A statistical approach: Predicting the Healthcare Costs to Be Incurred by Patients

Sayantan Chatterjee[1]

[1]Symbiosis Centre for Management and Human Resource Development, SCMHRD, SIU, Symbiosis International (Deemed University), SIU, Hinjawadi, Pune, Maharashtra, India

[1]sayantan_chatterjee@scmhrd.edu

ABSTRACT

Purpose: In a world where healthcare costs are always on the high and people do not have enough means to always get proper medical care, it is of paramount significance to predict the medical costs of patients. The motivation was to figure out which elements are related with predicting the average expenditure to be incurred by patients during their stay, based on electronic medical records, so as to oversee emergency clinic stay all the more proficiently. This will also help to allocate care management resources to those individuals at highest risk of incurring significant costs.

Proposed Design/Methodology/Approach:

The dataset contains data regarding the patient personal details, diagnosis, length of stay as well as the facility (medical provider) details. Data which were missing needed to be imputed or removed depending on the significance level of the factors. The data were broken down using descriptive and exploratory data analysis (EDA), data cleaning techniques of label encoding processes and features with significant importance for the predictive model was identified. A statistical linear regression model was created for forecasting the Total Charges (expenditure of patients during their hospital stay).

Practical/Theoretical implications:

There are a lot of theoretical and practical implications of this research, which are categorized in the following points:

- Revenue optimization for medical providers and patients
- Medical classification of patients for better treatment
- Effort/time optimization for medical providers
- Pre-emptive information for insurance companies
- Information to patients regarding payments and medical expenses incurred

Originality/value:

This research paper is mainly focused on how to use the expenditure prediction strategically by patients, and also by providers and other healthcare businesses firms to maximize their revenue and be more efficient.

Keywords

Total Charges, length of stay (LOS), Inpatient, medical providers, predictive model

Article Received: 10 August 2020, Revised: 25 October 2020, Accepted: 18 November 2020

Introduction

Healthcare costs are ever increasing. Reports from the World Bank shows a steady rise in the healthcare expenditure with approximately 9.989% of Total World's GDP being spent on it in 2016. Especially in times like these when we are going through a pandemic, the extent of clinical cost increments has been all around documented. But in any case, there have been limited efforts till date to evaluate the factors causing these increments. Also Health insurers have been grappling with claims as high as ₹7 lakhs as private hospitals seek higher package rates for treating a covid-19 patients. So our aim is to find a statistical approach to this problem in order to predict the Total Healthcare Charges for a patient beforehand using electronic medical data which will help both the patients as well as healthcare providers (insurance companies and hospitals).

In order to find a statistical approach for predicting the medical expenses of an in-patient, the dataset we will use is the New York State Department of Health public domain dataset for in-patients for the year 2016. The electronic medical data for all the individual patients are recorded in the dataset and made available for public domain for academic and research purposes. The dataset contains information regarding the medical provider (hospital) Facility Name, Facility Area, County, Permanent Facility Id etc as well as information regarding the in-patient such as their, Zip Code - 3 digits, Gender, Race, Ethnicity, Length of Stay, Type of Admission, Patient Disposition, Discharge Year, Age Group, CCS Diagnosis Code, CCS Diagnosis Description, CCS Procedure Code, CCS Procedure Description, APR DRG Code , APR DRG Description, APR MDC Code, APR MDC Description, APR Severity of Illness Code , APR Severity of Illness Description, APR Risk of Mortality, APR Medical Surgical Description, Payment Typology, Attending Provider License Number, Operating Provider License Number, Other Provider License Number, Birth Weight, Abortion Edit Indicator, Emergency Department Indicator, Total Costs, Total Charges and Ratio of Total Costs to Total Charges.

The data provided in the dataset have been in raw format with lots of missing values and also features irrelevant to our study. So we will feature engineer the data and find the right significant features that will help us create the statistical model for proper prediction of the Total Charges for an inpatient.

Literature Review

A. Identifying (Predict) Patients with Significantly High Health Care Expenses.

Countries all over the world are facing the uphill task of managing the ever increasing healthcare costs. Most statistical researches on patients with high healthcare expenditure depend on diagnosis information with extra cost and human services use information to better the forecasting. To improve model performance, analysts have been trying different new information sources, for example, self-revealed wellbeing status information. Three classifications of clinical check information, research facility tests, self-detailed clinical history, and self-revealed wellbeing conduct information can be used to evaluate and create models for predicting High Cost health care users. (Yeonkook J. Kim and Hayoung Park)

B. How does Health Care Costs Affect Insurers and Medical Providers?

Containing costs is of high need to insurers and patients, medicinal providers and suppliers would want to oppose such cost regulation. Since an enormous level of the patients rely upon employers for medical coverage, managers and insurance providers are basically liable for the expenses. Aside from human services costs, managers likewise acquire other wellbeing related uses because of the time taken by the patient to come back to work and the subsequent loss of profitability (Robert S. Kaplan and Michael E. Porter)

C. Length-Of-Stay and its Impact on Health Costs

Longer "Length-Of-Stay's, is exorbitant for suppliers, insurances companies, hospitals and patients alike since the expenditure increases multifold for both the patient as well as the provider and also greatly increases the risk for the patient to be affected by other diseases (for e.g. In the current scenario, a patient admitted for some treatment is at a greater risk of being affected by Covid-19 the longer the patient stays in the hospital). It also increases the Total Charges billed for the patient and not only creates a headache for the patient but also for the insurance companies providing medical covers for those patients. Thus LOS of patients is a very significant feature in estimating the potential cost charges for a patient and also planning the treatment procedures. (Alvaro J Riascos and Natalia Serna)

D. Transition to Risk-Adjustment Models

Over the last 10 years, risk-adjustment methodology have advanced a lot and are more and more used by researchers. Researchers have shifted their attention to change the data that they use for Risk-adjustment models. This change have resulted in incorporation of electronic medical patient data that provides much more insights into healthcare costs and thus improves in the prediction. Risk-adjustment models can all the more likely distinguish high cost expense patients with interminable ailments.

Research Methodology

The electronic in-patient medical reports for New York State Department of Health for the year 2016 contains detailed records of the patients, including their **Age Group, Zip Code - 3 digits, Gender, Race, Ethnicity, Length of Stay, Type of Admission, Patient Disposition, Discharge Year, CCS Diagnosis Code, CCS Diagnosis Description, CCS Procedure Code, CCS Procedure Description, APR DRG Code , APR DRG Description, APR MDC Code, APR MDC Description, APR Severity of Illness Code , APR Severity of Illness Description, APR Risk of Mortality, APR Medical Surgical Description, Payment Typology, Attending Provider License Number, Operating Provider License Number, Other Provider License Number, Birth Weight**, **Abortion Edit Indicator, Emergency Department Indicator, Total Costs, Total Charges and Ratio of Total Costs to Total Charges, as well as the medical provider details Facility Name, Hospital Service Area, Hospital County, Operating Certificate Number, Permanent Facility Id**. For our research using statistical models for predicting the healthcare charges for patients, we have tried different modelling techniques including SGDRegressor, Gradient Boosting Regressor, Linear Regression, K-Neighbors Regressor and Random Forest Regressor. After comparing the results, the decision to go ahead with Linear Regression was taken.

Our goal is to determine the relation on Total Charges for a patient with respect to several factors and thereby creating a model that can predict the same. My study considered several factors as independent variables that predicts the Total Charges (dependent variable). We aim to decrease the error between the actual data and the prediction (residuals). We need to select the factors properly based on their significance as many factors are related to each other and thus can increase multicollinearity in the regression model. So we need to check each and every factor based on their importance and drop irrelevant data to reducing multicollinearity.

A. Scalar response:

The scalar response is the dependent variable (Total Charges). We will use linear regression on the different independent variables we have in the dataset.

Total Charges: Total charges is the amount the patient is billed by the hospital

B. Independent variables:

There are several factors that have been considered as independent variables. The dataset used for the research contains data about different in-patient records (independent variables).

Each variable from the dataset have been explained below along with the methodology and reasoning behind selecting the final set of independent variables for the statistical model creation.

a.

ospital Service Area : Contains the area (location) of the

medical provider

b. **ospital County** : The county where the medical provider is located

c. **perating Certificate Number** : The operating certificate number of the provider

d. **ermanent Facility Id** : The unique ID for each hospital (provider)

e. **acility Name** : The name of the hospital

f. **ge Group**: The age group of the individual patient

g. **ip Code - 3 digits** : The Zip code of the individual patient

h. **ender** : Gender of the individual patient

i. **ace** : Race of the individual patient

j. **thnicity** : Ethnicity of individual patient

k. **ength of Stay** : The time period(in days) for which the patient stayed in the hospital

l. **ype of Admission :** Explains the type of admission (for eg : emergency, delivery of a new-born baby)

m. **atient Disposition**: Patient Disposition means where a patient is being discharged – i.e. home or self-care, or skilled nursing home or expired etc.

n. **ischarge Year** : The Year in which the patient is discharged (eg : 2016)

o. **CS Diagnosis Code and CCS Diagnosis Description** : CCS Diagnosis categorizes patient diagnoses into different categories

p. **CS Procedure Code and CCS Procedure Description**: Patients undergo different medical procedures. CCS Procedure categorizes the different medical procedures

q. **PR DRG Code and APR DRG Description**: All Patients Refined Diagnosis Related Groups (APR DRG) is used to categorize patients into different classes based on type of admission, their severity and risk of mortality

r. **PR MDC Code and APR MDC Description**: The Major Diagnostic Categories (MDC) are formed by categorizing the different diagnoses of patients

s. **PR Severity of Illness Code and APR Severity of Illness Description**: Explains the severity of the individual patient

t. **PR Risk of Mortality** : Explains the risk of mortality for individual patient

u. **PR Medical Surgical Description** : Explains whether patient is admitted for surgical purpose or general medical purpose

v. **ayment Typology** : The Payment type of the patient(e.g.: Insurance or Self-Pay)

w. **ttending Provider License Number** : It is a license number of the provider which the patient is attending

x. **perating Provider License Number** : It is a license number of the operating provider

y. **ther Provider License Number**

z. **irth Weight**: Birth weight of new-born baby

aa. **bortion Edit Indicator**: Indicator of whether the patient is admitted for abortion or not

bb. **mergency Department Indicator** : Whether the patient is admitted for emergency or not

cc. **otal Costs**: Total Costs indicate the total healthcare expenses of the provider. It does not include the doctor consulting charges though

dd. **Ratio of Total Costs to Total Charges**: Ratio of cost incurred to the cost billed by the provider

C. Methodology and Reasoning behind selection/removal of Independent Variables/Factors from our dataset for the Model

- The columns **Hospital Service Area, Hospital County, Operating Certificate Number, Facility Name** were dropped as they are correlated with column **Permanent Facility Id** (Permanent Facility Id uniquely identifies each hospital)
- he column **Discharge Year** contains only one value **2016** (all the records for inpatient discharges are from 2016). So this column doesn't have any impact in the prediction for the **Total Charges** billed for the patient
- he column **CCS Diagnosis Description** is already explained by the **CCS Diagnosis Code** (all Diagnosis contains unique codes) and hence are correlated. Hence, the column **CCS Diagnosis Description** is dropped
- imilarly, **CCS Procedure Description** is already explained by the CCS Procedure Code (all Procedures contains unique codes) and hence are correlated. Hence, the column CCS Procedure Description is dropped
- he column **APR DRG Description** is already explained by the **APR DRG Code** (all contains unique codes) and hence are correlated. Hence, the column **APR DRG Description** is dropped
- iimilarly, **APR MDC Description**, is already explained by the **APR MDC Code** (all contains unique codes) and hence are correlated. Hence, the column **APR MDC Description** is dropped
- he column **APR Severity of Illness Description** is already

explained by the **APR Severity of Illness Code** (contains unique codes) and hence are correlated. Hence, the column **APR Severity of Illness Description** is dropped

- he column **Zip Code - 3 digits** is important personal data for the patient but in our research, the zip code of a person cannot be a good predictor of the **Total Charges** for a patient billed at a hospital. Hence the column is dropped
- oth **Race** and **Ethnicity** kind of explains the same feature for an individual. The Race column explained it better in comparison to the Ethnicity. Both are not very good predictors for the Total Charges, so the column Race was kept and Ethnicity was dropped
- he column **Abortion Edit Indicator** is dropped as it is highly imbalanced(99.85% data saying N)
- he **Birthweight** column was also removed as weight of a new-born baby is an important data in terms for the baby but cannot be a good predictor of the Total Charges billed for an individual patient at a hospital. Hence the column is dropped
- he columns including the License numbers of Attending Provider and Operating Provider are just unique numbers of the providers and thus is not significant in terms of determining the Total Charges

D. Feature Engineering of the Data

- **Level Encoding** done on column **Patient Disposition** column for encoding purposes
- or the payment typology, there is a reason why three **Payment Typology** columns are given in the dataset, basically to show how a patient has paid the money. So if someone has paid in 2 installments, then the column Payment Typology 3 is blank for that patient. So values have been assigned for each of the payment types and 3 columns have been converted into one. (Details for encoding below)

Original Options	Payment	Encoding
Medicare	Insurance	4
Medicaid	Insurance	4
Department of Corrections	Non Self pay	1
Private Health Insurance	Insurance	4
Blue Cross/Blue Shield	Non Self pay	1
Miscellaneous/Other	Non Self pay	1
Federal/State/Local/VA	Non Self pay	1
Self-Pay	Self-Pay	0.75
Managed Care, Unspecified	Non Self pay	1
Unknown	NA	0

- All the columns like **Age, APR Medical Surgical Description Race, Type of Admission, APR Severity of Illness Code, APR Risk of Mortality, Gender, Patient Disposition, and Emergency Department Indicator** have been encoded for the purpose of Linear Regression as all contained just Categorical values

Independent Variables
Permanent Facility Id, Age Group, Gender, Race, Length of Stay, Type of Admission, Patient Disposition, CCS Diagnosis Code, CCS Procedure Code, APR DRG Code, APR MDC Code, APR Severity of Illness Code, APR Risk of Mortality, APR Medical Surgical Description , Payment Typology, Emergency Department Indicator, Total Costs

Dependent Variables
Total Charges

E. Linear Regression Model

Once the Independent Variables are determined, the linear model is created.

The above table shows the each independent variables and the dependent variable used to create the linear regression model. Multiple linear models are built similarly checking the significance level of variables and to cross-validate the model.

Results and Analysis

Parameters for estimating the Total Charges billed for each patient have been collected from New York State Department of Health (public domain open dataset) inpatient discharges electronic medical records dataset.

A linear regression model was built based on the independent variables to predict the dependent variable.

A. Model Accuracy

The accuracy of the model was 83%.

```
SS_Residual = sum((y_test-y_pred)**2)
SS_Total = sum((y_test-np.mean(y_test))**2)
r_squared = 1 - (float(SS_Residual))/SS_Total
adjusted_r_squared = 1 - (1-r_squared)*(len(y_test)-1)/(len(y_test)-X_test.shape[1]-1)
print ("The r-square is:",r_squared)
print("The adjusted r-square is: ", adjusted_r_squared)
```

```
The r-square is: 0.8352311382337642
The adjusted r-square is:  0.8352231195727408
```

B. Significance of Variables

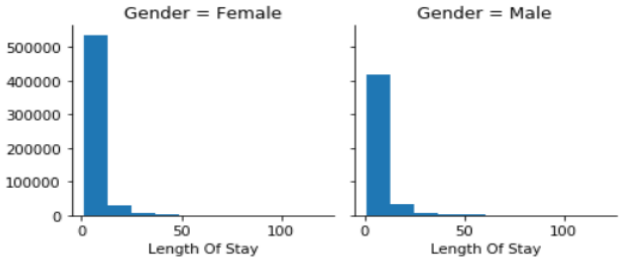
The below figure shows the importance of each factor in predicting the dependent variable (Total Charges). Each variable has particular significance in predicting of the model and the model performance is improved if we discard the less significant independent variables.

	importance
Total Costs	0.854855
Permanent Facility Id	0.091834
Length of Stay	0.011739
Type of Admission	0.009187
CCS Diagnosis Code	0.004710
CCS Procedure Code	0.004243
APR DRG Code	0.003980
Patient_Disposition_Encoded	0.003503
Payment	0.003416
Age	0.003404
APR MDC Code	0.002282
APR_Risk_of_Mortality	0.001946
APR Severity of Illness Code	0.001314
Race	0.001165
APR Medical Surgical Description	0.001052
Emergency_Department_Indicator	0.001015
Gender	0.000355

C. Visualizations of data with respect to Length of stay of patients:

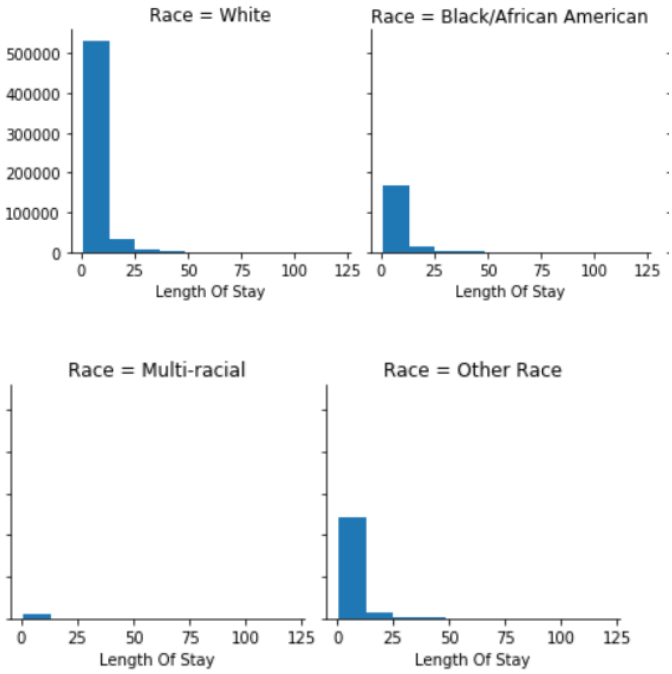
As Length of Stay has clearly been identified as one the most significant parameters to judge the Total Charges, we have identified some patterns in the data based on visualizations regarding the Length of Stay of Patients

1. Gender vs Length of Stay



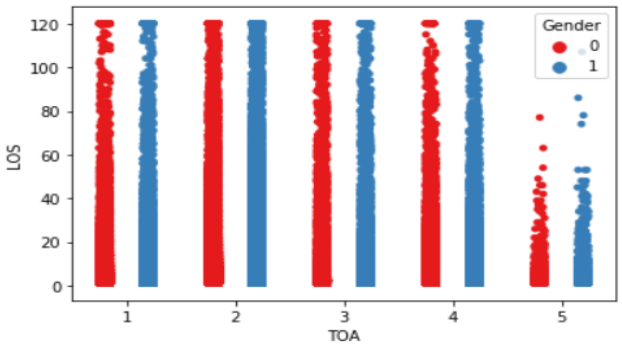
The above figure clearly shows that majority of female patients have stayed for a less amount of time in the hospital while Male patients in comparison have stayed longer. This can significantly impact the Total Charges of overall male patients as they have stayed for a longer period of time in the hospital.

2. Race of the patient vs Length of Stay



The above figure clearly shows that majority of patients in the dataset are white and have stayed for a longer amount of time in the hospital. There is hardly any patient records for Multi-racial records. Patients from Black/ African – American community have mostly been discharged pretty quickly. Patients belonging to the Other race in comparison have stayed less. The dataset here is a little imbalanced resulting in this kind of visualization.

3. Type of Admission of patients vs Length of Stay

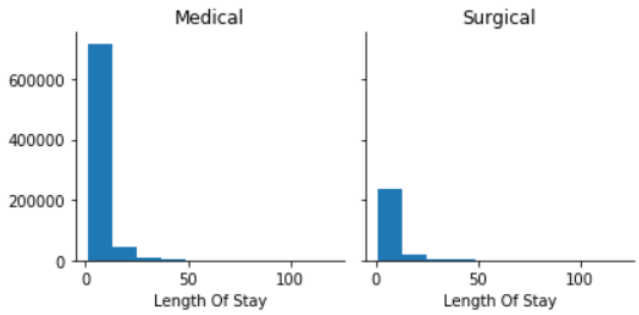


Type of Admission	
Elective	1
Emergency	2
Newborn	3
Urgent	4
Trauma	5

The above figure shows that both male and female patients who were admitted in Emergency have stayed longest at the hospital. Patients from both the genders admitted with Trauma have stayed the least or have been discharged to some other facility. Both male as well as female Newborns

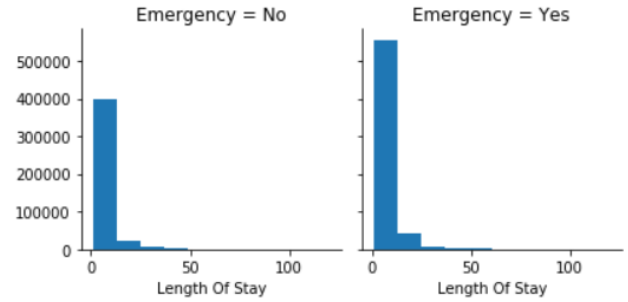
have stayed longer when compared to patients (both male and female) with Urgent admission. Male patients in the Elective type of admission have slightly stayed longer than the female patients.

4. Medical Surgical Description vs Length of Stay



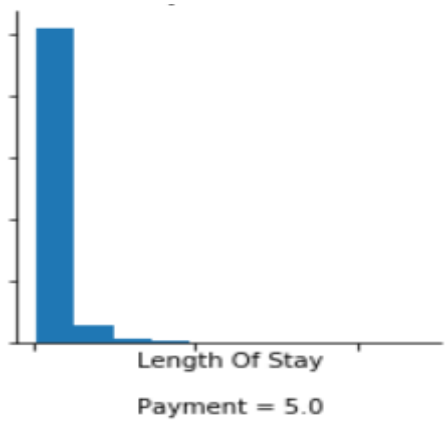
The above figure shows that most patients were admitted with normal Medical treatments have stayed 0-10 mostly at the hospital but some have stayed longer than even Surgical Patients. Patients with Surgical treatments are less in number and have stayed usually less days as well.

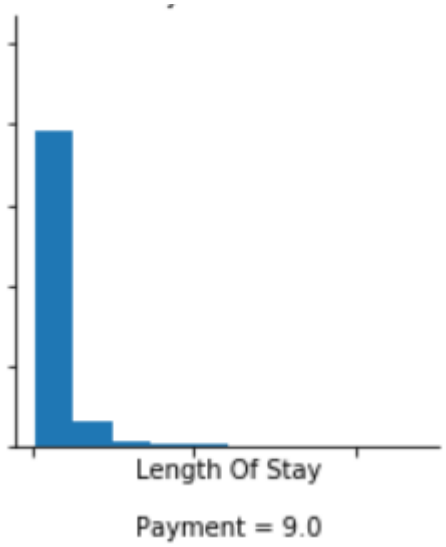
5. Emergency Department Indicator vs Length of Stay



The above figure shows that most patients who were admitted with Emergency condition have stayed for considerable length at the hospitals with some patients staying for more than 2 months even (60 days) at the hospital. Majority of patients without Emergency have been discharged within a week.

6. Payment Typology vs Length of Stay





The above figure shows that most patients who have paid using Insurance have stayed for considerable length at the hospitals with some patients staying for more than 2 months even (60 days) at the hospital. Majority of patients in the dataset have Insurance covers. Patients who have no insurance (self-paid) have stayed much less and are also much less in number.

Analysis and Discussion

The analysis of the results have led to many points of discussions that depict the centrality of our finding. The goal of the research was to find a statistical modelling approach for predicting the Total Charges a patient will be billed for by a medical provider (hospitals). Major discussing points are as follows:

A. The Importance of the Research Model Accuracy

Healthcare costs are ever increasing. Especially in times like these when we are going through a pandemic, the extent of clinical cost increments has been all around documented. But in any case, there have been limited efforts till date to evaluate the factors causing these increments. So our aim was to find a statistical approach to this problem in order to predict the Total Healthcare Charges for a patient beforehand using electronic medical data.

B. Pre-emptive information for insurance companies

Private Medical Providers have been charging extreme rates over the last decade and it is getting very difficult even for insurance companies as well to come up with policies that will benefit their clients as well as maintain a good revenue. Health insurers in India have been grappling with claims as high as ₹7 lakhs as private hospitals are seeking higher rates for treating a covid-19 patient. So having a predictive model for predicting the Total charges beforehand will go a long-way for the insurance companies to provide better future policies and covers for their clients

C. Information to patients and medical providers

Having a pre-emptive information regarding charges and payments always helps for in-patients and medical providers alike. The patients and their families get an idea of roughly how much they will be charged and can plan their payment procedures accordingly.
As for the medical providers, especially in countries like USA, Medicare legislation have standardized payments for patients based on the treatment, irrespective of the Length of Stay of a patient at the hospital. The lengthier Length of Stay of patients drains the hospital resources and revenue takes a hit. For medical providers this is a huge problem. Thus having information regarding the Total Charges of patients will definitely help the hospitals to better manage their revenue and plan treatment procedures accordingly.

D. Further scope of this research study

There is a lot of scope for further research on this topic as different countries have different medical providers and their rates vary depending on different factors like government policies, the overall facilities provided etc. Whether the provider is private or government aided (especially in countries like India) also makes a significant difference in the medical charges and this can be further researched using the statistical approach of our study.

Conclusions and Recommendations

We can conclude from our research study that having a model that can predict the Total Charges (expenses) for a patient can go a long way in benefitting not just the in-patients but also the insurance companies as well as medical providers.

- The predictive model will provide patients information which will help them in planning and payment of hospital bills
- The insurance companies can formulate better policy covers in order to serve their clients better as well as maintain their revenue
- Also for medical providers, it provides an unique opportunity to increase their revenue by better managing the treatment of patients who stay for a longer period (certain countries)

As part of recommendation (for Indian perspective), we feel the Government of India should look into the disparity of healthcare charges in private and government aided hospitals. With this study, we can have a pre-emptive information as for how much a particular hospital charges for the same treatment from a patient with respect to another provider. So the Government of India and the concerned authorities must come up with better healthcare budgets and policies so as to make sure this huge disparities of facilities and charges are not there and every citizen can be provided the best medical treatment possible

Limitations of the Research Study

Every research study and researchers face some challenges and limitations. Our study was no exception in this regard. The major limitations are listed below:

A. Availability of Healthcare Data

Healthcare information is usually not available in public domain as it is very private data. Hence to get data for the research was a huge challenge. The dataset for the study (New York State Department of Health dataset for in-patient discharges) also was not fit for the model creation. Hence we had to treat the data and make it fit for the best possible outcome of our study

B. Missing Values and Extreme values

The dataset also had a lot of missing values and random values. There were also some extreme outlier values and they had to be treated in order to create a better statistical model

C. Data Concentrated From One Particular Location

Even though our research was on a general overview of creating a model for predicting the Total Medical expense of a patient during a hospital stay, due to lack of data we had to concentrate on data from one particular State (New York). So providers from different regions charge patients differently and treatment rates also vary. But the statistical model we created can actually work on any dataset from any region thus overcoming this limitation.

References

[1] Alvaro J. Riascos and Natalia Serna. (n.d.). Predicting Annual Length-Of-Stay and its Impact on Health. Retrieved from https://pdfs.semanticscholar.org/7582/a9570c45a3636736b10de66d20689eb5ae37.pdf

[2] Yeonkook J. Kim and Hayoung Park. (n.d.). Mary Ann Liebert Inc Publishers.

[3] en.wikipedia.org

[4] https://data.worldbank.org/