Acta Scientific MEDICAL SCIENCES (ISSN: 2582-0931)

Volume 8 Issue 6 June 2024

Mini Review

Unlocking the Power of Clinical Notes: Natural Language Processing in Healthcare

Sarika Kondra, Wu Xu and Vijay V Raghavan*

University of Louisiana, Lafayette, USA

***Corresponding Author:** Vijay V Raghavan, University of Louisiana, Lafayette, USA.

DOI: 10.31080/ASMS.2024.08.1821

Received: April 15, 2024;

Published: May 09, 2024

Abstract

Electronic Health Records (EHRs) have become the backbone of modern healthcare, providing a comprehensive record of a patient's medical journey. However, a significant portion of this data resides in clinical notes, predominantly consist of unstructured text. While valuable for consumption by medical professionals, this format presents challenges for traditional data analysis methods.

Natural Language Processing (NLP) offers a powerful solution to structure the information presented and unlock the potential of clinical notes. This paper explores the application of NLP tasks within the healthcare domain, specifically focusing on EHR data. We delve into the NLP pipeline, which allows us to differentiate between essential upstream tasks like tokenization and downstream tasks like named entity recognition (NER) and relation extraction. We showcase how NLP can extract crucial clinical information through these tasks and also emphasize the importance of de-identification for maintaining patient privacy.

A major challenge in NLP for healthcare is the limited availability of labeled clinical data. We discuss this bottleneck and explore potential solutions like active learning and transfer learning. Finally, the paper highlights the transformative potential of NLP in healthcare data processing and paves the way for future advancements in this dynamic field.

Keywords: Natural Language Processing; Electronic Health Records; Clinical Applications

Abbreviations

NLP: Natural Language Processing; AI: Artificial Intelligence; EHR: Electronic Health Records; NER: Named Entity Recognition; RE: Relation Extraction; AI: Artificial Intelligence

Introduction

Natural Language Processing (NLP) is a branch of artificial intelligence (AI) that uses powerful techniques like machine learning and deep learning to understand human language. In recent years, NLP has made significant progress, especially with the rise of deep learning models, like Bidirectional Encoder Representations from Transformers (BERT) [12]. This has led to a surge in applying these models to various fields, including healthcare.

One of the biggest challenges in healthcare data analysis is the abundance of unstructured data, particularly in Electronic Health Records (EHRs). EHRs contain both structured data (like lab results) and unstructured data (like clinical notes written by doctors). This unstructured data, which makes up about 80% of the information in an EHR [1], is a goldmine of insights waiting to be unlocked.

Traditionally, extracting information from clinical notes was a labor-intensive process. Clinical experts would meticulously review notes by hand, a slow and error-prone approach. Later attempts at automation relied on rule-based methods and approaches involving and regular expressions. While these methods offered some improvement over manual alternative, they were limited in

scope, struggled with complex language nuances, and didn't scale well with large datasets.

The emergence of state-of-the-art Natural Language Processing (NLP) techniques has revolutionized clinical text processing. By leveraging big data and powerful AI architectures, NLP models can extract information from clinical notes with much higher accuracy and efficiency. These models only require minimal human intervention in their creation, further streamlining the process. This shift has excited researchers and industry professionals alike, leading to a surge in exploring and applying NLP algorithms to unlock the vast potential of clinical notes.

By applying NLP techniques to clinical notes, we can transform unstructured text into a structured format. This allows for the integration of information from clinical notes with other structured clinical information such as past medical history, current medications, allergies, and treatment plans. These insights can then be used to improve patient care, identify potential health risks, and even accelerate research efforts focused on healthcare outcomes.

In essence, NLP acts as a key for unlocking the hidden potential of clinical notes, transforming them from a burden into a powerful tool for improving healthcare outcomes.

Clinical notes

Electronic Health Records (EHRs) are the backbone of modern healthcare, providing a centralized repository for patient information. While EHRs often contain structured data like demographics, diagnoses, and medications, a significant portion of this information resides in unstructured format i.e., clinical notes. These notes, documented by physicians, nurses, and other healthcare professionals, capture the narrative of a patient's medical journey. They detail symptoms, physical exam findings, treatment plans, progress updates, and other crucial details.

Unlike structured data, which is organized in a predefined format, unstructured data lacks a rigid format. However, this wealth of unstructured text holds immense potential.

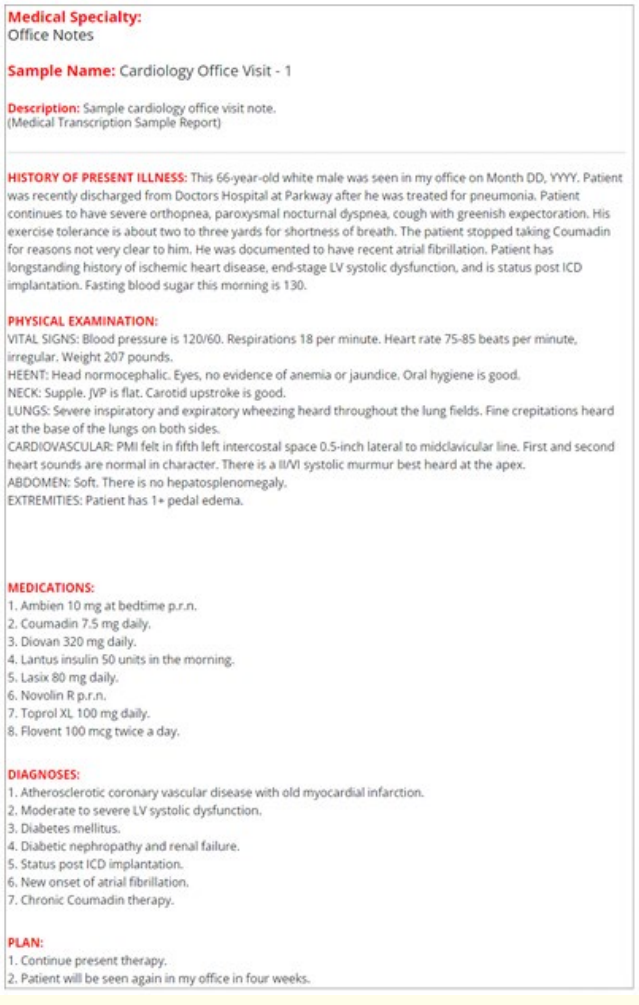
Clinical notes differ from general English language in terms of lack of format, proper grammar, and syntax. Clinical notes often use abbreviations, medical jargon, and specific coding systems like ICD-10 [10], CPT [11] codes for diagnoses and procedures. Figure 1 shows a sample clinical note obtained from publicly available dataset of clinical documents called MTSamples (Medical Transcription Samples) [9].

Clinical notes typically include:

Subjective Data: The patient's own perspective on their symptoms, concerns, and past medical history.

Objective Data: Physicians' observations during physical examinations, vital signs, and laboratory results.

Assessment: The physician's interpretation of the data and their clinical judgment.



Medical Specialty:
Office Notes

Sample Name: Cardiology Office Visit - 1

Description: Sample cardiology office visit note.
(Medical Transcription Sample Report)

HISTORY OF PRESENT ILLNESS: This 66-year-old white male was seen in my office on Month DD, YYYY. Patient was recently discharged from Doctors Hospital at Parkway after he was treated for pneumonia. Patient continues to have severe orthopnea, paroxysmal nocturnal dyspnea, cough with greenish expectoration. His exercise tolerance is about two to three yards for shortness of breath. The patient stopped taking Coumadin for reasons not very clear to him. He was documented to have recent atrial fibrillation. Patient has longstanding history of ischemic heart disease, end-stage LV systolic dysfunction, and is status post ICD implantation. Fasting blood sugar this morning is 130.

PHYSICAL EXAMINATION:
VITAL SIGNS: Blood pressure is 120/60. Respirations 18 per minute. Heart rate 75-85 beats per minute, irregular. Weight 207 pounds.
HEENT: Head normocephalic. Eyes, no evidence of anemia or jaundice. Oral hygiene is good.
NECK: Supple. JVP is flat. Carotid upstroke is good.
LUNGS: Severe inspiratory and expiratory wheezing heard throughout the lung fields. Fine crepitations heard at the base of the lungs on both sides.
CARDIOVASCULAR: PMI felt in fifth left intercostal space 0.5-inch lateral to midclavicular line. First and second heart sounds are normal in character. There is a II/VI systolic murmur best heard at the apex.
ABDOMEN: Soft. There is no hepatosplenomegaly.
EXTREMITIES: Patient has 1+ pedal edema.

MEDICATIONS:
1. Ambien 10 mg at bedtime p.r.n.
2. Coumadin 7.5 mg daily.
3. Diovan 320 mg daily.
4. Lantus insulin 50 units in the morning.
5. Lasix 80 mg daily.
6. Novolin R p.r.n.
7. Toprol XL 100 mg daily.
8. Flovent 100 mcg twice a day.

DIAGNOSES:
1. Atherosclerotic coronary vascular disease with old myocardial infarction.
2. Moderate to severe LV systolic dysfunction.
3. Diabetes mellitus.
4. Diabetic nephropathy and renal failure.
5. Status post ICD implantation.
6. New onset of atrial fibrillation.
7. Chronic Coumadin therapy.

PLAN:
1. Continue present therapy.
2. Patient will be seen again in my office in four weeks.

Figure 1: Sample clinical notes from a publicly available resource of clinical documents called MT samples (Medical Transcription Samples) [9]. This sample adheres to data privacy regulations by removing Personally Identifiable Information (PII) through a de-identification process.

Plan: The proposed treatment plan and recommendations for follow-up care.

The value of clinical notes lies in the:

- Comprehensive Patient History: Clinical notes provide a chronological record of a patient's encounters with the healthcare system, offering a more complete picture than structured data alone.
- Rich Context: Notes capture the reasoning behind diagnoses and treatment decisions, providing valuable context for understanding patient care.
- Temporal Insights: Notes allow for tracking changes in a patient's condition over time, aiding in treatment monitoring and disease progression analysis.
- Unveiling Hidden Patterns: Textual analysis techniques can uncover hidden patterns and relationships within clinical notes, leading to new research opportunities.

This rich tapestry of information provides a nuanced understanding of a patient's condition, but its unstructured nature presents challenges. Traditional methods for data analysis rely on structured data fields, leaving the valuable insights within clinical notes largely untapped. Some of the major challenges of unstructured data are:

- Manually extracting information from notes is time-consuming, labor-intensive, and prone to human error.
- Scalability issues arise and, as the volume of clinical data grows, manual analysis becomes impractical.
- Standardization issues arise due to the lack of standardization in note writing leading to inconsistencies and difficulties in data aggregation.

Natural language processing techniques for clinical text analysis

Natural Language Processing (NLP) tackles clinical notes in a two-step approach: upstream and downstream tasks. Upstream tasks lay the groundwork for understanding the specific language used in clinical notes. This includes fundamental tasks like tokenization (breaking text into words), stemming/lemmatization (reducing words to their base form), and part-of-speech tagging (identifying nouns, verbs, etc.). These tasks are essential for any NLP application, regardless of the domain and dictate the success of downstream tasks.

Downstream tasks, on the other hand, are specific to the researcher's goals. They leverage the foundation built by upstream tasks to perform more targeted analyses. In the clinical setting, common downstream tasks include named entity recognition (identifying medical entities like diseases or medications) and relation extraction (uncovering connections between entities). While a wider range of downstream tasks exist in NLP (like summarization or sentiment analysis), their applicability in the clinical domain is currently limited.

This limited scope is due to two main factors. First, some downstream tasks may not be as relevant to clinical research questions. Second, and more importantly, the availability of labeled clinical data sets is crucial for training effective downstream models. Unfortunately, such labeled datasets are scarce in the clinical domain.

Upstream tasks for clinical test analysis

Text preprocessing

Text preprocessing is the initial stage of the NLP pipeline and lays the foundation for downstream tasks. It involves several subtasks, each contributing to a more accurate understanding of the clinical text.

Tokenization: This fundamental step breaks down the text into smaller units, typically words and sentences. However, clinical notes pose a challenge due to variations in vocabulary and grammar compared to standard English. Tokenization needs to account for:

Abbreviations and codes: Medical abbreviations and codes (e.g., ICD-10 codes) should be treated as single tokens.

Special characters: Hyphens, slashes, and apostrophes within a term (e.g., "x-ray") should be preserved to maintain meaning.

Sentence segmentation: Identifying sentence boundaries in clinical notes can be more complex than in standard English [7]. Periods might not always indicate the end of a sentence, requiring rule-based or machine learning approaches for accurate segmentation. Here, leveraging domain-specific training data is crucial to capture the nuances of clinical note structure.

Here are a few examples showcasing the challenges of sentence segmentation in clinical notes, along with the desired output.

Example 1

Clinical Statement: *"Chief Complaint (CC): Difficulty breathing for 2 days. Hx of Asthma. ROS: Pt denies fever, chest pain, cough. SpO2 92% on RA".*

Desired sentence tokenization output:

Sentence 1: *"Chief Complaint (CC): Difficulty breathing for 2 days. Hx of Asthma".*

Sentence 2: *"ROS: Pt denies fever, chest pain, cough. SpO2 92% on RA".*

Explanation: Abbreviations ("Hx," "ROS," "Pt," "SpO2") and missing periods require separating the information into distinct sentences for better understanding.

Example 2

Clinical Statement: *"Labs: WBC 12,000, Platelets 250,000. Imaging: Chest X-ray shows bilateral infiltrates. Recommend CT scan for further evaluation".*

Desired sentence tokenization output:

Sentence 1: "Labs: WBC 12,000, Platelets 250,000".

Sentence 2: "Imaging: Chest X-ray shows bilateral infiltrates".

Sentence 3: "Recommend CT scan for further evaluation".

Explanation: Colons (":") introduce separate findings, requiring them to be treated as individual sentences for downstream tasks.

Example 3

Clinical Statement: "*Patient reports feeling tired for the past week. No recent travel or sick contacts. Denies fever, cough, or shortness of breath. Vital Signs: BP 120/80, HR 88, RR 18, Temp 98.6 F".*

Desired sentence tokenization output:

Sentence 1: "Patient reports feeling tired for the past week". Sentence 2: "No recent travel or sick contacts".

Sentence 3: "Denies fever, cough, or shortness of breath".

Sentence 4: "Vital Signs: BP 120/80, HR 88, RR 18, Temp 98.6 F".

Explanation: Each bullet point within the clinical note represents a distinct piece of information and should be separated for further processing.

The examples provided are illustrative and do not represent the actual patient data.

Lexical normalization

Lexical normalization aims to reduce the complexity of the vocabulary used in the text. This improves downstream task performance and reduces model size. Here are two common techniques:

Stemming: This process reduces words to their base form (stem) by removing suffixes (e.g., "treating" -> "treat"). While stemming is efficient, it can sometimes lead to loss of meaning (e.g., "running" vs. "racing").

Lemmatization: This technique goes beyond stemming by considering the morphological structure of the word and reducing it to its lemma (dictionary form). Lemmatization preserves more meaning compared to stemming (e.g., "treating" -> "treat").

Part-of-Speech (POS) tagging

POS tagging assigns grammatical labels (e.g., noun, verb, adjective) to each word in a sentence. While POS taggers achieve high accuracy on general English text, their performance can significantly drop on clinical notes due to domain-specific vocabulary and terminology. Therefore, training POS taggers on clinical datasets is essential for improved accuracy in the clinical NLP domain.

These preprocessing and normalization techniques prepare the clinical text for further analysis by downstream NLP tasks, such as Named Entity Recognition (NER) and Relation Extraction (RE). These tasks will be addressed in detail in subsequent sections.

Downstream NLP tasks for clinical text analysis

Having explored a few relevant text preprocessing tasks techniques, we now delve into the heart of clinical NLP:

downstream tasks. These tasks leverage the preprocessed text to extract valuable clinical information.

Named entity recognition (NER)

NER is a cornerstone task in clinical NLP, aiming to identify and extract critical clinical entities. These entities encompass a wide range of concepts, including:

- Diseases and conditions (e.g., cancer, diabetes)
- Procedures and surgeries (e.g., biopsy, X-ray)
- Medications and drugs (e.g., aspirin, morphine)
- Anatomical structures (e.g., heart, lungs)
- Genes and proteins.

State-of-the-art NER models can even capture entity modifiers, such as drug dosages and frequencies (e.g., "500mg" and "6 hours" from the sentence: "500mg of ibuprofen every 6 hours") or anatomical locations (e.g., "upper" from the sentence: "upper arm fracture").

Example

Consider the sentence: "*The patient complains of persistent headaches for the past month. They also report occasional nausea and vomiting. Blood pressure is slightly elevated at 140/90 mmHg*".

Here, a clinical NER model identifies entities like:

- Headaches (Symptom)
- Nausea (Symptom)
- Vomiting (Symptom)
- Blood pressure (Physiological Value)
- 140/90 mmHg (Measurement)

Negation detection

Clinical notes often contain negations that alter the meaning of a sentence. Accurately identifying these negations is crucial for downstream tasks.

Continuing with the previous example, consider the altered sentence: "The patient does not complain of persistent headaches for the past month. They report occasional nausea and vomiting. Blood pressure is slightly elevated at 140/90 mmHg".

Negation detection focuses on recognizing the negation ("does not") associated with the entity "headaches". This ensures the correct interpretation of the clinical information.

All the examples provided are illustrative and do not represent the actual patient data.

Relation extraction (RE)

While NER identifies individual entities, RE goes a step further by uncovering the relationships between them.

Consider the example: "There was a tumor in the ascending colon. A hot forceps biopsy was performed. A single medium-sized polyp was found in the descending colon," a relation extraction task might determine the relationship between "hot forceps biopsy" (procedure) and "polyp" (disease), indicating that the biopsy revealed the presence of a polyp.

RE can identify various relations of interest to clinical researchers, including:

- **Test-Problem:** Linking diagnostic tests to the identified problems.
- **Problem-Treatment**: Connecting medical problems to corresponding treatments.
- **Temporal Events:** Extracting the sequence of events within a clinical record.
- **Drug-related relationships:** Identifying interactions, effects, or protein interactions associated with medications.

All the examples provided are illustrative and do not represent the actual patient data.

De-identification

Maintaining patient privacy is paramount. De-identification techniques utilize NLP to anonymize clinical notes by identifying and masking sensitive information such as:

- Patient names
- Addresses
- Social security numbers
- Dates of birth

De-identification models can achieve this task with minimal human intervention, ensuring patient confidentiality while enabling research on anonymized data.

Example sentence: "*Ms. Sarah Jones with a date of birth of 12-10-1975 reports experiencing chest pain for the past few days. She has a history of hypertension*".

De-identified Sentence: "<LAST> <FIRST> with a date of birth <DATE> reports experiencing chest pain for the past few days. She has a history of hypertension".

This example changes

- Name: "Ms. Sarah Jones" to "<LAST> <FIRST>"
- Date of Birth: "12-10-1975" to "<DATE>"

Here, sensitive details are masked using placeholders like "<LAST>", "<FIRST"> and "<DATE>".

All the examples provided are illustrative and do not represent the actual patient data.

These downstream text extraction abilities are particularly valuable when dealing with EHRs containing minimal structured data and a predominance of free-text notes from nurses and care teams. For computer researchers and enthusiasts, Stanza Biomedical models provide both downloadable and online Demo versions that perform downstream tasks on clinical data [8].

Challenges of labeled clinical data

- **Domain Expertise:** Unlike general English text, clinical notes require medical knowledge for accurate labeling. Busy and scarce clinical experts are needed for this task, making data collection slow and expensive.
- **Privacy Concerns**: Strict regulations around patient privacy make it challenging to share labeled clinical data publicly. This limits the availability of training data for researchers.
- **Data Labeling Complexity:** Labeling clinical data is complex and requires nuance. Annotations might involve identifying specific entities, their modifiers (dosages, laterality), and relationships between them.

Techniques to mitigate data scarcity

Active Learning: This technique prioritizes the most informative data points for human labeling, maximizing learning efficiency with limited data. The model identifies the most uncertain examples and requests human input, focusing labeling efforts on the most impactful data [2,3].

Data Augmentation: Existing labeled data can be artificially expanded through techniques like synonym substitution, back-translation, or random noise injection. This creates synthetic data to train models without requiring entirely new annotations [2].

Transfer Learning: Pre-trained models on large, general language datasets can be fine-tuned for the clinical domain. This leverages existing knowledge and reduces the amount of domain-specific labeled data needed for good performance [2]. Peng., *et al*. explored transfer learning for clinical data [4].

Weak Supervision: Techniques like distant supervision or leveraging existing annotations for related tasks can provide weak labels for training. This might involve using existing structured data in EHRs or information retrieval methods to create noisy labels that can still guide model learning [2,5].

Unsupervised Learning: While unsupervised methods don't directly extract clinical entities, they can be used for tasks like topic modeling or anomaly detection. This can help identify interesting patterns in unlabeled clinical text, potentially guiding further investigation and focused labeling efforts [2,6].

These techniques offer promising solutions to overcome the data scarcity bottleneck in clinical NLP research. By combining these approaches, researchers can create robust NLP models even with limited labeled data, ultimately accelerating progress in the field.

Results and Discussion

This paper explored the application of Natural Language Processing (NLP) techniques for analyzing clinical text data, particularly Electronic Health Records (EHRs). We highlighted the potential of NLP in transforming unstructured clinical notes into a structured format, enabling efficient extraction of valuable clinical insights.

Key Findings

- The surge in deep learning architectures like BERT has

significantly improved the performance of language models, leading to a renewed interest in applying NLP to healthcare domains.

- Clinical notes, a rich source of textual information, pose challenges due to variations in vocabulary, grammar, and abbreviations compared to general English. This necessitates domain-specific NLP models for optimal performance.
- We categorized NLP tasks into upstream tasks (e.g., tokenization, stemming, lemmatization) and downstream tasks (e.g., named entity recognition, relation extraction, de-identification). While most upstream tasks are applicable in the clinical setting, downstream tasks require careful selection based on the specific research goals and the availability of labeled clinical datasets.
- State-of-the-art NLP methods offer promising results in various downstream tasks, including named entity recognition (NER) for identifying clinical entities, relation extraction (RE) for uncovering relationships between entities, and de-identification for ensuring patient privacy.
- Data scarcity is a major problem with healthcare data. Active learning, data augmentation, transfer learning, weak supervision, and unsupervised learning techniques offer promising solutions to address this challenge, paving the way for more powerful NLP models in healthcare.

Conclusion

Natural Language Processing (NLP) has emerged as a powerful tool for unlocking valuable insights from clinical notes, a rich source of unstructured textual data within Electronic Health Records (EHRs). This paper explored the application of NLP techniques in the clinical domain, highlighting its potential to transform healthcare by automating information extraction and analysis tasks.

Our discussion focused on two key NLP processing stages: upstream tasks and downstream tasks. Upstream tasks lay the foundation for downstream applications by performing essential text processing steps like tokenization, stemming, lemmatization, and part-of-speech tagging. These tasks, while well-developed for general English text, require specialized techniques and training data for optimal performance in the clinical domain due to the unique vocabulary and structure of clinical notes.

Downstream tasks leverage the preprocessed text to extract clinically relevant information. Named Entity Recognition (NER) plays a crucial role in identifying entities like diseases, procedures, medications, and anatomical structures. Relation Extraction (RE) goes beyond named entities by uncovering the relationships between them, providing a deeper understanding of the clinical context. De-identification ensures patient privacy by masking sensitive information before data analysis.

While NLP offers significant advantages for clinical research and practice, challenges remain. The availability of labeled clinical data, essential for training NLP models, is a major bottleneck. Additionally, the constantly evolving nature of medical language necessitates ongoing model adaptation and refinement.

Despite these challenges, the future of NLP in healthcare is promising. Advancements in deep learning and active learning techniques hold the potential to address data scarcity and improve model performance. As NLP research continues to evolve, we can expect the emergence of even more powerful tools that will revolutionize how we extract, analyze, and utilize clinical data to improve patient care, streamline healthcare workflows, and accelerate clinical research.

This paper has provided a high level understanding of NLP applications and challenges in the clinical domain. Further research is warranted to explore the potential of NLP in specific clinical use cases and to develop robust and scalable NLP pipelines for real-world healthcare applications.

Bibliography

1. Martin-Sanchez Fernando and Karin Verspoor. "Big data in medicine is driving big changes". *Yearbook of Medical Informatics* 23.01 (2014): 14-20.
2. Spasic Irena and Goran Nenadic. "Clinical text data in machine learning: systematic review". *Jmir Medical Informatics* 8.3 (2020): e17984.
3. Settles Burr. "Active learning literature survey". (2009).
4. Peng Yifan., *et al*. "Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets". arXiv preprint arXiv:1906.05474 (2019).

5. Snorkel AI. "Snorkel: A System for Crowdsourcing Weak Supervision".

6. Shen F., *et al*. "Unsupervised representation learning for electronic health records: A review and outlook". *Briefings in Bioinformatics* 21.2(2020): 544-553.

7. Campbell David A and Stephen B Johnson. "Comparing syntactic complexity in medical and non-medical corpora". *Proceedings of the AMIA Symposium. American Medical Informatics Association* (2001).

8. Zhang Yuhao., *et al*. "Biomedical and clinical English model packages for the Stanza Python NLP library". *Journal of the American Medical Informatics Association* 28.9 (2021): 1892-1899.

9. Medical Transcription Samples: www.mtsamples.com.

10. World Health Organization. ICD-10: international statistical classification of diseases and related health problems : tenth revision, 2nd ed. World Health Organization (2004).

11. Dotson Peggy. "CPT® codes: what are they, why are they necessary, and how are they developed?". (2013): 583-587.

12. Devlin Jacob., *et al*. "Bert: Pre-training of deep bidirectional transformers for language understanding". arXiv preprint arXiv:1810.04805 (2018).