# L'équipe-projet HeKA

Adrien Coulet, Stéphanie Allassonniere, François Angoulvant, Anita Burgun, Xiaoyi Chen, David Drummond, Nicolas Garcelon, Anne-Sophie Jannot, Sandrine Katsahian, Antoine Neuraz, et al.

AfIA
Association française
pour l'Intelligence Artificielle

# L'équipe-projet HeKA

*Équipe-projet HeKA / Équipe 22*
*Centre de Recherche des Cordeliers,*
*Inria Paris, Inserm et Université de Paris*
*team.inria.fr/heka*

**Adrien COULET,**
**Stéphanie ALLASSONNIERE,**
**François ANGOULVANT,**
**Anita BURGUN,**
**Xiaoyi CHEN,**
**David DRUMMOND,**
**Nicolas GARCELON,**
**Anne-Sophie JANNOT,**
**Sandrine KATSAHIAN,**
**Antoine NEURAZ,**
**Bastien RANCE,**
**Brigitte SABATIER,**
**Rosy TSOPRA,**
**Moreno URSINO,**
**Sarah ZOHAR**
*adrien.coulet@inria.fr,*
*anita.burgun@aphp.fr,*
*sarah.zohar@inserm.fr*

## Introduction

**HeKA** est une équipe-projet de recherche commune à Inria, l'Inserm et l'Université de Paris. Plus précisément, HeKA, dépend du Centre de Recherche des Cordeliers et du Centre Inria de Paris. En plus de deux chercheurs Inria et Inserm, HeKA est composé de chercheurs hospitalo-universitaires de l'AP-HP associés à des services de l'Hôpital Européen Georges Pompidou, l'Hôpital Necker et de l'Institut Imagine. Les thèmes de recherche de l'équipe sont l'**informatique médicale**, les **biostatistiques** et les **mathématiques appliquées** pour l'**aide à la décision clinique**. Le terme HeKA est à la fois une référence à la divité égyptienne de la médecine et un acronyme pour *Health data- and model- driven Knowledge Acquisition*. L'équipe HeKA fait suite à l'équipe 22 (*Information Sciences to support Personalized Medicine*) dirigée par Anita Burgun au Centre de Recherche des Corderliers (Inserm, Université de Paris). La responsable de HeKA est Sarah Zohar, elle est secondée par Adrien Coulet.

## Vers un système de santé apprenant

L'objectif partagé au sein de l'équipe est le développement de méthodes, modèles et outils pour un **système de santé apprenant** [8]. Ce paradigme, que nous développons en particulier pour les maladies rares dans le RHU C'IL-LICO [7, 1] tire parti des données générées au cours du soin pour apprendre de nouvelles connaissances, qui sont à leur tour utilisées pour guider la pratique clinique, de façon continue. Pour atteindre ce but, HeKA s'intéresse à 3 axes de recherche très liés : *(1)* l'extraction de connaissances à partir des données de santé et notamment le phénotypage profond ; *(2)* les approches stochastiques et supervisées pour l'aide à la décision ; et *(3)* les essais cliniques du futur et leur design, qui permettent l'évaluation des systèmes d'aide à la décision médicale. Les sections qui suivent décrivent brièvement ces axes et les illustrent avec des projets récents.

## Axe 1 : l'extraction de connaissances à partir de données de santé

Au sein de l'axe 1 nous nous attachons au développement de méthodes et d'outils pour tirer parti au mieux des données de patient·e·s, et ce malgré leur nature hétérogène et complexe (structurée *vs.* non-structurée, temporelle, incomplète, etc.). Nous étudions l'extraction et la transformation des données brutes, et notamment du texte clinique, en des descripteurs plus élaborés ou en des représentations apprises, qui facilitent le développement sous-jacent de systèmes d'aide à la décision ou de découverte de connaissances, comme ceux développés dans les Axes 2 et 3.

**Le projet Dr Warehouse** En collaboration avec l'institut Imagine et l'hôpital Necker-Enfants Malades, et pour faciliter la réutilisation des données hospitalières nous avons conçu l'entrepôt de données Dr Warehouse centré sur le document clinique [6]. A travers trois cas d'usage, nous avons tenté d'adresser les problématiques inhérentes aux données textuelles : *(i)* le recrutement de patient·e·s à travers un moteur de recherche adapté aux données textuelles (notamment en traitant négation et antécédents familiaux), *(ii)* le phénotypage automatisé à partir des données textuelles et *(iii)* l'aide au diagnostic par similarité entre patient·e·s, basée sur le texte [5]. Ce projet a fait l'objet de la création de la start-up Codoc, en charge de l'installation de Dr Warehouse dans d'autres hôpitaux (licence *open source*).

**Les projets TALONCO et TALREP** Il s'agit de deux projets d'extraction de connaissances, notamment à partir de textes en lien avec l'oncologie et le programme de recherche CARPEM (*Cancer Research and Personalized Medicine*) de l'AP-HP et Université de Paris (http://www.carpem.fr). **TALONCO** vise à l'extraction d'un ensemble d'informations (localisation, classification histologique, stade, etc.) pour la constitution automatique de "fiches de synthèses" de patient·e·s qui facilitent les réunions de concertations pluridisciplinaires en oncologie. **TALREP** vise quant à lui à l'extraction, à partir des dossiers patient·e·s informatisés (DPI), de la réponse clinique aux chimiothérapies (réponse positive, absence de réponse, effets indésirables). Cette première étape indispensable, nous permettra ensuite de développer des modèles prédictifs de la réponse à de tels traitements.

**Les projets PractiKPharma et PyMedExt** PractiKPharma est un projet ANR qui vise la comparaison des connaissances synthétisées dans l'état de l'art (bases de données expertes et littérature) avec celles qui peuvent être extraites des DPI [2]. Nous avons développé pour cela des méthodes pour assurer la qualité et la reproductibilité de l'extraction de connaissances à partir des DPI, notamment en proposant la réutilisation de librairies standards et indépendantes de l'environnement [3]. Le besoin de manipuler le texte clinique dans PractiKPharma et dans d'autres projets nous a amené à développer la librairie libre **PyMedExt** qui facilite leur transformation, échange et annotation [4].

## Axe 2 : les approches stochastiques et supervisées pour l'aide à la décision

L'axe 2 vise à proposer des méthodes originales d'apprentissage statistique ou automatique, notamment dans le contexte particulier de jeux de données de grande dimension, mais à faible effectif. En effet, même lorsque l'on a la chance d'avoir des données pour un grand nombre de patient·e·s, la sélection des cas pertinents dans un objectif d'aide à la décision diagnostique ou thérapeutique réduit

drastiquement le nombre d'individus. Dans ce contexte, nous prêtons une attention particulière à la modélisation du parcours de soin (ou trajectoire) des patient·e·s, ce qui intègre une dimension temporelle au sein de nos modèles. Enfin, la génération de données synthétiques et leur prise en compte dans nos modèles doit permettre de mieux les évaluer en augmentant artificiellement notre échantillonage.

**Le projet AntibioHelp®** Dans le cadre du projet RaMiPA ("Raisonner pour Mieux Prescrire les Antibiotiques"), financé par l'ANSM, nous avons développé AntibioHelp®, un système d'aide à la décision en antibiothérapie empirique. Ce système est capable de retrouver les antibiotiques recommandés dans les guides de bonnes pratiques cliniques, à partir des propriétés pondérées des antibiotiques. AntibioHelp® fournit des recommandations de prescription, notamment pour des situations rares ou non décrites dans les guides [9]. Afin de personnaliser les recommandations, nous travaillons à la mise en place de flux de données à partir d'un entrepôt de données cliniques, dans l'idée de générer, pour un·e patient·e donné·e, un score prédictif d'efficacité pour chaque antibiotique.

**Le projet Lights** Nous développons "Lights", un modèle de survie qui intègre les variables longitudinales de grande dimension en petit effectif dans le cadre du projet InCa "Thérapies personnalisées en oncologie pour les cancers métastatiques dans les populations asiatiques et caucasiennes : étude de transition réutilisant les dossiers patients informatisés". Lights permet d'identifier automatiquement les variables prédictives du pronostic à partir de données longitudinales de grande dimension, ainsi que des données expertes. Autrement dit, Lights permet d'identifier au cours d'un parcours de soin, les éléments les plus importants pour le pronostic des patient·e·s. Le développement théorique de cette méthode a déjà fait l'objet de deux communications orales, dont l'une au *Joint Statistical Meetings (JSM) - American Statistical Association* en 2020.

**Le projet Data Augmentation** Deux problèmes récurrents lorsque l'on travaille avec des données médicales sont le petit nombre d'exemples et leur grande dimension. En utilisant des *Variational Auto Encoders* (VAE) dont l'espace latent est muni d'une structure de variété riemanienne, nous pouvons résoudre les deux problèmes en un seul modèle génératif. Le VAE permet de réduire la dimension des données. La structure riemanienne permet d'apprendre une loi de probabilité sur l'espace latent qui permet de gérer de nouvelles données beaucoup plus informatives que les alternatives de la littérature. Ce travail fait l'objet d'une thèse de mathématiques appliquées débutée en 2020, mêlant statistiques computationelles et géométrie riemanienne.

## Axe 3 : les essais cliniques du futur et leur design

Le premier objectif de cet axe est de proposer des méthodes d'évaluation pour des outils logiciels en tant que dispositifs médicaux, et dans ce cadre de proposer des designs d'essais cliniques qui soient adaptés aux méthodes d'apprentissage continu. Son second objectif consiste à développer des modèles statistiques ou d'apprentissage automatique qui participent à la construction d'essais cliniques à partir de sources de connaissances extérieures comme des modèles de maladies, des modèles précliniques, des données expertes, des DPI, des patient·e·s synthétiques. Il s'agit dans ce cas d'optimiser les futurs essais cliniques pour faciliter l'acquisition de nouvelles connaissances biomédicales.

**Les projets européens ITFoC et PeCan** Dans le cadre des projets européens Flag-Era ITFoC *(Information Technology for the Future of Cancer Treatment)* et Era PerMed PeCan *(Parametrisation of large scale cancer models for personalised therapy of triple negative breast)*, nous développons une méthodologie de validation précoce des algorithmes d'IA à partir des données de vie réelle, avant leur implémentation dans le processus de soins. A partir de cette méthodologie, nous allons mettre en place un environnement d'expérimentation pour valider les algorithmes prédictifs de réponse au traitement des patientes atteintes de cancer du sein triple négatif. Cet environnement sera construit en collaboration avec le CEA, l'institut Curie, le CHU de Rennes, et le centre Unicancer Eugène Marquis.

**Le projet collaboratif européen FAIR** Dans le cadre du projet européen FAIR *(Flagellin Aerosol therapy as an Immunomodulatory adjunct to the antibiotic treatment of drug-Resistant bacterial pneumonia*, https://fair-flagellin.eu/), nous développons une plateforme de simulation translationnelle prenant en compte des modèles pré-cliniques (cellulaires et animaux) ainsi que des données expertes afin de construire de manière séquentielle un modèle physiologique. Cette approche de modélisation, par extrapolation et apprentissage par transfert, intègre différentes sources d'informations et va permettre de mettre en évidence le design optimal pour les essais cliniques évaluant les effets de la Flagellin chez l'humain.

En conclusion, HeKA est une équipe interdisciplinaire, en lien étroit avec le monde hospitalier, qui vise des contributions méthodologiques en science des données, capables d'impacter la pratique clinique.

## Références

[1] Xiaoyi Chen *et al.* Phenotypic similarity for rare disease : Ciliopathy diagnoses and subtyping. *Journal of Biomedical Informatics*, 100 :103308, 2019.

[2] Adrien Coulet and Malika Smaïl-Tabbone. Mining Electronic Health Records to Validate Knowledge in Pharmacogenomics. *ERCIM News*, 104 :56, January 2016.

[3] William Digan *et al.* Can reproducibility be improved in clinical natural language processing ? A study of 7 clinical NLP suites. *Journal of the American Medical Informatics Association*, December 2020.

[4] William Digan *et al.* Pymedext - a library to process clinical text : https://github.com/equipe22/pymedext_core/, 2020.

[5] Nicolas Garcelon *et al.* Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse : Dr. warehouse and the needle in the needle stack. *Journal of Biomedical Informatics*, 73 :51–61, 2017.

[6] Nicolas Garcelon *et al.* A clinician friendly data warehouse oriented toward narrative reports : Dr. warehouse. *Journal of Biomedical Informatics*, 80 :52–63, 2018.

[7] Nicolas Garcelon *et al.* Electronic health records for the diagnosis of rare diseases. *Kidney International*, 97(4) :676–686, 2020.

[8] Leigh Anne Olsen *et al.* The learning healthcare system : workshop summary, 2007.

[9] Rosy Tsopra *et al.* Helping GPs to extrapolate guideline recommendations to patients for whom there are no explicit recommendations, through the visualization of drug properties. The example of AntibioHelp® in bacterial diseases. *Journal of the American Medical Informatics Association*, 26(10) :1010–1019, 05 2019.