AI in Healthcare

Final Project Proposal

Emma SCHAPIRA

Elrich MIRANDA

Tonghan WEN

# Problem statement and tasks of our project

The problem we are investigating is the extraction and classification of medical decisions from discharge summaries.

In clinical practice, physicians make a vast array of decisions related to patient care, ranging from diagnostic conclusions to treatment plans and follow-up recommendations. These decisions are documented in clinical notes, but the unstructured nature of these texts makes it difficult to systematically analyze or retrieve this information. Our research focuses on automating the extraction of these medical decisions, enabling better understanding and structuring of clinical reasoning.

The task we are undertaking is crucial. We are introducing MedDec, a new dataset specifically designed for extracting and categorizing medical decisions within discharge summaries. This dataset consists of 451 discharge summaries containing more than 54,000 sentences, annotated based on the Decision Identification and Classification Taxonomy for Use in Medicine (DICTUM). DICTUM classifies medical decisions into ten categories, including drug-related decisions, therapeutic procedures, evaluation of test results, contact-related decisions, and legal/insurance-related decisions.

The complexity of our task lies in two fundamental challenges :

* First, we need to detect the exact spans of text that constitute a medical decision within a clinical note. These spans are often embedded in long, complex narratives that describe a patient’s medical history, symptoms, and treatment journey.
* Second, we must classify each decision according to the predefined DICTUM categories, ensuring that we capture the appropriate meaning despite variations in wording and medical terminology.

Also, the relevance of this research is multifaceted. First, it has the potential to transform clinical decision-making by allowing hospitals and healthcare providers to analyze medical decisions at scale. Extracting these decisions automatically from thousands of clinical notes would help identify trends, detect deviations from best practices, and enhance medical guidelines. Second, this work contributes to the advancement of BioNLP (Biomedical Natural Language Processing) by addressing a problem that has remained largely unexplored: the structured extraction of medical decisions. Unlike other NLP tasks, which focus on extracting diagnoses or symptoms, this research targets the reasoning and judgment behind clinical decisions, making it a critical step toward explainable AI in healthcare.

Beyond clinical applications, this research could also impact health policy and insurance regulations. By analyzing large-scale patterns in medical decision-making, policymakers could assess how treatment plans align with guidelines, identify potential inefficiencies in healthcare, and evaluate the impact of regulations on clinical practice.

In terms of technical challenges, the study identifies several issues that need to be addressed in future work. One of the main difficulties is class imbalance—certain categories, such as drug-related decisions, appear frequently in the dataset, while others, such as legal/insurance-related decisions, are much rarer. This creates a bias in model learning, making it harder to accurately predict underrepresented decision types. Additionally, handling long clinical texts remains a challenge, as most NLP models are limited to processing only 512 tokens at a time. To overcome this, we implement sequence chunking strategies, where clinical notes are divided into smaller segments before being analyzed.

Based on a brief preliminary reading that we carried out, we learnt that LLMs struggle with the task of extracting key decisions from medical reports when used without fine-tuning and domain-specific training, which makes the purpose of this project quite relevant.

# Data Overview

The dataset : MedDec (a newly developed resource for extracting medical decisions from clinical notes)

The data is sourced from : MIMIC-III (a publicly available database containing de-identified clinical records from intensive care units.)

We have in the dataset : 451 discharge summaries, covering more than 54,000 sentences.

Each summary contains multiple medical decisions categorized into ten different decision types : ex : drug-related decisions, therapeutic procedures, evaluation of test results, and legal/insurance-related decisions.

These decisions were manually annotated by two expert annotators, with disagreements resolved by a third senior annotator.

In our project we will focus on extracting and classifying medical decisions within discharge summaries using NLP and LLMs models. The specific aspects of the data that will be utilized include:

1. Decision spans and their categories for model training and evaluation.

2. Annotated summaries to analyze decision-making patterns across different diseases.

3. Metadata on patient demographics (sex, race, English proficiency) to explore potential biases.

There are multiple potentials obstacles such as :

- Class imbalance

- Long clinical texts

- Ambiguity in decision classification

# Planned Methodology

We plan to use MedDec: A Dataset for Extracting Medical Decisions from Discharge Summaries (Elgaar et al., 2024) as a baseline guide to extract key medical decisions from medical reports and decision summaries via span-detection methods.

There are several approaches that are under consideration:

1. Fine-tuning BERT-based models such as BioClinicalBERT, RoBERTa and ELECTRA and implementing a multi-class classification where each token is assigned one of several predefined medical decision categories
2. Evaluate zero-shot and one-shot LLM prompting methods to extract structured decisions from summaries and investigate their ability to g across different phenotypes of diseases
3. Combining rule-based filtering with deep learning models and then use unsupervised clustering to identify decision-making patterns across patient groups

The reference paper contains annotated medical decisions across ten categories and proposes a sequence chunking method that processes lengthy discharge summaries by breaking them into manageable 512-token segments. We can leverage the annotation schema and chunking approach into our pipeline to further fine-tune various LLMs for our task.

Besides MedDec, there are other studies that have developed annotated medical datasets which we can use to test our models, such as CLIP (Mullen-bach et al., 2021) which is a dataset of MIMIC-III summaries annotated with seven types of action items and MDACE (Cheng et. al., 2023), a dataset of clinical notes annotated with ICD codes (the International Classification of Diseases). This step can be carried out provided we gain access to these annotations.

Here are some potential improvements over existing work related to this task that we could study and try to implement:

* Improving the detection of span boundaries by incorporating unified medical language system (UMLS)
* Fine-tune models for specialized medical domains (e.g., cardiology, oncology)
* To reduce bias in the results towards specific disease phenotypes, apply data augmentation for low-represented categories.

# Evaluation of the Results

In general, we plan to evaluate our model’s performance across two main dimensions: token level and span level, as well as within the match to each decision category. We shall apply standard evaluation metrics for assessment: accuracy for token-level evaluation and F1 score for span-level and decision category evaluations.

* Span Exact Match: Measures the model’s ability to predict spans with both correct boundaries and categories.
* Token Accuracy: Assesses the prediction of decision categories at the token level. This metric is more flexible, allowing partial overlaps with true spans.
* Decision Categories: Medical decisions can be grouped into distinct categories, facilitating structured organization and enhancing patient comprehension of medical transcripts. Evaluating the model’s performance in each category helps determine its capability to accurately classify medical decisions.

In order to evaluate the model's capacity of generalization, we also plan to compare F1 score performance of span detection at seen phenotype with unseen phenotype. This comparison could allow us to check when a novel disease appears, whether our model could still show ideal performance in classifying their categorization with high accuracy.