# Diagnosis and risk level prediction from admission report in MIMIC-III database using Natural Language Processing Methods

**Limeng Liu, Shirley Wang**

## Abstract

Our project is going to build a patient oriented system to answer some frequently asked questions when people feel ill. We will conduct some NLP task to fulfill the project goal. We are going to use the admission information such as the history of present illness as our input to predict the diagnosis and the risk level of illness. The system is dedicated to provide people certain sense of the sickness and may help them to make some instant decisions such as whether they need to see a doctor right now.

## 1 Introduction

Public health is always an appealing domain to focus on. In our project, we hope to provide a patient oriented system that makes their lives more convenient. It is pretty common that people will search online for self-diagnosis before they actually decide to see a doctor in a hospital. Patients may ask what sickness I get? Should I go to the hospital? Is that a serious illness? Our system is designed to answer these and more questions that are frequently asked by people when they feel ill. Given the history of present illness from the admission notes and some personal information of the patient, we are going to predict the diagnosis, the risk level of the illness and more coming up in the future work. We hope people can relax to some extent after knowing the information provided by the system.

In our project, we are going to process the doctor notes as our main text data. Since different doctors may have various styles of writing the discharge summary report, the first task we need to do is to clean and organize the summary into admission information and ground truth output information. Our aim is to classify the doctor notes into several sections that the history of present illness and past medical history will be used as our admission information used as our input data. After that, we will build two NLP models to predict the diagnosis and the risk level of the illness separately. By the first model, we will provide the patient with the top three diagnoses with percentage. The second model will predict the risk level in the range of low, medium and high. The length of stay in the hospital and the diagnosis extracted from the discharge summary will be used to evaluate our model in the future work.

## 2 Data

The dataset we are going to use in this project is Medical Information Mart for Intensive Care III (shortened as MIMIC-III) which is a clinical database integrating comprehensive clinical data of patients who were admitted to the Beth Israel Deaconess Medical Center in Boston. The MIMIC-III database was constructed with data from several sources, which include archives from critical care information systems, hospital electronic health records, and SSA Death Master File (Johnson et al., 2016). This database contains information from the patient's admission to discharge with all procedures and lab tests results. The following shows the schema of the database (Figure 1). There are about 58,976 admission records in this database.

| Table | Children | Parents | Columns | Rows | Comments |
|---|---|---|---|---|---|
| admissions | 18 | 1 | 19 | 58,976 | Hospital admissions associated with an ICU stay. |
| callout | | 2 | 24 | 34,499 | Record of when patients were ready for discharge (called out), and the actual time of their discharge (or more generally, their outcome). |
| caregivers | 7 | | 4 | 7,567 | List of caregivers associated with an ICU stay. |
| chartevents | | 5 | 15 | 330,712,483 | Events occuring on a patient chart. |
| chartevents_1 | | | 15 | 38,033,561 | Partition of chartevents. Should not be directly queried. |
| chartevents_10 | | | 15 | 9,584,888 | Partition of chartevents. Should not be directly queried. |
| chartevents_11 | | | 15 | 470,141 | Partition of chartevents. Should not be directly queried. |
| chartevents_12 | | | 15 | 265,413 | Partition of chartevents. Should not be directly queried. |
| chartevents_13 | | | 15 | 39,090,570 | Partition of chartevents. Should not be directly queried. |
| chartevents_14 | | | 15 | 100,075,138 | Partition of chartevents. Should not be directly queried. |
| chartevents_2 | | | 15 | 13,116,197 | Partition of chartevents. Should not be directly queried. |
| chartevents_3 | | | 15 | 38,657,533 | Partition of chartevents. Should not be directly queried. |
| chartevents_4 | | | 15 | 9,374,587 | Partition of chartevents. Should not be directly queried. |
| chartevents_5 | | | 15 | 18,201,026 | Partition of chartevents. Should not be directly queried. |
| chartevents_6 | | | 15 | 28,014,688 | Partition of chartevents. Should not be directly queried. |
| chartevents_7 | | | 15 | 255,967 | Partition of chartevents. Should not be directly queried. |
| chartevents_8 | | | 15 | 34,322,082 | Partition of chartevents. Should not be directly queried. |
| chartevents_9 | | | 15 | 1,274,692 | Partition of chartevents. Should not be directly queried. |
| cptevents | | 2 | 12 | 573,146 | Events recorded in Current Procedural Terminology. |
| d_cpt | | | 9 | 134 | High-level dictionary of the Current Procedural Terminology. |
| d_icd_diagnoses | 1 | | 4 | 14,710 | Dictionary of the International Classification of Diseases, 9th Revision (Diagnoses). |
| d_icd_procedures | 1 | | 4 | 3,898 | Dictionary of the International Classification of Diseases, 9th Revision (Procedures). |
| d_items | 8 | | 10 | 12,487 | Dictionary of non-laboratory-related charted items. |
| d_labitems | 1 | | 6 | 753 | Dictionary of laboratory-related items. |
| datetimeevents | | 5 | 14 | 4,485,937 | Events relating to a datetime. |
| diagnoses_icd | | 3 | 5 | 651,047 | Diagnoses relating to a hospital admission coded using the ICD9 system. |
| drgcodes | | 2 | 8 | 125,557 | Hospital stays classified using the Diagnosis-Related Group system. |
| icustays | 8 | 2 | 12 | 61,532 | List of ICU admissions. |
| inputevents_cv | | 4 | 22 | 17,527,935 | Events relating to fluid input for patients whose data was originally stored in the CareVue database. |
| inputevents_mv | | 5 | 31 | 3,618,991 | Events relating to fluid input for patients whose data was originally stored in the MetaVision database. |
| labevents | | 3 | 9 | 27,854,055 | Events relating to laboratory tests. |
| microbiologyevents | | 3 | 16 | 631,726 | Events relating to microbiology tests. |
| noteevents | | 3 | 11 | 2,083,180 | Notes associated with hospital stays. |
| outputevents | | 5 | 13 | 4,349,218 | Outputs recorded during the ICU stay. |
| patients | 19 | | 8 | 46,520 | Patients associated with an admission to the ICU. |
| prescriptions | | 3 | 19 | 4,156,450 | Medicines prescribed. |
| procedureevents_mv | | 5 | 25 | 258,066 | Procedure start and stop times recorded for MetaVision patients. |
| procedures_icd | | 3 | 5 | 240,095 | Procedures relating to a hospital admission coded using the ICD9 system. |
| services | | 2 | 6 | 73,343 | Hospital services that patients were under during their hospital stay. |
| transfers | | 3 | 13 | 261,897 | Location of patients during their hospital stay. |
| **40 Tables** | | | 534 | 728,556,685 | |

Figure 1: MIMIC-III database schema

In this project, we will focus on the admission report, which will be summarized from the NOTEEVENTS table which contains 2,083,180 entries of notes. The NOTEEVENTS table contains a

large amount of plain text entered by doctors when the patient is discharged from hospital. It may includes information such as admission information (previous illness, medication background, family history, social history, etc.), procedure information (ordered lab tests, test results, services, medication used, etc.) and discharge summary (full diagnosis, follow-up plan, etc.). However, since the notes were entered by different doctors and every doctor had different styles, each note may contain distinguishable sections.

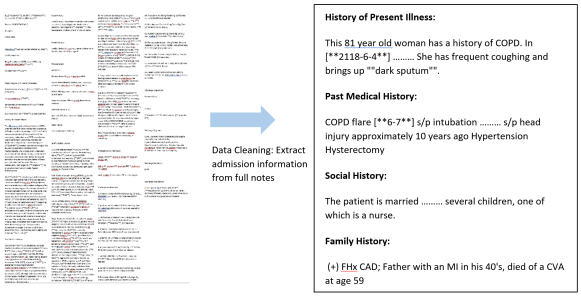


Figure 2: Sample admission information extraction from NOTEEVENTS table

For this project, we will extract specific admission information from the NOTEVENETS table, such as present illness, medical history, family history, social history, etc. and use the text as our input of the training dataset (Figure 2). The output of the training dataset is the ground truth of diagnosis for each patient which is represented as International Classification of Diseases Version 9 (ICD-9) code in the D_ICD_DIAGNOSES table. The other output of risk level is computed from mortality and length of stay (data distribution can be found in Table 1 and Table 2 ) from ADMISSION table, we will further classify into three ordinal categories: stay shorter than 5 days (low risk), stay longer than 5 days (medium risk), dead (high risk). Once the prediction is done, we will evaluate the estimated diagnosis and the risk level with the ground truth output as mentioned above.

## 3 Related Work

In Clinical Outcome Prediction from Admission Notes using Self-Supervised Knowledge Integration (van Aken et al., 2021), authors use the admission information from MIMIC-III dataset in order to solve the problem of preventing doctors from overlooking possible risks and help hospitals to plan capacities. In this paper, they conducted an admission to discharge prediction with admission

| Mortality | |
|---|---|
| **0** | **1** |
| 43,609 | 5,136 |

Table 1: Mortality distribution

| Length of Stay(in days) | |
|---|---|
| **<=3** | **>3 & <=7** |
| 5,596 | 16,134 |
| **>7 & <=14** | **>14** |
| 13,391 | 8,488 |

Table 2: Length of Stay distribution

information such as present illness when arrived at hospital, physical examination results when admission, family history and social history, and therefore predict the patient's diagnosis, doctor's procedures of treatment, possibility of in-hospital mortality and estimated length of stay. Their approaches are using the BERT and BioBERT to extract admission summary from discharge summary and pretrained patients information and article information (PubMed, Wikipedia, MedQuad) for prediction at a relatively high accuracy rate. Their evaluation were using the hyper-parameter from other authors and compare the results with the theirs while using the same model.

Another work done in November 2020 on ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission (Huang et al., 2019) also used MIMIC-III clinical notes to make predictions. In this paper, authors aimed to use admission and discharge summaries to estimate whether the patient will come back and be readmitted. Authors applied bidirectional encoder representations from transformers (BERT) to clinical text since the publicly-released BERT parameters are trained on standard corpora such as Wikipedia and Book Corpus, which differ from clinical text. They pretrained BERT use the clinical text and transformed as Clinical BERT to predict readmission which made more suitable for clinical prediction. This method was evaluated by comparing the efficiency with traditional models, such as BERT, Word2Vec and FastText.

Similarly, patient's length of stay can also be predicted using clinical information, such as drugs and prescriptions. In Using Clinical Drug Representations for Improving Mortality and Length of Stay Predictions (Bardak and Tan, 2021), authors were using drugs representations which ordered to

patients in hospital with over 15 years old. Authors conducted transformations from drug code to FDA drug name as plain text and their chemical representations as latent space to predict binary outcome of in-hospital mortality and length of stay (whether longer than 3 days). Their work of adding drug representations as additional features helps improve the length of stay prediction for AUROC around 6%. The proposed model is compared with a statistical time-series model to evaluate its accuracy of prediction.

Some other literature working on the diagnosis prediction were not using the MIMIC data but use relatively more massive Electronic Health Records. In the paper MNN: Multimodal Attentional Neural Networks for Diagnosis Prediction (Qiao et al., 2019), authors used a segment of continuous patient records including discrete medical codes and textual notes of discharge summary as their training data. They proposes Multimodal Attentional Neural Networks (MNN) in order to model the medical codes and clinical notes in a unified framework. The methods of using both the text feature selection and medical code feature selection helped to produce multimodal core features. The multimodel attentional neural networks shown that experimental results on real world EHR dataset were demonstrated the good performance. Their evaluation using comparison between their MNN model with traditional baseline models, such as Dipole, Retain, DoctorAI, PacRNN and RNN-multimodal.

Another paper using the Electronic Health Reord is from A Deep Learning Pipeline for Patient Diagnosis Prediction Using Electronic Health Records (Franz et al., 2020) which introduced a deep learning model in order to improve the prediction accuracy and efficiency for patient diagnosis. This paper aimed to solve the problem caused by COVID-19 situation, such as multi-reason mortality, and produced an advancement in both data representation and development of machine learning architectures. Their methods included transforming the medical records into FHIR format, grouping numerical observations into bins, etc. This helped authors to introduce an updated version of ClinicalBERT into ClinicalBERT_Multi. And it finally produced an outstanding results in diagnosis prediction. In this paper, the evaluation methods were used in comparing similar model results applying in the same dataset, such as SHiP, DeepObserver-FCNN, DeepObserver-RNN, and ClinicalBERT_Binary.

The above literature were focused on doctor oriented text memorization and for helping doctors to estimate the patient's length of stay. Since their work paid more attention in the medication background of the patients, we focus on to provide a patient oriented system with or without medication knowledge in necessary for all patient ages. This project will provide prediction only on the patient's information before going to hospital, and absolutely, without any lab test results, which will eliminate the risk of time-consuming and high cost of examination. Our approach will benefit patients and give them helps in over viewing their disease and risk level and relax them from anxiety.

## 4 Methodology

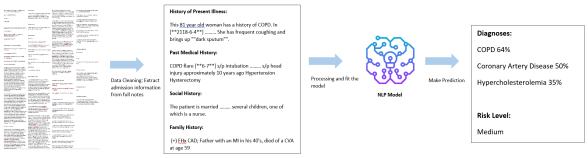


Figure 3: Sample workflow from original text to prediction

Since the data is too huge to process in the local computer, we will need to download the data and then upload to a cloud computer for combining several useful tables according to important columns. To better test the model, we will first randomly select 10,000 samples. While the random samples go smoothly in our model, we will implement our model in the full dataset which includes around 65,000 admission records. Having the necessary data in cloud, we will need to extract the notes from NOTEEVENT table and clean the data to retrieve admission information from doctors report. We will also need to extract diagnosis and length of stay as our ground truth for training and later evaluation. The admission information will be processed and transformed into a suitable format for fitting the model. We will also use the PATIENT table to match the subject id in the NOTEEVENT table. From this new table, we are able to extract more personal information of patients such as gender, ethnicity, and age. To deal with the noise data, we will first make some spelling checking on text datas to make sure the quality of them. To accomplish this, we will use the python package pyspellchecker to help to find the possible spelling errors to give the most likely correction at the same time. The limitation here is we can hardly make sure the pack-

age works well on our dataset and we may consider applying it on randomly sampling data. We will consider using a new spelling corrector depending on the performance of the current one. The next task is similar to a classification task. To organize the long verbose admission notes, we are going to extract several information from them by tracking the possible keywords. For example, we may look for words like "family history" to extract the record since this word is frequently used in the dataset. During this process, we came up with a question that since each patient will have similar personal information in each admission note, we wondered whether it is useful using only one admission note for each patient. After testing on this idea, we found out that they only left with around 10,000 unique patients which left us a pretty tiny dataset to analyze. We finally decided to keep all admission notes used to complete our model. However, having data which are correlated with each other, we may need to add random effects to the model to avoid the dependent drawbacks. After cleaning the admission notes, we are going to use the word2vec technique to build our own document-term matrix. To select the keywords as features, we will filter out some light important terms such as the stop words, rare words. We are still struggling to find the best method to accomplish this task. The current idea is to do the frequent word subsampling to implement the rare words. We will set up a minimum frequency threshold on cleaned data later and will adjust it according to the results. With the matrix, we may move to the next step to build our NLP model. After implementing the model, we will train the model using our training data. The prediction will return our model's estimated diagnoses and risk level (Figure 3). A test dataset will be used to calculate the accuracy and make evaluation. The model implementation will iteratively be improved based on the evaluation of model performance.

## 5 Evaluation and Results

To evaluate the model, we will use the information of the length of stay in the hospital and the diagnosis extracted from the discharge summary to evaluate our model. The first baseline model will be built by randomly generating the diagnosis. In our case, we will use the unique diagnosis list with ICD-9 code (which is a standard categorical method to represent disease name and description) and randomly generated conditional on the probability of each diagnosis. The second baseline for this classification task will be use the most frequent diagnosis to predict. To find out the most frequent diagnosis, we need to first count all patients' diagnosis and simply use it as our prediction output.

For the fist baseline (random), we extracted all unique diagnosis in the format of ICD-9 code and calculated the relative probability of occurrence for each diagnosis. The prediction is from a random choice and the probability. As a result of this baseline model, the accuracy score of the random prediction can reach to 1.4577%.

For the second baseline (most frequent class), we extracted all unique diagnosis in the format of ICD-9 code and count their frequency. The most frequent diagnosis is marked as "V3000", which represents "Single liveborn, born in hospital, delivered without mention of cesarean section". Simply use this most frequent class as our output, we calculate the accuracy score of around 6.5135%.

Below is the table showing true value and prediction values from two baseline models.

| | Truth_Diagnosis | baseline1_random | baseline2_most_freq |
|---|---|---|---|
| **0** | 40301 | 9802 | V3000 |
| **1** | 40301 | 0389 | V3000 |
| **2** | 53100 | 32381 | V3000 |
| **3** | 1915 | 431 | V3000 |
| **4** | 41401 | 51881 | V3000 |
| **...** | ... | ... | ... |
| **58924** | 566 | 4555 | V3000 |
| **58925** | 43411 | 9951 | V3000 |
| **58926** | 34680 | 430 | V3000 |
| **58927** | 0529 | 99859 | V3000 |
| **58928** | 7842 | 55229 | V3000 |

Figure 4: Truth and Baseline Outputs

## 6 Work Plan

Here is the weekly plan till the end of the semester:

3.7-3.13 Organize the admission notes, finish spelling checking

3.14-3.27 Clean admission notes to fit into model

3.28-4.3 Start to train model to predict diagnosis and risk level(separately for each group member)

4.4-4.10 Evaluate and improve model, start preparing presentation and writing final reports

4.11-4.17 Finish reports and presentation

## References


Batuhan Bardak and Mehmet Tan. 2021. Using clinical drug representations for improving mortality and length of stay predictions. In *2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–8. IEEE.

Leopold Franz, Yash Raj Shrestha, and Bibek Paudel. 2020. A deep learning pipeline for patient diagnosis prediction using electronic health records. *arXiv preprint arXiv:2006.16926*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Zhi Qiao, Xian Wu, Shen Ge, and Wei Fan. 2019. Mnn: multimodal attentional neural networks for diagnosis prediction. *Extraction*, 1:A1.

Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix A Gers, and Alexander Löser. 2021. Clinical outcome prediction from admission notes using self-supervised knowledge integration. *arXiv preprint arXiv:2102.04110*.