1 Title: Identifying Primary Medical Diagnoses and Underlying Conditions from Clinician Notes

Author: Vinayak (Vin) Kannan

2 Summary

Problem To Solve
In the current prior authorization, clinician notes need to be manually reviewed and parsed in order to derive medical understanding of the patient. Given that clinician notes are unstructured and lack a unified format, this process is tedious, time-consuming, and costly.

Potential Solution
Machine Learning approaches can be used to reduce complexity and waste in this domain. This report details an ensemble-based methodology leveraging NER (Named Entity Recognition), RE (Relationship Extraction), and LLM (Large Language Model) techniques to parse unstructured clinician notes to identify the primary medical diagnosis of a patient (e.g., heart disease) and the underlying factors for each diagnosis (e.g., type 2 diabetes). This approach has demonstrated promising results on synthetic data sources, laying the foundation for future development. Furthermore, this approach tries to provide as much feedback to the end-user as possible in order to facilitate explainability and 'trust' in the model

Next Steps
The team should invest resources validating the strategic / financial viability of solving this problem using a Machine Learning-based approach. Further work should also be done to gather 'ground truth' labeled data sets provided by medical professionals to support model development and facilitate explainability. Finally, resources should be dedicated to model deployment and monitoring.

3 Table of Contents

This report is divided into 3 parts:

4 Proposed Approach

There are two overarching problems to solve for each clinician note: identifying the primary medical diagnosis and determine the common underlying factors for each diagnosis

4.1 Problem 1: Identifying Primary Medical Diagnosis

Summary

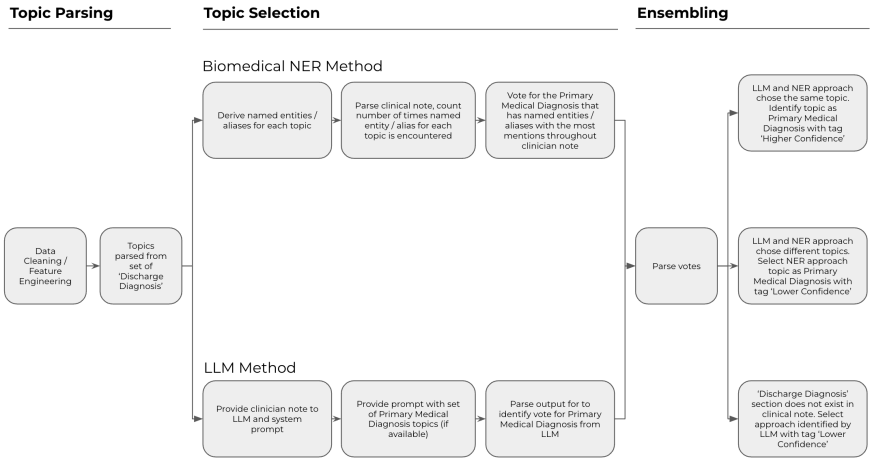
This approach looks to leverage unsupervised topic modeling to identify the primary medical diagnosis from an unstructured / unlabeled data set of clinician notes. The set of possible primary diagnoses can be reliably parsed from the 'Discharge Diagnosis' section within each clinician note (present in 91% of notes in data set). To select the correct diagnosis amongst this set, an ensemble approach is used leveraging both Biomedical NER powered by ScispaCy's Entity Linker and an LLM. Both methods then 'vote' on a topic for the clinician note and then tag their prediction with the label 'Higher Confidence Prediction' or 'Lower Confidence Prediction'.

By leveraging an ensemble approach, this method looks to both improve accuracy and potentially advise the end-user to manually review clinician notes where the Primary Medical Diagnosis is unclear.

Key Assumptions

1. Primary medical diagnosis information is often stored in the 'Discharge Diagnosis' section within each clinician note
2. A 'Primary Medical Diagnosis' is defined as the 'Main condition treated or investigated during relevant episode of healthcare'. Using this definition, the 'Discharge Diagnosis' that is discussed the most throughout the clinicians note is most likely the Primary Medical Diagnosis
3. Most entities ($\sim$ 99% percent) provided in supplied .ann files do not fall in the 'Discharge Diagnosis' section; hence .ann file cannot be relied on to identify Primary Medical Diagnosis
4. End-User has resources to use prediction tags manually review 'Lower Confidence Prediction' labels as needed

Visualization of Approach / Additional Details



Topic Parsing

Data cleaning consists of standard NLP handling (tokenization, lemmatization, text lowering, etc.) After cleaning, the 'Discharge Diagnosis' section is identified; each potential diagnosis is abstracted as a potential Primary Medical Diagnosis

Topic Selection

Biomedical NER Method

1. Each potential diagnosis is run through ScispaCy's EntityLinker and AbbreviationDetector to both identify relevant medical topics / aliases within each diagnosis
2. Every sentence, excluding the 'Discharge Diagnosis' section, is parsed by ScispaCy tooling. If the entity / alias encountered is in the set of aliases for a potential diagnosis, the 'score' for that diagnosis is incremented
3. The diagnosis with the highest 'score' is voted as the primary topic for the document and hence the Primary Medical Diagnosis

LLM Method
1. The clinician note is provided to the LLM in addition to the set of potential diagnoses. The LLM is then asked to identify the Primary Medical Diagnosis from the set for the document
2. The result is parsed and the selected value is voted as the Primary Medical Diagnosis

Ensembling
If the ensemble methods select the same diagnosis, then that is reported as the Primary Medical Diagnosis with the label 'Higher Confidence'. If they select different diagnoses, then the Biomedical NER method is reported with the label 'Lower Confidence' if a 'Discharge Diagnosis' section is present in the clinician notes. The LLM's output is discarded in this case as LLM's are prone to hallucinations, which are costly in this use-case. If a 'Discharge Diagnosis' cannot be found, then the LLM's prediction is reported with the label 'Lower Confidence'

Results
There is no set of labeled 'ground truth' data to compare the results of this approach against. As a proxy, an artificial data set has been generated that uses the first occurring potential diagnosis provided in the 'Discharge Diagnosis' as the 'ground truth' label to measure performance against. This approach was selected after observing several clinical notes, which identify the Primary Medical Diagnosis, list it first.

Performance Metrics using Artificial Data Set

Approach	Accuracy (% of Clinical Notes where 'Correct' Primary Medical Diagnosis Chosen)
NER	41%
LLM	87%
Ensemble	92%

Potential Approach Gaps and Further Steps to Mitigate
1. Additional models could be added to the ensemble method to improve performance and provide a wider range of confidence levels
2. A 'ground truth' set of labeled documents from a medical professional tagging the Primary Medical Diagnosis should be gathered in order to fairly assess methodology performance
3. This approach currently doesn't format the medical diagnoses from the clinician notes to be reader friendly. Further rules-based parsing could improve the end-user experience
4. Additional rules-based logic can be added to the NER approach to further increase accuracy
5. When deploying this method to production, optimizations should be made to run methods within ensemble concurrently, in order to improve performance
6. LLM should be migrated from OpenAI models to organization hosted / deployed models, in order to optimize costs, protect PHI data, and potentially leverage fine-tuned models better suited for task

4.2 Problem 2: Determine Common Underlying Factors for Diagnosis

Summary

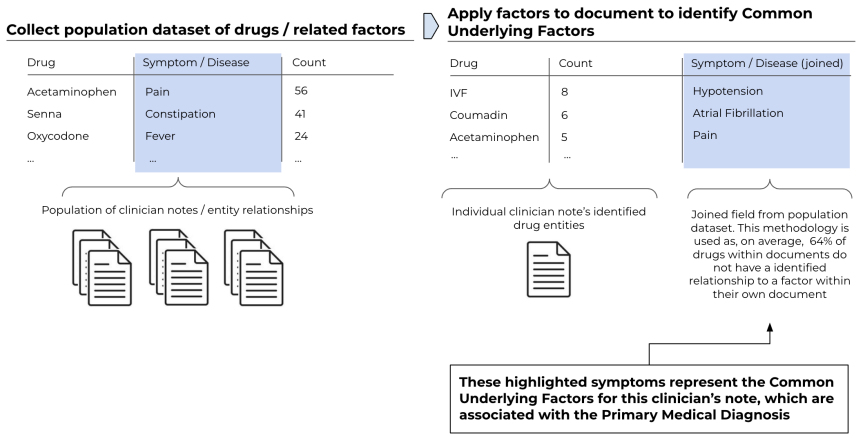
This approach uses relationships identified between named entities, such as drugs and symptoms / diseases (factors), to identify the most common underlying factors associated with a diagnosis. The set of drugs, factors, and their relationships, have already been identified for each clinical note. By aggregating this data, a data set of drugs and the most relevant factors they treat can be collected. After tallying the top 5 occurring drugs in a document, this data set can be used to relate each drug to a symptom / disease and identify them as the common underlying factors.

This approach benefits by using the population of related drugs and symptoms to support its unsupervised approach. Furthermore, with future development, this approach can help give the end-user confidence in the system by directly citing phrases / entities from the document and displaying them to the user.

Key Assumptions

1. A 'Common Underlying Factor' are the symptoms / diseases that appear in conjunction with the Primary Medical Diagnosis. They are not the primary focus for treatment
2. A 'Common Underlying Factor' can be identified by the drugs referenced in the clinician notes. The more often its related drug is referenced, the more pertinent the factor is to the Primary Medical Diagnosis
3. 'Common Underlying Factor' and Primary Medical Diagnosis are fungible (a factor in one case could be the Primary Medical Diagnosis in another)

Visualization of Approach / Additional Details



Collect population data set of drugs / related factors

A table is created leveraging the information from provided .ann entity / relationship files by summing the count of different 'Reason-Drug' category relationships across all files.

Apply factors to document to identify Common Underlying Factors

Using the provided .ann entity recognition files, a table of drug references in a particular document can

be creating using the 'Drug' category. Relevant symptoms / diseases can be joined to this table using the population data set of drugs / related factors. This step is critical as, on average, 64% of drug entities within a clinical note are not linked to a factor. The top 5 occurring represent the Common Underlying Factors for this document that are associated with the Primary Medical Diagnosis. To further validate that the right types of factors have been joined, the pipeline applies the Bio-Epidemiology-NER library to

Results
A potential concern with this approach is that the joined values from the population data set may override the patient's specific factors. For example, if the patient is treated with widely common drugs, the identified Common Underlying Factors may not mention the specific factors the patient is experiencing.

To measure this potential problem, the generated Common Underlying Factors were compared against an entity list of symptoms / factors present in the each clinician note. $\sim$ 89% of document Common Underlying Factors cited at least one factor directly mentioned in document. This suggests that the identified Common Underlying Factors are flexible enough to cite specific patient symptoms while also being able to leverage population knowledge about drug-to-factor relationships to correctly represent drug entities within a clinical note not linked to any factor.

Potential Approach Gaps and Further Steps to Mitigate
1. Additional rule-handling could be added to the method in order to ensure that the identified Common Underlying Factors do not restate the Primary Medical Diagnosis
2. A 'ground truth' data set of actual Common Underlying Factors labeled by medical professionals could be collected and used as part of a text classification model to predict Common Underlying Factors. It is important to verify that there is enough labeled data to pursue this path as this is a supervised learning technique
3. The model could retain direct references within the document corpus to cite where / how it derived the Common Underlying Factors, in order to explain results to end-user

5 Next Steps

There are several key next steps:
1. Financial cost-benefit analysis needs to be performed in order to assess if model is strategically viable
2. Given appropriate resourcing, this approach can be further strengthened by gathering labeled 'ground truth' data sources from medical professionals to fairly measure performance.
3. Further work can be done to improve model explainability for medical professionals. For example, the Common Underlying Conditions can be directly cite references within the clinician notes for the end-user to validate, giving them further confidence in the methodology
4. After further model refinement and testing, Cohere Health should explore model deployment and roll out in tandem with monitoring infrastructure, in order to evaluate model performance over time
5. Test cases leveraging unittest and linters should be implemented to ensure code quality