OFFRE DE STAGE niveau M2 / Ingénieur

Repérage et Phénotypage Automatique des Maladies Rénales Rares

11 septembre 2025

INFORMATIONS PRATIQUES

- Début : A partir de janvier 2026
- Durée : 6 mois
- Lieu : Centre de référence Maladies Rénales Rares, Brest
- Date limite : Si cette annonce est en ligne, le recrutement est ouvert

CONTEXTE

Le Centre de Référence Maladies Rénales Rares de Brest (CRMRB) a été labellisé en 2023 dans le cadre du plan national maladies rares. Il est rattaché au Centre de Référence des Maladies Rénales Héréditaires de l'Enfant et de l'Adulte (MARHEA). Le CRMRB suit environ 900 patients, pour une centaine de maladies rares. La prise en soin de ces patients nécessite une connaissance approfondie de la néphropathie d'origine afin de proposer un traitement adapté et d'anticiper son évolution.

En pratique, 20 à 40 % des patients suivis en néphrologie restent sans diagnostic étiologique clair et une proportion significative d'entre eux sont porteurs de maladies rénales rares. L'abondance des informations dans les dossiers patients et la diversité des maladies rénales rares font de l'intelligence artificielle un outil pertinent pour guider le diagnostic.

La faible fréquence des maladies rares et le diagnostic par défaut posé chez certains patients rendent difficile l'entrainement d'un algorithme d'intelligence artificielle spécialisé (manque de puissance, difficulté à généraliser, biais induit par le diagnostic posé). C'est pourquoi nous proposons dans le cadre du projet PHARE (Phénotypage Automatique des maladies REnales rares) d'utiliser des algorithmes de Natural Language Processing (NLP) pré-entrainés pour phénotyper notre base de données et ainsi identifier les patients porteurs de maladies rares.

DESCRIPTION DU STAGE

L'objectif principal du stage est d'identifier les patients porteurs de maladies rénales rares à partir des dossiers médicaux des patients du service de néphrologie à l'aide du Natural Language Processing (NLP).

Etat de l'art

Les comptes rendus cliniques sont riches en informations non structurées. Afin d'exploiter pleinement ces données, le NLP permet d'annoter automatiquement tout élément médical pertinent comme les symptômes, les pathologies et les antécédents. Cette tâche, connue sous le nom de reconnaissance d'entités nommées (NER), repose aujourd'hui sur des modèles spécialisés, notamment BioBERT ou en corpus français DrBERT et CamemBERT-bio. Par ailleurs, les grands modèles de langage (LLM) open source comme LLaMA 3 ou Mixtral montrent des capacités prometteuses en NER. Leur polyvalence, liée à leur entraînement multi-tâches, facilite leur adaptation à des domaines variés comme le médical.

Les informations extraites des comptes rendus présentent une forte variabilité lexicale. Une étape de normalisation est essentielle pour permettre l'interprétation des phénotypes cliniques. L'entity linking répond à cet enjeu en associant les entités extraites à des concepts standardisés issus d'ontologies médicales telles que UMLS, HPO ou Orphanet. Les approches d'entity linking s'étendent des méthodes associant les représentations vectorielles des entités extraites à celles des bases de connaissance (Phenotagger, RAG-HPO, SciSpaCy), à des modèles génératifs de terminologies standardisées (Pheno-GPT2 et PhenoBCBERT).

Tâches

1. Réalisation d'une veille sur les méthodes de NLP adaptées à l'extraction d'entités médicales et leur association aux bases de connaissances, en se concentrant sur les approches adaptées au français médical.
2. Participation à l'intégration du dossier patient de néphrologie au lac de données du Centre de Données Cliniques de Brest.
3. Pré-traitement des différentes sources de données.
4. Implémentation de plusieurs méthodes d'extraction d'entités et évaluation de leurs performances sur des patients atteints de maladies rares dont le phénotype est connu.
5. Expérimentation de plusieurs méthodes d'entity linking pour lier les entités aux concepts UMLS/HPO et évaluation de leur pertinence.
6. Proposition de pathologies probables à partir des descriptions phénotypiques. La performance de la méthode proposée sera évaluée par comparaison aux diagnostics de maladies rénales rares déjà posés.

Le stage se déroulera au Centre de référence des Maladies Rénales Rares à Brest. Le.a stagiaire bénéficiera de l'expertise du centre de référence, du service de néphrologie, de l'unité de recherche clinique et du centre de données cliniques du CHU de Brest.

PROFIL RECHERCHÉ

— Connaissances approfondies de **Python**, **SQL** et des librairies de traitement de données et d'apprentissage automatique comme Polars, Pandas, Scikit-learn, Tensorflow ou PyTorch.
— Capacité à lire des **articles scientifiques** en anglais.
— Bonnes pratiques de programmation et documentation pour un travail reproductible et collaboratif.
— Esprit d'initiative, rigueur scientifique et capacité à travailler en autonomie.

CANDIDATURE

Intéressé.e ? Adressez-nous votre CV par mail en précisant « Stage Projet PHARE » dans l'objet à :

nephrogenetique@chu-brest.fr