MEDICAL DIAGNOSIS ENGINE

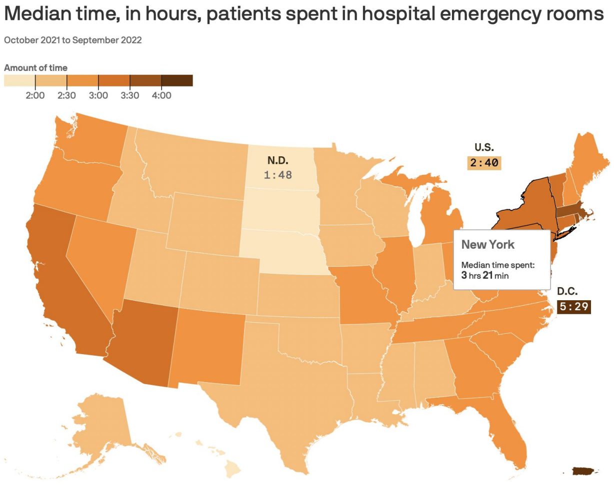
Group 7

Members: Louisa Ong, Siyuan Zhu, Benjamin Tan

Background and Problem Definition

The US healthcare system is inefficient.

The diagnosis process is inefficient because it often involves a soft diagnosis by a nurse first, followed by a final diagnosis from a doctor. While this ensures thoroughness, hospitals face long patient wait times, highlighting the need for faster diagnostics.



26 days

8% ⇧ Avg. patient appointment waiting time from 2017

201 mins

Avg. patient accident & emergency waiting time in NY

The Solution?

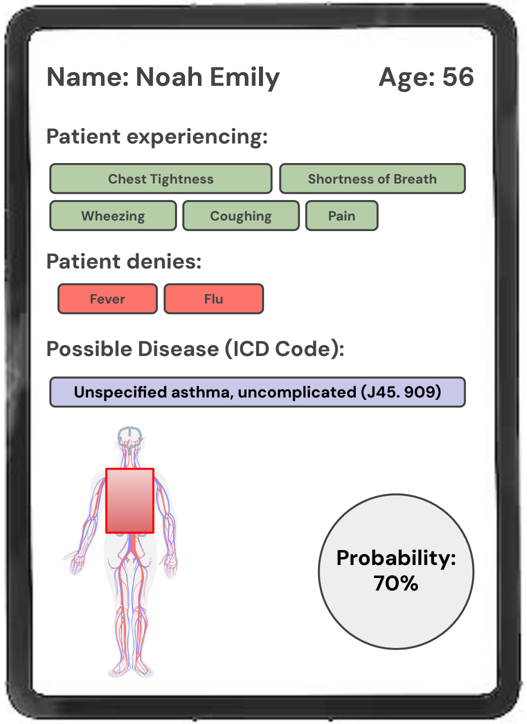


Figure 1: Doctor's View

Proposed use case: A medical search tool for nurses to assist doctors by performing preliminary diagnoses, saving time while waiting for the patient to be called.

1. Nurses inputs a detailed textual description of the patient's symptoms into the system during Triage.

2. The system captures the symptoms and matches them to potential diagnoses that the patient might have.
   a. The system provides a multi-label output indicating the presence of predicted ICD-10 codes used in healthcare, which provides the doctor assists the doctor on preliminary diagnosis.

WHY IS THIS IMPORTANT?

We believe that nurses play a critical role in the diagnosis process.

1. Nurses spend significant time on the triage process from taking down temperature, blood pressure to symptom collection.

2. This tool simplifies and speedup the triage process while capturing the symptoms without the doctor having to ask again during assessment. Quickly map symptoms to potential diagnoses, saving time for both doctors and patients.

Key Takeaway: Doctors can focus on patient instead of busy typing out what the patient is experiencing. Saves significant time in the process (Faster Diagnosis, lesser waiting time for patients)

Data source

MIMIC - IV

Published: Oct. 11, 2024. Version: 3.1

Dataset Description

Contains de-identified clinical data from over 40,000 patients admitted to critical care units.

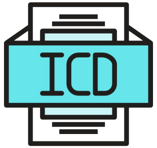
Includes clinical notes, ICD-9/10 codes, Clinical Terminology, and admission details.

Name and Source of Dataset

Link

Clinical Notes

Full note • Conversation • Summary



Link

ICD-10 Terminology

Code • Description

Procurement

- Data was accessed via PhysioNet after completing data usage agreements and certifications.
- Used for research purposes, ensuring compliance with ethical guidelines.

Methodologies used

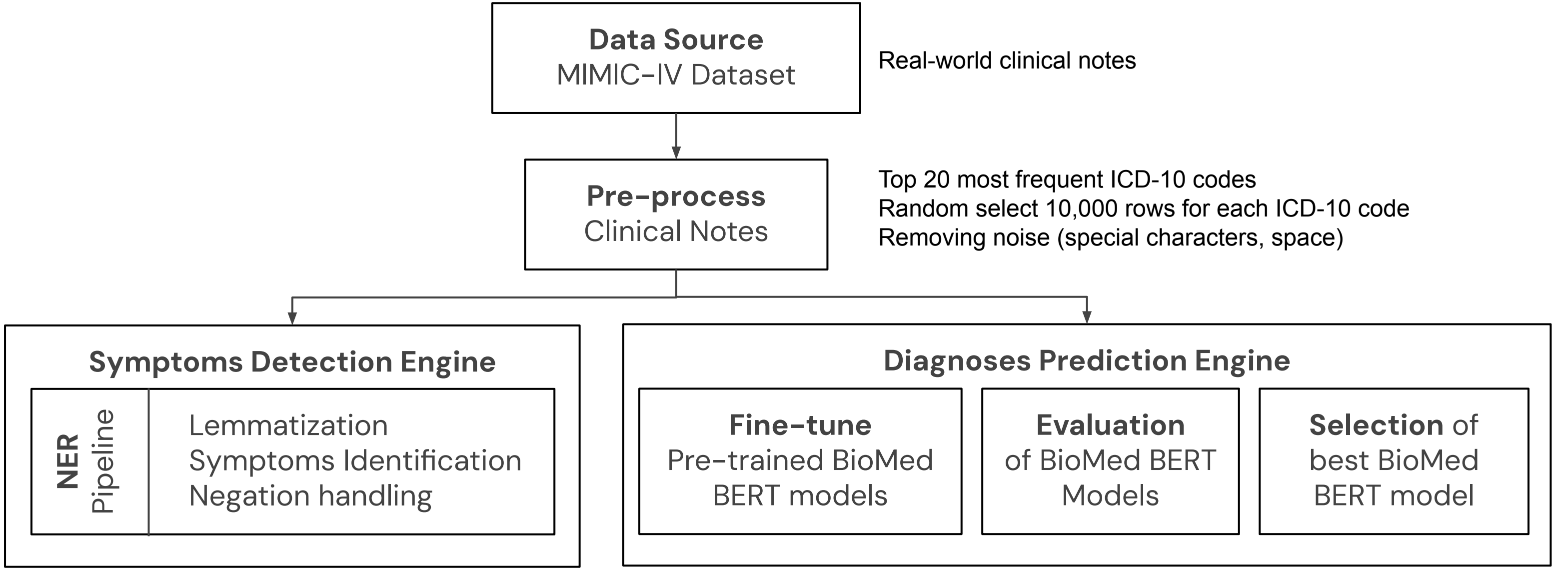
01

NER: Used SpaCy and SciSpaCy to extract features (symptoms, anatomical regions, etc)

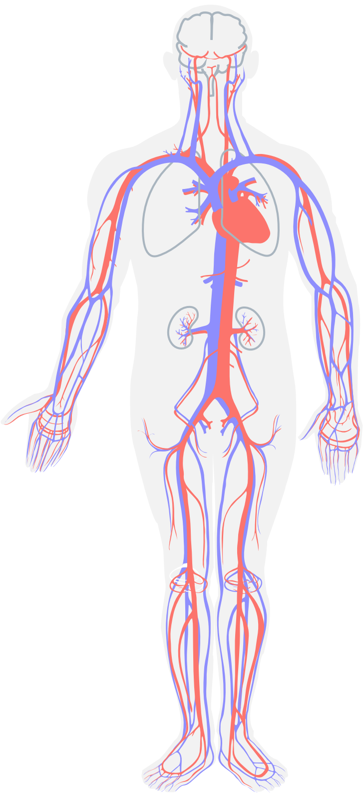
02

LLM: Fine-tuned **Three** pre-trained BERT model (BlueBert, ClinicalBert, PubmedBert)

Design Choice



Key Takeaway: Combination of NER for feature extraction and BERT-based multi-label classification provides a robust pipeline to quickly map symptoms to potential diagnoses, reducing the doctor's workload during patient assessment. The design ensures efficiency, scalability, and accuracy while highlighting the tool's potential to improve healthcare outcomes by enabling faster and more structured diagnostics.



Pre-Processing

For computational purposes, we only focused on the **top 20 most occurring ICD-10 codes** in the dataset

- Merged the main csv files from MIMIC-IV Dataset: ***admission.csv***, ***diagnoses_icd.csv***, ***d_icd_diagnoses***, ***discharge.csv*** using hospital admission id, ***hadm_id*** as the common identifier among the files.
  - The dataset was randomly sampled to have the top 20 most frequent ICD-10 codes with 200,000 rows (20,000 rows each)
  - Consolidated all the ***icd_code*** for each ***hadm_id***
  - Removing extra spaces, special characters
  - Using NER, stored the diseases in a new column

Key Takeaway: By focusing on the top 20 codes, we reduced the dimensionality of the label space, making models easier to run. Future iterations can expand to include more codes once the model's performance is validated

Ner using scispacy

spaCy
(Core library)

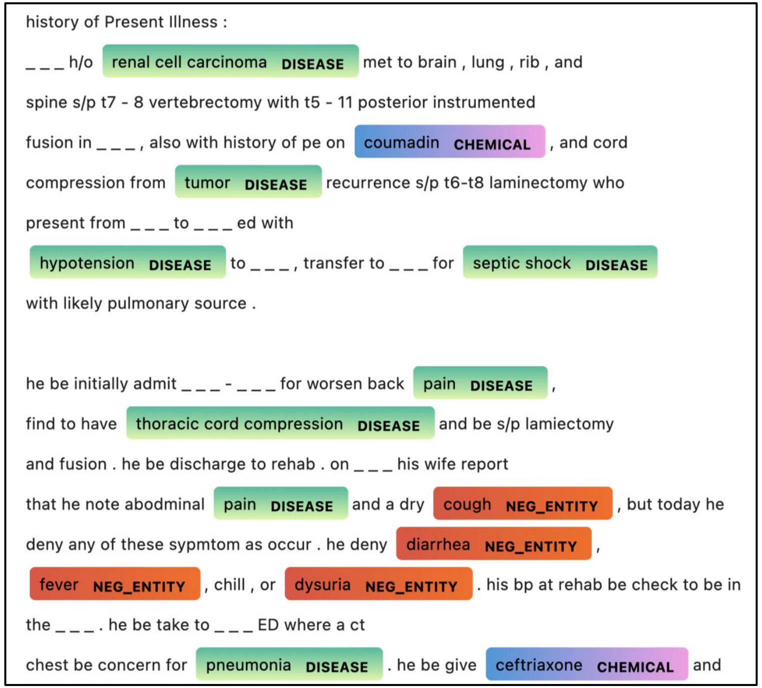
ScispaCy
(Specialized BioMedical Library)

Models

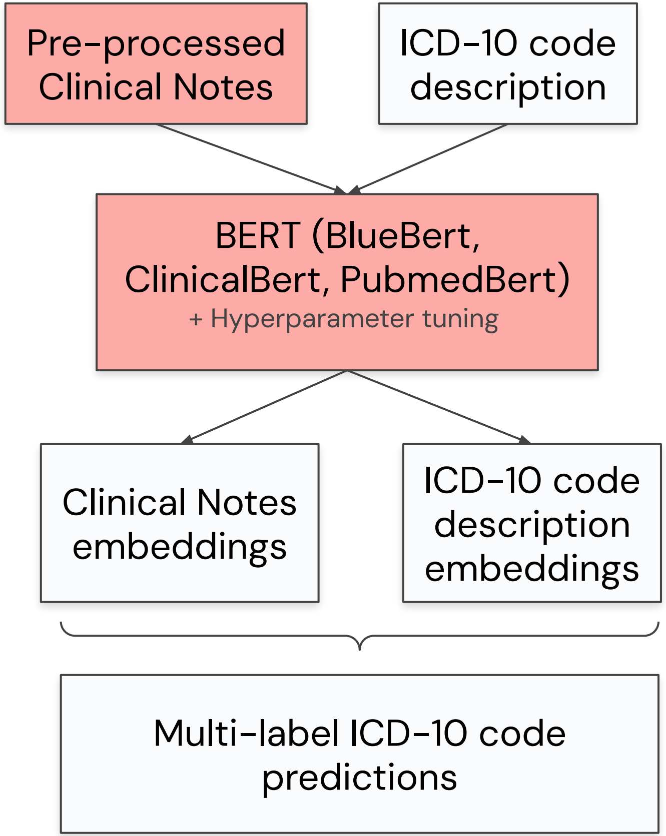
Lemmatization model: en_core_sci_sm

NER model: en_ner_bc5cdr_md

Negation terms
(for negation terms handling)
direct, pre_modifiers, post_modifiers related to the symptoms



Key Takeaway: Named Entity Recognition (NER) in clinical text involves identifying key medical terms like symptoms, diseases, and chemicals using specialized models such as ScispaCy for lemmatization and identification. Integrating negation terms to the pipeline enables detecting negated entities, providing critical insights for clinical decision-making.



BERT For Multi-Label Classification

- Pre-processed clinical notes with descriptions are fed as inputs to finetune the pre-trained BERT models for predicting the multi-label diagnoses.
- Medical Bert models are compared (Precision, Recall, F1-Score)

Transformer Model Comparison

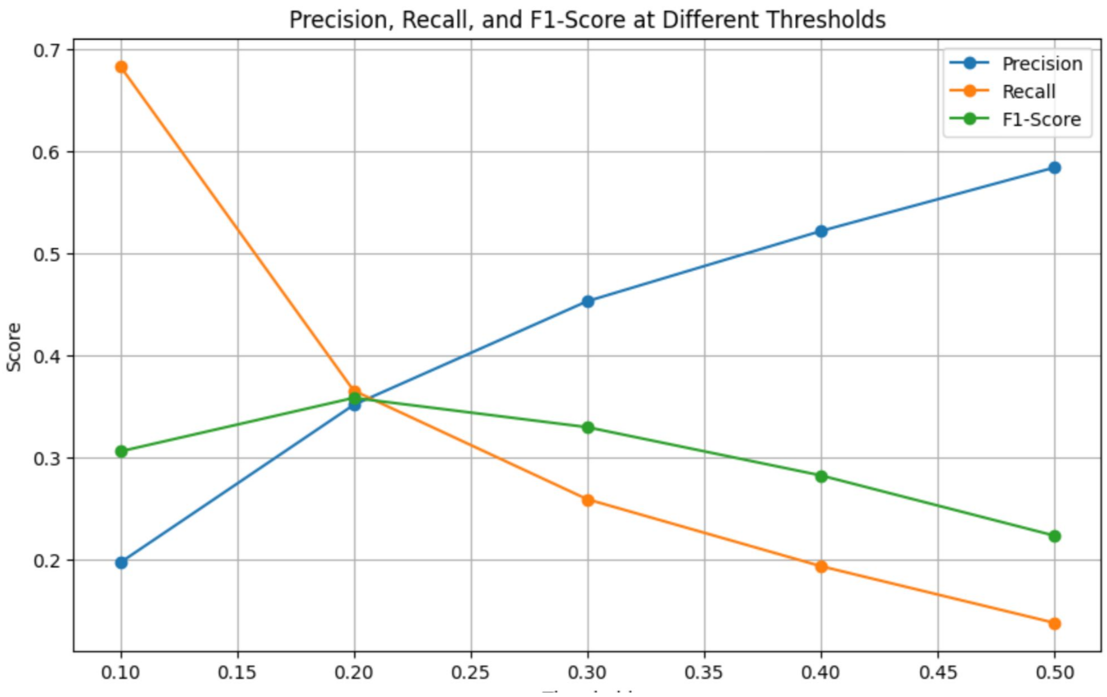
(Results)

	BlueBert (Binary, Default)	BlueBert (Binary, Fine-tuned)	ClinicalBert (Binary, Default)	ClinicalBert (Binary, Fine-tuned)	ClinicalBert (Multi-label, fine-tuned)	PubmedBert (Multi-label, fine-tuned)
Precision	0.13	0.08	0.01	0.17	0.32	**0.45**
Recall	0.14	0.08	0.05	0.17	0.28	**0.26**
F1-Score	0.12	0.08	0.01	0.16	0.29	**0.33**
Support	40,000	40,000	40,000	40,000	39,936	**39,936**

BlueBERT is trained on biomedical and clinical texts
ClinicalBERT is trained on all notes from MIMIC-III
PubMedBERT is pretrained from scratch using abstracts from PubMed and full-text articles from PubMedCentral

Key Takeaway (Conclusion): Precision score is so low despite efforts to process the data because Clinical notes often contain ambiguous, unstructured, or overlapping information, reference symptoms or historical conditions that are not directly related to the actual diagnosis. The model may incorrectly associate such terms with irrelevant ICD codes, increasing false positives. This ambiguity makes it harder for the model to correctly match notes to the exact set of ICD codes.

ClinicalBert (Multi-Labelled) Threshold Results Comparison



THANK YOU