Visualizing Factors Affecting Oregon Health Care Cost in 2015

Introduction:

The data within the healthcare industry is growing at a rapid pace. Each year the healthcare industry generates a large amount of data, driven by record keeping, compliance and regulatory requirements. This data holds the promise of supporting a wide range of medical and healthcare functions, including clinical decision support, disease surveillance, and clinical analytics.

The APAC Database: In 2009, aiming to measure the quality, quantity, and value of health care, the Oregon State Legislature established the Oregon All Payer All Claims (APAC) Database through a House Bill, "authorizing the formation of a healthcare data reporting program to measure the quality, quantity, and value of healthcare in Oregon". The two primary aims of the APAC is to provide a benchmark for the State to allow for improvement of the ***quality of care***, and the optimization of ***cost of care***.

In our project, we have tried to simulate the pathway an analyst would take to identify the pain points for the State's health care costs, and to find the underlying reasons.

Data Collection: For this project, we used the 2015 Episodes of Care (claims information submitted by the insurance companies for all insured patients in Oregon) Public Dataset (patient and insurance company names anonymized) that we requested from the Oregon Health Authority. We were kindly waived the $1,000 charge and provided with four 10GB files, for each of the four quarters of 2015.

Data Preparation: This was analyzed in MongoDB and then cleaned up using MySQL. Finally we extracted 1M rows of the 33M for each quarter, selecting randomly, and created a .csv file. All the visualization below is made using this reduced dataset and using Tableau. Therefore, any count value in the figures are multiplied by 33 to provide an actual count.

Explorative analysis:

Cost of care:

The journey starts with simple exploratory statistics to understand the trends and basic descriptive statistics to set the backdrop for any unusual finding. We started with the question:

Which disease types incur the most average cost per visit?

Fig 1 shows that Neonatal disorders, Injuries and iatrogenic diseases, Hematopoietic and Lymphatic diseases, Genetic disorders, and Gastrointestinal diseases are the five most expensive types of diseases based on average cost per visit. These categories all cost more than $300 per visit. However, this is biased by the fact that these diseases result in prolonged sickness and expensive medical and surgical interventions. So, we asked instead:

Which disease types incur the most expense annually?

Fig 2 not only identifies the major diseases, but also breaks them down by the system. It can be seen that the most expensive musculoskeletal disease is osteoarthritis- something we simply take for granted, accounting for almost a third. For Cardiovascular diseases, essential hypertension costs very little per patient, but a lot of people suffer from it- this inflates the proportional cost. Deliveries are most expensive for Female reproductive diseases, and multiple sclerosis is the most expensive neurological disorder. The autoimmune Crohn's disease is the most expensive gastrointestinal disorder, and asthma and depression are the leading respiratory and psychiatric disorders. However, the most important revelation is Diabetes mellitus being the most expensive of all diseases. Consequently, it makes sense for the

healthcare system to invest money and effort into its prevention through proper health education and regular physical check ups.

We further wanted to see the trend of these diseases, so we asked:

Is the cost incurred for these diseases increasing with time?

Fig 3 shows the quarter-by-quarter diseases like osteoarthritis, coronary artery disease, hypertension, asthma, depression and vaginal delivery increase- but they are again dwarfed by the increase in diabetes mellitus. It should be noted that diseases like asthma and osteoarthritis increase in Winter, and depression increases in Winter in Oregon specifically, however- since diabetes does not demonstrate seasonality- it is clear that the incidence is rising, showing the dire need for immediate preventive action.

Quality of care:

We now changed tack and looked at the other answer APAC data is expected to deliver- quality of health care delivery. APAC data does not include the result of a treatment- so we cannot tell if a visit to the provider led to the patient getting healthy or not. The closest metric we had was the length of stay, among others. A longer duration of stay usually depicts more serious illness or slower recovery, and actually increases morbidity. So, we asked the question:

Which diseases lead to the longest duration of stay?

Fig 4 shows that Neurological, Cardiovascular, Hematopoietic & Lymphatic, and Respiratory are the top 4 diseases required the longest average duration of stay, averaging 5.14 days, 4.34 days, 4.28 days, and 4.14 days respectively. We excluded 0 days (no stay), and this data needs looking into further- it is important to see the distribution of different lengths of stay- more cases of very long stays (as is common for neurological and hematological inpatient stays) can inflate the average compared to a disease where most diseases result in a longer stay and can benefit from medical intervention. So, we asked:

What is the breakdown of duration of stay for the diseases in Fig 4. ?

As we can see in Fig 5, the disease types with more preponderance of the longer duration are respiratory and gastrointestinal. Musculoskeletal, for example- has larger number of 1-5 days of stay. This was corroborated by Fig. 6, which shows the highest count for 6-10 days of stay are for respiratory and gastrointestinal diseases. Therefore, in the subsequent analysis we focused on Respiratory diseases. We asked: *Which respiratory diseases cause the longest stay?*

Figs 7 and 8 show the answer. While a straight bar chart in Fig 7 shows Respiratory Syncitial virus and Aspiration pneumonia (swallowing stomach content after surgery) to have the highest average duration of stay (0.25 and 0.20 days, respectively), Fig 8 shows the actual distribution to be heavily towards bacterial pneumonia. The other two diseases can have severe complications causing >1 m stay, which leads to a skewing of the distribution to the right. However, in general, it is the bacterial pneumonia that causes most length of stay and needs to be focused upon.

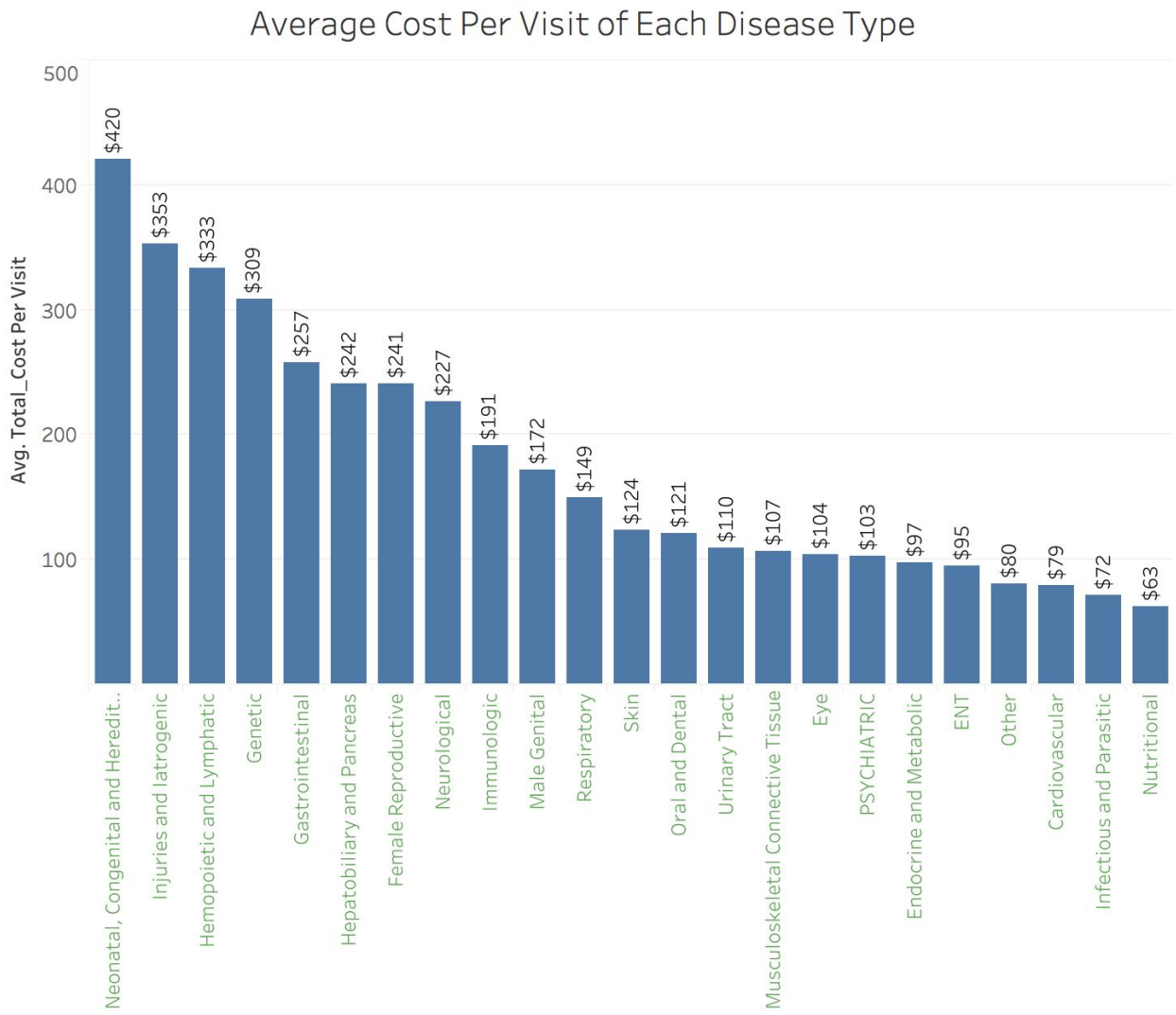


Figure 1 shows the *average total cost per visit* of each disease. For this figure, we broke down the dataset by the disease type and sorted them by average cost per visit, and visualized the ensuing bar chart.

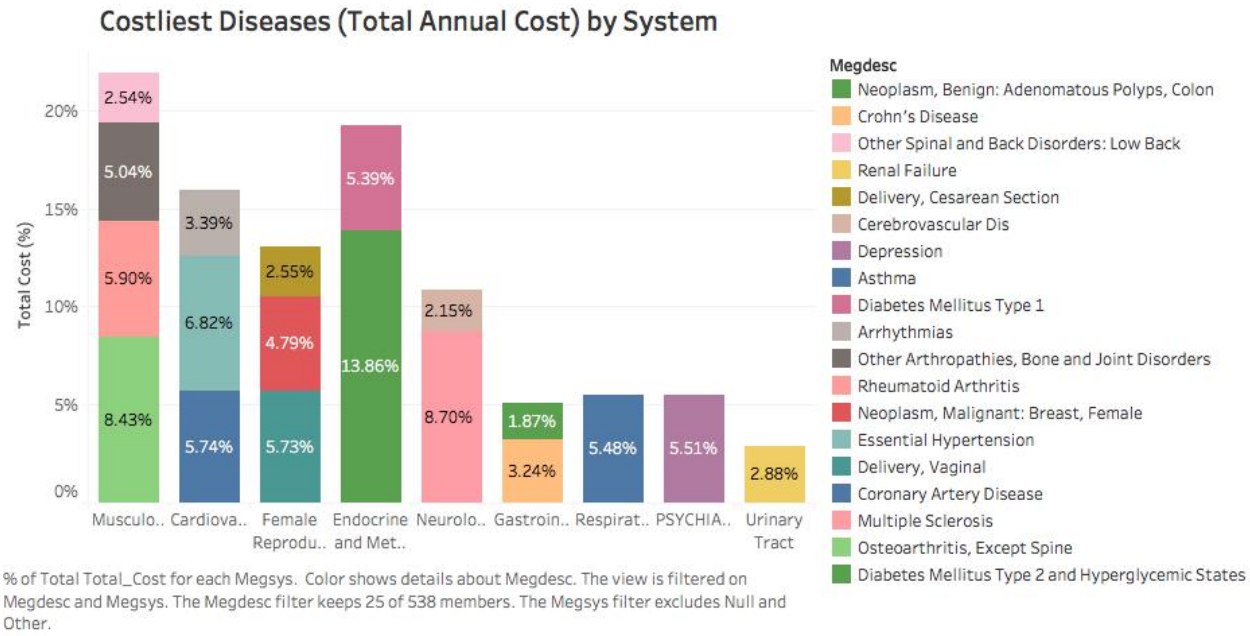
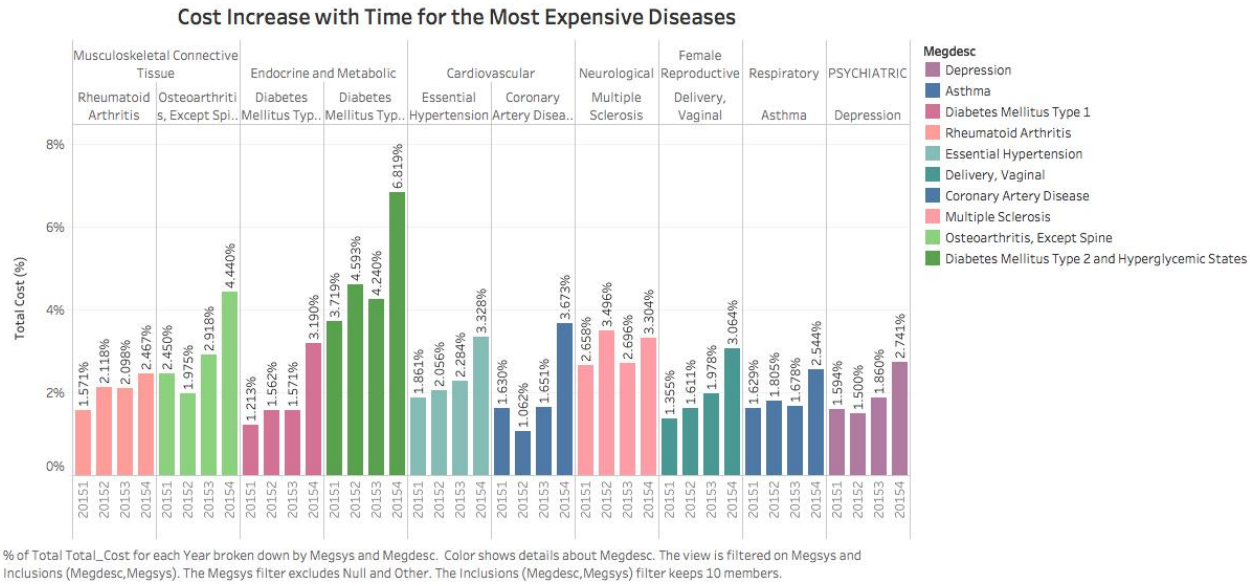


Figure 2 illustrates costliest diseases types by *total annual cost* incurred. We have broken down the dataset by the disease types (such as Musculoskeletal, Cardiovascular, etc.) and picked the ones with the maximum proportion in total medical cost. Next, we have broken each disease type by the diseases (such as osteoarthritis, diabetes mellitus). Finally, we have excluded any disease that cost less than $100M annually to narrow down to the most expensive diseases.



% of Total Total_Cost for each Year broken down by Megsys and Megdesc. Color shows details about Megdesc. The view is filtered on Megsys and Inclusions (Megdesc,Megsys). The Megsys filter excludes Null and Other. The Inclusions (Megdesc,Megsys) filter keeps 10 members.

Figure 3 shows the quarter-by-quarter expense incurred for the top 10 most expensive diseases extracted from figure 2. For this we filtered the previous figure for the top 10 diseases by total annual cost, and then broke down the expenses by the quarter. Using disease type for color allowed easier visualization of the bars, which otherwise would have been challenging to distinguish from one disease to another.

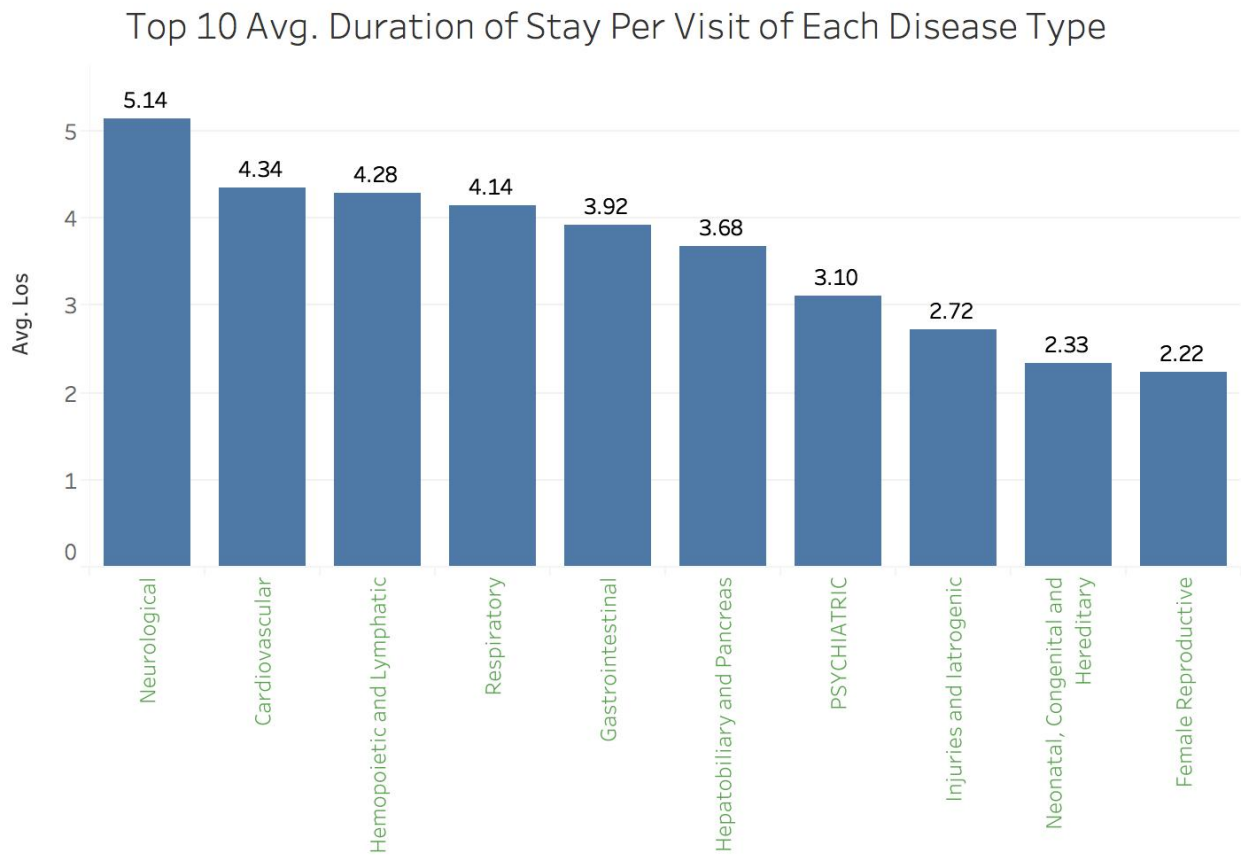
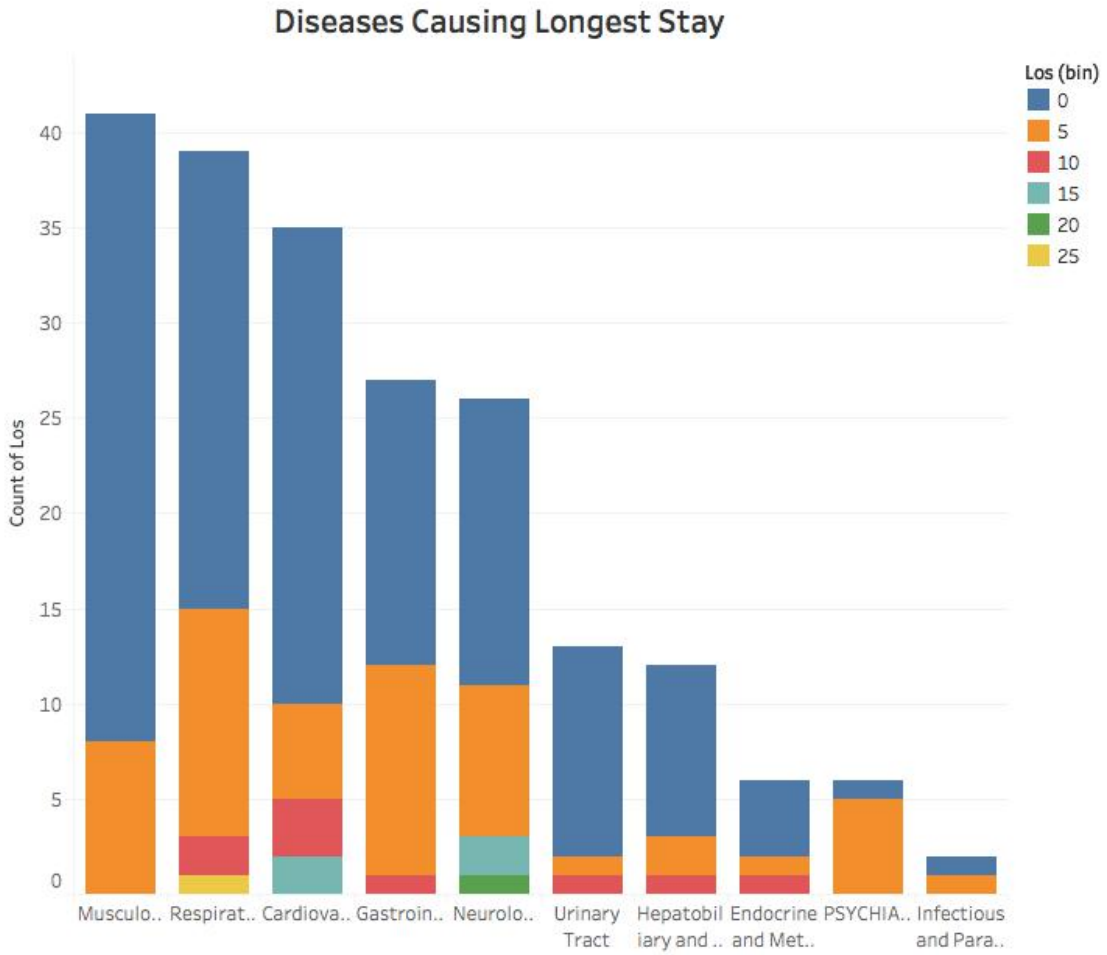
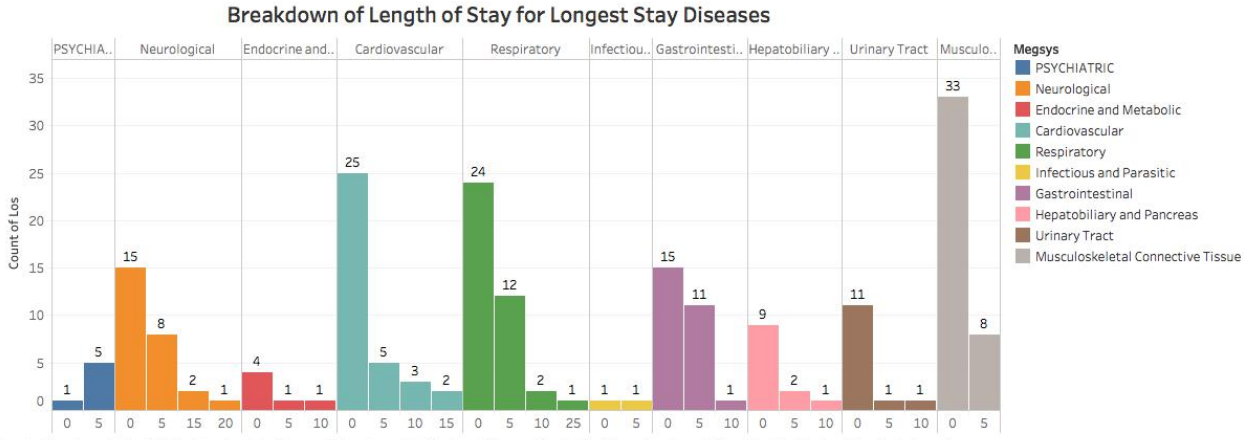


Figure 4 shows the top 10 diseases resulting in longest average length of stay. We filtered the dataset for 0 days (no stay) and only inpatient place of stay (such as inpatient hospital, hospice, inpatient psychiatric care, etc.). Then we sorted the data by average length of stay and selected the top 10.



Count of Los for each Megsys. Color shows details about Los (bin). The data is filtered on Los and Place Service. The Los filter includes values greater than or equal to 1. The Place Service filter keeps 10 of 41 members. The view is filtered on Megsys, which keeps 10 of 24 members.

Figure 5 shows the top 10 disease types (in terms of duration of stay) broken down in terms of count of days.



Count of Los for each Los (bin) broken down by Megsys. Color shows details about Megsys. The data is filtered on Los and Place Service. The Los filter includes values greater than or equal to 1. The Place Service filter keeps 10 of 41 members. The view is filtered on Megsys, which keeps 10 of 24 members.

Figure 6 shows the breakdown of longest stay diseases by bin sizes of 5. The data was filtered for >1 (no stay excluded) and color coded for the different bin sizes (1-5, 6-10, 11-15, etc.).

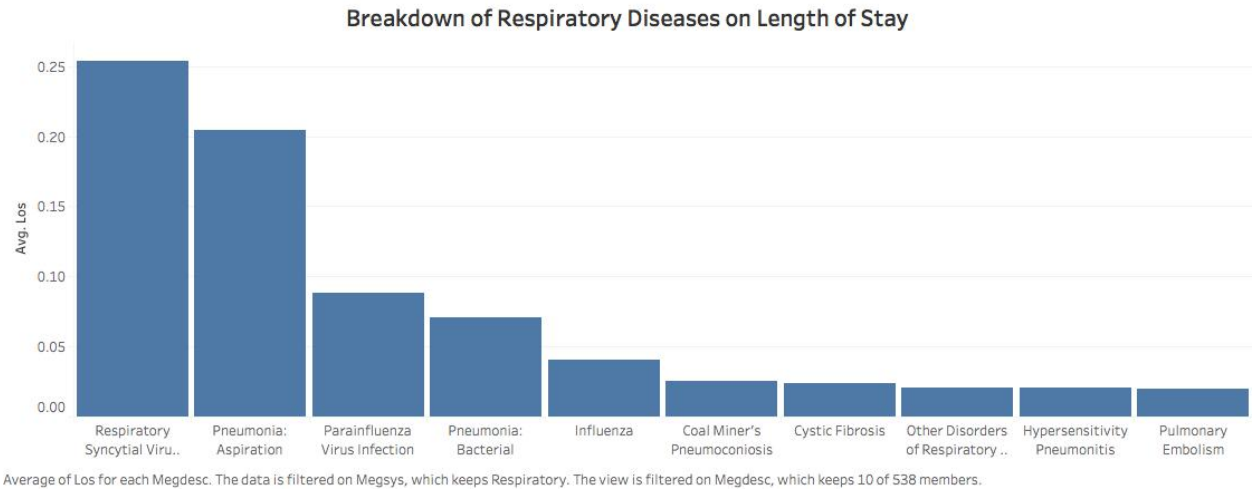


Figure 7 shows a breakdown of respiratory diseases by average duration of inpatient stay. Unusual diseases like RSV, Aspiration pneumonia and Parainfluenze lead the spectrum.

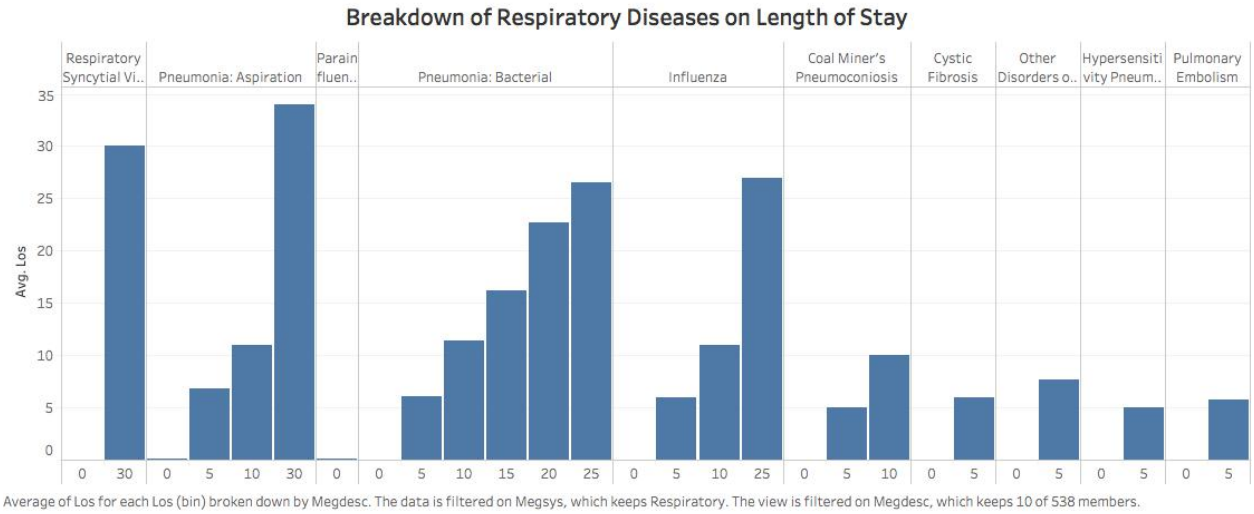


Figure 8 shows the underlying story. RSV and Aspiration pneumonia's average duration of stay is heavily skewed by the longer (>1m) duration of complicated cases. The average of parainfluenza is skewed by the very low number of very serious cases. Bacterial pneumonia is the one to look out for.