ICD Tagging and Data Exploration with the Mimic-III Dataset

Charlie Denhart
DS4900 Spring 2020
denhart.c@husky.neu.edu

April 18, 2020

1 Introduction

Over the past decade, the usage of electronic medical records has accelerated up to nearly 96% adoption [1]. Although this has increased some issues within the health care industry, it has also provided significant new opportunity to leverage technological advancements. The initial goal of this project was simply to identify an area in the health care space that could benefit from machine learning techniques and to implement a solution using publicly available data. Several possibilities were considered and are discussed in this paper. Ultimately, the task of ICD code tagging was selected as the central focus of this project.

Assigning ICD codes, classifications for diseases and procedures, to patient records is a significant cost for health care providers due to the manually intensive labeling process involved. To ease this burden, recent studies have attempted to apply deep learning models to the task in order to enable the assignment of ICD codes autonomously [1]. In this project, I attempted to reproduce some of the work done in such studies and expand upon them by utilizing more contemporary methods of embedding text. [2]. It should be noted, however, that this project was not absent of setbacks. Several false assumptions and misallocations of time perhaps attributable to my minimal research experience resulted in some more limited results than desired; results using bert embeddings were not obtained and models were only trained on a subset of the data. Regardless, there are some promising outcomes.

2 Related Work

As mentioned previously, the work in this project is largely inspired by previous studies that attempted to integrate deep learning with the task of ICD code tagging. Ayyar et al. and Huang et al. [1] [3] both took approaches similar to those in this paper in the sense that ICD tagging was treated as a multi-label classification task and that input data derived from the text in clinical notes. In this task, some arbitrary number of ICD codes can be assigned to each clinical note. The same baseline models and deep learning architectures are also utilized. A primary difference, however, is the selection of notes used. Both previous studies limited the clinical notes used to those that were labeled "Discharge Summaries" whereas I took random samples from the full dataset. In hindsight, utilizing the discharge summarizes may have been a wiser approach as it refined the dataset, making for more practical modeling. However, it would be interesting to explore if models can be improved by utilizing all notes involved.

Other work that was considered includes Gabriel et al. [4], which attempted to model patient trajectories as opposed to doing ICD tagging. An issue with the approach, however, is the fact that very few readmissions occur in the dataset used, resulting in a very limited dataset for attempting to forecast future patient events. Another study [5] attempted to learn a representation for patient care events. Although I did not choose to explore that route, I think it would make for very interesting future work.

3 Data

3.1 Source

I am using the MIMIC-III dataset [6], which consists of anonymous data from over 40,000 patients admitted to critical care units at Beth Israel Deaconess Medical Center from 2001 to 2012. It is structured as a relational database, consisting of 26 total tables. Each table represents a different portion of any individual patient's experience with the hospital. Events corresponding with a patient, such as laboratory results and measurements, are each tracked in its own table. There are also dedicated tables to track the stay of a patient. Most importantly, there are several tables devoted to recording ICD-9, CPT, and DRG codes associated with various procedures and diagnosis.

Category	Totals	Male	Female	Private	Medicare	Medicaid	Government	Self Pay
Patients	46520	26121	20399	19663	21002	4570	1614	600
Admissions	58976	32950	26026	22582	28215	5785	1783	611
ICD9 Codes	11501	5852	5649	5718	5023	3989	2657	1467
Deaths	5836	3141	2695	1372	3903	381	90	95

Table 1: Summary statistics of MIMIC-III dataset.

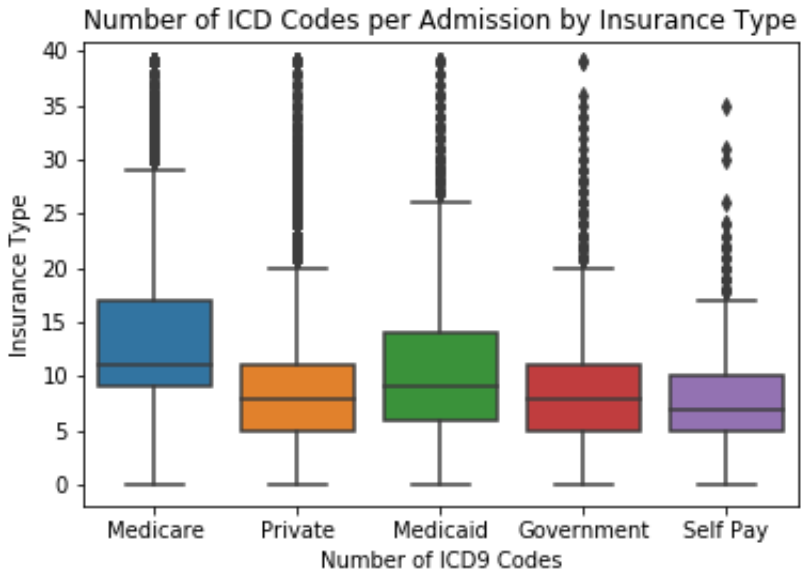

Figure 1: Number of ICD codes by insurance type.

3.2 Data Overview

This dataset is incredibly rich and, unfortunately, only the surface was scratched throughout the course of this project. For example, there is variation in ICD code assignment by insurance type, as depicted in Figure:1. Although there is certainly nothing to be concluded from this chart, it inspires some interesting questions about the role of insurance in code assignment and, ultimately, billing for procedures. This is another area of possible future work, particularly considering the relevance to current political climates.

3.2.1 Clinical Notes

Although there are many plausible indicators of diagnoses within the MIMIC-III dataset, features extracted from clinical notes are the primary independent variables for models. Clinical notes are suitable as they contain lots of rich, albeit unstructured, information about the patient and make for a more interesting academic experiment. It was originally desired to incorporate additional features, but this stage of the project was not reached in time.

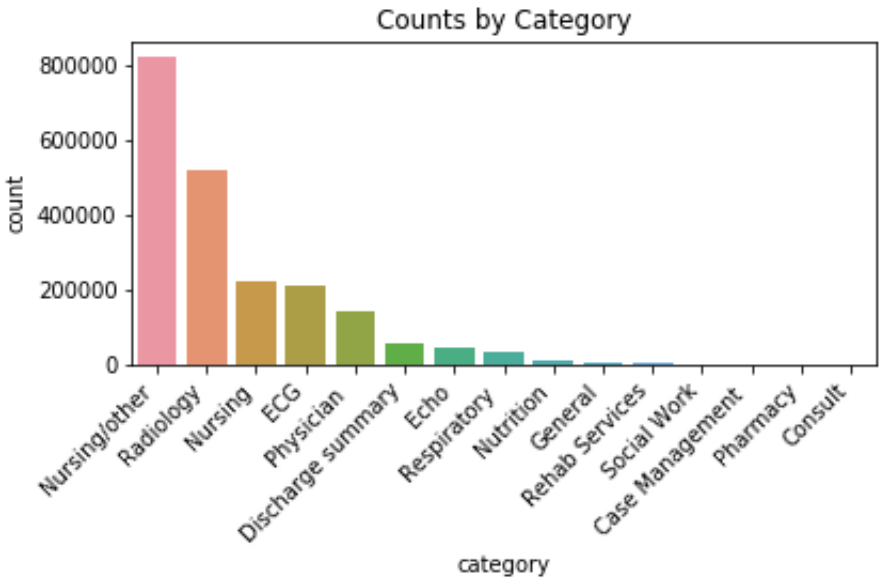

Figure 2: Distribution of clinical note categories.

In addition to joins with other tables that can be performed to associate additional data with the notes, each note is also assigned a category that classifies its subject. Although there are 15 different possible categories, "Nursing" related categories are predominant. It was not explored if these disproportions have any effect on model results, but it is a topic that could be investigated.

3.2.2 ICD Codes

As the items being forecasted, ICD Codes are likely the most important variable to be considered in the dataset. This data utilizes ICD9 codes, although Beth Israel Deaconness Medical Center changed over to ICD10 codes in 2015. Although the ICD9 codes are less abundant than the updated ICD10 codes, they are still incredibly granular [4]. A useful feature of the codes, though, is that they are structured in a hierarchy similar to the diagnoses and procedures that they classify. Therefore, it is possible to reduce the code set while still maintaining some of their relationships. This is discussed further in the methodology section. Despite this reduction, 2 also indicates that there is some significant class imbalance. Undersampling was considered, but the severe lack of a few classes would have lead to too small of a dataset. The best solution may be to simply drop out these classes, but class balances were ultimately left untouched for this study.

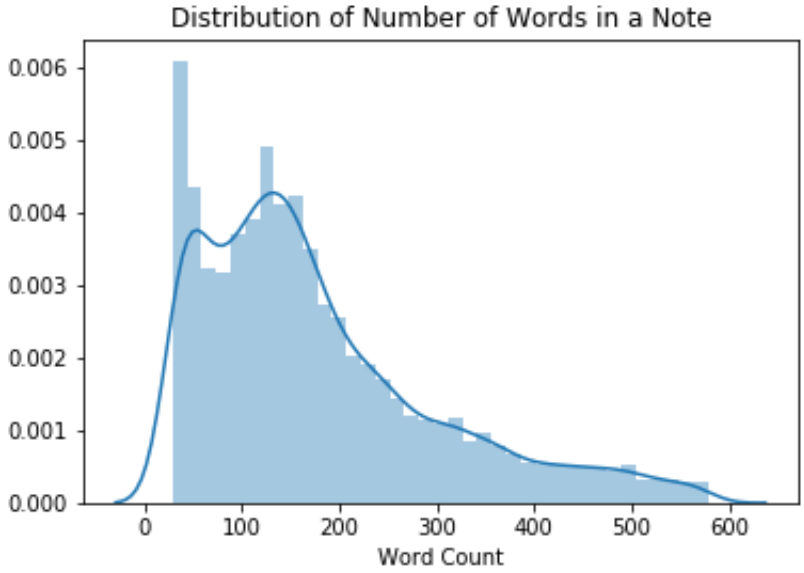

Figure 3: Distribution of word counts within clinical notes. Note that only the middle 80 percent of the distribution is shown to omit outliers.

	Code	Mimic-iii Counts	Number of Leaves	Description
0	001-139	173037	871	INFECTIOUS AND PARASITIC DISEASES
1	140-239	56220	660	NEOPLASMS
2	240-279	256828	278	ENDOCRINE, NUTRITIONAL AND METABOLIC DISEASES, AND IMMUNITY DISORDERS
3	280-289	0	10	DISEASES OF THE BLOOD AND BLOOD-FORMING ORGANS
4	290-319	106277	307	MENTAL DISORDERS
5	320-389	153848	1334	DISEASES OF THE NERVOUS SYSTEM AND SENSE ORGANS
6	390-459	334785	403	DISEASES OF THE CIRCULATORY SYSTEM
7	460-519	269565	217	DISEASES OF THE RESPIRATORY SYSTEM
8	520-579	192481	519	DISEASES OF THE DIGESTIVE SYSTEM
9	580-629	213352	382	DISEASES OF THE GENITOURINARY SYSTEM
10	630-679	63	327	COMPLICATIONS OF PREGNANCY, CHILDBIRTH, AND THE PUERPERIUM
11	680-709	77158	202	DISEASES OF THE SKIN AND SUBCUTANEOUS TISSUE
12	710-739	63633	417	DISEASES OF THE MUSCULOSKELETAL SYSTEM AND CONNECTIVE TISSUE
13	740-759	0	19	CONGENITAL ANOMALIES
14	760-779	112308	219	CERTAIN CONDITIONS ORIGINATING IN THE PERINATAL PERIOD
15	780-799	195940	308	SYMPTOMS, SIGNS, AND ILL-DEFINED CONDITIONS
16	800-999	221638	1558	INJURY AND POISONING
17	V01-V91	283792	985	SUPPLEMENTARY CLASSIFICATION OF FACTORS INFLUENCING HEALTH STATUS AND CONTACT WITH HEALTH SERVICES

Table 2: Breakdown of top level ICD codes.

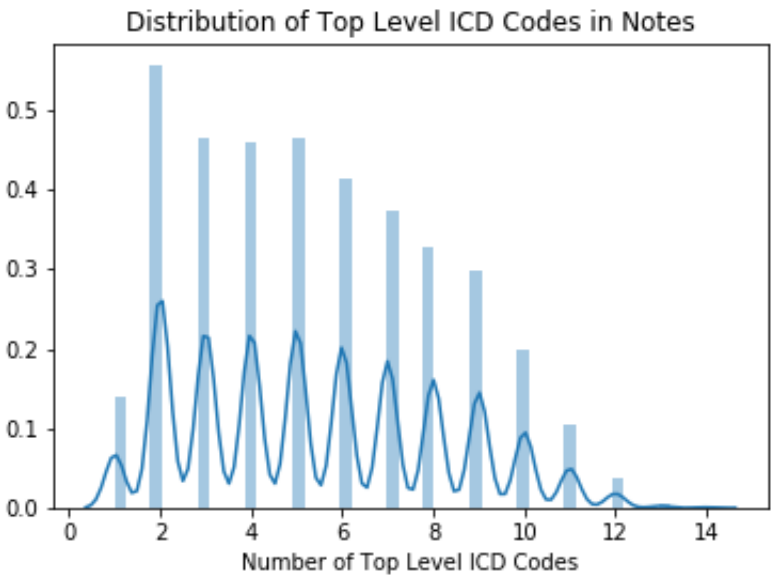

Figure 4: Distribution of ICD codes per note.

4 Methods

4.1 Preprocessing

4.1.1 Clinical Notes

Initial preparation of the clinical notes included tokenizing words and removing stop words. Tokens that filled redacted entities (for anonymization purposes) were also removed [3]. In order to embed these tokens, word2vec embeddings pretrained on Google News articles were utilized. For the baseline models, a document level embedding was obtained through element-wise averaging of the word embeddings per note. For sequence models, the word2vec word embeddings and bert embeddings pre-trained on the MIMIC-III dataset were implemented. A model trained with the bert embeddings, as mentioned above, was not trained in time for the submission of this paper, however.

4.1.2 ICD Codes

Although it was initially the goal to train models on the full set of ICD9 codes, there are over 11,000 distinct codes as shown in Table 1 and fewer than 60,000 admissions total to support them. Due to this granularity and lack of training data, only the top layer of the ICD9 code hierarchy are used for initial modeling [4]. As seen in Table 2, even the top layer of the codes are not very evenly distributed so future work may include either replicating underrepresented observations or dropping them out.

Figure 5: A sample of common words in a note from the 'Nursing' category.

4.2 Modeling

The baseline model is a logistic regression. Utilizing softmax, it assigns a probability to each top level ICD class. A prediction for a given note is formed, therefore, by taking the top level ICD codes that have a probability above a given decision threshold. A random forest was also implemented to supplement the baseline models. These baseline models in the context of this study are examples of problem transformation methods, which are required in order to convert the inherently binary (or multiclass) classifiers to multilabel classifiers [1]. To improve upon the baseline model, a multilayer perceptron is also trained that outputs a sigmoid layer. The network has two hidden layers, each with relu activation funtions. All three of these models were trained on 50 thousand randomly sampled observations from the dataset and were implemented using the scikit-learn library in python.

Finally, a long short-term memory model was implemented as an attempt to leverage sequential relationships (left-to-right) between words in clinical notes. The architecture was constructed using the PyTorch library and consists of a single LSTM gate with hidden layers of size 128. The hidden layer from the last sequence in a note is then fed into a linear layer that produces an output for each top level ICD code. Each of this outputs are additionally fed through a sigmoid function in order to obtain class-exclusive distributions. Due to the length of time that this model took to train and the trial and error involved with getting it functional, the model was only trained on a random sample of five-thousand observations.

Figure 6: Comparison of models through traditional metrics.

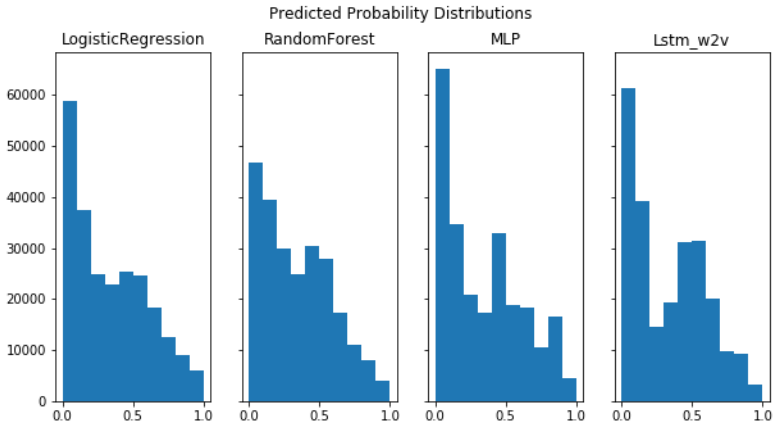

Figure 7: Distributions of model predicted probabilities.

5 Results

Ultimately, results were a mixed bag. The baseline models performed surprisingly well despite only being trained on a subset of the data. In fact, they performed very similarly to baseline methods implemented in Ayyar et al. and Huang et al. Another surprising factor, though, is that results between the three baseline models vary little. More in-depth investigation would be interesting in order to understand if these models are learning particular mappings between embeddings and ICD codes. Although not included in this paper due to the bulkiness of the table, performance metrics were also derived on a per-class basis. Although the overall metrics reported in Figure 6 were weighted by class, the absence of a few classes certainly appears to weight on performance.

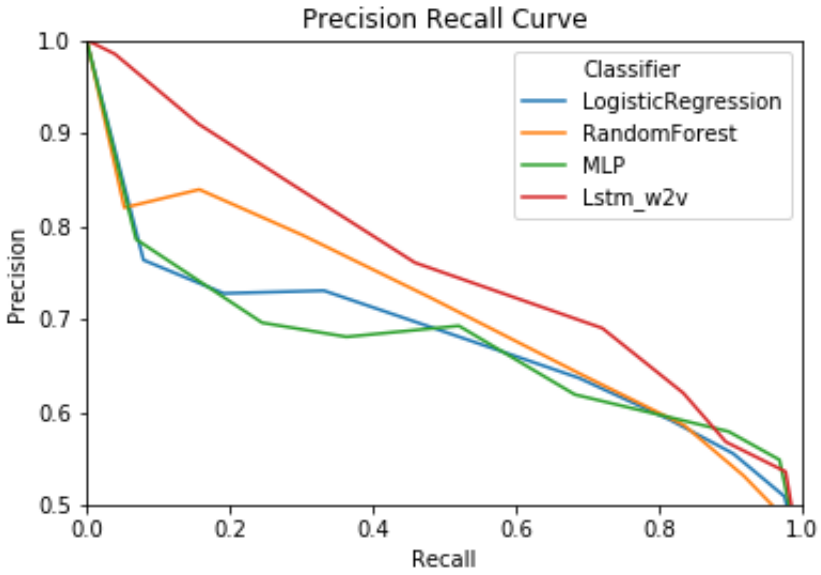

Figure 8: Plot of precision and recall scores per model at various decision thresholds.

The most discouraging aspect of the results came from the LSTM model. This perhaps should not be surprisingly as it was trained on such a limited set of the data, but it was hoped that its abilities to capture relationships between words in a note might be able to produce better outcomes.

6 Future Work

As has been mentioned several times previously, all of the desired results from this study were not obtained. The simplest improvement that will be attempted is to simply run the existing long short-term memory model with more observations, since it is ready to go and training time is no longer an obstacle. It would also be very interesting to have results from the LSTM with BERT embeddings integrated (as opposed to word2vec embeddings). This functionality is essentially implemented already and appears to work on limited runs, but has consistently failed when run due to memory failure, despite being executed on a server with 100Gb of memory available. I will continue to debug these issues.

Aside from the task of ICD code tagging, it would be interesting to explore some of the other topics mentioned throughout this paper. Only a small subset of the MIMIC-III data has been utilized throughout this project, so I am hoping to identify projects that might integrate more of this data and possibly join it with additional sources. One particular pursuit that would leverage deep learning techniques would to extend work done by

Beaulieu et al. and attempt to learn a representation for patient care events. Such a representation could be very useful when attempting to model patient outcomes and possibly even when comparing patient groups segregated by differences such as insurance type.

References

[1] Sandeep Ayyar and IV OliverBearDon'tWalk. Tagging patient notes with icd-9 codes. 2017.

[2] Artuur Leeuwenberg and Marie Francine Moens. Structured learning for temporal relation extraction from clinical records.

[3] Huang, Osorio, Cesar, and Luke Wicent. An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes, Jun 2019.

[4] Gabriel and Sihem. Patient trajectory prediction in the mimic-iii dataset, challenges and pitfalls, Nov 2019.

[5] Brett K. Beaulieu-Jones, Patryk Orzechowski, Jason H. Moore, Adriana Mihaela Coroiu, Alexandr A Kalinin, and Computational Genetics Lab. Mapping patient trajectories using longitudinal extraction and deep learning in the mimic-iii critical care database.

[6] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016.