Predicting Hospital Re-Admissions: Midterm Report

Cierra Beck, James Calixto, Dana Kim

Abstract—**Hospital re-admission is a large problem currently facing the health care industry. Often, when a patient is released from a hospital, they return within a short period of time because of the same issue. For example, for patients discharged from a hospital after treatment for heart failure, 23% of them return again within 30 days. This situation is not ideal, as it is more expensive for the patient, the hospital, and insurance companies to have the patient come back to the hospital again. In addition, if the patient has to come back to the hospital, the hospital has to re-collect their information and medical.**

Our goal with this project is to predict which patients are likely to return to the hospital. We would like to see what factors make a patient more likely to return, and what diseases are the most common for return risks. Based on this, we would be able to recommend certain patients for longer initial visits or alternative care methods to prevent the risk of return. These results will be beneficial to both the hospital and the patient, as the hospital will ideally be able to adapt their discharge process to prevent patients from having to return.

I. Introduction

For this project, we are trying to develop a predictive model that can assess a patient's likelihood of returning to the hospital within a given time. This will help cut costs for both patients and healthcare providers, since hospital admission is a relatively long process and it is much more efficient for patients in need of short-term subsequent treatment to stay in a short-stay unit within the hospital. With this model, we hope to be able to make decisions on whether a patient should be discharged (low probability of readmission) or be retained (high probability of readmission), as well as predict what treatment they will need in the near future.

II. Preparing the dataset

For this project, we will be using the MIMIC-III (Medical Information Mart for Intensive Care III) dataset, which collates data from over 40,000 patients at critical care units at the Beth Israel Deaconess Medical Center between 2001 and 2012. This dataset is very thorough; it includes information such as patient demographics, laboratory tests and results, prescribed medications, vital sign measurements, and so forth. Due to its thoroughness as well as the complicated nature of hospital records systems, the data is split up between 26 different .csv files which, together, total 43.3 GB. Additionally, text fields such as caregiver notes are not standardized and often contain misspellings as well as non-standard characters such as newlines. Thus, the MIMIC-III dataset constitutes a big and messy dataset and requires cleaning and preprocessing.

A. Data merging

For the initial data exploration, we chose not to look at the 34.5 GB CHARTEVENTS.csv file (which lists all of the charted data - i.e. hourly vital signs and patient status) for two reasons. First, the file is very large which would preclude us from performing the fast data analysis we need to get a good overview of the data. Even simple search and sorting operations would take a while given the 330 million rows of this file. More importantly, we did not think that hourly chart events were as relevant to our overall goal of predicting hospital readmission rates, as even in aggregate this would not yield much information about the hospital stay as a whole. Instead, we chose to perform joins on ADMISSIONS (data about individual hospital admissions), PATIENTS (identities and demographics of patients), and DIAGNOSES_ICD (patient diagnoses) to get a good overall picture of the characteristics of patients and their hospital stays. Importantly, using these three files we can see a patient's history and timeframe of admissions into the hospital and see for what reasons they've been admitted or readmitted.

B. *Missing and cleaning data*

While looking through the data we noticed that some fields were missing. Due to the nature of health care, patients come in with many different conditions and not all fields will be applicable to all patients. For example, if the patient is admitted directly into the emergency department, they will have an associated in/out time specific to the emergency department that other patients will not. Additionally, some patients are simply missing data; ex. marital status, religion, etc. We will have to ensure that when performing aggregate analyses over the whole dataset, we look at features common to all patients; using data specific to certain classes of patients is more helpful when analyzing that particular subgroup.

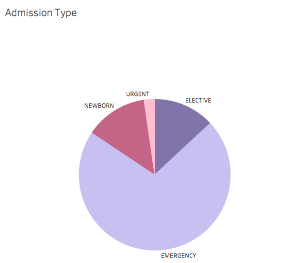
Our data also does not contain a feature that we think would be extremely helpful in analysis: length of stay. To add this to our dataset, we will be adding this field where the value is equal to the difference in time between admission time and discharge time. Our data does not explicitly contain age, so we will be adding a field to represent the patients age at the time of their admission.

C. *Initial Focus Features*

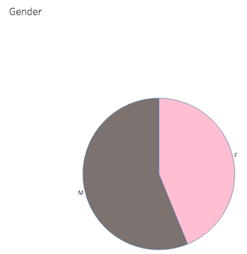
For the purposes of our project, there are several fields that we will be initially focusing on. One of the most important is the number of readmits for each individual patient, as that will allow us to separate the data into two categories: patients who do not get readmitted and patients who do get readmitted. This separation is critical for examining features that can increase risk of readmission. For those who have been readmitted, the reason for admission for each additional visit and the diagnoses are additionally important for our analysis. Age, gender, medical history, smoker vs non smoker, total length of stay, race and insurance provider.

III. Initial Summary Statistics, Visualized

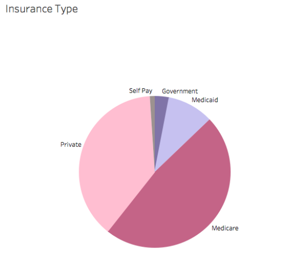
After preparing our dataset, we are examining the statistics for 58,976 total admissions, which includes 46,520 unique patients. Some patient breakdowns are shown below:



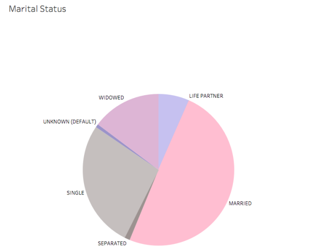
Admission Type. Color shows details about Admission Type. Size shows count of Admission Type. The marks are labeled by Admission Type.



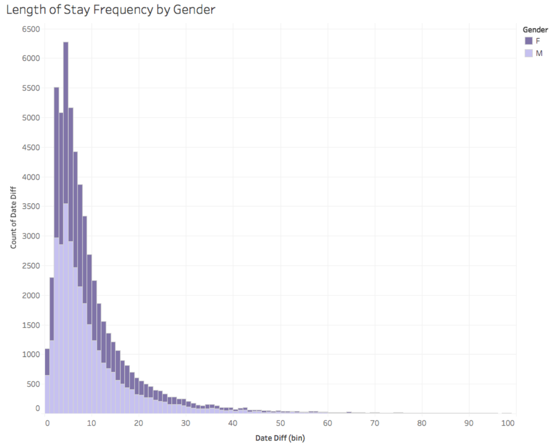
Gender. Color shows details about Gender. Size shows count of Gender. The marks are labeled by Gender.



Insurance. Color shows details about Insurance and Measure Names. Size shows % of Total Count of Insurance. The marks are labeled by Insurance. The view is filtered on Insurance, which keeps Government, Medicaid, Medicare, Private and Self Pay. Percents are based on the whole table .



Marital Status. Color shows details about Marital Status. Size shows count of Marital Status. The marks are labeled by Marital Status.



The trend of count of Date Diff for Date Diff (bin). Color shows details about Gender.

IV. Initial Analysis

To gain further insight into the factors that affect hospital readmission rates, we performed linear regression using the ADMISSIONS and PATIENTS files to determine factors that corresponded to intervals between hospital admissions. We first performed a left join on these two files on the SUBJECT_ID field, giving us a dataset of hospital admissions and information about the patients. We then used the fields in this combined dataset to generate a dataset with four features that we thought were of particular interest:

- Age - since older people often require more medical procedures and can often be in worse health than younger people, we thought that age would be a strong predictor of whether

clients required subsequent admissions sooner rather than later.
- Stay length - it could make sense that longer hospital stays are correlated with more complex procedures, with a correspondingly higher rate of complications and thus necessitating hospital readmission in the near future.
- Gender - we were interested in seeing the effects of gender on hospital readmission intervals.
- Expiration flag - this flag, which is 1 if a person expired after being released and 0 if they have not, was more of a sanity check to determine the relative strength of the other regression variables. Naturally, a deceased person is highly unlikely to be readmitted to the hospital.

After performing linear regression on the dataset, consisting of 56,360 hospital visits and their corresponding parameters, we discovered that the weights of our coefficients were:

Age	-0.323
Stay length	-9.780
Gender	7.881
Expiration	-1169.169

We note that both age and previous stay length are negatively correlated; the younger a person is and the shorter their previous stay is, the longer we expect them to stay out of the hospital. This makes sense. Interestingly we see a strong effect of gender; male patients have longer times between hospital admissions. Finally, we see an extremely strong effect of being deceased, which is what we expected.

V. Further progress

Our goal is to produce a binary indicator that will say whether a patient will be readmitted or not. We will be attempting several methods of classification to obtain this indicator, and then analyzing our different methods for general accuracy. We are also working to generate a prediction on the length of time before your revisit to the hospital.

We will be utilizing cross validation to improve our model accuracy, where we will divide our data into 80 percent training data and 20 percent testing data. The training data will be further divided into an initial training set and a validation set.

We expect to perform more sophisticated analyses on the dataset than regression with four variables. Some things we could use are the diagnosis codes to see the difference in readmission rates for different classes of patients. Some diagnosis codes, such as gunshot wounds, are more serious than others, such as mild flu symptoms. We could also use decision trees on continuous data such as age to separate the patients into different ”classes.” We can expect that, for instance, pregnant women have different risk factors than newborn infants, thus separating our patients into different cases could help improve our classification accuracy.