Université de Strasbourg

Corentin MEYER

École Doctorale des Sciences de la Vie et de la Santé STRASBOURG

Méthodes d'exploitation de données multimodales de patients par intelligence artificielle : application aux myopathies congénitales

Résumé

La croissance exponentielle des données générées par le séquençage, l'imagerie et les dossiers médicaux électroniques met en évidence le besoin d'outils pour l'exploitation des données biomédicales multimodales. L'objectif de ma thèse est de développer des méthodes basées sur l'intelligence artificielle pour explorer les données de patients atteints de myopathies congénitales, une famille de maladies génétiques rares difficiles à diagnostiquer. Dans un premier temps j'ai développé **IMPatienT** (*Integrated digital Multimodal PATIENt daTa*), une base de données permettant d'annoter et d'explorer les rapports et les images histologiques des patients. Ensuite j'ai développé **NLMyo** (*Natural Language Myopathies*), un outil basé sur les modèles linguistiques de grande taille comme *GPT-3.5* et *LLaMA*. **NLMyo** permet d'anonymiser et d'extraire de l'information de comptes rendus médicaux, de faire de l'aide au diagnostic et de créer un moteur de recherche de patients. Enfin, j'ai développé **MyoQuant** un outil utilisant des modèles d'IA pour quantifier automatiquement des marqueurs pathologiques dans les biopsies de fibres musculaires. **IMPatient**, **NLMyo** et **MyoQuant** sont disponibles de manière *open-source* et en version de démonstration en ligne.

Mots-clés : IA, quantification d'images, apprentissage profond, NLP, LLMs, myopathie congénitale, données biomédicales, rapports médicaux en texte libre, histologie, médecine translationnelle

Summary

The exponential growth of data generated by sequencing, imaging, and electronic medical records highlights the need for tools to exploit multimodal biomedical data. The objective of my thesis is to develop methods based on artificial intelligence to explore data from patients with congenital myopathies, a family of rare genetic diseases difficult to diagnose. First, I developed **IMPatienT** (Integrated digital Multimodal PATIENt daTa), a database for annotating and exploring patient reports and histological images. Then I developed **NLMyo** (Natural Language Myopathies), a tool based on large language models such as *GPT-3.5* and *LLaMA*. **NLMyo** allows to anonymize and extract information from medical reports, to do diagnostic assistance and to create a patient search engine. Finally, I developed **MyoQuant** a tool using AI models to automatically quantify pathological markers in muscle fiber biopsies. **IMPatient**, **NLMyo** and **MyoQuant** are available as open-source and online demo versions.

Keywords: AI, image quantification, deep learning, NLP, LLMs, congenital myopathy, biomedical data, free text medical reports, histology, translational medicine