Survey of NLP Techniques for Detecting Alport Syndrome in Electronic Health Records

Vilnius University

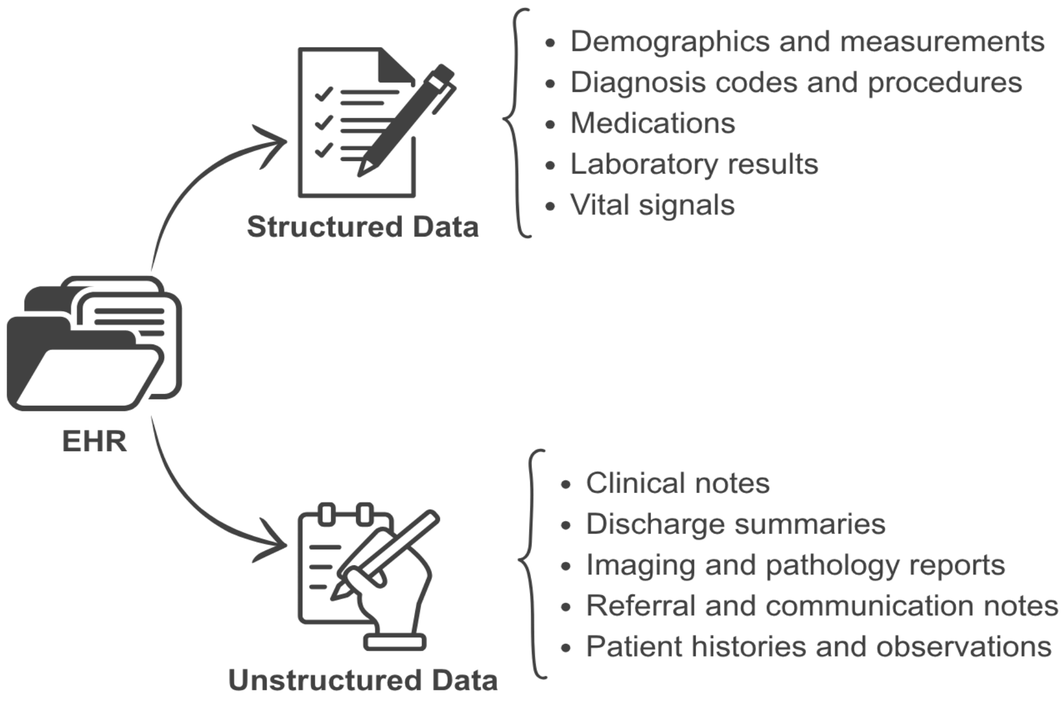
Gabrielė Skirmantaitė, Gražina Korvel

Vilnius university, Faculty of Mathematics and Informatics

1 Introduction

Rare diseases, such as Alport syndrome, present complex diagnostic challenges due to their low prevalence and heterogeneous clinical manifestations, often resulting in delayed recognition and treatment. The widespread adoption of electronic health records (EHRs) has facilitated the aggregation of extensive patient information, encompassing both structured datasets and detailed unstructured clinical documentation. While structured components are amenable to conventional computational analysis, a substantial proportion of clinically pertinent information is retained within free-text narratives. Natural language processing (NLP) methodologies offer effective approaches for extracting meaningful insights from these unstructured texts, thereby supporting earlier detection, precise stratification, and informed decision-making in the management of rare diseases.

2 EHR Data Structure



3 Challenges in Clinical NLP

Applying NLP to unstructured EHR text offers clear potential for supporting clinical analysis, but its use remains constrained by factors that complicate consistent and reliable extraction of meaningful information.

- Lack of Annotated Data
- Data Quality and Consistency
- Domain-Specific Language
- Data Privacy and Security
- Lack of Case Studies
- Generalizability and Portability
- Context and Semantic Understanding

4 Data Preprocessing

Preparing unstructured EHR data is a crucial first step in applying natural language processing to rare disease detection. By converting fragmented clinical notes into organized, machine-readable formats, preprocessing helps ensure that subsequent analyses are accurate and dependable.

5 Clinical NLP Tasks

Information Extraction

- Named Entity Recognition
- Relation Extraction
- Temporal Information Extraction
- Phenotype Extraction

Classification and Prediction

- Document Classification
- Diagnosis/Outcome Prediction
- Risk Prediction
- Patient Stratification

Retrieval and Summarization

- Information Retrieval
- Clinical Text Summarization
- Question Answering

6 NLP Models and Methods

Model family	Use Cases
Rule-Based Methods	Expert-defined rules, regular expressions, dictionaries for entity extraction and text classification.
Classical Machine Learning	SVM, Logistic Regression, Random Forest and CRF for classification, entity recognition and document categorization.
Deep Learning and Transformers	CNNs, RNNs, BERT, ClinicalBERT, BioBERT for entity recognition, phenotyping, information extraction.