Hospital Occupancy Prediction: Data and Visualization

Justin Ledford
Eric Smith

September 26, 2018

Abstract

In this paper, we analyze the relationships between various medical claim attributes in order to better understand what sort of health predictors can lead a person to hospital admission. The prediction of a person's number of days spent in a hospital the following year could be a valuable metric, especially for insurance companies that will be paying for these hospital visits.

We conclude that attributes like age, number of medical conditions, and number of claims are good indicators of whether or not people will be admitted into the hospital the following year. Furthermore, we show that severity of sickness (Charlson index) and the number of visits a patient makes to internal specialists, can help us predict how long patients will stay in the hospital.

Contents

1 Business Understanding

Given one year of a person's medical history from the Heritage Health Prize dataset, including hospital and routine visits, we can try to predict if that person will be admitted into a hospital the following year. There are multiple business applications that we can draw from this model, all of which pertain to preventing unnecessary costs for healthcare organizations.

For example, in ordinance with the Affordable Care Act, hospitals may be penalized if they have high readmission rates. According to the American Hospital Association, hospitals were fined $528 million in Fiscal Year 2017 [1]. The data and methodologies we describe could be used to help hospitals understand if they are in agreement with the ACA and if they are likely to be given a fine.

A financial burden for insurance companies is paying for the unnecessary admission of members into hospitals and emergency rooms. According to the Agency for Healthcare Research and Quality, over $30 billion was spent on the 4.4 million unnecessary hospital admissions in 2006 [2]. Although this dataset contains information on hospital admissions, we were not able to use it to conclude whether or not each hospital admission was necessary.

Although this dataset does not contain information on unnecessary hospital admissions, it can still be used to predict the amount of hospital visits from members of a given insurance company. This information can be used by insurance companies to assist in forecasting and budgeting for future hospital visits by its members. On average, it can cost $10,000/day per person to stay in the hospital [3]. By estimating the number of hospital claims, insurance companies can properly allocate funds and finances for the future.

The ability to accurately forecast costs is critical to any organization. In the healthcare sector especially, organizations like hospitals and insurance companies can get value from predicting if a person will be admitted to the hospital for multiple reasons. We use this dataset to find those variables that are most likely to lead to someone being admitted to a hospital the following year.

2 Data Understanding

2.1 Data Description

This dataset is comprised of members, claims and days in hospital data. The members data maps a unique member ID to the age and sex of a member. The claims data consists of one row per claim, and contains information about the member's condition, as well as where this claim took place, such as a doctor's office or hospital.

The data contains mostly categorical values, like IDs, conditions, places and specialities, with some quantitative variables, like age and days in hospital. See Table 1 for a detailed look at the attributes and possible values.

Attribute	Description	Scale	Values	Missing Values
Member ID	Unique identifier for member.	Nominal	25872 - 999999313	
Age at First Claim	Age in years at the time of the first claim's date of service computed from the date of birth; Generalized into ten year age intervals.	Ratio	0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+	
Sex	Biological sex of member	Nominal	M, F	
Provider ID	Unique identifier for insurance provider.	Nominal	59876 - 999923007	Yes
Vendor	Unique identifier for vendor.	Nominal	7666 - 9997631	Yes
Primary Care Physician	Unique identifier for the patient's primary care physician.	Nominal	982 - 998831	Yes
Year	Year in which the claim was made	Interval	Y1	
Speciality	Specialty of	Nominal	Internal, Laboratory, General Practice, Diagnostic Imaging, Emergency, Pediatrics, Surgery, Anesthesiology, Other, Pathology, Rehabilitation, Obstetrics and Gynecology	
Place of service	Place where member was treated	Nominal	Ambulance, Home, Independent Lab, Inpatient Hospital, Office, Outpatient Hospital, Urgent Care, Other	
Pay delay	Number of days between date of service and date of payment	Ratio	0 - 161, NA	Yes
Length of stay	Length of member's stay at place of service(discharge data - admission date + 1)	Ordinal	NA, 1 day, 2 days, 3 days, 4 days, 5 days, 6 days, 1- 2 weeks, 2- 4 weeks, 4- 8 weeks, 8-12 weeks, 12-26 weeks, 26+ weeks	Yes
Days since first claim	Days since first claim of that year	Ordinal	0-1 month, 1-2 months, 2-3 months, 3-4 months, 4-5 months, 5-6 months, 6-7 months, 7-8 months, 8-9 months, 9-10 months, 10-11 months, 11-12 months	
Primary Condition Group	Category of member's diagnosis, based on relative similarity of diseases and mortality rates	Nominal	See Table 2 for values and encodings	
Charlson Index	Measure of mortality, calculated as a sum of ranked scores for each disease the member has	Ordinal	0, 1-2, 3-4, 5+	
Days in Hospital	Number of days the member spends in the hospital during the next year	Ratio	0-15	

Table 1: Description of dataset

2.2 Data Quality

Overall the data seems to be of very high quality, with most attributes not having any missing values. However some attributes, including Speciality and Place of Service, have an "Other" value, which could have been used as a placeholder for a missing value, but given that some other attributes have actual empty values, we think the case is that this is truly "Other".

Provider, Vendor, and Primary Care Physician all have missing values, but these only account for 1% or less of all values in those fields. We will simply ignore the rows with missing values when making calculations that involve their respective fields, however for our task, we will likely drop these columns.

Pay Delay contains missing values, but the given data description states that there will be values labeled as 162+ to indicate values above the 95th percentile, but there are no values with this label, so it is possible NA was used in its place. However assuming this could introduce many outliers, we will likely drop the rows with NAs or impute the values by taking them from a distribution of the existing data or using the average when using this field for modeling.

Over 95% of the rows have an empty value for length of stay. The values are also inconsistent ranges, from days, to weeks, to multiple weeks. Because of this it will be difficult to impute these values, so we will choose to drop this feature.

There are not any duplicate rows in this data set, which makes sense, given the large space of the data set.

Every member in the days in hospital set has at least one claim in the claims set. This is good, otherwise we would not be able to map a set of patient data to the target.

2.3 Important Attributes

Next we will look at attributes we feel are important to predicting days in hospitals, and will likely be good indicators to this measure.

2.3.1 Age

Half of all members are age 50 or over, with the most populous category being the 70-79 age group (see Figure 1). This makes the distribution skewed towards older age compared to the population of the entire United States [4], which is expected, as older people tend to have more health problems. There's a significant drop off of age from 70-79 to 80+, which corresponds to average life expectancy in the United States [5].

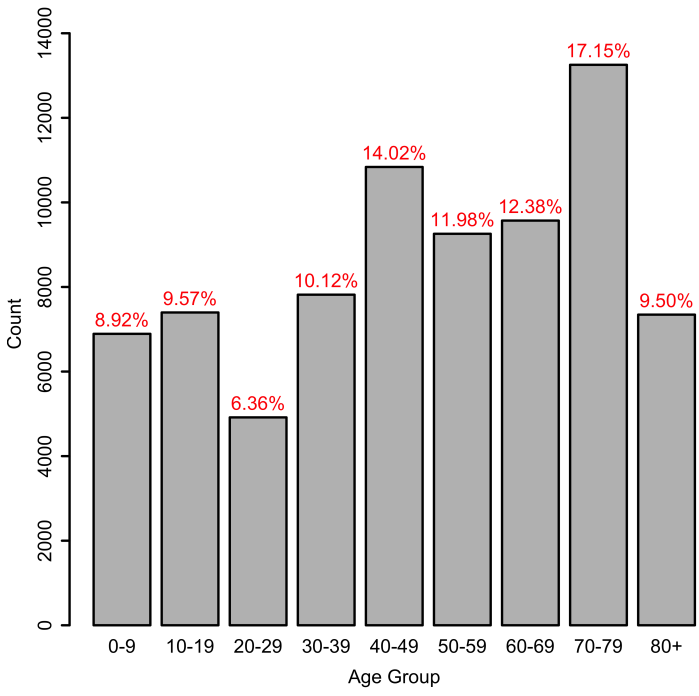

Figure 1: Distribution of Age

2.3.2 Sex

The gender distribution of members is scaled towards females compared to the 49/51% split of male to female in the US [6] with 42461 members being female compared to 34828 males (see Figure 2). This is probably related to observations that women seek health care more than men for various reasons [7].

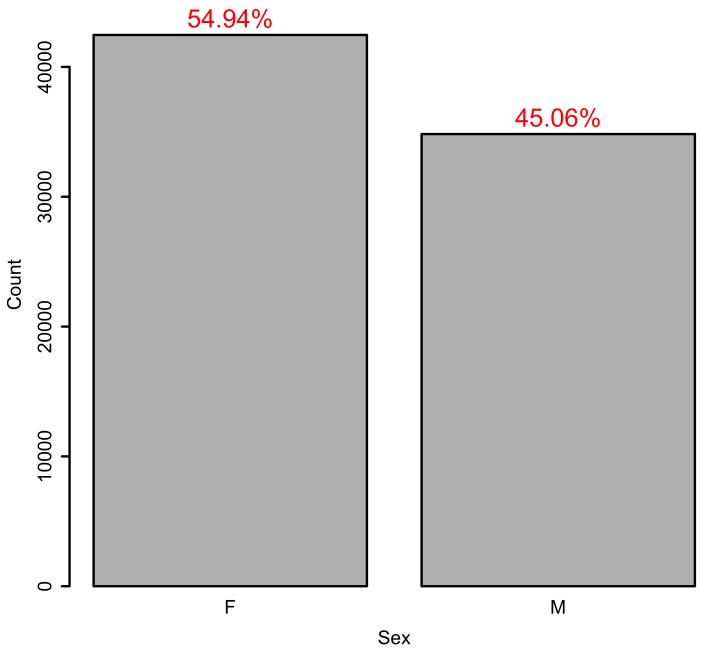

Figure 2: Distribution of Sex

2.3.3 Specialty

Specialty of the type of service provided is mostly made up of General Practice, Internal Medicine, and Laboratory work (see Figure 3). Surgery and emergency care (the specialty of care that tends to occur in hospitals) only make up for a combined 11% of all claims in the dataset. This is probably explained by the amount of regular visits to a general practitioner, as well as higher quantities of visits for internal issues.

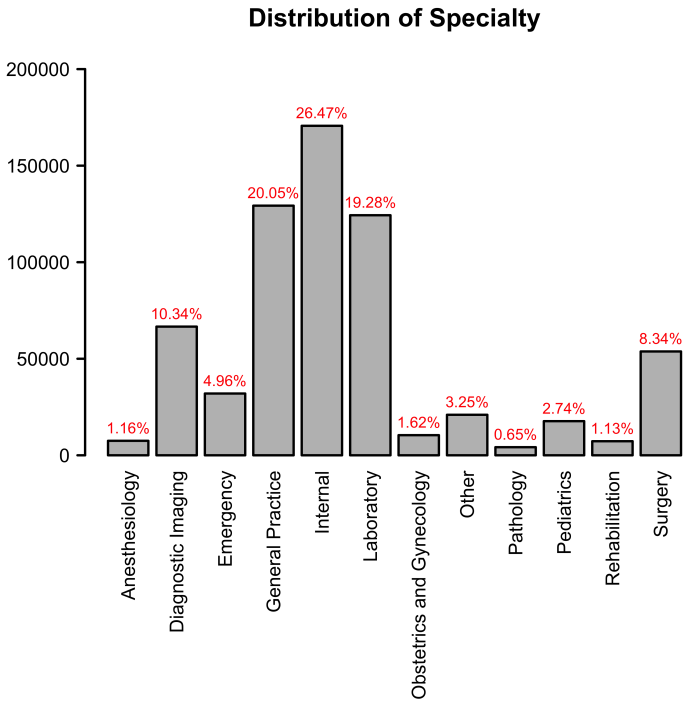

Figure 3: Distribution of Specialty

2.3.4 Place of Service

The majority of place of service is independent offices (see Figure 4). Meanwhile the places of service associated with hospitals (ambulances, inpatient hospitals, outpatient hospitals, and urgent care) only make up 13% of places listed in claims. This is expected, given that people routinely visit doctor's offices.

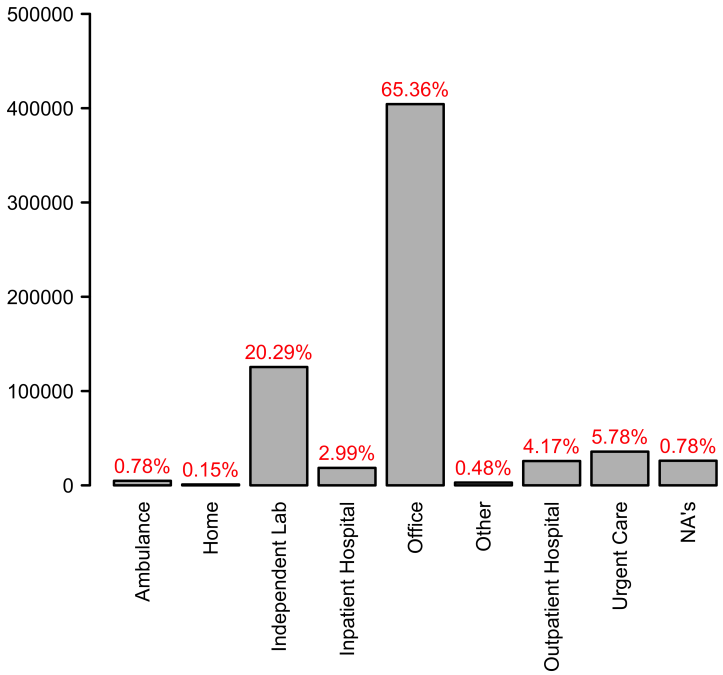

Figure 4: Distribution of Place of Service

2.3.5 Primary Condition Group

The type of condition a member has during a claim could be a very good indicator of their future health. For example some conditions are short term and aren't necessarily signals of any health problems, such as fractures, dislocations and pregnancies. Other conditions are long term, like diabetes, and some other conditions may not be actual diseases, but signs of a serious illness such as strokes.

From Figure 5 we can see the most common conditions. Because most of the common condition groups are "other" and "miscellaneous", it is hard to tell what are truly the most common conditions because the "other" groups could be made up of a majority of one condition, or many conditions with small counts. We think that it is probably the latter case because otherwise a condition taking the majority would most likely get its own group. Although the conditions seem to be loosely grouped, we still think the conditions are specific enough for there to be some correlation with days in hospital.

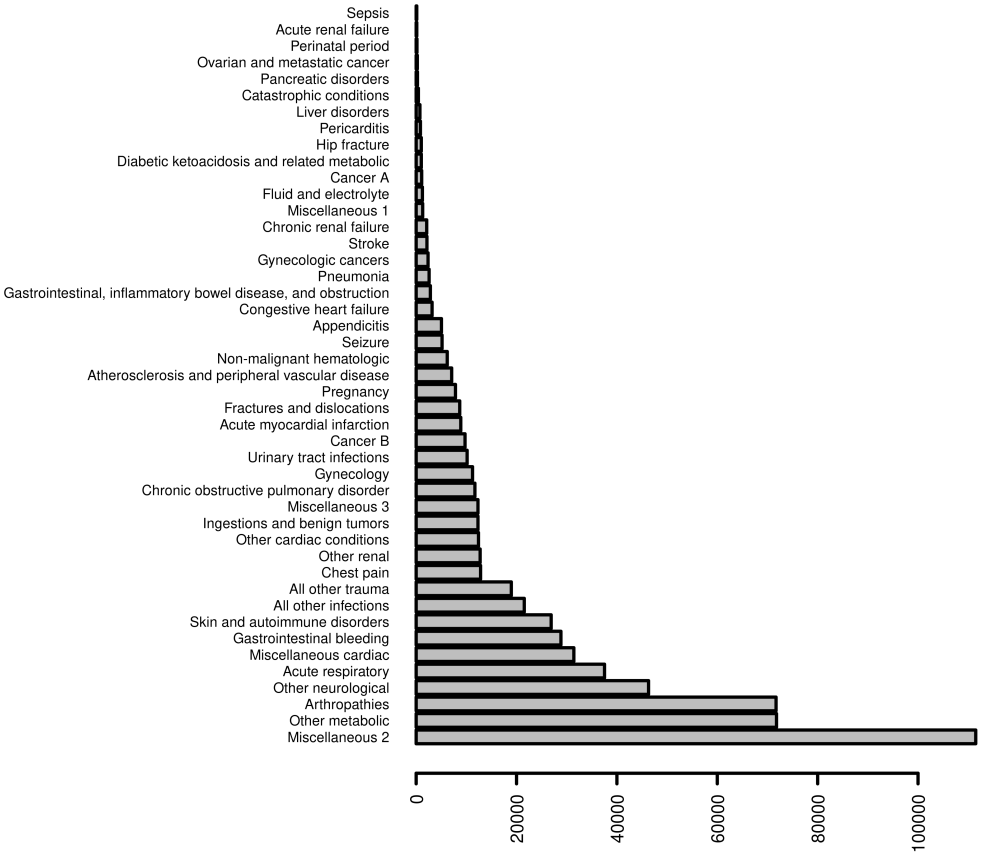

Figure 5: Frequency of Primary Condition Groups

2.3.6 Charlson Index

By definition Charlson Index seems to be a direct signal of future health. If the score is low, members are less likely to return to the hospital because they don't have life threatening diseases. If the score is very high, it may be likely that they have a short amount of time until death. From this we expect that there is some range in the middle that would be linked to more days in the hospital. Figure 6 shows that a little more than half of the claims are related to non life-threatening conditions (score of 0).

Below are the points that make up the Charlson Index [8]:

- 1 each: Myocardial infarct, congestive heart failure, peripheral vascular disease, dementia, cerebrovas-

cular disease, chronic lung disease, connective tissue disease, ulcer, chronic liver disease, diabetes.

- 2 each: Hemiplegia, moderate or severe kidney disease, diabetes with end organ damage, tumor, leukemia, lymphoma.
- 3 each: Moderate or severe liver disease.
- 6 each: Malignant tumor, metastasis, AIDS.

We would expect that a member with claims having a Charlson Index of 5+ would have a low survival rate into the next year.

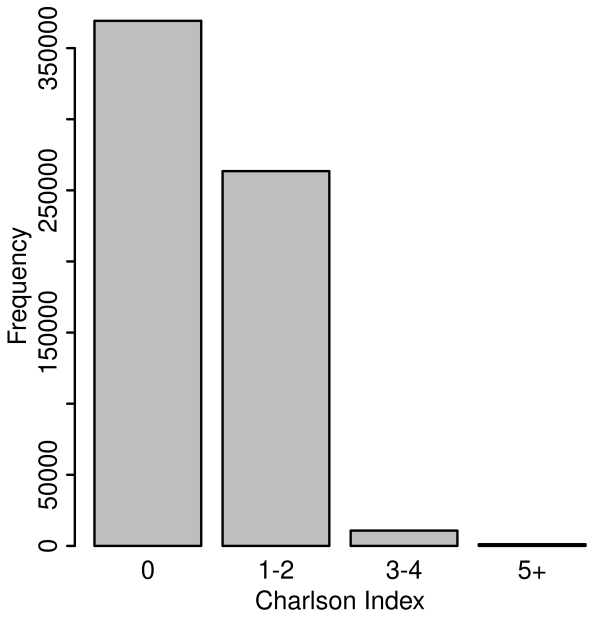

Figure 6: Distribution of Charlson Index

2.3.7 Days in Hospital

Days in hospital is important to us because it is our target, and important to understand the distribution. In Figure 7, we can see that a very large proportion ($\tilde{8}3\%$) of members in this data do not visit the hospital at all in the next year. This is most likely linked to the number of claims that are related to non life threatening conditions (with a Charlson Index of 0), as well as many claims that can be handled outside of a hospital in places like urgent care and doctor's offices. Because of the large amount of members not visiting the hospital in the next year, the mean is 0.665.

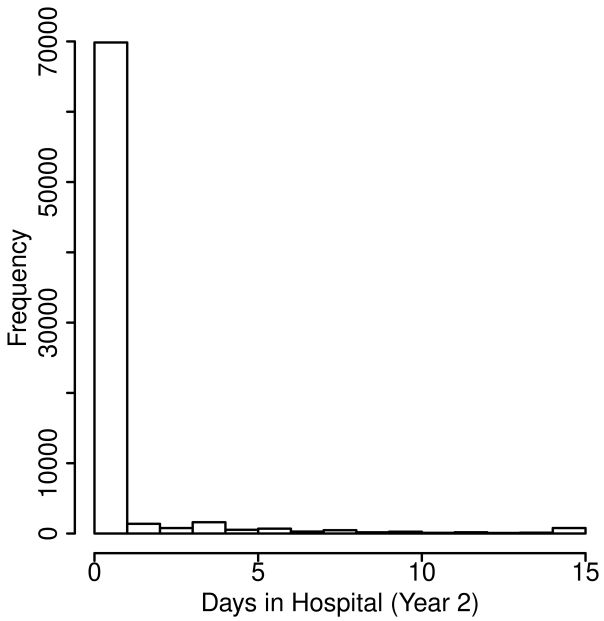

Figure 7: Histogram of Days in Hospital

2.3.8 Other Attributes

Days since first claim initially seemed to be an irrelevant attribute to predicting days in hospital. The main use of this attribute would be to measure the time elapsed between certain claims of a member in the same year. On its own it is not important, but new columns could be created using this attribute such as maximum claims a month, or longest time between successive claims. It could also be used to look at disease progression.

Provider, vendor and primary care physician can be important attributes for more granular tasks, by analyzing relationships between member's medical history and their primary care physicians or insurance providers, but we have chosen to focus on attributes that are more dependent on the member's health.

2.4 Attribute Relationships

2.4.1 Primary Condition Group and Charlson Index

Because Charlson Index gives a score based on conditions, it is interesting to see the relationship between the two. We would expect more severe conditions to have a higher Charlson Index.

In Figure 8 we can see that life threatening condition groups like those with types of Cancer have higher relative frequencies of Charlson Indexes, and groups like pregnancy have lower indexes.

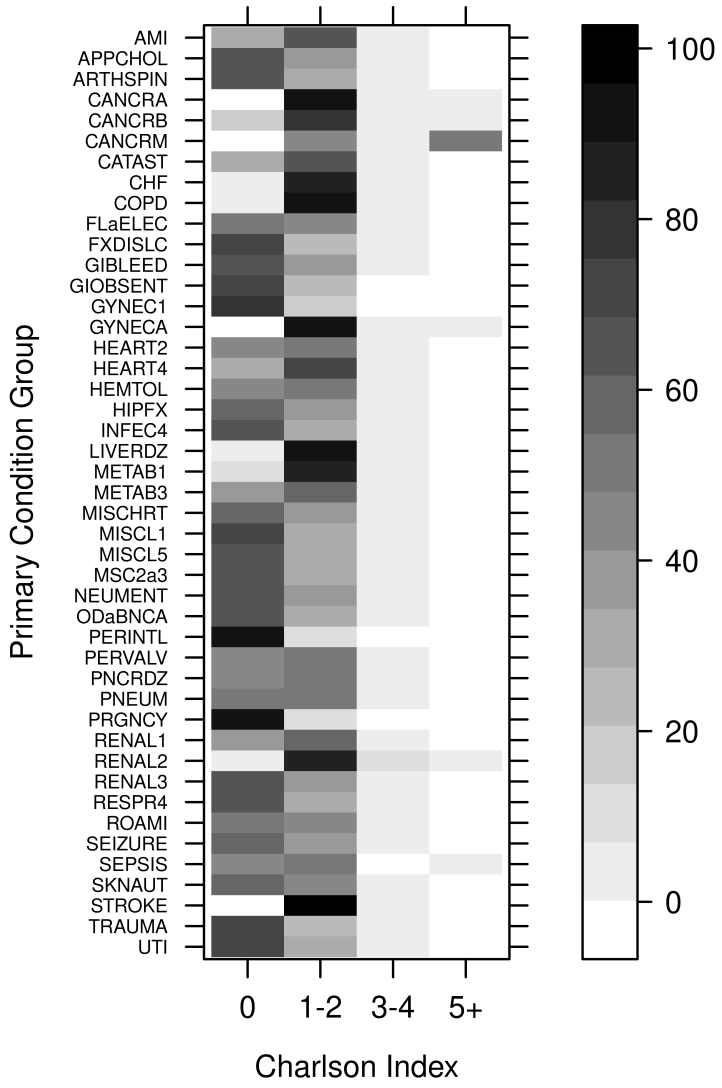

Figure 8: Relative Frequencies of Charlson Index by Primary Condition Group

2.4.2 Days in Hospital for Max Charlson Index

Of the 98% of members who had a documented Charlson Index, their max index in Year 1 is compared in Figure 9 with the average number of days that they spent in the hospital in Year 2. People who had a max Charlson Index of 3-4 spent the highest average number of days in the hospital in Year 2. And those with a high Charlson Index of 5 or more (despite making up less than 1% of all members), spent an average of 1.8 days in the hospital. This is likely due to higher Charlson Indexes resulting in a death.

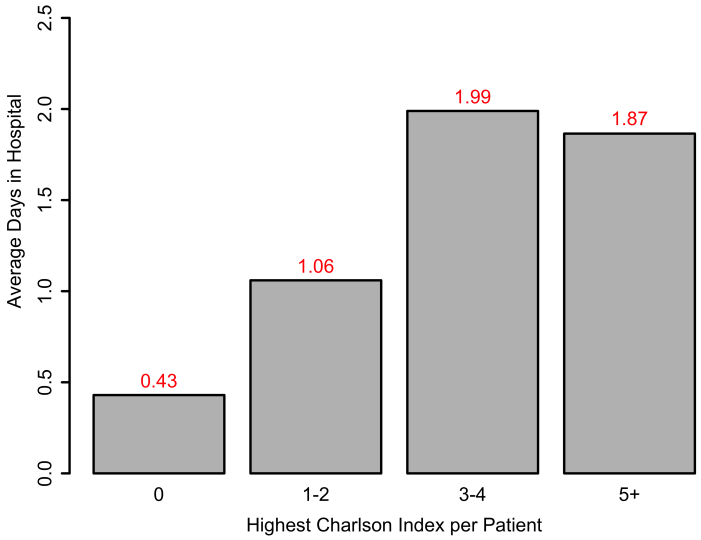

Figure 9: Charlson Index vs Days in Hospital

2.4.3 Pay Delay and Days in Hospital

Figure 10 shows a slight positive correlation between how long it takes for a member to pay and how long they spend in the hospital next year (r = 0.21). This could be explained by the possibility that people with lower income tend to have more health problems.

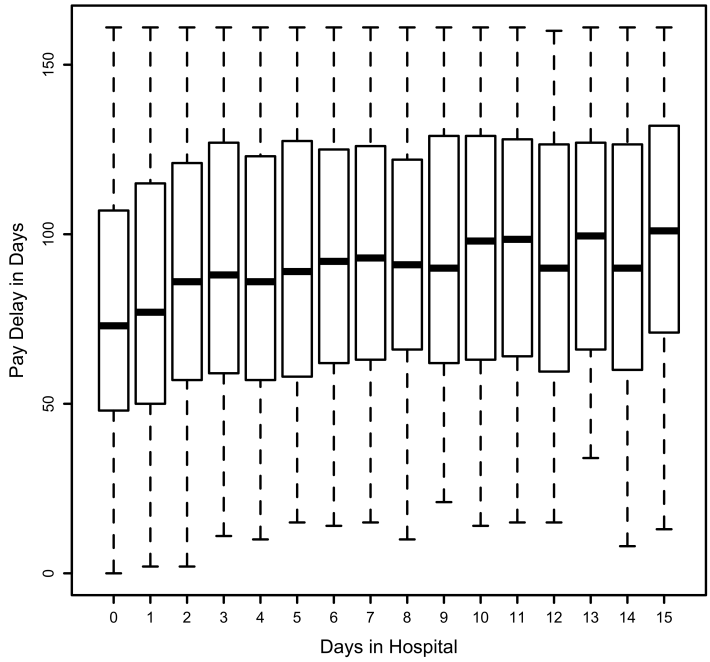

Figure 10: Longest Pay Delay vs Days In Hospital

2.4.4 Age and Days in Hospital

Looking at the days spent in the hospital by age group in Figure 11, we see a trend of higher ages leading to more days in the hospital. This makes sense, given that older people are generally more ill than younger people. For members of ages less than 50, little to no members stayed in the hospital more than 5 days.

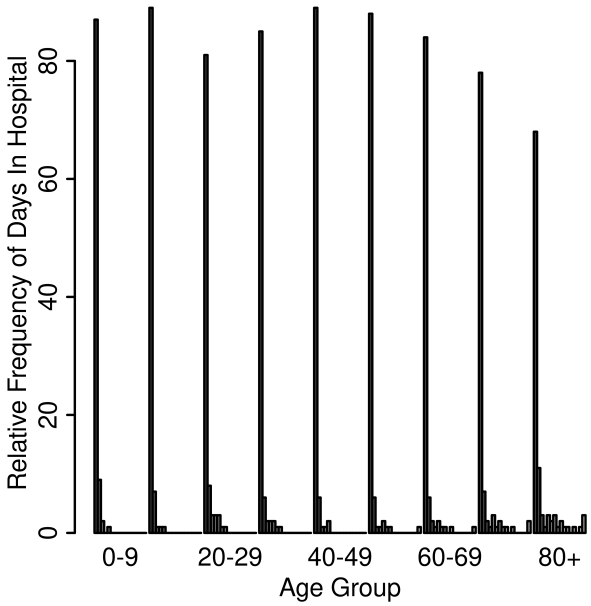

Figure 11: Age and Days in Hospital

2.4.5 Average Days in Hospital and Sex

While women tend to have a higher percentage of claims in the dataset, they also tend to have higher number average days in the hospital (see Figure 12). This is likely due to women visiting the hospital to give birth.

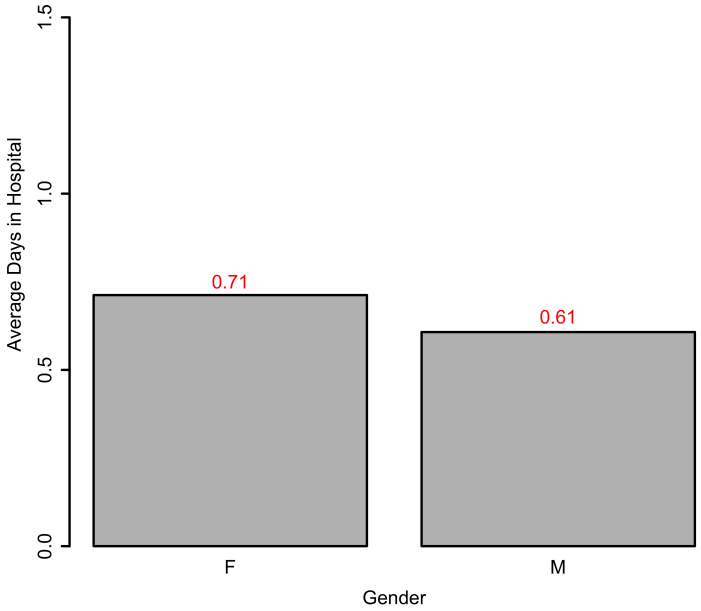

Figure 12: Average Days in Hospital by Sex

2.4.6 Number of Claims vs Days in Hospital

Figure 13 shows the number of claims that a person makes in Year 1 vs the number of days that they spend in the hospital the subsequent year. The distribution is divided by gender to see which we showed has a disproportionate number of days in the hospital. Those with the absolute highest number of claims tend to spend either 1 or less days in the hospital with low-claim women disproportionately spending at least one day. Patients of both sexes who spend 15 days in the hospital also have a history of a high number of claims the previous year.

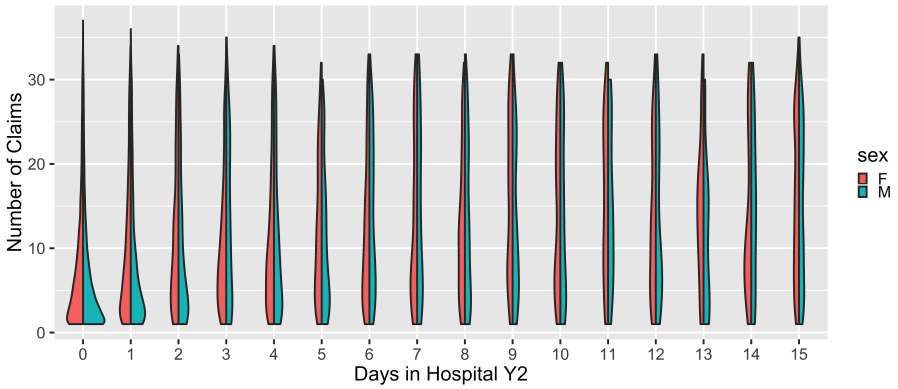

Figure 13: Number of Claims vs Days in Hospital

3 Data Preparation

3.1 Number of Conditions/Member per Year

The number of unique conditions a member has per year could be a good indicator to the number of days spent in this hospital. A person with multiple different conditions is more ill than someone with less.

The most common number of conditions is 1, and the frequency of each number decreases as the number increases as expected, given the probabilities of contracting diseases, and lower probability of combinations of those diseases (see Figure 14).

Looking at the number of conditions/year with days in hospital in Figure 15, we can see a trend, with more conditions leading to more days in the hospital. It is also interesting to see the large amount of members with 10 or more conditions that spent 0 days in the hospital. It is possible that these people passed away, but it is difficult to know for sure because of the large amount of members who spent 0 days in the hospital.

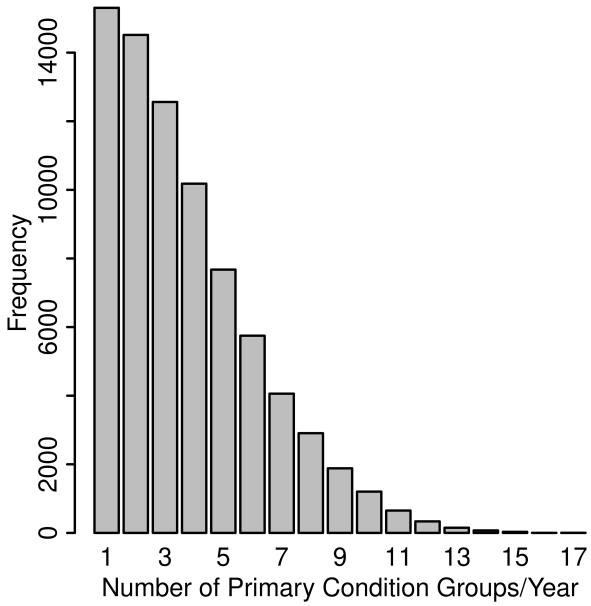

Figure 14: Frequency of Primary Condition Groups/Year Per Member

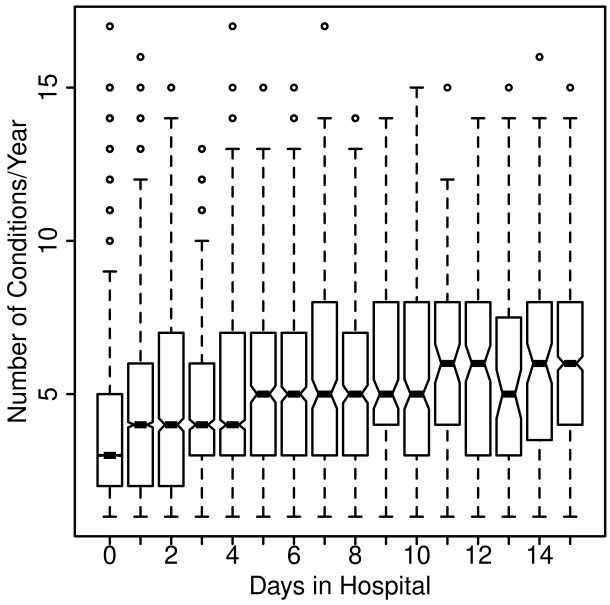

Figure 15: Number of Conditions/Year and Days in Hospital

3.2 Count Features

In order to flatten the dataset to one row per member, we created attributes for the number per year of each condition, place of service and specialty. This resulted in 65 new features that are more granular, and can give more information than number of claims, for example certain combinations of conditions can have more of an impact than other combinations.

We performed principal component analysis on these new features to find the ones that contribute most to the first principal component in order to see which are most important (see section 3.3 for details).

3.2.1 Number of Internal Specialty Claims

From Wikipedia's definition of internal medicine, "Internal medicine is the medical speciality dealing with the prevention, diagnosis, and treatment of adult diseases ... Because internal medicine patients are often seriously ill or require complex investigations, internists do much of their work in hospitals." [9].

In the feature contribution of the first principal component, number of internal specialty claims is one of the highest contributing features (see Figure 16). Taking the mean of the number of internal claims, we see a positive trend with days in hospital. This makes sense given that most internal specialists work in hospitals.

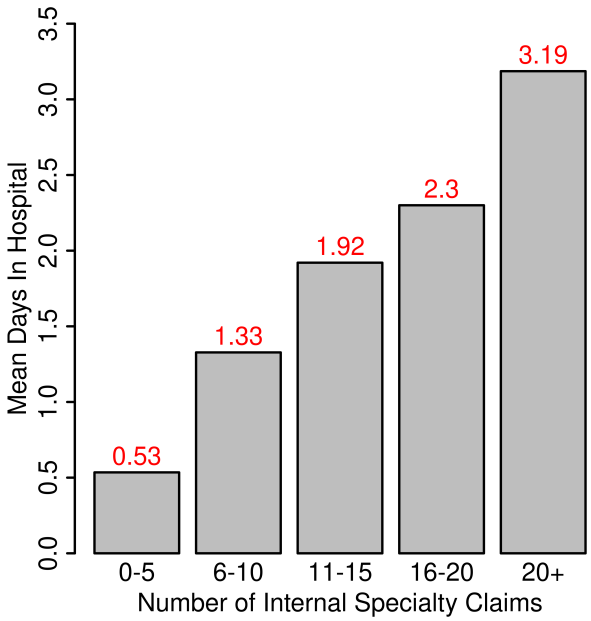

Figure 16: Mean Days in Hospital by Number of Internal Specialty Claims

3.2.2 Number of Pregnancy Condition Claims

Figure 17 shows the relationship between number of pregnancy claims and days in hospital.

Given that a member has 1 or more pregnancy condition claims, they are likely to spend about 1 day in the hospital, probably when giving birth. Note the mean for 0 pregnancy claims is approximately equal to the overall mean days in hospital.

Because chronic illnesses can decrease fertility [10], it is rare that a member would have pregnancy claims as well as life threatening conditions that could lead to more time in the hospital, so this could be a good measure of members who will not spend much time in the hospital.

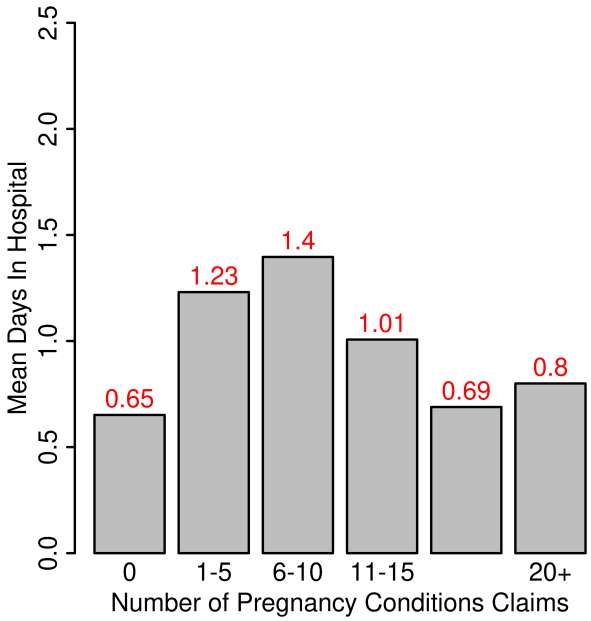

Figure 17: Mean Days in Hospital by Number of Pregnancy Condition Claims

3.3 Principal Component Analysis

Principal Component Analysis is a technique to convert a set of correlated variables into a set of uncorrelated variables called principal components. This can be used for dimensionality reduction, to reduce the number of features used in data mining algorithms by using some of these newly created attributes instead of the original features.

Performing PCA on the features created by the numbers of conditions, specialities and places of service, we can see in Figure 18 that the first component explains about 10% of the variance, with the next at 4%. By looking at the contributing features to the first component, we can get a sense of some important features, which will help decide which features to keep during modeling.

From Figure 19, we can see some of the most contributing features from specialities are internal laboratory, diagnostic imaging and surgery. This makes sense, given that these types of specialities are more related to serious illnesses.

Charlson Index and age are also important, as we previously saw when comparing to days in hospital.

At the bottom we can see conditions like pregnancy, which we analyzed in the last section. We also see metastatic cancer claims, which is the most advanced stage of cancer, at the bottom. This is most likely due to low survival rates of this type of cancer, which can lead to a member passing away.

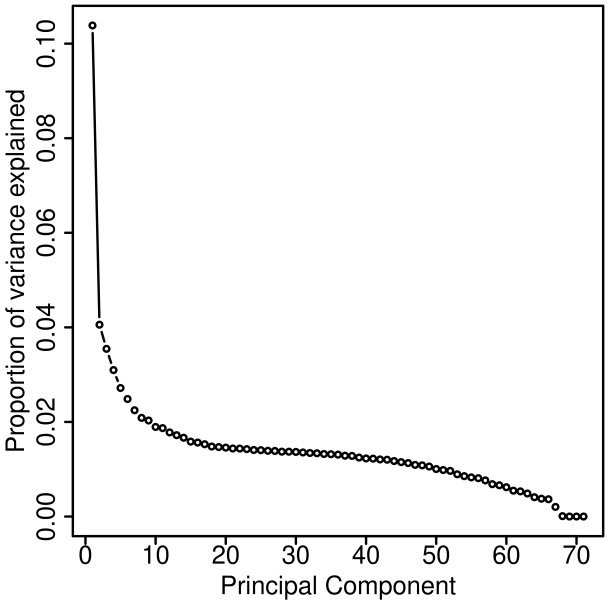

Figure 18: Variance Explained by Principal Components

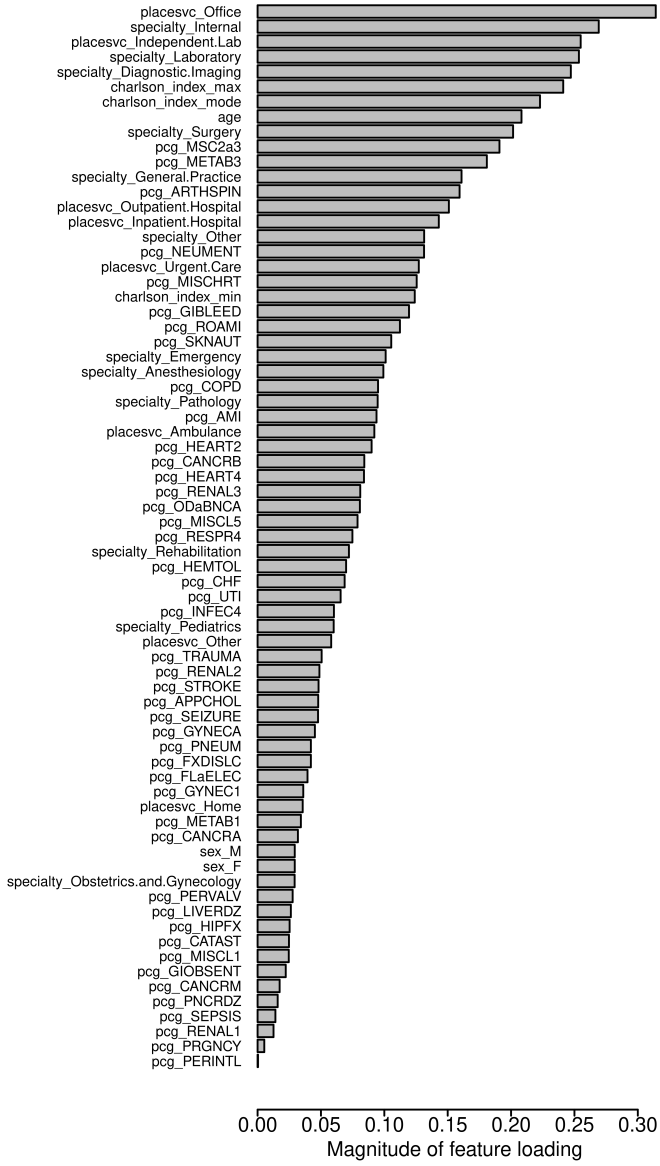

Figure 19: Feature Contribution of PC1

4 Conclusion

By looking at relationships between different attributes from someone's insurance claim history and days in hospital, we conclude that it is possible to predict days in hospital with this data. Attributes like age, number of conditions, Charlson index, and number of visits by specialty are good indicators of days in hospital, and would work well for simple estimates. At this point, we would recommended implementing a classification or regression model and use this data to make more accurate predictions. This would require flattening the claims dataset by member, and creating summary statistics such as the ones we have created, like number of types of conditions and specialties, so that the model could make predictions by a member's data.

Using this data and creating a prediction model, an insurance company could properly allocate funds and finances, and prepare for future costs, potentially saving millions of dollars.

Primary Condition Group	Description
AMI	Acute myocardial infarction
APPCHOL	Appendicitis
ARTHSPIN	Arthropathies
CANCRA	Cancer A
CANCRB	Cancer B
CANCRM	Ovarian and metastatic cancer
CATAST	Catastrophic conditions
CHF	Congestive heart failure
COPD	Chronic obstructive pulmonary disorder
FLaELEC	Fluid and electrolyte
FXDISLC	Fractures and dislocations
GIBLEED	Gastrointestinal bleeding
GIOBSENT	Gastrointestinal, inflammatory bowel disease, and obstruction
GYNEC1	Gynecology
GYNECA	Gynecologic cancers
HEART2	Other cardiac conditions
HEART4	Atherosclerosis and peripheral vascular disease
HEMTOL	Non-malignant hematologic
HIPFX	Hip fracture
INFEC4	All other infections
LIVERDZ	Liver disorders
METAB1	Diabetic ketoacidosis and related metabolic
METAB3	Other metabolic
MISCHRT	Miscellaneous cardiac
MISCL1	Miscellaneous 1
MISCL5	Miscellaneous 3
MSC2a3	Miscellaneous 2
NEUMENT	Other neurological
ODaBNCA	Ingestions and benign tumors
PERINTL	Perinatal period
PERVALV	Pericarditis
PNCRDZ	Pancreatic disorders
PNEUM	Pneumonia
PRGNCY	Pregnancy
RENAL1	Acute renal failure
RENAL2	Chronic renal failure
RENAL3	Other renal
RESPR4	Acute respiratory
ROAMI	Chest pain
SEIZURE	Seizure
SEPSIS	Sepsis
SKNAUT	Skin and autoimmune disorders
STROKE	Stroke
TRAUMA	All other trauma
UTI	Urinary tract infections

Table 2: Description of dataset

References

[1] Hospital Readmissions Reduction Program Factsheet
https://www.aha.org/system/files/2018-01/fs-readmissions.pdf

[2] Nationwide Frequency and Costs of Potentially Preventable Hospitalizations
https://www.hcup-us.ahrq.gov/reports/statbriefs/sb72.jsp

[3] Protection from high medical costs
https://www.healthcare.gov/why-coverage-is-important/protection-from-high-medical-costs/

[4] Population of the United States by sex and age 2017
https://www.statista.com/statistics/241488/population-of-the-us-by-sex-and-age/

[5] Life expectancy in North America in 2018
https://www.statista.com/statistics/274513/life-expectancy-in-north-america/

[6] Population Distribution by Gender (2016)
https://www.kff.org/other/state-indicator/distribution-by-gender

[7] The influence of gender and other patient characteristics on health care-seeking behaviour
https://bmcfampract.biomedcentral.com/articles/10.1186/s12875-016-0440-0

[8] A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation
https://www.sciencedirect.com/science/article/pii/0021968187901718?via%3Dihub

[9] Internal Medicine
https://en.wikipedia.org/wiki/Internal_medicine

[10] Chronic Illness Effect on Pregnancy
https://www.thebump.com/a/chronic-illness-effect-on-pregnancy