Does similarity in clinical notes mean similarity of patients insurance types?

Saman Jahangiri, Sean O'Connor

Abstract

The MIMIC- III (Medical Information Mart for Intensive Care) is a publicly available dataset created by the Massachusetts Institute of Technology Lab and includes health data for thousands of patients along with his or her accompanying attributes: gender, ethnicity, religion, language, etc. This dataset has been utilized in various areas of research ranging from academic research to making improvements in the healthcare field. Since this dataset is free, it has a seemingly never ending amount of analysis techniques. For our research, we will be utilizing a few different comma separated files to draw informed conclusions/determinations about the insurance type of patients only by looking at their clinical notes. We do this by checking the similarity of the note for the test patient and the note for other patients.

1 Introduction

With the utilization of the MIMIC- III dataset there are a varying amount of research techniques to be performed; however, our research will be conducted on specifically patient data and the corresponding clinical notes. With this in mind, our research question pertains to a patient's insurance: Can we take a patient's clinical notes and corresponding comparisons to draw conclusions about their form of insurance?

Furthermore, this question is dependent on five different forms of insurance provided by MIMIC-III. These forms are Self-Pay, Private, Medicare, Medicaid, and Government insurance. Self-Pay is clearly the most expensive, and private insurance is representative of companies such as Blue Cross and Blue Shield. Medicare and Medicaid are forms of public insurance where medicaid is the least expensive of the two because it is geared towards lower income families. Finally, government insurance covers areas such as the military by providing insurance plans such as Tricare Select and Tricare Plus.

Overall, this research question can be utilized to help improve the healthcare workforce by analyzing a specific attribute of a health database. We intend to answer this question to better understand if there is a different level of treatment dependent on a patient's insurance.

Since the care and treatment of patients is critical, this area of research could potentially trace corresponding relationships that could then be utilized in creating a better healthcare environment independent of what form of insurance him or her possesses. With this in mind, we would be able to make recommendations to doctors and other healthcare professionals who work in the ICU (Intensive Care Unit) about changing certain trends, actions, or treatment towards a patient with specific demographics/ attributes. The healthcare field will always have areas for improvement; therefore, it is important to ensure actively practicing healthcare professionals have the most information possible to achieve the highest success rate possible. Additionally, in conducting this research we want to know if Okapi BM25 (Okapi) is a good metric for predicting similarity in insurance. With this research, we will be able to assess the relation between a patient's insurance type and the clinical note provided by a doctor. From here, it is possible

to assess the clinical note to see if the level of urgency and/or level of attention for a patient is determined by a specific insurance type.

2 Literature review

As previously stated, the MIMIC- III dataset is a widely used and analyzed database, which means that an abundance of research has been conducted with it. After searching the internet for scholarly articles with ties to the MIMIC-III dataset, we were unable to find a paper that directly answered the question we intend on answering. Determining mortality rate with varying attributes seems to be the main research area when it comes to this dataset. Furthermore, there is research analyzing the type of health insurance a patient possesses; however, this research does not utilize clinical notes as its primary dataset in drawing conclusions. To the best of our knowledge, there is not a research paper in existence that attempts to take a patient's clinical notes and corresponding attributes to draw conclusions about his or her specific insurance type. The articles (Gentimis, 2017), (Nuthakki, 2019) and (Zhou, 2020) utilize MIMIC- III.

The three articles listed above serve as an example of research papers that have performed research on the MIMIC- III database; however, they are not representative of papers that answer our specific question nor do they apply the same attributes to their research in attempting to draw their corresponding conclusions.

3 Our method

As a high level and generalized plan of approach, there are several steps we intend on completing to achieve and fully answer the research question we have proposed. To reiterate, our main research question, which encompasses our subset of questions, is described below.

Can we take a patient's clinical notes and to draw conclusions about their form of insurance?

To answer this question, our first step was to gain access to the MIMIC- III database. There are several training modules on the ethics of medical research (which are about the do's and the don'ts of working with the data of real patients) that have to be completed prior to having the ability to download the dataset to your local machine. Once this training is complete, it will take a couple of days to go through processing before we receive a confirmation email with our credentials validating that we have gained access to the MIMIC- III dataset.

After gaining access, our next step was to browse/scan through the files and determine which files are best fit to help us answer our research question. This is a tedious task since the database contains around twenty-eight comma separated values files (.csv) which all contain different attributes. Additionally, some of these files relate to each other and it is highly difficult to conduct research if you do not have both files. For example, we will need the file that identifies all of the patients; however, we might not need the file that illustrates the particular caregivers in the healthcare sector.

Once we have selected our desired files, we will need to combine them into one file correlated by each individual patient. This means if one file contains a patient gender and another contains the patient ethnicity we will need to perform a merge to create a singular file which contains all our selected attributes. With this file, we will be able to perform our selected information retrieval strategy against it. The strategy we are going to choose is the Okapi BM25 ranking function.

The BM25 ranking function is a bag-of-words retrieval function that ranks a set of documents against a particular query. This formula takes into account document and query term frequency, document length normalization, and inverse document frequency. The score function being implemented is

presented below:

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})} \quad (1)$$

Our query will be represented by one note of a particular patient and our documents will be a larger set of clinical notes of different patients. From there we will rank the patients with respect to the similarity of their note with the query patient and hence draw conclusions about whether or not insurance plays a role in the attributes of their clinical notes. While this research is being conducted, we will provide a more in depth analysis and evaluation of our steps taken and research conducted. For this purpose, we took the patient notes data, and kept one note for each patient.

Our first step was to clean the data. We did this by removing the digits, punctuations and the stop words from the notes. After this was completed, we only kept the stem part of each word. When we are mentioning a stem word we mean we removed both, if applicable, the suffix and prefix of a word and were only left with the root word. For example, if the word we were looking at was "unrelenting" the prefix "un" and the suffix "ing" would be removed to leave us with the root word of "relent."

Once we have cleaned the data, we are ready to continue in our research process. We have divided this data into train and test subsets. Then, we selected a random patient from the test data, considered his/ her note as the query, and other documents as the document collection. Then we computed the BM25 score of the query and each of the other patient's notes. From there, we select the k patients whose notes got the highest scores. This is the subset of the data that we expected to have matching insurance type with the test patient. So, we have looked at the insurance type of this subset to see in how many cases the majority insurance type matches the insurance type of the test data. However, we realized there is a problem with this approach: The insurance type of the majority of the data would dominate the classification problem because it will constitute the majority of most of the samples; therefore, each sample will most likely have the most number of samples from the insurance which has the maximum number in the whole data. We decided to take this into account when looking at the distribution of the sample. In order to achieve this goal, we used the Chi-Squared test (Wiki). The Chi-Squared test is being utilized to compare the sample from a population against categorical data. It compares the population of each category in the sample with the original population and computes the the following statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i}$$

In which x_i is the sample population of category i, and m_i is the whole population of category i. For our problem, we did the same thing with each category being the patients having the same insurance type. So, we performed the Chi-Squared test for the subset selected for each test patient. We looked at the p-value of the test, and continued the prediction only if the p-value was less than 0.05, which was the case in 93 percent of the test patients. Then, we selected the insurance that contributed the most in the above mentioned summation as our prediction for the insurance type of the test patient.

4 Evaluation

After we have concluded our research, we determined that calculating precision and recall would create the best form of evaluation. Precision is the calculation of relevant results divided by total number of documents retrieved. Recall is defined as the total number of relevant documents in the result divided by the total number of relevant documents in the entire dataset. From precision and recall we are able to calculate the F_1 score. The formula for calculating the F_1 score is shown below:

Type of Insurance	Precision	Recall
Medicare	0.66	0.96
Private	0.22	0.73

Table 1: Top results

$$F_1 = \frac{2PR}{P+R}$$

where P represents precision and R represents recall. We have performed the experiment above on the test patients. Then we have looked into the precision of each insurance type. The best results were for Medicare insurance with precision of 0.58 and recall 0.96. With these values in mind, the subsequent F1 score is 1.1136 / 1.54 or 0.723. In terms of the other insurance types, the results were not conclusive and/or good enough. With this being said, the other calculations conducted did not provide sufficient evidence to create a supporting or opposing contribution in answering our research question.

As described earlier, topic four ("How do we plan to work out the answer to the question?") provides a high level illustration of our plan of approach in conducting our research. In this topic, we will provide a more in depth analysis of how exactly we have drawn our conclusions.To begin with, of the twenty eight comma separated values (.csv) files we were able to narrow it down to three. These three files are ADMISSIONS.csv, NOTEEVENTS.csv, and PATIENTS.csv. In PATIENTS.csv the patients were defined by their respective "SUBJECT ID". This file was the driver for trying to combine the three files because it contained the whole collection of patients. Furthermore, the ADMISSIONS.csv contained the remaining attribute we were looking for. This attribute was INSURANCE. For our research, INSURANCE is the most important attribute because our research is focused on correlations and patterns dictated by specific insurance types. Finally, the last file we desired to assess was NOTEEVENTS.csv. Within this file were the specific clinical notes written by doctors for each specific patient. Specifically, the analysis of the clinical notes will be conducted by implementing Okapi BM25 to determine similarities between notes which could potentially lead to similarities in types of insurance.

Once we selected our desired files, all of our implementation was conducted utilizing the Jupyter kernel.

5 Challenges Faced

The first challenge we faced was that the notes file was not available in the demo version. So we really did not know how it looked like. We needed to complete the prerequisite course to gain access to the whole data-set before we could take a look into the notes file.

Another challenge that we faced was dealing with the large csv files. Initially, we desired to take the comma separated values files, merge them, and output the resulting dictionary into a .json file. We were successful with this approach; however, we experienced difficulty in reading in such a larger .json file. Because of this, we resorted back to creating a large .csv file. Our basis for making this decision is that we were more familiar with inputting a large .csv file versus a .json file. Even so, loading the notes file for instance, was taking a very long time which made the compiler crash for several times. In this regard, we utilized our knowledge on how to load the data of a .csv file in chunks and add them to their corresponding dataset. For this purpose, we have used the pandas library in python and used its readcsv method and gave chunk sizes as input to read the data in chunks.

In addition to the previously stated challenge, we also faced another challenge when it came to accessing the desired .csv files. When trying to open these files, we were experiencing a user permission denied error. After researching this specific issue, we were able to realize that the solution was not to be found

in our program. Rather, the new updates released for Macintosh machines created an issue where users have to manually go into their system preferences and allow the IDE they are utilizing access to open specified files.
Last but not least, the date format of the patients was changed to protect the privacy of the data. the date was changed to a format that was not like any standard time format, so we could not take that into account in our research.

6 Further Improvements

Since the conclusion of our research, we have realized that there are a couple ways for our work to be improved. Firstly, we can improve our research by taking more notes into account when computing similarity. We could potentially compute text similarity for different patients visited by the same doctor to see how much of a role the doctor's writing style plays in our similarity function. Additionally, we could compare similarity on the basis of patients who are experiencing the same disease or illness. In terms of the MIMIC-III dataset, we can use the prescriptions of the patients to help make determinations on the patient's insurance. From the dataset, this would require us to access and utilize the two files named "DITEMS.csv" and "PRESCRIPTIONS.csv". This is just one example of how we could increase the amount of data interpreted into our similarity function which would allow us to make better predictions. Another way we could monitor discrimination against patients is to see if insurance type is a factor in how quickly doctors and nurses give a patient his or her treatment. From the dataset, this would require us to employ the DATETIMEEVENTS.csv file in our research.

Overall, there are several approaches to be taken into consideration when determining how we could make improvements to our research. The general synopsis is that evaluating more MIMIC-III data files will lead to better predictions of a patient's insurance type. From the analysis of these files we would be able to make better determinations if discrimination is present in the healthcare field on the basis of an individual's insurance type.

7 Conclusion

Regarding the way we did our experiment, which was to compare the distribution of the insurance type of samples and the original distribution, we found that in 93 percent of the test cases, the sample distribution was significantly different (p-value¡0.05). Also, the precision and recall for the Medicare insurance, suggest some correlation between the insurance type and BM25 similarity, meaning doctors clinical notes could be affected by the insurance type. However, before reaching a stronger conclusion, other attributes that could be effective in the notes should be taken into account. For instance similar diagnosis or similar doctor could each be the cause for similarity in the notes. However, speaking for all insurance types, our research cannot confidently assume any relation between clinical notes data and a patient's insurance type.

8 Team Contributions

For this research, both teammates, Sean O'Connor and Saman Jahangiri, were responsible for different aspects. Both have finished the course in ethics of research required for accessing the data, to gain access and legally work on the data. Sean was responsible for taking in the three .csv files (NOTEVENTS.csv, ADMISSIONS.csv, and PATIENTSS.csv) and grabbing the desired attributes. Once these attributes were selected, Sean was responsible for merging the data into one singular .csv file. On the other hand, Saman was responsible for taking this file, cleaning the data performing the BM25 ranking and the Chi-Squared test on the data. In terms of creating the presentation slides and report write up, both parties contributed in achieving the desired requirements.

References

Durante Steele Cook Gentimis, Alnaser. 2017. Predicting hospital length of stay using neural networks on mimic iii data.

Insurance	Total Percentage
Medicare	47
Medicaid	9
Private	38
Government	3
Self Pay	1

Table 2: The original distribution of different insurance types

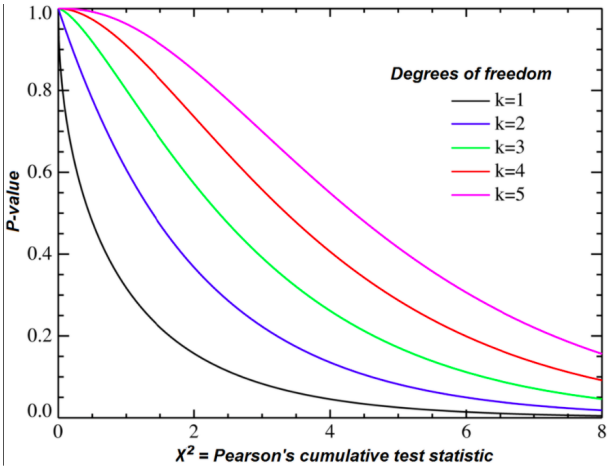
Gichoya Purkayastha Nuthakki, Neela. 2019. Natural language processing of mimic-iii clinical notes for identifying diagnosis and procedures with neural networks.

Okapi. Bm25.

Wiki. Chi-squared test.

Fu Chen Chen Zheng Zhou, Yang. 2020. Outcomes for patients with sepsis following admission to the intensive care unit based on health insurance status: A study from the medical information mart for intensive care-iii (mimic-iii) database.

A Supplemental Material



Here, you can see the Chi-squared distribution with different degrees of freedom. And also how the patients file looks like.

A.1 The tables

	SUBJECT_ID	GENDER	EXPIRE_FLAG	HADM_ID	INSURANCE	
0	249	F	0	116935	Medicare	
1	249	F	0	149546	Medicare	
2	249	F	0	158975	Medicare	
3	250	F	1	124271	Self Pay	BLACK/AFRIC
4	251	M	0	117937	Private	UNKNOWN/N
...	...	...	...	...	...	
58971	44089	M	0	165748	Medicare	
58972	44115	F	0	163623	Private	
58973	44123	F	1	116395	Medicare	
58974	44126	F	0	183530	Private	
58975	44128	M	0	141304	Private	

Figure 1: How the head of the created file looks like

Name		Size	Modified
ADMISSIONS.csv.gz		2.4 MB	2019-03-19
CALLOUT.csv.gz		1.1 MB	2019-03-19
CAREGIVERS.csv.gz		48.4 KB	2019-03-19
CHARTEVENTS.csv.gz		4.0 GB	2019-03-19
CPTEVENTS.csv.gz		4.7 MB	2019-03-19
DATETIMEEVENTS.csv.gz		52.5 MB	2019-03-19
DIAGNOSES_ICD.csv.gz		4.5 MB	2019-03-19
DRGCODES.csv.gz		1.7 MB	2019-03-19
D_CPT.csv.gz		3.9 KB	2019-03-19
D_ICD_DIAGNOSES.csv.gz		278.3 KB	2019-03-19
D_ICD_PROCEDURES.csv.gz		74.1 KB	2019-03-19
D_ITEMS.csv.gz		183.5 KB	2019-03-19
D_LABITEMS.csv.gz		11.2 KB	2019-03-19
ICUSTAYS.csv.gz		1.9 MB	2019-03-19
INPUTEVENTS_CV.csv.gz		402.6 MB	2019-03-19
INPUTEVENTS_MV.csv.gz		143.9 MB	2019-03-19
LABEVENTS.csv.gz		320.3 MB	2019-03-19
LICENSE.txt		2.5 KB	2019-04-09
MICROBIOLOGYEVENTS.csv.gz		7.3 MB	2019-03-19
NOTEEVENTS.csv.gz		1.1 GB	2019-06-14
OUTPUTEVENTS.csv.gz		55.7 MB	2019-03-19
PATIENTS.csv.gz		558.2 KB	2019-03-19
PRESCRIPTIONS.csv.gz		98.7 MB	2019-03-19

Figure 2: Size of some of the files

Important considerations

- The data is sourced from the admission, discharge and transfer database from the hospital (often referred to as 'ADT' data).
- Organ donor accounts are sometimes created for patients who died in the hospital. These are distinct hospital admissions with very short, sometimes negative lengths of stay. Furthermore, their `DEATHTIME` is frequently the same as the earlier patient admission's `DEATHTIME`.
- All text data, except for that in the `INSURANCE` column, is stored in upper case.

Table columns

Name	Postgres data type
ROW_ID	INT
SUBJECT_ID	INT
HADM_ID	INT
ADMITTIME	TIMESTAMP(0)
DISCHTIME	TIMESTAMP(0)

Figure 3: Information about the admission file provided by the data-set website

Important considerations

- `TEXT` is often large and contains many newline characters: it may be easier to read if viewed in a distinct program rather than the one performing the queries.
- Echo reports, ECG reports, and radiology reports are available for both inpatient and outpatient stays. If a patient is an outpatient, there will not be an `HADM_ID` associated with the note. If the patient is an inpatient, but was not admitted to the ICU for that particular hospital admission, then there will *not* be an `HADM_ID` associated with the note.
- Echos are generated using templates and in some cases there may be discrepancies in severity. For example one report may contain: "Mild PA systolic hypertension. Severe PA systolic hypertension." indicating that the caregiver may not have removed the appropriate item from the template.

Figure 4: Information about the notes provided by the website

Table columns

Name	Postgres data type
ROW_ID	INT
SUBJECT_ID	INT
HADM_ID	INT
CHARTDATE	TIMESTAMP(0)
CHARTTIME	TIMESTAMP(0)
STORETIME	TIMESTAMP(0)
CATEGORY	VARCHAR(50)
DESCRIPTION	VARCHAR(300)
CGID	INT
ISERROR	CHAR(1)
TEXT	TEXT

Figure 5: Information about the patient notes file, NOTEEVENTS