Healthcare Associated Infections Insights

Sukrut Shishupal, Govinda Dhakal, Kolton Hauck

Introduction

About 5% of all hospital admissions in the United States result from a Healthcare-Acquired Infection (HAI)[1]. This is a significant and surprising proportion considering hospitals are supposed to be places to treat infections and other ailments, not be a source of spreading them. These infections can be spread through devices such as central lines and urinary catheters, or from one patient to another after contact with an infected person or surface. They also increase the amount of hospitalised patients and healthcare costs[2].

There has been research into the investigation of various types of HAIs. For example, Matthew et al. investigated Surgical Site Infections (SSIs) which is "estimated to affect 2% to 5% of all surgical patients"-their causes, possible preventions and so on[3]. Cristina Valencia et al. investigated another type of HAI: central line-associated bloodstream infections (CLABSI) across multiple countries and what may contribute to this preventable infection[4].

The Center for Medicare and Medicaid Services (CMS) has implemented a program to help reduce HAIs in the United States: the Healthcare Acquired Condition Reduction Program (HACRP)[5]. With this implementation, HAIs have been tracked and these are available by hospitals: the CMS "Healthcare Associated Infections - Hospital" dataset[6]. The primary features this dataset provides for each HAI instance: what hospital the infection was found in, State, County, what kind of infection, the Healthcare Acquired Condition (HAC) score which CMS uses for HACRP, how the HAC score compares with the national average, and the start and end dates of the infection. CMS also provides general hospital information, such as MORT score of mortality and hospital ratings in their "Hospital General Information" dataset[7].

HAIs are an interesting and surprising healthcare topic. Our group is interested in analysing the dataset and predicting the likelihood of healthcare associated infections based on the hospital scores using statistical analysis. We attempt to extract the data developed by the CDC and collected through NHSN. Since we are amid emerging global pandemic due to infectious disease, our team found this topic interesting and relevant. The prediction outcome is expected to be useful for all of us to mitigate the infections during hospital admission. Healthcare facilities can be benefitted in terms of how they can improve their performance to reduce infections in humans.

The primary focus of our project is to identify primary predictors of the Standard Infection Ratio (SIR) score. Each infection type (6 in total) are tracked in the CMS dataset. The SIR score is the ratio between the predicted number of cases of an infection type and the actual number of cases of the same infection type. Are there correlations between the various predictors found in the dataset and the SIR score? Which of these predictors contribute most to better SIR scores? We would like to be able to describe and quantify these relationships between the predictors and the SIR score.

We hypothesise that location will be a strong predictor of SIR score: showing differences in HAI prevention measures with some regions or states performing better. We also hypothesise correlations between the SIR score and other hospital-ranked scores, such as the MORT score of mortality and the type of hospital ownership.

Methods

The datasets were obtained from these links:

Healthcare Associated Infections - Hospital:
- https://data.cms.gov/provider-data/dataset/77hc-ibv8

Hospital General Information:
- https://data.cms.gov/provider-data/dataset/xubh-q36u

Data Cleaning

After we downloaded the datasets, we performed the following steps to get a better view of the dataset. To clean the datasets, we removed the irrelevant columns to any analysis we would make (Footnote, Start Date, End Date, Phone Number, and Address). Once we removed these columns, we removed the rows that had a null Score value, as this Score value will be the focus of our project. Finally, we plotted a missmap plot to check for missing values.

Descriptive Statistics

Once the data was sufficiently cleaned, we performed basic descriptive statistics to get a better understanding of the dataset. First, measures of central tendency, such as mean, standard deviation, maximum, minimum and ranges were calculated for each of the SIR scores. Hospital rating means and ranges differentiated by different ownerships were also calculated.

We then plotted several histograms to understand the distributions of several columns: SIR scores across the whole dataset, hospital ratings by state, hospital type and ownership, and mean scores of CLABSI (one of the HAIs) by state.

Then, boxplots were created of the SIR scores, faceted by State, to see the distribution of SIR scores across states. This plot indicates the median, maximum and minimum values, and interquartile range and marked the outliers. This is the best way to understand how a particular state is performing.

QQplots, faceted by HAI type, were created to determine the normality of the distributions.

Results

Data Cleaning

Figure 1 shows the missingness map of the cleaned data.

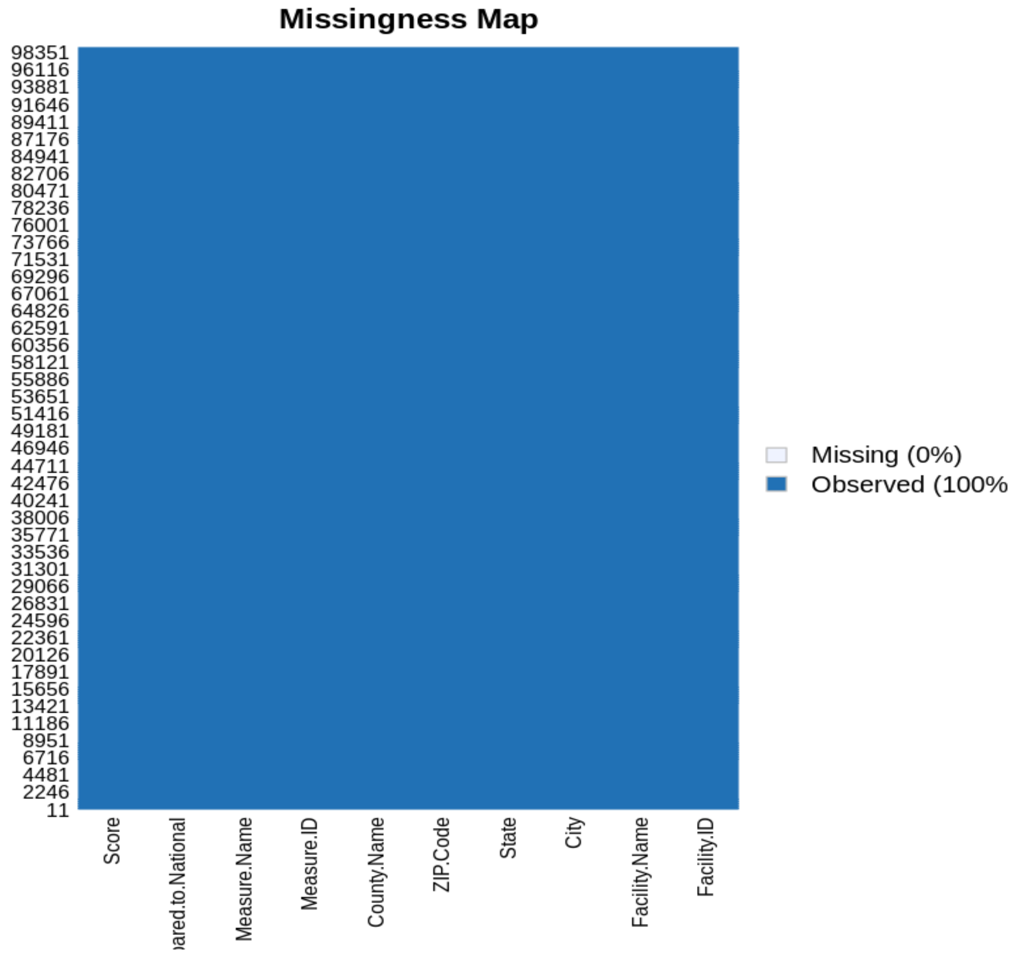

Fig. 1: Missingness map

We can see there is no missing data.

Descriptive Statistics

Table 1 shows the calculated measures of central tendencies of the SIR scores. SIR score changes as the city changes, so it is crucial to understand these cities' mean, maximum, minimum, and standard deviation values for the different infections, such as HAI-1, HAI-2, etc. This can help us understand the worst-performing cities based on the HAC (Hospital Acquired Conditions) scores.

Table 1: Central Tendencies of SIR Scores

HAI SIR	Mean	Std. Dev.	Max	Min	Range
1	1.108	1.009	18.755	0	18.755
2	0.873	0.772	6.164	0	6.164

3	**0.815**	**0.697**	**4.364**	**0**	**4.364**
4	**1.009**	**0.984**	**8.867**	**0**	**8.867**
5	**1.130**	**0.890**	**7.488**	**0**	**7.488**
6	**0.512**	**0.479**	**5.004**	**0**	**5.004**

As we can see, the mean varies significantly for HAI-2 and HAI-5 infections. Hence, these infections are common in city hospitals.

Using range and mean values, calculated and depicted in Table 2, we can understand the high and low values of the Ratings based on the hospital ownership column.

Table 2: Mean and Ranges of Hospital Ratings, by Hospital Ownership

Hospital Ownership	**Mean Rating**	**Rating Range**
Physician	3.684	1-5
Voluntary non-profit - Church	3.426	1-5
Voluntary non-profit - Other	3.396	1-5
Voluntary non-profit - Private	3.344	1-5
Government - State	3.054	1-5
Government - Federal	3.000	2-4
Government - Local	2.966	1-5
Government - Hospital District	2.962	1-5
Proprietary	2.899	1-5
Tribal	2.000	2-2

Figure 2 shows the distribution of the SIR scores. We plotted the score value and the number of times HAI infection has been reported. We can see that a score of around one is reported multiple times. The data is also right skewed, meaning there are a lot of higher scores compared with lower scores. This does not look like a normal distribution (QQplots were made - discussed later). So any analytical tests that are performed will need to be ones that don't require a normal distribution.

Another important parameter that we want to investigate is the hospital ratings. We used a bar graph to understand the distribution of the data, shown in Figure 3. We can see that the hospital average rating is 4.

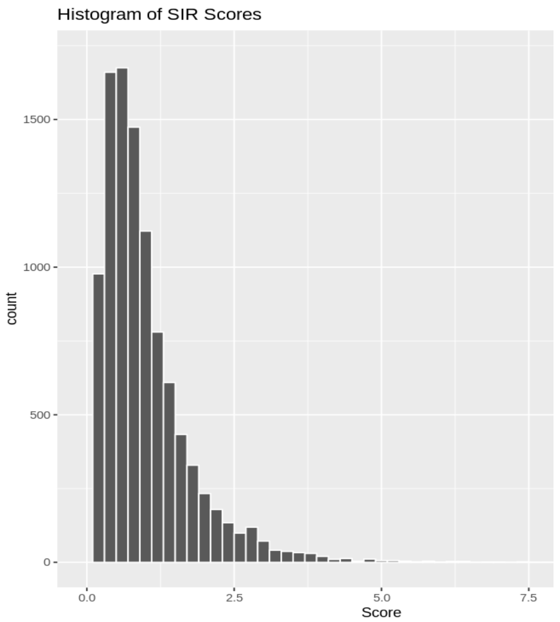

Fig. 2: Histogram of SIR Scores

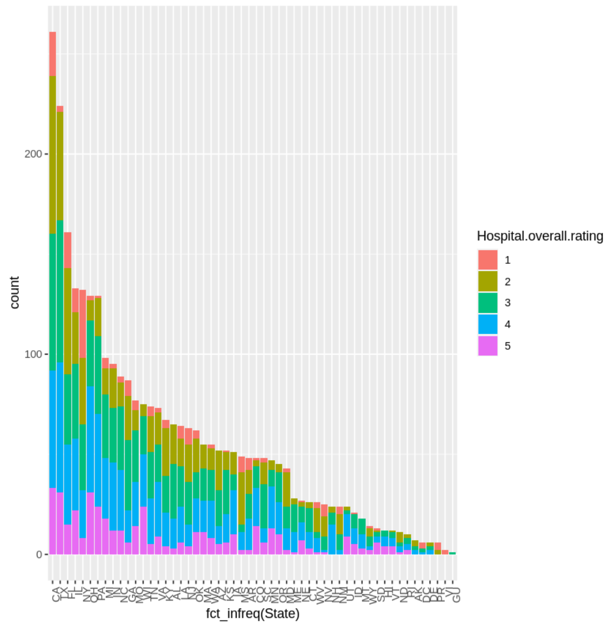

Fig. 3: Hospital rating by State

There are various types of hospitals, such as Acute care, Critical care, and other types of hospitals, that treat patients based on their symptoms. We check the number of such hospitals in the dataset using a bar graph, depicted in Figure 4. As we can see, acute care hospitals have the highest count as compared to other types of hospitals.

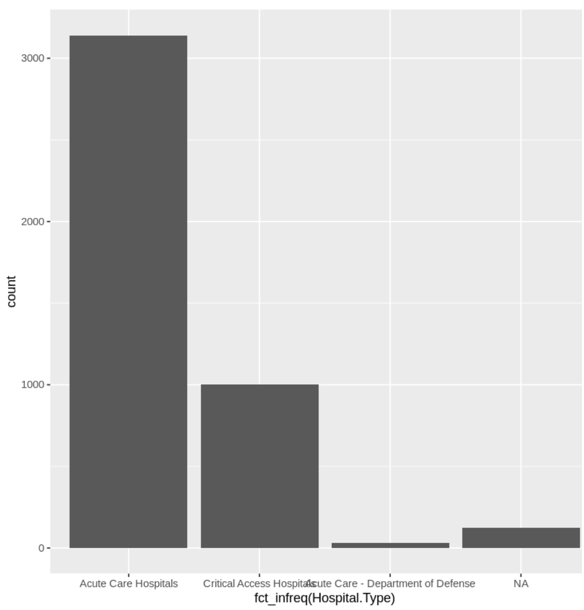

Fig. 4: Type of Hospitals

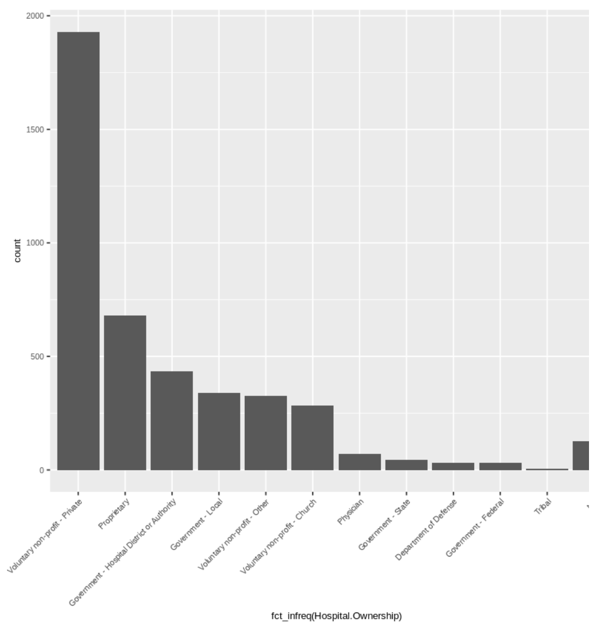

Fig. 5: Hospital Ownership

There are hospitals owned by different organisations such as voluntary, government, tribal, and so on. We are plotting the ownership graph - Figure 5 - to understand the distribution of ownership of the hospitals. Most of the hospitals are voluntary and non-profit owned. Even though Voluntary non-profit Private has the highest number of hospitals, when it comes to rating, they are below voluntary non-profit others and churches which have far less number of hospitals. The physician-owned hospitals have the highest mean rating, although their number is pretty low compared to other hospital owners.

Once we know the number of hospitals and their ownership, we can calculate the mean rating of that ownership. We tried to understand the HAC score distribution across different states, hence we calculated the mean score and plotted a graph in Figure 6 based on those values. Since type of hospital and hospital ownership may influence the HAC score of a hospital, these histograms give insight into any biases towards an disproportionate type or ownership.

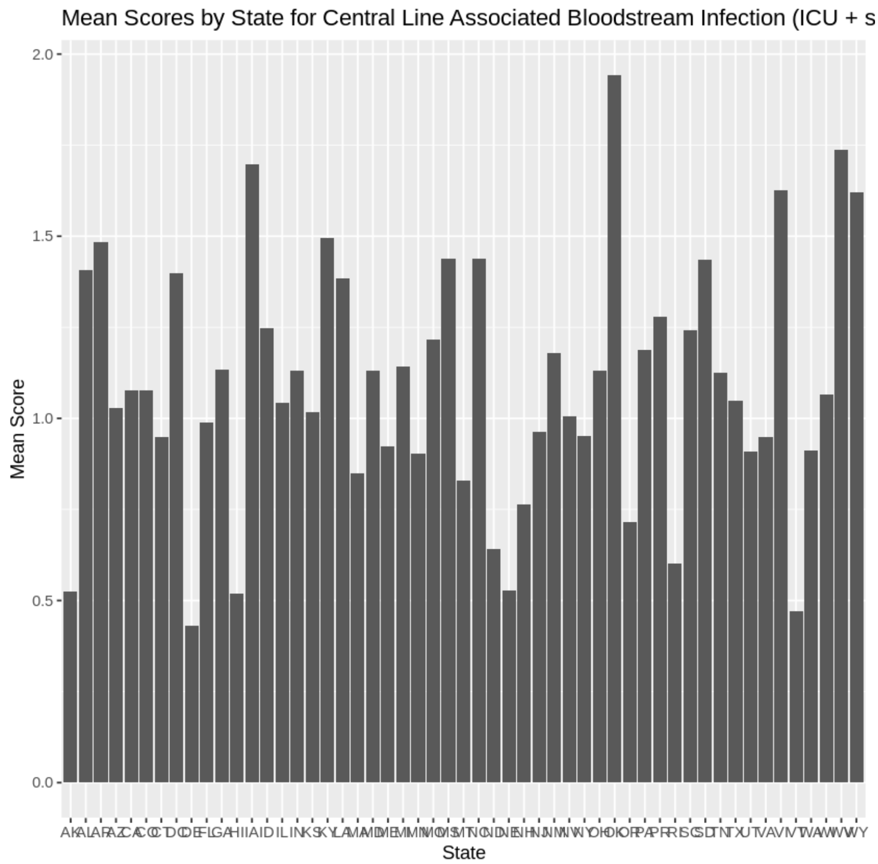

Fig. 6: Mean Score by State

Figure 7 shows the boxplots of SIR scores, grouped by state. Similar to the histogram of SIR scores, all of the states are skewed towards higher values. A lot of states have several outliers, but have similar means throughout.

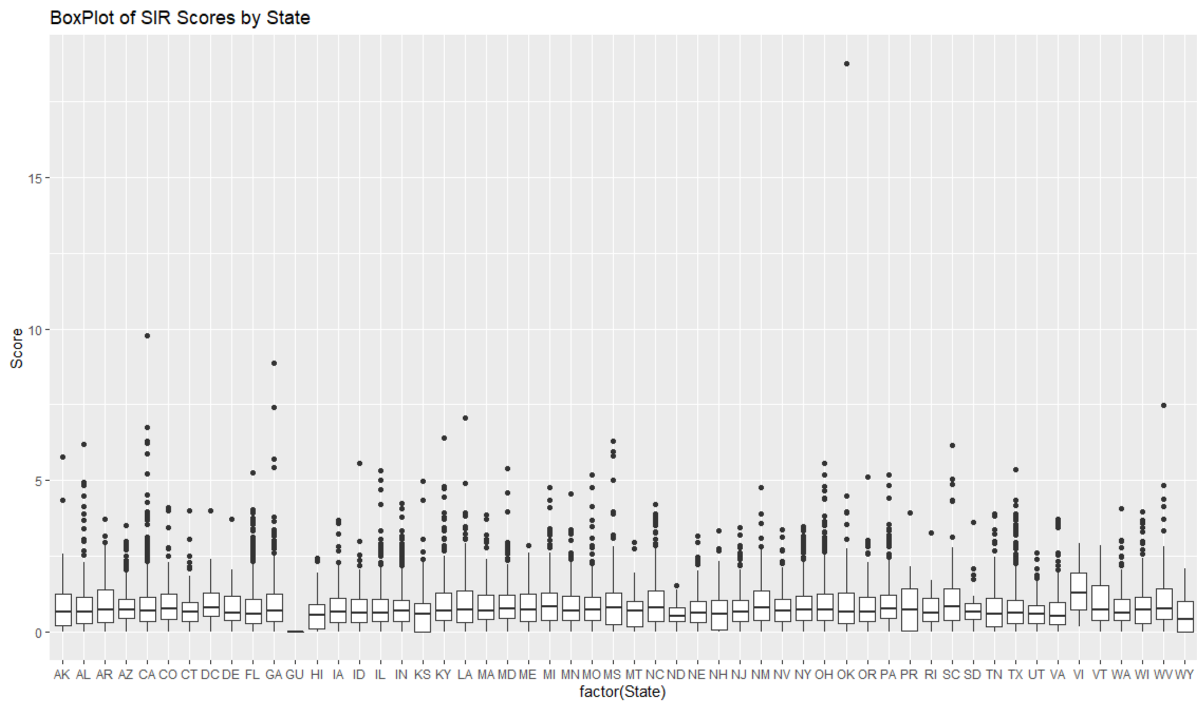

Fig. 7: Boxplot of SIR Scores by State

QQplots were made faceted by HAI score to determine the normality of the distributions. Since normal distribution is an assumption for different tests that we may like to use, such as ANOVA, we need to have an understanding of the normality of the scores.

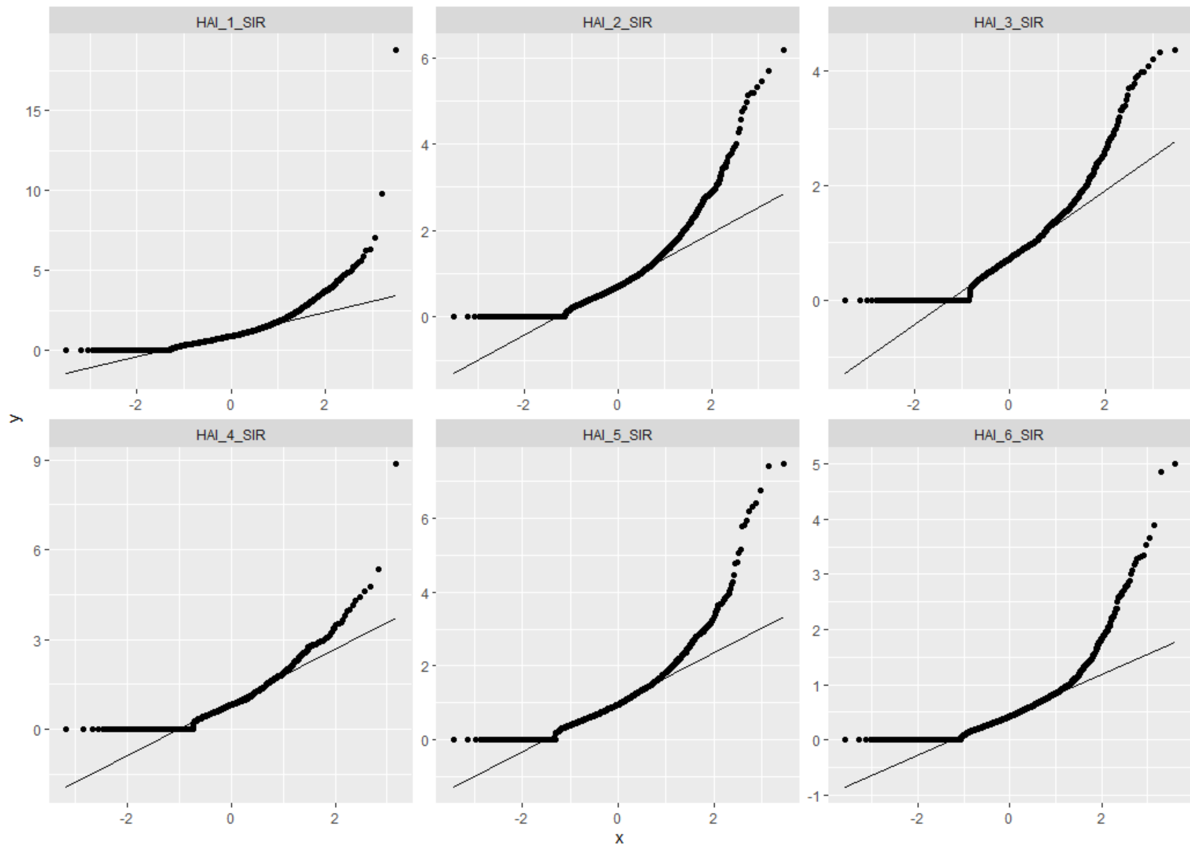

Fig. 8: QQ Plots of HAI Scores

The QQ plots show a lack of normality for each HAI score, supporting the previous interpretation of a lack of normality of the histograms of the HAI scores. This means that alternatives to tests that require normal distributions will be used. For example, instead of ANOVA tests, we will use Kruskal-Wallis.

Sources

[1] https://epi.dph.ncdhhs.gov/cd/hai/figures.html
[2] https://qualitynet.cms.gov/inpatient/measures/hai
[3] https://pubmed.ncbi.nlm.nih.gov/34774274/
[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5120566/
[5] https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/HAC-Reduction-Program
[6] https://data.cms.gov/provider-data/dataset/77hc-ibv8
[7] https://data.cms.gov/provider-data/dataset/xubh-q36u