

PROPOSITION DE SUJET DE THESE

TITRE DE LA THÈSE : MODÉLISATION DU PARCOURS DE SOINS ET DES DONNÉES CLINIQUES TEXTUELLES POUR LA CONSTITUTION DE COHORTES DE PATIENTS SIMILAIRES

CO-ENCADRANTE : GERARDIN CHRISTEL

1. Contexte scientifique du projet

La constitution de cohortes de patients similaires est un enjeu central en recherche clinique, en épidémiologie et en médecine de précision. Elle conditionne la capacité à identifier des sous-groupes homogènes de patients, à analyser l'hétérogénéité des trajectoires de soins et à adapter les stratégies thérapeutiques. Les approches classiques reposent principalement sur des variables structurées (diagnostics codés, actes, biologie) et négligent une part importante de l'information contenue dans les textes cliniques et dans la dynamique temporelle des parcours de soins. Cet enjeu est particulièrement aigu dans le champ des maladies auto-immunes systémiques (lupus érythémateux systémique, syndrome des anti-phospholipides, sclérodermie systémique et maladie de Takayasu) dont l'expression clinique hétérogène, la sévérité variable et les parcours de soins complexes et multidisciplinaires rendent la constitution de cohortes homogènes particulièrement difficile.

Les travaux récents en traitement automatique des langues (TAL/NLP) appliqué à la santé ont permis d'extraire automatiquement des concepts médicaux à partir de comptes rendus hospitaliers et de produire des représentations vectorielles de patients fondées sur les symptômes et les données biologiques. Ces représentations ont montré leur capacité à améliorer la recherche de patients similaires et la constitution de cohortes pertinentes [1,2].

Cependant, une dimension majeure reste insuffisamment modélisée : le parcours de soins, c'est-à-dire la succession des interactions d'un patient avec le système de santé. Ce parcours constitue un signal indirect mais riche de la sévérité, de la chronicité et de l'évolution des pathologies. Il reflète notamment :

- l'intensité du recours aux soins (hospitalisations répétées, passages aux urgences),
- la diversité des prises en charge (services, spécialités),
- la dynamique temporelle des épisodes de soins,
- la trajectoire diagnostique (évolution des codes CIM-10).

Dans la littérature, plusieurs approches ont exploré la modélisation des trajectoires patients via des modèles séquentiels (RNN, Transformers temporels) ou des représentations de graphes [3]. Toutefois, ces travaux ne sont pas centrés sur la question de la similarité entre patients et la constitution de cohortes interprétables. Par ailleurs, l'articulation entre données structurées, textes cliniques et description synthétique en langage naturel du parcours reste peu étudiée.

Cette thèse a pour objectif de concevoir des représentations patients intégrant explicitement la dimension trajectoire, et évaluer leur impact sur la qualité et la pertinence clinique des cohortes générées.

2. Questions posées

L'objectif général est de déterminer dans quelle mesure l'intégration explicite du parcours de soins permet d'améliorer la constitution de groupes de patients cliniquement cohérents. Les questions posées sont :

- **Représentation** : Comment modéliser le parcours de soins d'un patient de manière synthétique et exploitable pour la comparaison entre patients, en utilisant les informations issues des textes cliniques et des données biologiques ? Il s'agira de capturer la dynamique temporelle à gros grain, à travers l'intensité du recours au système de soins et la nature des épisodes cliniques.
- **Similarité** : Comment définir la similarité entre patients lorsque plusieurs dimensions cliniques peuvent être mobilisées (pathologie, organes concernés, gravité, dynamique du parcours) ? Quels choix de métriques et de pondération de ces dimensions permettent d'obtenir des regroupements cohérents au regard des objectifs cliniques, sachant que la notion de similarité peut varier selon le contexte médical, la spécialité ou la question posée ?
- **Validation** clinique des cohortes : Dans quelle mesure l'intégration explicite du parcours de soins améliore-t-elle la pertinence clinique des cohortes de patients similaires, évaluée par leur cohérence médicale interne et leur utilité pour l'exploration de sous-groupes de patients ?

ECOLE DOCTORALE 393

CENTRE BIOMÉDICAL DES CORDELIERS

15, RUE DE L'ECOLE DE MÉDECINE 75006 PARIS

[HTTPS://ED393.SORBONNE-UNIVERSITE.FR/](https://ed393.sorbonne-universite.fr/)

CONTACT : ED393@SORBONNE-UNIVERSITE.FR / TÉLÉPHONE : 01.44.27.24.35

3. Sources de données utilisées

Les données utilisées sont issues de l'entrepôt de données de santé (EDS) de l'AP-HP, dans le cadre du projet CSE 20-93 (accord du projet en 2020, renouvelé sur amendement en 2025, avec un rafraîchissement des données le 25 novembre 2025). Cette cohorte s'intéresse spécifiquement à des patients de diagnostics et de décision diagnostique complexe en contexte de médecine interne, et plus spécifiquement à quatre pathologies : le lupus érythémateux systémique, le syndrome des anti-phospholipides, la sclérodermie systémique et la maladie de Takayasu.

Les critères d'inclusion de la cohorte étaient les suivants : patients de plus de 15 ans hospitalisés au moins 1 jour dans un service de spécialité médicale de l'AP-HP avec une date d'entrée après le 31/07/2017, présentant un antécédent personnel de Lupus érythémateux ou de maladie de Takayasu ou de sclérodermie systémique ou de syndrome des anti-phospholipides.

À partir de cette requête, le nombre de patient(e) exporté(e) sur l'espace projet est de **93 064**.

Pour ces patient(e)s, nous avons à disposition, l'ensemble des données suivantes :

- L'ensemble des documents médicaux pseudonymisés (comptes-rendus d'hospitalisation, de consultation, ordonnances, compte-rendus d'examen etc.)
- Les données des tests biologiques réalisés à l'AP-HP
- Les données de prescription et d'administration des médicaments
- Les données du parcours patients : visites, hospitalisations, changement d'unités, ainsi que celles issues du PMSI

4. Méthodes

La méthodologie proposée repose sur une architecture modulaire articulant trois composantes : (i) l'extraction et la structuration de l'information clinique à partir des données textuelles et biologiques, (ii) la modélisation du parcours de soins sous forme de représentations vectorielles intégrant la dimension temporelle, et (iii) la constitution et l'évaluation de cohortes de patients similaires. Chaque composante est conçue de manière à permettre une évaluation indépendante ainsi qu'une intégration progressive des différentes sources d'information.

4.1. Extraction et structuration de l'information clinique : cette étape consiste à extraire, à partir des comptes rendus médicaux disponibles dans l'EDS de l'AP-HP, les informations clés nécessaires à la caractérisation clinique des patients. Cette extraction reposera sur des approches de TAL adaptées au français médical. Deux stratégies complémentaires seront explorées :

- **Extraction d'entités nommées médicales** : identification des diagnostics, symptômes, traitements et événements cliniques au sein des textes, en s'appuyant sur des modèles de type CamemBERT-bio ou DrBERT, éventuellement complétés par des outils de normalisation terminologique (vers les référentiels CIM-10, ATC, SNOMED-CT).
- **Génération de résumés cliniques structurés** : utilisation de modèles de langage de grande taille (LLM) pour produire, à partir de l'ensemble des documents d'un patient, une synthèse structurée de son parcours clinique couvrant les diagnostics retenus, les atteintes d'organes, les éléments de sévérité et les grandes étapes de la prise en charge.

Les données biologiques (résultats de laboratoire) seront intégrées sous forme de vecteurs de caractéristiques agrégées (valeurs médianes, extrêmes, tendances temporelles) pour les tests biologiques pertinents dans le contexte des pathologies étudiées (auto-anticorps, complément, marqueurs inflammatoires, marqueurs de la fonction rénale, etc.).

4.2. Modélisation du parcours de soins : Le parcours sera modélisé selon une granularité variable, combinant des descripteurs agrégés et des représentations séquentielles, à travers trois niveaux.

- **Descripteurs agrégés du parcours**. Le premier niveau repose sur des indicateurs synthétiques décrivant l'intensité et la nature du recours aux soins : nombre et durée des hospitalisations, passages aux urgences, services fréquentés, répartition par spécialité, fréquence des consultations et intervalles entre épisodes. Ces variables offrent une représentation simple, interprétable et exploitable pour mesurer la similarité entre parcours.
- **Représentation séquentielle du parcours**. Le deuxième niveau vise à modéliser la dynamique temporelle. Les séquences d'événements cliniques (hospitalisations, transitions, diagnostics CIM-10) seront encodées via des plongements séquentiels. Seront comparés : Med2Vec (word2vec appliqué aux événements médicaux [4]), des modèles récurrents (GRU [5]) et des architectures Transformer adaptées aux données cliniques [3,6]. L'objectif est d'obtenir un vecteur dense représentant la trajectoire de chaque patient.
- **Représentation hybride intégrant texte et parcours** : Un troisième niveau consistera à fusionner les représentations issues des textes cliniques (embeddings de résumés ou de concepts extraits) avec les représentations du parcours de soins. Plusieurs stratégies de fusion seront évaluées : concaténation simple avec réduction dimensionnelle, fusion par réseaux de type attention croisée (cross-attention), et fusion tardive avec pondération apprise ou paramétrable.

4.3. Mesure de similarité et constitution de cohortes

La similarité entre patients sera calculée dans l'espace des vecteurs obtenus précédemment. Plusieurs métriques seront comparées (cosinus, euclidienne, Mahalanobis) selon la structure des représentations. La constitution de cohortes reposera sur des méthodes de clustering (k-means, hiérarchique, HDBSCAN) et sur la recherche de plus proches voisins (k-NN, approximate nearest neighbors via FAISS ou Annoy).

ECOLE DOCTORALE 393

CENTRE BIOMÉDICAL DES CORDELIERS

15, RUE DE L'ECOLE DE MÉDECINE 75006 PARIS

[HTTPS://ED393.SORBONNE-UNIVERSITE.FR/](https://ED393.SORBONNE-UNIVERSITE.FR/)

CONTACT : ED393@SORBONNE-UNIVERSITE.FR/TÉLÉPHONE : 01.44.27.24.35

La similarité sera paramétrable afin d'adapter l'analyse à la question clinique : proximité diagnostique, sévérité, dynamique du parcours ou combinaison de ces dimensions. Cette flexibilité reposera sur une pondération des composantes, ajustable par le clinicien ou apprise à partir d'exemples de regroupements pertinents.

4.4. Évaluation et validation

L'évaluation sera conduite selon un protocole multi-niveaux combinant validation intrinsèque des représentations, validation prédictive sur des événements cliniques observables et validation extrinsèque par jugement expert.

1. *Évaluation intrinsèque des représentations* : Les représentations seront évaluées via la qualité des clusters (indices de silhouette, Calinski-Harabasz), la stabilité des regroupements (bootstrap) et leur capacité à distinguer des sous-groupes cliniques connus (ex. lupus avec ou sans atteinte rénale).
2. *Validation prédictive* : Pour vérifier la pertinence clinique des représentations, leur capacité à prédire des événements futurs issus de l'EDS sera testée. Si elles capturent correctement la dynamique de la maladie et le recours aux soins, elles doivent améliorer la prédiction de marqueurs d'évolution.
Deux tâches serviront de proxy : 1/ Délai jusqu'au prochain contact avec le système de soins (consultation, urgences, réhospitalisation), modélisé par régression ou analyse de survie à partir du vecteur patient. 2/ Modification thérapeutique dans les 30 jours post-hospitalisation, variable binaire évaluée par classification (régression logistique, gradient boosting) avec AUC, comparée à des modèles fondés uniquement sur les données structurées.
Ces tâches, entièrement dérivables de l'EDS, constituent des indicateurs interprétables de la qualité informationnelle des représentations. L'absence de gain prédictif indiquerait une valeur ajoutée limitée.
3. *Validation clinique experte* : Un panel d'internistes et de spécialistes évaluera la cohérence médicale des cohortes, les comparera à des regroupements basés sur les seules données structurées (CIM-10, données administratives) et examinera des cas d'usage visant l'identification de sous-phénotypes.
4. *Évaluation comparative* : Les cohortes obtenues avec et sans intégration du parcours seront comparées afin de mesurer l'apport de cette dimension, selon l'homogénéité intra-cluster, la performance prédictive (mortalité, réhospitalisation), la concordance experte et le gain observé sur les tâches prédictives.

Bibliographie

- [1] Gérardin C, Mageau A, Mékinian A, Tannier X, Carrat F. Construction of Cohorts of Similar Patients From Automatic Extraction of Medical Concepts: Phenotype Extraction Study. *JMIR Med Inform.* 2022 Dec 19;10(12):e42379.
- [2] Remaki A, Ung J, Pages P, Wajsburt P, Liu E, Faure G, Petit-Jean T, Tannier X, Gérardin C. Improving Phenotyping of Patients With Immune-Mediated Inflammatory Diseases Through Automated Processing of Discharge Summaries. *JMIR Med Inform* 2025;13:e68704
- [3] Shmatko, A., Jung, A.W., Gaurav, K. et al. Learning the natural history of human disease with generative transformers. *Nature* 647, 248–256 (2025).
- [4] Choi E, Bahadori MT, Searles E, Coffey C, Thompson M, Bost J, Tejedor-Sojo J, Sun J. Multi-layer Representation Learning for Medical Concepts. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, San Francisco, 2016, pp. 1495-1504
- [5] Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *Proceedings of the 1st Machine Learning for Healthcare Conference (MLHC), JMLR Workshop and Conference Proceedings* 56, 2016, pp. 301-318.
- [6] Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, Zhu Y, Rahimi K, Salimi-Khorshidi G. BEHRT: Transformer for Electronic Health Records. *Scientific Reports*, 2020;10:7155
- [7] Suo Q, Ma F, Yuan G, Wu M, Ganiz A, Gao J. Measuring Patient Similarities via a Deep Architecture with Medical Concept Embedding. *IEEE International Conference on Data Mining (ICDM)*, 2018.
- [8] Rodrigues-Jr JF, Spadon G, Brandoli B, Amer-Yahia S. Real-world Patient Trajectory Prediction from Clinical Notes Using Artificial Neural Networks and UMLS-Based Extraction of Concepts. *Journal of Biomedical Informatics*, 2022;127:104004.
- [9] Zhao Y, Nair N, Engelman CD, Gagliano Taliun SA, Sleiman P, Hakonarson H, et al. Language-model-based patient embedding using electronic health records facilitates phenotyping, disease forecasting, and progression analysis. *npj Digital Medicine*, 2024.
- [10] Sharafoddini A, Dubin JA, Lee J. Patient Similarity in Prediction Models Based on Health Data: A Scoping Review. *JMIR Medical Informatics*, 2017;5(1):e7.
- [11] Allam A, Feuerriegel S, Rebber M, Krauthammer M. Evaluation of data processing pipelines on real-world electronic health records data for the purpose of measuring patient similarity. *BMC Medical Informatics and Decision Making*, 2023;23:117.

PRÉREQUIS, FORMATION : INFORMATIQUE MÉDICALE, DATA SCIENCE, IA, BIostatistiques ou discipline proche.
DOUBLE COMPÉTENCE SANTÉ + DATA UTILE (MÉDECINE, PHARMACIE, SANTÉ PUBLIQUE).
TRÈS BONNE MAÎTRISE DE PYTHON (PANDAS, SCIKIT-LEARN, PYTORCH OU TENSORFLOW).
EXPÉRIENCE SOLIDE EN MACHINE LEARNING, CLUSTERING, MÉTRIQUES DE SIMILARITÉ.
EXPÉRIENCE EN NLP, IDÉALEMENT SUR TEXTES CLINIQUES.

ÉCOLE DOCTORALE 393
CENTRE BIOMÉDICAL DES CORDELIERS
15, RUE DE L'ÉCOLE DE MÉDECINE 75006 PARIS
[HTTPS://ED393.SORBONNE-UNIVERSITE.FR/](https://ed393.sorbonne-universite.fr/)

CONTACT : ED393@SORBONNE-UNIVERSITE.FR / TÉLÉPHONE : 01.44.27.24.35