

Predicting Discharge Disposition Using Patient Complaint Notes in Electronic Medical Records

Mohamad Salimi

Department of Computer Science
Queens College, CUNY
Krasnoff Quality Management Institute
Northwell Health
mohamad.salimi45@gmail.cuny.edu

Alla Rozovskaya

Department of Computer Science
Queens College, CUNY
arozovskaya@qc.cuny.edu

Abstract

Overcrowding in emergency rooms is a major challenge faced by hospitals across the United States. Overcrowding can result in longer wait times, which, in turn, has been shown to adversely affect patient satisfaction, clinical outcomes, and procedure reimbursements. This paper presents research that aims to automatically predict discharge disposition of patients who received medical treatment in an emergency department. We make use of a corpus that consists of notes containing patient complaints, diagnosis information, and disposition, entered by health care providers. We use this corpus to develop a model that uses the complaint and diagnosis information to predict patient disposition. We show that the proposed model substantially outperforms the baseline of predicting the most common disposition type. The long-term goal of this research is to build a model that can be implemented as a real-time service in an application to predict disposition as patients arrive.

1 Introduction

Studies show that wait times not only affect patient satisfaction, but also the perception of providers and quality of care (Chandra et al., 1981). Furthermore, the Center for Medicare and Medicaid Services is tying reimbursements to the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) scores. As a financial and patient experience priority, hospitals are focused on addressing issues that affect patient satisfaction. One common issue is long wait times in the emergency rooms, that are due to high volume

and overcrowding. Another issue is that of bed management and its effect on wait times. If no in-patient beds are available, admitted patients are kept in the emergency department until beds open. This is commonly referred to as patient boarding and has been shown to negatively affect outcomes and wait times. (Felton et al., 2011).

Improved bed management and resource utilization are necessary to achieve shorter wait times. This paper describes a first attempt at an experimental model which aims to predict discharge disposition based on chief complaint (i.e. symptoms description) and diagnosis information contained in clinical notes. The corpus that we use contains approximately 260,000 annotated emergency department records. The records contain free text of a complaint and admit diagnosis, and are labeled with the disposition information. The disposition, which is the destination after medical treatment, can be classified as *Admit*, *Discharge*, *Observation*, *Expire*, *Left Against Medical Advice (AMA)*, *Asthma Observation Unit (AOU)*, *Eloped*, or *Transfer*.

A model that predicts disposition type could be realized as an informational alert system integrated with electronic medical record software. Patient complaints are made available before discharge dispositions, allowing for an immediate prediction of disposition. In some cases the complaint is available hours before the discharge diagnosis or the disposition. Thus, such a model could provide integrated real-time forecasting on potential discharges and in-patient admissions.

The rest of the paper is organized as follows. Sec. 2 presents related work. Sec. 3 describes the corpus. Sec. 4 presents the experimental setup. Results are reported in Sec. 5. We present error analysis in Sec. 6 and conclude in Sec. 7.

2 Related Work

The Academic Emergency Medicine journal published preliminary results that attempt to predict emergency department in-patient admissions to improve same-day patient flow (Peck et al., 2012). They used three methods – expert opinion, Naive Bayes, and a generalized linear regression model – to analyze two months worth of emergency department data from the Boston VA healthcare System. However, Peck et al. (2012) focused strictly on predicting *admit* dispositions only, while we aim to predict all possible outcomes. Furthermore, their results focus strictly on structured fields, such as urgency level, age, sex, chief complaint, and the provider seen, while we work with free text in clinical notes. Another issue with the above model is that by including the provider seen to predict admission, they are tightly coupling the model to the Boston VA health care System.

Previous work has been done which aims to predict patient outcomes using unstructured text. Yamashita et al. (2016) analyzed admission records of 1,222 patients who had a clinical pathway of cerebral infraction. The goal was to develop a method for automatically performing clinical evaluations and to identify early interventions for cases that may have clinically important outcomes.

There has been a lot of other related work in the NLP area on unstructured electronic medical records and, in particular, in the clinical domain. For example, Jonnagaddala et al. (2015) developed a model to automatically identify smoking status using a SVM model. Jung et al. (2011) extracted events from clinical notes and used this information to construct a timeline of patient medical history. Both of the above mentioned works also used unstructured clinical notes, but focused on identifying patient history information. Cogley et al. (2012) used machine learning to determine whether a patient experienced a particular medical condition. However, while Cogley et al. (2012) looked at patient history and physical examination reports, we wish to predict disposition from complaint and admitting diagnosis alone.

3 Data

The data used in this project is provided by the Krasnoff Quality Management Institute of Northwell Health. Northwell Health is a not-for-profit healthcare network that includes 22 hospitals and

Disposition	Percentage (%)
Admit	30.88
AMA	0.89
AOU	0.05
Discharge	63.66
Eloped	0.27
Expired	0.08
Observation	3.56
Transfer	0.56

Table 1: Distribution of disposition labels in the training corpus.

over 500 medical offices. Krasnoff Quality Management Institute provides analytics support for the many facilities across Northwell Health. The dataset contains de-identified emergency department records from several facilities across the system. Each record includes information about the patient complaint, diagnosis, and the resulting disposition, filled out by clerical staff or nurses. We use a subset of the entire dataset in the present study, approximately 260,000 records. 215,000 are used to train the model, and 45,000 are used for testing.

There are eight possible values for the disposition outcome. Table 1 shows the distribution of the values in the corpus. Note that the outcome types are not evenly distributed. The most common disposition type, *discharge*, accounts for over 63% of all disposition types, and the two most frequent types, *discharge* and *admission* to the hospital, account for over 94% of all disposition types. The *observation* unit, which is an area in some emergency rooms which allows for extended evaluation for patients whose stays will likely be less than one day, follows as the third most common disposition (3.56%). Left against medical advice (*AMA*), asthma observation unit (*AOU*), left without notice (*eloped*), death in the ER (*expired*), and *transfer* to a different facility all account for less than 1% of total number of records.

Example Records Each instance in the dataset contains information about the symptoms, the diagnosis, and is annotated with its final disposition. The notes in the dataset do not contain information related to the treatment of the patient. Below we show several complaint instances from the corpus. As expected, since this information was entered by clinical staff, the text is quite noisy, contains a lot of specific medical abbreviations (“pt”), incomplete sentences, and typos (“cant”).

Data Point	Value
Complaint	flu-like symptoms
Admit diag.	fever
Discharge diag.	viral illness
Complaint	Abdominal pain & heart burn
Admit diag.	NULL
Discharge diag.	enteritis

Table 2: Complaint, admit diagnosis, and discharge diagnosis examples.

- “pt called EMS ‘I cant see’ pt says she cant open her eyes”
- “bite, animal pain in limb puncture wound of left thigh, initial encounter, observation”
- “head injury car passenger injured in collision with two- or three-wheeled motor vehicle in traffic accident, initial encounter mvc (motor vehicle collision)”

The records also contain admitting diagnosis and discharge diagnosis. The admitting diagnosis is entered shortly after the complaint and may be updated by staff. The discharge diagnosis is entered once the patient’s visit is complete. Table 2 shows two examples.

4 Experiments

Our aim is to create a prototype model that will be able to make predictions with the complaint and admit diagnosis extracted from clinical notes. Our model is trained with the Averaged Perceptron (Freund and Schapire, 1999) algorithm implemented with Learning Based Java (Rizzolo, 2011). While classical Perceptron comes with generalization bound related to the margin of the data, Averaged Perceptron also comes with a PAC-like generalization bound (Freund and Schapire, 1999). This linear learning algorithm is known, both theoretically and experimentally, to be among the best linear learning approaches and is competitive with SVM and Logistic Regression, while being more efficient in training. We do not use neural network approaches in this work both due to the moderate size of the dataset (neural models have been shown to have a steep learning curve (Koehn and Knowles, 2017) and also because our goal is to develop a model that would be as efficient as possible. We train the classifier on the training partition of the corpus and report results on the test partition. All the data was normalized by removing special characters, lowercased, and POS

Disposition	Rel. freq. (%)	Accuracy (%)
Discharge	63.4	75.7
Admit	31.1	77.9
Observation	3.6	96.3
Eloped	0.3	0.0
AMA	0.7	0.0
Transfer	0.6	0.0
AOU	0.1	0.0
Expire	0.1	99.9
Total	-	75.7

Table 3: Accuracy results by disposition type.

tagged with the NLTK tagger (Bird, 2006).

4.1 Features

The features include bag-of-word unigrams and bigrams, and collocations. To control for the vocabulary size, we only include the top unigrams and bigrams occurring in the training data. 75 unigram and bigram features are included.

The collocation features are based on a list of keywords extracted from the top 50 words occurring in the training data. Each collocation feature is a conjunction of the keyword, word tokens and part-of-speech tags occurring in the 2-word window around the keyword. Sample collocation features are shown below:

- $W_{i-2}, POS_{i-2}, W_{i-1}, POS_{i-1}, Infection, W_{i+1}, POS_{i+1}, W_{i+2}, POS_{i+2}$
- $W_{i-2}, POS_{i-2}, W_{i-1}, POS_{i-1}, Pain, W_{i+1}, POS_{i+1}, W_{i+2}, POS_{i+2}$

5 Results

We evaluate the model using both accuracy and F-score. Table 3 shows accuracy results by disposition type. We note that the most frequent class baseline that corresponds to selecting the discharge disposition, is 63.4. This is substantially lower than the overall accuracy of 75.7. The accuracy for the *discharge* class is 75.7%, while the accuracy for the second most frequent class, *admit*, is 77.9% (recall from Table 1 that the two labels account for over 94% of all instances in the training data). The performance on the least common disposition labels is poor, with the exception of *expire* (this is further discussed in the next section).

We further evaluate by computing precision, recall, and F-score for each class (Table 4). In general, the performance is higher for more frequent

Disposition	Precision (%)	Recall (%)	F-score (%)
Discharge	75.2	92.0	82.8
Admit	70.4	50.2	58.6
Observation	100.0	0.1	0.2
Eloped	0.0	0.0	0.0
AMA	0.0	0.0	0.0
Transfer	0.0	0.0	0.0
AOU	0.0	0.0	0.0
Expire	64.1	68.0	66.0
Total	73.2	74	70.8

Table 4: Precision, recall, and f-score results by disposition type.

classes, and very poor for the least common labels. The best F-score of 82.8% is achieved for the most frequent class, *discharge*. Again, one interesting exception here is the *expire* class.

6 Error Analysis

We analyze several cases on which the classifier’s predictions were incorrect. The first instance (shown below) had a prediction for *discharge* but the correct label was *admit*.

- abdominal pain pleural effusion in other conditions classified elsewhere pleural effusion associated with hepatic disorder

“Pleural effusions”, which is a condition in which excess fluid buildup is present around the lungs, is a potentially serious condition. In our corpus, pleural effusion cases were over five times more likely to be admitted than discharged. In this case, the addition of “abdominal pain” feature resulted in the classifier considering it a *discharge* record.

The next record was a *discharge* which was predicted to be an *admit*. This may be due to the presence of the word “bleeding”.

- abdominal pain diverticulitis of large intestine without perforation or abscess without bleeding

Finally, some notes are extremely short, such as the complaint “chest pain”, which was labeled as *admit* but the model classified it as *discharge*, due to it being the most common disposition for “chest pain”.

It is clear that this task is challenging, given the brief and noisy nature of the clinical notes, which contributes to data sparseness, and ambiguity of

features that may indicate multiple likely disposition outcomes.

Lastly, some dispositions were not classifiable by the model. In particular, we conjecture that leaving against medical advice (*AMA*) may be tied to factors not seen in symptoms such as social determinants. *Observation* and *transfer* classification may be improved with features that better target those dispositions. Clinical experts will need to be engaged for this task to better understand the feasibility of predicting those dispositions.

7 Conclusion

We presented a model for predicting emergency room disposition from clinical notes. We used a corpus of emergency room records that contains information on symptoms, diagnosis, and disposition labels, entered by medical staff. We showed that the proposed model significantly outperforms the baseline approach of selecting the most frequent class. The nature of the corpus is such that two most common classes account for over 94% of all cases. Although most machine learning problems have to do with label imbalance, we believe that our task is unique in that the imbalance is extreme. The performance of the model is better than the baseline in the most prevalent dispositions, as well as one very rare disposition of *expire*. The other least frequent classes are not classifiable by the model. We hypothesize that some dispositions may be tied to factors not reflected in symptoms, such as social determinants.

Although the results are promising, more work is needed to reach the level where such a model can be utilized in real-time applications. For example, text correction and text normalization of the clinical data might be helpful, given that the notes contain a lot of noise. However, we believe that the proposed experiment is an important step towards building a real-time system that can provide predictions as complaints come into emergency departments. Such a system can be utilized to assist clinical leadership in staffing and operational decisions.

Acknowledgments

We thank the anonymous reviewers for their helpful comments.

References

- Steven Bird. 2006. [Nltk: The natural language toolkit](#). In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72. Association for Computational Linguistics.
- Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. [Alternation](#). *Journal of the Association for Computing Machinery*, 28(1):114–133.
- James Cogley, Nicola Stokes, Joe Carthy, and John Dunning. 2012. [Analyzing patient records to establish if and when a patient suffered from a medical condition](#). In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, BioNLP '12*, pages 38–46, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Brent M. Felton, Earl J. Reisdorff, Christopher N. Krone, and Gus A. Laskaris. 2011. [Emergency department overcrowding and inpatient boarding: A statewide glimpse in time](#).
- Yoav Freund and Robert E. Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine Learning*.
- Jitendra Jonnagaddala, Hong-Jie Dai, Pradeep Ray, and Siaw-Teng Liaw. 2015. [A preliminary study on automatic identification of patient smoking status in unstructured electronic health records](#). In *Proceedings of BioNLP 15*, pages 147–151, Beijing, China. Association for Computational Linguistics.
- Hyuckchul Jung, James Allen, Nate Blaylock, William de Beaumont, Lucian Galescu, and Mary Swift. 2011. [Building timelines from narrative clinical records: Initial results based-on deep natural language understanding](#). In *Proceedings of BioNLP 2011 Workshop*, pages 146–154, Portland, Oregon, USA. Association for Computational Linguistics.
- P. Koehn and R. Knowles. 2017. [Six challenges for neural machine translation](#). In *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics.
- J S Peck, J C Benneyan, D J Nightingale, and S A Gaehe. 2012. [Predicting emergency department inpatient admissions to improve same-day patient flow](#).
- N. Rizzolo. 2011. Learning based programming.
- Takanori Yamashita, Yoshifumi Wakata, Hidehisa Soejima, Naoki Nakashima, and Sachio Hirokawa. 2016. [Prediction of key patient outcome from sentence and word of medical text records](#). In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 86–90, Osaka, Japan. The COLING 2016 Organizing Committee.