Chair of Junior Professor

Supporting institution/organization: Inria Research Center Lille
Head of the institution/organization: Mireille Régnier
Site concerned: *Centre INRIA de l'Université de Lille*
Academic Region: Hauts-de-France

Partner instiutions/organizations: *Université de Lille, Inserm, CHU Lille*

Corresponding person (for questions): Prof. Vincent Sobanski, vincent.sobanski@univ-lille.fr

Project name:
ENDOtyping Chronic Inflammatory Diseases: how to obtain a unified representation of patients from heterogeneous data?
Acronym: **ENDOMIC**

Keywords:
Health data ; omics analyses ; multi-dimensional analysis ; clinical heterogeneity ; chronic inflammatory diseases

Target duration: 6 years

Scientific topic: digital health

Section (s) CNU/CoNRS/CSS corresponding: CNU 26/27, CoNRS 6/41- CSS 5/6

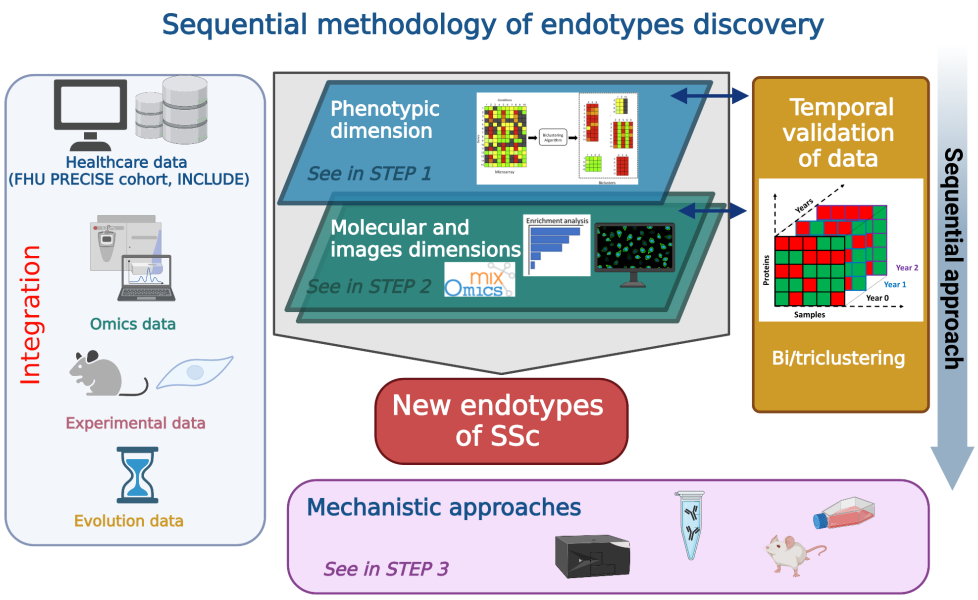
Establishment strategy:

Digital Life Sciences and Health is a major area of research and innovation at Inria, and the recent program to create **joint Inserm-Inria teams** is giving it new impetus. A strategic challenge is the structuring and processing of **massive health data** from hospitals and research centers while ensuring interoperability and preserving confidentiality.
This national scientific axis of Inria is consistent with the policy of the Lille campus:
- the Lille University Hospital has an institutional health data warehouse (EDS) **INCLUDE**. This EDS gathers 1.6 million patients received at the Lille University Hospital since 2008. INCLUDE is part of the national Health Data Hub dynamic, which consists of setting up **regional hubs** bringing together data producers and operating skills. A major asset is its membership of the **European Health Data & Evidence Network (EHDEN)**, which harmonizes data sources in OMOP interoperable format;
- an **Inserm ENDOMIC research team** (including 2 post-doctoral students, 2 PhD students, 1 data-scientist engineer) led by Pr Vincent Sobanski (member of the Institut Universitaire de France, co-founder of INCLUDE). This team is currently maturing to become a joint Inserm-Inria project team;
- we have just been awarded the ANR "skills and professions of the future" Digital Health call for expressions of interest with the CAPS'UL project (€3.8M);

- the Internal Medicine and Clinical Immunology team of the Lille University Hospital is a national and European reference centre for rare diseases (European Reference Network RECONNET).

Strategy of the host laboratory

The Inserm INFINITE U1286 research unit has set up the ENDOMIC team (Endotyper les maladies inflammatoires chroniques). The objective of this team is to analyze a clinico-biological cohort (FHU PRECISE, Internal Medicine Department of the Lille University Hospital) containing 1,000 highly-phenotyped patients with chronic inflammatory diseases (CID) for whom sera are available in a biobank at 0, 1, 2 and 3 years. This cohort is enriched with care data from the Lille University Hospital data warehouse (INCLUDE), omics profiles and routine diagnostic test images (100k images already standardized). All ethical and regulatory approvals are already available, with the team already performing unsupervised analyses on clinical and multi-omics data[1,2].



The profile sought in this CPJ is a **researcher specialized in computer science, statistics and/or artificial intelligence**, whose background would allow recruitment as an Inria research director within 3 to 6 years. He/she will complete a team comprising a university professor in internal medicine, two post-doctoral researchers (a researcher in computer science, a researcher in immunology), two PhD students (a medical doctor and a pharmacist) and a data-scientist engineer.

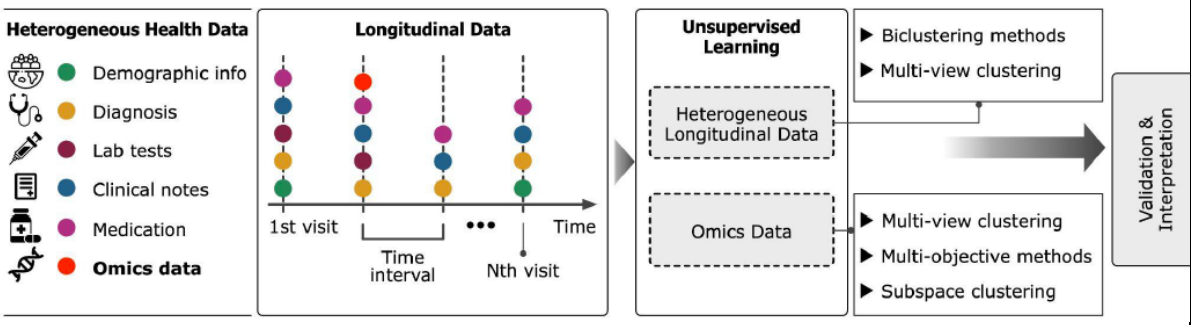
[1] Chepy A et al. IgG from Systemic Sclerosis Patients Induce a Profibrosing and Serotype-dependent Phenotype in Normal Dermal Fibroblast: A Multi-omics Study. *Frontiers Immunol* 2022.

[2] Sobanski V et al. Phenotypes Determined by Cluster Analysis and Their Survival in the Prospective European Scleroderma Trials and Research Cohort of Patients With Systemic Sclerosis. *Arthritis Rheumatol* 2019;71(9):1553-70.

The team actively collaborates with computer scientists from the University of Lille (ORKAD team of the **CRIStAL laboratory**) or from the **KU Leuven**. The candidate will be actively involved in projects with our partners.
No prerequisite in immunology is required for this position. Research experience in the field of health data would be an important asset but is not mandatory.

Summary of the scientific project:

Chronic inflammatory diseases (CID) are responsible for **significant morbidity and mortality** and can affect young individuals, causing a significant socio-professional impact. There is no curative treatment to date, and the results of recent clinical trials have been disappointing, probably due to the **significant heterogeneity** of the diseases, which accounts for a wide variety of clinical presentations and organ involvement.
The project aims to (i) identify **endotypes** (groups of similar patients) by performing a **multidimensional integrative analysis** combining health data from various sources with omics and imaging data; (ii) then validate these endotypes, study their stability over time and select them to explore pathophysiological mechanisms in a targeted manner.
Our team is developing an **unsupervised approach** (see figure)[3,4]. The candidate will drive this dynamic within the team by contributing his/her expertise. He/she will also be able to **develop other approaches** that could be promising. The computer science methodological developments are the subject of publications in dedicated journals and enable clinicians and researchers in biology to carry out innovative work that is valued in medical journals.



CID are models for studying chronic diseases. Thus, the innovations brought by this project can be **transposed to many other human pathologies**, contributing to the dynamics of personalized medicine.

[3] José-García et al. Multi-view Clustering of Heterogeneous Health Data: Application to Systemic Sclerosis. *Parallel Problem Solving from Nature – PPSN XVII*. 2022 Springer International Publishing, Cham. 352–367.

[4] José-García et al. What's in a distance? Exploring the interplay between distance measures and internal cluster validity in multi-objective clustering. *Nat Comput* 2022. https://doi.org/10.1007/s11047-022-09909-y

Summary of the teaching project:

The candidate will bring his or her expertise in data science and artificial intelligence to the **precision health**-related teaching delivered at the University of Lille (in connection with the "health" and "digital" hubs supported by the Initiative of Excellence). The Precision Health graduate program is open to biologists, physicians and pharmacists, as well as (bio)computer scientists, in order **to train the next generation of national and international research leaders** for academia, industry and the biotechnology world. It includes a Master's program in Biology and Health (M2) and a PhD program, **open to the entire scientific community**. It is structured around interactive seminars on current cutting-edge technologies and approaches (e.g. big data, omics, next generation sequencing, systems biology and integrated bioinformatics, AI, mathematical modelling of biological systems, innovative therapies, cohort management, health economics and ethics). He/she will be able to develop specific teaching or join the teaching team in order to structure and lead this **strategic training offer for our campus**.

Scientific diffusion:

The Chairholder will be required to promote the results obtained through publications in recognized **scientific journals** (target rank A) and in **international conferences** of reference for the computer science and/or health disciplines. The transdisciplinary nature of the research carried out will lead to the publication of articles in **generalist journals with a large audience** and to the writing of chapters or educational books. He/she will also have a certain appetite for **popularizing science**, particularly among the general public (open days, high school students, science festivals).

Open science:

The human-machine interface will allow navigation in the **virtual representation of endotypes**. Health professionals will be able to enter certain variables into a query module that will allow them to predict the potential assignment of a given patient to the endotypes. In addition, they will be able to implement their data in the form of an **interoperable database** that can be integrated into the algorithm and thus enrich the definition of endotypes (**open science**). Patients and health professionals will be asked about their needs. They will be involved in **co-creation activities** to finally obtain specifications that meet their **needs** and expectations. The specifications of this interface will also integrate the possibility for a health professional to position a given patient in the endotypes based on a few variables entered in the model, with the interface providing a **degree of confidence for this classification** (probabilistic projection).
The work carried out on the interoperability of data in OMOP format within the Lille University Hospital will enable the tool to be open "by design". The data necessary for the development of the algorithm will be available according to the **principles of open science** (we are members of the EHDEN initiative [European Health Data & Evidence Network]).

Science and society:

The appropriation of this new tool by the actors will be ensured by publications in **open access journals**, participation in **scientific and patient congresses**, documentation on the **website**, communication on **social networks** with professionals and patient associations (for example in the networks in which we are already involved: FAI2R, EUSTAR) and **scientific seminars**.
We are very involved in **scientific communication, especially with the general public**. For example, we have initiated a series of conferences for the general public on Covid-19. A YouTube platform has been set up[5], whose videos have been seen 27,000 times and widely relayed in the media and labelled by the Ministry of Higher Education and Research.

[5] https://www.youtube.com/channel/UCKv7KQLa0w2AVq6hzcY3q1w/featured

Indicators:

Primary indicators	
Short term (<2 years)	1. Recruitment of engineers, doctoral students and post-doctoral fellows constituting the Chair 2. Participation in the submission of national or international "health" projects (ANR, ERC, European network...) in the field of ENDOMIC topics
Mid term (2-4 years)	1. Publications in high level scientific journals 2. Establishment of a specific teaching program such as a university diploma or certificate graduate program (structured by the chairholder) 3. Submission by the chairholder of "computer science" projects of national or international scope (ANR, ERC, European network...) in the thematic of the ENDOMIC team
Long term (4-6 years)	1. Open science dynamic (actions in the framework of EHDEN and AEx FLAMED, making public datasets and codes available, integration of new FAIR data in the integrative classification algorithm) 2. Transposition of the classification tool to other medical fields (new collaborations, creation of a start-up)
Secondary indicators	
Short term (<2 years)	1. Publications in intermediate level computer or medical journals 2. Participation in international conferences of scientific societies
Mid term (2-4 years)	1. Integration into research networks, particularly European ones (international visibility) 2. Organization of an international scientific seminar
Long term (4-6 years)	1. Academic collaborations, in particular capacity of the ENDOMIC team to send/receive students/researchers in international mobility 2. Establishment of partnership contracts

Chaire de professeur junior - Fiche projet type

Établissement/organisme porteur : Inria
Nom du chef d'établissement/d'organisme : Bruno Sportisse
Site concerné : *Centre INRIA de l'Université de Lille*
Région académique : Hauts de France

Établissements/organismes partenaires : *Université de Lille, Inserm, CHU Lille*

Personne à contacter pour les renseignements : Pr Vincent Sobanski,
vincent.sobanski@univ-lille.fr

Nom du projet :
ENDOtyper les Maladies Inflammatoires Chroniques : comment obtenir une représentation unifiée des patients à partir de données hétérogènes ?
Acronyme : **ENDOMIC**

Mots-clés : *donner 5 mots-clés caractérisant le projet scientifique*
Données de santé ; analyses omiques ; analyses multi-dimensionnelles ; hétérogénéité clinique ; maladies inflammatoires chroniques

Durée visée : 6 ans

Thématique scientifique : santé numérique

Section (s) CNU/CoNRS/CSS correspondante (s) : CNU 26/27, CoNRS 6/41- CSS 5/6

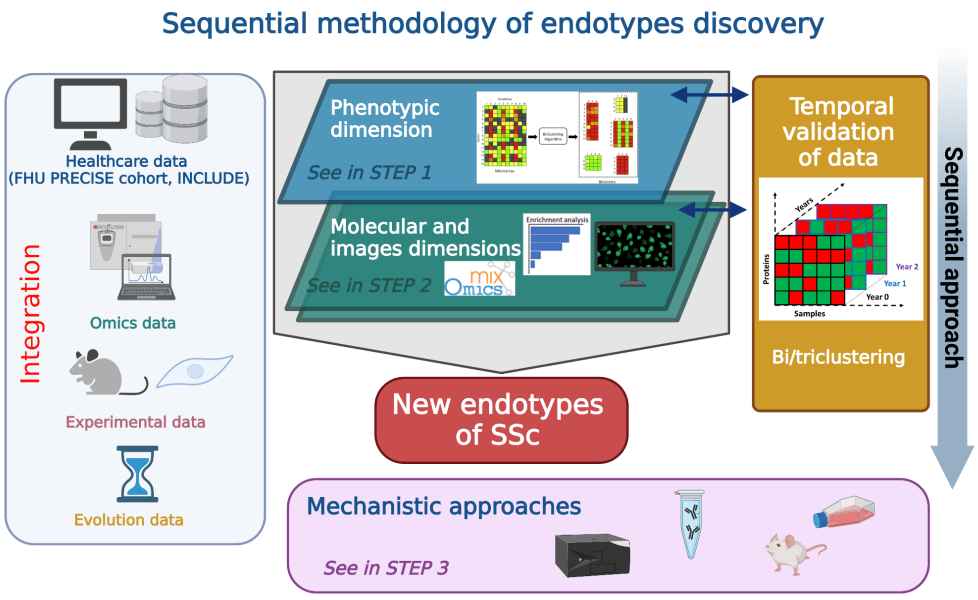
Stratégie d'établissement : *décrire en quoi le recrutement est en lien avec la stratégie de l'établissement (15 lignes maximum)*

Le Numérique dans les Sciences du Vivant et la Santé est un domaine de recherche et d'innovation majeur d'Inria, auquel le récent **programme de création d'équipes communes Inserm-Inria** donne une nouvelle impulsion. Un défi stratégique est la structuration et le traitement des **données de santé massives** issues des hôpitaux et centres de recherche en assurant l'interopérabilité et en préservant la confidentialité.
Cet axe scientifique national d'Inria est cohérent avec la politique du campus lillois :
- le CHU de Lille dispose d'un entrepôt de données de santé (EDS) institutionnel **INCLUDE**. Cet EDS rassemble 1,6 million de patients reçus au CHU de Lille depuis 2008. INCLUDE s'inscrit dans la dynamique nationale du *Health Data Hub* consistant à mettre en place des **Hubs régionaux** rassemblant les producteurs de données et les compétences d'exploitation. Un atout majeur est l'appartenance au **réseau Européen EHDEN** (European Health Data & Evidence Network – harmonisation des sources de données au format interopérable OMOP) ;
- une **équipe de recherche Inserm ENDOMIC** (comprenant 2 post-doctorants, 2 étudiants PhD, 1 ingénieur data-scientist) dirigée par le Pr Vincent Sobanski (membre Institut Universitaire de France, co-fondateur d'INCLUDE). Cette équipe est en cours de maturation pour devenir une équipe projet conjointe Inserm-Inria ;

- nous venons d'être **lauréats de l'appel à manifestation d'intérêt ANR « compétences et métiers d'avenir » Santé numérique** avec le projet CAPS'UL, financé à hauteur de 3,8M€ ;
- l'équipe de Médecine interne et immunologie clinique du CHU de Lille est centre de référence national et européen sur les maladies rares (European Reference Network RECONNET)

Stratégie du laboratoire d'accueil : *décrire en quoi le recrutement est en lien avec la stratégie du laboratoire d'accueil (15 lignes maximum)*

L'unité de recherche Inserm INFINITE U1286 a mis en place l'équipe ENDOMIC (Endotyper les maladies inflammatoires chroniques). Cette équipe a pour objectif d'analyser une cohorte clinico-biologique (FHU PRECISE, Service de Médecine interne du CHU de Lille) contenant 1000 patients atteints de MIC hautement phénotypés et pour lesquels les sérums sont disponibles dans une biobanque à 0, 1, 2 et 3 ans. Cette cohorte est enrichie des données de soin issues de l'EDS du CHU de Lille (INCLUDE), des profils omiques et des images de tests diagnostiques de routine (100k images déjà standardisées). Toutes les autorisations éthiques et réglementaires sont déjà disponibles, l'équipe réalisant déjà des analyses non supervisées sur des données cliniques et multi-omiques [1,2].



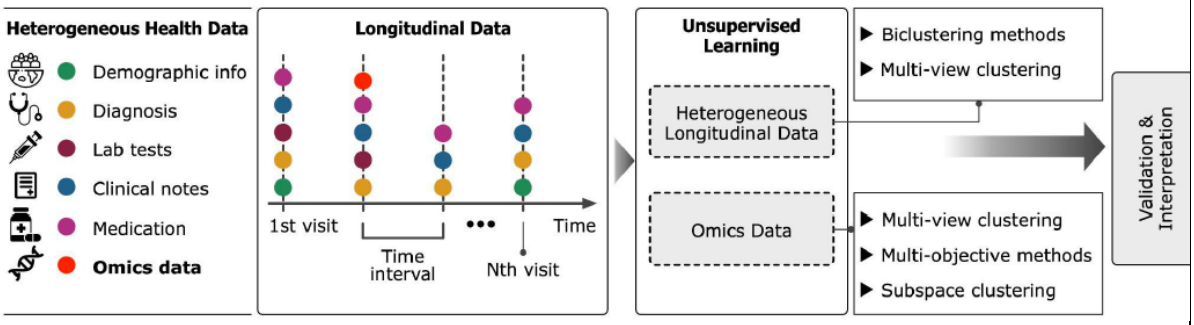
Le profil recherché dans cette CPJ est un **chercheur spécialiste en informatique, statistique et/ou intelligence artificielle**, dont le parcours permettrait un recrutement en tant que directeur de recherche Inria d'ici 3 à 6 ans. Il viendra compléter une équipe comprenant un professeur des universités-praticien hospitalier en médecine interne, deux chercheurs post-doctorants (un chercheur en informatique, une chercheuse en immunologie), deux étudiants PhD (un médecin et une pharmacienne) et un ingénieur data-scientist.
L'équipe collabore activement avec des chercheurs en informatique de l'Université de Lille (équipe ORKAD du **laboratoire CRIStAL**) ou de la **KU Leuven**. Le candidat sera activement impliqué dans les projets avec nos partenaires.

[1] Chepy A et al. IgG from Systemic Sclerosis Patients Induce a Profibrosing and Serotype-dependent Phenotype in Normal Dermal Fibroblast: A Multi-omics Study. *Frontiers Immunol* 2022.
[2] Sobanski V et al. Phenotypes Determined by Cluster Analysis and Their Survival in the Prospective European Scleroderma Trials and Research Cohort of Patients With Systemic Sclerosis. *Arthritis Rheumatol* 2019;71(9):1553-70.

Aucun pré-requis en immunologie n'est exigé pour ce poste. Une expérience de recherche dans le domaine des données de santé serait un atout important.

Résumé du projet scientifique : *15 lignes maximum*

Les maladies inflammatoires chroniques (MIC) sont responsables d'une **morbi-mortalité importante** et peuvent toucher des individus jeunes, causant un impact socio-professionnel significatif. Il n'existe pas de traitement curatif à ce jour, et les résultats des essais cliniques récents ont été décevants probablement en raison de **l'hétérogénéité importante** de la maladie qui explique une grande variété de présentations cliniques et atteintes d'organes.
Le projet vise (i) à mettre en évidence des **endotypes** (groupes de patients similaires) en réalisant une **analyse intégrative multidimensionnelle** combinant des données de santé de sources variées avec des données omiques et d'imagerie ; (ii) puis de valider ces endotypes, d'étudier leur stabilité dans le temps et les sélectionner pour explorer de façon ciblée les mécanismes physiopathologiques.
Notre équipe développe une **approche non supervisée** (cf. figure) [3] [4]. Le candidat impulsera cette dynamique au sein de l'équipe en apportant son expertise. Il pourra aussi **développer d'autres approches** qui lui semblent intéressantes. Les développements méthodologiques informatiques font l'objet de publications dans les revues dédiées et permettent aux cliniciens et chercheurs en biologie de réaliser des travaux innovants valorisés dans des revues médicales.



Les MIC sont des modèles d'études des maladies chroniques. Ainsi, les innovations apportées par ce projet pourront être **transposées à de nombreuses autres pathologies humaines**, contribuant à la dynamique de médecine personnalisée.

[3] José-García et al. Multi-view Clustering of Heterogeneous Health Data: Application to Systemic Sclerosis. *Parallel Problem Solving from Nature – PPSN XVII*. 2022 Springer International Publishing, Cham. 352–367.
[4] José-García et al. What's in a distance? Exploring the interplay between distance measures and internal cluster validity in multi-objective clustering. *Nat Comput* 2022. https://doi.org/10.1007/s11047-022-09909-y

Résumé du projet d'enseignement : *15 lignes maximum*

Le candidat apportera son expertise en sciences des données et intelligence artificielle dans les enseignements liées à la **santé de précision** délivrés à l'Université de Lille (en lien avec les hubs « santé » et « numérique » soutenus par l'Initiative d'excellence). Le programme gradué « santé de précision » est ouvert aux biologistes, médecins et pharmaciens, ainsi qu'aux (bio)informaticiens, afin de **former la prochaine génération de leaders nationaux et internationaux** de la recherche pour le monde universitaire, l'industrie et le monde des biotechnologies. Il comprend un parcours de Master en Biologie et Santé (M2) et un programme doctoral, **ouvert à toute la communauté scientifique**. Il est structuré autour de séminaires interactifs sur les technologies et les approches de pointe actuelles (par exemple les données en grand nombre, les omics, le séquençage de nouvelle génération, la biologie des systèmes et la bio-informatique intégrée, l'IA, la modélisation mathématique des systèmes biologiques, les thérapies innovantes, la gestion des cohortes, l'économie et l'éthique de la santé). Il pourra y développer un enseignement spécifique ou intégrer l'équipe pédagogique afin de structurer et animer cette **offre de formation stratégique pour notre campus**.

Synthèse financière : à réaliser à partir de la fiche financière jointe, décrire les besoins financiers et leur répartition pour mener à bien le projet scientifique (doctorant, post-doctorant, IT, équipement, …)

Total financé sur CPJ (dont package ANR)	558k€
Co-financement - Un chercheur post-doctorant 24 mois - Deux doctorants - Un ingénieur 48 mois	512k€
Total du projet	1070k€

Diffusion scientifique : préciser les résultats attendus en termes de diffusion scientifique (publications, communications, …)

Le titulaire de la chaire sera tenu de valoriser les résultats obtenus au sein de publications dans des **journaux scientifiques** reconnus (cible rang A) et dans les **congrès internationaux** de référence pour les disciplines informatique et/ou santé. Le caractère transdisciplinaire de la recherche menée amènera à la publication d'articles dans des **revues généralistes à forte audience** et à la rédaction de chapitres ou ouvrages pédagogiques. Il devra aussi avoir une appétence certaine pour la **vulgarisation scientifique** en particulier auprès du grand public (journées portes ouvertes, lycéens, fête de la science).

Science ouverte : le projet s'inscrit-il dans une démarche de science ouverte ? Si, oui décrire sa mise en œuvre.

L'interface homme-machine permettra la **navigation dans la représentation virtuelle des endotypes**. Les professionnels de santé pourront saisir certaines variables dans un module de requête qui leur permettra de prédire l'affectation potentielle d'un patient donné dans les endotypes. En outre, ils pourront implémenter leurs données sous la forme d'une **base de données interopérable** qui pourra être intégrée à l'algorithme et ainsi enrichir la définition des endotypes (**open science**). Les patients et les professionnels de santé seront interrogés sur leurs **besoins**. Ils seront impliqués dans des **activités de co-création** pour finalement obtenir des spécifications répondant à leurs besoins et attentes. Les spécifications de cette interface intégreront également la possibilité pour un professionnel de santé de positionner un patient donné dans les endotypes à partir de quelques variables entrées dans le modèle, l'interface fournissant un **degré de confiance pour cette classification** (projection probabiliste).
Le travail réalisé sur l'interopérabilité des données au format OMOP au sein du CHU de Lille permettra à l'outil d'être ouvert « by design ». Les données nécessaires au développement de l'algorithme seront disponibles selon les **principes de la science ouverte** (nous sommes membres de l'initiative EHDEN [European Health Data & Evidence Network]).

Science et société : *le projet envisage-t-il une communication auprès du grand public ? Si oui : préciser de quelle manière et à quelle échéance*

L'appropriation de ce nouvel outil par les acteurs sera assurée par des publications dans des **revues en accès libre**, la participation à des **congrès scientifiques et de patients**, la **documentation sur le site internet**, la communication sur les **réseaux sociaux** avec les professionnels et les associations de patients (par exemple dans les réseaux dans lesquels nous sommes déjà impliqués : FAI2R, EUSTAR) et des **séminaires scientifiques**.
Nous sommes très investis dans la **communication scientifique, en particulier auprès du grand public**. Par exemple, nous avons initié un cycle de conférences grand public sur la Covid-19. Une plateforme YouTube a été mise en place[5], dont les vidéos ont été vues 27 000 fois et largement relayée dans les médias et labélisée par le Ministère de l'Enseignement Supérieur et de la Recherche.

[5] https://www.youtube.com/channel/UCKv7KQLa0w2AVq6hzcY3q1w/featured

Indicateurs : *préciser les indicateurs de suivi du déploiement du projet et la méthodologie de leur suivi*

Indicateurs prioritaires	
Précoces (<2 ans)	1. Recrutement des ingénieurs, doctorants et post-doctorants constituant la chaire 2. Participation à la soumission de projets « santé » d'envergure nationale ou internationale (ANR, ERC, réseau européen…) dans la thématique de l'équipe ENDOMIC
Moyen terme (2-4 ans)	1. Publications dans des journaux scientifiques de haut niveau 2. Mise en place d'un enseignement spécifique type DU ou certificat programme gradué (structuration par le titulaire de la chaire) 3. Dépôt par le titulaire de la chaire de projets « informatique » d'envergure nationale ou internationale (ANR, ERC, réseau européen…) dans la thématique de l'équipe ENDOMIC
Long terme (4-6 ans)	1. Dynamique de science ouverte (mise à disposition de jeux de données et codes publics, intégration de nouvelles données FAIR dans l'algorithme de classification intégrative) 2. Transposition de l'outil de classification vers d'autres champs médicaux (nouvelles collaborations, constitution d'une start-up)
Indicateurs secondaires	
Précoces (<2 ans)	1. Publications dans des revues informatiques ou médicales de niveau intermédiaire 2. Participation aux congrès internationaux des sociétés savantes
Moyen terme (2-4 ans)	1. Intégration dans des réseaux de recherche notamment européens (visibilité internationale) 2. Organisation d'un séminaire scientifique international
Long terme (4-6 ans)	1. Collaborations académiques en particulier capacité de l'équipe ENDOMIC à envoyer/recevoir des étudiants/chercheurs en mobilité internationale 2. Mise en place de contrats partenariaux