# Fouille des documents contenus dans les entrepôts de données de santé pour identifier les facteurs de risques et prévenir les maladies allergiques respiratoires.

## 1/ Description scientifique

**Contexte** : Depuis l'informatisation de leur système d'information ces deux dernières décennies, les hôpitaux hébergent désormais de nombreux entrepôts de connaissances, partiellement exploités, notamment des documents textuels faiblement structurés, décrivant les cas passés, aux contenus et formats hétérogènes. Convertir cette masse d'informations sous une forme structurée est un enjeu majeur et constitue le point de départ du développement d'outils d'interrogation automatique adaptés.

Dans ce travail, nous nous intéresserons particulièrement au domaine des allergies et plus particulièrement à l'exploitation des comptes rendus de consultation, d'hospitalisation et les observations médicales qui contiennent des informations précieuses sur le suivi médical des patients. Les allergies représentent aujourd'hui la quatrième maladie chronique dans le monde. L'OMS prévoit qu'en 2050, 1 personne sur 2 dans le monde souffrira d'allergies avec une augmentation constante, notamment dans les pays industrialisés. En France, le nombre de personnes souffrant d'une allergie a doublé en 20 ans, et notamment chez les enfants et les adolescents. 8 % des enfants et 4 % de la population adulte souffrent d'une allergie alimentaire.

Ce projet permettra, grâce aux avancées récentes de l'intelligence artificielle et du Traitement Automatique de la Langue Naturelle, de convertir de grandes masses de données textuelles pour en extraire et déduire de nouvelles connaissances médicales, socio-économiques, environnementales dans l'objectif d'adapter la prise en charge des patients, de développer la prévention, la recherche, et d'améliorer le fonctionnement du système de santé. La question centrale consistera à proposer une approche d'annotation et de classification efficace et impliquant le moins possibles les professionnels de santé pour constituer des jeux de données annotés.

Ce projet aura un fort pouvoir structurant sur le site de Montpellier en impliquant deux équipes du LIRMM (Fado pour les annotations et ADVANSE pour la classification), trois équipes du Département d'Information Médicale (DIM) du CHU de Montpellier (l'unité Sciences des Données de Santé pour l'accès aux documents textuels et aux données cliniques, l'unité d'Analyse Médico-économique des Processus de Soins pour l'expertise sur les parcours, l'unité de Recherche Clinique, Biostatistiques & Epidémiologie pour la validation) ainsi que les cliniciens de la division d'allergologie du département de pneumologie et d'addictologie du CHU de Montpellier, les membres de l'unité de recherche INSERM en cours de constitution Desbrest' Institute of Epidemiology and Public Health (IDESP). L'objectif final sera d'identifier les facteurs de risques présents dans les données patients afin de prévenir les maladies allergiques respiratoires.

Le CHU de Montpellier dispose d'un entrepôt de données (entrepôt) indépendant du SIH de production. Cet entrepôt assure la centralisation et l'historisation des données issues des différents champs et domaines d'activité de l'hôpital, en s'assurant de l'intégrité et de la qualité des données. L'entrepôt est une base de données relationnelle. Les données sont consolidées depuis 1996. À ce jour, l'entrepôt regroupe 323 tables concernant 15 millions de séjours et 2,3 millions de patients.

Les informations sont issues notamment des domaines suivants : PMSI, mouvements, actes, médicaments, dispositifs médicaux implantables, facturation, comptes rendus médicaux. Les dossiers de spécialités représentent le plus gros volume de données structurées avec 22,3 millions de questionnaires et 186,4 millions de réponses. L'entrepôt contient 33 millions de documents textuels. Les données qui seront étudiées seront extraites du contenu des dossiers patients des cinq dernières années, soit environ 5 millions de documents.

**Verrous identifiés** : Si les approches d'extraction d'information dans les textes se sont avérées très efficaces dans le domaine général, ces dernières ont montré des limites dans des secteurs spécifiques comme la santé (Nguyen 2016). En effet, les chaînes de prétraitements sont difficiles à mettre en place pour traiter les informations mal placées, les oublis, les fautes d'orthographe, les acronymes, les abréviations, etc. On note également une variabilité d'expression inhérente à chaque professionnel de santé qui rend difficile la reconnaissance des termes du domaine dans les textes et par conséquent des concepts qui leur sont liés, ainsi que l'identification de relations les liant. Par ailleurs, les approches de *machine learning* actuellement utilisées nécessitent généralement beaucoup de données annotées par des experts humains, ce qui est difficile à obtenir dans le milieu médical car les annotations demandent parfois un important niveau d'expertise. Les exemples labellisés peuvent contenir des erreurs, des inconsistances due aux variations de l'attention des annotateurs. Par nature, ces données sont également très souvent déséquilibrées. Pour finir, ces approches de *machine learning* fournissent des résultats avec peu d'explications, qu'il est donc très difficile d'interpréter par la suite.

**État de l'art et Méthodologie envisagée** : Nous mettrons en place une chaîne de traitement classique en 4 étapes : 1/ prétraitement des textes, 2/ annotation pour l'identification d'entités cliniques ; 3/ classification des textes, 4/ catégorisation des patients pour les allergies respiratoires en fonction des facteurs de risques identifiés. Dans la suite, nous détaillons ces trois dernières étapes.

**Annotation** : La deuxième utilisation des données cliniques passe par une phase d'annotation des textes biomédicaux avec des terminologies et des ontologies (Tchechmedjiev 2019) via des ressources comme UMLS, SNOMED-CT, onzième version de la Classification Internationale des Maladies (CIM-11) etc. Il s'agit de reconnaitre des termes dans les textes et de leur associer une signification. L'annotation des textes médicaux est un processus complexe finement décrit par (Patel 2018) qui a listé les différents types d'entités cliniques recherchées dans les textes (e.g. Problèmes, Procédure, Dispositif médical, Structure anatomique, etc.), les modifieurs (négation, intensifieur) ainsi que les relations attendues entre ces entités. Généralement ces approches reposent soit sur des règles qui sont spécifiques et difficiles à mettre en œuvre, soit sur des approches de machine learning plus facilement généralisables mais toujours spécifiques à chaque type d'entité recherchée, soit sur des approches hybrides (Uzuner 2010). (Nguyen 2016) propose une approche itérative intéressante dans laquelle des petits jeux de textes sont annotés afin d'obtenir des exemples d'informations à extraire, puis un modèle d'apprentissage est développé pour utiliser ces exemples afin d'obtenir un modèle plus général qui sera évalué puis révisé à partir de nouveaux exemples jusqu'à un niveau de performance souhaité. Dans ce projet, nous nous baserons sur l'outil Bioportal Annotator (Tchechmedjiev 2019) développé dans l'équipe Fado du LIRMM pour annoter les

rapports cliniques et implémenterons une approche itérative.

**Classification** : Une fois l'annotation des textes libres effectuée, il est possible de mettre en place un processus de classification automatique de textes, qui consiste à assigner des catégories prédéfinies à des documents en texte libre. Ces catégories permettront la répartition des patients dans différentes sous-populations correspondants à différents niveaux de risques. La nouvelle section « Maladies allergiques et d'hypersensibilité » de la CIM-11 sera utilisée pour catégoriser les allergies. Cette nouvelle classification élaborée sous l'égide du CHU de Montpellier, a été présentée avec l'ensemble de la CIM-11 à l'Assemblée mondiale de la Santé en mai 2019, adoptée par les États Membres, et entrera en vigueur le 1er janvier 2022. Les facteurs de risques seront codifiés selon une terminologie à déterminer. Les approches traditionnelles de classification se concentrent sur le choix du meilleur classifieur (e.g. SVM ou régression logistique) et sur la définition des meilleures caractéristiques prises en entrée de ces classifieurs. La plupart des techniques sont basées sur les mots, les lexiques et sont spécifiques à une tâche particulière. Ces modèles ont été appliqués avec succès sur de très hautes caractéristiques dimensionnelles parfois éparses. Dernièrement, pour de nombreuses tâches de classification de textes, les méthodes d'apprentissage profond se sont révélées très efficaces. Cette tendance s'est confirmée avec le succès des *word embeddings* (Mikolov 2013) pour les données textuelles. Tout récemment, des méthodes comme ULMFiT (Howard 2018), ELMo (Peters 2018) ou BERT (Devlin 2018) généralisent avec succès le mécanisme des *word embeddings*. Au lieu d'associer un vecteur d'*embedding* statique à chaque mot, ces trois méthodes construisent des représentations plus élaborées, prenant en compte le contexte sémantique et syntaxique de chaque mot. Toutefois, une limite de ces approches supervisées est, comme pour la phase d'annotation, qu'elles nécessitent une forte implication des professionnels de santé, avec un important niveau d'expertise, pour l'étiquetage préalable des données utilisées l'apprentissage des modèles. Or, les experts ne sont pas toujours disponibles pour cette tâche rébarbative. Nous explorerons des approches d'augmentation de données (Abulaish 2019), d'*active learning* (Olsson 2009), des approches semi-supervisée (Li 2018a et Li 2018b) et de *self supervised learning* qui se sont avérées efficaces dans le cas de petits jeux de données annotés (Noroozi 2018). Dans ce projet, nous nous baserons sur les compétences sur la classification des petits jeux de données labellisés développées dans l'équipe Advanse du LIRMM pour classifier les rapports cliniques.

**Catégorisation des patients pour les allergies respiratoires en fonction des facteurs de risques identifiés** : Les approches développées devront être validées sur des données cliniques. Dans un premier temps, nous réaliserons une validation interne (par ré-échantillonnage *bootstrap*) pour comparer les résultats des approches aux expertises cliniques de dossiers médicaux annotés et classés par un pool de médecins allergologues. Puis, une validation externe sera menée avec l'application des approches sur de nouveaux dossiers prospectifs également expertisés par le pool de médecins. Un projet de recherche clinique pourra alors être déposé à l'appel d'offre interne du CHU pour passer à une validation clinique de l'approche retenue. Le but sera d'estimer la sensibilité, spécificité, valeurs prédictives positive et négative de l'approche dans le diagnostic de l'allergie. Nous évaluerons également chacune des approches sur leur facilité de mise en œuvre et de généralisation qui dépendra du nombre d'annotations nécessaires pour atteindre un niveau de qualité suffisant de classification. Une attention particulière sera mise sur la question de l'explicabilité des résultats de la classification, essentielle pour l'acceptabilité par les cliniciens. Elle

passe par l'interprétation de l'importance des caractéristiques impliquées dans les tâches d'annotation et de classification mais également par la visualisation des résultats proposés. Les classifieurs basés sur une approche *deep learning* sont généralement considérées comme des boites noire. Or, lorsque nous regardons une scène, nous focalisons notre attention sur certains objets, personnes car notre expérience nous a appris à identifier l'essentiel de l'information. Actuellement des systèmes (Ragheb 2019) intégrant des mécanismes d'attention parviennent à simuler ce phénomène pour améliorer leurs performances et surtout guider l'interprétabilité de modèles prédictifs.

**Originalité de l'approche** : Les approches d'annotation et de classification envisagées dans le cas de petits jeux de données sont particulièrement innovantes et font l'objet de travaux récents. Tout d'abord, il est difficile d'augmenter des données textuelles en raison de la grande complexité de la langue. Une technique consiste à utiliser des thésaurus pour remplacer par des synonymes ou par des mots similaires calculés à partir des *word embeddings* ou encore à partir de *word embeddings* contextualisés (Fadaee 2017). Une autre approche repose sur la *Back-translation* qui consiste à traduire la langue cible dans une langue source et à mélanger à la fois la phrase source originale et celle rétro-traduite pour entraîner un modèle (Sennrich 2016). Pour finir, une dernière approche proposée par (Kafle 2017) ne cherchent pas à remplacer un mot parmi quelques mots des textes mais génèrent une phrase entière.Un autre axe consiste à mettre en place des approches reposant sur de l'*active learning*. L'intuition principale est que si un algorithme d'apprentissage peut choisir les données qu'il veut pour apprendre, ses performances vont s'améliorer avec beaucoup moins de données pour l'apprentissage. En effet, dans ce type de tâche, il est important d'optimiser les informations disponibles afin que les systèmes de classification puissent les utiliser le plus efficacement pendant la phase d'apprentissage tout en préservant l'acquisition de nouveaux échantillons étiquetés. Si beaucoup d'approches d'*active learning* ont été proposées pour la classification d'images, on trouve dans la littérature moins de propositions pour les textes et encore moins combinée avec du *deep learning* (Siddhant 2018). Pour finir, nous investiguerons des approches semi supervisées (Li 2018a et Li 2018b) qui combinent des données non étiquetées avec un petit ensemble d'apprentissage de données étiquetées pour entraîner de meilleurs modèles ainsi que des approches dites de *self supervised learning* qui entraînent des modèles pour des tâches très simples pour lesquelles il est simple de générer des données, puis réalisent un transfert vers la tâche visée (Noroozi 2018).

## 2/ Interactions entre les partenaires

Le projet proposé ici applique des techniques modernes d'IA et de TAL à l'analyse de corpus textuels cliniques peu annotés. Il s'inscrit donc dans l'axe « Algorithmes et calculs » soutenus par NUMEV. Bien que notre objectif premier soit l'avancée des connaissances dans le domaine de l'allergologie, les outils d'IA et de TALN que nous implémenterons s'appuieront sur des techniques très récentes d'annotation, de deep learning, qui seront elles-mêmes l'objet de publications scientifiques. Finalement, le projet va au-delà des avancées technologiques et scientifiques attendues, car la communauté scientifique locale sur le site de Montpellier présente des compétences complémentaires mais doit se structurer et des liens doivent être construit (ici entre le

CHU et le LIRMM) afin de tirer parti de la richesse des données produites par les professionnels de santé.

La thèse sera supervisée par J. Azé et co-supervisée par S. Bringay et M. Servajean (ADVANSE). Une première collaboration a déjà été réalisée dans le cadre d'un stage de Master 2 DECOL (Données, Connaissances et Langage Naturel) dont le sujet était « Application de méthodes de classification à des textes médicaux en Cardiologie ». Le DIM a également accueilli un stagiaire de Master 2 en 2017 dirigé par A. Laurent sur le sujet « Mining Spatial Gradual Patterns: Application to the Measurement of Potentially Avoidable Hospitalizations ». Il accueille actuellement un thésard co-dirigé par A. Laurent (FADO) dont le sujet est « Conception d'un système décisionnel pour la réduction des hospitalisations potentiellement évitables ». Le DIM du CHU aura en charge la constitution des jeux de données de patients suivis par la division d'allergologie du département de pneumologie et d'addictologie du CHU de Montpellier et de patients dont le PMSI mentionne un code CIM10 d'allergie ou un acte de test allergique ou de désensibilisation. L'expertise clinique disponible au sein de l'équipe du département de pneumologie et d'addictologie du CHU de Montpellier et du DIM sera très précieuse pour évaluer la pertinence clinique de nos résultats pour le projet de prévention des allergies respiratoires. Nous travaillerons en collaboration avec des chercheurs de l'IDESP et en particulier avec M. Trzmielewski en thèse en Science de l'Information et de la Communication sur le sujet « Informations et données en allergologie : vers un modèle d'organisation des connaissances pour la conception de dispositifs infocommunicationnels » (ALLERGIDOC). Le thésard recruté sur ce projet travaillera sur le site du CHU 2,5 jours par semaine et les données resteront sur le site du CHU. Le reste du temps, le thésard sera présent au LIRMM.

**LIRMM** : J. Azé, S. Bringay et M. Servajean sont experts dans la fouille de données de santé. Ils encadrent actuellement deux thèses sur l'apprentissage automatique pour des applications médicales.  C. Jonquet est spécialiste des ontologies médicales.
**CHU** : L'unité Sciences des Données de Santé, codirigé par le Dr Grégoire Mercier et Caroline Dunoyer, est experte en entrepôt de données de santé, en évaluation économique et recherche sur les services de santé.
Le secteur statistique de l'Unité de Recherche Clinique et Epidémiologie, dirigé par N Molinari (membre de l'Institut Montpelliérain Alexander Grothendieck (IMAG)) apportera son expertise en analyse et statistique pour la partie évaluation.
L'Unité d'Analyse médico-économique des Processus de Soins, dirigée par le Dr I. Giraud apportera son expertise sur l'analyse des parcours de soins et des déterminants médicaux en collaboration avec les cliniciens allergologues sous la direction du Pr P. Demoly.
L'unité de Pneumologie et Allergologie du CHU a été labelliser en 2018 « Centre Collaborateur (CC) OMS pour la Classification Scientifique » par l'organisation Mondiale de la Santé (OMS). A ce titre, le Dr Luciana Tanno a coordonné l'élaboration de la section « Maladies allergiques et hypersensibilités » de la onzième version de la Classification Internationale des Maladies (CIM-11). Le projet bénéficiera donc de son expertise en classification.

## 3/ Impact sur la formation

Grand public : Un site web dédié au projet et présentant les principaux résultats sera créé. Les membres du projet profiteront de différentes occasions pour diffuser au grand public : journées portes ouvertes dans les universités, fête de la science, etc.
Vulgarisation à destination de l'enseignement supérieur : Nous utiliserons les vecteurs de diffusion des savoirs mis en place par les différents partenaires (Master 2 TIC Santé de l'Université de Montpellier, Master MIASHS Data Scientist), puis nous nous rapprocherons des facultés de médecine, des médecins généralistes en utilisant des supports tels que les TIC, simulations pédagogiques au CHU de Montpellier, avec une diffusion auprès des étudiants du DESC d'allergologie et du département universitaire de médecine générale en utilisant des supports tels que les TIC, ou les outils de simulations pédagogiques.

## 4/ Potentiel de Valorisation

**Publications** : Nos travaux seront publiés dans des conférences ou revues scientifiques internationales dans le domaine de l'informatique (e.g. ACL, LREC, MLNLP…) et de l'informatique médicale (e.g. Medinfo, MIE, JBI…) pour mettre l'accent sur les méthodes développées et dans les revues médicales de spécialité pneumologie pour leur application au domaine de la prévention des allergies afin d'en démontrer l'intérêt clinique. L'atelier annuel IC et Santé adossé à la conférence française IC sera un support régulier de diffusion des résultats. Des communications pourront être également intéressantes dans le cadre des journées/ateliers des GDR MaDICS sur les aspects masses de données, ou encore dans le cadre de journées liées aux recherches sur l'information médicale ou la e-santé. Les acteurs de ce projet, notamment du LIRMM, sont actifs dans le pilotage de ces conférences, ateliers et pourront proposer des présentations des travaux liés au projet.

**Partenariat** : L'intégration de l'outil d'annotation au logiciel de Dossier Patient Informatisé du CHU de Montpellier, DxCare de la société DEDALUS sera étudié pour permettre l'annotation de texte pour d'autres disciplines médicales.

**Innovation** : De façon plus générale, l'annotation de documents de l'entrepôt de données du CHU de Montpellier permettra d'envisager des partenariats avec des start-ups ayant besoin de développer ou de valider des outils basés sur l'intelligence artificielle.

## Bibliographie


- H. Nguyen and J. Patrick. 2016. Text Mining in Clinical Domain: Dealing with Noise. SIGKDD 2016. ACM, New York, NY, USA, 549-558.
- A. Tchechmedjiev, A. Abdaoui, V. Emonet, C. Jonquet. Enhanced Functionalities for

Annotating and Indexing Clinical Text with the NCBO Annotator+ Bioinformatics 34(11), 2018

- O. Uzuner, I. Solti, E. Cadag. Extracting medication information from clinical text. JAMIA 2010;17:514–8.
- J. Howard, S. Ruder, Universal Language Model Fine-tuning for Text Classification, ACL 2018
- M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer. Deep contextualized word representations. NAACL 2018.
- J. Devlin, M. Chang, K. Leen Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018
- M. Abulaish, A. Kumar Sah, A text data augmentation approach forimproving the performance of cnn, COMSNETS 2019
- F. Olsson, A literature survey of active machine learning in the context of natural language processing. SICS 2009.
- Y. Li, J. Ye Learning Adversarial Networks for Semi-Supervised Text Classification via Policy Gradient. SIGKDD 2018
- W. Ragheb, J. Azé, S. Bringay, M. Servajean: Language Modeling in Temporal Mood Variation Models for Early Risk Detection on the Internet. CLEF 2019
- K. Kafle, M. Yousefhussien and C. Kanan. Data Augmentation for Visual Question Answering. 2017
- M. Fadaee, A. Bisazza, C. Monz. Data Augmentation for Low-Resource Neural Machine Translation. ACL 2017
- R. Sennrich, B. Haddow, A. Birch; Improving Neural Machine Translation Models with Monolingual Data. ACL
- M. Noroozi, A. Vinjimoor, P. Favaro, H. Pirsiavash; Boosting Self-Supervised Learning via Knowledge Transfer. CVPR 2018, pp. 9359-9367