Domain Background

Mistakes in healthcare costing patients ‘lives and wellbeing have been reported and talked about on numerous occasions. The Institute of Medicine (IOM) report titled “To err is human” and "crossing the quality chasm" were landmarks report that triggered the alarm bell on medical errors and the need to revamp the way healthcare is being delivered to patients. It cited missteps in diagnosis, treatment, and prevention. It reported that between 50000 to 100000 people die each year from those errors.

Some of the recommendations were the implementation of electronic medical records and the substitution of written orders with computerized physicians' entry orders (CPOE) to avoid administration of the wrong dose or medicine due to poorly legible written orders.

I am personally motivated to discuss and explore this problem since I am a physician in Internal Medicine and I have been practicing at the bedside for more than 10 years.

Problem statement

Quality and effectiveness of care varies between different hospitals and this is reflected in the number of adverse events, rate of complications and mortality from certain conditions. It is suggested that some of those mishaps could be prevented by the digitization of medical records, creating a general surgery registry, completing a preoperative check list... Mishaps are things like increased rate of complications following surgeries such as a knee replacement, increased inpatients mortality from conditions like heart failure and pneumonia, increased incidence of inpatients acquired infections such as "Clostridium difficile diarrhea", line infections and surgical wound infections. It is believed that, to some extent, those problems arise from poor communications between physicians and nurses (poorly legible written orders), inadequate follow up on patient's lab results and imaging studies(lab report misplaced in the paper chart or placed in a chart belonging to a different patient, looking at imaging films requires a trip to the radiology suite to localize the film and look at it instead of being able to instantly download the image on the computer screen and look at it).

We can break up the hospitals group into 3 sets and use a chi square metric to see whether hospitals that have better infrastructure such as electronic medical records or general surgery registry fare better in terms of patient care and better outcomes.

Datasets and Inputs

Medicare compiled a set of different data tables on data.medicare.gov, I used a few data tables from the hospitals section. These are the following files that I downloaded as a csv format on my computer and then used pandas.read_csv to analyze them in a Jupyter notebook:

"Hospital General Information", "Complications and Deaths Hospital", "Healthcare Associated Infections Hospital".

I arranged the data in a data frame where each row corresponds to the identity of one hospital and the columns correspond to different outcomes and complications that arise in the hospital. I then used clustering to create three clusters (logically speaking those clusters should correspond to hospitals with good care, intermediate care and bad care).

"Structural_Measures_Hospital" is another file I used. It contains the variables that reflects how well each hospital is prepared in terms of Electronic medical records, general surgery registry and preoperative check list.

Solution Statement

I will look at the different clusters of performance and see if there is an association between performance on one hand (performance here refers to rate of complications and outcome of different medical conditions), and the ability to receive and track labs electronically and the presence of a general surgery registry, on the other hand.

If a significant association is found, it means that the use of electronic records and the creation of a general surgery registry (which implies that the hospital has an advanced informatics infrastructure to support the creation and management of a general surgery registry, we might infer from that, that this infrastructure might as well be used for other purposes related to storing, handling, analyzing and learning from health data) lead to a better outcome in hospitals.

Benchmark Model

There is no easy benchmark model to find for comparison. But it's been reported that certain rate of post op complications prior to using pre op check list and general surgery registry amounted to 6% . In my analysis, I will check if the current rate is less especially when we factor in the presence or absence of a general surgery registry.

Evaluations Metrics

I will use the Chi square test for independence to investigate the presence of such an association as mentioned in the Solution Statement or not. Chis square is used since after breaking the dataset into "best performance cluster" and "other performance clusters", the variables will be categorical.

The variable related to the electronical infrastructure and general surgery registry of the different hospitals will also be categorical.

Project Design

The datasets mentioned above will be first ridden of the outliers, then I will use most of the numerical variables related to complications in the hospital such as post-operative complications in the wounds, mortality and morbidities of a certain conditions (like Heart Failure and Pneumonia) that patients are admitted with, rate of infections acquired in the hospital... Those features will amount to a total of 16 features. Each row will be a single ID corresponding to the hospital ID in the dataset.

Once, I assemble all of that data from three files ("Hospital General Information", "Complications and Deaths Hospital", "Healthcare Associated Infections Hospital") together in one big set, I will preprocess it with a MinMaxScaler from sklearn library and then use KMeans() to break it down into 3 clusters. Why three?

I think it will be logical to separate these hospitals into best, intermediate and poor performers. Since there are 16 variables, increasing the number of clusters to more than three would be make it difficult to discern the pattern on a bar chart where all the variables from the different clusters are stacked side by side. After, identifying on the bar chart which cluster performed the best, I group the IDs of that "best performance" into the sample of best performance, and the rest will be grouped in the sample of "other performance".

Now regarding "Structural_Measures_Hospital" file, the data from that file will be arranged in such a way that each row corresponds to a hospital and the columns correspond to variables related to "Use of EMR", "presence of general surgery registry"… then this file is sampled into two group based on comprehensive use of EMRs, general surgery registry on one side and non-comprehensive use of EMRs and general surgery registry on the other side.

Then I will run a chi square test of independence between Performance on one side and Use of EMRs and or general surgery registry on the other side.