Case 2 Report

Yurong Li

- ## Introduction

The study is based on the synthetic dataset comprised of Electronic Health Records (EHR). There are 16 sub datasets, which record the medical history of 1171 patients, such as: demographic data, medications, diagnoses, care plans, lab results, etc.

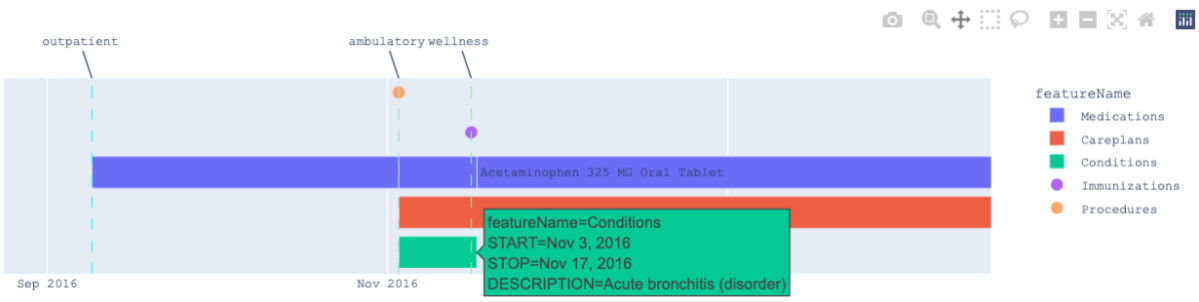
- ## Visualize Single Patient

Timeline is important to the patient's medical history. From the timeline, the patients' visiting records with corresponding conditions, care plans, lab observations, etc. can be easily collected and read. In this case, I utilize two approaches to visualize the single patient's medical history. The first and simplest way is to print the medical logs, e.g.

```
Christal240 Brown30
===========================
Race: white
Ethnicity: nonhispanic
Gender: F
Birth Date: 1982-09-01
Marital Status: S
Address: 1060 Hansen Overpass Suite 86, 2118 Boston, Massachusetts, Suffolk County
================================================================================================
Allergies: Latex allergy, Shellfish allergy
================================================================================================
Encounter:
1982-10-25: Encounter for problem  (class: ambulatory)
Medications:
1982-10-25 -- : diphenhydrAMINE Hydrochloride 25 MG Oral Tablet
Care Plans:
1982-10-25 -- : Self-care interventions (procedure)
------------------------------------------------------------------------------------------------
Encounter:
2000-06-14: Encounter for problem  (class: ambulatory)
Medications:
2000-06-14 -- : ferrous sulfate 325 MG Oral Tablet
------------------------------------------------------------------------------------------------
Encounter:
2010-03-27: Consultation for treatment  (class: outpatient)
Medications:
2010-03-27 -- 2011-03-22: Etonogestrel 68 MG Drug Implant
------------------------------------------------------------------------------------------------
```

From the log, we can read the conditions, lab observations and therapeutic schedules (e.g., medications, care plans, etc.) per patient per hospital visiting.

The second way is to plot the patient's medical history on a timeline. In order to plot timeline in Python, I called the *timeline()* function in a visualization library *plotly*. After pre-processing on raw datasets, an example of a part of patient's medical history can be shown as,



The plot can be dragged and zoomed up to the interesting periods by using the buttons on the top right corner, so the condition and therapeutic schedule information can be fastly checked and read from the timeline plot.

- **Explore the data to find and present patterns of patients with the same conditions such as:**

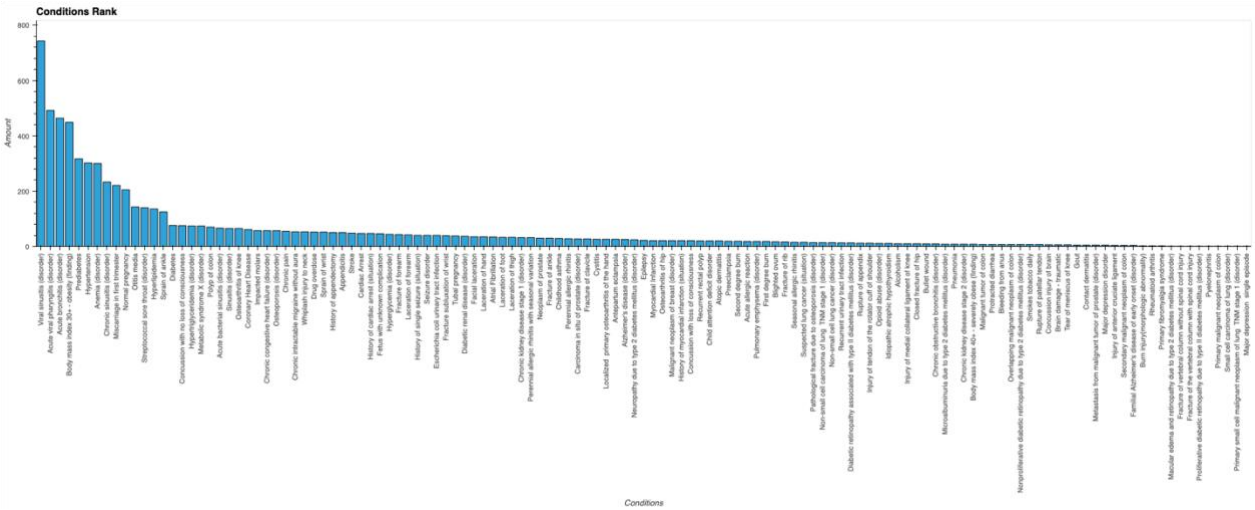
  - Which are the three most common conditions (present graphs and numbers)?

  Conditions dataset is used to get the three most common conditions. In order to get the total number for each condition, the patient with same condition but on different encounters should be seen as one case in the corresponding condition. After data pre-processing, we can get the three most common conditions and numbers are:

```
condition_rank.head(3)
```

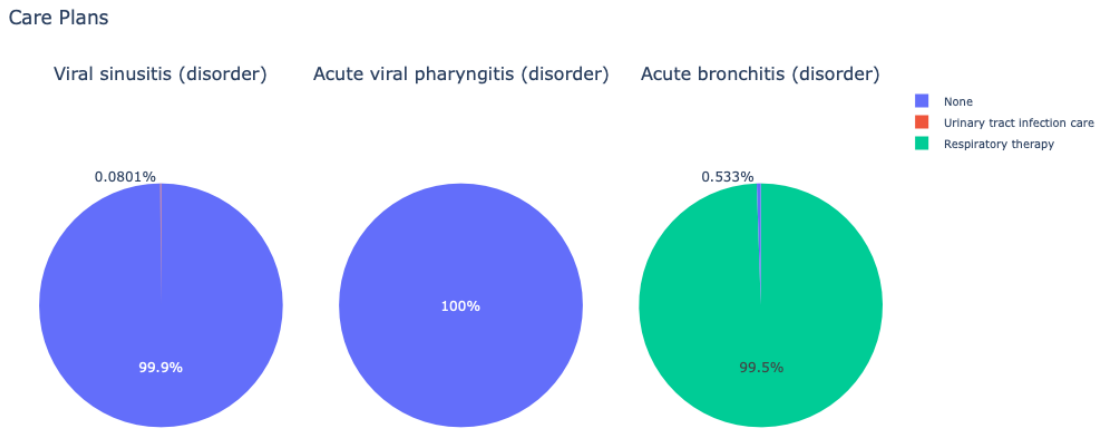
	Conditions	Freq
0	Viral sinusitis (disorder)	743
1	Acute viral pharyngitis (disorder)	492
2	Acute bronchitis (disorder)	464

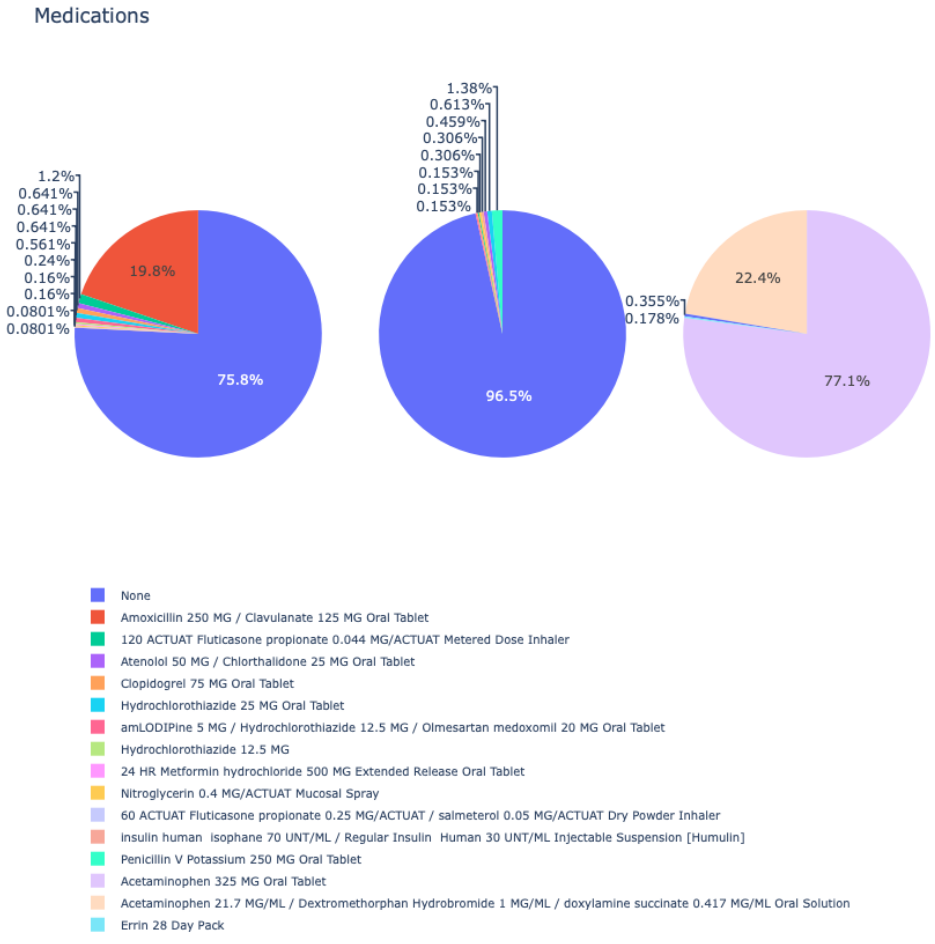
  And all of conditions and numbers can be visualized on the bar plot:



- Are there similarities in how the three conditions are treated? Showcase examples.

The care plans and medications of top 3 conditions can be plotted in the pie charts:

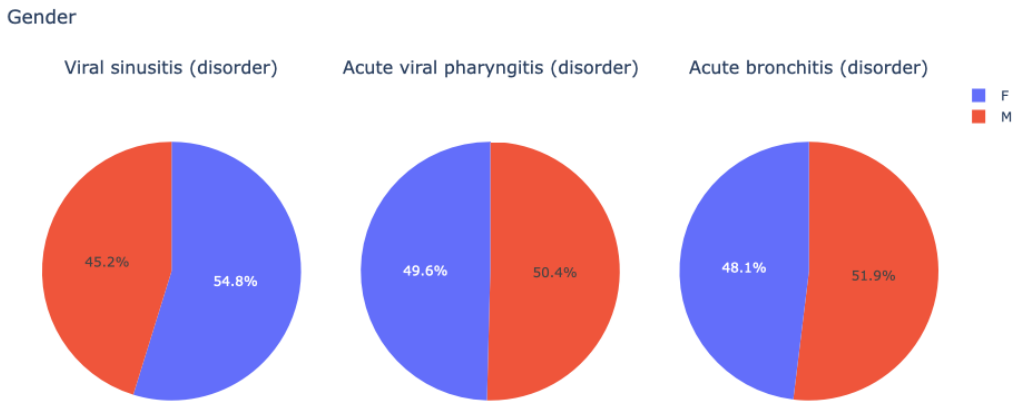


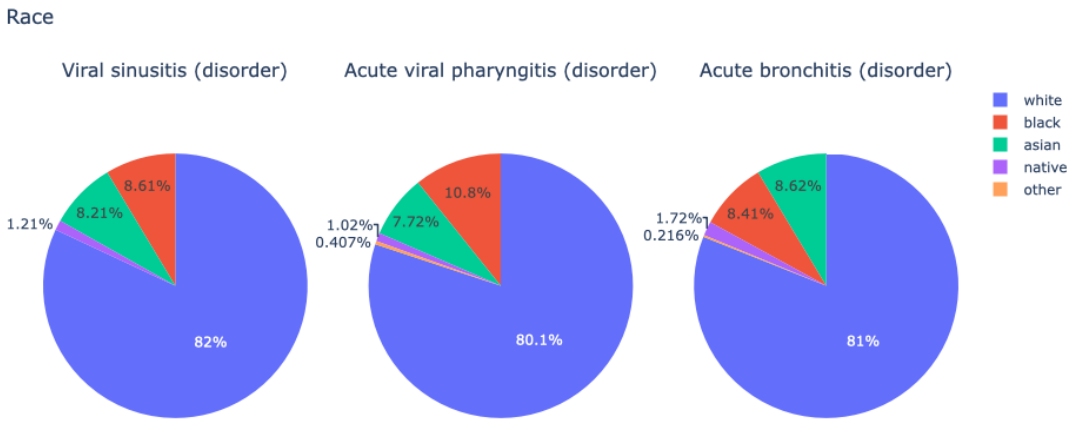


By observing the pie charts, the similarities in medication can be found in treating viral sinusitis and acute viral pharyngitis, e.g., 120 ACTUAT Fluticasone 0.044 MG/ ACTUAT Metered Dose Inhaler, Atenolol 50MG / Chlorthalidon 25 MG Oral Tablet, Clopidogrel 75 MG Oral Tablet, Hydrochlorothiazide 25 MG Oral Tablet, amlodipine 5MG / Hydrochlorothiazide 12.5 MG/ Olmesartan Medoxomil 20 MG Oral Tablet and Nitroglycerin 0.4 MG/ACTUAT mucosal Spray are the medications that in treating both viral sinusitis and acute viral pharyngitis.

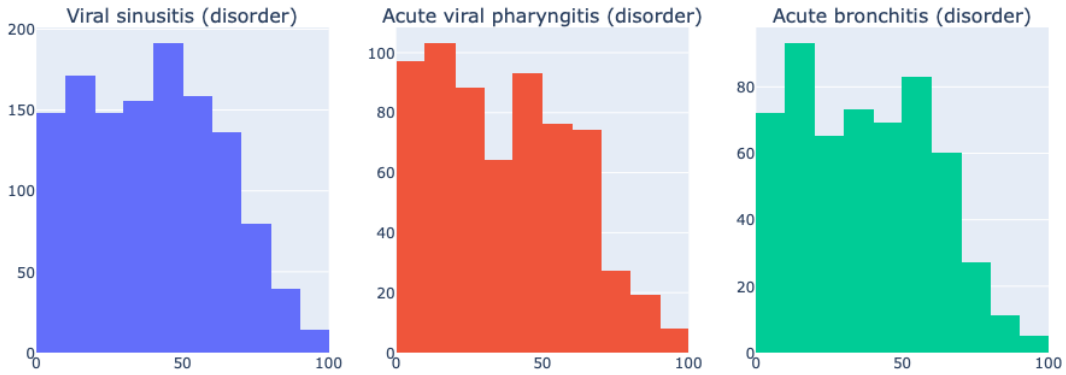
- What other common pattern characteristics can be found for the three groups of conditions?

In order to summarize the common pattern characteristics of the top three conditions, the demographic data are used to this task. The demographic data includes gender, race, marital status, age (which is the age of patient under corresponding condition), etc and they are also shown as pie charts,

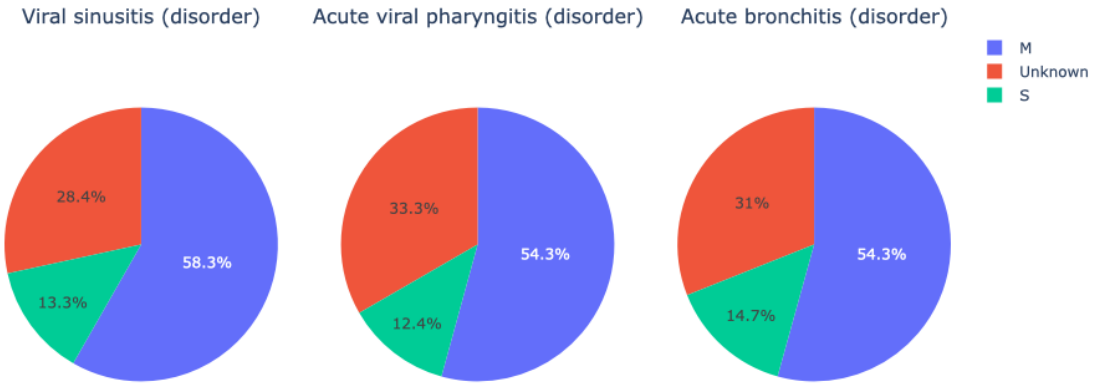




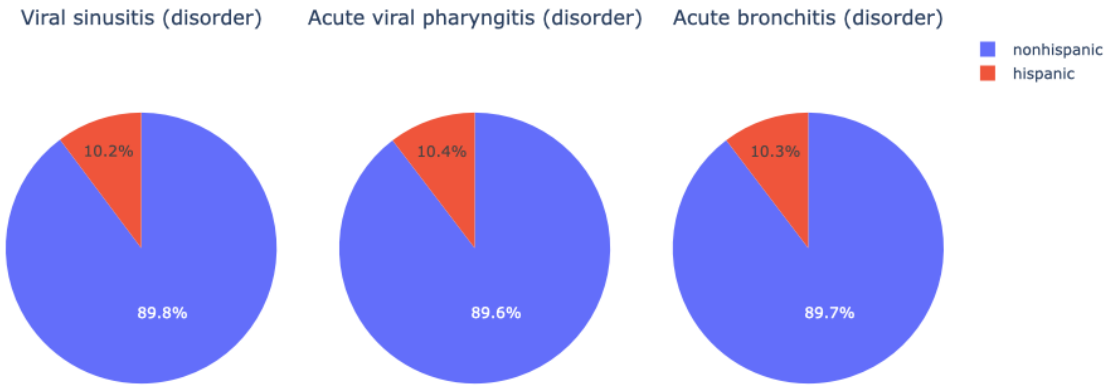
Age



Marital Status



Ethnicity



From the pie charts, teenagers in 10 -20 and people in 40 - 60 have the high risk in these three diseases. The similar patients' pattens can be also observed for race and ethnicity, but the resource of the patients should be considered, because the data are collected from the organizations located in Europe.

- **Formulate three other questions that could be interesting from a machine learning perspective using this data (that could potentially be used in a clinical setting to improve care)**

Based on the dataset in this case, three ML problems can be formulated:

- Question 1: Care-utility prediction, which aims to predict how many patients need to use a care/service in coming month.
- Question 2: Mortality prediction, which use classifiers to predict if a patient is dead or not based on engineering features.
- Question 3: Influencing factors of death, which aims to investigate the potential factors are causing death by using the explainable models based on the result of question 2.

For question 1, I choose one most common care plans: Respiratory therapy, which has the largest number. I sum up the total usage of this care plan for each month and I select the data from Jan 2005 to Apr 2020 based on the data quality. In order to predict the amount of this care utility in coming month, I use the six months' data before the coming month, so there are 184 samples in the amount dataset. I split the last year's data as testing and the rest are training. Since it is the time-series prediction, the LSTM model is selected in solving this problem. The result can be seen in the plot:

We can observe that the LSTM cannot reach a precise amount prediction but can catch the trend of care amount.

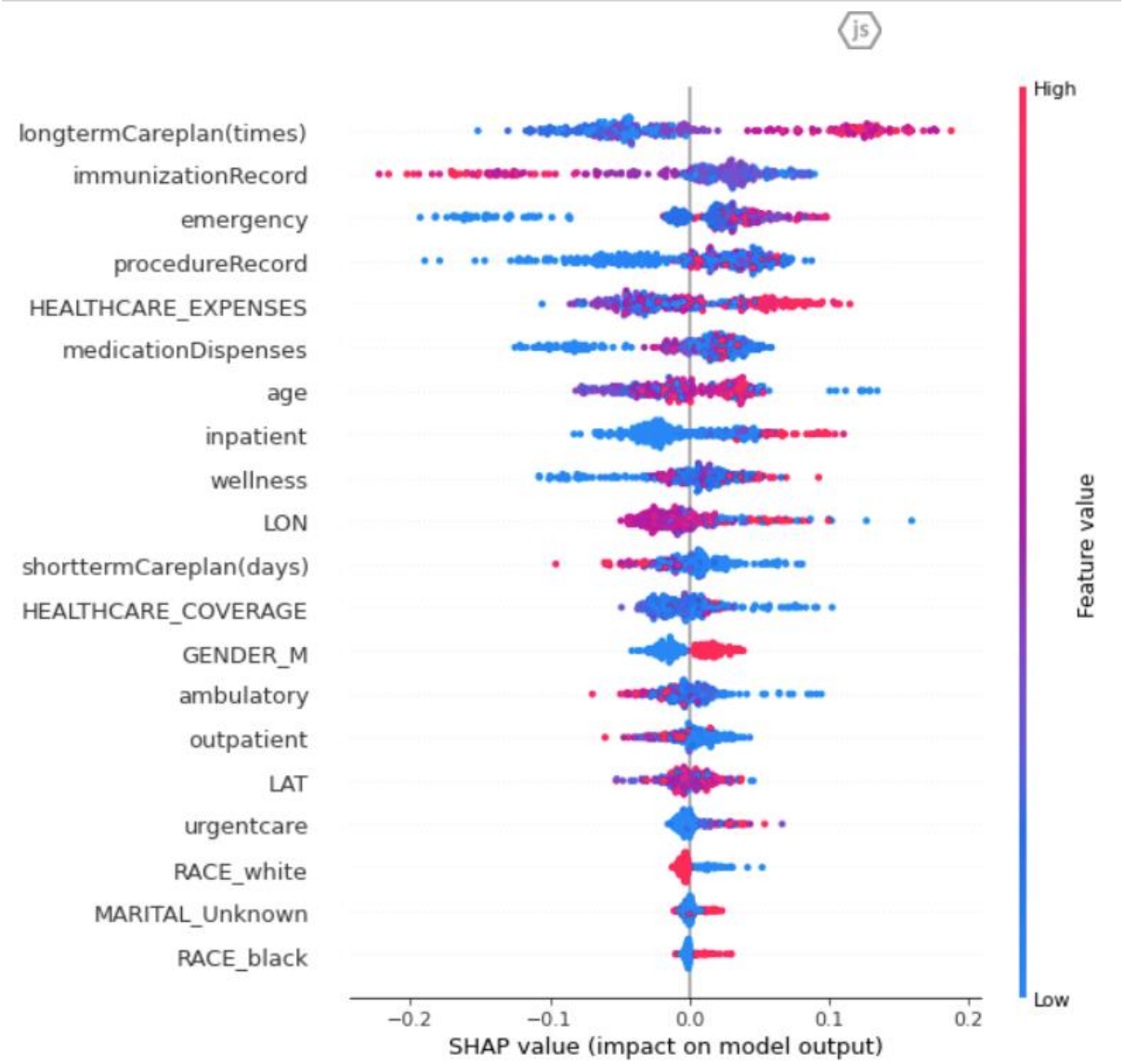
For question 2, I use the 171 patients with death date and randomly select same numbers of patients without death date in order to train unbiased classifiers. I used the demographic features of patients. In addition, some engineered features are also used. In this study, I generate following features: the total medication dispenses, the number of immunization and procedure records, the number of six encounter classes records, the amount of long-term care and the duration of short-term care. Since it is the classification problem, two classifiers are selected, which are random forest and artificial

neural network, and the random guess (should be around 0.5) are calculated as a baseline. The comparison of prediction accuracy from two classifiers can be observed by,

```
print("RF Accuracy:", metrics.accuracy_score(y_test, y_pred))
print("NN Accuracy:", test_acc)
print("RG Accuracy:", metrics.accuracy_score(y_test, y_guess))
```

```
RF Accuracy: 0.7428571428571429
NN Accuracy: 0.7142857313156128
RG Accuracy: 0.4857142857142857
```

For question 3, I apply the SHAP model on the result of random forest from question 3,



From the SHAP result, the number of long-term cares, immunization, emergency encounter and procedure as well as the healthcare expenses are top 5 potential indicators of death. To be more specific, the patient with high number of long-term cares, emergency encounter, procedure and healthcare expenses but with low number of procedures have the high risk of death.

- ## Future Work

  The EHR dataset include various information for healthcare, and numerous ML problems can also be formulated and solved based on this dataset, e.g. patient re-admission prediction, which aims to predict if a patient will visit a hospital in next 6 months, hospital-utility prediction, which aims to predict how many patients will visit a hospital, etc.