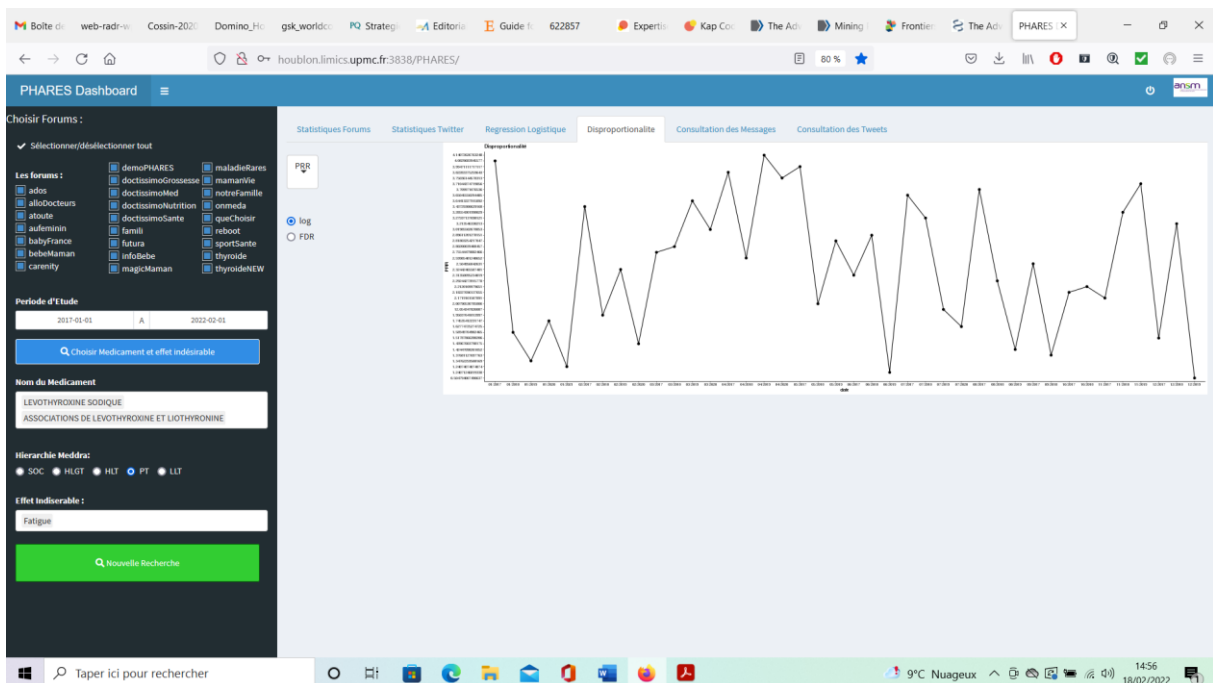


Figure 25 : Illustration du problème 5

De plus, il faudrait ajouter l'évolution des bornes inférieures et supérieures de l'intervalle de confiance à 95% de la statistique. Et il serait bien de pouvoir visualiser une ligne rouge représentant le seuil à partir duquel un signal est généré.

6. Ajout du Change Point Analysis

Chapitre 6

Discussion

1. Module d'analyse : Détection du signal

1.1. Cas d'usage du Levothyrox

Les principaux Els associés au Levothyrox dans notre étude sont consistants avec ceux trouvés par Audeh et al. sur des données similaires, mais sans utiliser l'interface [99]. Parmi les 1 554 notifications de patients entre le 1er Janvier 2017 et le 31 décembre 2017 au centre régional de pharmacovigilance de Nice, 1 372 concernaient la nouvelle formule du Levothyrox (pour 7 342 Els) [87], et les 10 Els les plus fréquemment rapportés étaient similaires aux Els que nous avons retrouvés. Ces 10 Els étaient : l'asthénie, les maux de tête, les vertiges, la perte de cheveux, l'insomnie, les crampes, le gain de poids, les nausées, les douleurs musculaires et l'irritabilité. Nos résultats sont donc cohérents avec la littérature.

11 340 messages contenaient une annotation concernant le Levothyrox. La plupart ont été publiés en 2017, l'année de la "crise du Levothyrox". Parmi ces messages, 1 841 couples Levothyrox-fatigue ont été retrouvés et le signal de pharmacovigilance a été généré par la méthode de régression logistique ainsi que le PRR, bien que cet EI ne figure pas sur le RCP du Levothyrox. Il est donc aussi possible de trouver des Els inattendus via la chaîne de traitement mise en place.

De plus, une augmentation significative du nombre d'Els retrouvés avec le Levothyrox a été détectée en août 2017, quelques mois après la commercialisation de la nouvelle formule (en mars 2017) en utilisant le CPA. Cette augmentation a été détectée après la première déclaration aux CRPVs et après une pétition lancée par les patients pour demander la disponibilité de l'ancienne formule (en juin 2017). Ainsi, il est possible de retrouver des dates d'importants événements concernant un médicament, mais la détection des signaux n'est pas plus précoce avec les médias sociaux qu'avec le circuit institutionnel de pharmacovigilance.

1.2. Cas d'usage des vaccins HPV

Dans notre étude, 29% des messages revus manuellement décrivaient un EI potentiel des vaccins contre le papillomavirus humain, et seulement 5% concernaient une expérience personnelle. Les principales limites de notre étude étaient qu'elle se concentrait uniquement sur des forums français et que seul un petit nombre de messages ont été évalués. Malgré ces limites, nous avons observé que les patients discutaient le plus souvent de faits lus sur d'autres médias plutôt qu'ils ne décrivaient des

expériences d'EI personnelles ou par des proches. Seul 14% des messages évalués correspondaient à une bonne opinion de ces vaccins, mais ce chiffre peut être sous-estimé car seuls des messages contenant à la fois une annotation de ces vaccins et d'un EI ont été évalués. On retrouvait significativement moins d'expériences personnelles d'EI dans les médias sociaux avec les vaccins HPV que pour d'autres médicaments mentionnés dans la littérature. Les messages étaient principalement l'illustration de croyances vis-à-vis de ces vaccins qu'ils sont la cause de graves EIs, croyances gagnant en popularité [100,101]. Les autorités de santé peuvent s'appuyer sur ce type d'observations pour communiquer sur les vaccins, notamment ceux contre le papillomavirus humain : connaître les « arguments » contre la vaccination permet d'y répondre plus efficacement.

2. Module d'analyse : Topic Models

Nous avons identifié 10 thèmes principaux et analysé manuellement 120 messages issus de 6 de ces 10 thèmes. Dans de futures analyses, les messages du quatrième thème devraient faire l'objet d'une analyse approfondie pour en apprendre davantage sur l'hésitation vaccinale, le refus et la méfiance de ce vaccin. Les troisièmes et cinquième thèmes pourraient eux aussi être explorés afin de connaître l'opinion des utilisateurs sur le gouvernement, la politique et les choix économiques menés face à cette épidémie.

Une des limites de notre étude était que les utilisateurs de médias sociaux peuvent inventer des néologismes ou bien orthographier des mots de différentes manières. Ainsi des mots pourtant identiques dans l'intention de communiquer des utilisateurs sont considérés comme différents par la méthodologie que nous avons suivie. Nous avons essayé la racinisation, mais les résultats devenaient très difficilement interprétables pour l'évaluateur. Une solution serait d'explorer la lemmatisation, qui permet au mot de garder un sens sémantique tout en éliminant par exemple le genre ou le pluriel. De plus, une analyse en topic models dynamiques pourrait être envisagée afin d'évaluer l'évolution dans le temps de ces thèmes.

3. Discussion : chaîne de traitement

Nous avons construit une chaîne de traitement pour analyser les mentions d'effets indésirables liées aux médicaments dans les forums de discussion. Cette chaîne de traitement inclut des étapes allant de l'extraction des messages à leur étude quantitative à travers une interface utilisateur. Nous avons montré qu'il était possible, en utilisant cette chaîne de traitement, de retrouver des informations similaires à la littérature ainsi que des EIs inattendus et des dates d'événements médiatiques importants concernant des médicaments, ce qui montre l'intérêt d'une telle chaîne de traitement.

Cette chaîne de traitement prend en compte un plus grand nombre d'exigences par rapport aux autres chaînes de traitement décrites dans l'état de l'art, et répond à tous les besoins que nous avons définis au moment de notre étape de spécifications (Tableau 3, page 44). Si la plupart de ces chaînes de traitement se concentraient sur l'étude des EIs liés aux médicaments (8 des 10 chaînes de traitement), seulement 4 étaient destinées à une utilisation en routine. Les scrapers utilisés dans ces chaînes de traitement réalisaient rarement la dé-identification (3/10), 6 extrayaient la totalité des messages, et 6 des messages issus de forums de discussion. L'annotateur fonctionnait avec le français pour 4 d'entre eux. Le CPA n'était jamais disponible dans ces 10 chaînes de traitement, tout comme l'évolution temporelle du PRR et une méthode de détection de signal basée sur la régression logistique. L'évolution temporelle du nombre de messages ainsi qu'une méthode de détection du signal étaient disponibles dans 5 des chaînes de traitement.

Des travaux ultérieurs pourront répondre à plusieurs limites de notre chaîne de traitement. Le paramétrage de Vigi4Med scraper dépend de la structure HTML des forums. Par conséquent, les fichiers de configuration doivent être mis à jour dès qu'un forum change de design. De plus, notre interface ne permet pas l'ajout de mots pour l'identification de médicaments ou d'EIs, par exemple le terme « baclo » en référence au baclofène, comme il est souvent appelé sur les forums. Il en découle que le nombre de messages concernant un médicament ou un EI peut être sous-estimé par notre chaîne de traitement.

Les temps de calculs pouvant être longs (notamment pour la régression logistique), il est parfois utile de sélectionner uniquement les forums dans lesquels le médicament et/ou EI d'intérêt sont présents, avant de commencer une étude. Cela entraîne un biais pour l'application des méthodes de détection du signal, spécialement pour la disproportionnalité, parce que la proportion du couple d'intérêt est surestimée.

Une limite de l'analyse qualitative des messages via la chaîne de traitement est que l'utilisateur ne peut pas modifier les annotations ni enregistrer des données utiles pour la pharmacovigilance via des formulaires.

L'hypothèse faite pour l'application du CISL que chaque médicament identifié dans un message a été pris par l'utilisateur mène de façon évidente à un biais.

Une limite à l'utilisation des données issues des médias sociaux est liée aux publications frauduleuses. En effet, la couverture permise par l'utilisation de pseudonymes donne aux utilisateurs malveillants l'opportunité de répandre de fausses informations. De plus, un patient peut publier le même message ou bien un message similaire sur plusieurs sites ou même plusieurs fois sur le même site. Cela devrait, dans le futur, être pris en compte si l'on veut diminuer les biais lors des analyses.

Notre hypothèse était que la difficulté pour les agences réglementaires d'utiliser les médias sociaux en pharmacovigilance était principalement associée à l'absence d'un outil approprié sur lequel s'appuyer pour prendre des décisions. Nous avons donc implémenté une chaîne de traitement composée de 4 modules indépendants et évolutifs permettant d'étudier de manière efficace les données issues de forums français de discussion pour une utilisation en pharmacovigilance. Nous avons montré qu'utiliser cette chaîne de traitement, au moyen d'une interface utilisateur, permet de simplifier la réalisation d'une étude quantitative, sans besoin de codage informatique en langage R, que les résultats sont consistants avec ce qui est décrit dans la littérature, et qu'on peut même détecter des EI inattendus.

Nos perspectives à moyen/court termes sont de modifier le module d'annotation pour que la méthode utilisée soit à jour de l'état de l'art, améliorer l'analyse qualitative en permettant à l'utilisateur de modifier/corriger les annotations, et d'ajouter davantage de méthodes (détection du signal, topic models) au module d'analyse que celles jusqu'alors implémentées.

L'annotation des indications présentes dans les RCPs des médicaments ainsi que des contre-indications pourraient nous permettre d'identifier des mésusages possibles ou d'identifier plus facilement de nouveaux EIs [102]. Ainsi, un terme pathologique retrouvé sur les forums serait considéré comme indication si présent dans la liste des indications du RCP, comme un EI si présent dans la liste des EIs du RCP, ou bien comme un mésusage ou un EI inattendu s'il n'est pas retrouvé dans ces deux listes. Mais, considérer tous les termes présents dans la liste des indications du RCP comme des indications pourrait cacher des cas d'inefficacité médicamenteuse.

La chaîne de traitement n'ayant pas été évaluée par des personnels d'agences réglementaires sur le médicament, nous ne pouvons que partiellement confirmer notre hypothèse que l'interface utilisateur répond aux besoins des évaluateurs et décideurs qui souhaitent connaître le contenu des médias sociaux sans avoir à coder dans un langage informatique. D'autres études seront donc nécessaires, notamment une évaluation ergonomique de l'interface.

4. Discussion générale

Nos perspectives au long terme sont d'étendre cette chaîne de traitement à d'autres langues étrangères et à d'autres thèmes, comme le mésusage, les compléments alimentaires ou l'utilisation illégale de médicaments et de drogues, un forum français fréquemment utilisé existant pour ces utilisations récréatives (<https://www.psychosocial.org/forum/forums.php>).

De plus, la surveillance des réseaux sociaux pourrait potentiellement apporter de la connaissance vis-à-vis des croyances New-Age dangereuses, proches d'être des pratiques illégales de la médecine, tels

que le Heally (https://www.youtube.com/watch?v=4fT-AZeVn1g&ab_channel=GMilgram), le « patch quantique » (<https://patchquantique-france.fr/>), les cristaux du « physicien » Nassim Hamein (<https://fr.resonancescience.org/ark>) ou encore les chambres à Tachyon (https://www.youtube.com/watch?v=NtO_UkKWr40&t=438s&ab_channel=GMilgram).

Une limite de l'analyse de ce type de données est liée aux faux avis et commentaires. En effet, dans une étude menée publiée en 2014, la Direction Générale de la Concurrence, de la Consommation et de la Répression des Fraudes a montré que 45% des opinions de consommateurs sur internet étaient « anormales » en 2013, et a recensé de nombreuses pratiques discutables [101] :

- Suppression de partie ou de tout avis à tendance négative.
- Publication très rapide des avis positifs et publication différée d'avis négatifs, ainsi une majorité d'avis positifs apparaissent parmi les avis les plus récents.
- Rédaction de faux avis rédigés par les professionnels, leur entourage ou des prestataires spécialisés contre rémunération.

« Ces sous-traitants possèdent un savoir-faire (style rédactionnel, fausses identités) et des techniques informatiques permettant une certaine discrétion. Les faux avis sont alors déposés en grand nombre, sur des supports variés pour plus d'efficacité (sites d'avis, forums de discussion, réseaux sociaux, etc.). » On peut légitimement penser que de telles fraudes sont également possibles dans le milieu de la santé.

Notre chaîne de traitement pourrait être utilisée en veille sanitaire par des agences telles que l'ANSM ou par des entreprises pharmaceutiques afin de détecter d'éventuels problèmes, et par des équipes de recherche pour des analyses rétrospectives.

Chapitre 7

Conclusion

Afin de répondre à notre objectif qui était de définir un cadre pour l'application des méthodes de détection du signal aux données issues des médias sociaux dans un cadre opérationnel, nous avons implémenté une chaîne de traitement qui dispose d'un plus grand nombre de fonctionnalités pour répondre aux besoins que nous avons identifiés au moment de notre étape de spécifications (Tableau 3, page 44), que toutes les autres solutions proposées dans l'état de l'art. Cette chaîne de traitement permet d'analyser les mentions d'effets indésirables liés aux médicaments dans les forums de discussion. Elle inclut quatre modules indépendants et modifiables. Il est possible, en utilisant cette chaîne de traitement, de retrouver des informations similaires à la littérature ainsi que des Els inattendus et des dates d'événements médiatiques importants concernant des médicaments, ce qui montre son intérêt.

Bibliographie

- [1] Weltgesundheitsorganisation, Collaborating Centre for International Drug Monitoring, editors. The importance of pharmacovigilance: safety monitoring of medicinal products. Geneva: WHO [u.a.]; 2002.
- [2] van Puijenbroek EP, Bate A, Leufkens HGM, Lindquist M, Orre R, Egberts ACG. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiol Drug Saf* 2002;11:3–10. <https://doi.org/10.1002/pds.668>.
- [3] Trinh NTH, Solé E, Benkebil M. Benefits of combining change-point analysis with disproportionality analysis in pharmacovigilance signal detection. *Pharmacoepidemiol Drug Saf* 2019;28:370–6. <https://doi.org/10.1002/pds.4613>.
- [4] Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, et al. A Bayesian neural network method for adverse drug reaction signal generation. *Eur J Clin Pharmacol* 1998;54:315–21. <https://doi.org/10.1007/s002280050466>.
- [5] Caster O, Sandberg L, Bergvall T, Watson S, Norén GN. *vigiRank* for statistical signal detection in pharmacovigilance: First results from prospective real-world use. *Pharmacoepidemiol Drug Saf* 2017;26:1006–10. <https://doi.org/10.1002/pds.4247>.
- [6] DuMouchel W. Bayesian Data Mining in Large Frequency Tables, with an Application to the FDA Spontaneous Reporting System. *Am Stat* 1999;53:177–90. <https://doi.org/10.2307/2686093>.
- [7] Szarfman A, Tonning JM, Doraiswamy PM. Pharmacovigilance in the 21st century: new systematic tools for an old problem. *Pharmacotherapy* 2004;24:1099–104. <https://doi.org/10.1592/phco.24.13.1099.38090>.
- [8] Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol Drug Saf* 2001;10:483–6. <https://doi.org/10.1002/pds.677>.
- [9] van Puijenbroek EP, Egberts AC, Heerdink ER, Leufkens HG. Detecting drug-drug interactions using a database for spontaneous adverse drug reactions: an example with diuretics and non-steroidal anti-inflammatory drugs. *Eur J Clin Pharmacol* 2000;56:733–8. <https://doi.org/10.1007/s002280000215>.
- [10] Dangoumau J, Evreux JC, Jouglard J. [Method for determination of undesirable effects of drugs]. *Therapie* 1978;33:373–81.
- [11] Bégau B, Evreux JC, Jouglard J, Lagier G. [Imputation of the unexpected or toxic effects of drugs. Actualization of the method used in France]. *Therapie* 1985;40:111–8.
- [12] Hazell L, Shakir SAW. Under-Reporting of Adverse Drug Reactions. *Drug Saf* 2006;29:385–96. <https://doi.org/10.2165/00002018-200629050-00003>.
- [13] Menard J. Benfluorex: analysis of a drug-related public health crisis. *Diabetes Metab* 2011;37:169–75. <https://doi.org/10.1016/j.diabet.2011.04.001>.
- [14] Abadie D, Chebane L, Bert M, Durrieu G, Montastruc J-L. Online reporting of adverse drug reactions: a study from a French regional pharmacovigilance center. *Therapie* 2014;69:395–400. <https://doi.org/10.2515/therapie/2014035>.
- [15] Liu F, Jagannatha A, Yu H. Towards Drug Safety Surveillance and Pharmacovigilance: Current Progress in Detecting Medication and Adverse Drug Events from Electronic Health Records. *Drug Saf* 2019;42:95–7. <https://doi.org/10.1007/s40264-018-0766-8>.
- [16] Warrer P, Hansen EH, Juhl-Jensen L, Aagaard L. Using text-mining techniques in electronic patient records to identify ADRs from medicine use. *Br J Clin Pharmacol* 2012;73:674–84. <https://doi.org/10.1111/j.1365-2125.2011.04153.x>.
- [17] Black C, Tagiyeva-Milne N, Helms P, Moir D. Pharmacovigilance in children: detecting adverse drug reactions in routine electronic healthcare records. A systematic review: Pharmacovigilance in Children: a Systematic Review. *Br J Clin Pharmacol* 2015;80:844–54. <https://doi.org/10.1111/bcp.12645>.

- [18] Bannay A. Bases médico-administratives : de l'épidémiologie à la pharmacovigilance. phdthesis. Université de Lorraine, 2021.
- [19] Chee BW, Berlin R, Schatz B. Predicting adverse drug events from personal health messages. *AMIA. Annu. Symp. Proc.*, vol. 2011, American Medical Informatics Association; 2011, p. 217.
- [20] Golder S, Norman G, Loke YK. Systematic review on the prevalence, frequency and comparative value of adverse events data in social media: Systematic review of social media and adverse events. *Br J Clin Pharmacol* 2015;80:878–88. <https://doi.org/10.1111/bcp.12746>.
- [21] 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:e171. <https://doi.org/10.2196/jmir.4304>.
- [22] Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, et al. Utilizing social media data for pharmacovigilance: A review. *J Biomed Inform* 2015;54:202–12. <https://doi.org/10.1016/j.jbi.2015.02.004>.
- [23] Sloane R, Osanlou O, Lewis D, Bollegala D, Maskell S, Pirmohamed M. Social media and pharmacovigilance: A review of the opportunities and challenges: Social media and pharmacovigilance. *Br J Clin Pharmacol* 2015;80:910–20. <https://doi.org/10.1111/bcp.12717>.
- [24] Caster O, Dietrich J, Kürzinger M-L, Lerch M, Maskell S, Norén GN, et al. Assessment of the Utility of Social Media for Broad-Ranging Statistical Signal Detection in Pharmacovigilance: Results from the WEB-RADR Project. *Drug Saf* 2018;41:1355–69. <https://doi.org/10.1007/s40264-018-0699-2>.
- [25] Bousquet C, Audeh B, Bellet F, Lillo-Le Louët A. Comment on “Assessment of the Utility of Social Media for Broad-Ranging Statistical Signal Detection in Pharmacovigilance: Results from the WEB-RADR Project.” *Drug Saf* 2018;41:1371–3. <https://doi.org/10.1007/s40264-018-0747-y>.
- [26] Karapetiantz P, Bellet F, Audeh B, Lardon J, Leprovost D, Aboukhamis R, et al. Descriptions of Adverse Drug Reactions Are Less Informative in Forums Than in the French Pharmacovigilance Database but Provide More Unexpected Reactions. *Front Pharmacol* 2018;9:439. <https://doi.org/10.3389/fphar.2018.00439>.
- [27] Candore G, Juhlin K, Manlik K, Thakrar B, Quarcoo N, Seabroke S, et al. Comparison of statistical signal detection methods within and across spontaneous reporting databases. *Drug Saf* 2015;38:577–87. <https://doi.org/10.1007/s40264-015-0289-5>.
- [28] MedDRA. Wikipédia 2022.
- [29] Classification ATC. Wikipédia 2022.
- [30] Eholié S, Tapi Nzali MD, Bringay S, Jonquet C. MuEvo, un vocabulaire multi-expertise (patient/médecin) dédié au cancer du sein. IC Ingénierie Connaiss., Montpellier, France: 2016.
- [31] Tapi Nzali MD, Azé J, Bringay S, Lavergne C, Mollevi C, Optiz T. Reconciliation of patient/doctor vocabulary in a structured resource. *Health Informatics J* 2019;25:1219–31. <https://doi.org/10.1177/1460458217751014>.
- [32] Hong Y, Ehlers K, Gillis R, Patrick T, Zhang J. A Usability Study of Patient-friendly Terminology in an EMR System. *MEDINFO 2010* 2010:136–40. <https://doi.org/10.3233/978-1-60750-588-4-136>.
- [33] Fage-Butler AM, Nisbeth Jensen M. Medical terminology in online patient-patient communication: evidence of high health literacy? *Health Expect Int J Public Particip Health Care Health Policy* 2016;19:643–53. <https://doi.org/10.1111/hex.12395>.
- [34] Zeng Q, Kogan S, Ash N, Greenes RA, Boxwala AA. Characteristics of consumer terminology for health information retrieval. *Methods Inf Med* 2002;41:289–98.
- [35] Keselman A, Smith CA, Divita G, Kim H, Browne AC, Leroy G, et al. Consumer health concepts that do not map to the UMLS: where do they fit? *J Am Med Inform Assoc JAMIA* 2008;15:496–505. <https://doi.org/10.1197/jamia.M2599>.
- [36] Smith CA, Wicks PJ. PatientsLikeMe: Consumer Health Vocabulary as a Folksonomy. *AMIA Annu Symp Proc* 2008;2008:682–6.

- [37] 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:e37. <https://doi.org/10.2196/jmir.1636>.
- [38] Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system | *BMC Medical Informatics and Decision Making* | Full Text n.d. <https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-6-30> (accessed November 1, 2022).
- [39] Morlane-Hondère F, Grouin C, Zweigenbaum P. Identification of drug-related medical conditions in social media. *Proc. Tenth Int. Conf. Lang. Resour. Eval. LREC16*, 2016, p. 2022–8.
- [40] Lafferty J, McCallum A, Pereira FC. Conditional random fields: Probabilistic models for segmenting and labeling sequence data 2001.
- [41] Kuo C-H, Nevatia R. How does person identity recognition help multi-person tracking? *CVPR* 2011, 2011, p. 1217–24. <https://doi.org/10.1109/CVPR.2011.5995384>.
- [42] Magge A, Tutubalina E, Miftahutdinov Z, Alimova I, Dirkson A, Verberne S, et al. DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. *J Am Med Inform Assoc JAMIA* 2021;28:2184–92. <https://doi.org/10.1093/jamia/ocab114>.
- [43] Rezaei Z, Ebrahimpour-Komleh H, Eslami B, Chavoshinejad R, Totonchi M. Adverse Drug Reaction Detection in Social Media by Deepm Learning Methods. *Cell J* 2020;22:319–24. <https://doi.org/10.22074/cellj.2020.6615>.
- [44] Hussain S-A, Sezgin E, Krivchenia K, Luna J, Rust S, Huang Y. A natural language processing pipeline to synthesize patient-generated notes toward improving remote care and chronic disease management: a cystic fibrosis case study. *JAMIA Open* 2021;4:ooab084. <https://doi.org/10.1093/jamiaopen/ooab084>.
- [45] Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Prepr ArXiv181004805* 2018.
- [46] Orre R, Lansner A, Bate A, Lindquist M. Bayesian neural networks with confidence estimation applied to data mining. *Comput Stat Data Anal* 2000;34:473–93. [https://doi.org/10.1016/S0167-9473\(99\)00114-0](https://doi.org/10.1016/S0167-9473(99)00114-0).
- [47] Norén GN, Bate A, Orre R, Edwards IR. Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. *Stat Med* 2006;25:3740–57.
- [48] Théorème de Bayes. *Wikipédia* 2022.
- [49] Ahmed I, Pariente A, Tubert-Bitter P. Class-imbalanced subsampling lasso algorithm for discovering adverse drug reactions. *Stat Methods Med Res* 2018;27:785–97. <https://doi.org/10.1177/0962280216643116>.
- [50] Xu Z, Kass-Hout T, Anderson-Smits C, Gray G. Signal detection using change point analysis in postmarket surveillance. *Pharmacoepidemiol Drug Saf* 2015;24:663–8.
- [51] Taylor W. Change-Point Analysis: A Powerful New Tool For Detecting Changes n.d.:19.
- [52] Azam R. Accessing social media information for pharmacovigilance: what are the ethical implications? *Ther Adv Drug Saf* 2018;9:385–7. <https://doi.org/10.1177/2042098618778191>.
- [53] Pierce CE, Bouri K, Pamer C, Proestel S, Rodriguez HW, Van Le H, et al. Evaluation of Facebook and Twitter Monitoring to Detect Safety Signals for Medical Products: An Analysis of Recent FDA Safety Alerts. *Drug Saf* 2017;40:317–31. <https://doi.org/10.1007/s40264-016-0491-0>.
- [54] Coloma PM, Becker B, Sturkenboom MCJM, van Mulligen EM, Kors JA. Evaluating Social Media Networks in Medicines Safety Surveillance: Two Case Studies. *Drug Saf* 2015;38:921–30. <https://doi.org/10.1007/s40264-015-0333-5>.
- [55] Patki A, Sarker A, Pimpalkhute P, Nikfarjam A, Ginn R, Oconnor K, et al. Mining Adverse Drug Reaction Signals from Social Media: Going Beyond Extraction, 2014.
- [56] Duh MS, Cremieux P, Audenrode MV, Vekeman F, Karner P, Zhang H, et al. Can social media data lead to earlier detection of drug-related adverse events. *Pharmacoepidemiol Drug Saf* 2016;25:1425–33. <https://doi.org/10.1002/pds.4090>.

- [57] Harpaz R, DuMouchel W, LePendu P, Bauer-Mehren A, Ryan P, Shah NH. Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. *Clin Pharmacol Ther* 2013;93:539–46.
- [58] Abou Taam M, Rossard C, Cantaloube L, Bouscaren N, Roche G, Pochard L, et al. Analysis of patients' narratives posted on social media websites on benfluorex's (Mediator®) withdrawal in France. *J Clin Pharm Ther* 2014;39:53–5. <https://doi.org/10.1111/jcpt.12103>.
- [59] Wu H, Fang H, Stanhope SJ. Exploiting online discussions to discover unrecognized drug side effects. *Methods Inf Med* 2013;52:152–9. <https://doi.org/10.3414/ME12-02-0004>.
- [60] Yang CC, Yang H. Exploiting Social Media with Tensor Decomposition for Pharmacovigilance. 2015 IEEE Int. Conf. Data Min. Workshop ICDMW, 2015, p. 188–95. <https://doi.org/10.1109/ICDMW.2015.163>.
- [61] Utilizing Text Mining on Online Medical Forums to Predict Label Change due to Adverse Drug Reactions | Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining n.d. <https://dl.acm.org/doi/10.1145/2783258.2788608> (accessed July 31, 2022).
- [62] Liu X, Chen H. AZDrugMiner: An Information Extraction System for Mining Patient-Reported Adverse Drug Events in Online Patient Forums. In: Zeng D, Yang CC, Tseng VS, Xing C, Chen H, Wang F-Y, et al., editors. *Smart Health*, Berlin, Heidelberg: Springer; 2013, p. 134–50. https://doi.org/10.1007/978-3-642-39844-5_16.
- [63] Bhattacharya S, Srinivasan P, Polgreen P. Social media engagement analysis of U.S. Federal health agencies on Facebook. *BMC Med Inform Decis Mak* 2017;17:49. <https://doi.org/10.1186/s12911-017-0447-z>.
- [64] Karapetiantz P, Audeh B, Lillo-Le Louët A, Bousquet C. Signal Detection for Baclofen in Web Forums: A Preliminary Study. *Stud Health Technol Inform* 2018;247:421–5.
- [65] Chen X, Faviez C, Schuck S, Lillo-Le-Louët A, Texier N, Dahamna B, et al. Mining Patients' Narratives in Social Media for Pharmacovigilance: Adverse Effects and Misuse of Methylphenidate. *Front Pharmacol* 2018;9:541. <https://doi.org/10.3389/fphar.2018.00541>.
- [66] Karapetiantz P, Audeh B, Faille J, Lillo-Le Louët A, Bousquet C. Qualitative and Quantitative Analysis of Web Forums for Adverse Events Detection: “Strontium Ranelate” Case Study. *Stud Health Technol Inform* 2019;264:964–8. <https://doi.org/10.3233/SHTI190367>.
- [67] Kürzinger M-L, Schück S, Texier N, Abdellaoui R, Faviez C, Pouget J, et al. Web-Based Signal Detection Using Medical Forums Data in France: Comparative Analysis. *J Med Internet Res* 2018;20:e10466. <https://doi.org/10.2196/10466>.
- [68] Lee J-Y, Lee Y-S, Kim DH, Lee HS, Yang BR, Kim MG. The Use of Social Media in Detecting Drug Safety-Related New Black Box Warnings, Labeling Changes, or Withdrawals: Scoping Review. *JMIR Public Health Surveill* 2021;7:e30137. <https://doi.org/10.2196/30137>.
- [69] Latent dirichlet allocation | The Journal of Machine Learning Research n.d. <https://dl.acm.org/doi/10.5555/944919.944937> (accessed October 18, 2022).
- [70] Sullivan R, Sarker A, O'Connor K, Goodin A, Karlsrud M, Gonzalez G. FINDING POTENTIALLY UNSAFE NUTRITIONAL SUPPLEMENTS FROM USER REVIEWS WITH TOPIC MODELING. *Pac Symp Biocomput Pac Symp Biocomput* 2016;21:528–39.
- [71] Cameron D, Smith GA, Daniulaityte R, Sheth AP, Dave D, Chen L, et al. PREDOSE: A semantic web platform for drug abuse epidemiology using social media. *J Biomed Inform* 2013;46:985–97. <https://doi.org/10.1016/j.jbi.2013.07.007>.
- [72] Daniulaityte R, Chen L, Lamy FR, Carlson RG, Thirunarayan K, Sheth A. “When ‘bad’ is ‘good’”: identifying personal communication and sentiment in drug-related tweets. *JMIR Public Health Surveill* 2016;2:e6327.
- [73] Sheth A, Jadhav A, Kapanipathi P, Lu C, Purohit H, Smith GA, et al. Twitris: A System for Collective Social Intelligence. In: Alhajj R, Rokne J, editors. *Encycl. Soc. Netw. Anal. Min.*, New York, NY: Springer New York; 2014, p. 2240–53. https://doi.org/10.1007/978-1-4614-6170-8_345.

- [74] Casperson TA, Painter JL, Dietrich J. Strategies for distributed curation of social media data for safety and pharmacovigilance. *Proc. Int. Conf. Data Sci. ICDATA, The Steering Committee of The World Congress in Computer Science, Computer ...*; 2016, p. 118.
- [75] Dietrich J, Gattepaille LM, Grum BA, Jiri L, Lerch M, Sartori D, et al. Adverse events in twitter-development of a benchmark reference dataset: results from IMI WEB-RADR. *Drug Saf* 2020;43:467–78.
- [76] Freifeld CC. Digital pharmacovigilance: The medwatcher system for monitoring adverse events through automated processing of internet social media and crowdsourcing. PhD Thesis. Boston University, 2014.
- [77] Nguyen AT, Lien J, Raff E, Mekaru SR. Improved Automatic Pharmacovigilance: An Enhancement to the MedWatcher Social System for Monitoring Adverse Events. *BioRxiv* 2019:717421.
- [78] Yeleswarapu S, Rao A, Joseph T, Saipradeep VG, Srinivasan R. A pipeline to extract drug-adverse event pairs from multiple data sources. *BMC Med Inform Decis Mak* 2014;14:13. <https://doi.org/10.1186/1472-6947-14-13>.
- [79] Cossin S, Lebrun L, Lobre G, Loustau R, Jouhet V, Griffiera R, et al. Romedi: an open data source about French drugs on the semantic web. *MEDINFO 2019 Health Wellbeing E-Netw. Proc. 17th World Congr. Med. Health Inform.*, vol. 264, IOS Press; 2019, p. 79.
- [80] Nikfarjam A, Ransohoff JD, Callahan A, Jones E, Loew B, Kwong BY, et al. Early Detection of Adverse Drug Reactions in Social Health Networks: A Natural Language Processing Pipeline for Signal Detection. *JMIR Public Health Surveill* 2019;5:e11264. <https://doi.org/10.2196/11264>.
- [81] Abdellaoui R, Schüick 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* 2017;3:e36. <https://doi.org/10.2196/publichealth.6577>.
- [82] Bousquet C, Dahamna B, Guillemin-Lanne S, Darmoni SJ, Faviez C, Huot C, et al. The Adverse Drug Reactions from Patient Reports in Social Media Project: Five Major Challenges to Overcome to Operationalize Analysis and Efficiently Support Pharmacovigilance Process. *JMIR Res Protoc* 2017;6:e179. <https://doi.org/10.2196/resprot.6463>.
- [83] Grosjean J, Merabti T, Soualmia LF, Letord C, Charlet J, Robinson PN, et al. Integrating the human phenotype ontology into HeTOP terminology-ontology server. *Stud Health Technol Inform* 2013;192:961.
- [84] Audeh B, Beigbeder M, Zimmermann A, Jaillon P, Bousquet C. Vigi4Med Scraper: A Framework for Web Forum Structured Data Extraction and Semantic Representation. *PLOS ONE* 2017;12:e0169658. <https://doi.org/10.1371/journal.pone.0169658>.
- [85] Poncet IA& A. PhViD: PharmacoVigilance Signal Detection 2016.
- [86] Courtois E, Ahmed I, Perdry H. adapt4pv: Adaptive Approaches for Signal Detection in Pharmacovigilance 2022.
- [87] Viard D, Parassol-Girard N, Romani S, Van Obberghen E, Rocher F, Berriri S, et al. Spontaneous adverse event notifications by patients subsequent to the marketing of a new formulation of Levothyrox® amidst a drug media crisis: atypical profile as compared with other drugs. *Fundam Clin Pharmacol* 2019;33:463–70. <https://doi.org/10.1111/fcp.12446>.
- [88] Les Français sont les plus sceptiques face aux vaccins, selon une enquête mondiale n.d. https://www.lemonde.fr/societe/article/2019/06/19/les-francais-sont-les-plus-sceptiques-face-aux-vaccins-selon-une-enquete-mondiale_5478259_3224.html (accessed October 18, 2022).
- [89] Cluster anxiety-related adverse events following immunization (AEFI): An assessment of reports detected in social media and those identified using an online search engine - PubMed n.d. <https://pubmed.ncbi.nlm.nih.gov/30172632/> (accessed October 18, 2022).
- [90] Lama Y, Hu D, Jamison A, Quinn SC, Broniatowski DA. Characterizing Trends in Human Papillomavirus Vaccine Discourse on Reddit (2007-2015): An Observational Study. *JMIR Public Health Surveill* 2019;5:e12480. <https://doi.org/10.2196/12480>.

- [91] Dr. Jekyll or Mr. Hyde? (How) the Internet influences vaccination decisions: recent evidence and tentative guidelines for online vaccine communication - PubMed n.d. <https://pubmed.ncbi.nlm.nih.gov/22472790/> (accessed October 18, 2022).
- [92] Karapetiantz P, Audeh B, Lillo-Le Louët A, Bousquet C. Discrepancy Between Personal Experience and Negative Opinion with Human Papillomavirus Vaccine in Web Forums. *Stud Health Technol Inform* 2020;272:417–20. <https://doi.org/10.3233/SHTI200584>.
- [93] Karapetiantz P, Audeh B, Bousquet C. Identification of COVID-19 Vaccines Concerns in Health-Related French Web Forums: A Topic Modelling Approach. *Stud Health Technol Inform* 2022;289:174–7. <https://doi.org/10.3233/SHTI210887>.
- [94] WHO Coronavirus (COVID-19) Dashboard n.d. <https://covid19.who.int> (accessed October 18, 2022).
- [95] Cathy. WHO'S top 10 threats to global health in 2019. ANMJ 2019. <https://anmj.org.au/whos-top-10-threats-to-global-health-in-2019/> (accessed October 18, 2022).
- [96] Grün B, Hornik K. topicmodels: An R Package for Fitting Topic Models. *J Stat Softw* 2011;40:1–30. <https://doi.org/10.18637/jss.v040.i13>.
- [97] LeBozec C, Jaulent MC, Zapletal E, Degoulet P. Unified modeling language and design of a case-based retrieval system in medical imaging. *Proc AMIA Symp* 1998:887–91.
- [98] Knape T, Hederman L, Wade VP, Gargan M, Harris C, Rahman Y. A UML approach to process modelling of clinical practice guidelines for enactment. *Stud Health Technol Inform* 2003;95:635–40.
- [99] Audeh B, Grouin C, Zweigenbaum P, Bousquet C, Jaulent M-C, Benkhebil M, et al. French Levothyrox® Crisis: Retrospective Analysis of Social Media. In: Publishing SI, editor. *Int. Soc. Pharmacovigil.*, Bogota, Colombia: Springer International Publishing; 2019.
- [100] A Constructed Reality? A Fairclough-Inspired Critical Discourse Analysis of the Danish HPV Controversy - Sanne Mohr, Kirsten Frederiksen, 2020 n.d. <https://journals.sagepub.com/doi/abs/10.1177/1049732320909098> (accessed November 1, 2022).
- [101] Luisi MLR. Kansan Guardian Perceptions of HPV and the HPV Vaccine and the Role of Social Media. *Kans J Med* 2020;13:9–18.
- [102] Campillos-Llanos L, Grouin C, Lillo-Le Louët A, Zweigenbaum P. Initial Experiments for Pharmacovigilance Analysis in Social Media Using Summaries of Product Characteristics. *Stud Health Technol Inform* 2019;264:60–4. <https://doi.org/10.3233/SHTI190183>.

Annexe 1 : Description des projets

Le projet ADR-Prism (Adverse Drug Reactions from Patient Reports in Social Media)

Partenaires :

Centre de Pharmacovigilance de l'Hôpital Européen Georges Pompidou ; CHU de Rouen - équipe CISMef ; INSERM U1138 eq 22 ; INSERM U1142 – LIMICS ; Institut de Santé Urbaine ; Kappa Santé ; Temis (devenu Expert System) ; Vidal

Résumé :

Financé par le Fond Unique Interministériel (FUI) en 2013, le projet ADR-PRISM avait pour objectif de mettre à disposition des équipes de pharmacovigilance une source de connaissances encore inexploitée en dehors de rares expérimentations par des équipes de recherche : les messages des patients dans les forums et autres lieux de discussions sur Internet. L'intégration de ces données pourrait permettre de générer de nouvelles hypothèses concernant les effets indésirables décrits par les patients qui seraient nouveaux ou mal documentés dans l'information officielle et/ou déjà existante sur les médicaments. Etant donné le volume d'information en provenance des patients collectée par millions de manière journalière depuis les différentes formes d'échange sur le web, ce cas d'usage était une application Big data.

Les méthodes mises en œuvre dans ADR-PRISM appartiennent à l'ingénierie des connaissances et à la fouille de texte. Les informations sont extraites à partir de données textuelles (messages des patients sur les plates-formes de discussion) au moyen de méthodes de traitement automatique du langage (TAL). Les données textuelles sont annotées et intégrées aux données structurées, comme la date, la source de donnée ou l'adresse, en considérant leur sémantique, pour en préserver le sens et permettre une intégration et interopérabilité améliorées entre le langage médical et le langage des patients. D'un point de vue technique, les résultats de l'intégration étaient consultables dans un environnement Web. Des méthodes de détection du signal ont pu être appliquées aux données.

Le projet Vigi4med (Vigilance dans les forums sur les médicaments)

Partenaires :

DSPIM Saint-Étienne ; INSERM U1138 eq 22 ; LIMSI ; ISCOD Ecole des Mines de Saint-Etienne ; INSERM U1142 – LIMICS ; Centres Régionaux de Pharmacovigilance partenaires du projet (Hôpital européen Georges Pompidou et CHU de Saint-Étienne)

Résumé :

L'hypothèse du projet Vigi4MED (Vigilance dans les forums sur le médicament), financé par l'ANSM en 2013, est que les commentaires des patients sur les effets indésirables susceptibles d'être liés aux médicaments rapportés dans les réseaux sociaux constituent une source complémentaire pour améliorer nos connaissances en pharmacovigilance. Les quelques études décrites dans la littérature montrent qu'on peut retrouver de façon rétrospective des effets indésirables des médicaments mais ne permettent pas de conclure sur l'intérêt des réseaux sociaux dans le cadre des processus de pharmacovigilance. Le projet Vigi4MED vise à combler ce manque.

Le projet PHARES (PHarmacovigilance dans les REseaux Sociaux)

Partenaires :

INSERM U1142 - LIMICS ; LIMSI ; Centre Régional de Pharmacovigilance de l'hôpital Européen Georges Pompidou (CRPV-HEGP)

Résumé :

Financé par l'appel à candidature décrite en 2018 de l'ANSM, le projet PHARES (Pharmacovigilance dans les réseaux sociaux) a pour objectif de mettre en œuvre une suite logicielle permettant d'interroger les réseaux sociaux, plus précisément les messages postés sur des forums, dans un objectif de pharmacovigilance. Le projet PHARES s'inscrit dans la continuité du projet de recherche Vigi4Med (Vigilance dans les forums médicaments) financé par l'appel à projet 2013 de l'ANSM et s'appuie sur un consortium plus réduit avec un objectif plus opérationnel. La suite logicielle qui a été développée comprend 4 modules : 1) un module de moissonnage des réseaux sociaux permettant de récupérer régulièrement les nouveaux messages pour les forums sélectionnés (module créé par Bissan Audeh de l'Ecole des Mines de Saint-Etienne) ; 2) un module d'annotation qui permet de détecter automatiquement les éléments d'intérêt dans les messages et plus particulièrement les noms de maladies et de symptômes d'une part et les noms de médicaments d'autre part (module créé par le

Laboratoire Interdisciplinaire des Sciences du Numérique, le LIMSI); 3) un module d'analyse statistique qui implémente les algorithmes classiques de détection du signal mais appliqués ici aux messages annotés ; 4) une interface d'interrogation à usage des professionnels de santé pharmacovigilants (module pour lequel j'avais défini les spécifications, rédigé un cahier des charges et un guide utilisateur et implémenté par Akram Redjda, ingénieur d'étude pour le LIMICS) .

Annexe 2 : Captures d'écran de la BNPV

The screenshot displays the 'Identifiants' section of the BNPV software. It includes fields for 'Centre *' (CRPV PARIS HEGP), 'Identifiant unique du cas', 'Identifiant unique mondial du cas', and 'Premier expéditeur' (Autorité / Autre). Below this is the 'Informations sur le cas' section, which is divided into 'Caractérisation du cas' and 'Résumé du cas'. The 'Caractérisation du cas' section contains fields for 'Date de réception *', 'Type de rapport *' (Rapport spontané), 'Cas médicalement confirmé', and 'Date de confirmation'. It also includes 'Identification du pays du rapporteur' with dropdowns for 'Pays où le réaction / l'événement s'est produit' and 'Département'. The 'Résumé du cas' section has checkboxes for 'Cas grave' (Cotés de gravité) and 'Mise en jeu du pronostic vital', along with a text area for 'Autre situation médicale grave : spécifier'.

Figure 26 : BNPV - Observations (date, gravité)

The screenshot displays the 'Notificateur #1' section of the BNPV software. It includes fields for 'Pays du notificateur' (FRANCE), 'Source principale à des fins réglementaires' (Oui), and 'Qualification E2B' (Spécialité). Below this is a section for 'Hôpital' with fields for 'Service', 'Titre', 'Prénom', 'Deuxième prénom', and 'Nom (en MAJUSCULE)'. There is also a section for 'Rue' with fields for 'Code postal', 'Ville', 'Département', 'Téléphone', 'Numéro de fax', 'Numéro de cellule', and 'Email'. The 'Type d'exercice' and 'Qualification PV (précision)' are also visible. At the bottom, there is a 'Contact avec le notificateur #1' section with a table for 'Question/Réponse'.

Figure 27 : BNPV - Déclarant

The screenshot shows the 'Information patient' form. It includes a search bar at the top right with 'Nouveau cas' and 'AI/NEW_ANSM' buttons. The main navigation bar contains tabs for 'Info cas', 'Observation', 'Notificateur', 'Patient', 'Antécédents', 'Info Parents', 'Effets', 'Examens complémentaires', 'Médicaments/Substances', 'Matrice Produits/Effets', and 'Commentaires'. The 'Antécédents' tab is active. The form fields include:

- Identifiant:** 'Trois premières lettres du nom (max.)' and 'Première lettre du prénom'.
- Informations:** 'Sexe', 'Taille (cm)', 'Poids (kg)', 'Surface corporelle (m2)', 'IMC', 'Date de naissance', 'Âge', 'Unité', 'Groupe d'âge', 'Département'.
- Tabagisme, Alcoolisme, Toxicomanie, Commentaire:** dropdown menus and a text area.
- Terme de la grossesse (aménorrhée) au moment de l'effet:** dropdown menu and 'Unité'.
- Date des dernières règles:** date field and 'Masqué' checkbox.

Figure 28 : BNPV - Patient

The screenshot shows the 'Antécédents médicaux (structurés) #1' form. It includes the same navigation bar as Figure 28. The form fields include:

- Antécédent médical tel que rapporté:** 'Terme MedDRA', 'PT', 'V.', 'SOC', 'Date de début', 'En cours', 'Date de fin', 'Commentaire', 'Antécédent familial'.
- Antécédents médicaux (texte):** 'Antécédent médical', 'Thérapies concomitantes'.
- Antécédents thérapeutiques #1:** 'Nom du médicament tel que rapporté', 'Nom du produit', 'Date de début', 'Date de fin', 'Indication telle que rapportée', 'Indication MedDRA', 'PT', 'V.', 'SOC'.

Figure 29 : BNPV - Antécédents

The screenshot shows the 'Effet #1' form. It includes the same navigation bar as Figure 28. The form fields include:

- Effet (tel que rapporté) dans la langue maternelle:** 'Langue maternelle', 'Effet (tel que rapporté) en anglais', 'Terme souligné par le notificateur', 'Pays où l'effet s'est produit', 'Médicament concerné'.
- Evolution:** dropdown menu.
- El grave:** 'Critères de gravité des EI' (Oui/Non), 'Décès', 'Issue en jeu de pronostic vital', 'Prolongation de l'hospitalisation', 'Incapacité ou invalidité', 'Anomalie congénitale', 'Autre situation médicale grave'.
- Date de début, Date de fin, Durée, Unité:** date and duration fields.
- Informations complémentaires:** 'Effet déjà survenu', 'Evolution (précisions)', 'Commentaire'.

Figure 30 : BNPV – Effets indésirables

Figure 31 : BNPV – Médicament (1)

Figure 32 : BNPV – Médicament (2)

Figure 33 : BNPV - imputabilité