Chaire de recherche AP-HP – INRIA – CentraleSupélec

Projet Scientifique et Technique

« Données Massives, Intelligence Artificielle et Santé »

Table des matières

1 Introduction

La santé est identifiée comme l'un des axes prioritaires de la stratégie nationale pour le développement de l'intelligence artificielle (IA) dans le rapport remis par Cédric Villani au président de la République en mars 2018 (1).

Dans ce contexte, l'AP-HP souhaite développer, en partenariat avec Inria et CentraleSupélec, un pôle de recherche dont l'objectif est de lever les verrous technologiques limitant l'émergence de l'IA en santé. Pour donner à ce pôle de recherche une visibilité internationale et pour renforcer les travaux de recherche technologique autour des données de santé, la mise en place d'une chaire de recherche adossée à l'Entrepôt de Données de Santé de l'AP-HP (EDS) est stratégique. Fondée sur le développement de travaux d'intégration, d'enrichissement et d'analyse des données massives de santé, la chaire aura pour objectifs d'accroître le potentiel de l'EDS pour la recherche médicale et l'innovation en IA et de valider l'utilité clinique de l'exploitation des données massives de santé au travers de projets de recherche en IA exploitant les capacités de la plateforme méga données de l'AP-HP.

2 L'Entrepôt de Données de Santé de l'AP-HP

L'EDS intègre les données administratives et médicales de près de 9 millions de patients hospitalisés ou venus en consultation dans l'un des 39 établissements de l'AP-HP. L'exploitation des données de l'EDS permet de soutenir la recherche scientifique et l'innovation dans le domaine de la santé.

Sur le plan réglementaire, la constitution de l'EDS a été autorisée par la CNIL le 19 janvier 2017 (autorisation n°1980120) et une déclaration de conformité à la méthodologie de référence CNIL concernant les recherches n'impliquant pas la personne humaine a été réalisée.

Au vu des enjeux éthiques et sociétaux soulevés par la constitution de l'EDS, l'EDS fait l'objet d'une gouvernance spécifique associant des professionnels de santé, des chercheurs et des représentants de patients. Des règles de fonctionnement ont été élaborées et approuvées en septembre 2016 par la Commission Médicale d'Établissement, le Directoire et la Direction générale de l'AP-HP.

Un Comité Scientifique et Ethique (CSE), labellisé IRB (IRB00011591), a été créé en décembre 2016 avec pour mission principale l'évaluation scientifique, éthique et réglementaire des projets de recherche.

Une plateforme performante, dotée des capacités de stockage et d'analyse dimensionnées pour les usages attendus, a été mise en place et sécurisée sur les plans matériel, logiciel et organisationnel. Un portail dédié à l'analyse de données massives (i2b2/Jupyter) permet de construire des environnements de travail sécurisés pour chaque projet de recherche.

Les usages se développent progressivement. A ce jour plus de 50 projets ont été soumis à l'avis du CSE. Près d'un tiers de ces projets ont pour objectif de développer des algorithmes d'IA dans des domaines variés (imagerie médicale, réanimation, cancérologie, codage de l'activité médicale…).

Par ailleurs, afin de favoriser le développement d'applications d'IA, l'AP-HP se mobilise en organisant différents évènements :

- Le datathon « DAT-ICU » s'est tenu en janvier 2018, en lien avec le *MIT lab for Computational Physiology* et a réuni plus de 150 médecins et data scientists autour de l'infrastructure EDS et d'une base de données ouverte rassemblant les données de santé dé-identifiées de 40 000 patients admis en réanimation (base de données MIMIC) https://www.aphp.fr/contenu/datathon-dat-icu-intensive-care-unit-4-projets-innovants-selectionnes-lissue-de-48h-danalyse
- Des ateliers entre les professionnels de santé de l'AP-HP et les chercheurs de CentraleSupélec et d'Inria se sont déroulées en novembre 2018 puis entre mars et avril 2019 et ont réuni près de 200 participants.

3 Les ateliers AP-HP – Inria – CentraleSupélec ont vu émerger de nombreuses idées de développement d'applications d'IA

L'exploitation des données de l'EDS permet l'essor d'études en vie réelle dont l'importance a été soulignée par le rapport Polton-Béguard-Von Lennep (2) et ouvre des perspectives d'innovation et de transformation numérique dans de nombreux domaines médicaux. L'organisation d'ateliers a permis d'identifier de nombreux exemples de projets de recherche parmi lesquels :

- Dans le domaine du traitement des signaux, le développement de modèles prédictifs sur des signaux issus de capteurs collectant des données à l'hôpital ou à domicile offrirait des perspectives d'amélioration de la surveillance et de la prise en charge dans un contexte de médecine d'urgence, de réanimation mais également de médecine ambulatoire. En anesthésie-réanimation, il pourrait s'agir, par exemple, d'aider au pilotage de respirateurs artificiels chez des patients grâce à des méthodes de traitement automatique du signal et au développement de modèles hybrides associant des modèles numériques à des modèles d'apprentissage. En diabétologie, il pourrait s'agir de personnaliser l'insulinothérapie grâce à l'exploitation des données de glucomètres.
- Dans le domaine de l'imagerie médicale, des projets collaboratifs AP-HP – Inria de développement d'algorithmes d'IA sont déjà en cours afin, par exemple, d'identifier automatiquement les organes et leurs principales lésions au sein d'échographies abdominales, d'assister le radiologue lors de la classification automatisée d'examens TEP-TDM, et la détection et l'évaluation pronostique des cancers de la prostate. En anatomie pathologique, l'essor des lames virtuelles représente une opportunité très intéressante de développement de solutions d'aide au diagnostic ou à l'évaluation pronostique des lésions, notamment dans le cancer.
- Dans le domaine de la modélisation des organes et tissus, des perspectives intéressantes pourraient par exemple concerner la cancérologie (modélisation de la réponse cellulaire durant la radiothérapie), la néphrologie et la chirurgie hépatique. En néphrologie, la création d'un jumeau virtuel (modélisation anatomique et fonctionnelle du rein) permettrait d'améliorer la prédiction de la perte de fonctionnalité rénale au décours d'une chirurgie, de tester l'intérêt d'un geste endovasculaire sur une sténose des artères rénales, de prédire l'évolution d'une maladie génétique comme la polykystose hépatorénale, d'anticiper la possibilité ou non de faire des transplantations rénales. Dans la chirurgie du foie, la création d'un jumeau virtuel permettrait d'anticiper la possibilité ou non de faire des hépatectomies. Dans la chirurgie des malformations cranio-faciales, la modélisation multimodale de la croissance cranio-faciale permettrait de concevoir un outil de diagnostic et de phénotypage automatique des malformations cranio-faciales, avec une aide

à la décision opératoire.

- Dans le domaine de la modélisation du parcours patient, les enjeux concernent par exemple la prédiction de l'observance et de l'acceptabilité du traitement par le patient, la prédiction de l'errance diagnostique chez des patients atteints de maladies rares, la prédiction de la rupture du parcours de soins, la mise en place de solutions automatisées de triage des patients aux urgences.
- Dans le domaine de la biologie médicale et de la biologie des systèmes, il pourrait s'agir, par exemple, de développer des modèles de détection de la puberté précoce en endocrinologie pédiatrique.

4 Thématiques de recherche

Un ensemble de thèmes de recherche potentiels de la chaire ont été identifiés.

4.1 Thème 1 : Evaluation de l'anonymat des données

L'AP-HP doit être en mesure de mettre à disposition des jeux de données qui peuvent servir à la mise en place d'outils d'analyse par des acteurs extérieurs tout en ayant des garanties sur la préservation de l'anonymat des patients.
L'anonymisation stricte des données personnelles consiste à supprimer ou modifier toutes les informations identifiantes rendant impossible toute ré-identification des personnes. S'il est relativement facile de supprimer les informations identifiantes au sein de données de santé structurées, cela l'est beaucoup moins lorsque ces informations se trouvent au sein de données non structurées (compte rendus médicaux, images médicales). Par ailleurs, les données longitudinales (ou appariées) de patients, lorsqu'elles sont suffisamment profondes et variées, permettent le plus souvent une ré-identification des personnes par inférence. Dans l'état de l'art actuel, on parle ainsi, le plus souvent, de dé-identification ou de pseudonymisation des données de santé. Leur traitement et analyse nécessitent un encadrement juridique conforme au règlement européen pour la protection des données limitant de fait leur accessibilité à la communauté de chercheurs.

Les technologies d'IA sont particulièrement prometteuses dans le cadre de l'anonymisation des données non structurées. Néanmoins la richesse des données

disponibles dans l'EDS rend complexe le problème de l'évaluation de l'anonymat des données.

Une thématique de recherche fondamentale, pour la mise à disposition de données de l'EDS dans un contexte sécurisé, éthique et respectueux de la vie privée des patients, est la définition de critères mesurables de l'évaluation de l'anonymat (ou d'une procédure d'anonymisation) des données, ainsi que la définition des propriétés de ces critères. Ces critères sont un prérequis indispensable à la production d'algorithmes efficaces d'anonymisation et doivent fournir des indices pour l'analyse de risques de diffusion des données.

4.2 Thème 2 : Génération de données synthétiques

La génération de données synthétiques peut répondre aux besoins de travaux méthodologiques, pour peu que l'on améliore les propositions actuelles. En effet, il faut synthétiser des données qui portent de l'information (la simple "ressemblance" n'est pas suffisante) pour pouvoir apparaître comme réalistes, toujours avec des garanties d'anonymat. Cependant, cette génération doit du coup être spécifique au regard d'une tâche. Par ailleurs, la génération de données synthétiques est actuellement principalement développée sur des données tabulaires.

Un axe de recherche consisterait à développer des méthodes de génération de données synthétiques pour des catégories de données non tabulaires ce qui pose des défis majeurs à la frontière entre *machine learning* et *privacy*.

Les recherches en cours sur données de l'EDS, dont l'objectif vise à développer des outils d'IA pourraient servir à faire des "études miroirs" sur la synthèse de données. Une étude miroir consiste à :

1) Générer des données synthétiques à partir des données utilisées lors d'une recherche
2) Tester les modèles d'IA développé lors de la recherche sur les données synthétisées (simulées)
3) Evaluer les méthodes de génération de données synthétiques sur leur capacité à refléter les résultats initiaux

4.3 Thème 3 : Standardisation des données

L'interopérabilité sémantique permet le partage du sens de données qui, dès lors qu'elles sont représentées sous une forme interprétable par la machine, peuvent être échangées et utilisées au-delà de leur lieu de production. Sa mise en œuvre repose sur la définition d'un cadre d'interopérabilité définissant les modèles d'information et terminologies permettant de structurer et standardiser les données. S'il existe plusieurs organismes de standardisation dans le domaine de la santé (HL7, DICOM, CDISC…) et de nombreuses autres initiatives proposant des standards, deux cadres de structuration et de standardisation des données de santé émergent actuellement : le standard FHIR (Fast Healthcare Interoperability Resources, https://www.hl7.org/fhir/) et le modèle commun de données OMOP (Observational Medical Outcomes Partnership).

L'AP-HP met en œuvre une base de données au format OMOP et intègre les données de l'EDS correspondant au périmètre de ce modèle. Par ailleurs, l'AP-HP développe des web services FHIR permettant d'accéder aux données de l'EDS dans le périmètre des profils FHIR définis.

L'adoption croissante de ces deux standards représente une opportunité unique de développement d'outils d'analyse de données opérant à grande échelle. En effet, dans la mesure où toute requête peut être exécutée sur n'importe quel site sans modification, toute solution développée sur un site unique est généralisable à d'autres sites. Les données cliniques, même structurées, étant rarement nativement codées au sein des SIS, l'interopérabilité sémantique repose, en pratique, sur des solutions de standardisation de l'information clinique lors de l'échange ou de l'exploitation de cette information par un système tiers. En ce qui concerne les données non structurées (texte, image, signal), des solutions d'annotation - manuelle ou automatique - doivent être développées afin d'enrichir ces données et les rendre intelligibles.

Par ailleurs, le format RDF du Web sémantique permet de spécifier un modèle de données formel standard non spécifique du domaine de la santé. Ce format est utilisé depuis plus de 10 ans pour représenter et intégrer des données de santé et est présenté par certains travaux comme le meilleur langage universel candidat pour l'échange de données de santé, que ce format rend auto-descriptives. Une représentation formelle des données et des connaissances ouvre des perspectives de fouille de données s'appuyant sur un raisonnement exploitant cette représentation

formelle. A ce titre il est préconisé de représenter les données à échanger soit directement en RDF soit dans un format d'échange aligné de manière standardisée à RDF. Il s'agit donc de définir une représentation RDF de toute nouvelle norme d'échange de données de santé. Plusieurs travaux proposent des représentations RDF de ressources FHIR.

Un axe de recherche consisterait à étendre la standardisation des données de l'EDS selon les modèles OMOP et FHIR en développant et évaluant des solutions d'alignement terminologique multilingues. En perspective, il s'agirait également de proposer une représentation RDF des profils FHIR d'accès aux données de l'EDS et à valider cette représentation dans le cadre d'un ou plusieurs projets de recherche sur données de l'EDS.

4.4 Thème 4 : Intégration de données hétérogènes

Les données de santé proviennent de sources multiples, hétérogènes, souvent interconnectées et potentiellement de grande dimension (texte, image, signal…). Chaque source peut présenter une structure spécifique complexe. Les données requièrent donc une analyse intégrée permettant de tirer profit des complémentarités qui existent entre les différentes sources. Cependant, les méthodes traditionnelles, pour être utilisées, requièrent d'altérer leur organisation naturelle au risque de perdre l'information pertinente. Ainsi, le développement de méthodes statistiques d'analyse de données capables d'épouser les structures globales et spécifiques est essentiel. Cette thématique, à l'interface de la statistique et du *machine learning*, pourrait chercher à développer un cadre statistique et informatique unifié pour l'analyse de données multi-sources. Ce cadre statistique s'appuie sur l'analyse canonique généralisée (RGCCA). Les principaux défis liés à l'exploitation de ce type de données multi-sources comprennent :

1) Extraction de l'information pertinente
2) Visualisation de données multi-source ; réduction de la dimension par projection et/ou sélection de variables
3) Prédiction de classes, de sorties vectorielles à partir de données hétérogènes
4) Prédiction d'un type de données en fonction d'un ou plusieurs autres types de données

5) Reproductibilité de la recherche

4.5 Thème 5 : Extraction d'informations d'intérêt à partir de données non structurées

Les solutions d'extraction d'informations d'intérêt à partir de données non structurées présentes dans les dossiers patients informatisés varient avec la nature de la donnée (documents textuels, image, signal…).

L'objectif de ces solutions est de faciliter l'analyse et la mise à disposition de données non structurées pour la recherche et le développement d'applications innovantes (phénotypage automatique de patients, identification de patients similaires, représentations vectorielles de dossiers patients…).

Un axe de recherche consisterait à concevoir et développer des chaînes de traitement automatique des données non structurées.

1) Les documents textuels sont omniprésents dans les dossiers patients et contiennent des informations essentielles sur les patients. L'EDS dispose aujourd'hui de plus de 50 millions de documents textuels. De nombreuses approches statistiques de traitement automatique des langues ont été proposées récemment dans la littérature pour exploiter l'information clinique des documents textuels. La plupart de ces méthodes tombent dans la catégorie de l'apprentissage supervisé, c'est-à-dire qu'elles nécessitent la création d'un jeu de données annoté (par des experts humains) pour permettre l'entraînement d'un modèle statistique, qui ensuite pourra être appliqué sur de nouvelles données. Ces approches demandent donc un investissement de départ considérable, et sont difficiles à généraliser car les annotations manuelles sont généralement spécifiques à un domaine clinique particulier. C'est la raison pour laquelle ces approches n'ont été appliquées que sur des cas d'usage relativement limités. Plus récemment, des approches dites semi-supervisées ont été proposées, permettant de s'affranchir partiellement de l'étape d'annotation manuelle, la remplaçant tantôt par un mécanisme d'amorçage (définition d'exemples ou de termes très discriminants dans le domaine considéré, permettant de servir d'amorces à un processus itératif de sélection de documents), tantôt par un apprentissage dit actif, dans lequel les dossiers proposés à l'expert sont

choisis automatiquement de manière à minimiser le nombre de patients à explorer avant d'obtenir un jeu de données d'entraînement de bonne qualité. Ces progrès récents ont été exclusivement effectués sur des documents de langue anglaise. Or, cette langue est dotée d'outils de traitement et de ressources terminologiques bien supérieures aux autres langues, et les approches ne sont pas directement transposables au français. En français, les travaux sont nombreux sur les textes du domaine général, beaucoup moins sur le domaine biomédical et doivent donc se développer.

2) La détection automatique d'anomalie sur des signaux issus de capteurs n'est pas une tâche aisée et nécessite, au préalable, l'utilisation d'algorithmes de traitement du signal permettant l'extraction de caractéristiques pertinentes. Le développement d'algorithmes prédictifs basés sur ces caractéristiques requiert des ressources de calcul haute-performance.
3) Les enregistrements audio peuvent également être des sources d'informations précieuses dans le cadre de la mise en place de solutions d'aide à la décision médicale.

4.6 Thème 6 : Phénotypage à haut débit et Représentation du patient

De nombreux travaux de développement d'algorithmes ont été réalisés pour automatiser l'identification au sein d'entrepôts de données cliniques des patients éligibles à des essais cliniques ou à des prises en charge personnalisées.

La base de données PheKB (Phenotype KnowledgeBase) contient des définitions de phénotypes à base de règles intégrant des concepts codés (en utilisant SNOMED, LOINC, ICD10…) ainsi que les performances des requêtes correspondantes implémentées au sein des entrepôts de données de santé des institutions partenaires.

L'identification de phénotypes à grande échelle (*high-throughput phenotyping*, *next generation phenotyping*) nécessite des méthodes permettant l'exploitation conjointe des données structurées et non structurées.

Utiliser l'ensemble des informations et des connaissances disponibles nous semble être une condition nécessaire pour s'approcher des performances d'un expert humain. L'exploitation de données si hétérogènes est un défi mais des progrès récents dans le domaine de la représentation d'informations, notamment grâce aux réseaux de

neurones, montrent que l'on peut représenter de façon jointe tous les types de structure dans un même espace, permettant la mise en œuvre d'algorithmes sur un seul mode de représentation issu de multiples sources. Ce type d'approches a été appliqué par exemple à la représentation jointe d'images et de textes, de bases de connaissances et de textes, et très récemment de dossiers patients avec des données structurées et des données textuelles.

Des travaux récents en IA permettent de faire des progrès considérables en proposant des modèles de représentation des données structurées des dossiers patients. Cela permet d'en faciliter l'analyse et les applications comme l'identification de patients similaires et l'analyse de parcours de soins.

4.7 Thème 7 : Modèles hybrides

La difficulté du problème de modélisation vient de la grande variabilité des signaux entre patients, d'où l'idée de développer des modèles hybrides associant des modèles numériques à des modèles d'apprentissage permis grâce notamment à la simulation de grands volumes de données pour représenter tous types de patient.

4.8 Thème 8 : Qualification des jeux de données

Le développement de solutions d'exploration et d'analyse de données afin de répondre à une problématique médicale particulière requiert la construction de jeux de données adaptés à la recherche envisagée.

Un axe de recherche consisterait à définir les méthodes d'optimisation de la construction des échantillons de données à extraire de l'EDS (sous-population de patients, type, nature et volumétrie des données souhaitées), de leur qualification et de leur prétraitement/enrichissement.

La méthode d'optimisation de la construction des jeux de données sera validée dans le cadre de projets de recherche sur données de l'EDS.

Un cadre de certification de base de données ou d'échantillons de données de santé pourrait être proposé.

5 Références

1. Villani Cédric. Donner un sens à l'intelligence artificielle. Pour une stratégie nationale et européenne. [Internet]. 2018 mars [cité 2 juin 2018]. Report No.: ISBN 978-2-11-145708-9. Disponible sur : https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf
2. Bégaud B, Polton D. Les données de vie réelle, un enjeu majeur pour la qualité des soins et la régulation du système de santé. 2017 mai p. 105. Disponible sur : http://solidarite-sante.gouv.fr/IMG/pdf/rapport_donnees_de_vie_reelle_medicaments_mai_2017vf.pdf