Création d'une intelligence artificielle capable de détecter les impacts de qualité de vie liée à la santé à partir de données de vie réelle

Tom Marty

UNIVERSITÉ CAEN NORMANDIE

U.F.R. Santé

Faculté des Sciences Pharmaceutiques

THESE

Pour l'obtention du Diplôme d'État de Docteur en Pharmacie

Préparée au sein de l'Université de Caen Normandie

CREATION D'UNE INTELLIGENCE ARTIFICIELLE CAPABLE DE DETECTER LES IMPACTS DE QUALITE DE VIE LIEE A LA SANTE A PARTIR DE DONNEES DE VIE REELLE.

Présentée par
Tom Marty

Soutenue publiquement le 14 Octobre 2022
devant le jury composé de

M. Michel Boulouard	Enseignant chercheur	Président du jury
M. Stéphane Schück	Médecin de santé publique	Directeur de thèse
Mme Catherine Schück - Beaufils	Pharmacien hospitalier	Co-directrice de thèse
M. Adel Mebarki	Spécialiste innovation & données de santé	Personne qualifiée

Thèse dirigée par Stéphane Schück et Catherine Schück

LISTE DES ENSEIGNANTS-CHERCHEURS

Directrice de la Faculté des Sciences Pharmaceutiques
Professeur Pascale SCHUMANN-BARD

Assesseurs
Professeur MALZERT-FREON Aurélie
Professeur Anne-Sophie VOISIN-CHIRET

Directrice administrative
Madame Sarah CHEMTOB

Directrice administrative adjointe
Madame Emmanuelle BOURDON

PROFESSEURS DES UNIVERSITES

BOULOUARD Michel ..... Physiologie, Pharmacologie
BUREAU Ronan ..... Biophysique, Chémoinformatique
COLLOT Valérie ..... Pharmacognosie
DALLEMAGNE Patrick ..... Chimie médicinale
DAUPHIN François ..... Physiologie, Pharmacologie
DELEPEE Raphaël ..... Chimie analytique
FABIS Frédéric ..... Chimie organique
FRERET Thomas ..... Physiologie, Pharmacologie
GARON David ..... Botanique, Mycologie, Biotechnologies
GIARD Jean-Christophe ..... Bactériologie, Virologie
MALZERT-FREON Aurélie ..... Pharmacie galénique
ROCHAIS Christophe ..... Chimie organique
SCHUMANN-BARD Pascale ..... Physiologie, Pharmacologie
SICHEL François ..... Toxicologie
SOPKOVA Jana ..... Biophysique, Drug design
VOISIN-CHIRET Anne-Sophie ..... Chimie médicinale

MAITRES DE CONFERENCES DES UNIVERSITES

ANDRE Véronique – HDR ..... Biochimie, Toxicologie
BOUET Valentine – HDR ..... Physiologie, Pharmacologie
BRIERE Joséphine ..... Biostatistique
CAILLY Thomas – HDR ..... Chimie bio-inorganique, Chimie organique
DENOYELLE Christophe – HDR ..... Biologie cellulaire et moléculaire, Biochimie, Cancérologie
DHALLUIN Anne ..... Bactériologie, Virologie, Immunologie
DUBOST Emmanuelle ..... Chimie organique
ELDIN de PECOULAS Philippe – HDR ..... Parasitologie, Mycologie médicale

GROO Anne-Claire ..... Pharmacie galénique
KIEFFER Charline..... Chimie médicinale
KRIEGER Sophie (Praticien hospitalier) **– HDR** ..... Biologie clinique
LAPORTE-WOJCIK Catherine ..... Chimie bio-inorganique
LEBAILLY Pierre – HDR..... Santé publique
LECHEVREL Mathilde – HDR..... Toxicologie
LEGER Marianne ..... Physiologie, Pharmacologie
LEPAILLEUR Alban – HDR ..... Modélisation moléculaire
N'DIAYE Monique ..... Parasitologie, Mycologie médicale, Biochimie clinique
PAIZANIS Eleni ..... Physiologie, Pharmacologie
POTTIER Ivannah ..... Chimie et toxicologie analytiques
PREVOST Virginie – HDR ..... Chimie analytique, Nutrition, Education thérapeutique du patient
QUINTIN Jérôme..... Pharmacognosie
RIOULT Jean-Philippe..... Botanique, Mycologie, Biotechnologies
SAINT-LORANT Guillaume (Praticien hospitalier) ..... Pharmacie clinique
SINCE Marc– HDR ..... Chimie analytique
THEAULT BRYERE Joséphine ..... Biostatistiques
VILLEDIEU Marie – HDR..... Biologie et thérapies innovantes des cancers

PROFESSEUR AGREGE (PRAG)

PRICOT Sophie..... Anglais

PERSONNEL ASSOCIE A TEMPS PARTIEL (PAST)

SEDILLO Patrick..... Pharmacie officinale
SEGONZAC Virginie..... Pharmacie officinale

Remerciements

Mes remerciements et pensées vont à …

Je tiens tout d'abord à remercier Monsieur le Professeur Boulouard, pour avoir accepté de présider mon jury de thèse. Monsieur le Docteur Stéphane Schück pour tes conseils et ta disponibilité, Madame le Docteur Catherine Schück - Beaufils pour ta bienveillance, ta philosophie de vie, un énorme merci à vous d'avoir accepté de m'accompagner sur ce projet et cette thèse durant cette période si particulière.

Mika, ton travail et investissement durant ces 6 mois et bien au-delà, ont rendu ce projet possible, merci binôme !

A Simon et Adel, pour nos échanges en fin d'année 2019 qui ont abouti à la création de ce stage, de ce projet et tous ceux qui ont suivi. Pour votre présence, votre confiance, votre bienveillance, vos conseils et votre amitié précieuse.

A mes amis et collègues de Kap Code, une Start-Up/entreprise à nulle autre pareil ! Merci de, chaque jour, créer les conditions d'épanouissement professionnel et personnel. Juliette, Wiem, Vanessa, Salma, Ginette, Pamela, Paul, Vincent, Blandine, Amélia, Pierre G, Pierre F, Anaïs, Manissa, Emma, Léa, Idir, Fakih, Patricia, Nathalie P, Nathalie T. Merci pour votre confiance, votre amitié et nos échanges toujours constructifs.

Je tiens également à remercier toute ma famille, mes parents pour votre amour et pour avoir été les supers parents que vous êtes, Florence pour le pilier que tu as été dans mon accompagnement et la définition de mon projet professionnel, à Nino et Noé mes petits frères, à Christiane et André, à Manou, à Majo, à Scott et Lewis.

Les amis sont la famille que l'on se choisit, merci à des groupes et individus si particuliers : Célia, Valentine, Maxime, Thomas, Florient, Louis. Mention spéciale à Clara, pour nos fous rires et autres échanges télépathiques. A ma famille d'accueil à Paris : Maylis, Benjamin R, Jordan R. A Jordan C, au Dr Florian. A Benjamin G, tu sais pourquoi.

A vous tous individuellement, merci d'être qui vous êtes.

Sommaire

Table des matières

Liste des abréviations

IA - Intelligence artificielle

TALN - Traitement Automatique du Langage Naturel

BMJ - British Medical Journal

HAS - Haute Autorité de Santé

PREMs - Patients Reported Experience Measure

PROMs - Patients Reported Outcome Measure

EQ-5D - EuroQoL - Five Dimensions

SF-36 - Short Form 36 Dimensions

API - Application Programming Interface

JAMA - Journal of American Medical Association

AMM - Autorisation de Mise sur le Marché

JMIR - Journal of Medical Internet Research

AUC - Area Under the Curve / Aire Sous la Courbe

AQoL - Assesment of Quality of Life

MedDRA - Medical

XGB - Extreme Gradiant Boosting

KNN - K Nearest Neighboors

SVM - Support Vector Machine

MLP - Multi Layer Perceptron

RF - Random Forest

SMOTE - Synthethic Minority Oversampling Technique

ROC - receiver operating characteristics

Liste des Figures

Liste des tableaux

Préambule

Qu'est-ce que l'Intelligence Artificielle (IA) ? Une entité abstraite à laquelle certains fantasmes prêtent une apparence, une conscience et donc, des intentions ?

Un mathématicien répondrait surement que l'IA est un champ de recherche conceptuelle, des équations et calculs probabilistes appliqués selon certaines règles bien précises.

Un *data scientist*, lui, la définirait surement par la capacité d'apprentissage d'une machine face à des données, dans le but de prédire un résultat sur d'autres données du même type.

Un professionnel de santé, pourrait la caractériser en fonction de l'espoir et de la promesse qu'un tel outil révolutionnairement innovant, peut incarner dans la prise en charge des patients.

Un philosophe, pourrait arguer (dans une interprétation certainement prométhéenne) qu'il s'agit d'une manifestation de *l'hubris* humain dont le développement est à encadrer par une garantie humaine, la dimension artificielle d'une solution synthétique risquant d'éloigner l'humain de son humanité.

Loin d'un visage numérique composé de lignes de codes vertes sur un écran, actuellement en 2022 l'appellation IA est démocratisée. Il s'agit même d'un argument marketing promettant des résultats et un retour utilisateur personnalisé.

Un autre mot que nous utilisons tous, parfois sans vraiment le comprendre, est le terme algorithme. Qu'est-ce qu'un algorithme ? C'est avec cette question qu'Aurélie Jean (scientifique et entrepreneuse française) démarre son livre « De l'autre côté de la machine » (1). D'origine perse, le terme algorithme vient du nom latinisé de Muhammad ibn Musa al-Khwarizmi, un mathématicien astronome géographe du IXè siècle. Euclide et son algorithme de division euclidienne a aussi sa part d'importance dans l'utilisation et la diffusion du terme (1).

Elle explique que selon une définition largement répandue, un algorithme serait comme une recette de cuisine, qui contient des ingrédients et une marche à suivre pour réaliser une opération sur un ordinateur. Cependant elle explique que sous un certain angle, cette image peut sembler correcte dans le sens où un algorithme est une séquence d'opérations visant à réaliser une tâche. Mais l'image est trop simpliste et induit en erreur. Selon elle, on qualifie de numérique un algorithme qui a été conçu pour être implémenté dans un code informatique destiné à faire tourner une simulation ou un calcul sur un ou plusieurs processeurs d'un ordinateur. La différence est fondamentale, là où Euclide appliquait son algorithme de division à la main, les algorithmes d'aujourd'hui sont spécifiquement pensés pour être utilisés par ordinateur. C'est encore plus vrai pour les algorithmes d'apprentissage profond, ou *deep learning*, qui permettent à un programme (après entraînement) d'identifier un chien ou une voiture sur une image. Ces réseaux sont inutilisables à la main, principalement en raison de leur complexité (1).

Qu'est-ce qu'est vraiment l'intelligence artificielle ? ou plutôt, qu'est-ce qu'elle n'est pas ? Car les personnes, quelle que soit leur profession d'origine, qui travaillent à ce jour dans cette discipline, la définirait en fonction de termes tels qu'algorithmes ou modèles, capacité d'apprentissage, approche statistique, prédiction, et d'autres. Ils pourraient aussi préciser, face aux fantasmes liés à l'IA, qu'actuellement, l'intelligence artificielle reste bien encore, artificielle.

Introduction

L'objet de ce travail était de réaliser un algorithme d'Intelligence artificielle, ou plus précisément, un modèle de traitement automatisé du langage, capable d'identifier dans des témoignages libres de patients, les verbatims reflétant un impact de qualité de vie.

La qualité de vie liée à la santé est un ensemble multi conceptuel qui profite de plusieurs définitions de différentes sociétés savantes. Un état de santé, une prise en charge ou encore une thérapeutique, peuvent avoir différents impacts sur la qualité de vie des patients. Depuis plusieurs dizaines d'années, les professionnels de santé monitorent l'évolution de la qualité de vie liée à la santé de leurs patients, grâce à différents questionnaires validés.

Depuis quelques années, un changement de paradigme appelé « patient centrisme », participe à l'évolution des pratiques médicales et à la prise en charge des patients. Ce changement de point de vue s'éloigne du paternalisme médical, afin de mieux appréhender la « réalité patient », et de remettre le patient au cœur de sa prise en charge.

A ce titre, il existe une synergie entre le patient centrisme et l'essor des données de vie réelle. Ensemble de données médicalement contextualisées, elles reflètent et objectivent la réalité patiente, par définition non captée lors d'essais cliniques. La qualité de vie rentre dans les données de vie réelle, au même titre qu'il peut s'agir d'indicateurs cliniques.

Sur les réseaux sociaux et forums médicaux, les patients forment des communautés en ligne dans un but de soutien, d'information et de partage d'expériences médicales. Ces commentaires publics sont récupérables informatiquement, dans le respect des réglementations pour la protection des données. C'est là que le traitement automatisé du langage entre en jeu. En étant capable d'analyser et de traduire le langage patient en ontologie médicale, il devient possible de gagner en information sur la réalité patiente, telle que décrite directement et sans filtre par eux-mêmes.

Ainsi, notre algorithme répond à l'enjeu de compréhension fine de ce qui peut réellement impacter la qualité de vie des patients, en vie réelle. La méthodologie adoptée est décrite au long du document.

L'organisation du projet a été définie par un binôme dont les connaissances et compétences étaient complémentaires. En tant que chargé de projet, un étudiant en pharmacie apportait ses compétences de conduite de projet, de bibliographie de la littérature médicale, ainsi que la dimension santé dans lequel l'algorithme, sa méthodologie et ses résultats futurs, seront contextualisés. En duo avec un *data scientist* (scientifique des données), qui apportait ses compétences de code, de création de modèle d'intelligence artificielle et de choix des meilleurs paramètres et variables à implémenter au projet. Les deux personnes ont ensemble annoté les données qui ont servi à entraîner l'algorithme, et ont ensemble interprété les résultats du modèle. Ce mode de fonctionnement a véritablement illustré la synergie qui existe entre les deux profils, un spécialiste des données et de l'intelligence artificielle, avec un pharmacien chef de projet médical.

Dans une première partie de cette thèse, les différentes définitions et concepts utiles à la compréhension du projet seront introduits. Ensuite, le projet et sa méthodologie seront développés. Puis dans une troisième partie, des exemples de cas pratiques sur différentes maladies seront présentés afin de comprendre comment cet algorithme peut être d'utilité dans la compréhension et la caractérisation fine des impacts médicaux sur la qualité de vie liée à la santé.

PARTIE 1 : DEFINITIONS

I. Intelligence artificielle, Traitement Automatisé du Langage Naturel, Machine Learning, Deep Learning

1. L'intelligence artificielle

Définie par le dictionnaire français le Larousse, l'intelligence artificielle est « l'ensemble des théories et des techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine » (2).

Plusieurs personnes sont considérées comme parentes de cette discipline. C'est John McCarthy qui, en 1949, baptise cette discipline alors qu'il était étudiant en mathématique à la *California Institute of Technology* (2,3). Il cherchait à développer un langage capable de traduire les raisonnements propres à l'humain dans des programmes informatiques. Une autre personnalité historique du domaine est Alan Turing avec son célèbre test du Turing : ce test consiste à mettre un humain en confrontation verbale, à l'aveugle, avec un ordinateur et un autre humain. Si la personne qui engage les conversations n'est pas capable de dire lequel des deux interlocuteurs est un ordinateur, on peut considérer que le logiciel de l'ordinateur a passé le test avec succès. C'est le cas de la célèbre intelligence artificielle ELIZA (1965), un agent conversationnel ayant pour but de simuler le discours d'un psychothérapeute.

L'intelligence artificielle est un grand concept qui manque encore de définition sur certains versants. Par exemple, des institutions telles que la Commission Nationale de l'Informatique et des Libertés (CNIL) notent le manque de précision de la définition de l'IA, qui est présentée comme le « grand mythe des temps moderne ». Le conseil de l'Europe précise même que malgré le fait d'être entré dans le langage commun et une utilisation devenue banale dans les média, il n'existe pas vraiment de définition partagée (4). Elle est souvent classée dans le groupe des mathématiques et des sciences cognitives, elle fait appel à la neurobiologie computationnelle (particulièrement aux réseaux neuronaux), à la logique mathématique, à la philosophie.

Il existe même plusieurs catégories d'intelligences artificielles (4). L'intelligence artificielle faible (ou étroite) constitue une approche pragmatique dans la résolution d'une tâche. L'IA faible vise à reproduire des facultés cognitives spécifiques, comme la reconnaissance d'une image ou le traitement du langage naturel. Il s'agit d'un algorithme simulant un comportement ou une faculté humaine, mais sans conscience. Son apport principal est l'automatisation d'une tâche pouvant libérer du temps humain puisqu'elle peut réaliser de manière rapide et précise, des tâches rébarbatives sans fatigue et aux performances souhaitées. A contrario, l'intelligence artificielle forte, qui est davantage théorique que pratique, emprunte à l'Homme des capacités cognitives, émotionnelles, sociales et psychomotrices avancées. La notion de conscience est au cœur de la réalisation d'IA fortes.

2. Le Traitement Automatisé du Langage

Le Traitement Automatique du Langage Naturel (TAL ou TALN) (5), est un domaine multidisciplinaire impliquant la linguistique, l'informatique et l'intelligence artificielle, qui vise à créer des outils de traitement de la langue pour diverses applications.

Les utilisations statistiques du traitement automatique du langage nature (TAL ou TALN) reposent sur des méthodes stochastiques, probabilistes ou statistiques pour résoudre certaines problématiques (phrases très longues, ambiguës, …) (6). La technologie pour le TAL statistique vient principalement de l'apprentissage automatique et de l'exploration de données, sous-tendant l'apprentissage à partir de données venant de l'intelligence artificielle.

Il existe plusieurs sous-catégories ou approches linguistiques pour entraîner un algorithme à reconnaître des informations textuelles. Au-delà des mots, c'est leur contexte et leur sens qui créé de l'information. A ce titre, on peut citer le « *word embedding* » aussi appelé « plongement lexical », en tant que technique d'apprentissage d'une représentation de mots. Chaque mot est transformé en un vecteur mathématique et chaque vecteur est placé dans la représentation d'un espace multidimensionnel. Les mots apparaissant dans des contextes similaires auront des vecteurs propres qui seront relativement proches. Cette tâche permet de faciliter l'apprentissage machine puisque les mots sont transformés en vecteurs mathématiques, lesquels sont formés à partir de l'information sémantique, linguistique et contextuelle desdits mots.

3. *Le machine Learning*

Le *machine learning*, ou apprentissage automatique, est une discipline qui est née dans les années 1940 (4). Elle connait deux phases d'essor, dans les années 80, puis dans les années 2010. L'accès à des volumes massifs de données, l'avancée technologique et la miniaturisation des machines ont rendu possible les calculs des algorithmes d'apprentissages. L'apprentissage automatique correspond à un réel changement de paradigme en informatique (figure 1). Il ne s'agit plus de faire appliquer des règles à un programme, mais de laisser un ordinateur découvrir par corrélation et classification, les règles et schémas régissant un ensemble de données. L'objectif de l'apprentissage automatique n'est pas d'acquérir des connaissances en tant que telles, mais de comprendre la structure des données, de l'intégrer dans des modèles (méthodes d'analyses de données) afin d'automatiser des tâches.

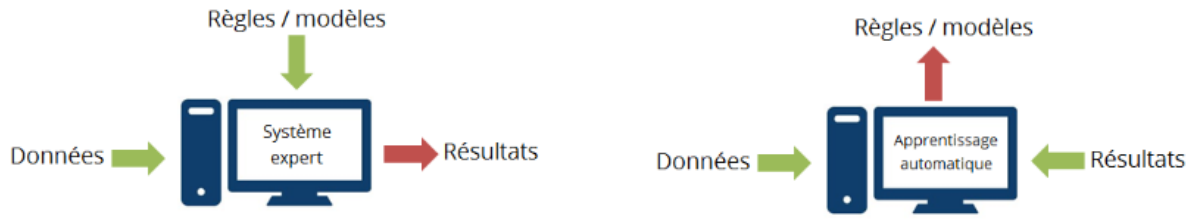


Figure 1 : Schématisation du changement de paradigme de l'apprentissage machine (Conseil de l'Europe, définition de l'intelligence artificielle)

4. *Le Deep Learning et réseaux de neurones*

Le *deep learning*, ou apprentissage profond, est, comme son nom l'indique, l'approfondissement de l'apprentissage automatique (4,7) . Il se distingue de ce dernier par l'utilisation de « réseaux de neurones ». Dans un modèle d'apprentissage profond, chaque « neurone » est en fait un programme, une unité de traitement à part entière avec la tâche d'appliquer une règle mathématique.

Chaque neurone est interconnecté à d'autres neurones. Ils sont ensuite organisés en « couches » où chaque couche du réseau aura une fonction dédiée. Ainsi, pour un réseau de neurones devant être capable d'identifier un chat sur une image, une couche sera entraînée à reconnaître la silhouette générale, une autre couche recherchera des oreilles, des pattes, le pelage, etc.

Cette organisation tire son inspiration du cerveau biologique. Elle permet d'augmenter la capacité de traitement. Le nombre de couches de neurones, leur densité, leur apprentissage, calibration et architecture permet d'augmenter la capacité de traitement des données et de toujours augmenter les attentes concernant les résultats.

Une autre similarité est la notion de seuil. Biologiquement, un neurone est activé par un potentiel d'action électrique. Ce signal doit être supérieur à un seuil pour générer une activation et la

transmission du signal. Les neurones artificiels ont un mode de fonctionnement similaire dans le sens où un « poids » (une pondération) est associé à l'information qu'ils délivrent. Des fonctions mathématiques permettent d'agréger et de combiner les informations transmises et ainsi de propager un signal, des informations, pour ensuite être traitées par les couches de neurones suivantes (figure 2) (7).

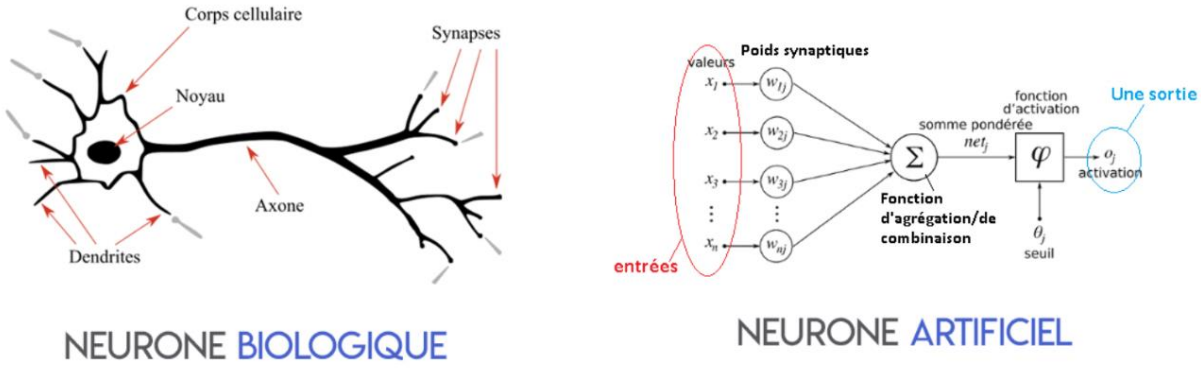


Figure 2 : Comparaison des représentations de neurone biologique et artificielle (7) (https://deeplylearning.fr/cours-theoriques-deep-learning/fonctionnement-du-neurone-artificiel/)

II. Qualité de vie, données de vie réelle et infodémiologie

L'amélioration du niveau de vie dans les sociétés modernes, industrielles et occidentales a transformé le concept de santé (8). Initialement définie comme « absence de maladie ou d'infirmité », l'OMS en 1948, élargit la définition a un « état de complet bien-être physique, mental et social » (9). Le concept de qualité de vie apparait dans les années 1960 aux Etats-Unis, comme « Conjonction d'une modification du pronostic des maladies, de la considération de l'autonomie du patient et d'un besoin d'évaluation médicale ». Depuis ce temps, la qualité de vie est considérée comme une dimension essentielle de la santé, le concept de « qualité de vie liée à la santé » en étant la conjonction. Elle est un paramètre indispensable pour évaluer les interventions et les prises en charge des patients.

Plus tard, dans les années 90, l'organisation mondiale de la santé se positionne sur la qualité de vie, comme étant un concept global et multidimensionnel. Elle l'a définie en 1993 (9) comme :

> *« La perception qu'ont les individus de leur place dans la vie, dans le contexte de la culture et du système de valeurs dans lesquels ils vivent, par rapport à leurs objectifs, attentes, normes et préoccupations. Il s'agit d'un vaste champ conceptuel, qui englobe de manière complexe la santé physique d'une personne, son état psychologique, son niveau d'indépendance, ses relations sociales, ses croyances personnelles et sa relation avec les spécificités du milieu environnant. Lorsque l'étude de la qualité de vie se limite aux effets sur la santé, on parle de qualité de vie liée à la santé (QVLS) »*

Ce concept est principalement utilisé en épidémiologie et en analyse médico-économique de coût-efficacité (10). En effet, depuis les années 1970, le critère de survie ou de morbidité n'apparait plus suffisant pour quantifier le bénéfice clinique d'un produit de santé (8).

1. Mesure de la qualité de vie et « patient centrisme »

La mesure de la qualité de vie est devenue incontournable dans l'évaluation des produits de santé, ainsi que dans l'évaluation des pratiques médicales des offreurs de soin (10,11). Un changement de perspective s'effectue, d'un point de vue pouvant être paternaliste dans les pratiques cliniques, vers la satisfaction/qualité perçue par les patients dans les différentes étapes du parcours de soin et leur perception des conséquences des décisions médicales dans leur vie.

La mesure subjective de la qualité de vie liée à la santé s'est imposée comme nécessaire dans l'évaluation du bénéfice des interventions de santé, en complément des mesures cliniques objectives, permettant d'évaluer l'impact d'une pathologie ou d'une intervention de santé du point de vue du patient (10,11).

La notion de subjectivité et de « point de vue patient » est ici importante. C'est à dire, il s'agit que le patient puisse, par son ressenti personnel, influer sur l'évaluation d'une pratique médicale ou d'un produit de santé. Là où il s'agissait d'obtenir des critères chiffrés, normés, les plus objectifs et standardisés possibles tels que l'allongement de la survie, ou un score de gravité d'effet indésirable, l'on incorpore un sentiment subjectif d'un usager en santé. Ce changement de paradigme a été évalué comme étant bénéfique à la pratique des soins, ainsi que dans les relations médicales (11). Les outils d'évaluation de qualité de vie utilisés par les professionnels de santé ont même été reconnus comme des aides au recueil de l'anamnèse, de la pratique de la consultation médicale et paramédicales (11).

Le patient est réinvesti d'une partie du pouvoir de décision et son ressenti participe davantage aux décisions médicales/thérapeutiques qui le concernent. C'est le principe de « décision partagée ».

En 2013, un article du *British Medical Journal* (BMJ) titre « *Let the patient revolution begin* » (12). L'article développe les efforts des patients pour sortir d'un paternalisme médical, pour chercher à renouveler les pratiques. Il évoque notamment les populations de patients présents sur internet, comme moteur de ce changement de paradigme. En particulier, les auteurs écrivent : « *Des années de paternalisme ont laissé les médecins et les patients démunis face à un autre type d'interaction […]. Le monde qui se dessine avec l'Internet participatif, où le virtuel l'emporte sur le réel, d'informations à la fois individualisée (où domine le point de vue de l'utilisateur) et communautaire (centré sur le partage, mais aussi sur le croisement statistique), va-t-il offrir l'écosystème capable d'accueillir – enfin – une médecine centrée sur les patients ?* » 8 ans plus tard, c'est exactement dans ce mouvement que le travail de cette thèse s'inscrit. Cependant les auteurs précisent aussi la nécessité d'une vigilance intellectuelle, où les acteurs du domaine de la santé pourraient chercher à exploiter la tendance du « patient-centrisme », où le danger serait que le patient soit captif d'une architecture commerciale et idéologique qui n'aurait rien à envier au paternalisme qu'il aurait dépassé. L'article se conclut d'une remarque :

> *Ne confondons pas « centré sur le patient » et « orientation client ». Être centré sur le malade, pour la médecine, n'est pas une stratégie. C'est la condition de son existence, la démarche d'où elle émerge : son origine.*

2. Les questionnaires d'exploration de qualité de vie liée à la santé

Une approche possible pour décrire la qualité de vie, pouvant être liée à la santé, d'un individu est d'obtenir des mesures rapportées par le patient concernant différentes dimensions de sa propre santé. Ces études ont pour but de fournir une description quantitative des états de santé expérimentés par un malade.

Actuellement, des méthodes de recueil de données de vie réelle qui s'intéressent à la réalité patient (en opposition à un produit de santé par exemple), sont les questionnaires de qualité de vie. Ils rentrent dans la catégorisation conceptuelle définie par la Haute Autorité de Santé (HAS) des *Patients Reported Outcomes Measures* (PROMs) et des *Patients Reported Experience Measures* (PREMs) (11).

Il existe différents questionnaires d'évaluation, explorant des ressentis patients pouvant être généralistes comme l'état moral, ou très spécifiques comme l'intensité d'un déficit fonctionnel ; ou encore un questionnaire spécifique à une maladie. S'il fallait préciser la dichotomie, l'on pourrait dire que les PROMs se concentrent davantage sur des « résultats cliniques » alors que les PREMs se concentrent plus sur le ressenti patient (11).

Ces questionnaires peuvent être transmis par un professionnel de santé, ou directement remplis par le patient. Ils se distinguent des autres résultats en santé par le fait que la mesure est évaluée par le patient, sans interprétation des réponses par un professionnel de santé (11,13). Les PREMs et les PROMs sont des instruments de mesure conceptuelle, selon la perspective du patient, dans sa subjectivité (relatif au sujet pensant), comme la douleur, la fatigue, l'anxiété, et plus largement la qualité de vie liée à la santé. On retrouve ici la notion de s'affranchir du filtre médical pour se rapprocher de la subjectivité, de la réalité patient, développée dans le patient-centrisme.

Il convient de préciser que ces outils doivent être validés de façon adéquate, d'évaluation des propriétés métrologiques et psychométriques. Dans le cadre des études en vie réelle, ils sont une ressource utile pour enrichir l'évaluation de l'impact des technologies de santé selon la perspective des patients.

Parmi les différents questionnaires validés et utilisés dans la pratique courante, nous pouvons citer les deux questionnaires d'évaluation de qualité de vie liée à la santé, l'EQ-5D (*Euro Quality of Life – 5 Dimensions*), et le SF-36 (*Short Form 36 items*). Ce sont ces deux questionnaires, en prenant en compte les faits qu'ils permettent l'exploration et la mesure de la qualité de vie, validés et reconnus par la HAS à ces fins (10,14), qui nous ont permis de commencer à définir les concepts à introduire dans notre approche.

2.1. Le questionnaire Short Form 36 (SF-36)

Le questionnaire générique *Short Form 36* (SF-36) (figure 3) est un exemple de mesure de la qualité de vie, composé de 36 *items* (15). Il permet d'évaluer un profil d'expérience d'état de santé d'un patient à travers 8 dimensions : Le fonctionnement physique, les limitations du rôle liées à la capacité physique, les douleurs physiques, la santé générale, la vitalité, le bien-être social, les limitations liées à la santé mentale, et la santé mentale en elle-même. Chaque dimension propose plusieurs choix au patient : « 1 – très limité », « 2 – légèrement limité », et « 3 – pas du tout limité ». *In fine* le patient obtient un score sur une échelle de 0 à 100.

1.- En général, diriez-vous que votre santé est : (cocher ce que vous ressentez)

Excellente__ Très bonne__ Bonne__ Satisfaisante__ Mauvaise__

2.- Par comparaison avec il y a un an, que diriez-vous sur votre santé aujourd'hui ?

Bien meilleure qu'il y a un an __ A peu près comme il y a un an __

Un peu meilleure qu'il y a un an __ Un peu moins bonne qu'il y a un an __

Pire qu'il y a un an __

3.- vous pourriez vous livrer aux activités suivantes le même jour. Est-ce que votre état de santé vous impose des limites dans ces activités ? Si oui, dans quelle mesure ? (entourez la flèche).

a.Activités intenses : courir, soulever des objets lourds, faire du sport.

Oui, très limité | oui, plutôt limité | pas limité du tout

b.Activités modérées :déplacer une table, passer l'aspirateur.

Oui, très limité | oui, plutôt limité | pas limité du tout

c.Soulever et transporter les achats d'alimentation.

Oui, très limité | oui, plutôt limité | pas limité du tout

d.Monter plusieurs étages à la suite.

Oui, très limité | oui, plutôt limité | pas limité du tout

e.Monter un seul étage.

Oui, très limité | oui, plutôt limité | pas limité du tout

f.Vous agenouiller, vous accroupir ou vous pencher très bas.

Oui, très limité | oui, plutôt limité | pas limité du tout

g.Marcher plus d'un kilomètre et demi.

Oui, très limité | oui, plutôt limité | pas limité du tout

h.Marcher plus de 500 mètres

Oui, très limité | oui, plutôt limité | pas limité du tout

i.Marcher seulement 100 mètres.

Oui, très limité | oui, plutôt limité | pas limité du tout

j.Prendre un bain, une douche ou vous habiller.

Oui, très limité | oui, plutôt limité | pas limité du tout

4.- Au cours des 4 dernières semaines, avez-vous eu l'une des difficultés suivantes au travail ou lors des activités courantes, du fait de votre santé ? (réponse : oui ou non à chaque ligne)

	oui	non
Limiter le temps passé au travail, ou à d'autres activités ?		
Faire moins de choses que vous ne l'espériez ?		
Trouver des limites au type de travail ou d'activités possibles ?		
Arriver à tout faire, mais au prix d'un effort		

5.- Au cours des 4 dernières semaines, avez-vous eu des difficultés suivantes au travail ou lors des activités courantes parce que vous étiez déprimé ou anxieux ? (réponse : oui ou non à chaque ligne).

	oui	non
Limiter le temps passé au travail, ou à d'autres activités ?		
Faire moins de choses que vous ne l'espériez ?		
Ces activités n'ont pas été accomplies aussi soigneusement que d'habitude ?		

6.- Au cours des 4 dernières semaines, dans quelle mesure est-ce que votre état physique ou mental ont perturbé vos relations avec la famille, les amis, les voisins ou d'autres groupes ?

Pas du tout | très peu | assez fortement | énormément

7.- Avez-vous enduré des souffrances physiques au cours des 4 dernières semaines ?

Pas du tout | très peu | assez fortement | énormément

8.- Au cours des 4 dernières semaines la douleur a-t-elle gêné votre travail ou vos activités usuelles ?

Pas du tout | un peu | modérément | assez fortement | énormément

9.- Ces 9 questions concernent ce qui s'est passé au cours de ces dernières 4 semaines. Pour chaque question, donnez la réponse qui se rapproche le plus de ce que vous avez ressenti. Comment vous sentiez-vous au cours de ces 4 semaines :

a. vous sentiez-vous très enthousiaste ?

Tout le temps | très souvent | parfois | peu souvent | jamais

b. étiez-vous très nerveux ?

Tout le temps | très souvent | parfois | peu souvent | jamais

c. étiez-vous si triste que rien ne pouvait vous égayer ?

Tout le temps | très souvent | parfois | peu souvent | jamais

d. vous sentiez-vous au calme, en paix ?

Tout le temps | très souvent | parfois | peu souvent | jamais

e. aviez-vous beaucoup d'énergie ?

Tout le temps | très souvent | parfois | peu souvent | jamais

f. étiez-vous triste et maussade ?

Tout le temps | très souvent | parfois | peu souvent | jamais

g. aviez-vous l'impression d'être épuisé(e) ?

Tout le temps | très souvent | parfois | peu souvent | jamais

h. étiez-vous quelqu'un d'heureux ?

Tout le temps | très souvent | parfois | peu souvent | jamais

i. vous êtes-vous senti fatigué(e) ?

Tout le temps | très souvent | parfois | peu souvent | jamais

10.- Au cours des 4 dernières semaines, votre état physique ou mental a-t-il gêné vos activités sociales comme des visites aux amis, à la famille, etc ?

Tout le temps | très souvent | parfois | peu souvent | jamais

11.- Ces affirmations sont-elles vraies ou fausses dans votre cas ?

a. il me semble que je tombe malade plus facilement que d'autres.

Tout à fait vrai | assez vrai | ne sais pas | plutôt faux | faux

b. ma santé est aussi bonne que celle des gens que je connais.

Tout à fait vrai | assez vrai | ne sais pas | plutôt faux | faux

c. je m'attends à ce que mon état de santé s'aggrave.

Tout à fait vrai | assez vrai | ne sais pas | plutôt faux | faux

d. mon état de santé est excellent.

Tout à fait vrai | assez vrai | ne sais pas | plutôt faux | faux

Wade JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). Medical Care 1992;30:473–483.

Figure 3 : Questionnaire Short Form 36 (SF-36)

2.2. Le questionnaire EuroQoL 5 Dimensions (EQ-5D)

L'EQ-5D est un autre questionnaire générique de qualité de vie dont l'utilisation est recommandée par la HAS (figure 4). Il est le fruit du travail de la société savante européenne sur la qualité de vie : EuroQoL (16).

Il se décompose en deux parties :

Une partie explore cinq dimensions évaluées par le patient selon sa perception propre : la mobilité, l'autonomie de la personne, les activités courantes, la/les douleur(s) et gêne(s) ressentie(s), ainsi que l'anxiété ou la dépression.

La deuxième partie correspond à une échelle visuelle analogique graduée de 0 à 100 sur laquelle le patient place son ressenti de « son état de santé aujourd'hui ».

Le codage du résultat est réalisé de la façon suivante : Chaque dimension a trois réponses possibles (respectivement codées de 1 à 3) : « pas de problème », « quelques problèmes », « problèmes majeurs ». 5 dimensions dont l'intensité de l'impact selon 3 codes résulte en 243 états de santé possibles ($3^5 = 243$).

Ainsi il est possible d'obtenir un code à 5 chiffres : 11111 lorsque le patient n'a « pas de problème » sur l'ensemble des cinq dimensions, ou 33333 s'il ressent des problèmes majeurs dans les différentes dimensions de sa qualité de vie.

L'EuroQoL a par la suite proposé une variante de ce questionnaire, avec cinq niveaux de réponse par dimension, afin d'éviter l'effet de plafonnement dû à un codage à trois chiffres, ainsi que pour améliorer la sensibilité de la détection de faibles variations de qualité de vie (16).

Mobilité

1. Je n'ai aucun problème pour me déplacer à pied.
2. J'ai des problèmes pour me déplacer à pied.
3. Je suis obligé(e) de rester alité(e).

Autonomie de la personne

1. Je n'ai aucun problème pour prendre soin de moi.
2. J'ai des problèmes pour me laver ou m'habiller tout(e) seul(e).
3. Je suis incapable de me laver ou de m'habiller tout(e) seul(e).

Activités courantes

1. Je n'ai aucun problème pour accomplir mes activités courantes (e.g. travail, études, travaux domestiques, activités familiales ou loisirs).
2. J'ai des problèmes pour accomplir mes activités courantes.
3. Je suis incapable d'accomplir mes activités courantes.

Douleurs/gêne

1. Je n'ai ni douleurs ni gêne.
2. J'ai des douleurs ou une gêne modérée(s).
3. J'ai des douleurs ou une gêne extrême(s).

Anxiété/Dépression

1. Je ne suis ni anxieux(se) ni déprimé(e).
2. Je suis modérément anxieux(se) ou déprimé(e).
3. Je suis extrêmement anxieux(se) ou déprimé(e).

Nous aimerions savoir dans quelle mesure votre santé est bonne ou mauvaise AUJOURD'HUI.

- Cette échelle est numérotée de 0 à 100.
- 100 correspond à la meilleure santé que vous puissiez imaginer.

0 correspond à la pire santé que vous puissiez imaginer.

- Veuillez faire une croix (X) sur l'échelle afin d'indiquer votre état de santé AUJOURD'HUI.
- Maintenant, veuillez noter dans la case ci-dessous le chiffre que vous avez coché sur l'échelle.

VOTRE SANTÉ AUJOURD'HUI =

https://euroqol.org/wp-content/uploads/2016/09/EQ-5D-5L_UserGuide_2015.pdf

Figure 4 : Questionnaire Euro Quality of Life 5 Dimension (EQ-5D)

3. *L'infodémiologie*

L'infodémiologie, un terme qui apparait dans les années 2000 suivant les travaux de Gunter Eysenbach (17), est définie comme :

> *« L'ensemble des méthodes et techniques conçues pour mesurer et suivre la « demande » d'informations sur la santé ainsi que « l'offre », dans les contenus numériques présents sur Internet, dans un but général de santé publique. »*

Depuis les années 2000, le champ de recherche de l'infodémiologie a accompagné l'augmentation des usages et l'essor d'internet. Le terme en lui-même a commencé à se démocratiser et l'on peut retrouver sa définition dans des articles de presse médicale. Le moniteur des pharmacies (2018) développe en particulier, qu'elle peut être définie comme la science basée sur des technologiques de l'information et des communications dont l'objectif est de surveiller l'état de santé des populations et d'orienter les politiques de santé publique (18). Les médias sociaux, qui regroupent réseaux de contact professionnels ou généralistes, forums, blogs, réseaux de contenus et autres, permettent aux patients et à leur entourage d'échanger directement, de rechercher de l'information médicale ou de décrire leurs symptômes sans passer par des professionnels de santé. Ils constituent pour des épidémiologistes une source informelle et complémentaire de données, apportant des informations de santé non recueillies par ailleurs ou révélant des points de vue sur des sujets liés à la santé (19). Par exemple les données de Twitter peuvent être employées pour surveiller l'émergence et l'évolution de maladies infectieuses comme la grippe (20).

Nous avons pu définir ce qu'est l'infodémiologie. Nous allons maintenant aborder la partie pratique, applicative d'un point de vue technique.

La condition *sine qua none* à la récupération des données est l'autorisation d'accès. En effet, si une page ou si un compte est privé, fermé ou verrouillé, ses données ne doivent pas être accessibles et récupérables par un tiers. Nous n'aborderons ici que les cas des données accessibles.

Il existe différentes méthodes pour extraire des données présentes sur internet (21), que ce soit une page normale ou un réseau social. Les deux principales sont les *Application Programming Interface* (API) et le *crawling* (que l'on pourrait traduire en français par robot d'exploration et de moissonnage de données). Dans le cas de l'API, le site ou réseau social lui-même, offre un accès direct à sa base de données par une interface dédiée, ce qui lui permet de contrôler les conditions d'accès d'une machine (ordinateur) à ses données. Ce fonctionnement permet de créer un marché de la donnée pour celui qui en est le propriétaire. C'est par exemple le cas du réseau social Twitter. Cette technique a l'avantage d'accéder aux données ciblées, mises à disposition par le propriétaire, sans avoir besoin d'accéder à la base de données sources.

Le *crawler* est un logiciel qui a pour fonction de lire le code d'une page internet et d'en extraire les données. Les *crawlers*, fonctionnent à l'inverse des API qui sont mis à disposition des sites internet. Ils sont développés par l'utilisateur qui cherche à récupérer les données présentes sur les sites. Afin de contrôler l'accès à leurs données, certains sites peuvent entraver l'utilisation de *crawlers*, obligeant à passer par leur API s'ils en disposent. De plus, un *crawler* exploite l'architecture d'une page, en termes de code, pour identifier et récupérer les données. Dès lors, dès la moindre modification de la page, le *crawler* sera rendu inutilisable.

Pour le cas des réseaux sociaux, blogs et forums, leur utilisation par des communautés d'internautes repose sur l'acceptation des conditions d'utilisations. Ces conditions, lorsqu'elles sont acceptées en l'état, garantissent le fait que les données générées par l'utilisateur sont exploitables par le site et ses partenaires. Ces paramètres peuvent par la suite être modifiés, notamment lors du passage à un compte « privé ». Par exemple, si l'on recherche des messages postés sur le réseau social Twitter, l'API du site livrera les publications publiques correspondant à la requête spécifique (ex : présence d'un mot clef), mais pas celles des comptes étant « privés ».

L'intérêt scientifique et médical concernant le champ de recherche qu'est l'infodémiologie augmente constamment. En effet, une revue (22) réalisée en 2020 montre que sur 338 études infodémiologiques incluses, le nombre d'articles augmente chaque année, les publications de 2017 et 2018 comptant pour plus de la moitié des publications de la décennie (22). La source de données la plus populaire était Twitter, suivie par Google, venaient ensuite les autres sites et forums. Le réseau social Facebook était marginal derrière ces sources. Concernant les sujets d'études, les études infodémiologiques abordaient de façon majoritaire les maladies (au sens général), ainsi que les épidémies. Suivies par les études centrées sur le soin et la santé publique, les médicaments, le tabagisme et l'alcool (22). La revue conclut sur le fait que l'infodémiologie est un moyen innovant de réaliser des évaluations sanitaires à différentes échelles. Les données issues d'internet ont aussi l'avantage d'être directement récoltables et analysables, là où d'autres méthodes traditionnelles peuvent être extrêmement chronophages (22).

Ainsi, l'on commence à discerner que l'infodémiologie, alliée au développement du mouvement du patient-centrisme, peut représenter une opportunité pour capter et analyser de nouvelles données, directement issues des patients, à propos de thématiques de santé.

4. Données de vie réelle

Cette notion de source de données exploitables et complémentaires pour les épidémiologistes, nous amène à évoquer les données de vie réelle (figure 5) (23).

Définies par la Haute Autorité de Santé (HAS), les données de vie réelle (ou de vraie vie), sont : « Les données qui par définition ne sont pas collectées dans un cadre expérimental, mais générées suite à la réalisation des soins et actes médicaux en routine pour les patients, reflétant, *a priori*, la pratique courante » (24). Elles peuvent être collectées et analysées de manière spécifique, comme pour des analyses de pharmacovigilance, pour constituer des registres ou des cohortes, ou encore dans le cas d'études *ad hoc* sur un sujet particulier. Leurs sources peuvent être multiples, comme les dossiers informatisés de patients ou encore les analyses des consommations de soins et produits de santé. La HAS et le ministère des solidarités et de la santé précisent qu'elles peuvent également provenir du web, des réseaux sociaux et des objets connectés (24).

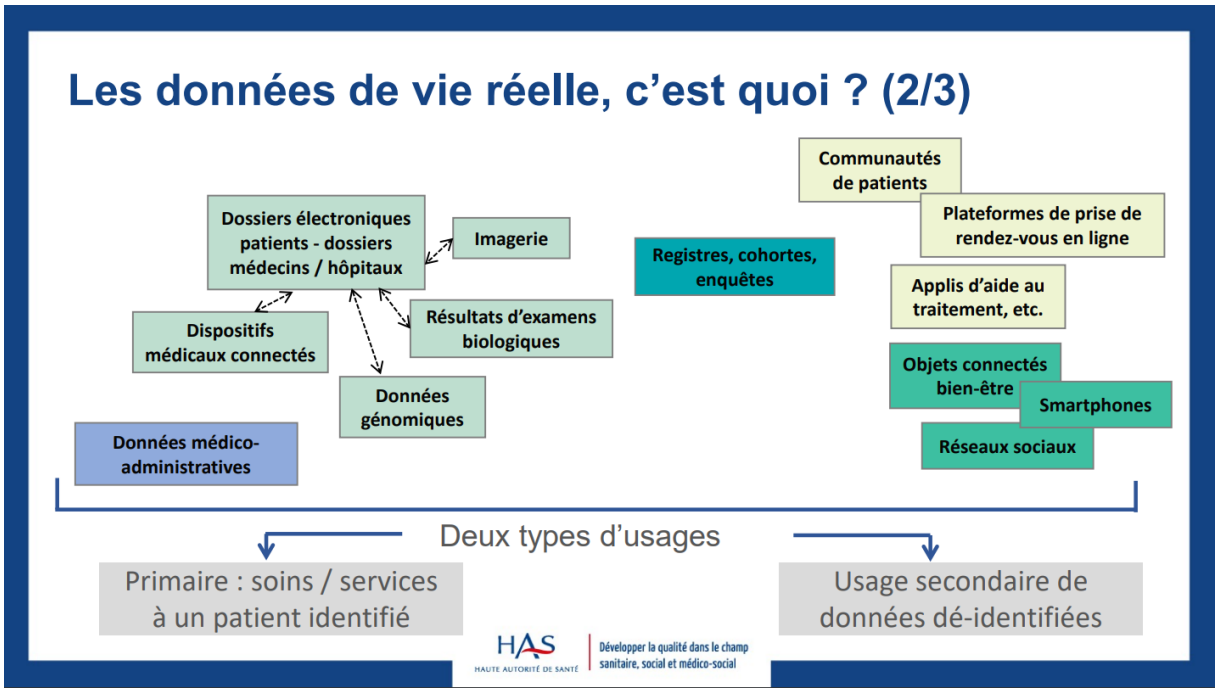


Figure 5 - Haute Autorité de Santé, Données de vie réelle : un enjeu majeur, une dynamique qui s'accélère – source des données de vie réelle. (2019) (24)

Un article publié dans le *Journal of American Medical Association* (JAMA), titre « Données de vie réelle, qu'est-ce que cela signifie vraiment ? » (25), il présente un tableau avec les différentes sources des données de vie réelle, leurs avantages et leurs limites. Le tableau 1, ci-dessous, est traduit et adapté de cette publication :

Tableau 1 : Limites et avantages des données de vie réelle (traduit et adapté des travaux de Nabhan et al. Real world evidence, what does it really mean ? (25))

Source	Description	Avantages	Limites
Bases de données médico-administratives	Dépenses médicales en ville et en hôpital, dépenses liées aux produits vendus en pharmacie, collectées dans un but de remboursement	• Recueillir les consultations ou hospitalisations dans de multiples établissements de soins • Echantillons de grandes tailles accessibles sous conditions • Sont généralement représentatives de la population cible • Contiennent des données relatives à la démographie des patients, aux diagnostics, aux procédures et aux médicaments, ainsi que les dates, le lieu de service et les coûts associés • Permettent de créer le parcours longitudinal de soins d'un patient donné dans le système de soins	• Les critères d'évaluation cliniques importants (facteurs de risque, progression, décès, autres) ne sont pas souvent disponibles, ce qui oblige à recourir à des approximations (ou « proxy ») • Données non initialement collectées à des fins de recherche • Ne tiennent pas toujours compte des opérations qui ne sont pas facturés ou non soumises à un remboursement • Perte de suivi possible lorsqu'un patient change d'employeur et donc de payeur
Comptes rendus médicaux	Données dérivées des dossiers médicaux des patients (ex : dossiers papiers, fichiers électroniques) incluant les	• Contiennent des données cliniques non disponibles parmi les demandes de remboursement • Peut contenir la raison sous-jacente à la consultation médicale, le besoin patient, lorsque l'examen est conduit par le prescripteur	• En général restreintes au lieu de soin (cabinet, service, centre ou établissement hospitalier) • Difficultés d'accès aux données non structurées (rapports scannés, notes médicales, résultats d'analyses biologiques), en raison de la protection des données de santé/données

	données démographiques, les critères cliniques et leurs évolutions, les médicaments, les procédures, les résultats de laboratoire, l'imagerie et les notes des professionnels de santé		personnelles, ou à cause des limites du traitement du langage naturel • Les données non structurées ne sont pas toujours facilement disponibles dans les fichiers électroniques et peuvent ne pas être récupérées en temps utile
Registres	Répertoire de caractéristiques cliniques, démographiques et des caractéristiques particulières basées sur une maladie spécifique ou sur l'utilisation d'un produit spécifique	• Sont une source standardisée, de caractéristiques de patients et de résultats cliniques uniformément collectés • Permettent de longs suivis historiques des patients • Les éléments correspondent généralement aux standards nécessaires aux objectifs et protocoles de recherche médicale • Améliorent la compréhension de l'histoire naturelle des maladies, de l'efficacité clinique et de la qualité de vie • Permettent de découvrir des éléments tels que des effets indésirables dans des populations de patients sous-étudiées	• Les données manquantes sont fréquentes • Représentent uniquement les échantillons de patients inclus • Sont coûteux à maintenir • Les traitements ne sont pas toujours standardisés entre les patients inscrits • Manque d'évaluation uniforme de la réponse au traitement de la progression de la maladie
Données générées par les patients	Données générées directement par les patients et leurs activités (enquêtes, questionnaires, objets connectés, …). Les résultats de ces données sont centrés sur les patients et leurs activités	• Informe sur le point de vue direct du patient, sans filtre/interprétation par un professionnel de santé • Permettent de fournir des informations sur la qualité de vie, souvent absente des autres sources	• Manque de validation des méthodes de recueil et d'analyse, ce qui peut limiter les interprétations • Manque souvent des critères importants relatifs aux patients et des données cliniques • Données déclaratives dont la subjectivité peut parfois être différente d'une réalité objective
Réseaux sociaux	Données de vie réelle générées par les patients qui partagent leurs expériences lorsqu'ils sont diagnostiqués avec une maladie ou lorsqu'ils évoquent le fardeau de leurs symptômes ou l'expériences d'effets indésirables	• Peut améliorer la compréhension des freins & obstacles à l'adhésion des patients • Permet d'évaluer la connaissance des patients sur un sujet de santé • Rapprocher les patients atteints de maladie rares malgré de distances géographiques fortes • Fournissent des informations représentatives des expériences de patients, directe et non filtrées	• Souvent limitées à des données qualitatives • Manque de report uniforme des informations • Peuvent manquer de caractéristiques importantes à propos des patients • Résultats cliniques non confirmés • Questionnement sur l'authenticité des informations

Depuis les années cinquante, la méthodologie des essais cliniques randomisés en double aveugle est à la base de la médecine fondée sur les preuves. Elle est considérée comme l'étalon en matière d'évaluation médico-scientifique des produits de santé en vue de leur autorisation de mise sur le marché (AMM). Cependant, ce qui constitue les atouts des essais cliniques randomisés est aussi un biais en soit. En effet, la rigueur d'un schéma expérimental est intrinsèquement une limite pour prédire l'usage et les effets d'un produit de santé une fois qu'il sera mis sur le marché (23). Après l'accès au marché, le produit de santé, et ce malgré le cadre de son autorisation de mise sur le marché, aura sa « vie propre ». C'est-à-dire que les populations de patients qui l'utiliseront/auxquelles il sera prescrit pourront différer des populations sélectionnées dans l'essai clinique avec ses critères stricts d'inclusion et d'exclusion, c'est le principe de population rejointe (14). De même, les dosages, durées de traitements prescrits pourront être adaptés par les professionnels de santé. Côté patient, cela ouvre aussi la porte à l'adhésion des usagers de santé, à leur compliance aux soins, à leur observance thérapeutique. Les anglo-saxons ont même des termes différents pour désigner l'efficacité évaluée dans les essais cliniques (*efficacy*), et l'efficacité dans les conditions réelle d'utilisation en pratique courante (*effectiveness*). Au même titre, le terme « *real world evidence* » est employé concernant les analyses scientifiques que l'on peut faire des données de vie réelle et aux conclusions que l'on peut en tirer.

Biais des études en vie réelle

Comme toute méthodologie d'étude médicale, les données et protocoles d'études sont inséparables de certains biais intrinsèques. Leur connaissance et leur prise en compte est un prérequis nécessaire avant toute méthodologie d'étude et interprétation de leurs résultats. Ces biais peuvent être spécifiques, propres au système de santé, ou encore technologiques. Au-delà de la question de la représentativité, d'autres biais influencent la robustesse et les interprétations liés à l'analyse des données de vie réelle (26). Ces biais sont parfois partagés avec ceux des études observationnelles :

- Les biais de sélection : Lors de la constitution des groupes d'individus, ou de leurs données, se pose la question de l'extrapolation des résultats à la population générale, ou en cas de comparaison entre deux groupes de patients comme dans le cadre d'études exposés/non exposés ou d'études cas-témoins. Ces biais sont corollaires de l'absence de randomisation des patients, cependant des méthodes statistiques existent pour équilibrer ce biais, tel que le score de propension.
- Les biais de confusion : l'absence de prise en compte d'un facteur d'imputabilité lors d'une analyse peut entraîner une erreur d'appréciation dans la causalité, entre un potentiel facteur étudié et l'évènement de santé observé.
- Les biais d'information : pouvant résulter d'une mauvaise définition, d'un manque de mémorisation des patients concernant les informations qu'ils délivrent, ou encore d'une subjectivité de l'enquêteur (intensification des recherches sur un sujet particulier, ou en raison de ses connaissances).

Cependant, la Cochrane collaboration conclut que lorsque les méthodologies d'études sont bien conçues, les résultats entre études observationnelles et essais cliniques randomisés sont comparables (27).

Pour aller plus loin, une étude américaine, aussi publiée dans le *JAMA*, s'est intéressée aux points de concordance entre des données de vie réelle, issues des données de remboursement de santé et des comptes rendus de consultation, et celles des essais cliniques (randomisés ou non) (28). 220 essais cliniques américains ont été inclus dans l'analyse. Les chercheurs ont évalué le nombre d'essais pour

lesquels l'indication, l'intervention, les critères d'inclusion et d'exclusion ainsi que les critères d'évaluation principaux, auraient pu être déterminés à partir de données d'assurance et/ou des comptes rendus médicaux. Les résultats de ces travaux montrent que 15% des essais cliniques inclus dans l'analyse auraient pu être reproduits à l'aide des données observationnelles/de vie réelle (28). Le fait d'utiliser des données préalablement recueillies en vie réelle, a un nom, il s'agit d'un essai clinique embarqué (29).

Ces études et consensus vont dans le sens d'une complémentarité des études en vie réelle avec les essais cliniques. En pratique, elle permettrait de confronter la population à l'essai avec la population en vie réelle, ou encore de définir des indicateurs à suivre dans l'essai clinique, qui auraient pu apparaître en vie réelle. Elles peuvent déjà permettre d'évaluer si les conditions de réalisation des essais cliniques sont validées, de définir la population rejointe (c'est-à-dire la population réellement traitée), posologies, conditions de prescription, observance, etc (23). Selon la capacité de recueil d'informations, elles peuvent aussi porter sur des populations numériquement plus importantes, pouvant permettre d'identifier des effets non repérables dans un essais clinique randomisé, du fait des effectifs limités de patients inclus. De plus, les données de vie réelles offrent une opportunité de suivi à long terme, là où les essais cliniques sont par définition réalisés sur une durée fixée et plus courte (23). Il est communément admis qu'en matière de sécurité et tolérance des produits de santé, la matério et pharmacovigilance sont essentielles et obligatoires et s'effectuent tout au long de la vie du produit, quelles que soient les modalités d'usages.

Les données de vie réelle se sont aussi illustrées par leur capacité à identifier des signaux faibles, précurseurs, ou encore avant-coureurs, utiles en terme de santé publique pour identifier une tendance ou des évènements importants (30). Un exemple concret, est l'étude réalisée en 2018 par la société Kap Code (31), qui a analysé des témoignages issus des réseaux sociaux concernant la molécule méthylphénidate (médicament psycho-actif, prescrit en France dans la prise en charge du Trouble Déficitaire de l'Attention avec Hyperactivité chez l'enfant). L'Agence Nationale de Sécurité du Médicament (ANSM) constate qu'une utilisation hors Autorisation de Mise sur le Marché (AMM) de ce médicament continue d'être observée (32). L'étude infodémiologique de Kap Code a inclus 3 443 commentaires publics postés sur les réseaux sociaux de 2007 à 2016. 61 effets secondaires ont été identifiés dans le corpus d'analyse des messages. Les résultats ont été évalués par deux experts en pharmacovigilance, 67% des effets indésirables détectés dans les messages étaient de réels signaux de pharmacovigilance. Les patients reportaient principalement des symptômes neuropsychiatriques ainsi que des palpitations cardiaques. Au-delà des symptômes ressentis, l'analyse des messages (par méthode de Traitement Automatisé du Langage), a identifié des cas de mésusages du méthylphénidate. En effet, un usage hors AMM, du méthylphénidate a été observé dans les populations étudiantes et adultes, où ce composé et ses effets étaient comparés à ceux des amphétamines, dans un but récréatif ou pour augmenter les capacités cognitives. Le poster médical (figure 6) adapté de cette étude est disponible ci-dessous.

Cet exemple pratico-pratique qui s'intéresse à l'usage en vie réelle d'un médicament, est une des nombreuses études qui démontrent l'opportunité que les réseaux sociaux représentent pour capter des données de vie réelle. En objectivant des tendances et des comportements sanitaires, l'analyse médicale des réseaux sociaux peut être un outil utile aux pilotages des politiques publiques en santé.

ÉTUDE DE L'USAGE DU MÉTHYLPHÉNIDATE SUR LES RÉSEAUX SOCIAUX

Pierre Foulquié[1], Paméla Voillot[1], Carole Faviez[1],
Adel Mebarki[1], Xiaoyi Chen[2], Nathalie Texier[1], Stéphane Schück[1]

[1]Kap Code, Paris, France - [2]INSERM, UMRS1138, équipe 32, Centre de Recherche des Cordeliers, Paris, France

KAP CODE

INTRODUCTION

Le **méthylphénidate**, prescrit en France à plus de **49 000 patients** atteints de troubles du déficit de l'attention avec hyperactivité (TDAH) âgés de 6 à 17 ans, a fait l'objet de plusieurs évaluations par l'EMA et l'ANSM concernant son usage hors AMM.

Parallèlement, la FDA et l'EMA ont reconnu les réseaux sociaux comme nouvelle source de données pour renforcer la surveillance des médicaments [1][2].

Dans ce contexte, l'objectif de cette étude consistait en **l'exploration du comportement** des patients vis-à-vis du **méthylphénidate** à travers les réseaux sociaux.

MATÉRIEL & MÉTHODES

Les messages ayant trait au **méthylphénidate** ont été extraits à partir de cinq forums français généralistes via l'outil Detec't [3], développé par la société Kap Code.

L'analyse s'est déroulée en deux étapes :

1. Premièrement, un modèle de sujet (*topic model*) était appliqué pour identifier les thèmes de discussions (*topic*) relatifs au **méthylphénidate**. Ces derniers permettaient de classer les messages selon les thématiques qu'ils contenaient.

2. Les messages associés aux thématiques qui se rapportaient aux effets secondaires, aux mésusages et aux abus du **méthylphénidate** faisaient ensuite l'objet d'une classification hiérarchique descendante (CHD), directement appliquée aux mots. Différentes sous-thématiques étaient ainsi obtenues.

RÉSULTATS

Les corpus d'analyse méthylphénidate était constitué de **3 343 messages** publiés **entre 2007 et 2016**.

L'évolution temporelle du volume de messages postés (fig.1) n'a pas permis en général pas de mettre en évidence de pics aux dates clés associées au profil de sécurité du **méthylphénidate**. Ces dates consistaient en des décisions à l'échelle française (2006, 2011, 2012 et 2013) et européenne (2007 et 2009). Seule la décision de la HAS (2012) semble avoir eu une influence sur le volume de messages observé.

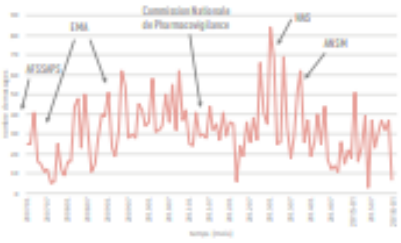


Figure 1 – Évolution temporelle des messages associés au méthylphénidate

La création d'un **nuage de mots** a permis d'observer les termes les plus présents dans le **corpus**.

Plusieurs champs lexicaux tels que ceux de la TDAH, de la prescription et de la prise, de l'enfance ainsi que des inquiétudes concernant les **effets indésirables** ont pu être identifiées.

Le nuage de mots (fig.2) indique une plus grande fréquence d'utilisation du méthylphénidate chez les enfants, en accord avec l'indication du médicament.

Figure 2 – Nuage de mots des messages méthylphénidate

Par la suite, l'application d'un topic model a permis de mettre en avant 14 thématiques de discussions principales (fig. 3).

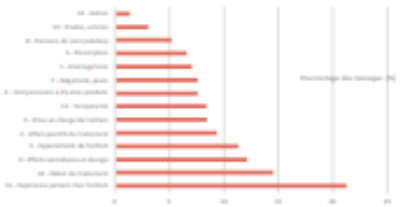


Figure 3 – Thématiques du topic model

Parmi les topics principaux, un topic était centré sur les « **effets secondaires** » (**12,15%** des messages), deux sur le mésusage (intitulés « **Comparaison aux autres produits** » et « **Négativité et peurs** », **7,6%** des messages chacun) et un sur l'usage chez les adultes (**5,2%** des messages). Les autres concernaient des usages cohérents dans le cadre de l'AMM.

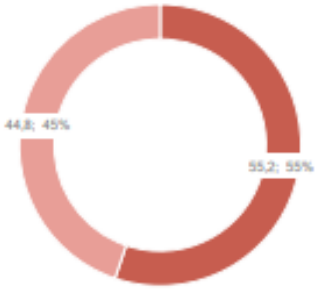
La CHD appliquée au topic d'« **effets secondaires** » (fig.4) a mis en évidence plusieurs troubles liés à la prise du traitement tels que la perte de poids ou encore les vomissements.

effet du traitement · posologie · effet chez l'enfant · durée du traitement · effets secondaires · début du traitement

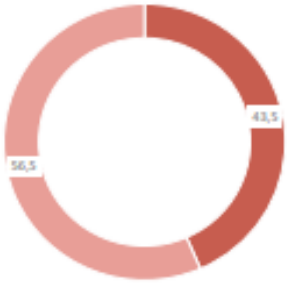
Figure 4 – Thématiques de la CDH du topic « effets secondaires »

Concernant les topics du mésusage (fig. 5), la méthode a permis d'observer l'utilisation de la molécule par les adultes ou les étudiants pour un usage récréatif ou pour augmenter leurs capacités.

De plus, les patients comparaient les effets de la molécule à certaines drogues telles que les amphétamines. Enfin, le topic sur l'usage par les adultes a mis en avant des erreurs ou des retards de diagnostic de TDAH.



relations avec l'Abilify · autisme



les dangers du methylphénidate

comparaison aux stupéfiants

Figure 5 – Thématiques de la CDH des deux topics « mésusage »

CONCLUSION

Cette étude a permis d'identifier :

- Des **effets indésirables**, ce thème constituant un thème d'importance au sein des discussions ;
- Des pratiques non identifiables via les systèmes de soins classiques :
  - l'usage du **méthylphénidate** par des populations spécifiques telles que les adultes et les étudiants.
  - des utilisations hors AMM : comme **substitution aux drogues** et en tant que **psychostimulant**.

La surveillance des réseaux sociaux apparait comme un outil d'alerte complémentaire dans la surveillance des médicaments.

RÉFÉRENCES

[1] www.fda.gov/ScienceResearch/SpecialTopics/RegulatoryScience/ucm452304

[2] www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000258

[3] Abdellaoui R et al. 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 and surveillance, 2017, vol. 3, no 2.

Figure 6 : Poster médical - Etude de l'usage du méthylphénidate sur les réseaux sociaux.

PARTIE 2 : LE PROJET EN PRATIQUE

Maintenant que les différentes notions clefs ont été introduites, nous allons pouvoir nous attacher à développer le projet en lui-même, sa conception et méthodologie, ses objectifs et les moyens d'évaluations qui ont été utilisés. Avant toute chose, ce travail spécifique a fait l'objet d'une publication dans le *Journal of Medical Internet Research* en 2022 (19), disponible en annexe. Le propos développé ci-après, rédigé avec un angle moins technique que dans l'article, s'attachera à retracer la méthodologie globale du projet, en décrivant la place du pharmacien et sa valeur ajoutée, ainsi que la synergie de son duo avec un *data scientist* dans la conduite du projet.

1. Problématisation du sujet

La société Kap Code est une start-up française dédiée à l'analyse médicale des réseaux sociaux. Ce travail de thèse est issu d'un stage de 6 mois de recherche et développement au sein de cette société. Kap Code utilise des algorithmes de TAL et des ontologies médicales et thérapeutiques permettant d'identifier les médicaments (dénomination commune internationale, molécule), effets indésirables, concepts médicaux, ou encore de classifier les différents sous-sujets liés à un thème médical abordés par les patients. L'algorithme d'évaluation des impacts de QdV était un des nombreux projets d'algorithmes en phase de recherche et développement. La caractérisation du profil d'impact de QdV était un enjeu majeur pour la start-up, dont les analyses servent à mieux comprendre la réalité patiente, leurs difficultés, besoins, dans un but de proposer des services et informations les plus utiles possibles.

La qualité de vie et ses impacts sont étudiés et mesurés. Cependant l'analyse des résultats est conditionnée aux objets évalués. L'analyse des verbatims libres présents sur les réseaux sociaux est complémentaire, dans le sens où l'on ne pose pas de question standardisée aux patients, mais que l'on analyse ce qui est décrit de façon naïve. Le pendant de cette approche est la difficulté à standardiser les résultats, dans le sens où il existe différentes façons de faire référence, ou décrire, une même chose. Par exemple, il est possible de décrire un impact sur le moral par la pose d'un diagnostic de trouble de l'humeur, par la présence de certains symptômes, ou encore par des expressions courantes (ex : avoir le moral dans les chaussettes). De même, un impact physique peut être décrit par la description d'un symptôme (exemple : mes jambes sont faibles), ou par la description d'une incapacité à réaliser une action (exemple : j'ai besoin d'aide pour prendre ma douche).

La première étape du projet, a été de définir comment traduire le langage patient et ses différentes expressions possibles, en impact de qualité de vie. Parce que la Haute Autorité de Santé recommande l'utilisation des questionnaires EQ-5D et SF-36, ce sont ces outils qui ont servi à établir les bases de notre approche, sur les dimensions de la qualité de vie à appréhender. La définition de 1993 de l'OMS a aussi grandement contribué à établir le cap du projet.

La partie suivante développe comment une approche mixte de QdV a été créée pour analyser les impacts décrits dans les commentaires libres de patients. En reprenant les différentes définitions et identifiants les consensus des différentes parties prenantes.

2. Étapes clefs

L'ensemble des étapes clefs a été prédéfini avant de démarrer le projet (figure 7).

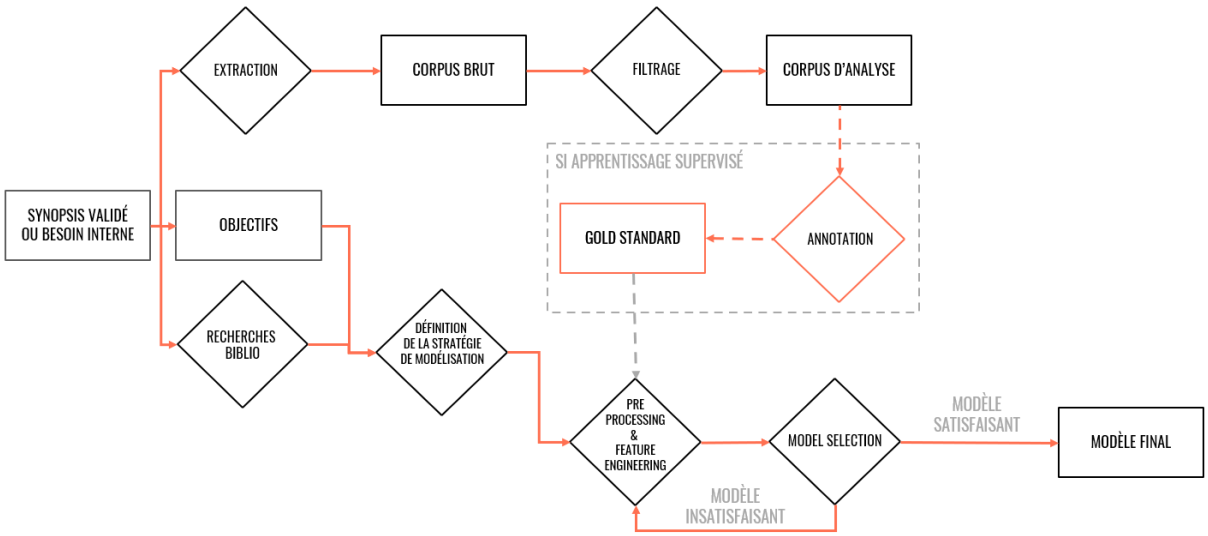


Figure 7 - Etapes clefs du projet.

La partie Synopsis ou besoin interne, correspond ici à la nécessité de développer un algorithme de détection des impacts de qualité de vie. A partir de la définition du besoin, 3 étapes parallèles suivent. La définition des objectifs associés, qu'ils soient quantitatifs (performances) ou qualitatifs (quels types d'analyse sont à réaliser). L'extraction correspond à la récupération de messages pertinents, évoquant au moins une thématique de santé, à partir des réseaux sociaux, blogs et forums. Bien sûr, la partie recherche bibliographique correspond à une recherche de la littérature scientifique, avec comme objectif d'identifier des projets semblables à celui en cours de réalisation afin de profiter des connaissances et de l'état de l'art. Cette partie a concerné à la fois des travaux d'analyse médicale de données textuelles, l'analyse médicale des réseaux sociaux, ainsi que les définitions et consensus portant sur la qualité de vie liée à la santé. Concernant les analyses médicales des réseaux sociaux, une emphase particulière a été donnée à l'identification des méthodes/outils algorithmiques utilisés ainsi qu'à leurs performances associées. Ceci ayant pour but de comparer les méthodes entre elles et de tester celles les plus pertinentes pour notre projet, tout en visant les meilleures performances possibles.

La partie annotation correspond à la labellisation humaine des messages identifiés et nettoyés pour être exploitables par ordinateur. Cela correspond à l'annotation, par message, de la présence ou de l'absence d'un impact de qualité de vie, qui permet d'aboutir à un *Gold standard*, ou standard étalon servant de base de référence pour l'ensemble de la partie algorithmique.

L'ingénierie des variables correspond à l'identification des paramètres les plus pertinents, jouant un rôle dans l'identification des schémas clefs (verbatims spécifiques, temps des verbes, pronoms personnels, etc.).

Cette étape peut bénéficier de boucles d'amélioration continue, où il est possible de tester différentes combinaisons de variables entres elles, et de voir l'influence de ces différentes combinaisons sur les performances de l'algorithme. Cette étape peut aussi s'appeler « optimisation des hyperparamètres ».

L'évaluation des performances peut se faire selon différents indicateurs, certains propres à l'intelligence artificielle, et d'autres communs avec le monde de la santé comme l'Aire Sous la Courbe (AUC), la sensibilité et la spécificité, ainsi que les valeurs prédictives positives et prédictives négatives. La sélection des indicateurs d'évaluation fait partie intégrante d'un projet de développement d'algorithme d'apprentissage automatique, car il permet d'aboutir au modèle final.

3. Approche et méthodologie médicale de la QdV

2.1. Différentes définitions cohabitent

A l'heure actuelle, des standards validés servent à explorer la qualité de vie liée à la santé et les ressentis patients (11). Cependant la définition même de la QdV se heurte toujours à certains problèmes conceptuels (33,34). En effet, des notions clefs cohabitent, beaucoup de concepts sont retrouvés dans différentes définitions, alors que d'autres semblent en être absents.

Exceptée pour la définition de l'OMS, il ne semble pas y avoir de consensus établi pour définir la qualité de vie de manière précise, tous s'accordant à dire qu'il s'agit d'une définition complexe, propre à chaque individu en fonction de sa propre perception, propre définition, dans le contexte socio-culturel dans lequel il vit.

En effet, selon les auteurs et sociétés savantes, il existe plusieurs différences au niveau des définitions (34): la qualité de vie serait définie en termes de position de vie, de fonctionnement, de sentiment à propos du fonctionnement, d'existence et de différence entre le soi actuel et le soi idéal. Lorsque la qualité de vie est liée à la santé, elle est davantage définie en termes de fonctionnement, de sentiment à propos du fonctionnement, de santé et de valeur accordée à la durée de vie.

A titre d'exemple, on peut trouver plusieurs définitions publiées et validées, de la qualité de vie (34) :

- *« La satisfaction ressentie par un sujet dans les différents domaines de sa vie* » Patrick & Erickson, 1987)
- « *Sous l'angle individuel, c'est ce que l'on se souhaite au nouvel an : non pas la simple survie, mais ce qui fait la vie en bonne santé, amour, succès, confort, jouissance, bref le bonheur* » (Fagot-Largeault, 1991)
- « *La perception qu'un individu a de sa place dans l'existence, dans le contexte culturel et du système de valeurs dans lequel il vit, en relation avec ses objectifs, ses attentes, ses normes et ses inquiétudes* » (OMS, 1993)
- « *Un état d'équilibre : équilibre entre plaisirs et contraintes ; équilibre entre aspirations et possibilités du moment* » (Kemoun et al. 1996)
- « *La vie est de qualité quand la vie fait sens* » (Corten, 1998)
- « *La combinaison du bien-être objectivement et subjectivement indiqué dans de multiples domaines de la vie considérés comme saillant dans la culture et le temps tout en adhérant à des standards universels des Droits de l'Homme* » (Wallander et al. 2001)

Des comparatifs entre les différents outils validés d'évaluation de la qualité de vie, montrent que des dimensions sont communes à plusieurs outils, certains étant plus spécifiques sur des points particuliers (Tableau 2).

Tableau 2 : Comparaison des différentes dimensions de qualité de vie, telles qu'explorées par différents questionnaires validés (10)

Instrument	*Instrument domains*								
	Physical function	Symptoms	Global judgement of health	Psychological well-being	Social well-being	Cognitive functioning	Role activities	Personal constructs	Satisfaction with care
Preference-based measures									
15-D	x	x		x	x	x	x		
EQ-5D	x	x	x	x	x		x		
HUI3	x	x		x		x			
SF-6D	x	x		x	x		x		
Generic measures									
FSQ	x		x	x	x		x	x	
NHP	x	x		x	x		x		
SF-36	x	x	x	x	x		x		
SF-20	x	x	x	x	x		x		
SF-12	x	x	x	x	x		x		
SIP	x	x		x	x	x	x		

De manière générale, les dimensions de la qualité de vie les plus retrouvées, à la fois en termes de composante empirique et dans les définitions conceptuelles, concernent les dimensions sociale et relationnelle, le bien-être psychologique et l'humeur, l'emploi, la capacité d'un individu à s'autodéterminer, son autonomie en tant qu'acteur de sa vie propre, la compétence personnelle et l'intégration communautaire.

Lorsque la qualité de vie s'allie à la santé, quatre dimensions principales sont retrouvées :

- La dimension physique : capacité physique, autonomie, gestion des activités quotidiennes, …
- La dimension psychologique : dépression, anxiété, gestion des émotions, …
- La dimension somatique : symptômes, douleurs, asthénie, sommeil, …
- La dimension sociale : environnement familial, professionnel, amical, activités de loisirs, vie sexuelle, …

Ainsi, ces 4 dimensions au moins devaient figurer comme grandes catégories de notre approche. De par l'expertise de Kap Code dans l'analyse du vécu patient via les réseaux sociaux, des réunions d'idéation ont été réalisées avec les différents chefs de projets amenés à analyser les messages. L'objectif était de confronter les grandes catégories identifiées, à la fois la réalité terrain de ce que les patients peuvent décrire dans leurs commentaires, ainsi qu'à l'adéquation du besoin en termes d'analyse.

Suite à ces discussions, les quatre dimensions ont été évaluées comme pertinentes à inclure dans l'algorithme. Une autre dimension a été identifiée au cours de ces réunions : la dimension financière. En effet, celle-ci peut apparaître accessoire par rapport aux quatre précédentes que l'on peut juger plus « cliniquement pertinentes », mais l'expertise des chefs de projets a identifié que dans certains cas, la maladie peut avoir un impact financier. En effet, certains patients témoignent du fait qu'à force d'invalidité, des pensions ou allocations financières sont nécessaires. C'était par exemple le cas de certaines populations de patients qui discutaient de la difficulté de faire reconnaître leur maladie comme Affection Longue Durée (ALD), garante d'un meilleur remboursement de leurs frais de santé. De même, la recherche d'approches holistiques ou de médecines « douces / alternatives » non remboursées peuvent peser sur le budget d'une personne qui cherche à alléger le fardeau de sa maladie. De façon beaucoup plus classique, l'achat de médicaments ou produits non remboursés en pharmacie participe aussi à l'impact financier de la maladie. De manière générale, dans les études déjà menées par Kap Code et sur plusieurs maladies différentes, des commentaires de patients témoignent de la nécessité d'avoir un « budget maladie » ou encore d'une difficulté à « boucler les fins de mois »

en lien avec la gestion de la maladie. Ces réunions, analyses préexistantes et messages patients, ont motivé l'ajout de la dimension financière comme cinquième dimension de notre approche. Il s'agit d'un exemple où le ressenti, les témoignages patients, ont permis de définir une méthodologie d'analyse de l'expérience patiente la plus adaptée possible.

2.2. Les dimensions de Qualité de vie retenue

L'objectif de l'algorithme est de reconnaître dans les verbatims patients, les impacts de qualité de vie sur les cinq dimensions retenues, physique, psychique, activités quotidiennes, relationnelle, financière. Ainsi, il convient de prédéfinir tous les impacts possibles qui pourraient rentrer dans ces catégories. Par exemple, une douleur peut à la fois être physique et psychique. Ne plus pouvoir effectuer une action peut être le reflet d'un impact physique et/ou d'une activité quotidienne devenue impossible.

Cette phase de caractérisation fine des impacts, était nécessaire à l'étape suivante du projet qui était l'annotation des messages. L'annotation et son but sera développée dans la partie suivante, cette partie se concentrant sur les raisons sous-jacentes, décrites dans les messages, comme révélatrices d'une dimension impactée.

Le développement d'une grille d'impact de qualité de vie a été permis grâce à différents points de comparaisons avec les questionnaires précédemment identifiés.

Cette méthodologie a été inspirée par différents arbres décisionnels (figure 8) établis pour définir les objets (ou items) investigués par des questionnaire de qualité de vie, en l'occurrence les questionnaires AQoL-8D et AQoL-6D (Assesment of Quality of Life – six or eights dimensions) (35).

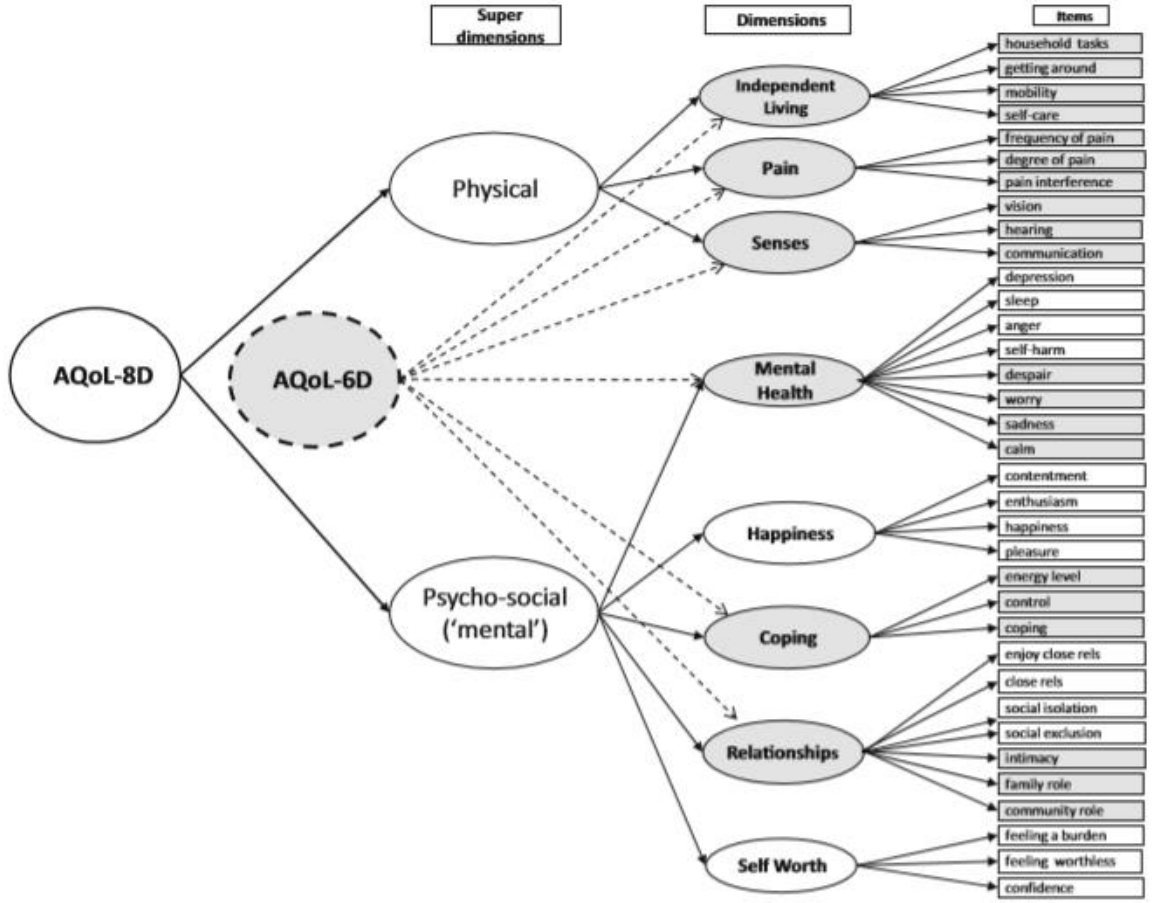


Fig. 1 AQoL-8D and AQoL-6D instruments*.
*AQoL-6D is shown as the shaded items and dimensions. It does not map into psychometrically valid 'super-dimensions' shown for AQoL-8D

Figure 8 : Arbres décisionnels et sous objets des questionnaires d'évaluation de qualité de vie AQoL-6D et AQoL-8D (35)

Afin de définir précisément chaque sous-objet (ou *items*) inclus dans notre méthodologie, nous avons procédé à une classification descendante similaire. La différence principale avec les questionnaires, est qu'ils posent au moins une question par items. Pour notre méthodologie, parce que nous nous plaçons d'un point de vue analytique des témoignages patients, il s'agit des *items* qui seront possiblement détectés dans les messages.

Chaque intitulé d'*item* est à comprendre en termes d'impact :

- « Social » → la maladie a un impact social sur le patient.
    - « Isolement » → L'impact de la maladie est sur la dimension social et génère un isolement du patient.
- « Alimentation », de la catégorie « Activité » est à comprendre dans le sens où cette activité quotidienne est à moduler par le patient en raison de sa maladie → éviction de certains aliments, mesures hygiéno-diététiques potentiellement contraignantes, etc.

De plus, la catégorie « Activité » a été subdivisée en différentes catégories afin de couvrir les activités quotidiennes/domestiques, et d'inclure également les activités professionnelles et scolaires. En effet, un état de santé peut impacter à différents âges (enfant ou adulte) et sur différentes sphères. A ce titre, il nous paraissait important d'inclure les potentiels impacts professionnels et scolaire. Les différentes sous catégories et l'ensemble des *items* associés sont présentés ci-après, et visualisés dans une cartographie sous forme d'image à la suite (figure 8).

- Dimension Physique
  - Symptôme physique
  - Effet indésirable
  - Douleur
  - Amaigrissement
  - Fatigue
  - Motricité
    - Motricité – Aide extérieur
    - Motricité – Handicap
    - Motricité – Soin personnel

- Dimension Psychique
  - Acceptation de la maladie
    - Acceptation de la maladie – Adaptation
  - Image de soi
  - Dépression
  - Détresse
    - Détresse – Solitude
  - Espoir / désespoir
    - Espoir / Désespoir- stabilisation de la maladie
  - Peur
    - Peur – Anxiété
  - Stress
  - Douleur psychique/morale
  - Trouble du sommeil
    - Trouble du sommeil – Insomnie
    - Trouble du sommeil – Hypersomnie
  - Fardeau psychique des symptômes
  - Fardeau psychique de la thérapeutique/prise en charge
  - Perte de chance – accès au traitement
  - Perte d'appétit / Anorexie mentale
  - Trouble cognitif
    - Trouble cognitif – Mémoire
    - Trouble cognitif – Attention / Concentration
    - Trouble cognitif – Langage
    - Trouble cognitif – Réalisation de tâches simples
    - Trouble cognitif – Cognition sociale

- Dimension Activité
  - Activité quotidienne
    - Impact du parcours de soin dans la vie quotidienne
    - Autonomie
      - Autonomie – soin personnel
      - Autonomie – domestique
        - Autonomie – Domestique – Cuisiner
        - Autonomie – Domestique – Faire les courses
        - Autonomie – Domestique – Tâches ménagères
        - Autonomie – Conduite voiture
    - Alimentation
    - Loisirs

        - Vacances
        - Sport
        - Activités sexuelles
        - Organisation matérielle du foyer
            - Organisation matérielle du foyer – Aménagement
            - Organisation matérielle du foyer – Hospitalisation à domicile
            - Organisation matérielle du foyer – Intrusion des soins
    - Activités professionnelles
        - Embauche
        - Arrêt de travail / Maladie
        - Adaptation du temps de travail
        - Handicap
        - Pression de l'employeur
        - Impossibilité de travailler
        - Réalisation de tâche rendue difficile
    - Activités scolaires
        - Absence
        - Mauvais résultats
        - Arrêt des études
        - Réalisation de tâche rendue difficile
        - Impossibilité de travailler

- Dimension relationnelle
    - Isolement
    - Social
    - Familiale
        - Famille – conjugal
        - Famille – Enfant
        - Famille – Enfant – Grossesse
        - Famille – Fratrie
        - Famille – Parents
        - Famille – Equilibre familial perturbé

- Dimension financière
    - Coût
    - Frais annexe
    - Reste à charge
    - Difficulté avec assurance
    - Difficulté avec prêt d'argent
    - Péril financier

Cette organisation en dimensions et catégories avait un but double : L'annotation des données (qui est développée dans la partie suivante), ainsi que d'anticiper les futurs versions de l'algorithme, qui seront capables d'identifier de plus en plus en détails les spécificités des impacts de qualité de vie, en accord avec le niveau de précision de notre méthodologie (figure 9).

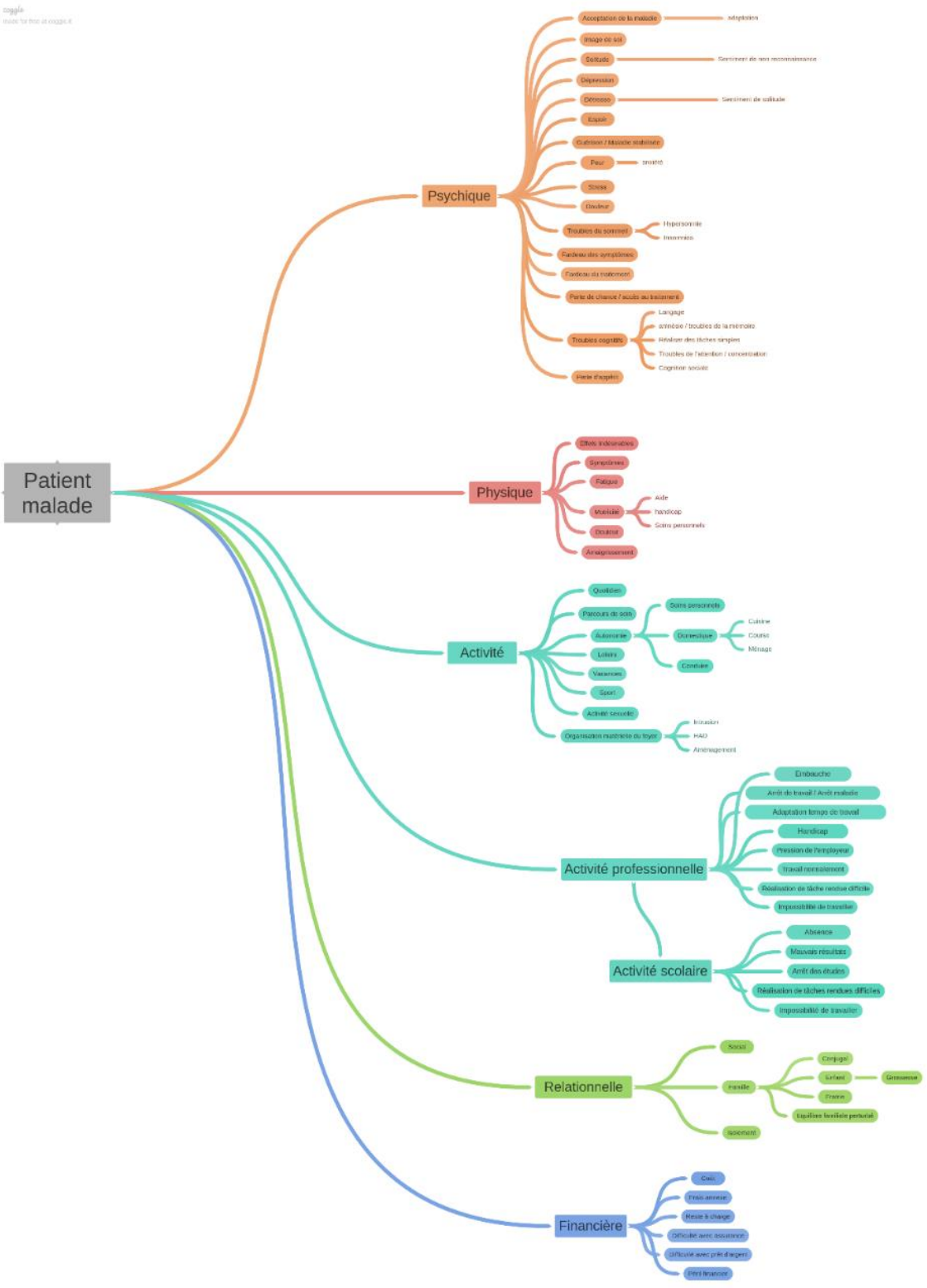


Figure 9 - Organigramme des objets inclus par dimensions de qualité de vie

4. Formation & annotation des messages - coefficient de kappa

Il existe différentes façons d'entraîner un algorithme d'Intelligence artificielle. Le but de son entraînement est d'identifier les schémas et caractères distinctifs propres à une catégorie de données.

Pour exemple, un algorithme dont le but est la reconnaissance faciale, pourra se servir de paramètres comme l'écart entre les yeux, la position des oreilles par rapport au visage, l'insertion des mandibules et la distance jusqu'au menton, etc. Pour cela, deux méthodes sont principalement utilisées : l'entraînement supervisé et l'entraînement non supervisé. Pour l'entraînement non supervisé, l'algorithme doit se débrouiller seul pour identifier les variables et schémas qui régissent un ensemble de données. Dans notre exemple de reconnaissance faciale, seules les photos de visage seront fournies à l'algorithme. Dans le cas de l'entraînement supervisé, les points morphologiques importants précités seront fournis avec les photos des visages, des valeurs seront associées aux photos. Chaque méthode a ses avantages et défauts, l'apprentissage non supervisé pourra identifier une ou plusieurs nouvelles variables non connues et ainsi gagner en connaissance sur un sujet. Par exemple, pour la reconnaissance faciale, l'algorithme pourrait identifier et utiliser la variable « longueur de cheveux » mais risquerait de mal catégoriser des femmes aux cheveux courts ou des hommes aux cheveux longs. L'avantage de l'apprentissage supervisé, est la définition à l'avance des catégories dans lesquelles nous bornons l'algorithme à fonctionner. Par définition méthodologique de catégories et de leurs variables, il devient possible d'influer sur la spécificité des résultats. Par exemple pour notre algorithme de reconnaissance faciale, l'algorithme se bornerait à utiliser les variables prédéfinies et ainsi la longueur des cheveux risquerait moins d'induire de mauvais classements.

Le revers de la médaille est qu'il faut fournir à l'algorithme des données annotées (ex : visage d'homme, visage de femme). Cela implique une méthodologie rigoureuse d'annotation des données et donc, du temps humain.

De manière générale, les performances de l'algorithme sont dépendantes de la quantité des données annotées et de la qualité de leur annotation.

D'autres catégories d'entraînement existent, tel que l'apprentissage profond qui utilise des réseaux de neurones artificiels. Là où les deux premières catégories d'apprentissage exploitent des règles mathématiques et/ou statistiques, l'apprentissage profond permet d'aller plus loin en augmentant la capacité de calcul. Cette méthode est plus sophistiquée et plus coûteuse en ressources (notamment informatiques) mais permet de véritables avancées en matière de performances. Le type d'entraînement est à définir en fonction des données et de la façon dont il est possible de les exploiter, des ressources humaines, des ressources informatiques.

Dans notre projet, c'est l'apprentissage supervisé qui a été retenu. Pour sa capacité à définir les variables à exploiter, la maîtrise de la méthodologie du projet et en fonction des ressources qui lui étaient allouées.

Dans le but d'identification des mentions d'impacts de qualité de vie, cela s'est traduit par l'annotation des expressions, ou verbatims, utilisés par les patients et leurs proches lors de la mention d'un impact de qualité de vie. De manière concrète, c'est ce qui permettra au modèle de comprendre que « j'ai le moral dans les chaussettes », correspond à un impact sur la dimension psychique, ou que « ça brûle » correspond à un impact physique.

Parce que l'étape d'annotation est critique et conditionne les futurs résultats du modèle, une formation ainsi qu'une méthode d'évaluation ont été réalisées.

Un corpus de messages a été constitué, par tirage au hasard, dans la base de données propriétaire de Kap Code, contenant 20 000 messages évoquant diverses maladies, états de santé ou traitement. Les dates d'écritures de ces messages s'étendaient de 2000 à 1019.

Les sources de données étaient 19 forums communautaires en ligne en France, généralistes ou liés à la santé (*Atoute* (36)*, Doctissimo* (37)*, AuFéminin* (38)*, Journal des femmes* (39)*, Psychoactif* (40)*, Forum.hardware* (41)*, Lesimpatientes* (42)*, Laxophobie* (43)*, Magic maman* (44)*, thyroïde* (45)*, Forum ado/public.fr* (46)*, Onmeda* (47)*, Psychologies* (48)*, MeaMedica* (49)*, Futura-sciences* (50)*, Allodocteurs* (51)*, Vulgaris Medical* (52)*, Lymphome espoir* (53)*, Maman pour la vie* (54). Ni Facebook ni Twitter n'ont été inclus dans les sources de données, car les tweets sont limités à 240 caractères, ce qui limitait la probabilité de développement narratifs de l'histoire de la maladie et des témoignages d'impact. Facebook a également été écarté pour des questions de confidentialité et de difficultés d'accès aux données.

Le spectre d'états de santé allait des cancers, au diabète, à l'endométriose, aux affections gastro-entérologiques, ainsi qu'aux affections psychiatriques (dépression) ou encore des thérapeutiques (LEVOTHYROX®, antidépresseurs). L'objectif était de constituer un panel représentatif des atteintes à la santé : physique et psychologique, reconnues et plus rares, légères ou lourdes. Ce corpus regroupait des messages aléatoirement sélectionnés mentionnant 1 280 termes médicaux (au moins un par message, maladie ou médicament). Les maladies et termes de traitements ont été identifiés respectivement par des méthodes de correspondance exacte sur le dictionnaire MedDRA® (le dictionnaire médical validé pour les activités de régulation de la santé) (55) ainsi que la base de données OpenMedic de l'Assurance Maladie française (56).

Une stratégie d'annotation a été définie par le binôme en charge du projet, le *data scientist* responsable du développement du modèle, conjointement avec l'étudiant en pharmacie en charge de la dimension médicale et du projet en lui-même.

La taille du corpus d'annotation (nombre de messages) a été définie en fonction du temps prévu pour l'annotation, alloué au projet. Le corpus d'annotation comprenait 1 400 messages, dont 1 000 messages mentionnaient au moins une maladie et 400 au moins un traitement. La taille des messages allait de quelques lignes, à plusieurs pages. Ces messages longs reprenaient l'histoire de la maladie et le vécu patient, alors que les messages courts se concentraient à des points plus précis.

Parce que deux personnes étaient chargées d'annoter les messages, il était essentiel de s'assurer de l'homogénéité de l'annotation. Une méthodologie en deux étapes a été mise en place afin de contrôler au maximum le risque d'annotation hétérogène. Une formation médicale portant sur la qualité de vie, ainsi que des exemples de messages annotés, a été dispensée par l'étudiant en pharmacie à son binôme, avant de commencer l'annotation.

Des exemples concrets de messages, préparés par l'étudiant en pharmacie, ont été présentés et discutés par le binôme. Le but de cette étape était de s'assurer d'une compréhension homogène de ce qu'est un impact de qualité de vie, avant de commencer l'annotation des messages. Il s'agissait de la fin de la formation sur la qualité de vie, afin de garantir le plus possible l'homogénéité des annotations. Les messages présentés sont disponibles en annexe, un exemple suit (figure 10).

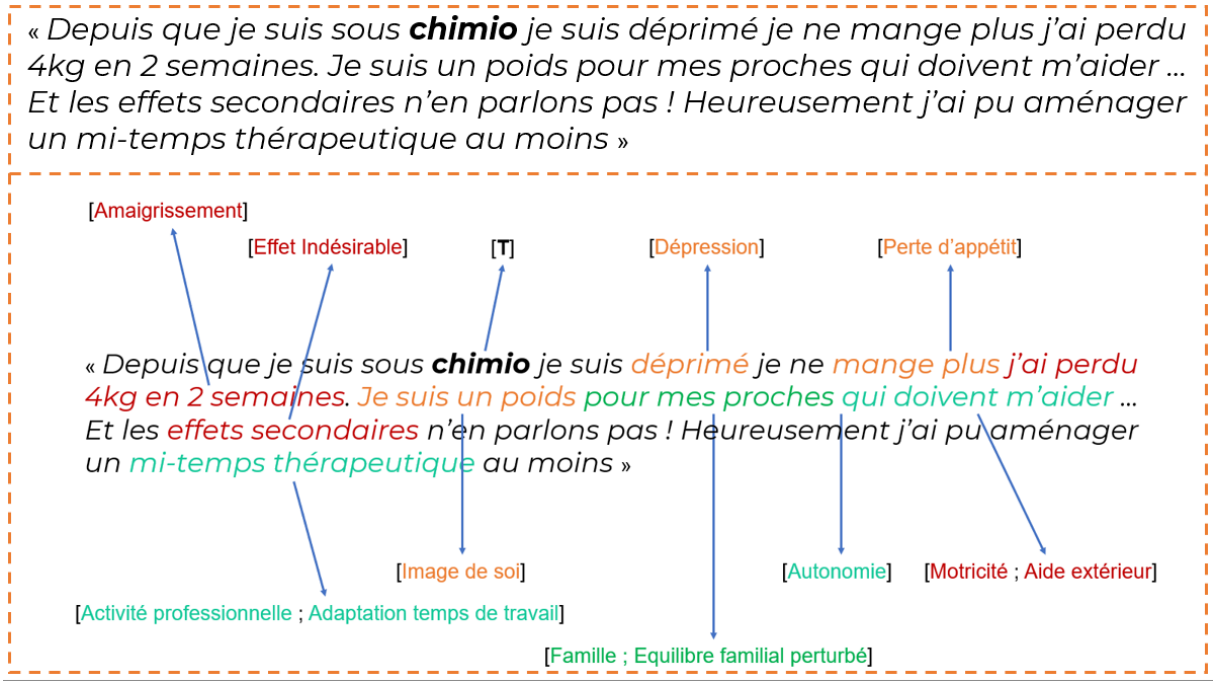


Figure 10 - Exemple d'annotation de message

Excel® a été le logiciel utilisé pour l'annotation. Bien que des outils d'annotation textuelles existent et présentent de nombreux avantages, le revers de la médaille est souvent leur prix. Excel a été choisi car c'est cet outil de la suite Microsoft™ qui était déjà utilisé en pratique chez Kap Code.

Sur le logiciel Excel®, une ligne était associée à un message issu de la base. Les différentes colonnes correspondaient aux différentes données à annoter (figure 11) :

- Présence d'un impact : Oui ou Non
- Auteur : Patient ou Proche de patient
- Présence d'au moins un impact sur la dimension Physique : Oui ou Non
- Présence d'au moins un impact sur la dimension Psychique : Oui ou Non
- Présence d'au moins un impact sur la dimension Activité : Oui ou Non
- Présence d'au moins un impact sur la dimension Relationnelle : Oui ou Non
- Présence d'au moins un impact sur la dimension Financière : Oui ou Non
- Expression utilisée exprimant une causalité entre l'état de santé et l'impact : copier-coller

Figure 11 - Colonnes d'annotation

Comme vu dans l'exemple de message annoté plus haut, il est possible que plusieurs expressions relatives au même type d'impact soient décrites dans un message. En plus d'indiquer la présence d'un impact dans une dimension de la qualité de vie, les deux annotateurs avaient comme consigne de relever les expressions utilisées par les auteurs pour décrire cet impact. Si plusieurs impacts dans la même dimension étaient présents, tous les verbatims utilisés étaient relevés.

Cette étape a permis, au fil de l'annotation, de constituer un champs lexical spécifique des témoignages d'impacts de qualité de vie, utilisés par les patients. Ces différents champs lexicaux (catégorisés par dimension) ont permis une analyse par ordinateur des schémas et des variables

linguistiques pouvant être spécifiques à chaque impact de dimension. Les variables identifiées sont présentées dans la partie suivante propre au développement de l'algorithme en lui-même.

Bien qu'il existe déjà des dictionnaires proposant des mots, expressions, champs lexicaux, propres à la qualité de vie et à différentes dimensions, le fait de constituer notre propre champ lexical, directement issu des verbatims patients, avait l'avantage de garder une certaine spécificité par rapport aux autres champs lexicaux plus généralistes.

En tout, ce sont 1 366 expressions qui ont été relevées durant l'annotation. Des exemples dans les dimensions Psychique et Activité sont disponibles ci-après (figure 12).

Dimension	Catégorie	Mot / expression
Psychique	[Dépression] #	vouloir mourir
Psychique	[Dépression] #	vouloir disparaître
Psychique	[Dépression] #	ne pas sortir du lit
Psychique	[Dépression] #	rester au lit
Psychique	[Dépression] #	envie de mourir
Psychique	[Dépression] #	envie de disparaître
Psychique	[Dépression] #	plus de plaisir
Psychique	[Dépression] #	ne rien ressentir
Psychique	[Dépression] #	se sentir triste
Psychique	[Dépression] #	plein de tristesse
Psychique	[Dépression] #	moral à zéro
Psychique	[Dépression] #	moral dans les chaussettes
Psychique	[Dépression] #	n'avoir envie de rien
Psychique	[Dépression] #	pleurer tout le temps
Psychique	[Perte d'appétit] #	ne plus manger
Psychique	[Trouble du sommeil - Insomnie] #	ne plus dormir
Psychique	[Trouble du sommeil - Hypersomnie] #	trop dormir
Psychique	[Trouble du sommeil - Insomnie] #	insomnie
Psychique	[Trouble du sommeil - Hypersomnie] #	hypersomnie
Psychique	[Détresse] #	ne plus savoir quoi faire
Activité	[Autonomie] #	prendre mon bain
Activité	[Autonomie] #	me laver
Activité	[Autonomie] #	être aidé pour
Activité	[Autonomie - Domestique] #	faire le ménage
Activité	[Autonomie - Domestique] #	m'occuper de
Activité	[Autonomie - Domestique] #	m'occuper de la maison
Activité	[Autonomie - Domestique] #	m'occuper de l'appart
Activité	[Autonomie - Domestique] #	m'occuper de l'appartement
Activité	[Autonomie - Domestique] #	entretenir
Activité	[Autonomie - Domestique] #	jardiner
Activité	[Autonomie - Domestique] #	le jardinage
Activité	[Autonomie - Domestique] #	entretenir le jardin
Activité	[Autonomie - Domestique - Cuisiner] #	cuisiner
Activité	[Autonomie - Domestique - Cuisiner] #	faire à manger
Activité	[Autonomie - Domestique - Cuisiner] #	préparer le repas
Activité	[Autonomie - Domestique - Cuisiner] #	préparer à manger
Activité	[Autonomie - Domestique - Cuisiner] #	faire le repas
Activité	[Autonomie - Domestique - Cuisiner] #	ne peux plus faire à manger
Activité	[Autonomie - Domestique - Cuisiner] #	ne peux plus faire le repas
Activité	[Autonomie - Domestique - Cuisiner] #	ne peux plus faire la cuisine

Figure 12 - Exemples d'expressions et vocabulaire utilisés dans la mention des impacts de qualité de vie.

Au-delà d'identifier des schémas linguistiques, récurrences de certains termes, verbes, et autres, utiles à l'élaboration de nos modèles d'apprentissage automatique, nous avons développé une approche lexical complémentaire. En effet, à partir des verbatims patients identifiés, des règles linguistiques ont permis d'enrichir l'approche de détection d'impact. Plusieurs verbatims propres à une catégorie étaient rangés dans cette catégorie et des synonymes étaient ajoutés afin d'enrichir l'éventail des

expressions détectables. De même, d'autres expressions étaient ajoutées. Par exemple dans la dimension activité, des listes d'activités étaient couplées, par des règles d'associations, à des termes exprimant une limitation. Par exemple « sport » était couplé à « abandonner le », « ne peux plus », « ne pas arriver », etc. De même, d'autres règles permettaient d'éliminer certaines expressions. Par exemple « pincement » peut refléter un impact physique si l'expression est « pincement d'un nerf », mais qui serait un faux positif si l'expression est « pincement au cœur », une expression plus généraliste non spécifique d'un impact de qualité de vie.

Des exemples pour les dimensions Activités, Psychique et Physique sont développés ci-après (figure 13). Pour des raisons de respect de secret industriel, les règles linguistiques ne seront pas révélées dans les exemples.

Catégorie d'impact	Sous-catégorie	nom
Physique	Douleur	Douleur
Physique	Douleur	Douloureu*
Physique	Douleur	Avoir mal
Physique	Douleur	faire mal
Physique	Douleur	Mal
Physique	Douleur	Mal
Physique	Symptome	vomir
Physique	Symptome	crampe*
Physique	Douleur	Maux
Physique	Douleur	Maux
Physique	Douleur	Pincement*
Physique	Symptome	Migraine*
Physique	Douleur	souffrir
Physique	Symptome	Brulure*
Physique	Symptome	bruler
Physique	Symptome	Compression*
Physique	Symptome	Compresser
Physique	Douleur	plier en deux
Physique	Douleur	plier en 2
Physique	Altération de l'état général	Perte de poid*
Physique	Altération de l'état général	Poid*
Physique	Altération de l'état général	Kilo
Physique	Altération de l'état général	Maigrir
Physique	Altération de l'état général	Grossir
Physique	Altération de l'état général	sac d'os

Catégorie d'impac	Colonne1	nom
Activité	Activité physique	nager
Activité	Activité physique	sport
Activité	Activité physique	muscu*
Activité	Tâches quotidiennes	shopping
Activité	Tâches quotidiennes	vaisselle
Activité	Activité physique	foot
Activité	Activité physique	rugby
Activité	Activité physique	courrir
Activité	Activité physique	courir
Activité	Alimentation	dégluti*
Activité	Alimentation	boire
Activité	Alimentation	manger
Activité	Dépendance	aider a domicile
Activité	Dépendance	aide menagere
Activité	Dépendance	dependan*
Activité	Dépendance	autonom*
Activité	Etude	classe
Activité	Etude	cours
Activité	Etude	etude
Activité	Etude	etudie*
Activité	Etude - Collège	college
Activité	Etude - Ecole	ecole
Activité	Etude - Lycée	lycee
Activité	Etude - Université	fac
Activité	Etude - Université	universite
Activité	Etude - Université	cours magistral
Activité	Hygiène personnel	douche
Activité	Hygiène personnel	bain
Activité	Hygiène personnel	laver

Catégorie d'impact	Colonne1	nom
Psychique	Anxieté et stress	anxiete OR stress OR anxieux OR angoisse OR terrifie OR terreur OR nerveux OR nervosite OR sur les nerfs OR parano OR p
Psychique	Choc diagnostique	choc OR annonce
Psychique	Combat	epreuve OR combat OR dur a vivre OR ingerable OR sens depasse OR sentir depasse
Psychique	Consultations	psy OR psychologue OR psychiatre
Psychique	Dépression	depression OR deprime OR depressif OR deprimant OR à bout OR au bout du rouleau OR suicide OR suicidaire OR pensees
Psychique	Epuisement	fatigue OR epuiser OR epuise OR epuisement OR asthenie OR insupportable OR ne supporte pas OR perdu
Psychique	Fin de vie	mourir OR fin de vie
Psychique	Image de soi	mal OR cheveux OR miroir OR regarder
Psychique	Médicaments	anxiolytique OR antidepresseur OR somnifere OR hypnotique
Psychique	Nuits et sommeil	dormir OR sommeil OR nuit OR coucher
Psychique	Perte de mortal	moral OR mental
Psychique	Reconnaissance	ignorance OR incompris OR non compris OR ne me comprend pas OR ne connaisse pas
Psychique	Resignation	ne plus pouvoir OR n'en peux plus OR perdre pied OR perdre ses reperes
Psychique	Solitude	isole OR isolement OR seul OR solitaire OR difficile à en parler
Psychique	Tristesse	triste OR tristesse OR pleurs OR pleurer OR larmes OR bouleverser OR bouleversement
Psychique	Acceptation	accepter OR acceptation OR vivre avec OR survivre
Psychique	Fardeaux des soins	fardeau OR ingerable OR difficile OR lourds OR acharnement

Figure 13 - Exemple d'approfondissement des champs lexicaux, synonymes, sans les règles linguistiques associées.

Sur les 1 400 messages composant les messages à annoter, 1 000 messages ont été attribués à l'étudiant en pharmacie et les 400 autres, à son binôme *data scientist*. Au tout début du projet, avant la formation qualité de vie et l'établissement de « *guidelines* » concernant l'annotation, 100 messages

ont été aléatoirement sélectionnés dans le jeu de données et annotés en aveugle par chacun des binômes. Puis, 100 nouveaux messages ont été annotés après la standardisation de l'annotation.

Cette étape a permis de calculer l'accord d'annotation entre les annotateurs. L'annotation des mêmes 100 messages tirés au sort a permis de comparer mathématiquement leurs annotations. Cela s'est traduit par le calcul du coefficient Kappa de Cohen, permettant une évaluation de la méthodologie, pouvant expliquer les résultats futurs de l'algorithme.

De manière concrète, le risque est une interprétation personnelle des impacts présents dans les messages. La relation entre le patient et son professionnel de santé doit être empreinte de compréhension, d'empathie et parfois de la faculté à savoir lire entre les lignes, ce qui est propre à l'être humain. Or, nos données vont servir à entraîner un algorithme, un robot qui par définition est dépourvu de ces capacités. Il était donc nécessaire d'annoter de façon franche, ce qui est un impact objectif, sans interprétation de ce que le patient peut ressentir, c'est-à-dire en se bornant à son strict déclaratif car c'est ce qui sera analysé par l'algorithme.

Standardiser l'annotation était le but de la formation sur la qualité de vie, ainsi que d'échanger sur les exemples de messages précités. Le calcul du coefficient Kappa de Cohen était le moyen de mesurer le niveau de standardisation, à travers les messages doublement annotés en aveugle.

Avant la standardisation des pratiques d'annotation, l'accord inter-annotateurs était faible sur la présence d'un impact, ainsi que sur les dimensions impactées. En effet, pour chaque catégorie le score κ_1 était compris entre 0 et 0,4. Après la formation et l'établissement des guidelines, l'évaluation des 100 nouveaux messages doublement annotés a révélé une harmonisation des annotations à travers le κ_2 (Accord fort au minimum ou accord presque parfait), à l'exception de la catégorie financière, car aucun des 100 messages aléatoirement sélectionnés ne mentionnaient cette dimension. Les seuils d'interprétation Kappa de Cohen sont détaillés dans le tableau 3, et l'évolution des scores κ inter-annotateurs est détaillée dans le tableau 4.

Tableau 3 - Grille d'interprétation des scores Kappa de Cohen

κ	**Interprétation**
< 0	Désaccord
0,0 – 0,20	Accord très faible
0,21 – 0,40	Accord faible
0,41 – 0,60	Accord modéré
0,61 – 0,80	Accord fort
0,81 – 1,00	Accord presque parfait

Tableau 4 - Score Kappa de Cohen inter-annotateurs, avec (κ1) et après (κ2) formation médicale et standardisation des pratiques d'annotation des impacts de qualité de vie.

Dimension	**κ_1 (sur 100 messages)**	**κ_2 (sur 100 messages)**
Impact	0,403	0,723
Physique	0,237	0,871
Psychique	0,278	0,663
Activité	0,313	0,639
Relationnelle	0,217	0,649
Financière	0,0	0,0

Au final, sur les 1 400 messages analysés et annotés, 818 messages témoignaient d'au moins un impact de qualité de vie, 442 montrant un impact physique, 519 psychique, 363 liés à l'activité, 193 au relationnel et 69 à la dimension financière. Plusieurs impacts par dimension étaient souvent mentionnés par les patients. Les résultats sont présentés dans le tableau 5.

Tableau 5 - Nombre de messages exprimant un impact sur la qualité de vie, au moins un impact et par dimension.

Catégorie	**Au moins un impact**	**Physique**	**Psychique**	**Activité**	**Relationnel**	**Financier**
Nombre de messages	818	442	519	363	193	69

La sélection, quantité, ainsi que la qualité des données, sont des paramètres à maîtriser pour tout projet d'apprentissage automatique supervisé. Parce que l'on attend d'une machine d'identifier les schémas régissant ces données, afin d'apprendre à les reconnaître sur d'autres données, la mesure et le contrôle des différentes étapes sont primordiaux. C'est ce que nous a permis d'évaluer le coefficient Kappa de Cohen. Si les deux annotateurs avaient annoté sans concertation, mise au point, formation, recommandations, le manque d'homogénéité aurait induit ce que l'on peut appeler du « bruit » dans les données, impactant négativement les futures performances de l'algorithme.

Au-delà de l'importance de l'annotation des données, cette étape illustre la synergie positive qui a existé entre l'étudiant en pharmacie et son binôme *data scientist*. Le *data scientist* étant un expert des phases de traitement des données et de leur labellisation. L'étudiant en pharmacie, par sa formation et ses connaissances médicales, ainsi que par l'étude bibliographique sur la qualité de vie liée à la santé effectuée, a pu dispenser une formation et établir des recommandations sur la conduite des annotations, ainsi qu'une méthode de maîtrise statistique du procédé.

La complémentarité de l'approche interprofessionnelle est un argument supplémentaire en faveur de l'organisation en mode projet. En effet, dans les projets mêlant intelligence artificielle en santé, il a été remarqué que les projets avaient davantage de chance de réussite et d'impact positif, si le chef de projet était une personne issue du milieu de la santé, travaillant avec des experts des domaines de l'intelligence artificielle.

Sur l'ensemble de ce processus, le binôme a pu profiter des conseils d'un médecin de santé publique spécialiste des études de qualité de vie, qui a validé l'ensemble des étapes.

5. Machine learning - sélection des variables et différents modèles

Ingénierie des paramètres algorithmiques

Les 1 400 messages annotés ont été répartis selon des proportions 70%/30% dans deux jeux de données distincts, respectivement le jeu de données « apprentissage » et le jeu de données « test ». Il s'agit d'une répartition classique dans l'apprentissage automatique, dont le nom est validation croisée. L'objectif de la validation croisée est de donner une partie des données à l'algorithme sur laquelle il peut apprendre (70%). Pour ensuite être évalué sur la partie restante (30%) qu'il n'a jamais vu, sur laquelle ses prédictions seront comparées aux annotations humaines.

Tous les messages témoignant d'impact ont été utilisés pour générer des caractéristiques spécifiques à chaque dimension, appelées « variables ». D'autres caractéristiques, ou variables, étaient basées sur la structure du message, comme le sentiment exprimé (positif, négatif, colère, dégoût, peur, joie, tristesse, surprise), la grammaire (nombre de pronoms, auteur, phrases négatives...) et la conjugaison (temps des verbes).

Un score de champ lexical correspondant à chaque dimension de la qualité de vie a été calculé, en comptant les expressions associées préalablement collectées lors de l'étape d'annotation. Les variables lexicales ont été générées en utilisant un outil interne propre à l'outil Detec't© (57). Cette phase a permis de développer des modèles spécifiques de détection d'impact, pour chacune des cinq dimensions. Le rationnel de ce processus était de pouvoir s'adapter aux nombreuses expressions des patients. En effet, les impacts psychiques et les impacts physiques sont différents et les expressions utilisées pour les décrire le sont également, même si des points de similitudes peuvent exister. Ainsi, disposer d'un modèle spécifique par dimension est un moyen de minimiser un biais d'interprétation.
Nous avons abouti à un ensemble de données, ou corpus, composé de caractéristiques quantitatives telles que les sentiments, la grammaire, la conjugaison et les champs lexicaux des caractéristiques liées à la qualité de vie.

Sélection du modèle

Nous avons ensuite utilisé des technologies d'exploration de données et d'apprentissage automatique pour catégoriser et analyser les données extraites de notre corpus final. Comme nos caractéristiques ne présentent pas de valeurs négatives, nous avons normalisé nos données en divisant toutes les valeurs des caractéristiques par leur maximum respectif de sorte que toutes les valeurs se situent entre 0 et 1, minimisant ainsi les variances interclasse et intraclasse. Toutes les valeurs manquantes ont été remplacées par la médiane afin de ne pas influencer la variance intraclasse, ce qui est une pratique courante et reconnue en *data science* et approche statistique pour gérer les données manquantes.
Nous avons obtenu un premier algorithme de classification dont le but était de déterminer la présence d'un impact de qualité de vie. Ensuite, un algorithme de classification propre à chaque dimension a été développé afin d'évaluer si l'impact détecté par le premier algorithme, concernait la dimension correspondante (figure 14).

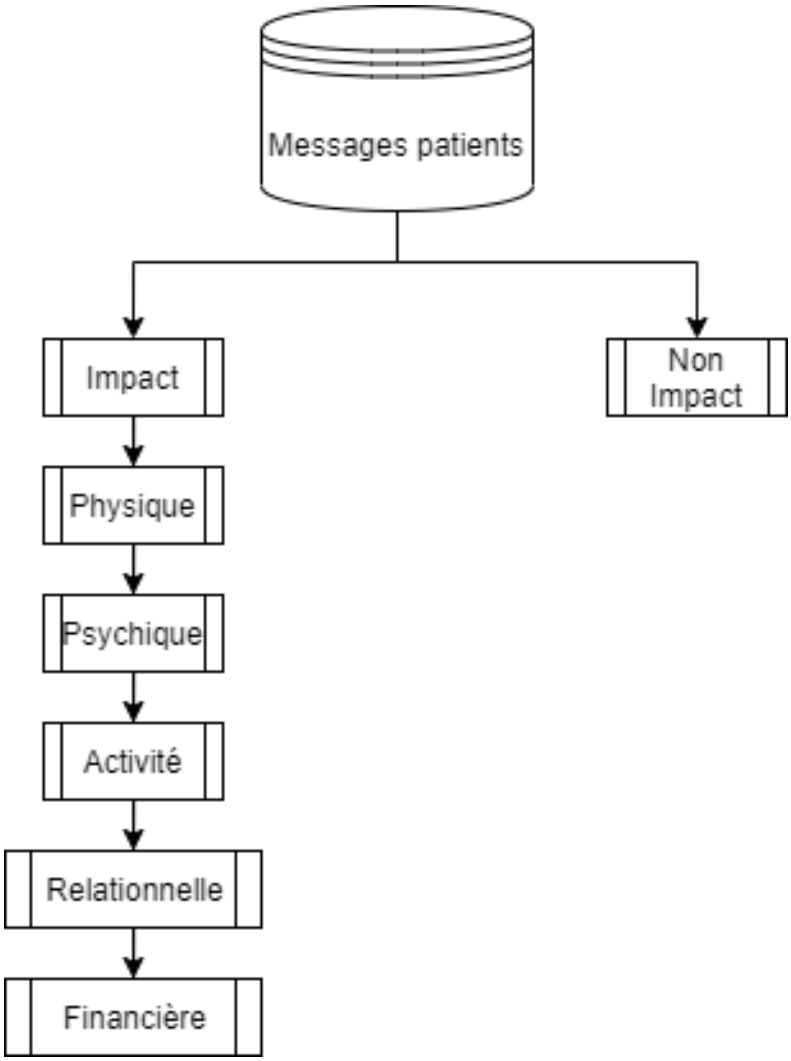


Figure 14 - Processus de détection des impacts selon les modèles algorithmiques choisis.

Différentes approches de traitement automatisé du langage ont été identifiées dans la phase de bibliographie. L'objectif de cette thèse n'est pas de développer chacune des méthodologies identifiées, elles seront uniquement nommées. Un résumé des informations à retenir après la lecture des articles est disponible ci-après (figure 15). La partie *machine learning* (apprentissage machine), encadrée dans l'image ci-après est celle pour laquelle différentes méthodes pertinentes ont été identifiées comme pertinents. La lecture des publications médicales a notamment mis en évidence que la combinaison de différents classificateurs était meilleure que leur utilisation seule.

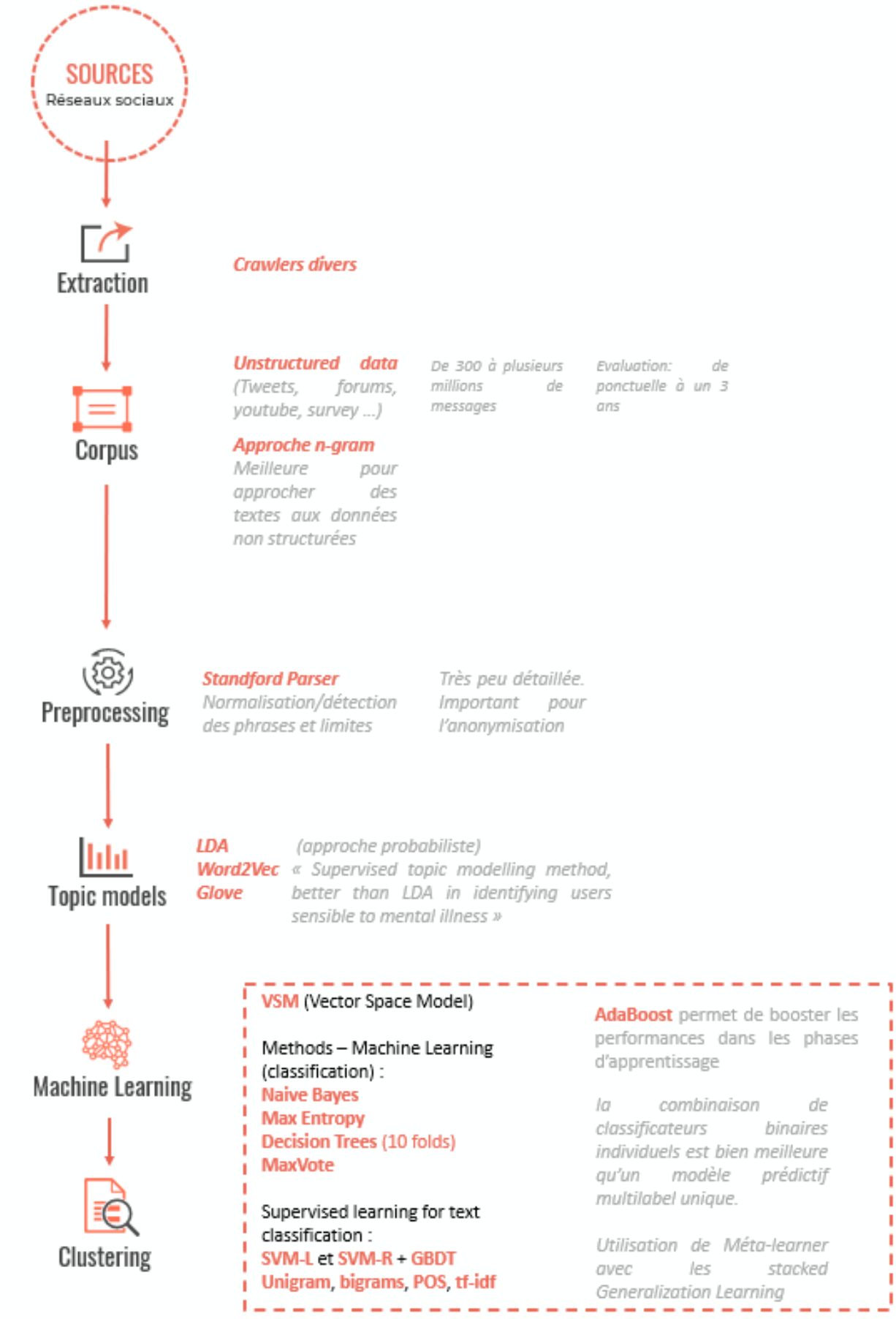


Figure 15 - Différentes méthodes algorithmiques et synthèses des informations importantes à retenir, identifiés durant l'étape de bibliographie.

Tout d'abord, nous avons utilisé un *sélecteur flottant séquentiel avant 5 plis* avec un algorithme *eXtreme Gradient Boosting* pour sélectionner la meilleure combinaison parmi tous les paramètres et variables à notre disposition. Nous avons d'abord essayé de maximiser la précision du modèle, mais nous nous sommes retrouvés avec un grand nombre de faux négatifs. Nous avons finalement choisi l'Aire Sous la Courbe (AUC) comme méthode de notation pour maximiser le taux de vrais positifs, car nous préférons avoir un nombre légèrement plus élevé de messages contenant un impact, même avec des faux positifs, plutôt que de manquer certains d'entre eux.

Nous avons choisi le *Sélecteur flottant séquentiel avant* par rapport au *Lasso* pour maximiser la valeur du ROC (*Receiver Operating Characteristics*), alors que le *Lasso* essaie de minimiser la fonction de coût. Cela a permis d'obtenir les meilleures performances pour toutes les classes au lieu de la classe majoritaire.
Nous avons ensuite essayé plusieurs algorithmes d'apprentissage automatique, les *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM), *Multi-Layer Perceptron* (MLP), *Random Forest* (RF) et enfin *XGBoost* (XGB).

À l'exception de la dimension psychique, XGB était bien supérieur aux autres méthodes en termes d'AUC (tableau 6). Plus l'AUC est proche de 100 plus la performance est bonne.

Tableau 6 - AUC des différentes méthodes d'apprentissage automatique

Algorithme	Impact*	Physique	Psychique	Activité	Relationnel	Financier
KNN	69.9	66.5	64.9	64.4	68.6	65.6
SVM	67.9	63.3	56.3	55.3	57.7	56.9
MLP	74.6	67.9	61.6	64.5	64.5	58.6
RF	75	70.5	70.7	71.8	70.9	69
XGB	78.5	75	71	76	71.7	76.5

**au moins un impact sur au moins une des cinq dimensions.*

Nous avons ensuite effectué un *GridSearch à validation croisée 5 plis* sur les caractéristiques sélectionnées afin d'ajuster nos hyperparamètres. Nous avons divisé notre ensemble d'entraînement en 5 échantillons et entraîné l'algorithme successivement sur quatre de ces échantillons tandis que le dernier était utilisé comme ensemble de validation. Cette méthode a permis de minimiser le sur-apprentissage de l'algorithme et de s'assurer que les modèles se généralisent bien d'un jeu de données à un autre. Nous avons fait varier le taux d'apprentissage, le nombre d'arbres de décision et leur profondeur maximale, le poids minimum nécessaire dans un nœud enfant, la réduction de perte minimum nécessaire pour effectuer une partition supplémentaire sur un nœud feuille et la régularisation L1. La régression Lasso était préférable pour la sélection des variables dans le cas d'un grand nombre de variables, rendant les variables non importantes encore plus insignifiantes, en termes de poids.

Ce processus a permis d'élaborer un modèle capable de détecter un impact général. L'algorithme développé a filtré le corpus de messages en deux catégories : Présence d'un impact de qualité de vie ou non. Pour chaque modèle, nous avons sélectionné les variables pertinentes en appliquant le sélecteur flottant séquentiel avant et choisi la combinaison qui permettrait de mieux discriminer un message impacté d'un message non impacté. Cette étape permet de supprimer ou ajouter une caractéristique sur le classificateur et tester les performances jusqu'à atteindre le meilleur score possible. Les mêmes étapes ont ensuite été reproduites dans chaque dimension en fonction de leurs caractéristiques spécifiques afin d'obtenir des algorithmes spécifiques adaptés à chaque dimension.

Le dictionnaire (LIWC) (58) propose des expressions pour divers sentiments, comme la positivité, la négativité, la joie, la tristesse, le dégoût, la surprise, la peur et la colère. Il s'agit initialement d'un programme d'analyse qui compte les mots d'un texte en les attribuant à des catégories qui ont un sens psychologique. Il offre une analyse (en proportion) de 80 dimensions du langage (mots fonctionnels, thèmes, ponctuations). Ce dictionnaire nous a permis de compter les émotions contenues dans les messages. Ces variables ont été sélectionnées en fonction des modèles identifiés lors de la phase d'annotation. En effet, nous pouvons supposer que pour décrire les actions et les difficultés quotidiennes, le présent est le plus approprié. A l'inverse, pour parler d'un impact au sein de la famille, le "nous" est plus souvent utilisé.
Un exemple non exhaustif de certains variables identifiées comme pertinente, par les algorithmes précédemment cités, est présenté ci-après (tableau 7).

Tableau 7 - Variables les plus importantes par modèle

Dimension	Variable 1	Variable 2	Variable3
Impact	Nombre de verbes à l'infinitif	Première personnel du singulier	Expressions de tristesse comptées
Physique	Nombre d'expressions physiques	Nombre d'expressions négatives	Nombre de négations
Psychique	Nombre d'expressions liées à la dimension psychique	Expression de la colère	Expression de la peur
Activité	Nombre d'expressions professionnelles, académiques ou quotidiennes liées à l'activité	Compte des verbes au présent	Nombre de pronoms
Relationnelle	Nombre d'expressions relationnelles	Première personne du pluriel	Verbes au participe passé
Financière	Expressions financières	Nombre de "pas"	Compte des verbes au passé

En raison d'une variation des proportions de témoignages d'impact de qualité de vie entre les dimensions, certaines classes étaient déséquilibrées les unes par rapport aux autres ; afin de corriger cela, nous avons créé une classe artificiellement équilibrée en utilisant la méthode de suréchantillonnage synthétique des minorités *Synthethic Minority Oversampling Technique* (SMOTE) (59). Basée sur la structure mathématique des messages sous-représentés, cette méthode crée artificiellement des itérations similaires qui correspondent au même schéma de caractéristiques, afin d'équilibrer les catégories. Nous avons utilisé cette méthode pour les algorithmes d'impact lié aux dimensions d'activité, relationnelle et financière. Cette méthode est une des approches validées pour gérer les données manquantes et éviter que des catégories sous-représentées n'influencent l'apprentissage machine. Il existe d'autres méthodes de gestion des données manquantes.

6. Résultats et performances de l'algorithme

Nous avons utilisé la sensibilité (définie comme l'identification correcte d'un impact sur la qualité de vie lorsque notre algorithme le classe ainsi) et la spécificité (définie comme l'identification correcte d'un message sans impact lorsque notre algorithme le classe ainsi). La courbe ROC (*receiver operating characteristics*) et l'AUC (l'aire sous la courbe) ont été considérées pour mesurer la performance globale de l'algorithme. La courbe ROC représente le taux de vrais positifs (sensibilité) en fonction du taux de faux positifs (spécificité). La mesure F est une métrique utilisée dans l'évaluation des modèles d'apprentissage automatique, combinaison de différents indicateurs (60) : la précision et le rappel. La précision mesure la capacité de l'algorithme à écarter les vrais négatifs. Le rappel mesure la capacité de l'algorithme à identifier les vrais positifs. La mesure F est la moyenne harmonique de la précision et du rappel, il mesure la capacité de l'algorithme à identifier les vrais positifs et écarter les vrais négatifs.

Cela nous a permis de détecter un impact général sur la qualité de vie avec une sensibilité de 0,8 et une spécificité de 0,7 (tableau 8). Au total, 818 messages présentaient un impact et 581 non. Pour l'impact physique, la sensibilité était de 0,56 et la spécificité de 0,857, pour l'impact psychique de 0,58 et 0,82, pour l'impact lié à l'activité de 0,71 et 0,79, pour l'impact relationnel de 0,675 et 0,73, et pour l'impact financier de 0,77 et 0,814, respectivement.

Tableau 8 - Résultat globaux de l'algorithme dans la détection d'un impact, puis l'attribution par dimension

Dimension	Mesure F	Courbe ROC
Au moins un impact	0.78	0.78
Physique	0.70	0.75
Psychique	0.68	0.71
Lien avec l'activité	0.75	0.76
Relationnel	0.70	0.72
Financier	0.76	0.76

L'algorithme a été capable de détecter différents types de maladies et de traitements ayant un impact sur la qualité de vie avec une bonne sensibilité et spécificité. L'algorithme a obtenu un score ROC de

0,78 pour la détection d'au moins un impact sur au moins une des cinq dimensions : 0,75 pour la dimension physique, 0,71 pour la dimension psychique, 0,76 pour la dimension liée à l'activité, 0,72 pour la dimension relationnelle et 0,76 pour la dimension financière. Par rapport à d'autres études (61,62), ces indicateurs étaient élevés et robustes ; en comparaison notamment avec les travaux de Caster et al. où l'AUC (Aire Sous la Courbe) variait entre 0,43 et 0,67. Cependant, les objectifs et les approches de ces études étaient différents des nôtres et il est assez difficile de comparer les résultats. Les performances peuvent varier en fonction de la source de données Web, étant donné que nous avons pu accéder à un grand ensemble de données et utiliser un sous-ensemble d'entraînement satisfaisant, cela pourrait expliquer nos meilleures performances. Néanmoins, Facebook et Twitter ont été écartés de nos sources en raison des messages courts de Twitter et des difficultés d'accès aux données de Facebook.

PARTIE 3 : UTILISATION DE L'ALGORITHME, EXEMPLE D'ANALYSE D'IMPACT DE QUALITE DE VIE

Suite à sa réalisation, l'algorithme a pu être utilisé dans le cadre de différentes études portant sur différentes aires thérapeutiques, maladies, et leurs impacts associés sur la qualité de vie liée à la santé. Pour chaque étude, une méthodologie mixte d'analyse a été réalisée. Dans un premier temps, l'algorithme identifiait et quantifiait par dimension, les impacts de qualité de vie dans les messages. Ensuite, l'analyse manuelle des messages identifiés permettait une caractérisation fine des informations, des raisons associées aux impacts.

Cas pratique : Etude de l'impact de cinq dermatoses visibles sur la qualité de vie liée à la santé

Dans le cadre d'une étude portant sur le vécu des patients et de leurs proches, atteints de cinq dermatoses (eczéma, rosacée, vitiligo, psoriasis, acné), une extraction, sur une période temporelle de 3 ans (septembre 2018 à septembre 2021) a été réalisée. Cette extraction a permis de récolter 20 282 messages (tableau 9), tous issus de patients ou de leur proche, sur lesquels le modèle a pu être utilisé.

Tableau 9 - Distribution des effectifs en fonction des cinq dermatoses

Corpus	Messages (n)
Eczéma	4 522
Rosacée	252
Vitiligo	771
Acné	12 799
Psoriasis	1 938
Total	20 282

Les résultats de l'algorithme sont présentés dans le tableau 10. En raison des effectifs différents entre les tailles des jeux de données (nombre de message par corpus), seules les proportions relatives à chaque dermatose ont été analysées dans un but de comparaison (figure 16).

Tableau 10 - Nombre et proportions de messages présentant un impact de qualité de vie, par maladie et par dimension

Category	**Eczéma** (n, %)	**Rosacea** (n, %)	**Vitiligo** (n, %)	**Acne** (n, %)	**Psoriasis** (n, %)
Impact post	1 944 43%	68 27%	107 14%	3 455 27%	639 33%
Physical	1 518 58,63%	24 38,71%	77 50,99%	2 859 63,55%	341 41,43%
Psychological	688 26,57%	14 22,58%	51 33,77%	1 471 32,70%	309 37,55%
Activity	332 12,82%	11 17,74%	15 9,93%	0 0%	141 17,13%
Social	1 0,04%	8 12,90%	8 5,30%	1 0,02%	2 0,24%
Financial	50 1,93%	5 8,06%	0 0%	168 3,73%	30 3,65%

Ce type d'analyse est réalisé sur un partage libre et naïf de l'expérience d'une maladie. Par conséquent, seul ce qui est verbalisé est détectable. Cette analyse est limitée, par conception, aux verbatims des

patients et de leurs proches. L'absence de mention d'un impact sur un message de réseau social n'est pas une preuve d'absence d'impact dans la vie réelle.

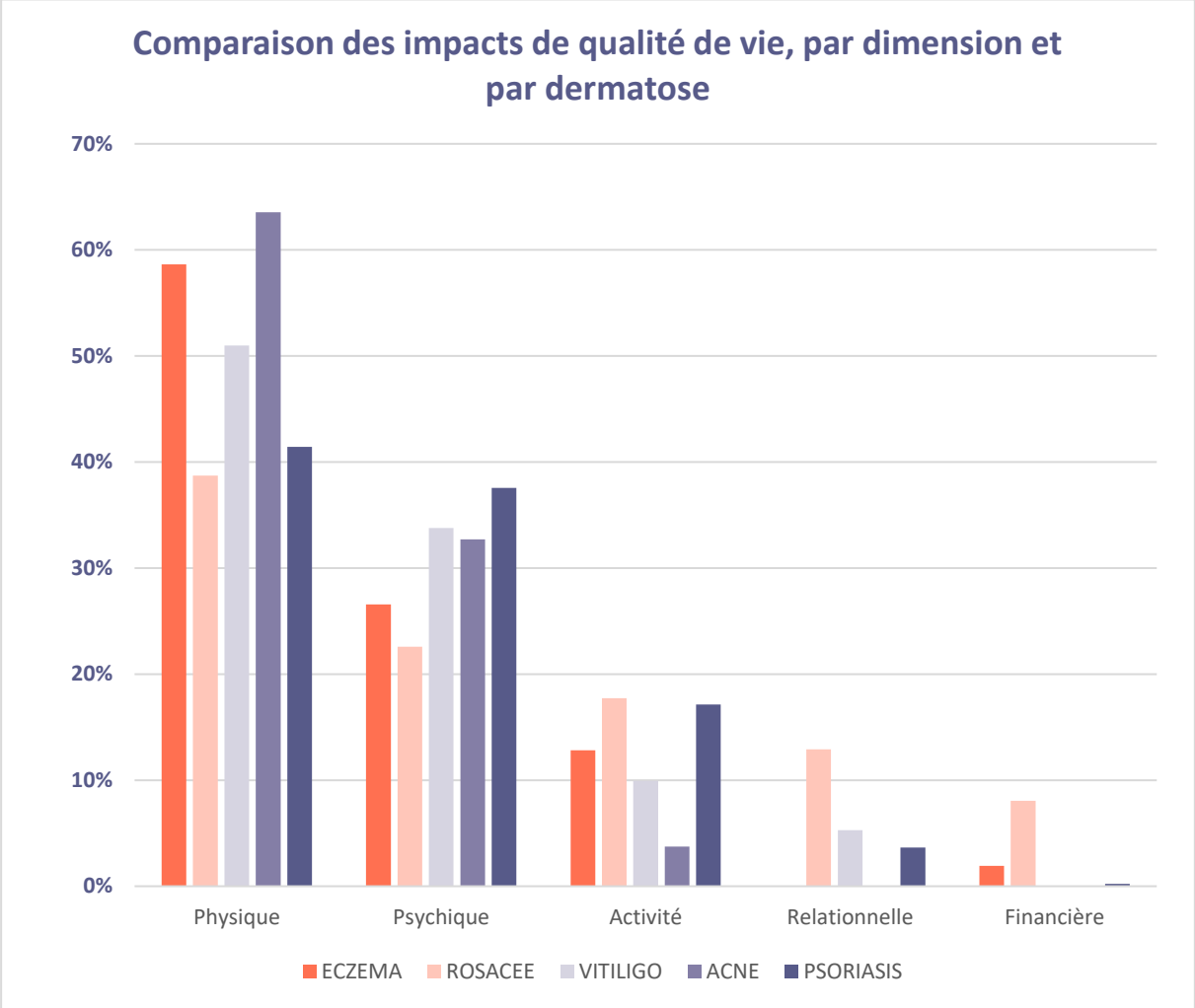


Figure 16 - Comparaison des impacts de qualité de vie, par dimension et par dermatose

L'analyse de la figure 16 permet d'illustrer que la qualité de vie liée à la santé dans les maladies cutanées visibles est principalement impactée sur la dimension physique, puis psychologique. La lecture manuelle des messages a permis de caractériser ce que sont les impacts décrits.

Sur le plan physique, les impacts correspondent à des symptômes dermatologiques, tels que " boutons ", " pustules ", " plaques érythémateuses " et des sensations telles que des prurits, la douleur, sensation de brûlure. Les impacts psychologiques sont associés principalement au caractère récurrent, aux poussées, à la maladie évoluant en crise et aux rechutes associées. Une fatigue psychologique est identifiée, associée dans les verbatims à l'errance thérapeutique, à l'inefficacité des traitements, au manque de sommeil provoqué par le prurit nocturne. Les patients décrivent comme lourd, le poids du regard des autres, affectant négativement l'image de soi des personnes atteintes de maladies de la peau. Les activités impactées concernent principalement les patients qui décrivent être obligés de faire face aux variations de temps et de température extérieure ; ils disent éviter de sortir de chez eux pour éviter une poussée ou une aggravation de la maladie (le type d'activité n'est pas toujours mentionné). De manière surprenante, l'impact social est peu abordé par rapport aux autres proportions de la dimension impactée ; cela pourrait s'expliquer par un biais comportemental sur les réseaux sociaux, les patients et les proches recherchant principalement des conseils et des solutions aux problèmes

qu'ils décrivent. Peu d'impacts financiers ont été détectés, l'analyse des messages a montré que l'achat de produits dermatologiques semble normal, dans la quête de produits dermatologiques efficaces. Les dépenses évoquées concernent surtout l'achat récurrent de produits dermatologiques jugés chers, le non-remboursement de certaines crèmes ou préparations pharmaceutiques, ou encore les séances de laser.

L'analyse par maladie dermatologique fait apparaître des spécificités : L'impact des activités sur l'eczéma est lié à la nutrition, comme la nécessité d'éviter certains aliments pouvant contenir des allergènes, en particulier pour l'eczéma du nouveau-né. Le sommeil est également une activité quotidienne impactée par l'eczéma et son prurit, chez l'adulte comme chez l'enfant. L'eczéma est aussi une source de pleurs dans la population pédiatrique, ayant un impact moral sur le patient et ses proches, comme en témoignent les mères dans leurs messages.

Pour la rosacée, une anxiété sociale associée à une peur de rougir est décrite dans les verbatims des patients, source d'altération sociale. Les patients rapportent qu'ils doivent équilibrer leurs activités extérieures avec le soleil et la chaleur. Ils restent à la maison pour éviter l'aggravation ou les poussées. L'impact financier est plus élevé dans la catégorie rosacée, en raison du prix élevé des séances de traitement au laser, où un coût non remboursable de 200€ et plusieurs séances, sont décrits.

Dans le cas du vitiligo, la dépigmentation est le principal symptôme physique à être décrit. C'est une maladie mal reconnue et les patients témoignent ressentir une peur du regard des autres, ce qui les inhibe dans la réalisation d'activités de plein air, comme aller à la plage en vacances. Le vitiligo atteignant les zones intimes, génitales, a un impact sur la relation du patient qui témoigne d'une peur de développer des relations intimes, pendant plusieurs années.

Le profil d'impact de l'acné est axé sur les dimensions physiques et psychologiques. Les boutons et les comédons sont décrits comme étant difficiles à cacher, ce qui a un impact sur la confiance en soi. Pour les patients atteints d'acné adulte, la persistance d'une maladie juvénile à l'âge adulte a un impact moral important. L'acné est aussi une maladie qui peut laisser des cicatrices sur le visage, source d'un impact physique persistant ou éternel. L'acné est également source de douleurs dans les verbatims et associée à des kystes et des boutons sur le visage.

Pour les personnes atteintes de psoriasis, les proportions des impacts physiques et psychologiques sont très proches. Une explication a été identifiée par l'analyse manuelle des verbatims, les patients décrivent les symptômes et la charge psychologique associée ; en effet, ils sont à la fois sources et conséquences de stress et d'anxiété, un cercle vicieux de symptômes dermatologiques et psychologiques auto-entretenus, où le stress provoque une poussée, la progression du psoriasis étant en retour une source d'anxiété. L'errance thérapeutique et l'évolution des poussées sont décrites comme étant les plus impactantes sur le plan psychologique, car elles obligent à être en permanence sur des montagnes russes d'espoir et de désespoir ; Les changements saisonniers, les variations de température et le temps sec sont à l'origine d'activités impactées, les patients évitant de sortir pour prévenir les poussées de psoriasis. Le sommeil est également altéré par les plaques prurigineuses et la desquamation.

Des messages d'intérêt, révélateurs des impacts de qualité de vie, sont disponibles en annexe.

Parce que la reconnaissance du fardeau psychologique des patients atteints de dermatoses visibles est un sujet de recherche, un travail de bibliographie a été réalisé afin de contextualiser ces résultats. Une emphase particulière a été d'identifier des études qui ont utilisé les questionnaires SF-36 et/ou EQ-5D,

dans le but d'explorer les impacts de qualité de vie des patients atteints de nos cinq dermatoses d'intérêt.

Bibliographie : Qualité de vie impactée par les dermatoses d'intérêt

Il existe de nombreuses façons d'explorer l'impact des maladies de la peau sur la qualité de vie liée à la santé (QVLS). Les mesures génériques de santé Euro-QoL 5D et SF-36 ont été utilisées pour évaluer l'impact de l'acné (63,64) et du psoriasis (65,66) sur la QVLS. Les résultats pour l'acné étaient que l'impact de la QVLS n'était pas fonction de la sévérité de l'acné cliniquement évaluée, les patients rapportaient des niveaux de problèmes sociaux, psychologiques et émotionnels aussi importants que ceux rapportés par les patients souffrant d'asthme chronique invalidant, d'épilepsie, de diabète, de douleurs dorsales ou d'arthrite (63). Pour le psoriasis, les patients ont signalé une réduction du fonctionnement physique et du fonctionnement mental comparable à ceux observé dans d'autres maladies chroniques telles que le cancer, l'arthrite, l'hypertension, les maladies cardiaques, le diabète et la dépression (65). La relation entre le psoriasis et le stress psychosocial est complexe. Il peut être la cause principale du développement initial du psoriasis, ou même être à l'origine de poussées et de rechutes, entraînant la progression de la maladie, ce qui aboutit à un cercle vicieux où le stress et le psoriasis s'auto-perpétuent (67,68). On a constaté que les patients dont le stress lié au psoriasis est plus important présentent un psoriasis davantage défigurant sur le plan esthétique. Les patients souffrent d'un sentiment de stigmatisation, de rejet social, de honte, d'embarras et d'un manque de confiance en soi, ce qui a un impact significatif sur la dimension sociale de leur qualité de vie et une plus grande détresse émotionnelle. La dépression et les pensées suicidaires ne sont pas rares et la défiguration esthétique ainsi que l'impact du psoriasis sur la qualité de vie des patients sont des cofacteurs importants de la dépression et du suicide (69,70). D'autres aspects du fonctionnement social ont été identifiés comme étant altérés chez les patients atteints de psoriasis, notamment en ce qui concerne les activités professionnelles (école ou lieu de travail), ce qui se traduit par l'absence totale de travail ou la perte de jours de travail (71). Les activités de loisirs sont également affectées, les patients déclarent éviter de se baigner, de prendre des bains de soleil et de participer à des sports collectifs (72). Le psoriasis affecte également le choix des vêtements, en évitant les vêtements d'été à manches courtes ou les shorts et en préférant les vêtements sombres pour minimiser l'exposition de la peau au regard des autres (72). En outre, le psoriasis est identifié comme étant la cause d'un impact dévastateur sur la vie sexuelle des patients, entraînant l'inhibition de l'activité sexuelle (72).

La population touchée par l'eczéma présenterait également une QVLS inférieure et une détresse psychologique accrue ; une autre évaluation du SF-36 a montré que la QVLS des patients était affectée dans les dimensions de vitalité, de fonctionnement social et mental (73). Une autre étude souligne que l'image de soi impactée, provenant du sentiment d'embarras par leur apparence, conduit même à la colère (74). Parmi la population pédiatrique atteinte d'eczéma, des problèmes de comportement sont identifiés, tels qu'une dépendance accrue, de la peur et des difficultés de sommeil (75).

Le vitiligo affecte environ 1% de la population mondiale et est identifié comme étant un problème plus important pour la population pigmentée et comme étant la cause d'un plus grand handicap (76,77). Environ 75% des personnes souffrant de vitiligo trouvent leur apparence modérément à sévèrement intolérable (78) ; les symptômes du vitiligo, qui entraînent une défiguration visible, sont une cause importante d'impact sur la qualité de vie, en particulier sur les dimensions psychologiques et sociales. Les patients peuvent être affectés de nombreuses manières émotionnelles, comme une faible estime de soi, la peur, l'anxiété, le stress et un sentiment de honte dans les interactions sociales (79–81). Les activités quotidiennes peuvent également être affectées, en ce qui concerne la vie sociale, les loisirs

et l'habillement, où les patients peuvent ressentir une pression pour s'habiller d'une certaine manière afin d'échapper au regard des autres, ou même les activités sexuelles où les patients rapportent un sentiment de gêne et la peur de dévoiler leur corps (82,83). La qualité de vie des patients était plus altérée que celle des hommes et était égale à celle du psoriasis (82).

La rosacée est une affection cutanée courante et chronique, dont la prévalence varie de 0,1 % à 22 % (84–87). Cependant, l'errance du diagnostic est une des difficultés identifiées concernant la reconnaissance de la maladie, une étude souligne que lorsqu'un instrument de dépistage validé est utilisé pour diagnostiquer spécifiquement la rosacée, sa prévalence varie de 5% à 12% (87). En tant que dermatose visible, les implications psychosociales sur la qualité de vie des patients sont documentées d'un point de vue clinique. Les déterminants de la QdV dans la rosacée comprennent des facteurs physiques tels que la douleur, l'irritation, la sensation de brûlure, la sécheresse et les symptômes oculaires (88,89). Des facteurs psychosociaux tels que la colère, la dépression, la baisse de l'estime de soi, la stigmatisation, l'inquiétude, la gêne, la phobie sociale, l'anxiété ou la frustration (90–92). Les activités quotidiennes sont également affectées, comme les jours d'absence au travail, la diminution des possibilités d'emploi ; 50 % des patients ont délibérément manqué le travail en raison de l'impact psychosocial de leur maladie, et 8,3 % des patients atteints de rosacée ont estimé que leur productivité au travail était affectée par leur maladie. Chez les patients présentant une rougeur faciale sévère associée à la rosacée, la moitié d'entre eux ont indiqué que la rougeur faciale interférait avec leur vie professionnelle (90,93). Les comorbidités psychosociales associées à la rosacée les plus courantes sont la dépression, l'anxiété sociale et la phobie sociale (94). L'anxiété sociale et la phobie sociale entraînent un stress qui peut exacerber physiologiquement les rougeurs du visage et les poussées de rosacée, en raison de la libération d'agents pro-inflammatoires (94–97). Les mesures génériques de la qualité de vie telles que le SF-36 et l'EQ-5D sont couramment utilisées pour évaluer le fardeau psychosocial des personnes souffrant de rosacée (90). Dans une étude où le SF-36 a été utilisé pour mesurer le paramètre des résultats dans un essai clinique sur la rosacée, le groupe atteint de rosacée présentait une qualité de vie inférieure dans les domaines SF-36 des perceptions générales de la santé, de la vitalité, de la sphère émotionnelle, du fonctionnement physique, de la santé mentale et de la douleur corporelle (98). Dans une métanalyse utilisant l'EQ-5D, les domaines les plus touchés étaient la douleur ou la gêne, et l'anxiété ou la dépression ; 1 624 sujets atteints de rosacée (26,4 %) ont déclaré être modérément ou extrêmement anxieux ou déprimés et ont cité l'anxiété et la dépression comorbides comme la principale raison de la diminution de la QVLS (93). Les résultats économiques ou l'impact financier des maladies de la peau n'ont pas été recherchés au cours de l'analyse documentaire, bien que la rosacée soit la seule maladie de la peau pour laquelle un impact économique a été évalué parmi les paramètres classiques de la qualité de vie, dans le profil d'impact de la qualité de vie de la maladie. Le coût annuel moyen de la prise en charge de la rosacée aux États-Unis a été évalué à 347 dollars, dont 56 dollars pour les soins médicaux et 291 dollars pour les soins pharmaceutiques (94).

Ainsi, les informations identifiées dans la littérature scientifique à propos de la qualité de vie des patients atteints de dermatoses, trouvent un écho sur les réseaux sociaux. En effet, notre étude et sa contextualisation bibliographique nous permettent d’illustrer que les réseaux sociaux peuvent être une source de données de vie réelle, à la fois concordante et complémentaire aux études classiques. Le côté qualitatif des études infodémiologiques est à la fois un vecteur d’illustration des données scientifiques existantes, mais aussi un vecteur de développement du patient-centrisme, dans le sens où les partages des difficultés et des besoins patients sont des informations utiles à l’évolution de leurs prises en charge en santé.

CONCLUSION

Les communautés de patients qui se forment, participent au changement de paradigme de la relation patient-soignant. C'est aussi vrai pour les communautés en ligne, où internet devient une source d'émancipation et d'information médicale, outil de démocratie sanitaire. Au-delà de l'intérêt pour les patients, leurs échanges en eux-mêmes sont sources d'informations utiles à la médecine. L'analyse de leurs échanges permet de comprendre des enjeux parfois plus profonds ou spécifiques, utiles à la personnalisation de l'approche de soin et des programmes patients.

Les avancées technologiques et la démocratisation de la *data science* permettent à ce jour de faire ce que l'infodémiologie n'avait probablement pas anticipé à sa création. L'évolution des algorithmes de Traitement Automatisés du Langage permet de toujours en attendre davantage. Le passage du *machine learning* au *deep learning*, le développement de l'intelligence artificielle pour s'adapter à toutes les formes de verbalisation, la détection et l'interprétation des contextes verbalisés dans les messages seront nécessaires pour standardiser les analyses. A ce jour, seules les plus grosses entreprises comme Google™ disposent de modèles linguistiques suffisamment développés. La santé est un sujet très spécifique, le développement de nouveaux algorithmes spécialisés (ou généralistes pouvant être appliqués aux écrits médicaux), l'amélioration continue des modèles existants, dans le TAL médical, seront des vecteurs de réussite.

La complémentarité des données de vie réelle présentes sur les réseaux sociaux ainsi que leur adéquation avec les données présentes dans la littérature médicale sont des preuves de l'utilité de l'infodémiologie. Cependant il existe à ce jour un fossé entre les approches classiques standardisées d'étude du vécu patient et les protocoles d'études infodémiologiques qui sont encore très qualitatifs. L'essor technologique doit permettre d'augmenter la robustesse des études sur les réseaux sociaux, de produire d'avantage d'études investigant leur crédibilité comme source d'informations utiles, optimisant leur recevabilité.

De même, l'accès aux données, leurs traitements, analyses, réutilisations, devront profiter de cadres législatifs, éthiques et méthodologiques favorables au développement de ce champ de recherche et des intelligences artificielles associées, dans des objectifs de santé publique. De plus, la recevabilité des réseaux sociaux comme source de données de vie réelle, et l'intérêt des études infodémiologiques, sont encore à prouver. La robustesse des analyses et la confiance dans les résultats, seront des enjeux prochains. Le développement de l'infodémiologie et la reconnaissance des réseaux sociaux comme source exploitable de données de vie réelle, sont permis par l'acculturation du numérique en santé par les pouvoirs publics et les laboratoires pharmaceutiques.

L'essor de l'innovation médicale et du numérique en santé en France sont des volontés, à la fois des professionnels de santé de terrain, ainsi que des pouvoirs publics. La Stratégie Nationale 2020 pour la e-santé a défini 4 grands axes principaux, dont l'objectif est d'accompagner les acteurs du système de soins dans le virage numérique et de permettre à la France de rester à la pointe en matière d'innovation. Le 4ème axe en particulier aborde les données de santé dont celles accessibles sur internet (*Big Data*), favorise la modernisation des outils de régulation de notre système de santé et des informations et données associées. Un sous-point en particulier développe le numérique au service de la veille et de la surveillance sanitaire, avec le développement de nouveaux outils de modélisation, pouvant faciliter l'anticipation des menaces épidémiques, de représentation des données environnementales pour faciliter les interventions en santé. Cet axe aborde aussi la nécessité de lever les freins au développement du *Big Data* au service de la santé, avec la loi de modernisation de notre système de santé, qui simplifie le cadre juridique lié à la circulation de l'information de santé. De plus,

l'ouverture des données (*Open Data*) est devenue une mission officielle du ministère chargé de la santé. Cela s'inscrit dans une volonté de transparence des pouvoirs publics et dans une obligation de retransmission des données générées par les citoyens et collectées par les institutions publiques. Si la volonté politique aborde ici principalement des données médico-sociales, le point sur les nouveaux outils permettant une veille numérique, justifie le développement de l'infodémiologie. Internet en tant que source de données de santé/vie réelle, alliée à des outils de récolte des données, et d'intelligence artificielle pour traiter les volumes de données, est l'un des outils, qui s'est déjà illustré dans la prédiction d'apparition et de modélisation (suivi) d'épidémies de maladies infectieuses (fièvre hémorragique, grippe, zika).

L'extraction et l'utilisation des données ouvertes et accessibles présentes sur les réseaux sociaux (messages), sont régulées et encadrées. Par définition, nos projets d'analyse des verbatims à des fins médicales se placent dans la réutilisation secondaire de données accessibles et ouvertes sur internet, car les utilisateurs ont donné leur consentement au réseau social en contrepartie de leur utilisation. Cependant des questionnements éthiques se posent, dans le sens où une fois les données extraites et présentes dans une base de données, le message restera présent même si l'utilisateur retire son consentement (le droit à l'oubli). Le Règlement Général sur la Protection des Données (RGPD) précise l'encadrement et les traitements autorisés pour les données à caractère personnel. Cependant si d'un point de vue strictement juridique les analyses infodémiologiques sont possibles, éthiquement, la question se pose du consentement patient, précisément pour l'analyse médicale de leurs verbatims. Actuellement en 2022, les citoyens français sont interrogés sur leur volonté de partage de leurs données personnelles. Une ambivalence existe entre l'intérêt du partage et le fait d'être propriétaire de leurs données. La finalité sous-jacente au partage de la donnée est un élément crucial, lorsque le but sert une cause de santé publique ou d'intérêt commun, où l'individu et ses données sont noyées dans une masse, les usagers sont davantage favorables ; cependant le rapport bénéfice/risque devient en défaveur du partage lorsque les finalités deviennent commerciales et personnalisées à l'échelle de l'individu. La confiance des usagers dans le système de partage des données de santé est liée à la communication des objectifs et des résultats de recherche. La transparence en tant que telle est un vecteur de réussite de création d'un cercle vertueux où *open data* et informations utiles au plus grand nombre s'auto-alimentent. Cela rejoint la citation « *Ne confondons pas « centré sur le patient » et « orientation client ». Être centré sur le malade, pour la médecine, n'est pas une stratégie. C'est la condition de son existence, la démarche d'où elle émerge : son origine.* »

Ces notions d'encadrement, de transparence et de confiance, sont au cœur de l'essor de l'intelligence artificielle en médecine. En effet, l'effet « boite noire » de l'IA est un phénomène connu ; discipline brumeuse, aux termes compliqués et formules mathématiques pointues, source de fantasmes, la transparence liée au fonctionnement d'un modèle d'intelligence artificielle est cruciale. C'est aussi un enjeu car tout un chacun n'est pas capable d'être juge vis-à-vis du fonctionnement d'un algorithme. L'effet boite noire autour de l'IA est le phénomène par lequel des données d'entrée sont fournies au système, qui produit des données de sortie de manière autonome sans vraiment que l'on comprenne la totalité du cheminement. Cet effet boite noir a un risque double : le danger de produire de faux résultats, et l'un des risques associés à ces conséquences, l'augmentation de la défiance dans l'utilisation de systèmes d'intelligence artificielle pour baser des décisions médicales. Actuellement en France, seul un système dispositif médical a été validé, certifié et est remboursé ; il s'agit d'un algorithme de *machine learning* présent dans un pancréas artificiel, distribuant l'insuline de manière personnalisée au patient. Ce sera le travail de la Haute Autorité de Santé de statuer sur les différents modèles possibles et leurs performances, contextualisés dans la santé. Actuellement, beaucoup de sociétés développent des algorithmes, entraînent des modèles, dans un objectif de médecine

personnalisée. Cependant les critères sur lesquels ils seront évalués sont en train d'être pensés et définis par les institutions qualifiées. Si l'accès au marché des médicaments et des produits de santé, dispositifs médicaux, est encadré, évalué et défini, dans un futur proche les systèmes d'évaluations adéquats des algorithmes, thérapies numériques, bases de données, devront eux aussi être construits. Au-delà de l'essor de la e-santé, de l'intelligence artificielle en médecine, en France, le facteur de réussite sera l'adéquation des solutions numériques aux standards d'évaluation et de compliances aux règlements et juridictions en vigueur.

1. Jean A. De l'autre côté de la machine: Voyage d'une scientifique au pays des algorithmes. Humensis; 2019. 145 p.

2. Larousse É. intelligence artificielle - LAROUSSE [Internet]. [cité 28 mai 2022]. Disponible sur: https://www.larousse.fr/encyclopedie/divers/intelligence_artificielle/187257

3. Perez JC. Père d'intelligence artificielle… John McCarthy [Internet]. Histoire-cigref.org. [cité 28 mai 2022]. Disponible sur: https://www.cigref.fr/archives/histoire-cigref/blog/pere-d-intelligence-artificielle-john-mccarthy/

4. Conseil de l'Europe. L'IA, c'est quoi ? [Internet]. Intelligence artificielle. 2022 [cité 28 mai 2022]. Disponible sur: https://www.coe.int/fr/web/artificial-intelligence/what-is-ai

5. Tanguy L. Traitement Automatique de la Langue Naturelle et interprétation:Contribution à l'élaboration d'un modèle informatique de la Sémantique Interprétative. :208.

6. Manning C, Schutze H. Foundations of Statistical Natural Language Processing. MIT Press; 1999. 657 p.

7. Maurice. Fonctionnement du neurone artificiel. Deep Learn [Internet]. 15 sept 2018 [cité 28 mai 2022]; Disponible sur: https://deeplylearning.fr/cours-theoriques-deep-learning/fonctionnement-du-neurone-artificiel/

8. Brousse C, Boisaubert B. La qualité de vie et ses mesures. Rev Médecine Interne. 1 juill 2007;28(7):458-62.

9. World Health Organisation. Study protocol for the World Health Organization project to develop a Quality of Life assessment instrument (WHOQOL). Qual Life Res Int J Qual Life Asp Treat Care Rehabil. avr 1993;2(2):153-9.

10. Haute Autorité de Santé. Évaluation des technologies de santé à la HAS : place de la qualité de vie [Internet]. 2018. Disponible sur: https://www.has-sante.fr/jcms/c_2883073/fr/evaluation-des-technologies-de-sante-a-la-has-place-de-la-qualite-de-vie

11. Haute Autorité de Santé. Qualité des soins perçue par le patient - Indicateurs PROMs et PREMs : panorama d'expériences étrangères et principaux enseignements [Internet]. [cité 28 mai 2022]. Disponible sur: https://www.has-sante.fr/jcms/p_3277049/fr/qualite-des-soins-percue-par-le-patient-indicateurs-proms-et-prems-panorama-d-experiences-etrangeres-et-principaux-enseignements

12. Richards T, Montori VM, Godlee F, Lapsley P, Paul D. Let the patient revolution begin. BMJ. 14 mai 2013;346:f2614.

13. U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Biologics Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Devices and Radiological Health. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. Health Qual Life Outcomes. 11 oct 2006;4:79.

14. Haute Autorité de Santé. Choix méthodologique pour l'analyse de l'impact budgétaire à la HAS. 2016.

15. Aaronson NK, Acquadro C, Alonso J, Apolone G, Bucquet D, Bullinger M, et al. International Quality of Life Assessment (IQOLA) Project. Qual Life Res Int J Qual Life Asp Treat Care Rehabil. oct 1992;1(5):349-51.

16. Borms Matthias. Utilisation des outils de mesure de la qualité de vie liée à la santé dans l'accès au marché des médicaments. Lille; 2018.

17. Eysenbach G. Infodemiology: The epidemiology of (mis)information. Am J Med. 15 déc 2002;113(9):763-5.

18. Gauthier LM des P. Qu'est-ce que l' infodémiologie ? Monit Pharm [Internet]. 2018 [cité 28 mai 2022];3245(Le Moniteur des Pharmacies n° 3245 du 02/11/2018). Disponible sur: https://www.lemoniteurdespharmacies.fr/revues/le-moniteur-des-pharmacies/article/n-3245/qu-est-ce-que-l-infodemiologie.html

19. Renner S, Marty T, Khadhar M, Foulquié P, Voillot P, Mebarki A, et al. A New Method to Extract Health-Related Quality of Life Data From Social Media Testimonies: Algorithm Development and Validation. J Med Internet Res. 28 janv 2022;24(1):e31528.

20. Amir Hassan Zadeh, Hamed M Zolbanin. Social Media for Nowcasting Flu Activity: Spatio-Temporal Big Data Analysis | SpringerLink [Internet]. 2019 [cité 19 mars 2020]. Disponible sur: https://link.springer.com/article/10.1007/s10796-018-9893-0

21. Renaud C, Fernandez V, Puel G. Mèmes internet et réseaux sociaux chinois : état des lieux et perspectives d'analyse. :25.

22. A M. Infodemiology and Infoveillance: Scoping Review. J Med Internet Res [Internet]. 28 avr 2020 [cité 6 juin 2022];22(4). Disponible sur: https://pubmed.ncbi.nlm.nih.gov/32310818/

23. Ministère des solidarités et de la santé. Les données de vie réelle, un enjeu pour la qualité des soins et la régulation du système de santé. 2017.

24. Haute Autorité de Santé, Polton Dominique. Colloque HAS - Données de vie réelle : un enjeu majeur, une dynamique qui s'accélère. In 2019.

25. Nabhan C, Klink A, Prasad V. Real-world Evidence-What Does It Really Mean? JAMA Oncol. 1 juin 2019;5(6):781-3.

26. Khozin S, Blumenthal GM, Pazdur R. Real-world Data for Clinical Evidence Generation in Oncology. J Natl Cancer Inst. 1 nov 2017;109(11).

27. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev. 29 avr 2014;(4):MR000034.

28. Bartlett VL, Dhruva SS, Shah ND, Ryan P, Ross JS. Feasibility of Using Real-World Data to Replicate Clinical Trial Evidence. JAMA Netw Open. 2 oct 2019;2(10):e1912869.

29. Zhu M, Sridhar S, Hollingsworth R, Chit A, Kimball T, Murmello K, et al. Hybrid clinical trials to generate real-world evidence: design considerations from a sponsor's perspective. Contemp Clin Trials. juill 2020;94:105856.

30. Signaler un risque pour la santé publique - Ministère des Solidarités et de la Santé [Internet]. [cité 18 juin 2022]. Disponible sur: https://solidarites-sante.gouv.fr/prevention-en-sante/securite-sanitaire/article/signaler-un-risque-pour-la-sante-publique

31. 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.

32. Actualité - Méthylphénidate : données d'utilisation et de sécurité d'emploi en France - ANSM [Internet]. [cité 18 juin 2022]. Disponible sur: https://ansm.sante.fr/actualites/methylphenidate-donnees-dutilisation-et-de-securite-demploi-en-france

33. EuroQol Group. EuroQol--a new facility for the measurement of health-related quality of life. Health Policy Amst Neth. déc 1990;16(3):199-208.

34. Heusse C, publique (EHESP) E des hautes études en santé. La qualité de vie : un indicateur pertinent pour l'évaluation d'impact des programmes d'intervention de Handicap International [Internet]. 2014. 111p. Disponible sur: http://documentation.ehesp.fr/memoires/2014/shps/heusse.pdf

35. Maxwell A, Özmen M, Iezzi A, Richardson J. Deriving population norms for the AQoL-6D and AQoL-8D multi-attribute utility instruments from web-based data. Qual Life Res. déc 2016;25(12):3209-19.

36. Forums - Atoute.org [Internet]. [cité 14 sept 2021]. Disponible sur: https://www.atoute.org/n/rubrique1.html

37. Doctissimo. Santé et bien être avec Doctissimo [Internet]. Doctissimo. [cité 14 sept 2021]. Disponible sur: https://www.doctissimo.fr/

38. Forum aufeminin [Internet]. [cité 14 sept 2021]. Disponible sur: https://forum.aufeminin.com/forum/

39. Forum Journal des Femmes Santé - Journal des Femmes [Internet]. Journal des Femmes Santé. [cité 14 sept 2021]. Disponible sur: https://sante-medecine.journaldesfemmes.fr/forum/

40. Psychoactif [Internet]. Psychoactif, l'espace solidaire entre usagers de drogues. [cité 14 sept 2021]. Disponible sur: https://www.psychoactif.org

41. Forum HardWare.fr : Discussions Informatiques & Généralistes [Internet]. [cité 14 sept 2021]. Disponible sur: https://forum.hardware.fr/

42. Le forum cancer du sein des Impatientes, la communauté des femmes contre le cancer du sein [Internet]. [cité 14 sept 2021]. Disponible sur: http://www.lesimpatientes.com/

43. Laxophobie et Colopathie fonctionnelle • Page d'index [Internet]. [cité 14 sept 2021]. Disponible sur: http://www.laxophobie.fr/

44. Forum de discussion des parents et futurs parents : grossesse, femme, bébé, enfant, éducation, maman, papa, famille, recettes, maison, loisirs [Internet]. [cité 14 sept 2021]. Disponible sur: http://forum.magicmaman.com/

45. « Vivre sans thyroïde » - Forum de discussion [Internet]. [cité 14 sept 2021]. Disponible sur: https://www.forum-thyroide.net/

46. Forum Ados [Internet]. Public.fr. [cité 14 sept 2021]. Disponible sur: https://www.public.fr/Forum-Ados

47. Onmeda [Internet]. Perles des forums. [cité 14 sept 2021]. Disponible sur: http://perles-des-forums.fr/forum/onmeda

48. Forum Psychologies [Internet]. les Forums de Psychologies.com. [cité 14 sept 2021]. Disponible sur: https://forum.psychologies.com/

49. Expériences avec des médicaments | meamedica.fr [Internet]. [cité 14 sept 2021]. Disponible sur: https://www.meamedica.fr/

50. Forum FS Generation [Internet]. [cité 14 sept 2021]. Disponible sur: https://forums.futura-sciences.com/

51. AlloDocteurs : Actualités santé, émission, maladies, symptômes [Internet]. AlloDocteurs. [cité 14 sept 2021]. Disponible sur: https://www.allodocteurs.fr

52. Forum Santé, Forum Médical [Internet]. Vulgaris Médical. [cité 14 sept 2021]. Disponible sur: https://www.vulgaris-medical.com/forum-sante

53. France Lymphome Espoir - Association de malades du lymphome [Internet]. France Lymphome Espoir. [cité 14 sept 2021]. Disponible sur: https://www.francelymphomeespoir.fr/

54. Forum | Mamanpourlavie.com [Internet]. [cité 14 sept 2021]. Disponible sur: https://www.mamanpourlavie.com/forum/

55. MedDRA [Internet]. [cité 10 juill 2022]. Disponible sur: https://www.meddra.org/

56. Open Medic : base complète sur les dépenses de médicaments interrégimes - data.gouv.fr [Internet]. [cité 10 juill 2022]. Disponible sur: https://www.data.gouv.fr/fr/datasets/open-medic-base-complete-sur-les-depenses-de-medicaments-interregimes/

57. Abdellaoui R, Foulquié P, Texier N, Faviez C, Burgun A, Schück S. Detection of Cases of Noncompliance to Drug Treatment in Patient Forum Posts: Topic Model Approach. J Med Internet Res. 14 mars 2018;20(3):e85.

58. La version française du dictionnaire pour le LIWC : modalités de construction et exemples d'utilisation - EM consulte [Internet]. [cité 15 juill 2022]. Disponible sur: https://www.em-consulte.com/article/658282/la-version-francaise-du-dictionnaire-pour-le-liwc-

59. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 1 juin 2002;16:321-57.

60. Apprentissage artificiel : Évaluation de l'apprentissage – Précision, Rappel et F-mesure | Blog Onyme [Internet]. [cité 15 juill 2022]. Disponible sur: http://blog.onyme.com/apprentissage-artificiel-evaluation-precision-rappel-f-mesure/

61. Caster O, Dietrich J, Kürzinger ML, 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. déc 2018;41(12):1355-69.

62. Colilla S, Tov EY, Zhang L, Kurzinger ML, Tcherny-Lessenot S, Penfornis C, et al. Validation of New Signal Detection Methods for Web Query Log Data Compared to Signal Detection Algorithms Used With FAERS. Drug Saf. mai 2017;40(5):399-408.

63. Mallon E, Newton JN, Klassen A, Stewart-Brown SL, Ryan TJ, Finlay AY. The quality of life in acne: a comparison with general medical conditions using generic questionnaires. Br J Dermatol. avr 1999;140(4):672-6.

64. Klassen AF, Newton JN, Mallon E. Measuring quality of life in people referred for specialist care of acne: comparing generic and disease-specific measures. J Am Acad Dermatol. août 2000;43(2 Pt 1):229-33.

65. Rapp SR, Feldman SR, Exum ML, Fleischer AB, Reboussin DM. Psoriasis causes as much disability as other major medical diseases. J Am Acad Dermatol. sept 1999;41(3 Pt 1):401-7.

66. van de Kerkhof PCM. The impact of a two-compound product containing calcipotriol and betamethasone dipropionate (Daivobet/ Dovobet) on the quality of life in patients with psoriasis vulgaris: a randomized controlled trial. Br J Dermatol. sept 2004;151(3):663-8.

67. Gupta MA, Gupta AK, Kirkby S, Schork NJ, Gorr SK, Ellis CN, et al. A psychocutaneous profile of psoriasis patients who are stress reactors. A study of 127 patients. Gen Hosp Psychiatry. mai 1989;11(3):166-73.

68. Seville RH. Psoriasis and stress. Br J Dermatol. sept 1977;97(3):297-302.

69. Ginsburg IH, Link BG. Feelings of stigmatization in patients with psoriasis. J Am Acad Dermatol. janv 1989;20(1):53-63.

70. Gupta MA, Schork NJ, Gupta AK, Kirkby S, Ellis CN. Suicidal ideation in psoriasis. Int J Dermatol. mars 1993;32(3):188-90.

71. Finlay AY, Coles EC. The effect of severe psoriasis on the quality of life of 369 patients. Br J Dermatol. févr 1995;132(2):236-44.

72. Ramsay B, O'Reagan M. A survey of the social and psychological effects of psoriasis. Br J Dermatol. févr 1988;118(2):195-201.

73. Anderson M, Mckee M, Mossialos E. Covid-19 exposes weaknesses in European response to outbreaks. BMJ. 2020;

74. Kiebert G, Sorensen SV, Revicki D, Fagan SC, Doyle JJ, Cohen J, et al. Atopic dermatitis is associated with a decrement in health-related quality of life. Int J Dermatol. mars 2002;41(3):151-8.

75. Solomon CR, Gagnon C. Mother and child characteristics and involvement in dyads in which very young children have eczema. J Dev Behav Pediatr JDBP. août 1987;8(4):213-20.

76. Nordlund JJ. The epidemiology and genetics of vitiligo. Clin Dermatol. déc 1997;15(6):875-8.

77. Kent G, al-Abadie M. Factors affecting responses on Dermatology Life Quality Index items among vitiligo sufferers. Clin Exp Dermatol. sept 1996;21(5):330-3.

78. Salzer BA, Schallreuter KU. Investigation of the personality structure in patients with vitiligo and a possible association with impaired catecholamine metabolism. Dermatol Basel Switz. 1995;190(2):109-15.

79. Van Moffaert M. Psychodermatology: an overview. Psychother Psychosom. 1992;58(3-4):125-36.

80. Porter J, Beuf AH, Nordlund JJ, Lerner AB. Psychological reaction to chronic skin disorders: a study of patients with vitiligo. Gen Hosp Psychiatry. avr 1979;1(1):73-7.

81. Porter JR, Beuf AH, Lerner A, Nordlund J. Psychosocial effect of vitiligo: a comparison of vitiligo patients with « normal » control subjects, with psoriasis patients, and with patients with other pigmentary disorders. J Am Acad Dermatol. août 1986;15(2 Pt 1):220-4.

82. Ongenae K, Van Geel N, De Schepper S, Naeyaert JM. Effect of vitiligo on self-reported health-related quality of life. Br J Dermatol. juin 2005;152(6):1165-72.

83. Porter JR, Beuf AH, Lerner AB, Nordlund JJ. The effect of vitiligo on sexual relationships. J Am Acad Dermatol. févr 1990;22(2 Pt 1):221-2.

84. McAleer MA, Fitzpatrick P, Powell FC. Papulopustular rosacea: prevalence and relationship to photodamage. J Am Acad Dermatol. juill 2010;63(1):33-9.

85. Berg M. Epidemiological studies of the influence of sunlight on the skin. Photodermatol. avr 1989;6(2):80-4.

86. Elewski BE, Draelos Z, Dréno B, Jansen T, Layton A, Picardo M. Rosacea - global diversity and optimized outcome: proposed international consensus from the Rosacea International Expert Group. J Eur Acad Dermatol Venereol JEADV. févr 2011;25(2):188-200.

87. Tan J, Schöfer H, Araviiskaia E, Audibert F, Kerrouche N, Berg M, et al. Prevalence of rosacea in the general population of Germany and Russia - The RISE study. J Eur Acad Dermatol Venereol JEADV. mars 2016;30(3):428-34.

88. Karimkhani C, Dellavalle RP, Coffeng LE, Flohr C, Hay RJ, Langan SM, et al. Global Skin Disease Morbidity and Mortality: An Update From the Global Burden of Disease Study 2013. JAMA Dermatol. 1 mai 2017;153(5):406-12.

89. Two AM, Wu W, Gallo RL, Hata TR. Rosacea: part I. Introduction, categorization, histology, pathogenesis, and risk factors. J Am Acad Dermatol. mai 2015;72(5):749-58; quiz 759-60.

90. Oussedik E, Bourcier M, Tan J. Psychosocial Burden and Other Impacts of Rosacea on Patients' Quality of Life. Dermatol Clin. avr 2018;36(2):103-13.

91. Garnis-Jones S. Psychological aspects of rosacea. J Cutan Med Surg. juin 1998;2 Suppl 4:S4-16-9.

92. Dirschka T, Micali G, Papadopoulos L, Tan J, Layton A, Moore S. Perceptions on the Psychological Impact of Facial Erythema Associated with Rosacea: Results of International Survey. Dermatol Ther. juin 2015;5(2):117-27.

93. Bewley A, Fowler J, Schöfer H, Kerrouche N, Rives V. Erythema of Rosacea Impairs Health-Related Quality of Life: Results of a Meta-analysis. Dermatol Ther. juin 2016;6(2):237-47.

94. Huynh TT. Burden of Disease: The Psychosocial Impact of Rosacea on a Patient's Quality of Life. Am Health Drug Benefits. juill 2013;6(6):348-54.

95. Reich A, Wójcik-Maciejewicz A, Slominski AT. Stress and the skin. G Ital Dermatol E Venereol Organo Uff Soc Ital Dermatol E Sifilogr. avr 2010;145(2):213-9.

96. Klaber R, Wittkower E. The Pathogenesis of Rosacea: A Review with Special Reference to Emotional Factors.*. Br J Dermatol. 1939;51(12):501-24.

97. Beerman H. A re-evaluation of the rosacea complex. Am J Med Sci. oct 1956;232(4):458-73.

98. Tan J, Almeida LMC, Bewley A, Cribier B, Dlova NC, Gallo R, et al. Updating the diagnosis, classification and assessment of rosacea: recommendations from the global ROSacea COnsensus (ROSCO) panel. Br J Dermatol. févr 2017;176(2):431-8.

Annexes

Article original : A new method to extract health-related quality of life data from social media testimonies : algorithm development and validation

Original Paper

A New Method to Extract Health-Related Quality of Life Data From Social Media Testimonies: Algorithm Development and Validation

Simon Renner[1*], PharmD; Tom Marty[1*], PharmD; Mickaïl Khadhar[1*], MSc; Pierre Foulquié[1], MSc; Paméla Voillot[1], MSc; Adel Mebarki[1], MSc; Ilaria Montagni[2], DPhil; Nathalie Texier[1], PharmD; Stéphane Schück[1], MD

[1]Kap Code, Paris, France
[2]Bordeaux Population Health Research Center, UMR 1219, Bordeaux University, Inserm, Bordeaux, France
[*]these authors contributed equally

Corresponding Author:
Simon Renner, PharmD
Kap Code
4 Rue de Cléry
Paris, 75002
France
Phone: 33 9 72 60 57 63
Email: simon.renner@kapcode.fr

Abstract

Background: Monitoring social media has been shown to be a useful means to capture patients' opinions and feelings about medical issues, ranging from diseases to treatments. Health-related quality of life (HRQoL) is a useful indicator of overall patients' health, which can be captured online.

Objective: This study aimed to describe a social media listening algorithm able to detect the impact of diseases or treatments on specific dimensions of HRQoL based on posts written by patients in social media and forums.

Methods: Using a web crawler, 19 forums in France were harvested, and messages related to patients' experience with disease or treatment were specifically collected. The SF-36 (Short Form Health Survey) and EQ-5D (Euro Quality of Life 5 Dimensions) HRQoL surveys were mixed and adapted for a tailored social media listening system. This was carried out to better capture the variety of expression on social media, resulting in 5 dimensions of the HRQoL, which are physical, psychological, activity-based, social, and financial. Models were trained using cross-validation and hyperparameter optimization. Oversampling was used to increase the infrequent dimension: after annotation, SMOTE (synthetic minority oversampling technique) was used to balance the proportions of the dimensions among messages.

Results: The training set was composed of 1399 messages, randomly taken from a batch of 20,000 health-related messages coming from forums. The algorithm was able to detect a general impact on HRQoL (sensitivity of 0.83 and specificity of 0.74), a physical impact (0.67 and 0.76), a psychic impact (0.82 and 0.60), an activity-related impact (0.73 and 0.78), a relational impact (0.73 and 0.70), and a financial impact (0.79 and 0.74).

Conclusions: The development of an innovative method to extract health data from social media as real time assessment of patients' HRQoL is useful to a patient-centered medical care. As a source of real-world data, social media provide a complementary point of view to understand patients' concerns and unmet needs, as well as shedding light on how diseases and treatments can be a burden in their daily lives.

(J Med Internet Res 2022;24(1):e31528) doi: 10.2196/31528

KEYWORDS

health-related quality of life; social media use; measures; real world; natural language processing; social media; NLP; infoveillance; quality of life; digital health; social listening

Introduction

Most people use the internet regularly to research and discuss health-related topics. Patients give and receive advice on their diseases and treatments in online forums and social media platforms [1]. These messages are massive, continuously generated, and easy to access [2]. This type of information is direct, genuine, and authentic, offering access to new real-world

data, which can facilitate the understanding of patients' perspectives. As the internet offers anonymity, patients talk about their fears and concerns and share details about their diseases and treatments, which can inform health public authorities, pharmaceutical companies, and other health professionals and institutions [3]. Thus, social media are a large and diverse source of information nurtured by continuous exchanges and interactions, ranging from commenting on posts to sharing of opinions.

The World Health Organization defines quality of life (QoL) as individuals' perception of their place in life in the context of the culture and the value system in which they live, as well as in relation to their objectives, expectations, standards, and concerns. This is a broad conceptual field, encompassing, in a complex way, a person's physical health, psychological state, level of independence, social relationships, personal beliefs, and relationship with the specificities of the surrounding environment [4]. When the study of QoL is restricted to health-related effects, one can refer to them as health-related quality of life (HRQoL) [5]. Therefore, HRQoL is a multidimensional concept focusing on the impact health and diseases have on QoL [6,7]. This concept is mainly used in epidemiology and cost-effectiveness analysis [8].

Several instruments have been developed to quantitatively measure individuals' HRQoL [9]. Among them, the EQ-5D (Euro Quality of Life 5 Dimensions) and SF-36 (Short Form Health Survey) have been used in medical practice for more than 20 years [10,11]. They are designed to be self-completed by patients. Nonetheless, these surveys are not adapted to the amount of qualitative information on QoL contained within the free speech and various testimonies of patients' populations on social media.

It has been suggested that the measurement of HRQoL can benefit from machine-driven, quantitative analysis of patient-generated data, which expands hypothesis testing based on patient input regarding disease experience, lifestyle preferences, functioning, and more [12]. Opinions and advice shared on social media can provide insights on HRQoL directly from patients in real-life conditions [13].

Social media listening is the collection and interpretation of all patients' social media conversations, which can help discover what really impacts patients' lives [14]. Social media listening aggregates large amounts of unstructured patient-centered data points to identify behavioral patterns and obtain medical insights without infringing privacy policy and personal rights. Social media listening uses text mining and the natural language processing (NLP) approach as an algorithmic toolbox for identifying and managing texts of interest [15].

Against this background, the objective of this study was to develop an algorithm that is able to detect and measure the mentions of impact of diseases and treatments on 5 HRQoL dimensions in patient's testimonies through the scope of social media listening.

Methods

This study was conducted through several main steps: QoL definition, literature review, data extraction and manual treatment, annotation, preprocessing and feature engineering, modeling, and statistical analysis.

Health-Related Quality of Life

The European Knowledge Society on Quality and HRQoL has compared the many definitions of HRQoL and discussed the existing confusions between health, QoL, HRQoL, and well-being [8]. The EQ-5D tool is recommended by the French public institute that regulates recommendations toward health products, their uses, and efficacy measurement (Haute Autorité de Santé [16,17]). The SF-36 is another validated generic medical survey investigating HRQoL, broadly used by practitioners for years. Three dimensions are always at the heart of the definitions or surveys: physical, psychological, and social. However, exploring HRQoL (especially on social media, with the spontaneous discussions of patients) can shed light on other views and aspects of an individual, including economic, spiritual, or even political matters. Therefore, in addition to the 3 constant dimensions (physical, psychological, and social dimension), 2 more dimensions were added to the methodology for their important role in one's life, which can especially be impacted in the case of diseases. The dimension of generic activity is unavoidable in one's life and can be limited in some health states, from taking a shower to professional activities; therefore, the aim of analyzing a 4th dimension is to detect mentions of impact on patients' activity and autonomy, which are complementary to the physical dimension that focuses on body impairments. The 5th dimension is the financial one; according to the definitions developed by the European Knowledge Society on Quality and HRQoL, economic and personal finances are important contextual factors to patients [8]. Some can encounter bad or no insurances toward treatment costs or must pay for parallel cares or products that are not covered by their insurance. Patients can express specific health expenses or the necessity to have a specific budget because of their disease; therefore, the financial dimension covers this relation between health state and the impact on one's finances as expressed by patients in their messages.

A previous work by Cotté et al [13] showed that posts from social media could be used to assess the impact of a disease or a treatment on HRQoL. This study, focused on the narratives of cancer patients treated with immunotherapies, highlighted that posts from patients could provide additional information on HRQoL to conventional QoL measurement instruments (ie, QLQ-C30 [Quality of Life Questionnaire] and FACT-G [Functional Assessment of Cancer Therapy—General]).

Literature Review

We searched on PubMed and Google Scholar for articles responding to the following keywords: (natural language processing[MeSH Terms] OR processing natural language[MeSH Terms]) AND (quality of life[MeSH Terms] OR health related quality of life[MeSH Terms] OR healthrelated quality of life[MeSH Terms] OR HRQOL[MeSH Terms] OR cost of illness[MeSH Terms] OR disease burden[MeSH Terms]

OR sickness impact profile[MeSH Terms]). The selected results were based on NLP, social media, patients' messages, QoL, diseases, and side effects. About 40 articles were found and used with the aim of establishing the best method and modeling to adopt (Multimedia Appendix 1). A focus was made on articles that developed machine learning techniques over neural network because of their lower cost in resources and correspondence with our database. The takeaway from literature review is that some machine learning methods, tools, or approaches were highlighted for their good performance in the literature review, such as Naive Bayes, Max Entropy, Decision Tree (10 folds), and MaxVote. AdaBoost has been used for its performance boost in the learning phases. The overall performances showed that the combination of a binary classifier was better than the use of only 1 predictive model. Concerning the supervised learning for text classification, a stacked generalization method, such as SVM-L (support vector machine light), SVM-R (support vector machine regression), GBDT (gradient boosting decision tree), Unigram, bigrams, POS (part of speech), and TF-IDF (term frequency-inverse document frequency), has proven interesting for obtaining state-of-the-art results [18].

Data Sources and Manual Treatment

The sources of data were 19 online general or health-related community forums in France, which are as follows: Atoute [19], Doctissimo [20], AuFeminin [21], Journal des femmes [22], Psychoactif [23], Forum.hardware [24], Lesimpatientes [25], Laxophobie [26], Magic maman [27], thyroide [28], forum ado/public.fr [29], Onmeda [30], Psychologies [31], MeaMedica [32], Futura-sciences [33], Allodocteurs [34], Vulgaris Medical [35], Lymphome espoir [36], and Maman pour la vie [37]. Facebook and Twitter were not included because tweets are limited to 240 characters, which limited the probability of disease history development and impact testimonies. Facebook was also discarded for data privacy questions and difficulties of access. Messages were extracted using a web crawler technology [38,39]. Health-related messages were selected based on a named entity recognition (NER) module. NER is a process where a sentence is parsed through to find entities (names, organizations, locations, and quantities). The NER module was used here to identify drug or disease mentions using an approximate matching algorithm. These messages were then preprocessed and stored. The metadata extracted along with the text were the date and hour of post.

Raw data sets were composed of randomly selected health-related messages according to the presence of treatment or disease in it. Preprocessing of the extracted data included a code attribution to every message as identifier, the detection of sentences, normalization, and deduplication; since the extracted data were unstructured, this was a necessary first step to process patients' posts.

Annotation

The corpus (n=1399 posts), with 1000 (71%) posts with disease mentions and 399 (29%) posts with treatment mentions, was first manually annotated. Manual annotation was performed by 2 individuals: a health-specialized data scientist and a health care professional specialized in social media listening, both sensitized and trained about the medical field of QoL, following guidelines in accordance with the methodology of HRQoL. The 2 annotators' profiles worked in synergy in the approach of data annotation with a medical finality. Medical insight toward patients' testimonies was brought by one of the annotators, with an expert eye toward the variables to be included in the future models by the other. The 1399 health-related messages extracted from forums were split into 2 sets for labelling; respectively, 900 and 499 messages for the 2 annotators. The aim of this step was to classify the messages according to 5 specific dimensions corresponding to 5 different types of impact: physical, psychic, activity-related, relational, and financial. The labels data were either "not impacted" or "impacted." If "impacted," the concerned dimensions were characterized through annotation, and the patients' expressions of the said impact were extracted. This collection allowed the identification of specific features for each dimension being impacted, capturing the patients' vocabulary when mentioning the impact. To evaluate the annotation homogeneity, a subset of 100 messages coming from the data scientist's data set was blindly annotated by the health care specialist, allowing to calculate the kappa coefficient. The kappa coefficient for interrater reliability for the presence of a general HRQoL impact was 0.724; for the physical impact, it was 0.871; for the psychic impact, 0.663; for the activity-related impact, 0.639; and for the relational impact, 0.649; this is while no messages mentioned a financial impact in this subset (κ=0). Thus, agreement ranged from strong to very strong according to the kappa Cohen coefficient scale for 4 of the dimensions, but not for the financial one because no financial impact was mentioned in the subset of messages. This high interrater reliability for 4 of the 5 dimensions suggests that the used guidelines and training about the HRQoL ensured a homogeneous annotation of the messages.

Preprocessing and Feature Engineering

All impact-related messages were used to generate dimension-specific features. Other features were based on the message structure, such as expressed sentiment (eg, positive, negative, anger, disgust, fear, joy, sadness, and surprise), grammar (eg, count of pronouns, who is writing, and negative sentences), and conjugation (eg, count of verb tenses). A lexical field score corresponding to each HRQoL dimension was computed by counting the associated expressions previously collected during the annotation stage. We used the R packages of the Detec't extractor [39,40] to create lexical variables. This phase enabled the development of specific models of impact detection per dimension. The rationale behind this process was to be able to adapt to the many expressions of the patients. Psychic impacts and physical impacts are different, and so are the expressions used to describe them. Hence, having specific models by dimension is a way to minimize an interpretation bias.

We ended the process with a data set or corpus composed of quantitative features such as expressed sentiments (from the Linguistic Inquiry and Word Count dictionary), grammar, conjugation, and lexical fields of HRQoL-related features.

Model Selection

We used data mining and machine learning technologies to categorize and analyze retrieved data of our final corpus

according to our predefined objective. As our features do not exhibit negative values, we normalized our data by dividing all feature values by their respective maximum so that all values would be somewhere between 0 and 1, thus minimizing interclass and intraclass variances. All the missing values were replaced by the median so as not to influence intraclass variance.

We obtained a first classification algorithm to determine if there was an impact on HRQoL (corresponding to the first step of manual annotation). Subsequently, we created a classification algorithm for each dimension to assess whether the impact concerned the related dimension (second step).

We used a 5-fold sequential forward floating selector with an extreme gradient boosting algorithm to select the best features combination. We tried first to maximize the model accuracy, but we ended up with several false negative cases. We finally chose the area under the curve (AUC) as our scoring method to maximize the true positive rate because we would rather have a slightly larger number of posts containing an impact, even with false positive, than missing some of these.

We chose sequential forward floating selector over LASSO (least absolute shrinkage and selection operator) to maximize the ROC (receiver operating characteristic) value, while LASSO is trying to minimize the cost function. This allowed to obtain the best performances for all classes instead of the majority class.

We then tried several machine learning algorithms, the K-Nearest Neighbors, SVM, Multi-Layer Perceptron, Random Forest, and finally XGBoost.

Except for the psychic dimension, XGBoost was far above the other methods in terms of AUC (Table 1).

We then performed a 5-fold cross-validated grid search on our selected features to tune our hyperparameters. We split our training set into 5 samples and trained the algorithm successively on 4 of these samples, while the last sample was used as validation set. This method allowed minimizing overfitting and making sure that the models generalize well. We varied the learning rate, the number of epochs, the number of trees and their maximum depth, the minimum weight needed in a child node, the minimum loss reduction required to make a further partition on a leaf node, and the L1 regularization. LASSO regression was preferable for feature selection in case of a great number of features, making nonimportant features even more insignificant in term of weights.

Table 1. AUC (area under the curve) values of the different machine learning methods.

Algorithm	Impact[a]	Physical	Psychic	Activity	Relational	Financial
KNN[b]	69.9	66.5	64.9	64.4	68.6	65.6
SVM[c]	67.9	63.3	56.3	55.3	57.7	56.9
MLP[d]	74.6	67.9	61.6	64.5	64.5	58.6
RF[e]	75	70.5	70.7	71.8	70.9	69
XGB[f]	78.5	75	71	76	71.7	76.5

[a]At least 1 impact on at least 1 of the 5 dimensions.

[b]KNN: K-Nearest Neighbors.

[c]SVM: support vector machine.

[d]MLP: Multi-Layer Perceptron.

[e]RF: Random Forest.

[f]XGB: XGBoost.

This process allowed elaborating a model that can detect a general impact. The developed algorithm filtered the corpus of messages into 2 categories: HRQoL impacted or not. For each model, we selected the relevant variables by applying the sequential forward floating selector and chose which combination could better separate an impact message from a nonimpact message. In a nutshell, it removes or adds one feature at a time on the classifier and test performances until it reaches the best possible score. The same steps were then reproduced in each dimension according to their specific features in order to obtain specific algorithms fitted for each dimension.

Features of patient expressions specific to each impact were identified with the Linguistic Inquiry and Word Count dictionary, which provides expressions for various feelings, such as positivity, negativity, joy, sadness, disgust, surprise, fear, and anger. The frequency of these expressions within the posts was used to select the relevant variables for each impact domain (Table 2). Patterns identified during data labelling were also used to select relevant variables. We can assume than to describe daily actions and difficulties, the present tense is the most appropriate tense. Conversely, to talk about an impact within the family, "we" is more often used.

Due to the lack of a specific dimension's impact mention, some classes were imbalanced regarding one another; in order to correct that, we created an artificially balanced class by using the oversampling method SMOTE (synthetic minority oversampling technique) [41]. Based on the mathematical structure of the under-represented messages, this technique artificially creates similar examples that fit the same feature pattern in order to balance the categories. We used this method for the activity-related, relational, and financial impact algorithms.

Table 2. Most important features by model.

Dimension	Feature 1	Feature 2	Feature 3
Impact	Number of infinitive verbs	Count of first person of singular markers	Counted sadness expressions
Physical	Counted physical related expressions	Counted negative expressions	Number of negations
Psychic	Counted psychic-related expressions	Counted anger expressions	Counted fear expressions
Activity-Related	Counted professional, academic, or daily activity-related expressions	Count of verbs in present tense	Number of pronouns
Relational	Counted relational expressions	Count of first person of plural markers	Count of past participle verbs
Financial	Count of financial expressions	Number of "not"	Count of verbs in past tense

Statistical Analysis

We used sensitivity (defined as correctly identifying an HRQoL impact when classified as so by our algorithm) and specificity (defined as correctly identifying a message without impact when classified as so by our algorithm). The ROC curve and the AUC were considered to measure the overall performance of the algorithm. The ROC curve represented the true positive rate (sensitivity) plotted in function of the false positive rate (100-specificity) for different thresholds of the metric.

Results

Corpus

We extracted 20,000 messages from health-related forums mentioning diverse and different diseases such as cancers, diabetes, endometriosis, and psychological afflictions, from defined diagnosis to syndrome name (eg, nausea, "feeling blue/depressed"). Treatments such as vaccines, Levothyrox (thyroid hormones) and psychiatric drugs were also mentioned. The goal was to constitute a representative panel of health impairments, including physical, psychological, frequent, rare, light, and heavy afflictions. This corpus merged random messages mentioning 1280 medical terms (at least 1 term per message, disease, or medication). The diseases and treatment terms were identified with exact matching methods on MedDRA (Medical Dictionary for Regulatory Activities). Of the 20,000 extracted messages posted from 2000 to 2019, we randomly selected 3000 (15%) messages, which were split into 1000 and 2000. We removed duplicate entries so that we finally annotated 1399 messages: 1000 (71%) related to diseases and 399 (29%) to treatments. In the end, we had 818 (58%) messages showing at least 1 impact on QoL, 442 (31%) showing physical impact, 519 (37%) psychic, 363 (25%) activity-related, 193 (13%) relational, and 69 (4%) financial (Table 3). Many impacts on more than 1 dimension can be expressed in messages by patients.

The final corpus was then composed of 1399 French forum messages extracted from 19 conversation threads. These messages were written by users in an informal style. The length ranged from a few words to narratives longer than 1000 characters, the average message length being 905 (SD 1041) characters.

Table 3. Number of messages showing health-related quality of life impact, at least 1 impact, and by dimension.

Dimension	Message, n (%)
At least 1 impact	818 (58)
Physical	442 (31)
Psychic	519 (37)
Activity-Related	363 (25)
Relational	193 (13)
Financial	69 (4)

Modeling

From our 1399 annotated messages, we chose to split them in a 70:30 ratio where 70% of the messages were used for the training phase and the rest as validation. Out of the 1399 messages, 420 (30%) were used to evaluate the model. Among these 420 messages, 203 (48%) were predicted with an impact.

We searched for lexical fields in order to evaluate the attribution of a score per dimension. We tested the different machine learning algorithms to optimize the parameters and the results. Extreme gradient boosting was the chosen model for both impact detection and specific dimension identification. The final HRQoL impact detection algorithm was composed of several models, including a model that identified the presence of an impact and all the impact-flagged messages, which went through each specific dimension model. The models were trained using cross-validation and hyperparameter optimization. Oversampling was used to augment infrequent dimensions. This allowed us to detect a general impact on HRQoL with a sensitivity of 0.8 and a specificity of 0.7 (Table 4). Overall, 818 messages presented an impact and 581 did not. For physical impact, sensitivity was 0.56, and specificity was 0.857; for psychic impact, 0.58 and 0.828; for activity-related impact, 0.71 and 0.79; for relational impact, 0.675 and 0.73; and for financial impact, 0.77 and 0.814, respectively.

Table 4. Overall results of the different HRQoL[a]-impacted dimensions.

Dimension	F-measure	ROC[b] curve
At least 1 impact	78.6	78.5
Physical	70	75
Psychic	68	71
Activity-Related	75	76
Relational	70.6	71.7
Financial	76	76.5

[a]HRQoL: health-related quality of life.

[b]ROC: receiver operating characteristic.

Discussion

Principal Findings

We developed an algorithm to evaluate the impact diseases and treatments can have on patients' HRQoL based on their emotions and opinions shared on social media. The algorithm was based on an adaptation for the social media listening approach, of the EQ-5D and SF-36 scales, which are recommended by several national and international institutions for assessing HRQoL and whose psychometric proprieties are well known [7,42,43]. Five dimensions of impact on HRQoL were then covered and identified in a filtered corpus of 1399 messages. The algorithm was able to detect different types of disease and treatment impact on HRQoL with good sensitivity and specificity. The algorithm had an ROC score of 0.785 for detecting at least 1 impact on at least 1 of the 5 dimensions (0.75 for physical dimension, 0.71 for psychic, 0.76 for activity-related, 0.717 for relational, and 0.765 for financial). Compared to other studies [44,45], these indicators were high and robust; for example, with Twitter and Facebook data, the area under the curve of Caster et al [44] varied between 0.43 and 0.67. For patient forum posts, sensitivity was 0.14 (and specificity was 0.88); and for Twitter and Facebook, sensitivity was 0.08 or lower. However, the objectives and approaches of these studies were different from ours, and it is thus quite difficult to compare the results. Performance might vary according to the data source. Considering that we were able to access a large data set and to use a satisfying training subset, this might explain our better performance. Nonetheless, Facebook and Twitter were discarded from our extracted sources due to the short messages of Twitter and the difficulties of access to Facebook data.

Social media listening allows direct monitoring of patients' messages capturing "live" their opinions and feelings compared to a punctual "fixed" self-administered questionnaire. This approach corresponds more to the evolutive nature of HRQoL.

Our study adds to the literature on the use of NLP and text mining concerning medical care from web-based data. This approach relies on the potential strength of large and real time web-based data, which are complementary to classic medical reporting systems. This work contributes to the need for an improvement in methodologies that can produce more sophisticated joint models of user and message-level information or the use of syntactic structure as their features.

A similar study was conducted to outpredict baselines of popular happy and hedonistic lexica through the satisfaction with life scale over Facebook volunteers [46]. The findings of this study were also encouraging by demonstrating the effectiveness of machine learning algorithms to detect users' health-related emotions.

Another study carried out in France [47] showed a good performance in terms of sensitivity and specificity of an NLP method to detect self-reported signals of issues with treatments. Our results confirm the same success of established statistical detection algorithms in social media for a wide range of diseases and treatments.

Strengths and Limitations

This methodological study contributes to the growing research on social media listening and machine learning in general as a technique to develop and train tools to measure broad constructs such as HRQoL. Our work is among the first research projects proving that a social media listening tool can provide a sound and efficient measurement of impacts on HRQoL directly accessible from patients to health professionals. In this sense, it highlights some of the promises of social media and forums as data sources. One of the strengths of this study was the quality of preprocessing and processing of the data extracted. Several cleansing and validation steps were performed to ensure the quality of the messages. Furthermore, we used medically validated (general) scales, the EQ-5D and SF-36, as a strong scientific basis and gold standard for the detection of 5 specific HRQoL dimensions (ie, physical, psychological, activity-related, financial, and social). Different diseases or treatments would differently affect patients; therefore, our generalist approach of the machine learning model, which has been trained based on the patients' free speech on various diseases and treatments, is able to detect different expressions of impact on our 5 common dimensions.

However, an algorithm does not have the human sensitivity to understand very specific and subjective ways of expressing a HRQoL impact (such as sarcasm), despite the constant improvement of the work. Sentimental analysis can complement such algorithms, and manual review remains strongly required. Additionally, our approach lacks flexibility in the feature extraction process; impact-specific features are not exhaustive because the expression of impact can vary. This also requires improvement in order to complete the lexical fields.

Limitations also include the data sources. More analysis is needed to prove that insights from social media are complementary to a patient-centric repository. Furthermore, Twitter and Facebook were discarded as sources due to short message format and accessibility issues; however, this does not mean that these social platforms are irrelevant resources for analyzing health testimonies from patients.

Our data were randomly extracted from a large sample of French messages coming from French forums and social media. The fact that our sample selection was random should ensure a certain representativity of the internet message population. The proportion of women speaking about their health in forums is higher than the proportion of men (difference of 6%) [48], which introduces a possible bias when exploring HRQoL. However, our algorithm is designed to work on data coming from French forums and social media with similar gender proportions.

Future work is needed to continue training the algorithm and to further study the differences on HRQoL between internet users and patients not posting messages on social media or forums.

Implications

We provided evidence that social media listening can be used to assess the impact and burden of one or more diseases and treatments on patients' HRQoL. These findings can provide public health experts, health care professionals, and pharmaceutical companies with patient-generated information on their experiences with treatments, burden of diseases, and needs for appropriate medical care in a timely manner and in real-life conditions. For instance, the generated data coming directly from patients can inform potential changes of a treatment and development of new pharmaceutical products. The use of social media listening might be recommended to monitor HRQoL constantly and consistently in patients under a new treatment or experiencing a severe disease.

Conclusion

We developed an algorithm that can translate social media patient messages into the identification of an impact on HRQoL. Based on medically validated questionnaires, this is a patient-centered approach using machine learning and NLP to better understand how diseases and treatments can represent a burden for patients.

Acknowledgments

None declared.

Conflicts of Interest

SR, TM, MK, PF, PV, AM, NT, and SS are members of Kap Code. IM is a member of the Bordeaux Population Health Research Center, UMR 1219, Bordeaux University, Inserm.

Multimedia Appendix 1

Literature references.
[DOCX File , 24 KB-Multimedia Appendix 1]

References

1. Cline RJ, Haynes KM. Consumer health information seeking on the Internet: the state of the art. Health Educ Res 2001 Dec;16(6):671-692. [doi: 10.1093/her/16.6.671] [Medline: 11780707]
2. Morlane-Hondère F, Grouin C, Zweigenbaum P. Identification of Drug-Related Medical Conditions in Social Media. 2016 May Presented at: Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC; 2016; Portoroz, Slovenia.
3. Bansal G, Zahedi F, Gefen D. The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online. Decision Support Systems 2010 May;49(2):138-150. [doi: 10.1016/j.dss.2010.01.010]
4. Nussbaum M, Sen A. The Quality of Life. Oxford, UK: Clarendon Press; 1993.
5. Measuring Quality of Life.: World Health Organization; 1997. URL: https://tinyurl.com/bvf746fa [accessed 2022-01-21]
6. Karimi M, Brazier J. Health, Health-Related Quality of Life, and Quality of Life: What is the Difference? Pharmacoeconomics 2016 Jul 18;34(7):645-649. [doi: 10.1007/s40273-016-0389-9] [Medline: 26892973]
7. Hays RD, Reeve BB. Epidemiology and Demography in Public Health. In: Measurement and Modeling of Health-Related Quality of Life. San Diego, USA: International Encyclopedia of Public Health; 2010.
8. Koonal KS. A brief review of concepts: health, quality of life, health-related quality of life and well being. Rotterdam, the Netherlands: EuroQol Research Foundation; 2017:9.
9. Lorente S, Vives J, Viladrich C, Losilla J. Tools to assess the measurement properties of quality of life instruments: a meta-review protocol. BMJ Open 2018 Jul 23;8(7):e022829 [FREE Full text] [doi: 10.1136/bmjopen-2018-022829] [Medline: 30037880]
10. EuroQol Group. EuroQol - a new facility for the measurement of health-related quality of life. Health Policy 1990 Dec;16(3):199-208. [doi: 10.1016/0168-8510(90)90421-9]
11. Ware J, Snow KK, Kosinski M, Gandek B. SF-36 Health Survey Manual Interpretation Guide. Boston, USA: The Health Institute, New England Medical Center; 1993.

12. Quality of Life Research Group. Pushing the boundaries: frontiers of quality of life research. Qual Life Res 2012 Jan;20 Suppl 1:2-117. [doi: 10.1007/s11136-011-0097-z] [Medline: 22298117]
13. Cotté FE, Voillot P, Bennett B, Falissard B, Tzourio C, Foulquié P, et al. Exploring the Health-Related Quality of Life of Patients Treated With Immune Checkpoint Inhibitors: Social Media Study. J Med Internet Res 2020 Sep 11;22(9):e19694 [FREE Full text] [doi: 10.2196/19694] [Medline: 32915159]
14. Powell GE, Seifert HA, Reblin T, Burstein PJ, Blowers J, Menius JA, et al. Social Media Listening for Routine Post-Marketing Safety Surveillance. Drug Saf 2016 May 21;39(5):443-454. [doi: 10.1007/s40264-015-0385-6] [Medline: 26798054]
15. Dreisbach C, Koleck TA, Bourne PE, Bakken S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. Int J Med Inform 2019 May;125:37-46 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.02.008] [Medline: 30914179]
16. New update of official guide to the methods of economic evaluation for France recommends use of EQ-5D-5L to derive a utility score. EQ-5D. 2020 Sep. URL: https://tinyurl.com/bdd2ywz5 [accessed 2022-01-21]
17. Haute Autorité de Santé. Choix méthodologiques pour l'évaluation économique à la HAS. Saint-Denis La Plaine, France: HAS; 2020:8.
18. Young IJB, Luz S, Lone N. A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. Int J Med Inform 2019 Dec;132:103971. [doi: 10.1016/j.ijmedinf.2019.103971] [Medline: 31630063]
19. Forums. Atoute. URL: https://forum.atoute.org/ [accessed 2022-01-21]
20. Forums Santé. Doctissimo. URL: https://forum.doctissimo.fr/ [accessed 2022-01-21]
21. Forum. aufeminin. URL: https://forum.aufeminin.com/forum/ [accessed 2022-01-21]
22. Forum Journal des Femmes Santé. Journal des Femmes Santé. URL: https://sante-medecine.journaldesfemmes.fr/forum/ [accessed 2022-01-21]
23. Psychoactif, l'espace solidaire entre usagers de drogues. Psychoactif. URL: https://www.psychoactif.org/ [accessed 2022-01-21]
24. HardWare.fr. URL: https://forum.hardware.fr/ [accessed 2022-01-21]
25. Le forum cancer du sein des Impatientes. Les Impatientes. URL: http://www.lesimpatientes.com/ [accessed 2022-01-21]
26. Laxophobie et Colopathie fonctionnelle. Laxophobie. URL: http://www.laxophobie.fr/ [accessed 2022-01-21]
27. Forum. magicmaman. URL: http://forum.magicmaman.com/ [accessed 2022-01-21]
28. Forum de discussion. Vivre sans thyroïde. URL: https://www.forum-thyroide.net/ [accessed 2022-01-13]
29. Forum Ados. Public.fr. URL: https://www.public.fr/Forum-Ados [accessed 2022-01-21]
30. Perles des forums. Onmeda. URL: https://www.onmeda.fr/ [accessed 2022-01-21]
31. Les Forums. Psychologies.com. URL: https://forum.psychologies.com/ [accessed 2022-01-21]
32. Expériences avec des médicaments. meamedica. URL: https://www.meamedica.fr/ [accessed 2022-01-21]
33. Forum FS Generation. Futura. URL: https://forums.futura-sciences.com/ [accessed 2022-01-21]
34. AlloDocteurs. URL: https://www.allodocteurs.fr/ [accessed 2022-01-21]
35. Forum Santé. Vulgaris Médical. URL: https://www.vulgaris-medical.com/forum-sante [accessed 2022-01-21]
36. Informer et soutenir Lutter contre le lymphome. France Lymphome Espoir. URL: https://www.francelymphomeespoir.fr/ [accessed 2022-01-21]
37. Forum de discussion. Maman pour la vie. URL: https://www.mamanpourlavie.com/forum/ [accessed 2022-01-21]
38. Cro spécialiste en real world evidence. Kappa Santé. URL: https://www.kappasante.com/ [accessed 2022-01-21]
39. Abdellaoui R, Foulquié P, Texier N, Faviez C, Burgun A, Schück S. Detection of Cases of Noncompliance to Drug Treatment in Patient Forum Posts: Topic Model Approach. J Med Internet Res 2018 Mar 14;20(3):e85 [FREE Full text] [doi: 10.2196/jmir.9222] [Medline: 29540337]
40. Abdellaoui R, Schück S, Texier N, Burgun A. Filtering Entities to Optimize Identification of Adverse Drug Reaction From Social Media: How Can the Number of Words Between Entities in the Messages Help? JMIR Public Health Surveill 2017 Jun 22;3(2):e36 [FREE Full text] [doi: 10.2196/publichealth.6577] [Medline: 28642212]
41. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. jair 2002 Jun 01;16:321-357. [doi: 10.1613/jair.953]
42. Chen T, Li L, Kochen MM. A systematic review: How to choose appropriate health-related quality of life (HRQOL) measures in routine general practice? J. Zhejiang Univ. Sci 2005;6B(9):936-940. [doi: 10.1631/jzus.2005.b0936]
43. Lins L, Carvalho FM. SF-36 total score as a single measure of health-related quality of life: Scoping review. SAGE Open Med 2016 Oct 04;4:2050312116671725 [FREE Full text] [doi: 10.1177/2050312116671725] [Medline: 27757230]
44. Caster O, Dietrich J, Kürzinger ML, 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 Dec;41(12):1355-1369 [FREE Full text] [doi: 10.1007/s40264-018-0699-2] [Medline: 30043385]
45. Colilla S, Tov EY, Zhang L, Kurzinger M, Tcherny-Lessenot S, Penfornis C, et al. Validation of New Signal Detection Methods for Web Query Log Data Compared to Signal Detection Algorithms Used With FAERS. Drug Saf 2017 May 2;40(5):399-408. [doi: 10.1007/s40264-017-0507-4] [Medline: 28155198]

46. Schwartz HA, Sap M, Kern M, Eichstaedt JC, Kapelner A, Agrawal M, et al. Predicting individual well-being through the language of social media. Pac Symp Biocomput 2016;21:516-527 [FREE Full text] [Medline: 26776214]
47. Kürzinger ML, 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 Nov 20;20(11):e10466 [FREE Full text] [doi: 10.2196/10466] [Medline: 30459145]
48. Parler de sa santé en ligne : une pratique loin d'être marginale et qui peut aider la recherche. Odoxa. URL: http://www.odoxa.fr/sondage/parler-de-sante-ligne-pratique-loin-detre-marginale-aider-recherche/ [accessed 2022-01-21]

Abbreviations

AUC: area under the curve
EQ-5D: Euro Quality of Life 5 Dimensions
FACT-G: Functional Assessment of Cancer Therapy—General
GBDT: gradient boosting decision tree
HRQoL: health-related quality of life
LASSO: least absolute shrinkage and selection operator
MedDRA: Medical Dictionary for Regulatory Activities
NER: named entity recognition
NLP: natural language processing
POS: part of speech
QLQ-C30: Quality of Life Questionnaire
QoL: quality of life
ROC: receiver operating characteristic
SF-36: Short Form Health Survey
SMOTE: synthetic minority oversampling technique
SVM-L: support vector machine light
SVM-R: support vector machine regression
TF-IDF: term frequency-inverse document frequency

Edited by R Kukafka; submitted 24.06.21; peer-reviewed by A Trifan, D Huang; comments to author 11.08.21; revised version received 05.10.21; accepted 29.10.21; published 28.01.22

Please cite as:
Renner S, Marty T, Khadhar M, Foulquié P, Voillot P, Mebarki A, Montagni I, Texier N, Schück S
A New Method to Extract Health-Related Quality of Life Data From Social Media Testimonies: Algorithm Development and Validation
J Med Internet Res 2022;24(1):e31528
URL: https://www.jmir.org/2022/1/e31528
doi: 10.2196/31528
PMID:

Annexe 2 : Le questionnaire Short Form 36 items

1.- En général, diriez-vous que votre santé est : (cocher ce que vous ressentez)

Excellente__ Très bonne__ Bonne__ Satisfaisante__ Mauvaise__

2.- Par comparaison avec il y a un an, que diriez-vous sur votre santé aujourd'hui *?*

Bien meilleure qu'il y a un an __ A peu près comme il y a un an __

Un peu meilleure qu'il y a un an __ Un peu moins bonne qu'il y a un an __

Pire qu'il y a un an __

3.- vous pourriez vous livrer aux activités suivantes le même jour. Est-ce que votre état de santé vous impose des limites dans ces activités ? Si oui, dans quelle mesure ? (entourez la flèche).

a.Activités intenses : courir, soulever des objets lourds, faire du sport.

Oui, très limité oui, plutôt limité pas limité du tout

b.Activités modérées :déplacer une table, passer l'aspirateur.

Oui, très limité oui, plutôt limité pas limité du tout

c.Soulever et transporter les achats d'alimentation.

Oui, très limité oui, plutôt limité pas limité du tout

d.Monter plusieurs étages à la suite.

Oui, très limité oui, plutôt limité pas limité du tout

e.Monter un seul étage.

Oui, très limité oui, plutôt limité pas limité du tout

f.Vous agenouiller, vous accroupir ou vous pencher très bas.

Oui, très limité oui, plutôt limité pas limité du tout

g.Marcher plus d'un kilomètre et demi.

Oui, très limité oui, plutôt limité pas limité du tout

h.Marcher plus de 500 mètres

Oui, très limité oui, plutôt limité pas limité du tout

i.Marcher seulement 100 mètres.

Oui, très limité oui, plutôt limité pas limité du tout

j.Prendre un bain, une douche ou vous habiller.

Oui, très limité oui, plutôt limité pas limité du tout

4.- Au cours des 4 dernières semaines, avez-vous eu l'une des difficultés suivantes au travail ou lors des activités courantes, du fait de votre santé ? (réponse : oui ou non à chaque ligne)

	oui	non
Limiter le temps passé au travail, ou à d'autres activités ?		
Faire moins de choses que vous ne l'espériez ?		
Trouver des limites au type de travail ou d'activités possibles ?		
Arriver à tout faire, mais au prix d'un effort		

5.- Au cours des 4 dernières semaines, avez-vous eu des difficultés suivantes au travail ou lors des activités courantes parce que vous étiez déprimé ou anxieux ? (réponse : oui ou non à chaque ligne).

	oui	non
Limiter le temps passé au travail, ou à d'autres activités ?		
Faire moins de choses que vous ne l'espériez ?		
Ces activités n'ont pas été accomplies aussi soigneusement que d'habitude ?		

6.- Au cours des 4 dernières semaines, dans quelle mesure est-ce que votre état physique ou mental ont perturbé vos relations avec la famille, les amis, les voisins ou d'autres groupes ?

Pas du tout très peu assez fortement énormément

7.- Avez-vous enduré des souffrances physiques au cours des 4 dernières semaines ?

Pas du tout très peu assez fortement énormément

8.- Au cours des 4 dernières semaines la douleur a-t-elle gêné votre travail ou vos activités usuelles ?

Pas du tout | un peu | modérément | assez fortement | énormément

9.- Ces 9 questions concernent ce qui s'est passé au cours de ces dernières 4 semaines. Pour chaque question, donnez la réponse qui se rapproche le plus de ce que vous avez ressenti. Comment vous sentiez-vous au cours de ces 4 semaines :

a. vous sentiez-vous très enthousiaste ?

Tout le temps | très souvent | parfois | peu souvent | jamais

b. étiez-vous très nerveux ?

Tout le temps | très souvent | parfois | peu souvent | jamais

c. étiez-vous si triste que rien ne pouvait vous égayer ?

Tout le temps | très souvent | parfois | peu souvent | jamais

d. vous sentiez-vous au calme, en paix ?

Tout le temps | très souvent | parfois | peu souvent | jamais

e. aviez-vous beaucoup d'énergie ?

Tout le temps | très souvent | parfois | peu souvent | jamais

f. étiez-vous triste et maussade ?

Tout le temps | très souvent | parfois | peu souvent | jamais

g. aviez-vous l'impression d'être épuisé(e) ?

Tout le temps | très souvent | parfois | peu souvent | jamais

h. étiez-vous quelqu'un d'heureux ?

Tout le temps | très souvent | parfois | peu souvent | jamais

i. vous êtes-vous senti fatigué(e) ?

Tout le temps | très souvent | parfois | peu souvent | jamais

10.- Au cours des 4 dernières semaines, votre état physique ou mental a-t-il gêné vos activités sociales comme des visites aux amis, à la famille, etc ?

Tout le temps | très souvent | parfois | peu souvent | jamais

11.- Ces affirmations sont-elles vraies ou fausses dans votre cas ?

a. il me semble que je tombe malade plus facilement que d'autres.

Tout à fait vrai | assez vrai | ne sais pas | plutôt faux | faux

b. ma santé est aussi bonne que celle des gens que je connais.

Tout à fait vrai | assez vrai | ne sais pas | plutôt faux | faux

c. je m'attends à ce que mon état de santé s'aggrave.

Tout à fait vrai | assez vrai | ne sais pas | plutôt faux | faux

d. mon état de santé est excellent.

Tout à fait vrai | assez vrai | ne sais pas | plutôt faux | faux

Wade JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). Medical Care 1992;30:473–483.

Annexe 2 : Le questionnaire Short Form 36 items

Mobilité

1. Je n'ai aucun problème pour me déplacer à pied.
2. J'ai des problèmes pour me déplacer à pied.
3. Je suis obligé(e) de rester alité(e).

Autonomie de la personne

1. Je n'ai aucun problème pour prendre soin de moi.
2. J'ai des problèmes pour me laver ou m'habiller tout(e) seul(e).
3. Je suis incapable de me laver ou de m'habiller tout(e) seul(e).

Activités courantes

1. Je n'ai aucun problème pour accomplir mes activités courantes (e.g. travail, études, travaux domestiques, activités familiales ou loisirs).
2. J'ai des problèmes pour accomplir mes activités courantes.
3. Je suis incapable d'accomplir mes activités courantes.

Douleurs/gêne

1. Je n'ai ni douleurs ni gêne.
2. J'ai des douleurs ou une gêne modérée(s).
3. J'ai des douleurs ou une gêne extrême(s).

Anxiété/Dépression

1. Je ne suis ni anxieux(se) ni déprimé(e).
2. Je suis modérément anxieux(se) ou déprimé(e).
3. Je suis extrêmement anxieux(se) ou déprimé(e).

Nous aimerions savoir dans quelle mesure votre santé est bonne ou mauvaise AUJOURD'HUI.

- Cette échelle est numérotée de 0 à 100.
- 100 correspond à la meilleure santé que vous puissiez imaginer.

0 correspond à la pire santé que vous puissiez imaginer.

- Veuillez faire une croix (X) sur l'échelle afin d'indiquer votre état de santé AUJOURD'HUI.
- Maintenant, veuillez noter dans la case ci-dessous le chiffre que vous avez coché sur l'échelle.

VOTRE SANTÉ AUJOURD'HUI =

https://euroqol.org/wp-content/uploads/2016/09/EQ-5D-5L_UserGuide_2015.pdf

Annexe 4 : Poster médical : Etude de l'usage du méthylphénidate sur les réseaux sociaux

ÉTUDE DE L'USAGE DU MÉTHYLPHÉNIDATE SUR LES RÉSEAUX SOCIAUX

Pierre Foulquié[1], Pamela Voillot[1], Carole Faviez[1],
Adel Mebarki[1], Xiaoyi Chen[2], Nathalie Texier[1], Stéphane Schück[1]

[1]Kap Code, Paris, France - [2]INSERM, UMRS1138, équipe 22, Centre de Recherche des Cordeliers, Paris, France

KAP CODE

INTRODUCTION

Le **méthylphénidate**, prescrit en France à plus de **49 000 patients** atteints de troubles du déficit de l'attention avec hyperactivité (TDAH) âgés de 6 à 17 ans, a fait l'objet de plusieurs évaluations par l'EMA et l'ANSM concernant son usage hors AMM.

Parallèlement, la FDA et l'EMA ont reconnu les réseaux sociaux comme nouvelle source de données pour renforcer la surveillance des médicaments [1][2].

Dans ce contexte, l'objectif de cette étude consistait en **l'exploration du comportement** des patients vis-à-vis du **méthylphénidate** à travers les réseaux sociaux.

MATÉRIEL & MÉTHODES

Les messages ayant trait au **méthylphénidate** ont été extraits à partir de cinq forums français généralistes via l'outil Detec't [3], développé par la société Kap Code.

L'analyse s'est déroulée en deux étapes :

1. Premièrement, un modèle de sujet (*topic model*) était appliqué pour identifier les thèmes de discussions (*topic*) relatifs au **méthylphénidate**. Ces derniers permettaient de classer les messages selon les thématiques qu'ils contenaient.

2. Les messages associés aux thématiques qui se rapportaient aux effets secondaires, aux mésusages et aux abus du **méthylphénidate** faisaient ensuite l'objet d'une classification hiérarchique descendante (CHD), directement appliquée aux mots. Différentes sous-thématiques étaient ainsi obtenues.

RÉSULTATS

Les corpus d'analyse méthylphénidate était constitué de **3 343 messages** publiés **entre 2007 et 2016**.

L'évolution temporelle du volume de messages postés (fig.1) n'a pas permis en général pas de mettre en évidence de pics aux dates clés associées au profil de sécurité du **méthylphénidate**. Ces dates consistaient en des décisions à l'échelle française (2006, 2011, 2012 et 2013) et européenne (2007 et 2009). Seule la décision de la HAS (2012) semble avoir eu une influence sur le volume de messages observé.

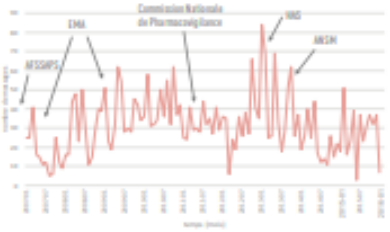


Figure 1 – Évolution temporelle des messages associés au méthylphénidate

La création d'un **nuage de mots** a permis d'observer les termes les plus présents dans le **corpus**.

Plusieurs champs lexicaux tels que ceux de la TDAH, de la prescription et de la prise, de l'enfance ainsi que des inquiétudes concernant les **effets indésirables** ont pu être identifiées.

Le nuage de mots (fig.2) indique une plus grande fréquence d'utilisation du méthylphénidate chez les enfants, en accord avec l'indication du médicament.

Figure 2 – Nuage de mots des messages méthylphénidate

Par la suite, l'application d'un topic model a permis de mettre en avant 14 thématiques de discussions principales (fig. 3).

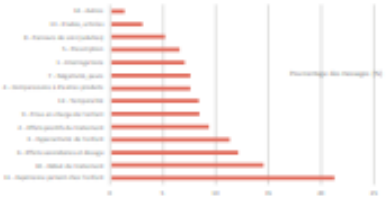


Figure 3 – Thématiques du topic model

Parmi les topics principaux, un topic était centré sur les « **effets secondaires** » (**12,15%** des messages), deux sur le mésusage (intitulés « **Comparaison aux autres produits** » et « **Négativité et peurs** », **7,6%** des messages chacun) et un sur l'usage chez les adultes (**5,2%** des messages). Les autres concernaient des usages cohérents dans le cadre de l'AMM.

La CHD appliquée au topic d'« **effets secondaires** » (fig.4) a mis en évidence plusieurs troubles liés à la prise du traitement tels que la perte de poids ou encore les vomissements.

Figure 4 – Thématiques de la CDH du topic « effets secondaires »

Concernant les topics du mésusage (fig. 5), la méthode a permis d'observer l'utilisation de la molécule par les adultes ou les étudiants pour un usage récréatif ou pour augmenter leurs capacités.

De plus, les patients comparaient les effets de la molécule à certaines drogues telles que les amphétamines. Enfin, le topic sur l'usage par les adultes a mis en avant des erreurs ou des retards de diagnostic de TDAH.

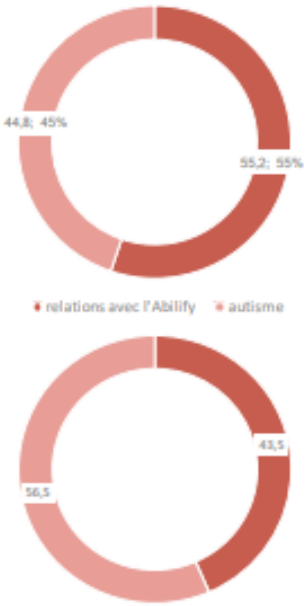


Figure 5 – Thématiques de la CDH des deux topics « mésusage »

CONCLUSION

Cette étude a permis d'identifier :

- Des **effets indésirables**, ce thème constituant un thème d'importance au sein des discussions ;
- Des pratiques non identifiables via les systèmes de soins classiques :
  - l'usage du **méthylphénidate** par des populations spécifiques telles que les adultes et les étudiants.
  - des utilisations hors AMM : comme **substitution aux drogues** et en tant que **psychostimulant**.

La surveillance des réseaux sociaux apparait comme un outil d'alerte complémentaire dans la surveillance des médicaments.

RÉFÉRENCES

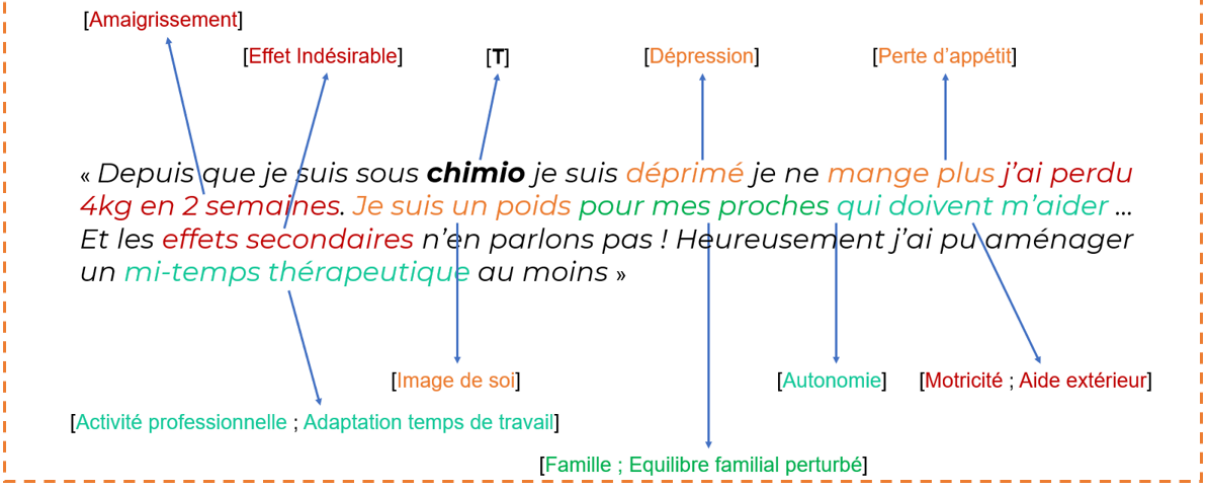
[1] www.fda.gov/ScienceResearch/SpecialTopics/RegulatoryScience/ucm452304

[2] www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000258

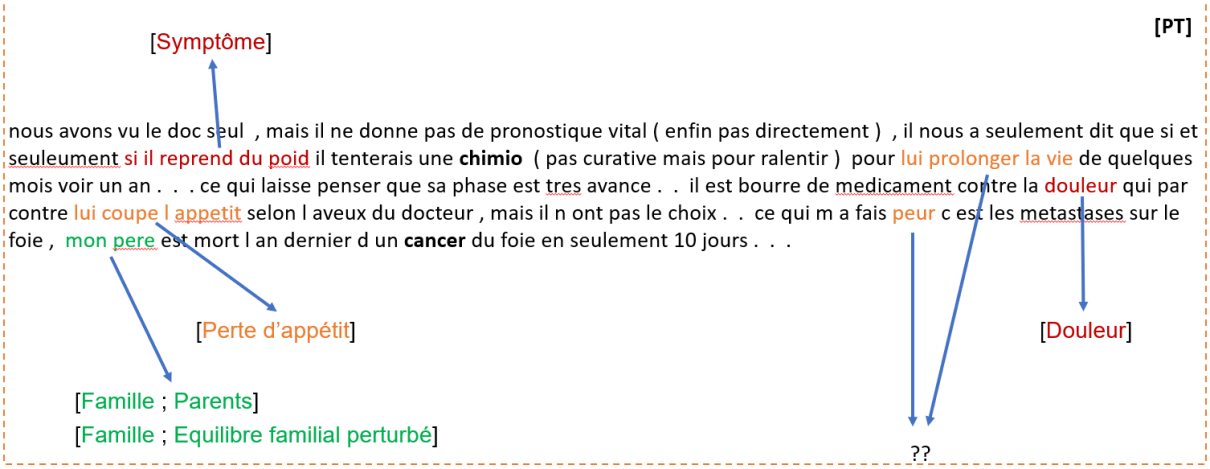
[3] Abdellaoui R et al. Filtering Entities to Optimize Identification of Adverse Drug Reaction From Social Media: How Can the Number of Words Between Entities in the Messages Help?. JMIR public health and surveillance, 2017, vol. 3, no 2.

Annexe 5 : exemple de messages annotés, issue de la formation annotation impact de qualité de vie

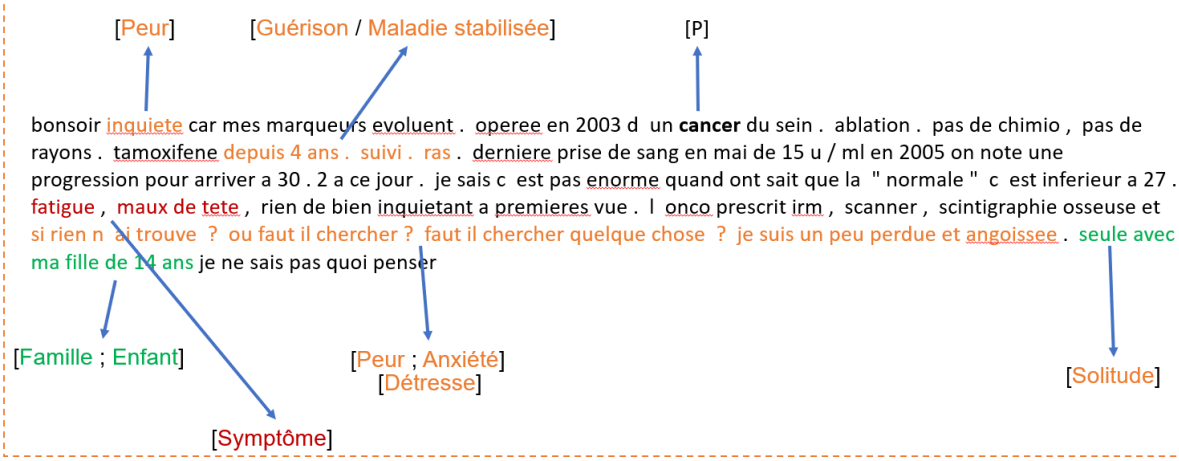
« *Depuis que je suis sous **chimio** je suis déprimé je ne mange plus j'ai perdu 4kg en 2 semaines. Je suis un poids pour mes proches qui doivent m'aider … Et les effets secondaires n'en parlons pas ! Heureusement j'ai pu aménager un mi-temps thérapeutique au moins* »



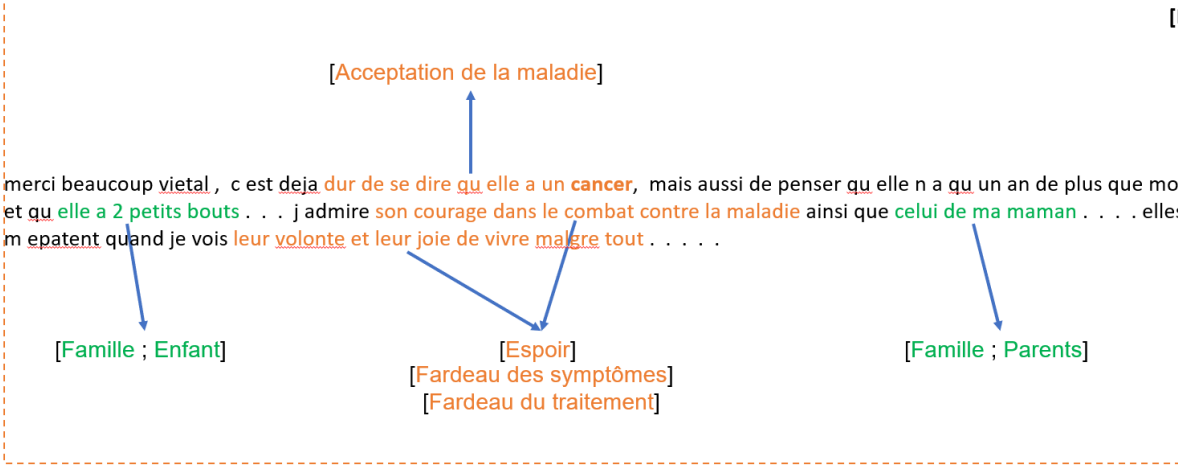
nous avons vu le doc seul , mais il ne donne pas de pronostique vital (enfin pas directement) , il nous a seulement dit que si et seuleument si il reprend du poid il tenterais une chimio (pas curative mais pour ralentir) pour lui prolonger la vie de quelques mois voir un an . . . ce qui laisse penser que sa phase est tres avance . . il est bourre de medicament contre la douleur qui par contre lui coupe l appetit selon l aveux du docteur , mais il n ont pas le choix . . ce qui m a fais peur c est les metastases sur le foie , mon pere est mort l an dernier d un cancer du foie en seulement 10 jours . . .



bonsoir inquiete car mes marqueurs evoluent . operee en 2003 d un cancer du sein . ablation . pas de chimio , pas de rayons . tamoxifene depuis 4 ans . suivi . ras . derniere prise de sang en mai de 15 u / ml en 2005 on note une progression pour arriver a 30 . 2 a ce jour . je sais c est pas enorme quand ont sait que la " normale " c est inferieur a 27 . fatigue , maux de tete , rien de bien inquietant a premieres vue . l onco prescrit irm , scanner , scintigraphie osseuse et si rien n ai trouve ? ou faut il chercher ? faut il chercher quelque chose ? je suis un peu perdue et angoissee . seule avec ma fille de 14 ans je ne sais pas quoi penser

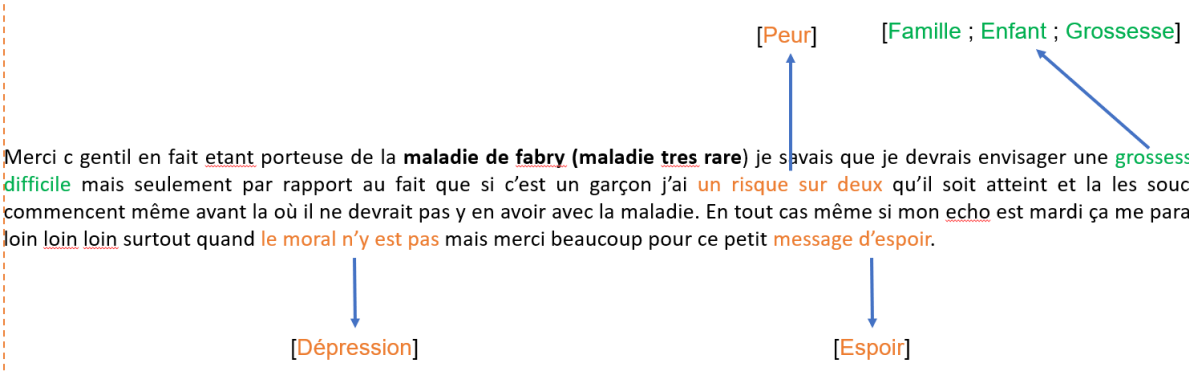


merci beaucoup vietal , c est deja dur de se dire qu elle a un cancer, mais aussi de penser qu elle n a qu un an de plus que moi et qu elle a 2 petits bouts . . . j admire son courage dans le combat contre la maladie ainsi que celui de ma maman . . . . elles m epatent quand je vois leur volonte et leur joie de vivre malgre tout . . . . .

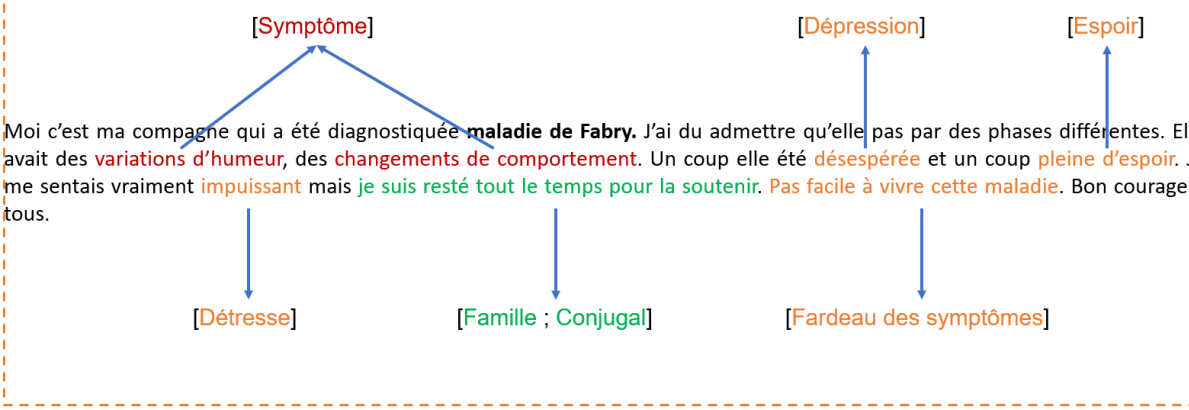


Merci c gentil en fait etant porteuse de la maladie de fabry (maladie tres rare) je savais que je devrais envisager une grossesse difficile mais seulement par rapport au fait que si c'est un garçon j'ai un risque sur deux qu'il soit atteind et la les soucis commencent mpeme avant la où il ne devrait pas y en avoir avec la maladie. En tout cas même si mon echo est mardi ça me parait loin loin loin surtout quand le moral n'y est pas mais merci beaucoup pour ce petit message d'espoir.

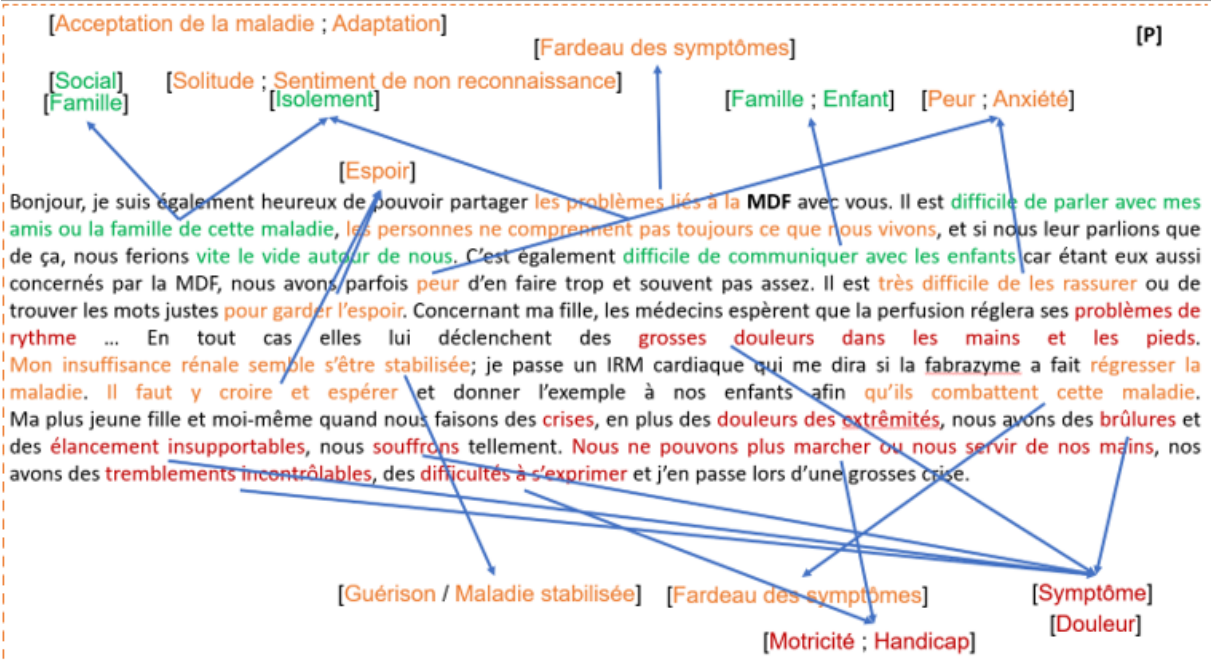
[P]



Moi c'est ma compagne qui a été diagnostiquée maladie de Fabry. J'ai du admettre qu'elle pas par des phases différentes. Elle avait des variations d'humeur, des changements de comportement. Un coup elle été désespérée et un coup pleine d'espoir. Je me sentais vraiment impuissant mais je suis resté tout le temps pour la soutenir. Pas facile à vivre cette maladie. Bon courage à tous.



Bonjour, je suis également heureux de pouvoir partager les problèmes liés à la MDF avec vous. Il est difficile de parler avec mes amis ou la famille de cette maladie, les personnes ne comprennent pas toujours ce que nous vivons, et si nous leur parlions que de ça, nous ferions vite le vide autour de nous. C'est également difficile de communiquer avec les enfants car étant eux aussi concernés par la MDF, nous avons parfois peur d'en faire trop et souvent pas assez. Il est très difficile de les rassurer ou de trouver les mots justes pour garder l'espoir. Concernant ma fille, les médecins espèrent que la perfusion réglera ses problèmes de rythme … En tout cas elles lui déclenchent des grosses douleurs dans les mains et les pieds. Mon insuffisance rénale semble s'être stabilisée; je passe un IRM cardiaque qui me dira si la fabrazyme a fait régresser la maladie. Il faut y croire et espérer et donner l'exemple à nos enfants afin qu'ils combattent cette maladie. Ma plus jeune fille et moi-même quand nous faisons des crises, en plus des douleurs des extrémités, nous avons des brûlures et des élancement insupportables, nous souffrons tellement. Nous ne pouvons plus marcher ou nous servir de nos mains, nos avons des tremblements incontrôlables, des difficultés à s'exprimer et j'en passe lors d'une grosses crise.



Annexe 6 : Résumé des variables d'importances impliquées dans l'optimisation de chaque modèle algorthmique, propre à chaque dimension

Algo - Impact

29 variables crées et sélectionnées par le SequentialFeatureSelector
Inclut toutes les variables lexicales

1400 messages. 818 avec impact, 581 sans.
Split Train 70 / 30 Test

Cet algo sert de filtre, s'il retourne 1 pour un message, on applique les autres algorithmes.

Paramètres Sequential Feature Selector :

Xgboost100 arbres
AIC Stepwise(forward)
Accuracychoisie pour scoring
5 folds

Meilleurs paramètres GridSearchCV 5-folds:

Learning_rate: 0,1
Max_depth: 8
Min_child_weight 3
Min_split_loss:2
N_estimators:100
Reg_alpha: 1
Epochs: 35

Variables :

'disgust',
'first_person_plur',
'first_person_sing',
'joy',
'n_ne_ar',
'n_other_drugs',
'n_pas',
'n_verbs_aux_ar',
'n_verbs_inf',
'n_verbs_inf_ar',
'n_verbs_past_ar',
'n_verbs_past_participle',
'n_verbs_past_participle_ar',
'n_verbs_present',
'n_verbs_present_2nd_ar',
'n_verbs_present_3rd',
'n_verbs_present_3rd_ar',
'n_verbs_present_ar',
'pronouns',
'sadness',
'second_person_plur',
'third_person',
'third_person_sing',

Variables CL:
'count_psy',
'count_pro',
'count_rel',
'count_fin'

Kap•Code

15

Algo - Psychique

28 variables crées et sélectionnées par le SequentialFeatureSelector
Inclut variables lexicales psychique, activité professionnelle et scolaire

1400 messages, 519 avec impact, 880 sans.
Split Train 70 / 30 Test

472 Expressions d'impact psychique relevées

Paramètres Sequential Feature Selector :

Xgboost1000 arbres
AIC Stepwise(forward)
Accuracychoisie pour scoring
5 folds
scoring=rocaucscore (courbe sensibilité/spécificité)

Meilleurs paramètres GridSearchCV 5-folds:

Learning_rate: 0,1
Max_depth: 9
Min_child_weight 4
Min_split_loss:2
N_estimators:100
Reg_alpha: 1
Epochs: 470

Variables :

'anger',
'fear',
'first_person_plur',
'joy',
'n_ne_ar',
'n_other_drugs',
'n_pas',
'n_verbs_aux',
'n_verbs_inf_ar',
'n_verbs_past',
'n_verbs_past_participle_ar',
'n_verbs_present',
'n_verbs_present_2nd',
'n_verbs_present_2nd_ar',
'n_verbs_present_3rd_ar',
'negative',
'positive',
'pronouns',
'second_person',
'second_person_plur',
'second_person_sing',
'third_person_sing',

Variables CL:
'count_psy',
'count_act',
'count_pro',
'count_sco',
'count_rel',
'count_fin'

Kap•Code

19

Algo - Activité

40 variables crées et sélectionnées par le SequentialFeatureSelector
Inclut variables lexicales relationnelle, activité professionnelle et scolaire

1400 messages, 363 avec impact, 1036 sans.
Split Train 70 / 30 Test

Oversampling smote appliqués sur bases de Train et Test pour équilibrer les classes avant la sélection de variables

205 Expressions d'impact sur l'activité relevées.
64 sur l'activité professionnelle
48 sur l'activité scolaire

Paramètres Sequential Feature Selector :

Xgboost1000 arbres
AIC Stepwise(forward)
Accuracy choisie pour Scoring
5 folds
Scoring $F1\ Score = 2 * \frac{Precision\ *Recall}{Precision\ +Recall}$

Meilleurs paramètres GridSearch CV 5-folds:

Learning_rate: 0,001
Max_depth: 8
Min_child_weight 3
Min_split_loss:2
N_estimators:1000
Reg_alpha: 20
Epochs: 453
Scoring F1 score

Variables :

'anger',
'disgust',
'fear',
'first_person_plur',
'first_person_sing',
'joy',
'n_ne',
'n_ne_ar',
'n_other_drugs',
'n_pas',
'n_pas_ar',
'n_verbs_aux',
'n_verbs_inf',
'n_verbs_inf_ar',
'n_verbs_past',
'n_verbs_past_ar',
'n_verbs_past_participle',
'n_verbs_past_participle_ar',
'n_verbs_present',
'n_verbs_present_2nd',
'n_verbs_present_2nd_ar',
'n_verbs_present_3rd',
'n_verbs_present_3rd_ar',
'n_verbs_present_ar',
'negative',
'positive',
'sadness',
'second_person',
'second_person_plur

Variables CL:
' count_phy',
'count_psy',
'count_act',
'count_pro',
'count_sco',
'count_rel',
'count_fin',

Kap•Code

21

Algo – Relationnel

13 variables crées et sélectionnées par le SequentialFeatureSelector
Inclut toutes les variables lexicales sauf physique

1400 messages, 193 avec impact, 1206 sans.
Split Train 70 / 30 Test

Oversampling smote appliqués sur bases de Train et Test pour équilibrer les classes avant la sélection de variables

57 Expressions d'impact relationnel relevées

Paramètres Sequential Feature Selector :

Xgboost1000 arbres
AIC Stepwise(forward)
Accuracy choisie pour Scoring
5 folds
Scoring $F1\ Score = 2 * \frac{Precision\ *Recall}{Precision\ +Recall}$

Meilleurs Paramètres GridSearchCV 5-folds:

Learning_rate: 0,1
Max_depth: 20
Min_child_weight 3
Min_split_loss:2
N_estimators:100
Reg_alpha: 50
Epochs: 17
Scoring F1 score

Variables :

'disgust',
'n_other_drugs',
'n_pas',
'n_verbs_aux',
'n_verbs_past_ar',
'n_verbs_past_participle',
'n_verbs_present',
'n_verbs_present_2nd_ar',
'negative',
'pronoun',
'third_person',

Variables CL:
'count_phy',
'count_psy',

Kap•Code

23

Algo – Financier

35 variables crées et sélectionnées par le SequentialFeatureSelector
Inclut variables lexicales sur l'activité scolaire, l'impact relationnel et financier

1400 messages, 69 avec impact, 1330 sans. Split Train 70 / 30 Test

Oversampling Smote appliqués sur bases de Train et Test pour équilibrer les classes avant la sélection de variables
Régression Lasso pour réduire le sur apprentissage (overfit)

71 Expressions d'impact financier relevées

Paramètres Sequential Feature Selector :

Xgboost 1000 arbres
AIC Stepwise (forward)
Accuracy choisie pour Scoring
5 folds
Scoring: Roc auc score

Paramètres finaux Algo GridSearch CV:

Learning_rate: 0,1
Max_depth: 8
Min_child_weight 4
Min_split_loss: 2
N_estimators: 100
Scoring Roc auc score
Reg_alpha 10
Epochs: 300

Variables :
'anger',
'disgust',
'fear',
'first_person_plur',
'first_person_sing',
'n_ne',
'n_other_drugs',
'n_pas',
'n_pas_ar',
'n_verbs_aux',
'n_verbs_aux_ar',
'n_verbs_inf',
'n_verbs_inf_ar',
'n_verbs_past',
'n_verbs_past_ar',
'n_verbs_past_participle',
'n_verbs_past_participle_ar',
'n_verbs_present',
'n_verbs_present_2nd',
'n_verbs_present_2nd_ar',
'n_verbs_present_3rd',
'n_verbs_present_3rd_ar',
'n_verbs_present_ar',
'positive',
'pronouns',
'second_person_sing',
'third_person',
'third_person_sing

Variables CL:
'count_phy'
'count_psy'
'count_act'
'count_pro'
'count_sco',
'count_rel,
'count_fin'

Kap Code

Annexe : Exemples de messages relatifs à l'analyse d'impact de qualité de vie dans les cinq dermatoses

« Bon bah je viens de prendre rdv avec mon médecin pour la semaine pro **Mon eczéma s'est étendue** (c'était pas arrivé depuis 2 ans) J'espère qu'on va pouvoir trouver une solution parce que **marcher et courir** alors qu'on a **littéralement les pieds à vif** je recommande pas du tout 😭 »

« Bonsoir, J'ai déjà poster un sujet similaire mais voila le topo : Je suis avec ma conjointe depuis 5 ans maintenant. Celle-ci a une **maladie chronique, qui lui cause des problèmes de peau (type eczéma)** sur tout le corps. Ce qui n'est **pas du tout facile à vivre au quotidien**. Ce **problème est si grave qu'il l'empêche d'avoir une vie "normale".** Du coup, beaucoup d'aspect de sa vie sont atteints : » **sa confiance en soi, sa motivation en générale, et sa libido.** »

« […] **je me sens seul et insatisfait**. Surtout que nous faisons **chambre à part** avec ma conjointe, car c'est plus pratique pour elle (**son eczéma** la démange régulièrement), et moi ça me permet de ne pas avoir d'envies. D'un côté **je me sens moins aimé**, même si elle me dit le contraire. […] Depuis quelques mois, je pense a certaines choses qui me font du mal : - **La quitter? […] Attendre?** Cela est en train de me rendre fou je ne pense plus qu'a ça **je n'arrive pas a vivre ma relation pleinement** […] »

« J'ai jamais réussi à lui faire boire le allernova. La je suis au althera de nestle mais il ne le prend que sous forme de bouillie. **Son eczema a diminue mais est tjrs présent**. Ils passe son temps a **se gratter a sang**. Il est **tres difficile a faire manger et dormir**. Je suis épuisée et je dois encore gerer les 2 grands. J'en pleure tt les soirs. »

« **Vraie nuit de sommeil cette nuit, diminution de l'eczéma**. Il semblerait que j'ai compris (et régler) un des messages que mon corps voulait me faire passer.. »

« Mon fils prend modilac riz AR apres echec de l allernova depuis plus de 3 semaines et ca a était un miracle pour **lui qui souffrait d atroce colique, rgo et surtout eczéma et boutons d allergies +++** En quelques jours **sont eczéma** avait disparu et en 3 jours fini les coliques et rgo qui a bien diminué Je le commande sur 1001 pharmacie car dans ma pharmacie **une boîte coûte 32,80 €** alors que sur ce site frais de port compris **ça me revient à 20€** »

« Et j'ai pris du diprosone et ce truc t'enlève l'eczéma rapidement comme il te le ramène aussi sec et bien pire !! Le mieux que j'ai eu c'est un médecin qui m'a prescrit une **préparation, un peu cher je sais mais efficace**.. »

«mon petit garçon de 9 mois **fait de l eczéma** depuis ses 2 mois... On **a tt essayé** : pédiatre allergologue, dermatologue, micro kiné... On a aussi **essayé ts types de crèmes** : Avene, la roche posay, aderma,... On a fait **installé un adouciceur** d eau, je fais les lessives moi mêmes... Et pourtant les crises de démangeaisons s intensifient. »

« J'ai visiblement d'après ma dermato de la "**rosacée**", après un traitement de 15 jours sous "rosex" c'est encore pire **la plaque est encore plus rouge et avec des sensations de brulures**... La plaque

se situe sur des taches brunes au niveau du haut de la joue sous l'oeil et c'est ce qui m'inquiétait à l'origine (peur du cancer de la peau...) mais d'après la dermato aucun signe dangereux... connaissait vous un moyen pour lutter contre cette affection?»

« Avoir de la rosacée vasculaire localiser sur le visage **ça te bouffe la vie** de 1, mais en plus de ça tu devrais éviter à fond le soleil pas manger épicer pas boire d'alcool ect **laissez moi vivre non** ?? »

« Bon je me suis lancée dans des compléments alimentaires pour ma peau car **ma rosacée je ne la supporte plus** J'espère que ça va fonctionner »

«je suis atteint de **"rosacée" aux stades des "flush"**, (c'est à dire des "crises" de rougeur sur mes joues notamment). Mon dermato m'a dit que c'était + ou - incurable, il m'a filé une crème qui ne fonctionne pas, et m'a dit que seul le laser pouvait "améliorer" les choses. En parallèle, **je ne supporte pas la chaleur, j'ai des picotements un peu partout une fois que la température** (et en fait ma température à moi) augmente. En plus à cause de la rosacée, mon visage devient très chaud et c'est insupportable . Et c'est vraiment handicapant (**j'évite de sortir le plus possible**). Mon médecin dit que ce n'est rien, et n'a pas de solution. **Comment faire pour me sentir mieux par forte chaleur et au soleil ?** »

« Je te fais pars de l'évolution de ma **rosacée** […] je deviens tout rouge une fois à l'exterieur sinon BEAUCOUP de boutons sont apparus ces 4 derniers jours au niveau de mes joues des gros papules quoi, le genre de buton qui fait mal au touché **c vraiment horrible moralement** parlant j imagine que tu sais de quoi je parle **je minimise mes sorties le maximum possible pour éviter le regard des autres** mais jusqu'à quand »

« Bonjour, voilà c'est la première fois que j'ouvre un forum pour expliquer mon problème qui me pourrit la vie.. j'ai à peine 19 ans et depuis 5-6 ans j'ai des problèmes de peau au visage (**erythro-couperose**), ça se traduit par des grosses plaques rouges au niveau des joues (rouge vin) et un peu au niveau du nez. J'ai fait **5 séances de laser**, c'est parti je dirais à 10-15% pas plus (**165€ la séance, cher pour le résultat obtenu**), pas de suivi ni rien, je fais ma séance et je pars. Et donc j'ai pris rdv dans un autre centre de laser où là il y a un suivi du dermato etc..»

« 30 ans que je vis **avec le vitiligo. Je ne l'ai jamais accepté,** je fais avec 😩 en 2018 de gros ennuis de santé, 5 opérations en 4 mois et grosse poussée de vitiligo. Depuis mes mains, mes pieds sont complètement de pigmentés et de plus en plus de zones sont atteintes. **Pas toujours facile à vivre**. Courage à vous tous»

« Je sois atteinte de **vitiligo sur tous le corps** et donc **je brule au moindre soleil**. Je suis très angoissée par rapport à la radiothérapie et aux brulures que cela va engendrer. Pouvez vous me donner les coordonnées d'un barreur de feu près de chez moi s'il vous plait ? »

« bonjour, j'ai 17 ans et **sa fait 7 ans que j'ai le vitiligo**. je n'en peu plu. C'est trop dur. Je n'ai personne a qui me confier et j'aimerai connaitre des personnes atteinte aussi de cette maladie. **Je**

ne sort plus je n'arrive plus à faire face au regard exterieur et plus précisément aux garçons ! »

« Atteinte du vitiligo depuis l'âge de 13 ans, j'en ai 35 aujourd'hui et c'est pas facile tous les jours, le regard des autres. En vacances mes filles m'ont vu **pleurer parce que je ne supportait plus les gens qui bloquaient sur moi** à la piscine du camping. comme si j'étais un extraterrestre. »

« Bonjour ou bonsoir, je viens écrire ce poste pour vous dire que j'en ai marre. Ca fait maintenant 6 ans que **je lutte avec détermination contre mon acné**. Depuis 6 ans je jongle entre les produits naturelle (dentifrice, aloe vera, citron, masque, argile, ail ...), changer d'alimentation (arrêter la mal-bouffe, les sauces ...), les médicaments (curacné pendant un peu plus d'un an, toléxine donc les antibiotiques et le Zinc) et les crêmes pour le matin et soir. Et rien ne fonctionne. **J'ai réellement envie de mourir, de me couper la peau ou même de me mettre de l'acide sur le visage. Déjà il y a 6 ans ca m'avait atteint psychologiquement mais alors là c'est pire...** [...] »

« **Je souffre d'exactement la même acné**, plus ou moins. J'ai eu un peu d'acné étant au collège mais c'était rien de très alarmant, puis j'ai eu une peau pareil vraiment parfaite jusqu'à mes 19/20 ans où j'ai commencé à en avoir de plus en plus sous la même forme que la tienne c'est à dire **Kystes et c'était hyper douloureux**, d'autant plus que j'en avais également dans le dos (hormonal). J'ai été voir un dermato qui m'a prescrit l'exact même traitement au zinc.... Ca a marché un peu, je n'ai plus rien eu du tout jusqu'à septembre/octobre 2018 et maintenant c'est la catastrophe et je ne sais plus quoi faire, je pense même avoir quelques **cicatrices** et malheureusement j'ai beau penser que ça se calme, ça revient par pics.... et mentalement c'est difficile, tout le monde y va de ses hypothèses et avis aussi, les gens regardent mal dans la rue »

« Je suis tombé dans tous les pièges possibles à l'adolescence, tous les comportements jugés normaux par la société voir encouragés, qui sont en réalité très malsains et/ou déviants : porno, malbouffe, passivité, etc. Résultat : 0 confiance en moi, **acné lourde**, **antisocial** (alors que j'étais très sociable enfant). Tout ça j'ai mis plus de 10 ans à le comprendre et à me régénérer et c'est pas fini, sachant que tout ça aurait pu être évité tout simplement avec un père pour me guider.»

« Moi **je trouve ça juste ABUSÉ que l'acné n'est pas considérée comme une véritable maladie**. Je trouve ca injuste que **les pillule contre l'acné ne sois pas remboursé (je paye ma pillule 40€ tout les deux mois)** sachant que d'autre ont des opérations de chirurgie plastique GRATUITE, et je commence a en avoir clairement marre, que les gens en général se foutent de la gueule des gens ayant de l'acné alors que c'est honteux de rire de gens ayant un cancer ou autre? Alors pourquoi vous moquez vous de ça, c'est une maladie on ne le choisi pas. Enfin bref désolé pour ce pavé mais je comprend tout a fait ce que tu traverse, j'ai 20 ans et tjr de l'acné comme j'en avait a 14 ans, et j'en souffre également! »

« J'ai les **paupieres entièrement desséchées (psoriasis**) et ça commence à être **vraiment très douloureux**. C'est quoi vos trucs / soins / astuces de grand-mère les plus hydratants ? Merci d'avance 🙏 #help #helpme »

« Bonjour , je m'appelle [nom] j'ai 22ans en couple est maman d'une petite fille de 3ans **je suis atteinte d'un psoriasis generalise** , j'en est de la tete aux pied apres plusieurs rendez vous avec ma dermatologue , elle m'a dis que la seul choses qui pouvait m'aider c'etait d'aller voir un psychologue cars je suis une personne angoissee stresse est j'ai vecu des choses assez dur a vivre . Je suis sous locatop creme hydratante , **c'est tres complique a vivre je ne sais plus quoi faire , je me sens pommer..** »

« bonjour, excusez moi d'entrer dans votre conversation, mais moi aussi, comme homme, je suis touché par le psoriasis là, ce n'est pas très beau car cela donne de micros sillons sur la verge, pas sur le gland, mais je me demande, si, pour faire l'amour ça ne ferait pas mal ? et si, malgré qu'on me disent que ce n'est pas contagieux, il n y aurait pas des risque pour ma partenaire, je serais aussi sensé soigner ça a la diprozone, seulement, la, moi, je ne veux pas y mettre de corticoide, la peau est beaucoup plus douce que sur le reste du corps, je ne me sens pas de le faire, ça a commence , après la première fois que j'ai été rase en hospitalisation et franchement, j ai été long à le montrer à ma généraliste car pudique et surtout j'avais peur que ce soit un cancer. **je ne fais pas l'amour, donc pour le moment et depuis 1989, ça ne risque rien pour personne**. »

« y a pas que les maladies rares ou les gens doivent subir les regards désobligeantes. **J'ai du psoriasis et beaucoup de coiffeurs ne veulent pas toucher mes cheveux** de peur que ce soit contagieux . »

SERMENT DE GALIEN

Je jure, en présence des maîtres de la Faculté, des conseillers de l'Ordre des pharmaciens et de mes condisciples :

D'honorer ceux qui m'ont instruit dans les préceptes de mon art et de leur témoigner ma reconnaissance en restant fidèle à leur enseignement.

D'exercer, dans l'intérêt de la santé publique, ma profession avec conscience et de respecter non seulement la législation en vigueur, mais aussi les règles de l'honneur, de la probité et du désintéressement.

De ne jamais oublier ma responsabilité et mes devoirs envers le malade et sa dignité humaine ; en aucun cas, je ne consentirai à utiliser mes connaissances et mon état pour corrompre les mœurs et favoriser des actes criminels.

Que les hommes m'accordent leur estime si je suis fidèle à mes promesses.

Que je sois couvert d'opprobre et méprisé de mes confrères si j'y manque.

VU, LE PRESIDENT DU JURY

CAEN, LE

VU, LE DIRECTEUR DE LA FACULTE

DES SCIENCES PHARMACEUTIQUES

CAEN, LE

L'université n'entend donner aucune approbation ni improbation aux opinions émises dans les thèses et mémoires. Ces opinions doivent être considérées comme propres à leurs auteurs.

Année de soutenance
2022

TITRE

CREATION D'UNE INTELLIGENCE ARTIFICIELLE CAPABLE DE DETECTER LES IMPACTS DE QUALITE DE VIE LIEE A LA SANTE A PARTIR DE DONNEES DE VIE REELLE.

Résumé

L'objet de ce travail était de réaliser un algorithme d'Intelligence artificielle, ou plus précisément, un modèle de traitement automatisé du langage, capable d'identifier dans des témoignages libres de patients, les verbatims reflétant un impact de qualité de vie. Il existe une synergie entre le patient centrisme et l'essor des données de vie réelle. Ensemble de données médicalement contextualisées, elles reflètent et objectivent la réalité patiente, par définition non captée lors essais cliniques. Sur les réseaux sociaux et forums médicaux, les patients forment des communautés en ligne dans un but de soutien, d'information et de partage d'expériences médicales. Ces commentaires publics sont récupérables informatiquement, dans le respect des réglementations pour la protection des données. C'est là que le traitement automatisé du langage entre en jeu. En étant capable d'analyser et de traduire le langage patient en ontologie médicale, il devient possible de gagner en information sur la réalité patiente, telle que décrite directement et sans filtre par eux-mêmes. Ainsi, notre algorithme répond à l'enjeu de compréhension fine de ce qui peut réellement impacter la qualité de vie des patients, en vie réelle.

TITLE

CREATING AN ARTIFICIAL INTELLIGENCE CAPABLE OF DETECTING HEALTH-RELATED QUALITY OF LIFE IMPACTS FROM REAL-LIFE DATA.

Summary

The purpose of this work was to develop an artificial intelligence algorithm, or more precisely, an automated language processing model, capable of identifying in free patient testimonies, the verbatims reflecting an impact on quality of life. There is a synergy between patient centrism and the rise of real-life data. Medically contextualized data sets reflect and objectify patient reality, which by definition is not captured in clinical trials. On social networks and medical forums, patients form online communities for support, information and sharing of medical experiences. These public comments can be retrieved electronically, in compliance with data protection regulations. This is where automated language processing comes in. By being able to analyze and translate patient language into medical ontology, it becomes possible to gain information about patient reality, as described directly and unfiltered by themselves. Thus, our algorithm meets the challenge of fine understanding of what can really impact the quality of life of patients, in real life.

Nom et Prénom de l'auteur : Tom Marty

Mots-clés

Intelligence artificielle, Qualité de vie, Données de vie réelle, Machine Learning, Infodémiologie, Réseaux sociaux, Patient centrisme

Année de soutenance
2022

Nom et Prénom de l’auteur : Tom Marty

Année de soutenance
2022

Nom et Prénom de l’auteur : Tom Marty