PREDICTING LENGTH OF STAY OF IN-PATIENTS

Statistical Data Mining Project

Abstract

Analysis of factors affecting length of stay of a patient at the time of admission using statistical tools and methods.

Apeksha Bhise
David Lam
Kirti Tiwari
Nikhitha Sama

Executive Summary

The aim of this report is to determine the factors that affect the duration of patient stay at a healthcare facility. The motivation comes from the thought of improving the quality of healthcare amongst people. The healthcare quality can be determined by multiple factors like assessing patient satisfaction, improving hospital facilities, reducing or optimizing operational costs for a patient and so on.

The report focuses on predicting the length of stay of a patient at the time of admission to a hospital and analyze the various factors like age, gender, severity of illness, admit type, diagnosis category, and how these factors impact the patient stay. The analysis is performed using regression techniques, exploratory techniques and evaluating some previous studies. The data contains 29,500 patient level details and was collected using multiple tables from different database systems obtained from one of the largest healthcare systems in Tampa Bay area.

The key findings of the analysis show that the diagnosis category of the patient is a stronger influencer of the patient's length of stay than age. The report also suggests some actionable recommendations about how controlling or better optimizing the length of stay of a patient can help in reducing length of stay as well as increasing the hospital capacity and provide better care to the patients.

Table of Contents

Problem Definition and Significance

Hospital inpatient care constitutes almost one-third of all health care expenditures in the United States. As per a study from the Institute for Health Metrics and Evaluation (IHME), the average in-patient stay cost in the US is close $22,000. Financial costs at hospitals are a growing concern for people these days. The average length of stay for patients is approximately 4.8 days in the US. This is an important factor in assessing the quality of healthcare facility. While there are many different definitions of quality improvement, the Health Resources and Services Administration (HRSA) defines it as "systematic and continuous actions that lead to measurable improvement in health care services and the health status of targeted patient groups."

Predicting the accurate length of stay is important from a patient as well as hospital point of view. It could help in improving resource utilization at hospitals, minimizing the nonvalue added care time and better prioritizing patient discharge.

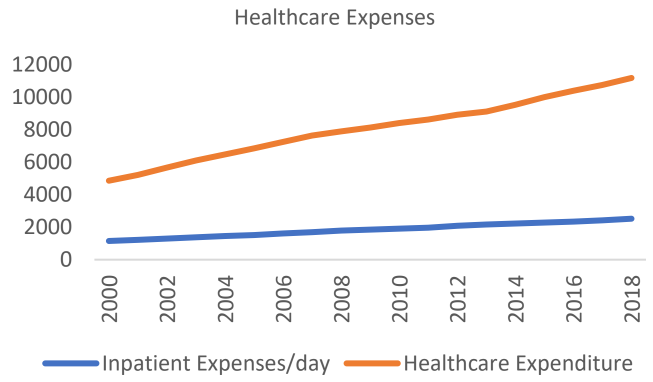


Figure 1: Total National Health Expenditure 2018 in US

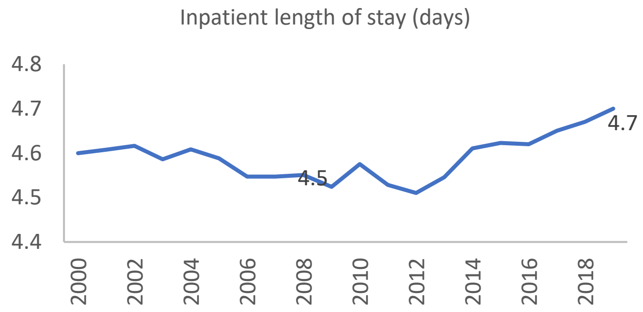


Figure 2: Inpatient LOS in days in US

While the average length of stay has remained stable, the health expenditure has increased at an alarming rate. [Reference Figure 1 and Figure 2] The type of insurance provider can hugely impact the healthcare expenditure.

Hospitals are challenged to provide timely patient care while maintaining high resource utilization. Many hospitals are working on the Real-Time demand capacity management system (RTDC) to better allocate their resources and prioritize the discharging of patients based on their predictions. If patients are classified accurately based on their length of stay at a healthcare facility, their bed allocations can be planned in a better way, which would save time to relocate patients from one ward to another, in case of a prolonged stay. Long term hospitalization patients can be allocated special wards and their treatment can be customized.

Predictive capabilities around LOS are an extremely important area for innovation. It is a big area if interest for Insurance companies and hospitals, since even patients getting discharged on the same day of treatment are charged a flat fee, irrespective of the number of hours spent. This helps in increasing hospital profits. Therefore, hospitals are focused to make optimum use of their capacity and resources.

Prior Literature

Study 1: Analysis of length of hospital stay using electronic health records: A statistical and data mining approach, Published online 2018 Apr 13 (NIH)

The basis of this study was to improve the efficiency of hospital management, and length of stay was considered to be an important indicator. The approach was to reduce the length of stay of patients which would result in decreased risk in infection and medication side effects, better quality treatment, and increased hospital profit.

This research was done on a database of patients admitted to a tertiary general university hospital in South Korea between January and December 2013. Patients were analyzed according to the following three categories: descriptive and exploratory analysis, process pattern analysis using process mining techniques, and statistical analysis and prediction of LOS. The database was not available for public viewing because of ethnicity restrictions.

The results of the study were based on multiple analysis, which included analysis of Diagnosis results, Predictive model of LOS, Performance analysis of LOS and Analysis of long-term patients. The key results of this study were:

- 55% patients were discharged within 4 days.
- Highest average LOS was 15.9 days for the department of rehabilitation medicine (RH).
- The patients with diagnostic category of cerebral infection, middle cerebral artery had the longest average length of stay, followed by myocardial infraction.
- Transfer, discharge delay time, operation frequency, frequency of diagnosis, severity, bed grade, and insurance type were strongly correlated with the LOS.

Study 2: Systematic, Data-Driven Approach Lowers Length of Stay and Improves Care Coordination

This study was done by the Memorial Hospital, at Gulfport. It is one of the most comprehensive healthcare systems in Mississippi with a capacity of 303 acute care beds and handling over 17,000 patient discharges annually. Their goal was to improve their revenue and patient satisfaction with a sustainable long-term cost-effective quality care plan.

The challenges with their approach was the absence of interdisciplinary teams who would facilitate communication and handle discharge planning. They also had inconsistent documentation about consultations and patient records. Therefore, they incorporated a systematic, data-driven, and multi-pronged approach to identify patterns in patient records and come up with actionable solutions. They formed designated committees and teams who took charge of very specific tasks.

Their data-driven approach resulted in the below outcomes:

- $2 million in cost savings, the result of decreased LOS and decreased utilization of supplies and medications.
- 0.47-day percentage point reduction in LOS through improved care coordination and physician engagement. The 30-day readmission rate has remained stable.
- 3% increase in the number of discharges occurring on the weekend.

Data Source and Preparation

The data is collected from one of the biggest healthcare systems located in Tampa Bay area. The data was extracted and put together by combining multiple backend tables from different sources including Cerner EMR (Electrical Medical Record system) and Soarian Finance (Enterprise Financial System of Cooperation). The dataset contains 1 year of non-sensitive (none-PHI information) adult inpatient data from a hospital in the area.

The dataset had basic inpatient demographic information such as gender, age, race, ethnicity, level of illness, admit type, emergency room flag, diagnosis category as well as facility ID, delay in admission and physician time seen. The length of stay was measured in hours. The demographic details were obtained from the hospital's patients records which were collected at the time of admission to the healthcare facility. The facility id here is the hospital's name id which is its unique key value in the healthcare database system.

The details of the dataset are described below:

- The master dataset had details from 5 hospital facilities with more than 100,000 patient records. Since the records were not in a standard format, we decided to conduct our analysis for 1 hospital, with 29,500 records, which would give more hospital-centric results and recommendations.
- Based on our studies of the topic, we observed that the demographics like age, gender were important in determining the recovery time of a patient. Females have more length of stay than males like in cases of arthritis, hypertension etc. while males have longer LOS in cases or cancers, heart diseases Therefore, these variables were taken into consideration for the analysis.
- The category of illness in which a patient has been admitted to, ranging from minor to extreme, were also considered important factor for the analysis since the length of stay would depend on whether the patient needs a surgery or not based on the level of illness. The type of admit, Urgent, Routine, or Emergency, were also taken in the analysis to determine which admit type would have the longest hospital stay.
- Out of all the details that were obtained, the Diagnosis category was assumed to be the most critical one. Based on our study of the prior works, majority of the analysis were based on the category of diagnosis. The diagnosis-category coding system was developed by CDC for use in all US based healthcare systems. For the purpose of our analysis, we have combined the diagnosis codes into 6 broad categories.
- The records with unknown data or blank values('NA') were deleted for the purpose of keeping a clean dataset and removing bad records.
- The admit type had some redundant values like 'emergency', 'Emergency', etc. These values were cleaned and put into 3 standard categories: Urgent, Routine and Emergency.
- If there is a delay in admission of a patient, it could impact his/her medical condition and would affect the recovery time and stay at the hospital. We have considered the admission delay variable as well for our analysis.
- The stay was also seen to be dependent on the amount of time a physician spends with a patient with the initial diagnosis at admission.

Variable Choice

Based on the data preparation, prior work, and some exploratory analysis (described in the next section), the following variables were selected as dependent and independent variables to analyze and predict the length of stay of a patient.

Dependent Variable	Independent Variable
Length of Stay (In Hours)	Gender
	Age (at the time of admission)
	Severity of Illness (1: Minor, 2: Moderate, 3: Major, 4: Extreme)
	Admit Type (Routine, Urgent, Emergency)
	Diagnosis Category
	Admission delay
	Physician time seen

Figure 3: Dependent and Independent Variables

- **FACILITY_ID (Facility) - No:**
  It contains data only for one facility, so it won't help much in our analysis.
- **ADMISSION_DELAY - Yes:**
  If there is any delay in admitting a patient. The longer the process will take it will ultimately increase the LOS for the patient.
- **PHYSICIAN_TIMESEEN - Yes:**
  This tells how much time a physician spends with the patients. This basically focus on the time spent in having face-to-face interactions with patients, gathering information about their health, developing a relationship and monitoring them. The more time they spend with patients, the more they will know about the patient condition and if there is anything which needs to be taken care immediately, sometimes fastening the administrative work related to visits. All of this can further help in reduction of LOS.
- **GENDER - Yes:**
  Gender should not have impact, but we would like to see the how it helps in determining LOS. There are some cases/ illness where Females have more length of stay than males like in cases of arthritis, hypertension etc while males have longer LOS in cases or cancers, heart diseases
  So, we should have interaction term between gender and diagnosis, to determine where the LOS is more for female.
- **AGE - Yes:** Older people tend to fall more ill and they might suffer from multiple ailments, so their length of stay will be more compared to younger patients.
- **SVRTY_OF_ILLNESS (Severity of illness) - Yes:**
  More severe illness is, the more LOS will be. (This field is numeric where 1 indicates Least Severe or minor while 4 is the Extreme level.
- **SVRTY_OF_ILLNESS_DESC (Severity of illness Description) - No:**
  This field is simply description of the severity of illness, we are already considering the field above, so we will skip this field.
- **ADMT_TYPE (Admission Type) - Yes:**
  Patients admitted under Routine should have less LOS than patients admitted as urgent or emergency.
- **RACE - No:** Race will not have an impact on patient length of stay
- **ETHNIC - No:** Ethnic will not have an impact on patient length of stay
- **EMERGENCY_ROOM_FLAG - No:**

It should have impact, but as we already have Admission type, so using this variable in the model will cause multicollinearity.

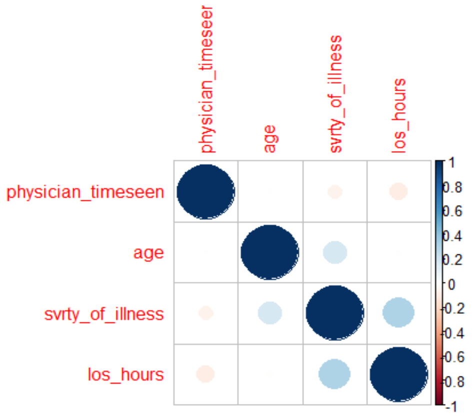
- **DIAGNOSIS_CATEGORY - Yes:**
  LOS will vary based on the diagnosis for which a patient is admitted in hospital. It is more for the patients suffering from Cancer or for the patients suffering for some mental/mood disorders as their treatment process can go long in comparison to the patients having some fractures or minor injury.

Exploratory Data Analysis & Visualizations

- Data Details:

Demographic Predictors	29488 Patients (Initially we had 29499 records, out of which few records where having NA , after omission of those 29488 records are remaining)
Gender	54.6 % Females
Age (in years)	Mean : 66 , median : 68, IQR : 25 Min : 19, Max : 108
Clinical Predictors Diagnosis	AMI-STEMI/HeartAttack - 12.06%, Fracture/Hand - 8.68%, Hema-Lymph/Breast/Lung - 6.70%, Pneumoconioses/OtherLungDiseases - 30.73%, Spine/peripheralNervous/Cerebrovascular 28.95% , Suicide/Addictive&MoodDisorders - 12.88%
Severity	MINOR - 16.34%, EXTREME - 8.58%,MAJOR - 31.86%,MODERATE - 43.22%
Admit Type	EMERGENCY - 99.75%, ROUTINE - 0.08%, URGENT - 0.17%
Admission Delay	74.5 % delays
Physician visit time(in hrs)	Mean : 26 , median : 8, IQR : 17 Min : 1, Max : 1022 (In hours)
Outcome Length of Stay in hrs	Mean : 115 , median : 82, IQR : 90 Min : 1, Max : 5374

- Collinearity:

	LOS seems to have negative correlation with physician_timeseen which implies more time spend by physicians in visiting his patients, soon they can leave, and LOS will be reduced. LOS is positively correlated with severity of illness, higher the severity, longer will be LOS duration. There is slight multicollinearity between age and severity of illness but < 0.2, so should not have much impact on our model results.

- Distribution of dependent variable LOS: log(los_hours) is more normally distributed.

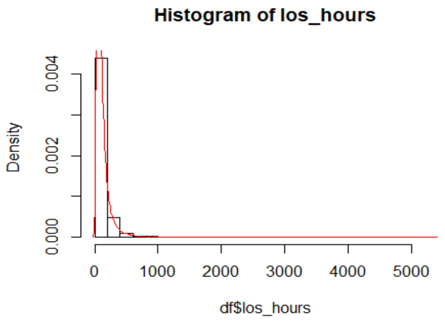


Figure 1

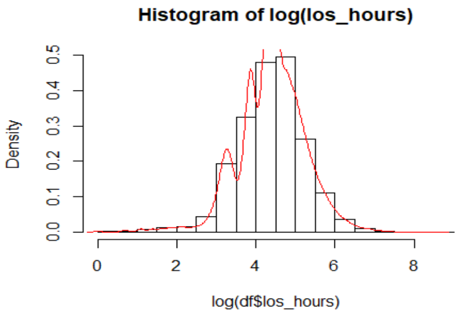


Figure 2

- Distribution of dependent and independent continuous variables

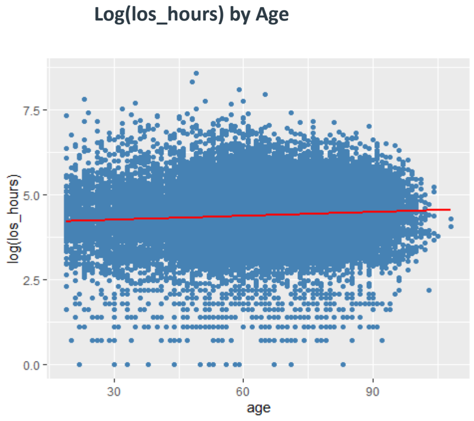


Figure 3

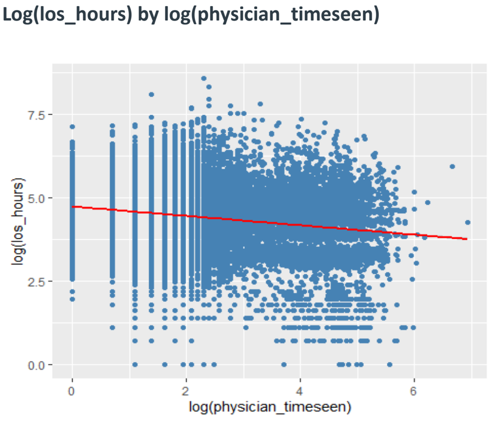


Figure 4

Both above figures are non-linear, Fig 3 seems to have zero slope while Fig 4 has negative. But both relationships seem to be weak, so they violate linearity assumption.

- Distribution of dependent and independent categorical variables

The following graphs were plotted to better understand the data and infer the relationship of the variables to the length of stay. Statistical tools were used to plot these figures.

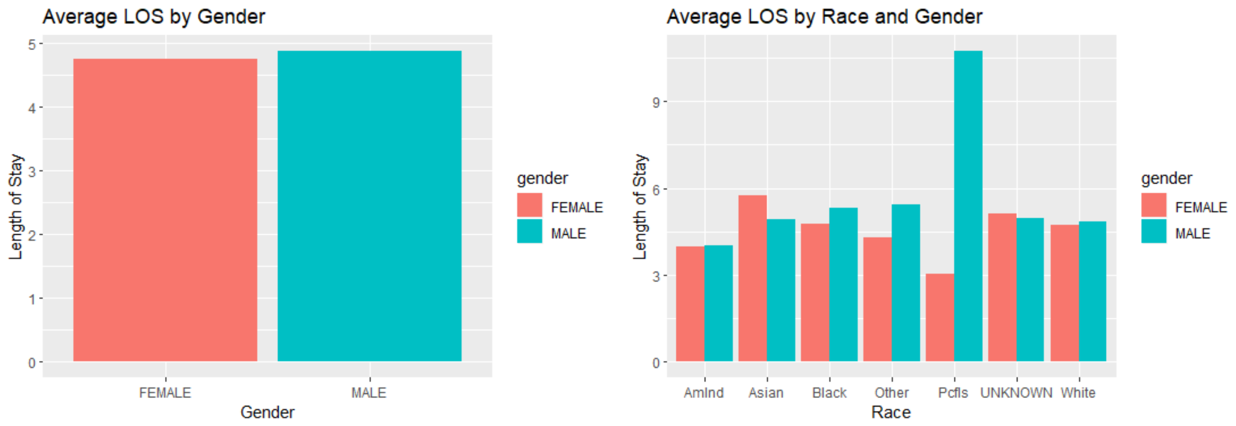


Figure 4: Length of Stay dependent on Gender and Race

- The length of stay does not vary much if only gender is considered. However, the stay varies when Race is considered for different Race. Asian Females have a higher length of stay than Asian Males.
- There is also a difference in the length of stay for males and females for the Pacific race community.

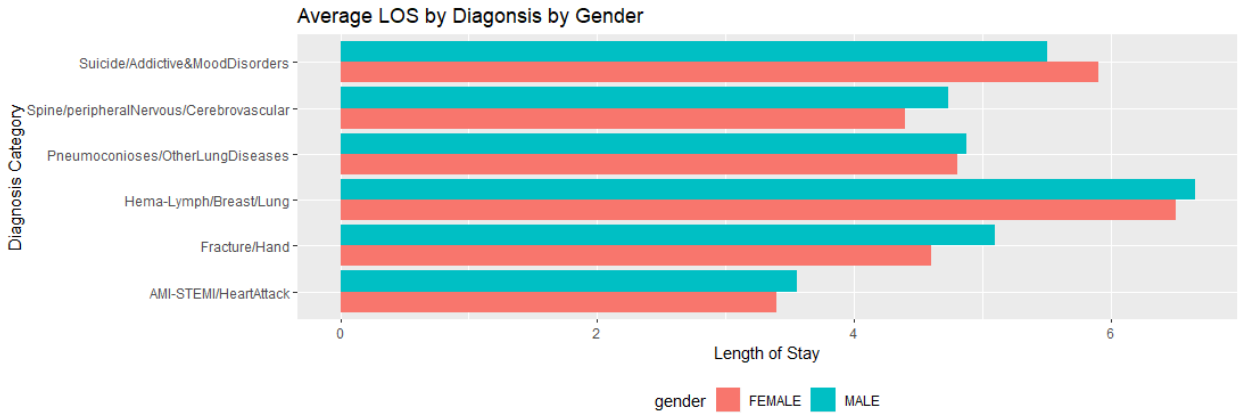


Figure 5: Length of Stay dependent on Gender and Diagnosis Category

- The average length of stay is the highest for males with lung diseases, followed by females with Psychiatric diseases.
- The relationship of diagnosis category changes with gender and should be analyzed further.

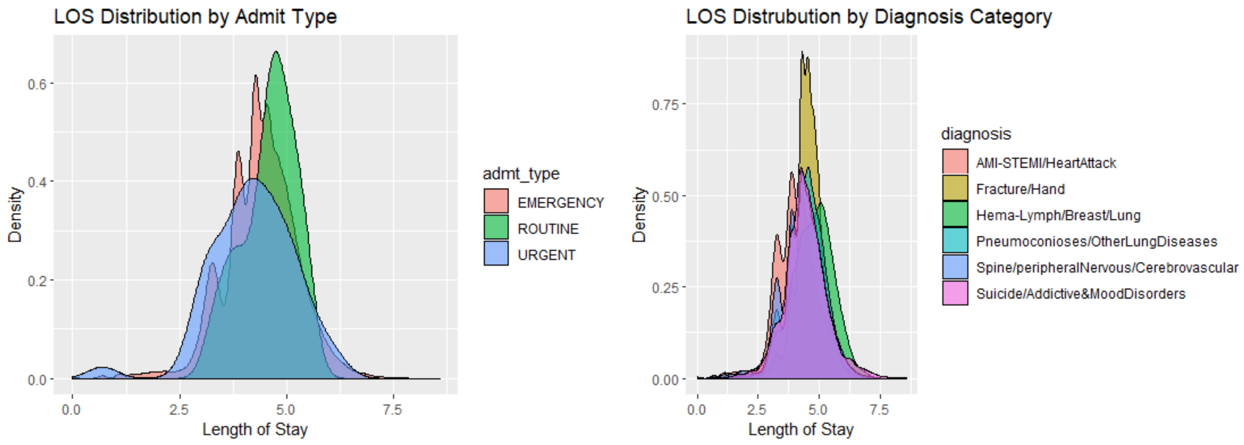


Figure 6: Distribution of Length of stay by admit type and diagnosis category

- LOS is longest for Emergency admission type. There are multiple instances where count is higher but peak is at around 4.7 (110 hours). For Urgent the spread is more which indicates higher variation in LOS than others, but maximum length of stay is almost same with Emergency Type. Routine admit type has highest peak around 5.5 which is close to 250 hours.
- Hema-lymph/Breast/Lung had longer LOS than others while for Fracture cases as the spread is thin and peak is high, so most of the cases had LOS around the peak.

Models

Based on the data exploration and our study of the subject, the following models were built to predict the length of stay. Since a relationship needs to be evaluated between the length of stay of patients and the factors impacting are the same, Regression Analysis techniques were used in the process. The Analysis was done using R language.

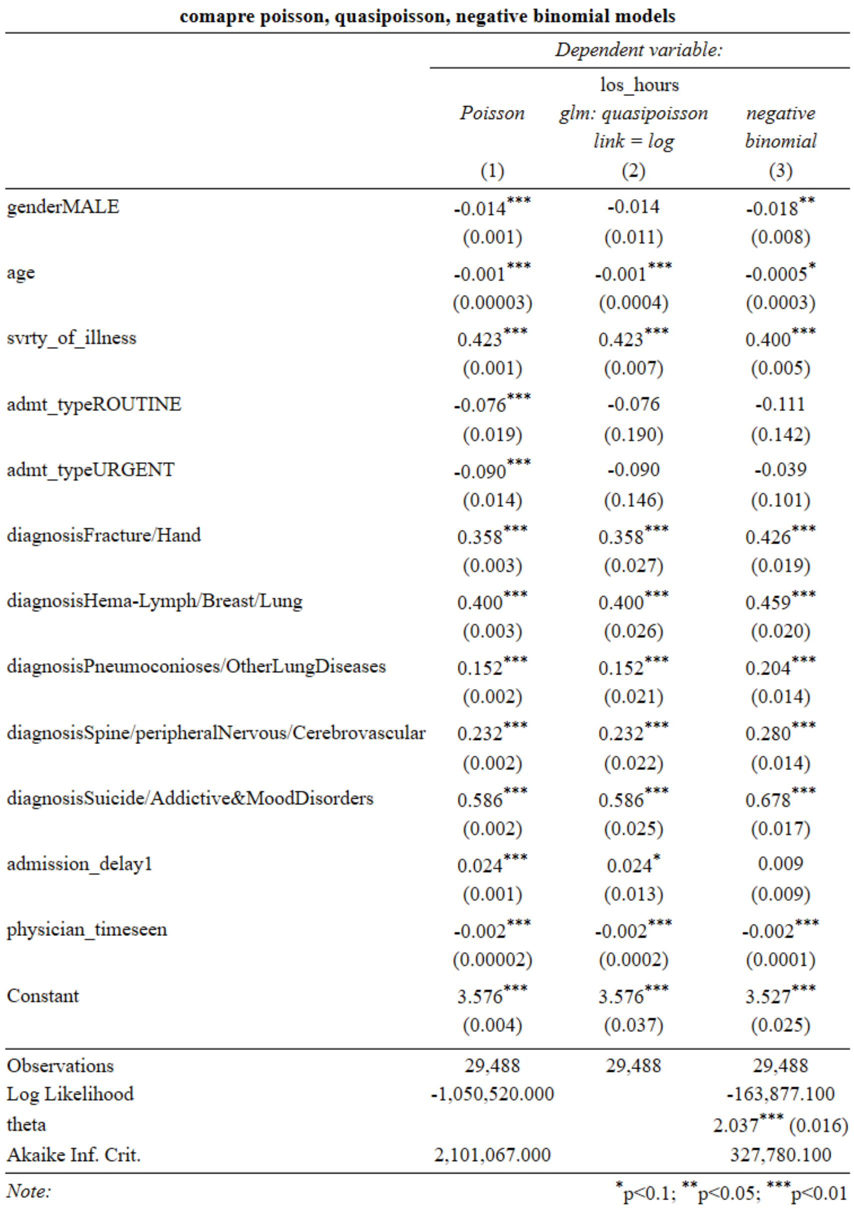
Three models were built using different regression techniques such as Poisson, Negative Binomial and Quasi-Poisson since the variable to be predicted is a count variable of hours.

Same set of predictors were used for all the three models, and the type of regression technique was modified.

Model = f(Gender, Age, Severity of Illness, Diagnosis Category, Admit Type, Admission Delay, Time seen by Physician)

Final Equation:

Log(los_hours) ~ 3.527 + (-0.18)* genderMale + (-0.0005)*age + (0.4)*svrty_of_illness + (-0.111)*admt_typeRoutine + (-0.039)*admt_typeUrgent + (0.426)*diagnosisFracture/Hand + (0.459)*diagnosisHema-Lymph/Breast/Lung + (0.204)*diagnosisPneumoconioses/OtherLungDiseases + (0.280)*diagnosisSpine/peripheralNervous/Cerebrovascular + (0.678)*diagnosisSuicide/Addictive&MoodDisorders + (0.009)*admission_delay1 + (-0.002)*physician_timeseen



comapre poisson, quasipoisson, negative binomial models

	Dependent variable:		
		los_hours	
	Poisson	*glm: quasipoisson*	*negative*
		link = log	*binomial*
	(1)	(2)	(3)
genderMALE	-0.014***	-0.014	-0.018**
	(0.001)	(0.011)	(0.008)
age	-0.001***	-0.001***	-0.0005*
	(0.00003)	(0.0004)	(0.0003)
svrty_of_illness	0.423***	0.423***	0.400***
	(0.001)	(0.007)	(0.005)
admt_typeROUTINE	-0.076***	-0.076	-0.111
	(0.019)	(0.190)	(0.142)
admt_typeURGENT	-0.090***	-0.090	-0.039
	(0.014)	(0.146)	(0.101)
diagnosisFracture/Hand	0.358***	0.358***	0.426***
	(0.003)	(0.027)	(0.019)
diagnosisHema-Lymph/Breast/Lung	0.400***	0.400***	0.459***
	(0.003)	(0.026)	(0.020)
diagnosisPneumoconioses/OtherLungDiseases	0.152***	0.152***	0.204***
	(0.002)	(0.021)	(0.014)
diagnosisSpine/peripheralNervous/Cerebrovascular	0.232***	0.232***	0.280***
	(0.002)	(0.022)	(0.014)
diagnosisSuicide/Addictive&MoodDisorders	0.586***	0.586***	0.678***
	(0.002)	(0.025)	(0.017)
admission_delay1	0.024***	0.024*	0.009
	(0.001)	(0.013)	(0.009)
physician_timeseen	-0.002***	-0.002***	-0.002***
	(0.00002)	(0.0002)	(0.0001)
Constant	3.576***	3.576***	3.527***
	(0.004)	(0.037)	(0.025)
Observations	29,488	29,488	29,488
Log Likelihood	-1,050,520.000		-163,877.100
theta			2.037*** (0.016)
Akaike Inf. Crit.	2,101,067.000		327,780.100
Note:			*p<0.1; **p<0.05; ***p<0.01

Negative Binomial Model was considered to be the best model since it had the least dispersion amongst the three models (i.e. the difference in mean and the variance of the values). The lesser the dispersion, the more stable the values would be.

Also, the AIC value for the Negative Binomial Model is the least, indicating that it is the most stable model.

Notes

- The base case for the various predictors were taken as: Female for Gender, Minor for Severity of Illness, Emergency for Admit type, AMI-STEMI/Heart Attack for Diagnosis Category, and No for Admission Delay.
- All the assumptions and interpretations are made by comparing the base case scenario with the other scenarios.

Results

- The length of stay decreases by 1.8% for males as compared to females, when other factors are kept constant. For a female if LOS is 50 hours, it will be 49 hours for a male.
- With the increase in age by 1 year, LOS will decrease by 0.05%. For example, if a person with age 60 has LOS 50 hours then for a person with age 51 it will be 2.5 hours less when keeping other factors constant, which is opposed to our initial assumption that with increase in age LOS will increase.
- The length of stay will increase by 40% if the severity of illness changes from Minor to Moderate and so on.
- If the admit type changes to Routine from Emergency then LOS will reduce by 11% and if changed to urgent, LOS will reduce by 4% when other factors are kept constant.
- For Fracture/Hand cases, LOS will increase by 43% as compared to Heart Attack.
- For Hema-Lymph/Breast/Lung diseases, LOS will increase by 46%, 20% for Pneumoconiosis and Other Lung related Diseases, 28% for Spine/peripheral Nervous/Cerebrovascular and 68% for Suicide/Addictive & Mood Disorders cases by 68% compared to AMI-STEMI/Heart Attack.

  The results seem to be in line with our initial assumption that Mood Disorder and Cancer patients will have higher LOS.
- In case of any delay in admission process, LOS will increase by 0.9%.
- In case time spend by physician with the patient increases by 1 minute then it will help in reducing LOS by 0.2%.

Quality Checks

From the results which we got for Poisson, qpoission and negative binomial, these estimates are relatively stable in magnitude, sign, and significance across models, and the results also satisfy most of our assumptions.

We did some test as well for Robustness checks of the model.

Linearity Assumptions Test: Although by the nature of data plots generated the model doesn't fall under linear category but we ran a linear model to test the assumptions to be sure that all the assumptions are violated.

- ○ **Linearity:** As mentioned above based on figures 3 & 4, Linearity assumption is violated. From the Figure 9 below between actual vs fitted values, we can confirm the same.

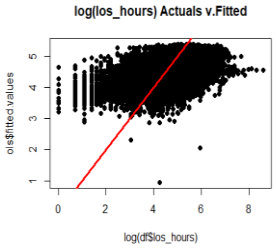


Figure 9

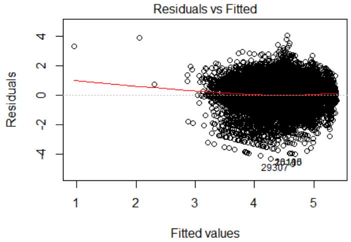


Figure 10

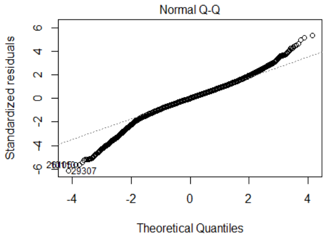


Figure 11

- **Homoskedasticity:** Figure 10, shows homoskedasticity is violated and we ran Bartlett test, p < 0.05, so 2 samples don't have equality of variances.

```
bartlett.test(list(ols$res, ols$fit))

##  Bartlett test of homogeneity of variances
## data:  list(ols$res, ols$fit)
## Bartlett's K-squared = 13243, df = 1, p-value < 2.2e-16
```

- **Normality :** From the plot in Figure 11 we can say that Normality is violated as qqnorm is not linear (except a bit in the middle but not on the edges) and deviates from overlapping with qqline.Also ran Kolmogorov-Smirnov Test: As p<0.05, we can reject H0 and infer data is not normally distributed and violates normality

```
norm <- rnorm(29488)
ks.test(norm, ols$res)

##  Two-sample Kolmogorov-Smirnov test
## data:  norm and ols$res
## D = 0.092161, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Poisson models are robust to violations in normality of residuals, homoscedasticity, and linearity. However, we still need to test for multicollinearity and autocorrelation on negative binominal model, which is our final model:

- **Multicollinearity:** None of the predictors have VIF > 5. Hence, we conclude that there is no multicollinearity in our model. We initially ran model along with emergency flag as well to test the multicollinearity in data and it gave us vif > 5, so we dropped it to avoid multicollinearity in the model.

```
vif(nb)

##                         GVIF Df GVIF^(1/(2*Df))
## gender              1.030228  1        1.015002
## age                 1.259436  1        1.122246
## svrty_of_illness    1.128742  1        1.062423
## admt_type           1.003612  2        1.000902
## diagnosis           1.390830  5        1.033540
## admission_delay     1.001942  1        1.000971
## physician_timeseen  1.022718  1        1.011295
```

- **Autocorrelation:** Since DW = 1.9884, p-value = 0.1604, so there is no autocorrelation in the model

```
dwtest(nb)##
##  Durbin-Watson test
##
## data:  nb
## DW = 1.9885, p-value = 0.1625
## alternative hypothesis: true autocorrelation is greater than 0
```

Insights and Recommendations

1. From the analysis, we can see that the patients with delay in admission have 0.9% more in length of stay than the patients who don't have any delay during admission. Thus, we would recommend to the hospital try to reduce delays in admission in order to reduce patient length of stay. For example, there are multiple ways in which hospital can improve admission efficiency in general by improving bed assignment process, adding operator/transport staff, or improve housekeeping service for a much more efficient process of admitting patients.

2. Another thing that we would recommend to the hospital administrators and management team is to add more medical staff such as physician on the daily basic shift. We can see from the analysis that increasing time of physician spend with a patient by 1 minute can help to reduce patient length of stay by 0.2%. Instead of having a physician take care of ten patients, hospital can add two or more physicians on a daily shift to take care of those patients. If physicians spend more time with their patients, there will be a highly chance of physicians making better treatment decisions and much more accurate diagnosis for the patients. Physician can carefully study patient conditions, and therefore patients will receive more of appropriate care as soon as possible from the medical staff in the hospital.

3. Based on the patient information such as gender, age, admit type, severity of illness, and diagnosis as well as hospital control variables including delay in admission, physician's time spend with patients, the hospital management team can predict the length of stay of the patient and therefore the hospital can opt for proper treatment plan for patients which can help reduce length of stay. In addition, with knowing length of stay beforehand, we would recommend the hospital use this extremely valuable insight and come up with a much more efficient operation plan/schedule such as unit capacities, beds, rooms, operation staff to ensure smoothly patient flow which can help reducing length of stay as well.

Conclusion

The quality of healthcare is dependent on a lot of factors, and getting access to such datasets is always challenging, keeping in mind the sensitivity of the data. A lot of work has been ongoing in this field, the goal is to provide affordable and quality healthcare to people in need. We would want to explore more on the various other factors which are under control of the hospitals and analyze further to help reduce the length of stay of patients and other related issues.

References

Editors, Health Catalyst. “The Top 6 Examples of Quality Improvement in Healthcare.” *Health Catalyst*, 29 Oct. 2019, www.healthcatalyst.com/insights/top-examples-quality-improvement-healthcare

OECD (2020), "Health care utilisation", *OECD Health Statistics* (database), https://doi.org/10.1787/data-00542-en (accessed on 02 May 2020).

Baek, H., Cho, M., Kim, S., Hwang, H., Song, M., & Yoo, S. (2018). Analysis of length of hospital stay using electronic health records: A statistical and data mining approach. *PloS one*, *13*(4), e0195901. https://doi.org/10.1371/journal.pone.0195901

Offerings, O. and Plat, D., 2020. *Reducing Length Of Stay Improves Outcomes And Lowers Costs*. [online] Health Catalyst. Available at: <https://www.healthcatalyst.com/success_stories/reducing-length-of-stay-memorial-hospital-at-gulfport>.

Peterson-Kaiser Health System Tracker. 2020. *Per Person Spending - Peterson-Kaiser Health System Tracker*. [online] Available at: <https://www.healthsystemtracker.org/indicator/spending/per-capita-spending/> [Accessed 2 May 2020].

Appendix

R Code :

```
####---------------Create ds-------------------------###
#########################################################
rm(list =ls())
df <- read.csv("LOS_2020_04_29_Cleaned.csv")
#colnames(mydf)=tolower(make.names(colnames(mydf)))
str(df)


#------------------Data Cleaning---------------------#
#########################################################
which(! complete.cases(df))                              # Checking for missing values
df <- na.omit(df)

df$ADMISSION_DELAY <- as.factor(df$ADMISSION_DELAY)
df$EMERGENCY_ROOM_FLAG <- as.factor(df$EMERGENCY_ROOM_FLAG)
df$GENDER <- relevel(df$GENDER,    "FEMALE")                     # Set baseline for factor variables
df$ADMT_TYPE <- relevel(df$ADMT_TYPE,     "EMERGENCY")                     # Set baseline for factor variables
df$SVRTY_OF_ILLNESS_DESC <- relevel(df$SVRTY_OF_ILLNESS_DESC, "MINOR")

colnames(df)=tolower(make.names(colnames(df)))
#unique(df$ADMISSION_DELAY)
str(df)
attach(df)

#------------------Data Visualization------------------#
#########################################################

# Distribution of dependent variable â��los_hoursâ��
hist(df$los_hours)                                # Not normally distributed
hist(log(df$los_hours))                             # Much better

# hist of Y variable
hist(df$los_hours, breaks=20, prob=T, main="Histogram of los_hours")          # Not normally distributed
den <- density(df$los_hours)
lines(den, col="red")

hist(log(df$los_hours), breaks=20, prob=T, main="Histogram of log(los_hours)")  # Much better
den <- density(log(df$los_hours))
lines(den, col="red")

# Distribution of dependent and independent continuous variables
plot(log(los_hours) ~ age, data=df)
plot(log(los_hours) ~ log(age), data=df)

plot(log(los_hours) ~ physician_timeseen, data=df)
plot(log(los_hours) ~ log(physician_timeseen), data=df)

plot(log(los_hours) ~ svrty_of_illness, data=df)
plot(log(los_hours) ~ log(svrty_of_illness), data=df)

library(ggplot2)
ggplot(df, aes(x=age, y=log(los_hours))) +
  geom_point(color= "steelblue") +               # Fit polynomial plot
  geom_smooth(method="lm", formula = y ~ poly(x, 2), color="red")
```

```
# Distributions between Y and X Factor variables
ggplot(df, aes(log(los_hours), fill=admission_delay)) +          # much better
  geom_density(alpha = 0.6) +
  ggtitle("los_hours Distributions by admission_delay") +
  theme_gray()

ggplot(df, aes(log(los_hours), fill=gender)) +          # much better
  geom_density(alpha = 0.6) +
  ggtitle("los_hours Distributions by gender") +
  theme_gray()


ggplot(df, aes(log(los_hours), fill=admt_type)) +          # much better
  geom_density(alpha = 0.6) +
  ggtitle("los_hours Distributions by admt_type") +
  theme_gray()

ggplot(df, aes(log(los_hours), fill=diagnosis)) +          # much better
  geom_density(alpha = 0.6) +
  ggtitle("los_hours Distributions by diagnosis") +
  theme_gray()


ggplot(data=df, aes(x=diagnosis,y=los_hours, fill=gender)) +
  geom_bar(stat = "summary", fun.y = "mean", position=position_dodge())+
  ggtitle("Average hours of LOS by Diagonsis by Gender") +coord_flip()



#-------------------------Models-----------------------#
##########################################################

#MLE Regression Models
mle <- glm(log(los_hours)   ~ gender + age + svrty_of_illness + admt_type  + diagnosis + admission_delay + physician_timeseen, data=df,
family=gaussian)

#poisson models
poisson = glm(los_hours     ~ gender + age + svrty_of_illness + admt_type  + diagnosis + admission_delay + physician_timeseen, data=df,
family=poisson (link=log))
summary(poisson)
#There is some dispersion in the data (deviance > df), Residual deviance: 1917100  on 29475  degrees of freedom
#hence, we can try quasi-poisson and negative binomial regression

#quasipoisson timeseen_to_admit
qpoisson = glm(los_hours    ~ gender + age + svrty_of_illness + admt_type  + diagnosis + admission_delay + physician_timeseen, data=df,
family=quasipoisson (link=log))
summary(qpoisson)
#' Given that the dispersion parameter (lambda) is 104.3475 and quite far from 1, negative binomial regression may be more appropriate.

#negative binomial models : best model
library(MASS)
nb  <- glm.nb(los_hours ~ gender + age + svrty_of_illness + admt_type  + diagnosis + admission_delay + physician_timeseen, data=df)
summary(nb)

#compare models
#poisson and qpoisson have the same results

library(stargazer)
stargazer(poisson, qpoisson, nb, type="text", title = "comapre poisson, quasipoisson, negative binomial models ")
#Negative binomial appears to be the best model for los_hours, given dispersion in data.
```

```
##########Test models#############

#Poisson models are robust to violations in normality of residuals, homoscedasticity, and linearity.
#However, we must still test for multicollinearity between predictors and independence(autocorrelation).

#' VIF test for multicollinearity
library("car")
vif(nb)

#admt_type & emergency_room_flag have VIF > 5. Hence, we must exlcude emergency_room_flag to avoid multicollinearity in our model.
#After remove emergency_room_flag predictor, None of the predictors have VIF > 5. Hence, we conclude that there is no multicollinearity in our model.

#' Durbin-Watson test for independence
#' DW value range: [0, 4], values between 1.5 and 2.5 suggest no autocorrelation
#' H0: Autocorrelation = 0. Reject H0 if p<0.05
library(lmtest)
dwtest(nb)

#Since DW = 1.9884, p-value = 0.1604, no autocorrelation
#=> pass all the test assumption


############################## For test of assumptions #######################
#OLS Model
ols<- lm(log(los_hours)   ~ gender + age + svrty_of_illness + admt_type  + diagnosis+ admission_delay + physician_timeseen, data=df)

# LINE Assumption test
#par(mfrow=c(2,2))
plot(ols)
#par(mfrow=c(1,1))

# Linearity:
plot(log(df$los_hours),ols$fitted.values,pch=19,main="log(los_hours) Actuals v.Fitted")
abline(0,1,lwd=3,col="red")
#' VIF test for multicollinearity
library("car")
vif(ols)

#' Durbin-Watson test for independenc
#library(lmtest)
#dwtest(ols)


norm <- rnorm(29488)
ks.test(norm, ols$res)
#ks.test(norm, nb$res)

bartlett.test(list(ols$res, ols$fit))
#bartlett.test(list(nb$res, nb$fit))
```