Exploring Vital Signs and Medications

Yoon Choi, Kelvin Encarnacao, Zhao Zhang

---

1. Introduction

Patient clinical notes are an important piece of information that we can use to analyze the health data. Our objective in this paper is to determine whether we can gain insights from the dataset in regard to vital signs/physical exam readings and medication across the dataset as a whole and across the dataset at a patient level. We got the data from Kumar et al. paper, extracted relevant information for the health data analysis, and finally created charts and figures to analyze it.

2. Methods

We were given dataset in XML format from Kumar et al. paper titled "Creation of a new longitudinal mix of clinical narratives." The dataset is a mix of free texts and tags. With the data in our hands, we analyzed the vital signs / physical exam readings through the free-text portions and the medications through the xml tagged portion. Our group decided the most efficient method for parsing and analyzing the dataset was to use Python as our programming language and write scripts that iterate through all the relevant information.

For analyzing vital sign/physical examination, we converted the xml file to single CSV file utilizing python's csv library. There would be several key words in the csv file about the vital signs/physical examination such as chest, neck, abdomen, heart, lungs, etc. These data helped us to query all the file to count the frequency of the words. Thus, our final data for question 1 includes vital signs on the first column and frequency distribution on the second column. In order to avoid double counting, we will only calculate the relevant information once in each file.

For analyzing medications (questions 2-5), we likewise converted the XML tags to a single CSV file utilizing Python's csv library. Putting all medication information in a single CSV file allowed for easier data parsing which saved both time and effort [1]. The relevant headers for the columns in the CSV file was patient_id, visit_id, medication_id, start, end, time, text, type1, type2, and comment. Patient_id

and visit_id are integers that are split by '-' from the name of each XML file. For example, patient_id and visit_id would be 100 and 01, respectively, if the file name was 100-01.xml. This way we are able to determine the patient and the separate visit to the doctors. The rest of the headers are the same as from the medication tags. The output process was a simple iterative function that took each row in the CSV file and outputting relevant results in a dictionary, which was later put in another separate CSV file for easier analysis. All rows that didn't match the criteria for each related question were skipped. The most important thing for us to understand was that each row has medication_id that started with 'DOC' or 'M', which we will call as row in '**first category**' and row in '**second category**', respectively. This information is important because the first category encompasses the second category and also has less information. The second category has more information but usually contain the same information with each other if they are encompassed by the same first category.

For question 6a we explored the data Investigation of the Frequency of Patient's Blood Pressure Range. For question 6b we looked at risk factors related to each medication category on the dataset as a whole. Specifically, we compared the frequency of risk factors (diabetes, coronary artery disease (CAD), hyperlipidemia, and hypertension) present in the entire dataset to the number of medication categories given to the patients. The purpose of this question was to see the number of mentions of medications for the risk factors associated with the evidence that the risk factor is present. To do the analysis, we first had to reference the data we had from question 2b (frequency distribution of the medications categories taken). The medication categories from this data all had risk factor/diseases associated with diabetes, coronary artery disease, hyperlipidemia, and hypertension, as stated above. We were able to get the correlations from the Kumar et al. paper annotation guidelines. [2]

2.1 Question 1: frequency distribution of the vital signs

The method used is to use python to return a list of tuples corresponding to the vital sign keywords and frequency, and then loop the xml file to find the keywords and frequency of occurrence. Vital sign keywords include blood pressure, pulse, weight, temperature, and more. Then we cycle each file query keyword and when a single file appears one or more times, we increment the count. Then we use the first column in the table to store vital sign information, and the second column for the frequency of its information in the xml file. We use the data in csv file to make the graph for the result.

2.2 Question 2a: Freq. Distribution of the Medications Taken

This method returns a list of tuples where first element of the tuple is a medication name, and the second element is a frequency of the medication. We loop through only the rows that fit the second category (medication_id starts with 'M' as stated in the methods section) in medication.csv. This is because the ones from the first category does not have the 'text' field which indicates the medication name. We utilize the list 'intermerdiate_medicines' to store 'text' field from each row in the second category until the next row's medication_id starts with 'DOC' (first category), which then we will flush out the elements in the list. This means that this method doesn't double-count the medication name for the frequency if they are 1) the same name and 2) under the same first category.

2.3 Question 2b: Freq. Distribution of the Medications Categories Taken

This method returns a list of tuples where first element of the tuple is a medication category, and the second element is a frequency of the medication category. We loop through only the rows that fit the first category (starts with 'DOC') in medication.csv. This is because the ones from the first category encompasses the ones from the second category that follows. If we counted both rows in the first category as well as the rows in the second category, we will be double counting the frequency and thus inflate our resulting frequency distribution (we safely skipped over the second category because the first category rows encompass the second category rows that follow). We store the 'type1' and 'type2' field from each row without regard to which type it is. We disregarded the different types and simply aggregated them together because the medicines with two types still mean that the medicines are within the bounds of these two and we cannot ignore one another since it doesn't mean the first type is more important than the second type or vice-versa.

2.4 Question 3 and 4: 10 Individuals taking the Greatest / Least Number of Medication Types

These methods return a list of tuples where first element of the tuple is a patient_id and the second element is a frequency of the medication types for the said patient. Like question 2b, we loop through only the rows that fit the first category (starts with 'DOC') in medication.csv. This is because the ones from the first category encompasses the ones from the second category that follows. For each patient, we aggregate the type1 and type2 field of the rows if we haven't counted the medication type

inside the patient's medication type taken. After we gathered all the medication types from a patient, we simply count the number of them and thus we get our result. We put sections for question 3 and 4 together because they are similar, and we can obtain results from simply sorting by second element of the tuple either ascending or descending.

2.5 Question 5: Freq. Distribution of the Medications Taken

This method returns a list of tuples where first element of the tuple is a patient_id and the second element is a frequency of the medication for the said patient. We loop through only the rows that fit the second category (starts with 'M') in medication.csv. This is because the ones from the first category does not have the 'text' field that indicates the medications taken by the said patient. Similarly with question 2a, we utilize the list 'intermerdiate_medicines' to store 'text' field from each row in the second category until the next row's medication_id starts with 'DOC' (first category), which then we will flush out the elements in the list. This means that this method doesn't double-count the medication if they are the same name and under the same first category but does not prevent from double-counting if they are from different first category.

2.6 Question 6a: The blood pressure frequency distribution

This method returns a list of tuples corresponding to the vital sign keywords and frequency, and then loop the xml file to find the keywords and frequency of occurrence. Vital sign keyword is blood pressure. We cycle each file query keyword and get the value for each patient’s blood pressure. Then we use the first column in the table to store blood pressure high information, and the second column to store blood low pressure. We use the data in csv file to make the graph for the result.

2.7 Question 6b: Risk Factors and Medication Categories

The preliminary method returns a list of tuples where first element of the tuple is a risk factor/disease (diabetes, CAD, hyperlipidemia, hypertension, obesity), and the second element is a frequency of the said risk factors in the documents. We can get the first element in each XML tag because the XML tag text is associated with the risk factors. To get the frequency of risk factors, we traversed through the files in the XML folder and picked up XML tags whose id started with ‘DOC’ (first category). This is to keep the data discovery consistent with data from question 2b as they are both in

the 'category type' level. After we get this list of tuples, we mapped them with the number of medication categories from data in 2b. The result is our final data.

For a quick reference, Kumar et al. paper annotation guidelines state the correlations as following:

Risk Factor / Disease	Medication Category
Diabetes	Metformin, insulin, sulfonylureas, thiazolidinediones, GLP-1 agonists, Meglitinides, DPP-4 inhibitors, Amylin, anti-diabetes medications, combinations including these.
CAD	: Aspirin, Thienopyridines, beta blockers, ACE inhibitors, nitrates, calcium-channel blockers, combinations including these.
Hyperlipidemia	statins, fibrates, niacins, ezetimibes, combinations including these.
Hypertension	beta-blockers, ACE inhibitors ARBs, Thiazide diuretics, calcium-channel blockers, combinations including these.
Obesity	orlistat (xenical), Lorqess (Loriciferan)

The first two columns from the table below are the results we had from question 2b. Then, based on each medicine category from column 1, we created additional column to map each medication category to risk factors.

Medicine Category	Frequency	Risk Factor/Diseases
beta blocker	2246	2, 4
statin	2118	3
aspirin	2081	2
ace inhibitor	1583	2, 4
insulin	1029	1
calcium channel blocker	930	2, 4
metformin	930	1
sulfonylureas	759	1
nitrate	607	2
diuretic	593	4

thienopyridine	576	2
arb	481	4
thiazolidinedione	185	1
fibrate	154	3
ezetimibe	72	3
niacin	45	3
dpp4 inhibitors	7	1
anti-diabetes	3	1

Then, we got the frequency of tags associated with each risk factor / disease and combined the tables.

Risk Factor / Disease	Medicine Category Frequency	Number of tags
Diabetes	2913	2875
CAD	8023	1970
Hyperlipidemia	2389	1813
Hypertension	5833	3219

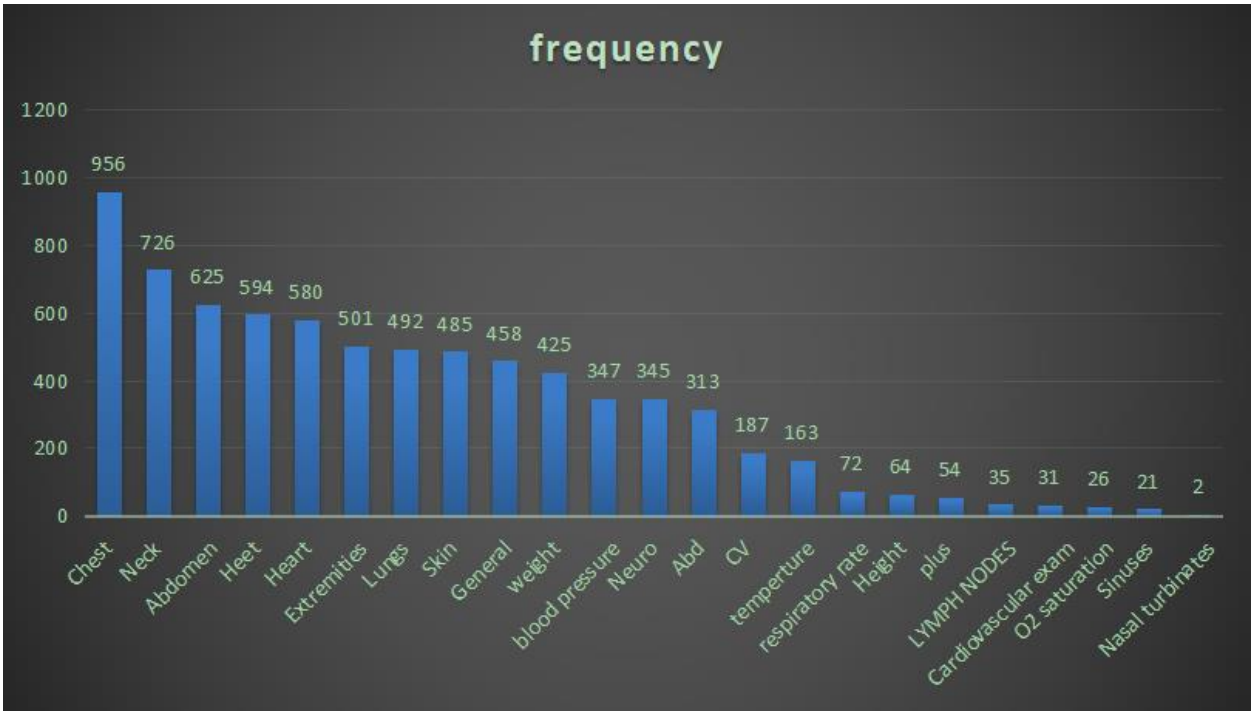
Each of the functions that returned dictionaries as outputs were saved to CSV files using the custom python module named csvModule. This allowed for an easier graph creation and more intuitive result of our analysis. We also created a custom python module named print Module in order to get a sense of our data representative[2] [3]. The final data analysis was done under Excel.

3. Results

3.1 Question 1 Result: frequency of the vital signs/physical examination

The graph below has the vital signs/physical examination on the x-axis and the frequency of them on the y-axis. From the dataset graph, it presents the frequency distribution of the vital signs/physical examination. The most frequently occurring vital signs is chest data, which frequency is

around 956. Its average frequency distribution is around 400. As we can see from the graph, the top 10 vital signs are Chest, Neck, Abdomen, Heet, Heart, Extremities, Lungs, Skin, General, and Weight.



3.2 Question 2a Result: Freq. Distribution of Medications Taken

Figure 1 below represents the frequency distribution of medications taken. There were a lot of medication in our returned result, so we are showing the top ten medications in the figure 1. As we can see from the graph, aspirin is the most frequent medication with the frequency of around 1200, with Lipitor and lisinopril in the second and third place with the frequency of around 1000-1100. There

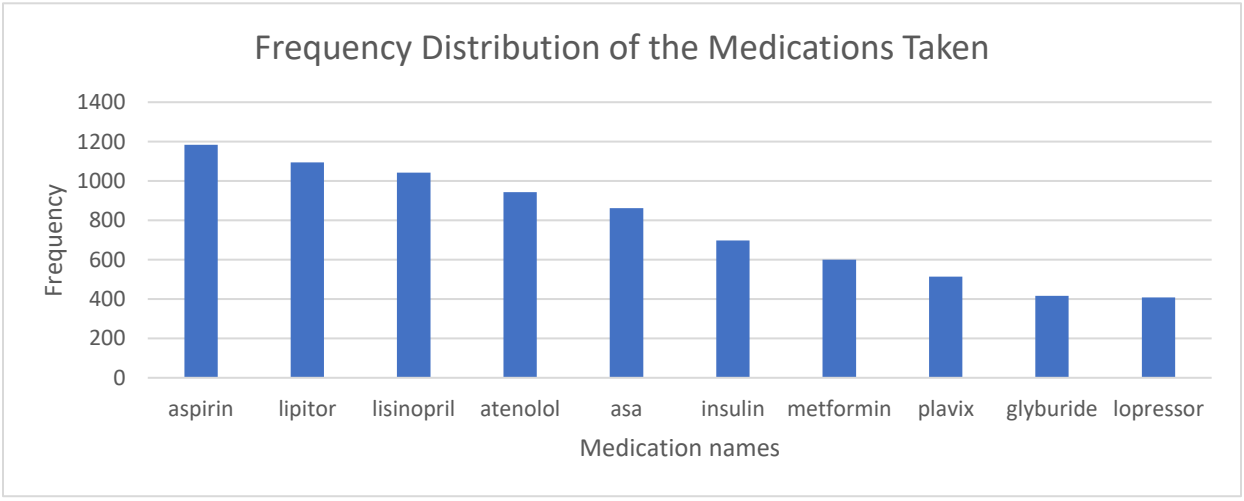


Figure 1: Frequency Distribution of the Medications Taken

doesn't seem to be a dramatic difference in the frequency of the medications taken from one another. Rather, the differences are slight and very gradual.

3.3 Question 2b Result: Freq. Distribution of the Medications Categories Taken

Figure 2 below represents the frequency distribution of medication categories taken. We returned all the medication categories as there were only 18 from our results. As we can see from the graph, beta blocker is the most frequent medication category with the frequency of around 2250, with statin and aspirin in the second and third place with the frequency of a little north of 2000. After that, the frequency of the categories of medications taken sort of fiddles out. One thing to note is that three of the top five medication types are CAD, which stand for coronary artery disease. This seems to suggest that there are a lot of patients visiting the doctors for CAD than one might have expected. However, this is far from surprising because heart disease is "the leading cause of death for men, women, and people of most racial and ethnic groups in the United States" [4].

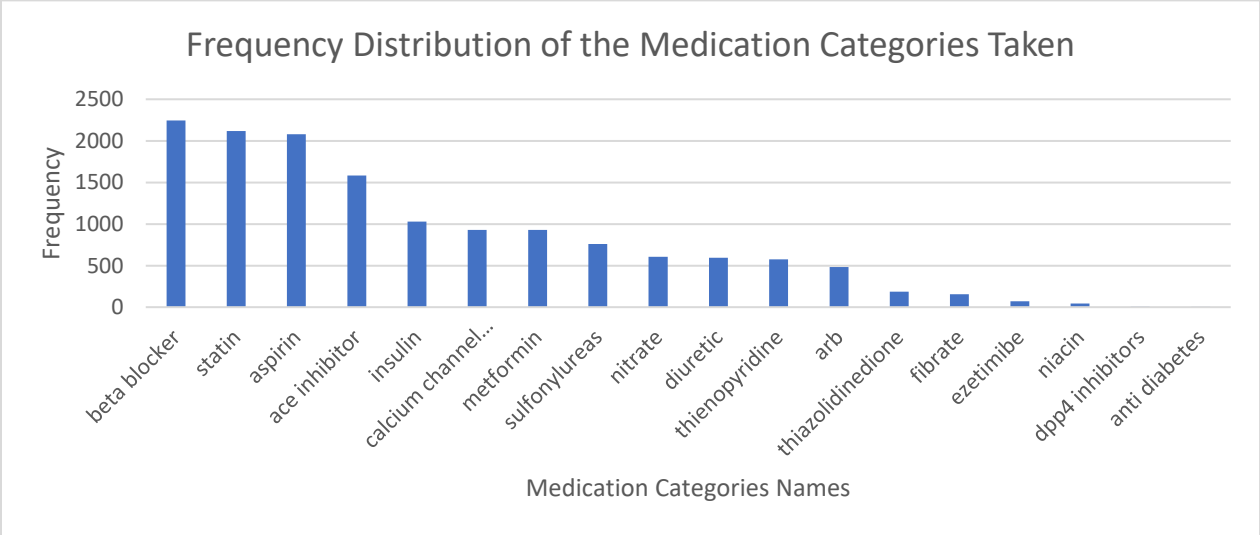


Figure 2: Frequency Distribution of the Medication Categories Taken

3.4 Question 3 Result: Individuals taking the Greatest Number of Medication Types

Below table displays individuals taking the greatest number of medication types. Using our methods as described from the methods section, there weren't 10 individuals cut off, but rather four patients with 13 medication types and nine patients with 12 medication types, bringing the individuals taking the greatest number of medication types to thirteen patients in the table shown.

patient_id	num of medication types
125	13
216	13
281	13
400	13
100	12
106	12
115	12
156	12
177	12
184	12
196	12
202	12
237	12

3.5 Question 4 Result: Individuals taking the Least Number of Medication Types

Below table displays individuals taking the least number of medication types. Using our methods as described from the methods section, there weren't 10 individuals cut off, but rather one patient with 2 medication types, one patient with 3 medication types, and nine patients with 4 medication types, bringing the individuals taking the least number of medication types to eleven patients in the table shown.

patient_id	num of medication types
176	2
160	3
142	4
174	4
246	4
251	4
259	4
307	4
318	4
326	4
369	4

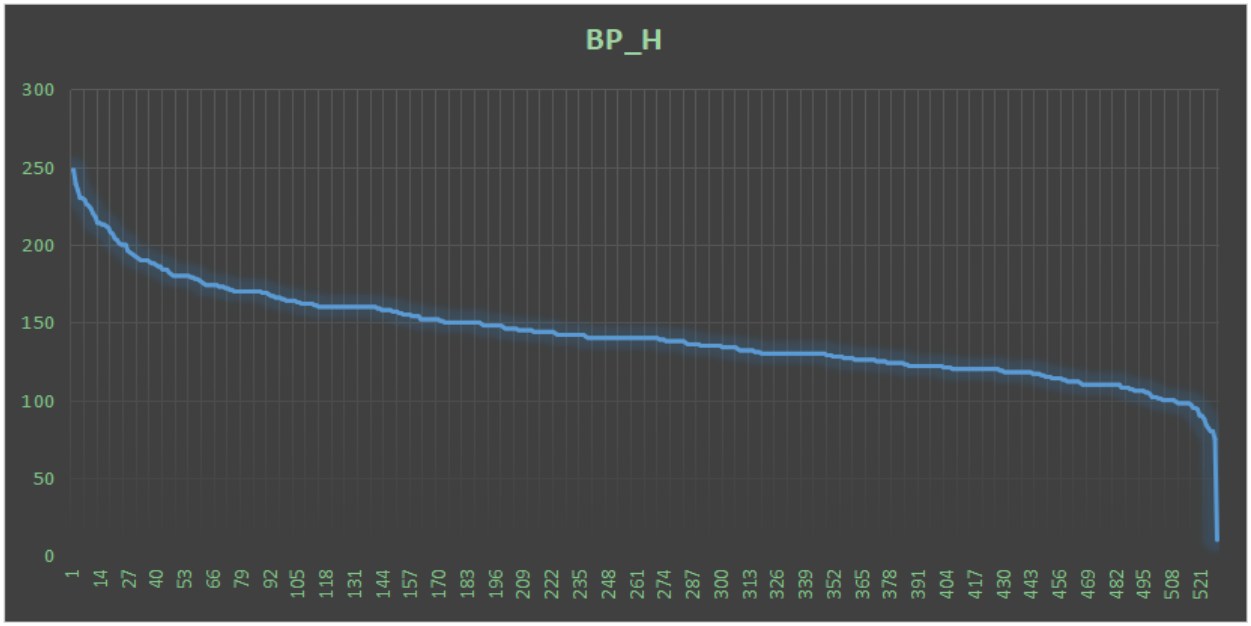
3.6 Question 5 Result: 10 Individuals taking the Least Number of Medications

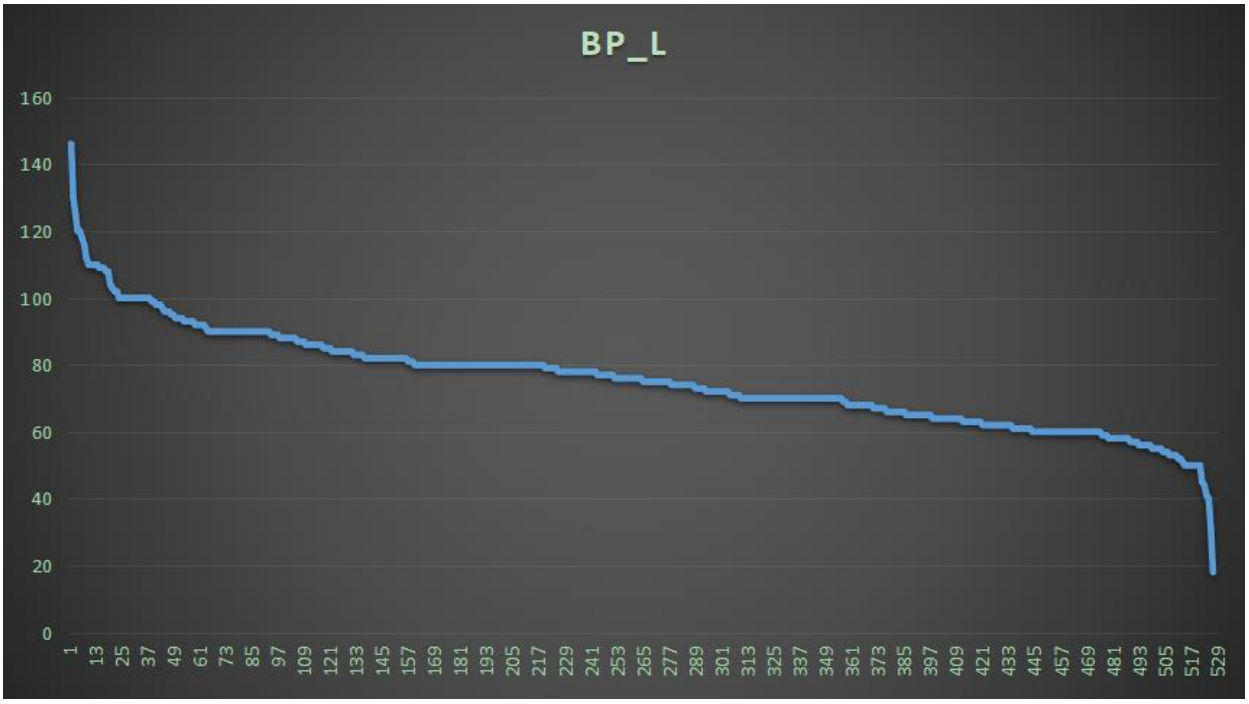
Below table displays individuals taking the least number of medications. Using our methods as described from the methods section, there were three patients with 3 medications, four patients with 4 medications, and three patients with 3 medications, bringing the individuals taking the least number of medications to ten patients in the table shown.

patient_id	num of medication
174	3
176	3
251	3
246	4
259	4
264	4
307	4
119	5
275	5
326	5

3.7 Question 6a Result: frequency distribution of the blood pressure

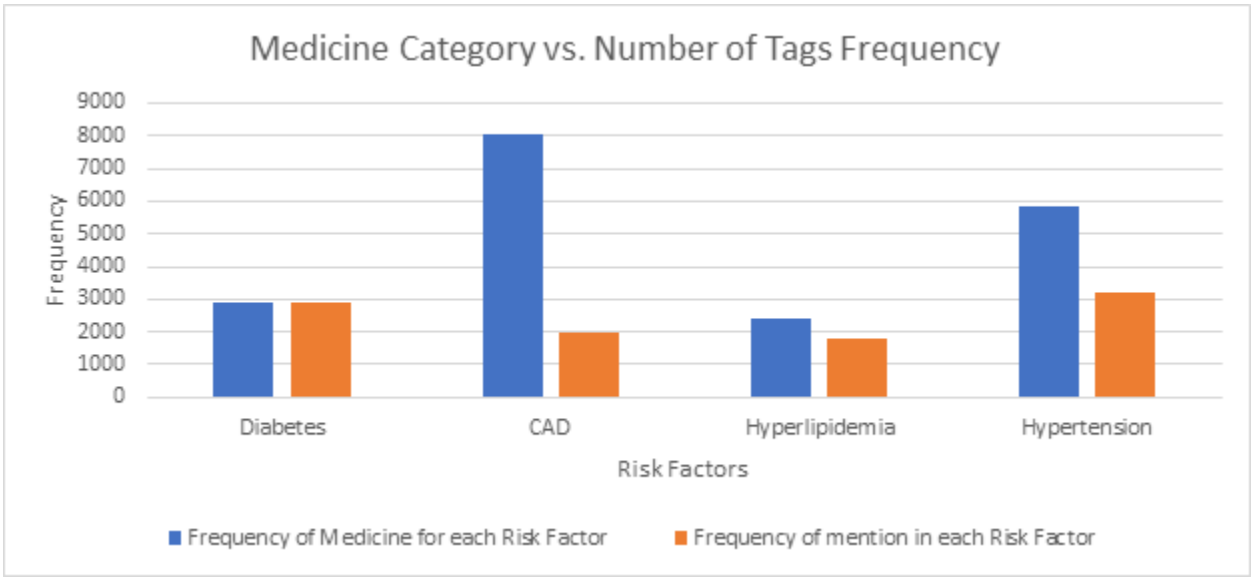
From the figures below, we could know the average of the blood pressure high is around 145, the average of the blood pressure low is around 75. Apparently, the blood pressure high 160-130 is most frequently of the blood pressure high. The blood pressure low 85-70 is the more frequently of the blood pressure low.





3.8 Question 6b Result: Risk Factors and Medication Categories

From the figure below, we can see that diabetes and hyperlipidemia have mostly similar risk factor mention and the medicine mention, while CAD and hypertension have more frequency of medicine than the risk factor mention. Both CAD and hypertension suggest that more medications may have been mentioned than suggested as an evidence of each risk factor looking at the figure.



4. Limitations

Double-counting of the same medications is possible for the frequency questions even if the only difference between the medications is ‘time’. This is because tags whose medication_id starts with 'DOC' differ from one another only by the 'time' field while medication tags inside are identical. However, we left it as is because time is an important aspect of the medication frequency due to the possibility of getting on/off the drugs from before/during/after the DCT.

Secondly, we constructed the frequency results based on their relevant strings, so if the 'text' field had any discrepancies from the representatives due to spelling errors or long strings, the medicine didn't match well enough to contribute to the frequency. For example, we couldn't make sense of 'dilt (30mg TID)' from 114-04.xml so we left them out. This could mean that we were missing some relevant medicines. However, our dataset was large enough to justify filtering through the anomalies and we are confident that we had enough data to make a proper analysis.

5. Conclusions

There is various information available through patient clinical notes dataset that we can use for health data analysis. From the dataset, we know the most frequently occurring vital signs is chest data with the frequency of around 950, and aspirin is the most common medication with the frequency of around 1200. Thirteen medication types were the highest number of medications types the patients were taking, while two were the lowest. Blood pressure high of 160-130 and blood pressure low of 85-70 are the most common blood pressure for each and we can finally see that diabetes and hyperlipidemia have mostly similar evidence of each risk factor and the medicine, while CAD and hypertension have more frequency of medicine than the mention.

Bibliography

1. “csv — CSV File Reading and Writing”. [Online]. Available: https://docs.python.org/3/library/csv.html. [Accessed: 12-March-2021]
2. “Creation of a new longitudinal corpus of clinical narratives”. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26433122/. [Accessed: 12-March-2021]
3. “7. Input and Output — Python 3.9.2 documentation”. [Online]. Available: https://docs.python.org/3/tutorial/inputoutput.html. [Accessed: 20-March-2021].

4. “os.path — Common pathname manipulations”. [Online]. Available: https://docs.python.org/3/library/os.path.html. [Accessed: 20-March-2021].
5. "About Underlying Cause of Death, 1999-2019”. [Online]. Available: https://wonder.cdc.gov/ucd-icd10.html. [Accessed 28-March-2021].