**Module 7: Portfolio Project**

**Hospital Inpatient Discharge Data**

Willis Silliman

Management Information Systems, CSU Global

MIS581: Capstone: Business Intelligence and Data Analytics

Dr. Justin Bateh

July 23, 2023

**Abstract**

The state of New York had over two million hospital patients in the year 2013. Hospital beds are a precious resource. Data released by the state of New York detailing the lengths of stay and costs of 2.4 million patients were analyzed to determine whether there were patterns in the lengths of stay that could be exploited to help identify when patients stayed longer than necessary. Costs were also analyzed to figure out if the hospital benefits financially from longer stays. Two systematic biases, region and race, were tested to figure out if certain patients were staying longer. Severity of illness increased the length of stay the longest, and treating extreme illnesses earned the hospital the most money. The length of stay was shown to increase hospital profits. Additionally, both sources of bias were found to contribute to the length of stay for patients. These indicate the need to evaluate the discharges of patients in New York, as there are patients with unnecessarily long stays.

*Keywords*: hospitals, patient stay, patient cost, race, New York

**Introduction**

One of the most important resources in a hospital setting can be the beds for the inpatients. As seen during the surge of COVID-19, the number of hospital beds that are available for patients is critical for care. According to Gong et al. (2022) the number of patients admitted to public hospitals increased by 5.8% in China in 2019. This was before the pandemic, so the number of patients has only increased since then.

Hospital flow rate is a difficult problem to manage. Many models and frameworks have been proposed to manage the complex issue. And there continue to be many ethical and clinical challenges that are faced by hospitals who are looking to be more efficient for the sake of their patients (Jankowski et al., 2009). Time and money are extremely important in the hospital, and any waste of these are unacceptable for an industry that is expected to serve everyone.

Thus, the problem is twofold. The first part is that there may be patients who are being released too late. The second part of this problem is the resource cost analysis. If patients are staying too long in the inpatient beds, what is the cost to the hospital, and to them, that is being wasted. In addition to these, bias may play a role in keeping certain patients in hospital beds for longer that should be addressed by the hospitals.

**Objectives**

There are a few different tasks that can be performed with the data available. The length of stay is available, which leads to a prediction task. The length of stay might be able to be predicted, which could help with the scheduling of beds in the inpatient section. This variable could be based on many of the factors that come with the data, such as whether the admission was urgent or elective. However, since the other variables that could be potentially used to predict the values are categorical, and they are fairly unbalanced, this leaves a mixed linear model as perhaps the best approach to this data.

In addition to this, perhaps some cases that were staying much longer than expected could be found. These could serve as a proposal to have been discharged earlier, though with publicly available data it is impossible to say for sure. Using these predictions could result in seeing how much inpatient stay could have been saved for the hospital, which may have been allocated to other incoming patients. This will likely come in the form of a statistical model that can be interpreted statistically, such as a linear regression curve, or a logistic regression with thresholds to determine the class.

There are numeric values for the financial cost to the hospital that each of these patients had, and how much was charged to recoup the cost. This data could also be analyzed to determine whether there is an effect on the length of the stay. In addition, with the last model proposed, this could serve to estimate how much revenue could have been saved by discharging patients on time. This could also be a regression problem, or it could be an analysis of the distribution to look for outlier values.

**Overview of the Study**

In this study on predicting hospital patient discharge, and the causes of long stays and high bills in the hospital, a dataset from the state of New York hospitals will be analyzed. This dataset contains many variables that can be used to try to predict the length of stay of a patient in the hospital. Being anonymized, many of these values are categorical, leading to a method that predicts mean values of the patients within a subgroup. Hospital costs are also reported, which can be fit to with a linear model. This will allow for the prediction of patient costs, allowing increases in prices dynamically, or, hopefully, with the help of the first project reducing the number of patients staying overtime in the hospital. Finally, an analysis of systemic biases in the hospital systems will shed light on areas of improvement for hospital stay in non-white and rural patients.

**Research Questions**

The hospital inpatient discharge dataset leads to many questions that can be answered with the data. The main question has to do with the amount of time that a patient is in the hospital. This is given by a variable that gives the number of days. Minimizing this value without causing patients unnecessary burden is key to managing a hospital. There are a few different categorical variables that the length of the stay can be related to. The most likely ones to cause a difference between classes are likely: age group, severity of illness, and procedure received. The question we are asking is: how do these three variables affect the predicted expected value of the length of stay?

The second analysis to be performed has to do with the cost of the patient to the hospital. It is important to try to minimize these costs, as there are unnecessary expenses. Patients who are predicted to cost the hospital more should have the bill raised for their stay. Since cost is a continuous variable and is going to be affected by another continuous variable, the length of stay, this will be a linear model based on it and several other variables, such as if they were admitted to the emergency room. The question is then, how do each of the variables contribute to the cost of stay?

Finally, there can be confounding factors that explain outliers in the data. There might be an effect of race, ethnicity, and hospital service area on the prior analyses that will not be tested for. These could be analyzed separately in a t-test to address systemic bias in hospital care. These variables will likely dictate the quality of care and thus the time spent in the hospital. There may also be a difference in the costs associated with different healthcare plans, such as Medicare. These can contribute to bias in the model, but it is difficult to address them in the model itself if one group is under-represented.

**Hypotheses**

The first set of hypotheses are about the relationship between age, severity of illness, and procedure to the number of days stayed in the hospital. Age is a large factor here, since it is known that many obstacles to patient transfer are increased by age (Smith et al., 2022). The alternate hypothesis (Ha) is that age, severity of illness, and the severity of the procedure will all have a negative effect on the time spent in the hospital. Increased age, a more severe illness, and a most serious surgery will all increase the mean stay in the hospital. The null hypothesis (Ho) is that there is no effect of these variables on the mean, or that at least one of them will decrease the mean.

The second set of hypotheses refer to the linear model of the cost to the hospital. It refers to several variables all making a linear model. These are: number of days in the hospital, type of admission, the type of procedure, and whether the patient was admitted to the emergency department. The alternate hypothesis (Ha) is that these variables will all have a statistically significant effects on predicting the cost. This means, each coefficient will have a statistically significant F statistic. The null hypothesis is that one or all of these will not be significant to the linear regression model.

The final sets of hypotheses have to do with testing a few variables against the length of stay. There will be two tests, so two sets of hypotheses will be needed. The first hypothesis has to do with race and ethnicity. These two categories will be combined since ethnicity only has two values. Then, the test is a comparison between groups, the average number of days spent in the hospital. The alternate hypothesis (Ha) is that non-white patients will stay in the hospital longer than other racial groups. The null hypothesis (Ho) is that there will be no difference between the groups, or white patients will stay less time in the hospital on average.

Last, there will be a comparison of different regions in New York to test whether the funding different regions get is affecting the patient stay. The alternate hypothesis (Ha) is that there will be a significant difference between the average patient stay between regions. The null hypothesis (Ho) is that there will be no difference in the patient stay of the different regions.

**Research Review**

Many papers have focused efforts on predicting patient stay using machine learning techniques, or complex stochastic models. A paper by Barnes et al. (2016) formulates the problem in a way to use classification rather than predicting a continuous variable. This is done by creating a model that should predict whether a patient should be released on the given day. There were also two time frames per day it was tested on: 2pm releases and end of day releases. This allows for prediction in advanced, if the patient does not have any variables change between days but will grow more error prone the further out the prediction. Due to this, it does not allow for long term prediction for scheduling in advanced.

A stochastic network model was proposed by Shi et al. (2016). This model seeks to optimize the waiting time to get a bed, but it proposes a policy that affects discharges, since the two are interlinked (patient flow). The model created a policy that discharges patients early in the morning, and as a result there were more open beds in the morning to accept patients. This policy is in opposition to most hospital’s discharge policies, which perform most discharges in the early afternoon. This policy relieves administrative stress since fluctuating discharge times are more difficult to handle for the staff.

These papers both propose solutions to the problem of prediction, which are good to schedule patient beds. However, they do not allow for the analysis of outlier patients. Since many machine learning models are prone to overfitting data and do not allow for statistical analysis. However, the goal of this project has to do with determining which patients could have a shorter stay. The papers on machine learning prediction both take doctor recommendations as complete truths, but there may be mistakes in the scheduling that could be corrected to reduce patient stay.

Part of the interest in reducing stay is also a cost analysis. Longer stays cost patients more. Longer stays also are not as good for hospitals, since care expenses are not the majority of most bills. Additionally, there is an opportunity cost when turning down other patients due to a lack of space. The length of stay variable is likely a part of reducing this variable.

Cost analyses are common for hospitals. Academically, they are not common to pair with length of stay, since that is an obvious causative value, and may be used to scale rather than as a predictor. The article by Kalakoti et al. (2019) tries to use similar variable to this study to predict both cost and length of stay, specifically for patients undergoing cervical spinal surgery. These variables are linked, but they are not always the same. So, procedures might be more expensive to the patient than others, which would bias their cost towards being more expensive, even if they stayed in the hospital for less time. Several variables from the dataset will be included to account for this bias.

There is evidence across multiple cases that there is a systematic bias in the length of stay between different races or ethnicities. When compared with White patients, non-White patients have a higher average length of stay. This is supported by both Zheng et al. (2022) and Kameg & Lee (2023) for both cardiac and psychology admissions. This systemic bias should be addressed to make more beds available for patients. Some patients are being given a longer stay than necessary because of their race, which should be adjusted. This is an ethical problem that also costs the hospital more bed space than necessary.

**Methodology**

The dataset being analyzed is from a public release of hospital inpatient data from the state government of New York. It includes all inpatient admissions for the year 2013 and was made publicly available in December 2014. The data has over 2 million entries from hospitals in New York state. As this dataset contains no text data to analyze, the analysis will be completely quantitative. Statistical analysis methods of regression, ANOVA, and t-tests will be used to test hypotheses.

**Methods**

The first question requires a between group comparison of means. Because there is a large class imbalance between the intersections of these groups, this leaves a mixed linear model as the best approach. These classes will not be very balanced if the entire dataset is used. It is important to not perform multiple comparisons, as this could lead to erroneous results purely by chance (Westfall, 2015). This could require a power calculation and sampling to avoid an unbalanced design, however, there are so many rows in the data that this is unlikely. There are a few datasets from different years available to the public, so there can be some data added to the dataset to create a powerful test with many samples. Regarding tools, a mixed linear model is a form of generalized linear model, which can be fit in R. There are contested ways to determine statistical significance of the results.

The second test requires fitting a linear regression model. A linear regression model can be fit automatically in R. The coefficients, and additional bias terms, can then be tested using the F statistic. These are done automatically when creating a linear model in R. If one of the variables is not significant, it can be dropped, and the model can be refit. The other variables will also need to be checked in this model because the model will be different after fitting it again.

The last two are variables that can be analyzed with a t-test and an ANOVA test. These variables will likely be unbalanced, again. This can be fixed either with adding data or using a non-parametric test. These can be weaker than parametric tests, so it may be better to sample larger numbers of data from the different years provided by the state of New York. The data may also not be normal. In fact, it is likely very skewed, since more patients are expected to stay little time, whereas the time in the hospital can reach over 120 days. It was found in other papers that studied a similar metric that a log transformation was at least used to normalize the data (Kameg & Lee, 2023). These could be able to be fixed with a transformation or might require the use of a non-parametric t-test. It is possible no transformation would be able to normalize the data, such as if there is a bimodal distribution. However, it is also possible that the data will have enough rows for a non-parametric test to be enough for comparison in the t-test. R defaults to this non-parametric test.

**Limitations**

Since the health data provided by the New York government is a part of a dataset containing a lot of sensitive information, much of the data cannot be released to the public. There are many pieces of information that would be useful to predict patient outcomes that cannot be shared in public data sources. Specific health conditions, previous medical history, individual biological measures like blood pressure, or personal information like amount of exercise all likely contribute to the individual length of stays. These cannot be released to the public to preserve the anonymity of the patients. However, hospitals could run a study on this information by consenting many patients, or using the historic data they retain.

**Ethical Considerations**

With any health data, there are ethical considerations about revealing too much of the data through a research paper. If identifiable information about a patient is released through a research paper, even with no names attached, individuals can be tracked down through this information. IRBs have requirements to release data on patients involved in a research study. The data being used in this paper is anonymized and released by a governmental institution. There are pieces of information in the dataset that could perhaps be used to identify the patient, such as the total of their bill, insurance information, and approximate age. However, these would require prior knowledge of the patient to find them. By releasing the dataset, New York has confirmed that these satisfy New York HIPPA standards.

Another ethical consideration is the harm that following the results of this analysis could bring. Often, the job a data analyst is performing is trying, in some way, to maximize the profits of the company. This can run counter to the health of patients in the case of a hospital. In the case of this analysis, if it is indeed most profitable to keep patients longer, this would be a way of earning money the hospital could exploit. This would run counter to the goals of the hospital but would also make the most profit, which is another incentive hospitals often follow. There is, of course, a risk of jeopardizing public relations with this, but even in the meantime, it is not okay to risk the health of patients for profit. In addition, the profitability of the hospital is not a major concern, as it is an inflexible market, and hospitals can make up for expenses in other areas (Luan, Shao, & Dou, 2020). This means that there is not really an excuse to focus on maximizing profit when patient safety is at stake.

**Results**

The results of each analysis were conclusive. A mixed linear model gave very significant results for almost each of the categorical variable tried on the data. The only variables that did not show extreme significance (<2e-16) were the ages 18 to 29, 30 to 49, 50 to 69, and the “Unknown” severity of illness category, which was used to fill in unknown values in the table. Surprisingly, the group with the longest stays were in the 18 to 29 age brackets when the surgery and illness were marked as extreme. This is likely because older patients are much more likely to die in the hospital when they have an extreme illness. Of these four variables, each were significant except for the ages 30 to 69 range. Summary tables of these values can be viewed in Tables 1 and 2 below. Ultimately, the null was rejected, and these variables are significant to the model.

A linear model was fit to predict the cost to the hospital, which was defined as the charges the hospital brought minus the cost of the patient. This was based on the length of stay, the type of admission, the illness severity, and if the patient was admitted to the emergency room or not. In this analysis, each of the variables used to fit the linear model were deemed significant. Again, the only variable that had a higher p-value was the “Unknown” category for the severity of illness. As such, the null is rejected. As feared in the ethics section, the length of stay increases the profits of the hospital by about $1100, the highest increase of any variable. This could be used to increase the profits of the hospital. Although, the costs of surgery are not known in this data, so perhaps more expensive surgeries could result in longer hospital stays. The coefficients of each variable and the p-value results can be seen in Table 3 below.

Finally, both hospital area and race were found to impact the number of days spent in the hospital. White patients were more likely to have a slight increase in the number of days spent in the hospital compared to non-white patients. The histograms for this data can be seen in Figure 1. White patients had a slightly larger left side in the histogram. Additionally, rural patients seemed to have shorter times in the hospital, though results between areas fluctuated too much to be conclusive. The means for this analysis can be seen in Table 4. As a result, we reject the null for both tests.

**Table 1.**

*Collective results of the significance levels of each variable used to fit the mixed linear model.*

|  |  |  |
| --- | --- | --- |
| **Variable** | **F value** | **p-value** |
| Age Group | 605.951612 | < 2.2E-16 |
| Severity of Illness | 35839.0734 | < 2.2E-16 |
| Risk of Mortality | 944.576283 | < 2.2E-16 |
| Surgical or Medical | 5905.93939 | < 2.2E-16 |

**Table 2.**

*Results of each category ‘coefficient’ in the mixed linear model.*

|  |  |  |  |
| --- | --- | --- | --- |
| **Category Name** | **Estimate** | **t value** | **p-value** |
| (Intercept) | 2.2660 | 23.473 | < 2.2E-16 |
| Age: 18 to 29 | 0.0891 | 4.583 | 4.58E-06 |
| Age: 30 to 49 | 0.0038 | 0.225 | 0.82161 |
| Age: 50 to 69 | 0.0027 | 0.161 | 0.8719 |
| Age: 70 or Older | -0.6128 | -33.339 | < 2.2E-16 |
| Severity of Illness: Moderate | 1.8920 | 156.893 | < 2.2E-16 |
| Severity of Illness: Major | 4.3030 | 236.502 | < 2.2E-16 |
| Severity of Illness: Extreme | 12.9300 | 368.928 | < 2.2E-16 |
| Severity of Illness: Unknown | 3.1530 | 2.773 | 0.00556 |
| Risk of Mortality: Moderate | 0.3661 | 24.166 | < 2.2E-16 |
| Risk of Mortality: Major | 1.1570 | 52.641 | < 2.2E-16 |
| Risk of Mortality: Extreme | 1.1840 | 31.046 | < 2.2E-16 |
| Surgical or Medical: Surgical | 0.9025 | 76.850 | < 2.2E-16 |

**Table 3.**

*Results of the fitting a linear model to the defined hospital profit variable.*

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Estimate** | **t value** | **p-value** |
| (Intercept) | 5001.924 | 49.296 | < 2.2E-16 |
| Length of Stay (Days) | 3513.600 | 1102.580 | < 2.2E-16 |
| Type of Admission: Urgent | -5087.371 | -49.215 | < 2.2E-16 |
| Type of Admission: Emergency | -4648.292 | -49.379 | < 2.2E-16 |
| Type of Admission: Trauma | 8179.578 | 15.238 | < 2.2E-16 |
| Type of Admission: Newborn | -10057.813 | -104.973 | < 2.2E-16 |
| Type of Admission: Unknown | -6698.119 | -7.762 | 8.26E-15 |
| Severity of Illness: Moderate | 341.536 | 5.922 | 3.18E-09 |
| Severity of Illness: Major | 5774.920 | 81.312 | < 2.2E-16 |
| Severity of Illness: Extreme | 33636.678 | 277.678 | < 2.2E-16 |
| Severity of Illness: Unknown | 13041.009 | 2.302 | 0.0213 |
| Emergency Department: No | 3379.370 | 41.252 | < 2.2E-16 |

**Table 4.**

*Health Service Area means, standard deviations, and counts.*

|  |  |  |  |
| --- | --- | --- | --- |
| **Health Service Area** | **Mean** | **SD** | **Count** |
| Capital/Adirond | 5.317031 | 7.27047 | 164757 |
| Central NY | 4.990639 | 6.724189 | 162488 |
| Finger Lakes | 5.24734 | 8.004066 | 151205 |
| Hudson Valley | 5.933361 | 8.464908 | 251985 |
| Long Island | 5.368399 | 7.312038 | 352289 |
| New York City | 5.571184 | 8.653229 | 1138713 |
| Southern Tier | 5.02685 | 6.719273 | 31210 |
| Western NY | 5.318163 | 8.032926 | 170328 |

**Figure 1.**

*White vs Non-White Patients, Days in the Hospital Histograms.*A graph of patients with white and black text

Description automatically generated

**Conclusion**

There are several factors that can be used to predict the length of stay of a hospital patient. Severity of illness tends to have the most effect on a patient’s length of stay with the risk of mortality being the second highest. Surprisingly, older patients are likely to stay a shorter period, though this is most likely because of the increased mortality of patients who are older. Although not many variables were able to be analyzed due to the limited nature of public datasets, this model gives a rough estimate of how long to expect a patient to stay in the hospital. Anything outside this norm may have other contributing factors that cannot be identified in this analysis.

Hospitals are interested in reducing their overhead, and this was turned into a method to measure the hospital profit using linear approximation based on the patient. This analysis revealed that as the number of days in bed increases, the more the hospital profits. This is troubling but does make sense. When a patient is in bed, there are not many expenses for keeping them there, but the hospital can continue to charge. This analysis was not able to take opportunity cost into account, however. There are certainly procedures that make the hospital a lot of money that could not be performed if the hospital was full. This would create an incentive to discharge patients on time. In addition, this variable could have been dependent on the type of surgery a patient got, meaning a more expensive surgery could require a longer hospital stay. More analysis on this topic should be done by individuals that have access to such data.

Finally, there are biases between service areas and towards certain patients. There are certain areas in the state of New York in which the average stay in the hospital is unusually long. In order, the top three are: Hudson Valley, Long Island, and Western New York. Long Island is an outlier, but there seem to be longer stays in areas that serve rural populations. This could be due to the funding they receive, or it could be due to more accidents happening in rural areas. Whatever the case, this should be noted by the state to figure out why patient stays are longer in these areas.

Another systematic difference that should be reviewed by New York is the difference in stays between White and Non-White patients. There appears to be only a slight difference between the groups, however this slight difference in means can add up when it is applied to the thousands of Non-White patients that are hospitalized in New York each year. Unless there is some other explanation, this appears to be a systemic discrimination against non-white patients that costs them more money on average, according to the profit model.

The results of these analyses provide a framework to check for patients who are outliers. These could be outliers because of one of the issues with bias raised here, or for another reason their chart might reveal. However, there are indeed biases in the system that should be addressed to free up more hospital beds.

**References**

Barnes, S., Hamrock, E., Toerper, M., Siddiqui, S., & Levin, S. (2016). Real-time prediction of inpatient length of stay for discharge prioritization. *Journal of the American Medical Informatics Association: JAMIA, 23*(e1), e2-e10. <https://doi.org/10.1093/jamia/ocv106>

Gong, X., Wang, X., Zhou, L., & Geng, N. (2022). Managing hospital inpatient beds under clustered overflow configuration. *Computers and Operations Research*, 148. <https://doi.org/10.1016/j.cor.2022.106021>

Jankowski, J., Seastrum, T., Swidler, R., & Shelton, W. (2009). For Lack of a Better Plan: A Framework for Ethical, Legal, and Clinical Challenges in Complex Inpatient Discharge Planning. *HEC Forum*, 21(4), 311–326. <https://doi.org/10.1007/s10730-009-9117-6>

Kalakoti, P., Gao, Y., Hendrickson, N. R., & Pugely, A. J. (2019). Preparing for Bundled Payments in Cervical Spine Surgery: Do We Understand the Influence of Patient, Hospital, and Procedural Factors on the Cost and Length of Stay? *Spine (Philadelphia, Pa. 1976), 44*(5), 334-345. <https://doi.org/10.1097/BRS.0000000000002825>

Kameg, B. N., & Lee, H. (2023). Factors influencing length of hospital stay among veterans admitted to an inpatient psychiatric unit. *Archives of Psychiatric Nursing*, 43, 106–110. <https://doi.org/10.1016/j.apnu.2022.12.001>

Luan, M., Shao, X., & Dou, F. (2020). Financial conditions, health care provision, and patient outcomes: Evidence from Chinese public hospitals. Economics Letters, 186. <https://doi.org/10.1016/j.econlet.2019.108876>

Smith, H., Grindey, C., Hague, I., Newbould, L., Brown, L., Clegg, A., Thompson, C., & Lawton, R. (2022). Reducing delayed transfer of care in older people: A qualitative study of barriers and facilitators to shorter hospital stays. *Health Expectations*, 25(6), 2628–2644. <https://doi.org/10.1111/hex.13588>

Shi, P., Chou, M. C., Dai, J. G., Ding, D., & Sim, J. (2016). Models and Insights for Hospital Inpatient Operations: Time-Dependent ED Boarding Time. *Management Science, 62*(1), 1-28. <https://doi.org/10.1287/mnsc.2014.2112>

Westfall, P. H. (2015). Multiple Comparisons, Statistics of. *International Encyclopedia of the Social & Behavioral Sciences*, 81–87. <https://doi.org/10.1016/B978-0-08-097086-8.42053-2>

Zheng, J., Tisdale, R. L., Heidenreich, P. A., & Sandhu, A. T. (2022). Disparities in Hospital Length of Stay Across Race and Ethnicity Among Patients with Heart Failure. *Circulation. Heart Failure, 15*(11), e009362-e009362. <https://doi.org/10.1161/CIRCHEARTFAILURE.121.009362>