CLIX - a configurable framework for CLinical Information eXtraction

Motivation

The secondary use of real-world clinical data holds the opportunity to enhance disease understanding and improve clinical decision-making. For example, electronic health records (EHR) can define cohorts, stratify patients, derive patient trajectories, and build clinical predictive models. Traditionally, the data used for these tasks are those that are stored in a structured format. However, a significant proportion of the medically useful information is stored in unstructured (clinical) notes. Leveraging this information resource promises to contribute to an improved performance of these approaches. Specifically, extracting data regarding symptoms, clinical progress or disease staging is a laborious undertaking that requires considerable engineering efforts and has a high likelihood of being non-reproducible.

Thus, in this master project, you will develop a pipeline to enable easy extraction of information from clinical notes. The pipeline will complement FIBER, a framework to automatically extract and preprocess structured EHR data.[1] It may further integrate with the AIR.MS (Artificial Intelligence-Ready Mount Sinai) platform under development at the Hasso Plattner Institute for Digital Health at Mount Sinai.[2] Using natural language processing (NLP) methods, the unstructured text will be automatically transformed into structured data for subsequent analyses (e.g., machine learning, predictive modeling, phenotyping) (Figure 1).

You will evaluate the pipeline on real-world clinical use cases in the context of kidney disease and Inflammatory Bowel Disease (IBD). A primary scientific output of the project will be an investigation of the impact on predictive performance in these use cases provided through the integration of information derived through NLP. The underlying data will come from the MSHS, one of the largest healthcare networks in the Greater New York area.

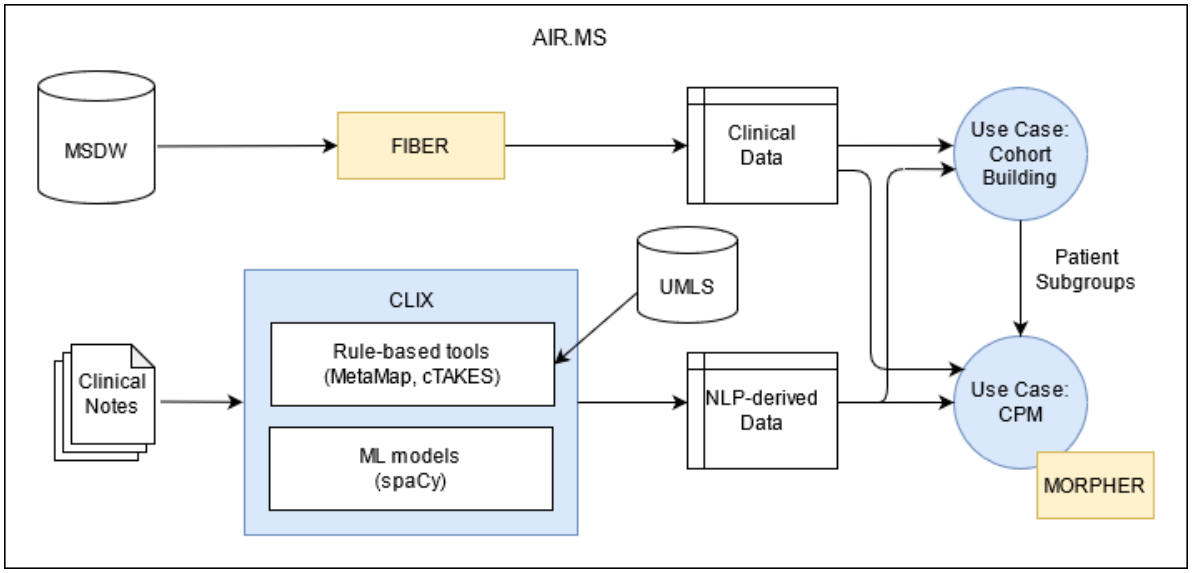

Figure 1: *Extraction of structured EHR data from the Mount Sinai Data Warehouse (MSDW) and clinical notes to build cohorts and clinical predictive models (CPM) within the FIBER framework, linkable with the AI Ready Mount Sinai (AIR.MS) platform. Existing components are highlighted in yellow, components to be developed in blue.*

[1] https://github.com/hpi-dhc/fiber
[2] https://www.hpims.org/research-projects/airms-ai-ready-mount-sinai

Methodology

CLinical Information eXtraction (CLIX) framework

The developed text processing pipeline within the CLIX framework will be designed to provide easy access and a unified wrapper interface to validated, best-of-breed rule-based NLP tools, like Apache cTakes or MetaMap, for common clinical information extraction tasks, such as section classification, named entity recognition and normalization, or detection of negation and other modifiers. The pipeline should be configurable with domain-specific ontologies from the Unified Medical Language System (UMLS) and semantic types of interest to the use case [1]. Moreover, the framework should allow the integration of custom, specialized deep-learning-based NLP models, e.g., based on the spaCy library.

Use Case: Cohort Building / Definition and Extraction of (Sub-)phenotypes

IBD is a relapsing immune-mediated disorder with heterogeneous clinical presentation. One aim of this project is to automatically apply the Montreal Classification system [2] to analyze and describe the MSHS IBD cohort and extract labels for clinical predictive modeling tasks. In a second use case, we would like to use the extracted information from clinical notes to determine the cause of Acute Kidney Injuries (AKI) as well as sub-phenotype AKI patients based upon extracted symptoms and imaging findings.

Use Case: Clinical Predictive Modelling

You will build clinical predictive models using features from structured/unstructured EHR data alone and in combination, for instance to forecast the progression of disease. Predictive models could be built from scratch or using the MORPHER toolkit.[3]

Project Goals

- Build and validate an NLP-based pipeline for information extraction from clinical notes
- Build a Python interface (Jupyter Notebook) for researchers
- Define patient (sub-)cohorts (e.g. based on the Montreal Classification) using structured and unstructured EHR data
- Track disease development over time
- Build clinical predictive models (e.g. for disease progression) using structured EHR data + extracted features from unstructured clinical notes and evaluate the performance differences compared to traditional models using just structured EHR data

What You Will Learn

- Building, validating, and documenting a pipeline in python
- Obtaining a deep understanding of working with domain-specific textual data and how to apply NLP methods to extract meaningful information
- Working with large scale real-world health data
- Understanding the medical background of the IBD and AKI use cases
- Project management and teamwork skills in an international and multidisciplinary team

[3] https://github.com/hpi-dhc/morpher-toolkit

About You

You should be interested in working in a multidisciplinary team with backgrounds ranging from computer science to bioinformatics and medicine, in exploring real-world data and using unstructured clinical notes to extract information that will be used for patient classification and/or predictive clinical modeling. You should be able to apply or be willing to learn the following skills:

- Text processing, Named Entity Recognition
- Python programming, working with Jupyter notebooks
- Working on a virtual machine within a secured IT infrastructure
- Developing a data (pre-)processing pipeline
- Machine learning fundamentals: predictive modeling, clustering
- Data visualization
- Medical domain knowledge

Contact

HPI Potsdam Team

Susanne Ibing, M.Sc.
Research Assistant, PhD candidate
Phone: +49 331 5509 3938
E-Mail: susanne.ibing@hpi.de

Florian Borchert, M.Sc.
Research Assistant, PhD candidate
Phone: +49-(0)331 5509-4839
E-Mail: florian.borchert@hpi.de

Prof. Dr. Erwin Böttinger
Head of DHC and Professor for Digital Health - Personalized Medicine
E-Mail: erwin.boettinger@hpi.de

Mount Sinai Team

Girish Nadkarni, MD
Professor Medicine
The Mount Sinai Hospital
girish.nadkarni@mountsinai.org

Ryan Ungaro, MD
Asst. Professor Medicine
The Mount Sinai Hospital
Email: ryan.ungaro@mssm.edu

References

[1] Demner-Fushman, Dina, Noémie Elhadad, and Carol Friedman. "Natural language processing for health-related texts." *Biomedical Informatics*. Springer, Cham, 2021. 241-272.

[2] Silverberg, Mark S., et al. "Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: report of a Working Party of the 2005 Montreal World Congress of Gastroenterology." *Canadian journal of gastroenterology* 19.Suppl A (2005): 5A-36A.