Group 3: Xuhui Bai, Jiajie Yuan, Ying Tung Lau, Yunyi Wang, Zoey Yang

# Healthcare Data and Analytics

## Case study of AI-Opening

## Introduction

In 2020, this special year, people around the world are responding to a pandemic of coronavirus disease 2019 (COVID-19) caused by a novel coronavirus, SARSCoV-2, that is spreading from person to person. The question everyone in the world wants answered is how far the new coronavirus will spread and when the pandemic will begin to ebb. To know that, epidemiologists, public health authorities and policymakers rely on models to predict and make decisions. Compared to influenza, modeling the current COVID-19 outbreak is much more challenging, simply because researchers know very little about the disease and have little history record to rely on. Therefore, scientists can only rely on records that we currently have to build the model. However, it is a little bit challenging to only rely on constantly changing variables to return an accurate prediction, since everything might change overnight due to policymakers.

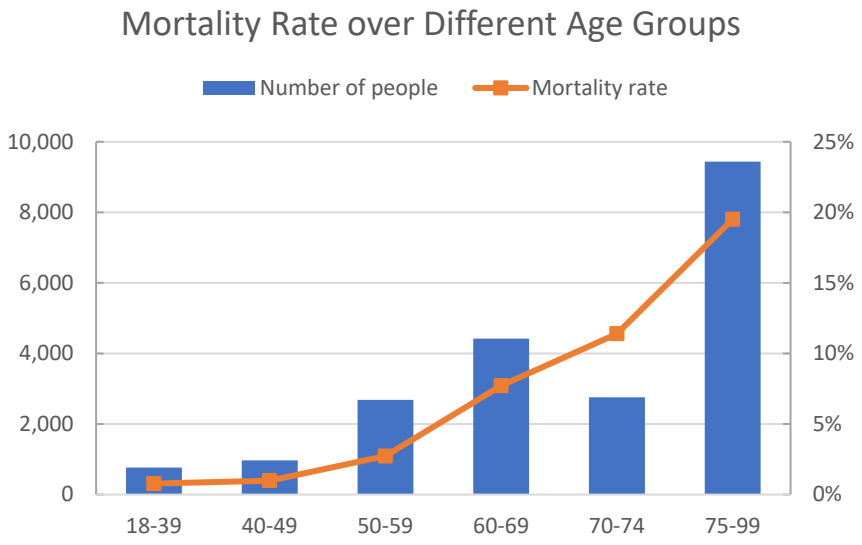
As the White House has decided to reopen the market at the beginning of May in sake of the economy, models project a sharp rise in deaths as states reopen. At this time, forecasting COVID-19 Deaths in the US is critical. Forecasts of deaths will help inform public health decision-making by projecting the likely impact in coming weeks. More importantly, we can identify the high-risk population based on features by analyzing the COVID-19 data from the last several months, and suggest policymakers and the federal government offer some level of paid leave benefits to the high risk population to stay at home with less financial burden. Furthermore, if a vaccine is made available, we can suggest prioritizing the identified at-risk

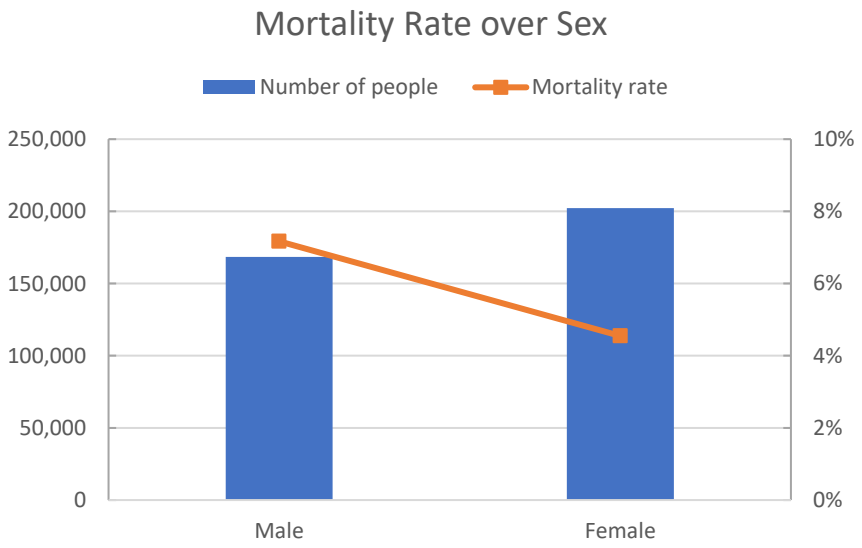
population for early reception of the vaccine. In this project, we would show you the whole process of building a best-performed prediction model on COVID-19 mortality rate based on COVID-19 dataset, and share our insights with you about the high-risk population identified by our team based on the analysis of our model.

# I. Data Procession & Model Development

## Statistical Analysis

Before training our models, we made relevant statistical analysis of our overall data. Age and gender are the two factors we consider first, and we guess that age plays a large role in mortality. In order to verify the rationality of our guess, we counted the number of people in each age group and the corresponding mortality rate and drew the following plot. This plot is a combination of a bar chart and a line chart. The bar chart shows the number of people in each age group, and the line chart shows the mortality rate of this age group. We can clearly notice from the graph that as age increases, mortality rate also shows an increase. And the age of the highest mortality rate is 75 to 99. From this we can conclude that the older you are, the greater the probability of death after infection.






Through gender analysis, we found that although our sample contains more data on females, the mortality rate of males is about 7%, while that of females is only about 4%. The mortality rate of females  is lower than that of males. From this we can conclude that men are more likely to die after infection than women.

## Data Processing

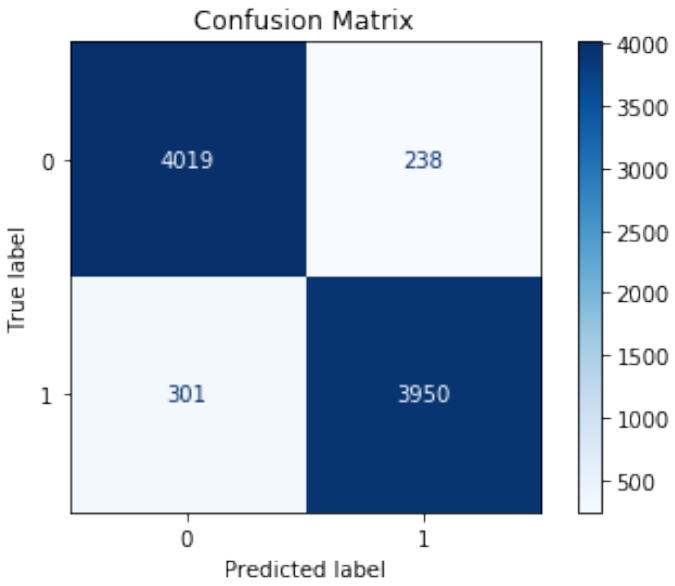
After mapping the DX1-DX20 to DGL with the taxonomy tree, we first checked the distribution of the target variable, 'mortality'. As most machine learning algorithms are sensitive to unbalanced data, that is, unequally distributed dataset (such as 10 death verses 90 live) will bias the prediction model towards the more common class. The problem also exists in our dataset. To fix the unbalanced dataset, we used an under-sampling method to match the number of each class: randomly collect samples of class 0 types matching the number of class 1 types. In total 42540 data points were selected for model development, half of the points were class 0 and half were class 1.

## Model Development

Five algorithms were initially developed and compared: Logistic Regression, Linear Support Vector Machine, Decision Tree Classifier, Random Forest Classifier, and Gradient Boosting Classifier. For each algorithm we first split the whole dataset into training and testing sets and selected the features that have significant impact on mortality by running models on training data sets. Then we used grid search and cross validation methods to select the optimal parameter set that can achieve the highest score on the test data set for each model. The performance of each model has been presented below:

| Model | Score on Testing Set | Score on Training Set |
|---|---|---|
| **Logistic Regression** | 0.940 | 0.941 |
| **Linear SVC** | 0.939 | 0.937 |
| **Gradient Boosting Classifier** | 0.930 | 0.928 |
| **Random Forest Classifier** | 0.910 | 0.910 |
| **Decision Tree Classifier** | 0.868 | 0.874 |

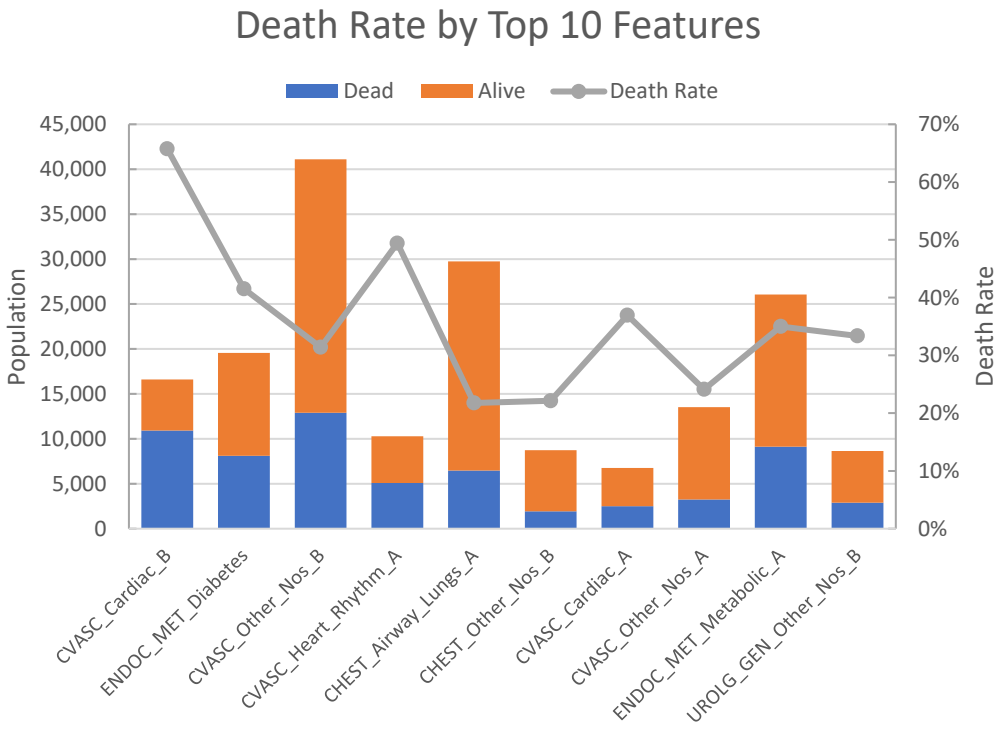
Based on the overall performance and the confusion matrix score, we selected Logistic Regression as our main algorithm for risk analysis on COVID-19. The confusion matrix generated from test dataset is shown below:




| | |
|---|---|
| **Accuracy** | 0.937 |
| **Recall** | 0.929 |
| **Precision** | 0.943 |
| **F1-Score** | 0.936 |

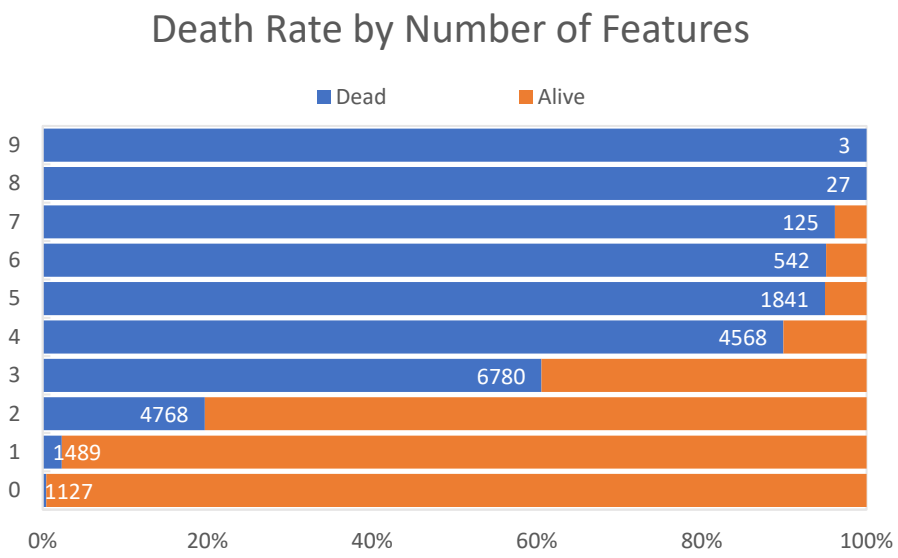
After model development, 160 features were selected as the risk factor leading to high mortality rate of COVID-19. In order to better understand the risk factors, we generated a word cloud of the DGL category. We first visualized the initial capital letters of each DGL code: some features share the same initial capital letters and can be regarded as in the same category. The weight of each word was the averaged coefficient for features within the category. The features that have a large impact on mortality rate have been visualized with word clouds.


As we observed in the plot, there are some categories that have a significant impact on the mortality rate: Age, ENDOC, CVASC, CHEST. Then we will dig into the categories to find the health conditions that may make COVID-19 more fatal.




According to the graph above it can be observed that the patients who have certain pre-existing diseases are suffering from higher mortality in these pandemics than other people. The latest

overall death rate is 7% globally[1], while for the patients who have certain underlying diseases the figure can rise to 20%-70%. These patients are the so-called High-Risk population in the pandemics.




We researched the impact of features selected by the model on the whole population, and the plot presented above summarizes the death rate for people having part or all diseases emphasized by our model. There is a general trend: the more diseases the patient has, the higher mortality rate is. In the data set there are 3 people who have all the 9 diseases indicated in our model, and all of them passed away due to the infection of COVID-19. On the other hand, for those who don't have any diseases listed by our model, only a very small proportion of them died during the pandemic. Therefore, it is reasonable to infer that multiple diseases will compound the risk of COVID-19 for patients.

[1] Max Roser et al., *Mortality Risk of COVID-19* (Our World in Data, 2020) https://ourworldindata.org/mortality-risk-covid#country-by-country-data-on-mortality-risk-of-the-covid-19-pandemic

## Odds Ratio

Age: Our results show an overall increase in the odds of fatality of covid-19 patients as one ages. Our referent category is age under 18 or unknown. If the patient is between 70 and 74, the odds of death increase with an odds ratio of 1.296, which means an increase of 29.6% in the odds compared to the referent category. If the patient is 75 or older, the odds of death increase significantly with an odds ratio of 2.639. Therefore, elderly individuals (>70 y) are at greater risk of death from covid-19. The explanation could be that the population is more likely to suffer from medical problems like lung disease, heart disease, and kidney disease.

| | | **Coefficient** | **Odds Ratio** |
|---|---|---|---|
| **Age** | <18 or Unknown (ref. cat.) | 0.000 | |
| | 18-39 | -0.374 | 0.688 |
| | 40-49 | -0.304 | 0.738 |
| | 50-59 | -0.223 | 0.800 |
| | 60-69 | -0.004 | 0.997 |
| | 70-74 | 0.260 | 1.296 |
| | 75-99 | 0.970 | 2.639 |

Gender: Our results show higher odds of fatality of male covid-19 patients compared to female. Specifically, with females being the referent category, the odds ratio calculated for male is 1.461, meaning an increase of 46.1% in the odds of death. A greater fatality rate of male patients could be due to biological and other innate differences between sexes, such as genes, immune system, and behavior.

| | | **Coefficient** | **Odds Ratio** |
|---|---|---|---|
| **Gender** | Female (ref. cat.) | 0.000 | |
| | Male | 0.379 | 1.461 |

Clinical Chapters: Our results show that covid-19 patients diagnosed with diseases in certain clinical chapters have markedly higher odds of fatality. For instance, the calculation shows that the odds ratio for CVASC_Cardiac_B is 8.756, meaning that there are about eight times higher odds of fatality for patients who are diagnosed with any disease in the chapter than the other patients. The top five clinical chapters ranked by odds are CVASC_Cardiac_B, CVASC_Other_Nos_B, CVASC_Other_Nos_B, CVASC_Heart_Rhythm_A, and CHEST_Airway_Lungs_A, which are mostly associated with heart, lung, or vascular diseases.

| | | **Coefficient** | **Odds Ratio** |
|---|---|---|---|
| | CVASC_Cardiac_B | 2.170 | 8.756 |
| | CVASC_Other_Nos_B | 1.444 | 4.240 |
| **DGL** | ENDOC_MET_Diabetes | 1.427 | 4.164 |
| | CVASC_Heart_Rhythm_A | 1.192 | 3.294 |
| | CHEST_Airway_Lungs_A | 1.027 | 2.792 |

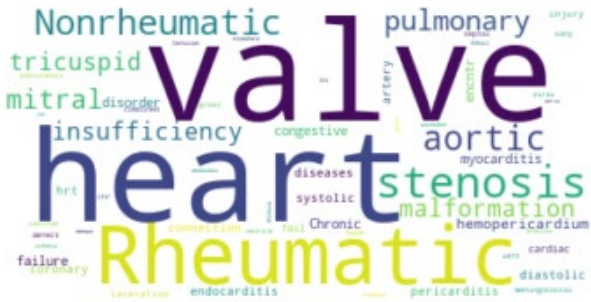
# II. Feature Analyses

In previous analyses, we found that Age, ENDOC, CVASC, and CHEST are the five main categories which could be the determinants leading to a higher risk of mortality. While previous points started from looking at effecting factors separately, in the following part, we will conduct horizontal joint analysis, meaning that considering several factors, like a subset of Age, ENDOC, and Sex, simultaneously to dig into their correlations or other patterns. In addition, we may explore the symptoms themselves presented by the specific category to see why this category will rank first among the top-list of feature importance. Specifically, in each main category, we will first look into the mortality of each sub-category. Then we will conduct joint analysis of the most significant subgroup selected by the model.

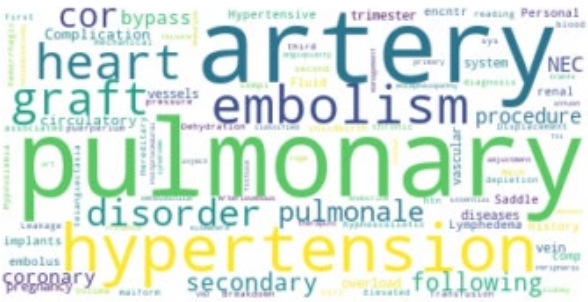
## CVASC Analysis

In the top 10 risk factors there are 5 factors belonging to the CVASC category. The word clouds for each factor are presented here:

**CVASC_Cardiac_B**



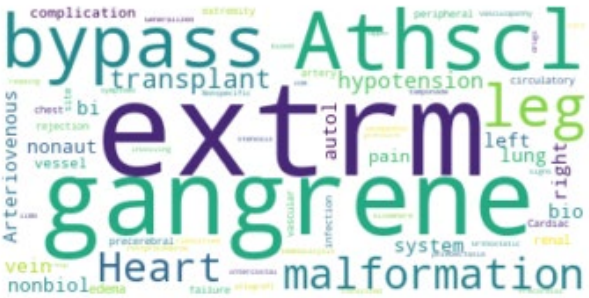

**CVASC_Other_Nos_B**




**CVASC_Heart_Rhythm_A**


**CVASC_Cardiac_A**


**CVASC_Other_Nos_A**

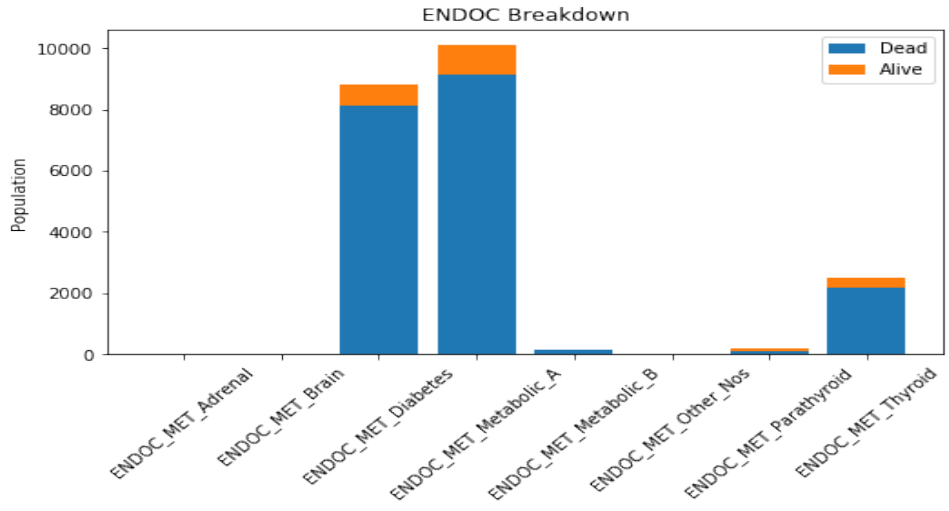


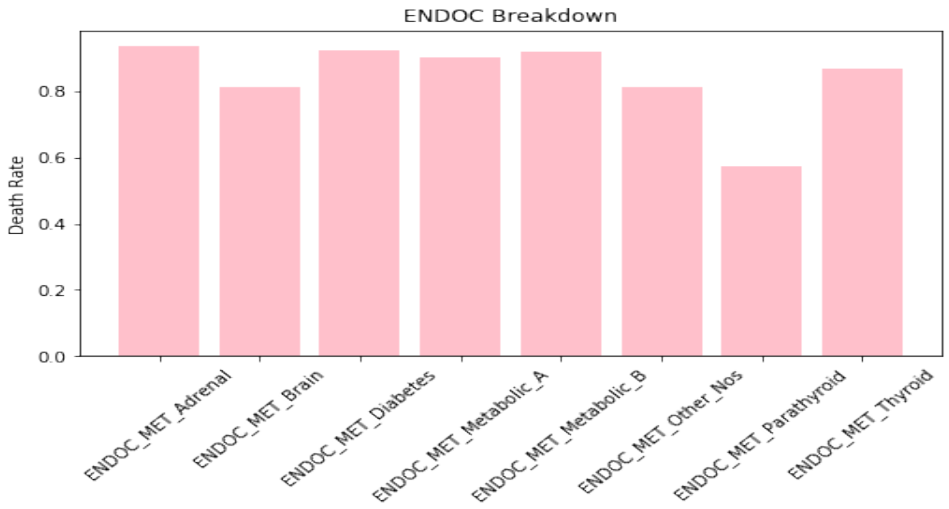
According to these plots we can find some words frequently appear: Cardiac, Heart, Valve, Atherosclerotic (Athscl for short), which indicate CVASC is a category for cardiovascular diseases. Based on the high importance of CVASC features in our model, it is reasonable to infer that underlying cardiovascular diseases can make COVID-19 more fatal for patients, that is, the social reopen in the US will expose patients with cardiovascular diseases to deadly danger. Another concern is the population of patients with cardiovascular diseases. A latest report released by American Heart Association estimated that there are 121.5 million adults in the U.S. having cardiovascular disease, including heart disease, hypertension, high cholesterol, and stroke. An article by Harvard Medical School[2] suggests four ways that the COVID-19 can affect the cardiovascular disease patients. First, for people with preexisting heart disease, the virus may load extra burden on heart and increase cardiovascular mortality. Second, the undiagnosed cardiovascular disease may be unmasked by COVID-19. Fever and inflammation render the blood more prone to clotting, while also interfering with the body's ability to dissolve clots and sparking the cardiovascular diseases. Third, the virus can directly damage heart by starving heart for oxygen even though the patients do not have any cardiovascular disease. The mismatch between oxygen supply and oxygen demand triggered by breathing difficulties and inflammation within body can further diminish oxygen supply to the heart muscle. Finally, fulminant inflammation of the heart muscle is resulted from virus directly affecting the heart. This type of inflammation may cause heart rhythm disturbances and negatively impact the heart's ability.

[2] Ekaterina Pesheva, *Coronavirus and the heart* (The Harvard Gazette, 2020) https://news.harvard.edu/gazette/story/2020/04/covid-19s-consequences-for-the-heart/

## ENDOC Analysis

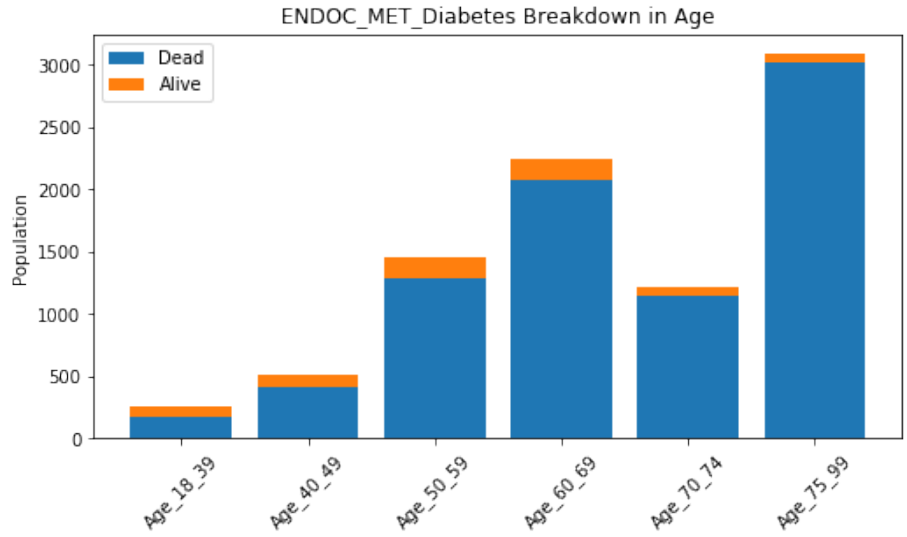
ENDOC is the second important clinical chapter category in our model as a predictor of the mortality rate of confirmed Covid-19 cases. The figures below show the break-down of the ENDOC clinical chapters and compare death and alive cases for each ENDOC clinical chapter. We see that ENDOC_MET_Diabetes, ENDOC_MET_Metabolic_A, ENDOC_MET_Thyroid are the top three chapters in regards of the number of patients in the population as well as dead cases. The chapters match those important predictors in our model based on their estimated coefficients, which are 1.427, 0.855, and 0.398 respectively. These chapters will be the ones we further investigate.

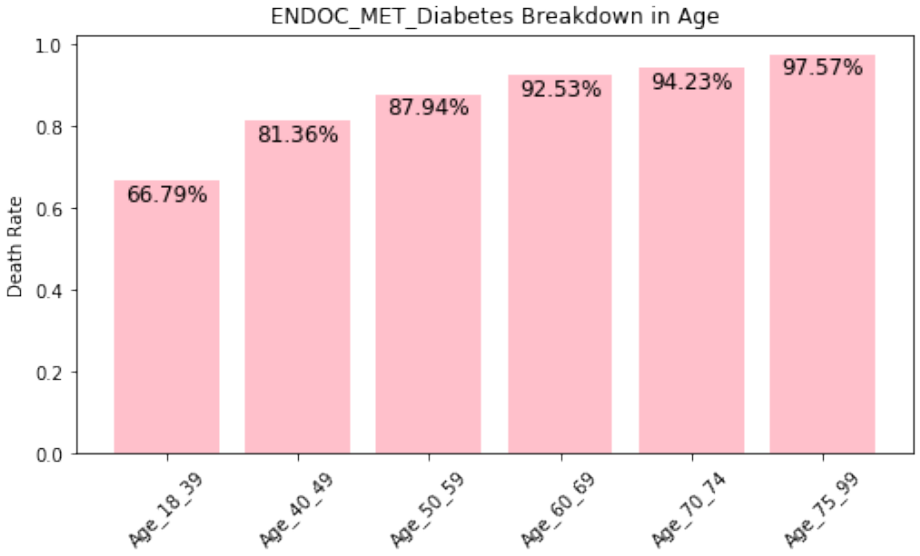


## TOP 1: ENDOC_MET_Diabetes

The figures below show the impacts of having diseases in the chapter ENDOC_MET_Diabetes on patients of different age groups. Basically, the number of cases increases with age, with the group of people aging between 75 and 99 having the largest number of cases. The death rate is quite high for each group and also increases with age, with the young population (18-39 y) having a lowest death rate at 66.79% and the elderly population (75-99 y) having a highest death rate at 95.57%.

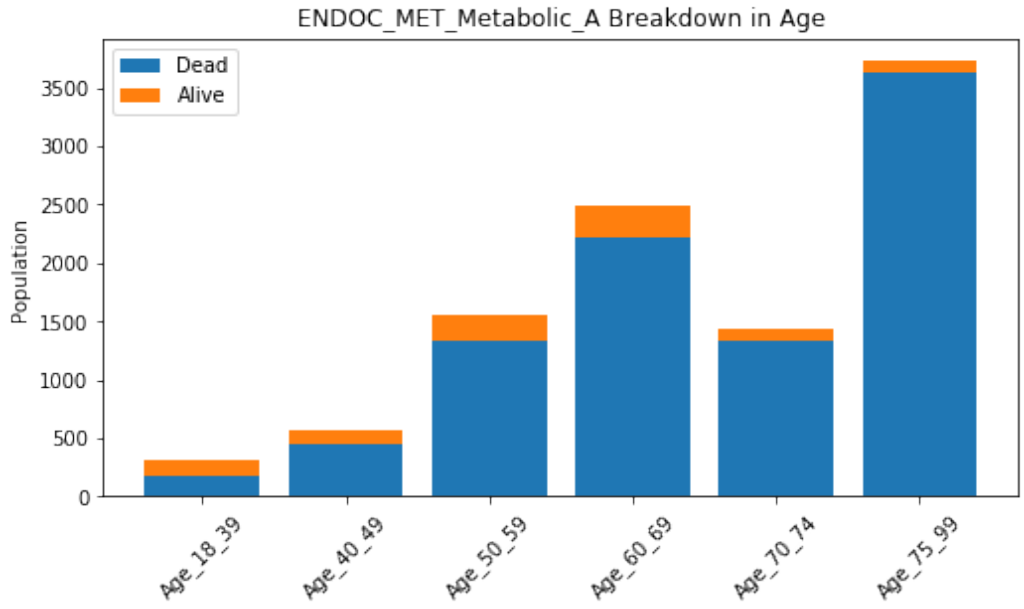


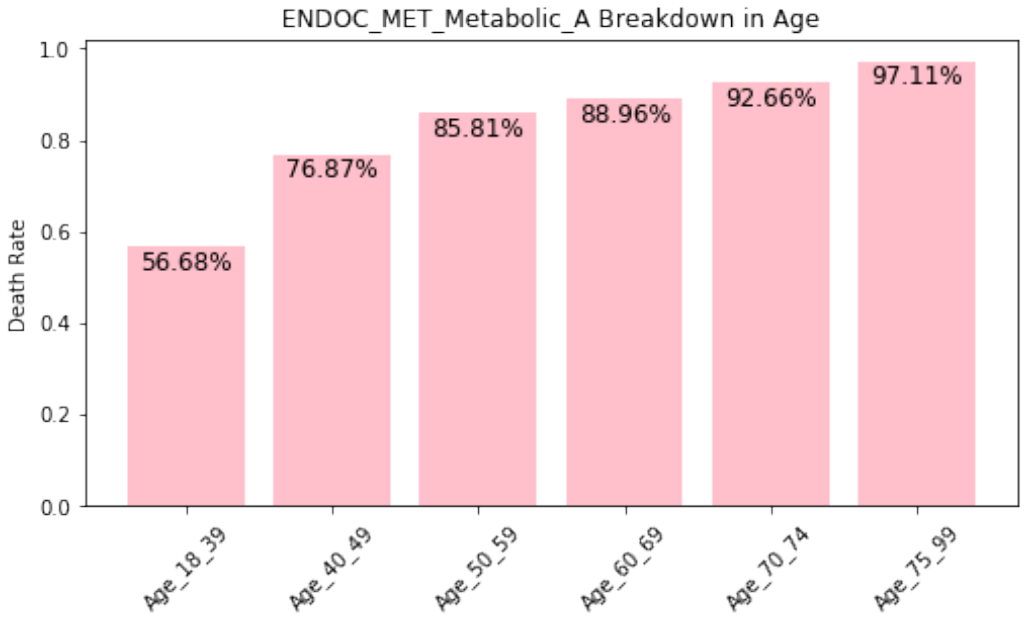



The word cloud below shows the words appearing most frequently in diagnoses belong to this chapter. We can see that ‘diabetes’ occurs most frequently as the name of this chapter suggests. The results from previous analyses are consistent with the fact that elder people are at a greater risk of developing diabetes and have less robust immune systems to confront diseases.


## TOP 2: ENDOC_MET_Metabolic_A

The figures below show the impacts of having diseases in the chapter ENDOC_MET_Metabolic_A on patients of different age groups. As we can see, the trends are very similar to those of the TOP 1 chapter. Basically, the number of cases and death rate increase with age. The death rates of all age groups are also very high, with the lowest being 56.68% and the highest being 97.11%.

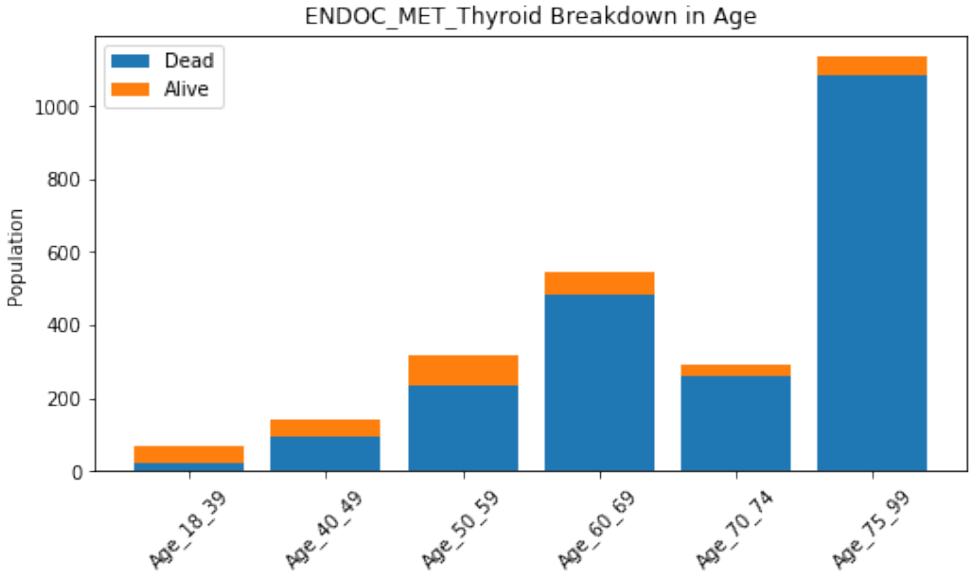


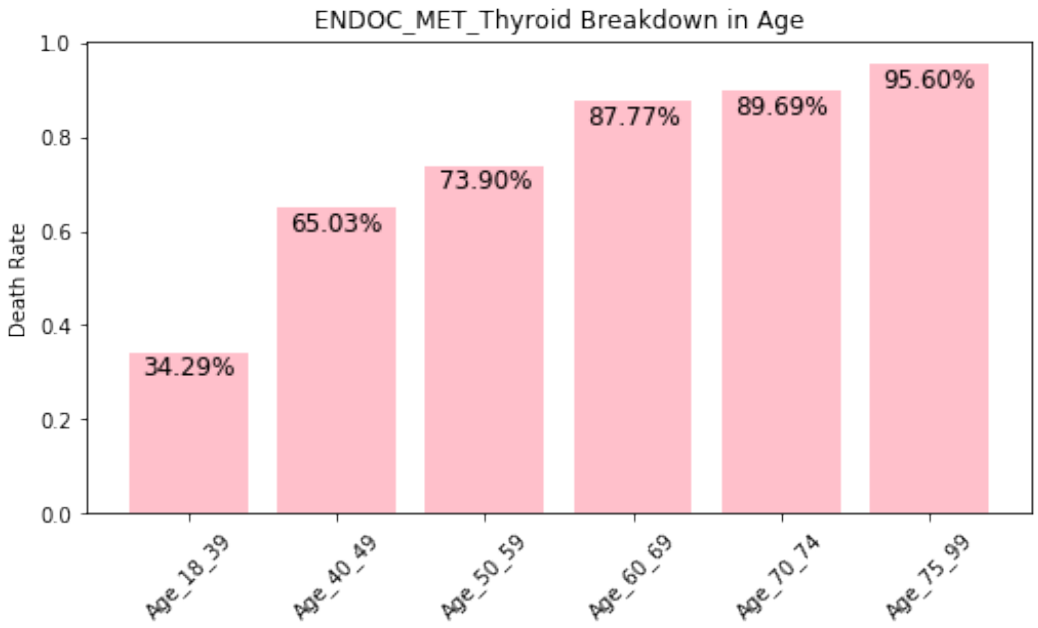




The word cloud below shows that “gout”, “chronic”, “tophus”, “metabolism”, and “disorder” are among the most frequently appearing words in diagnoses belong to this chapter. This suggest that elder people are more likely to suffer from chronic gout and metabolism disorder and are also at a greater risk of death with the concurrence of covid-19.

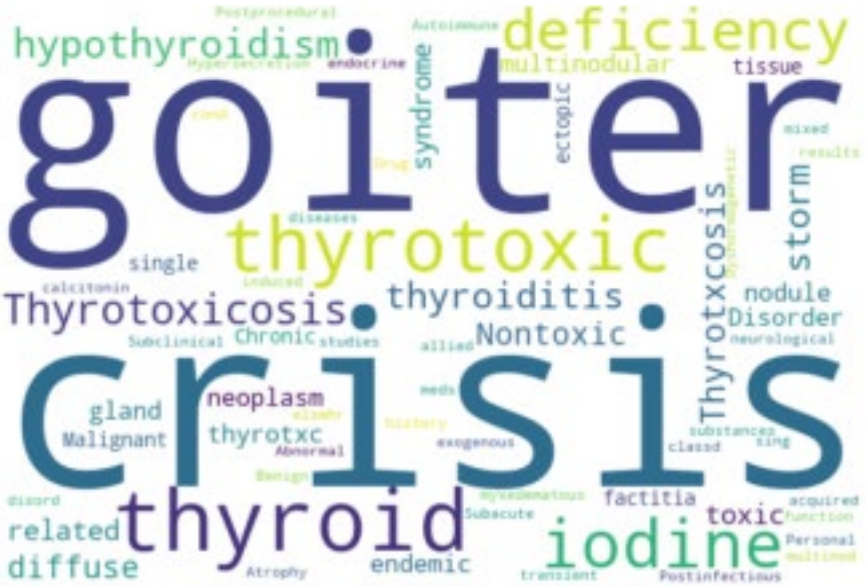

## TOP 3: ENDOC_MET_Thyroid

The figures below show the impacts of having diseases in the chapter ENDOC_MET_Thyroid on patients of different age groups. There is still a similar trend that number of cases and death rate increase with age. For the breakdown of population with diseases in this chapter, we see a clearer difference among age groups, as the number of patients aging between 75 and 99 is considerably higher. It is also noticeable that the death rate of people aging between 18 and 39 is much lower than the rest of the population. The death rate is 34.29%, compared to a death rate of 95.60% of people aging between 75 and 99s.

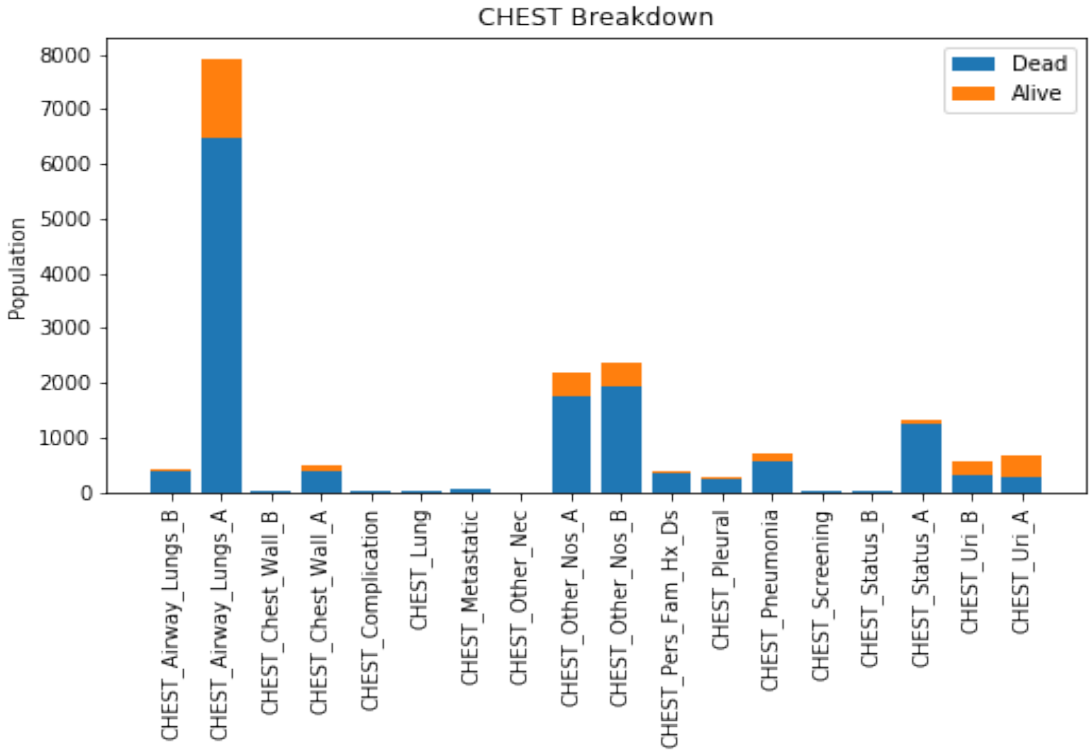


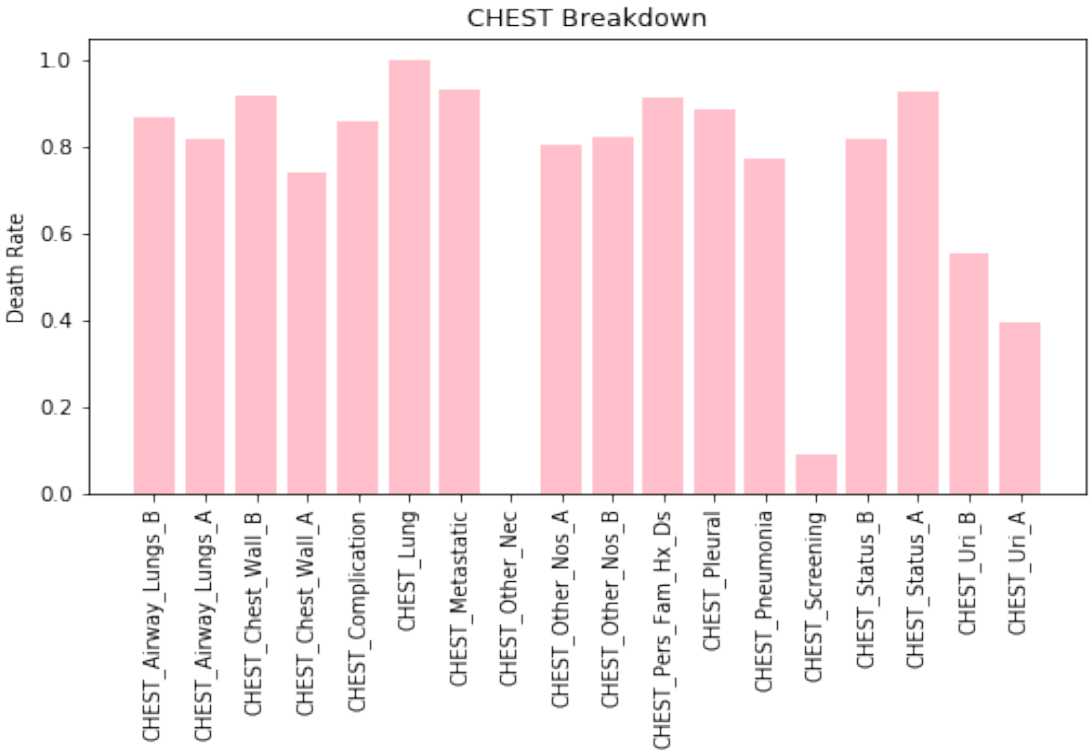



The word cloud below shows that “goiter”, “thyroid”, “thyrotoxic”, “iodine”, and “hypothyroidism” are among the most frequently appearing words in diagnoses belong to this chapter. Scientific studies have suggested that thyroid disorders are more common in older people and treatments for them are not easy. Therefore, it is reasonable to see this trend in our sample of covid-19 patients.



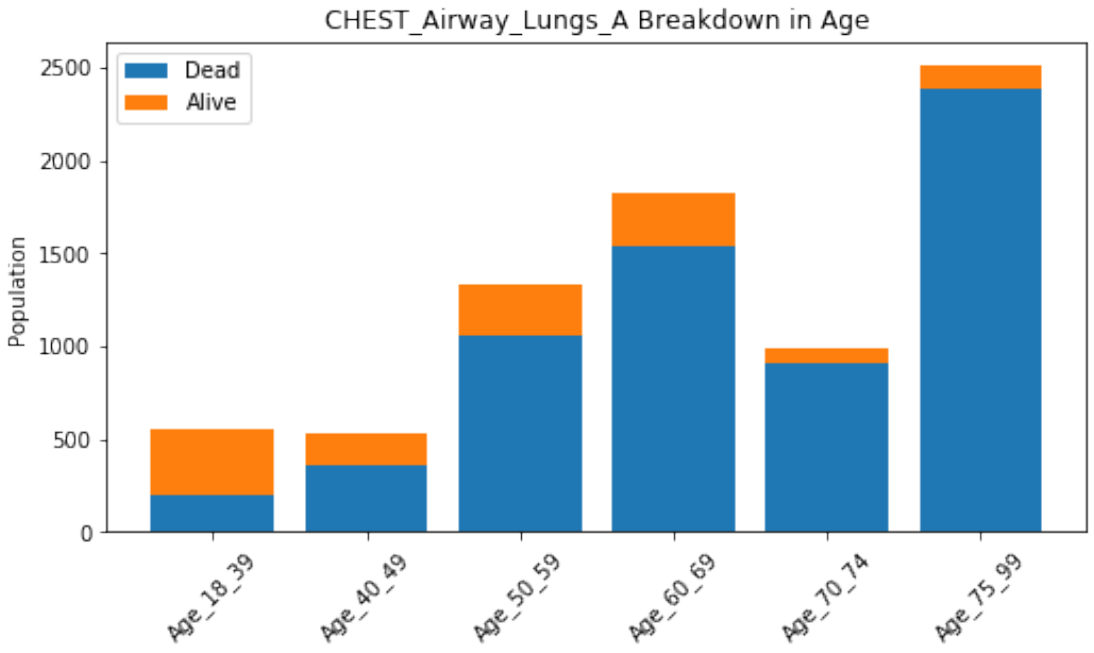
# CHEST ANALYSIS

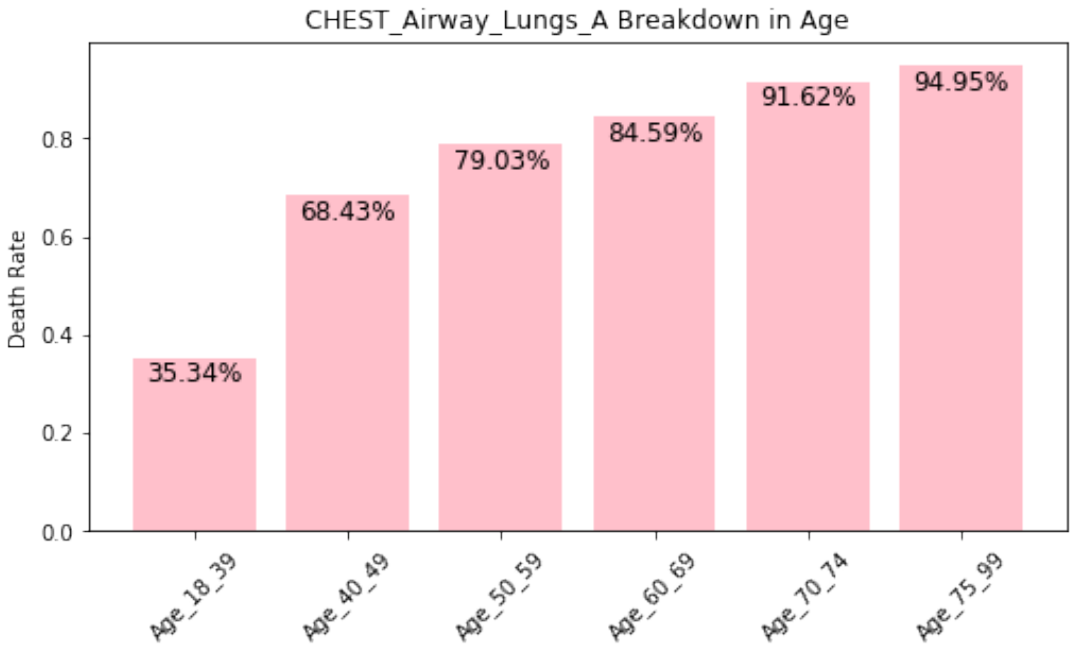


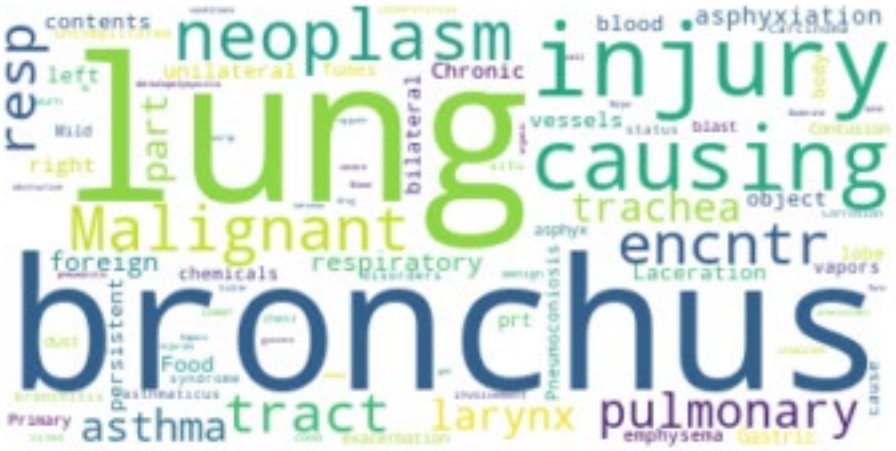
As we can see from our word cloud, CHEST is one of the most important categories to impact the mortality rate of confirmed Covid-19 cases. Therefore, we break down the CHEST category to analyze the subcategory of CHEST. The chart above in Figure CHEST Breakdown shows the comparison  of death and alive population for each CHEST subcategory, the top 4 subcategories based on total population perfectly match the feature importance based on the result of our model. Based on the feature importance, the top 4 important subcategories of CHEST is CHEST_Airway_Lungs_A (coef = 1.01), CHEST_Other_Nos_B (coef=0.88),  CHEST_Status_A( coef = 0.72),  CHEST_Other_Nos_A(coef = 0.66). Therefore, we want to dig into each of these subcategories and see if there are some insights coming out from the data.

## TOP 1: CHEST_Airway_Lungs_A





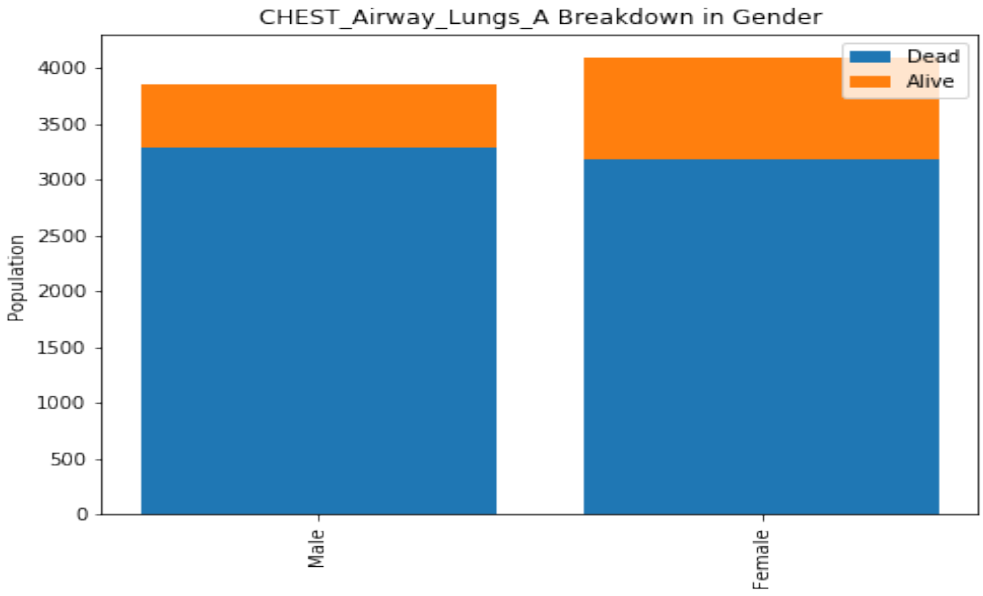

CHEST_Airway_Lungs_A would be the most important feature in the CHEST subcategory. We then break down this subcategory to see how this subcategory impacts different age groups of people. From figure above, we can see Age 18-39 group have slightly greater confirmed cases than Age 40-49. However, we can clearly see from the graph that the dead population of Age 18-49 is much lower than the dead population from Age 40-49. The graph below confirms this perspective, we can see the death rate of Age 18-39 group (35.34%) is much lower than that of Age 40-49 group (68.43%). After that, we can see that the death rate slightly grows as the age group changes to older. The result makes sense to me, since elder people have less robust immune systems to confront COVID-19 plus disease from CHEST_Airway_Lungs_A subcategory. However, the huge gap (33%) between the death rate of Age 18-39 and death rate of Age 40-49 implies that CHEST_Airway_Lungs_A subcategory may increase the risk of death from COVID-19 in 40-49-year-old adults.

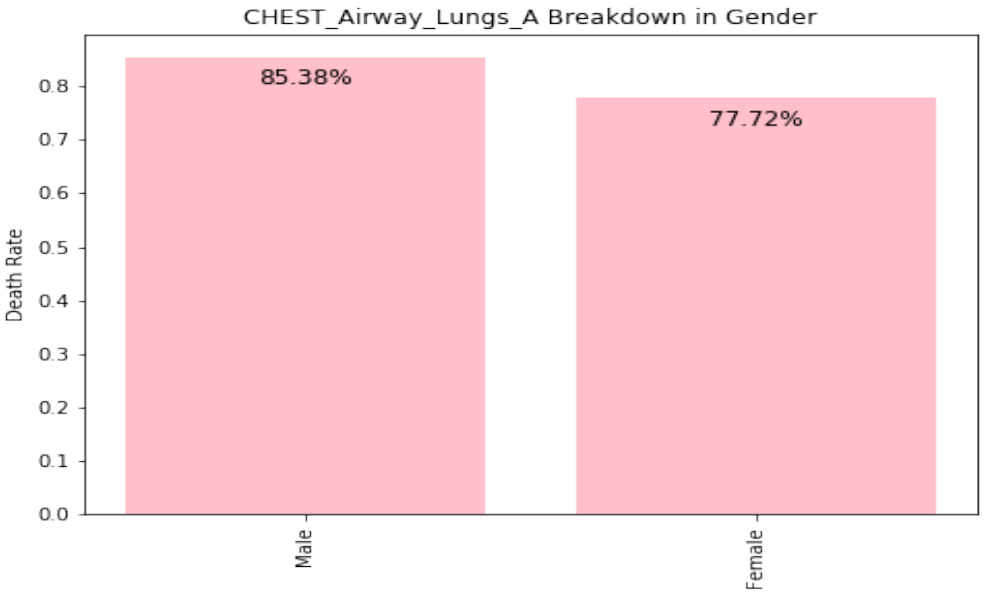



When we look at the ICD10 Dx Code Description within CHEST_Airway_Lungs_A subcategory, we can see the major diseases are asthma in all levels and malignant neoplasm of lung, etc. Take Asthma as an example. Asthma is a pre-existing lung condition affecting 1 in 13 people in the U.S. It can cause wheezing, chest tightness, coughing, and shortness of breath. COVID-19 is a respiratory disease caused by the SARS-CoV-2 virus. The virus affects cells in the airways, from the nose and throat down to the deepest parts of the lungs. In the nose and throat, it might cause symptoms of a cold. In the upper airways, it might cause some breathlessness and cough. When the coronavirus lodges itself deep in the lungs, this is when things can start to get serious. Here, the coronavirus commonly causes a double lung infection, or bilateral pneumonia, resulting in shooting up the death rate. In this way, we can infer that asthma and lung disease might be main factors to exacerbate the death rate from COVID-19 in 40-49-year-old adults and above.

Also, as the word cloud shows above, bronchus, lung and asthma are the most frequently appeared words in CHEST_Airway_Lungs_A subcategory's ICD10 code description. Therefore, even though the 40-49 year old population does not have a high death rate as older groups do, this specific population with asthma or lung disease would be highly recommended to stay at

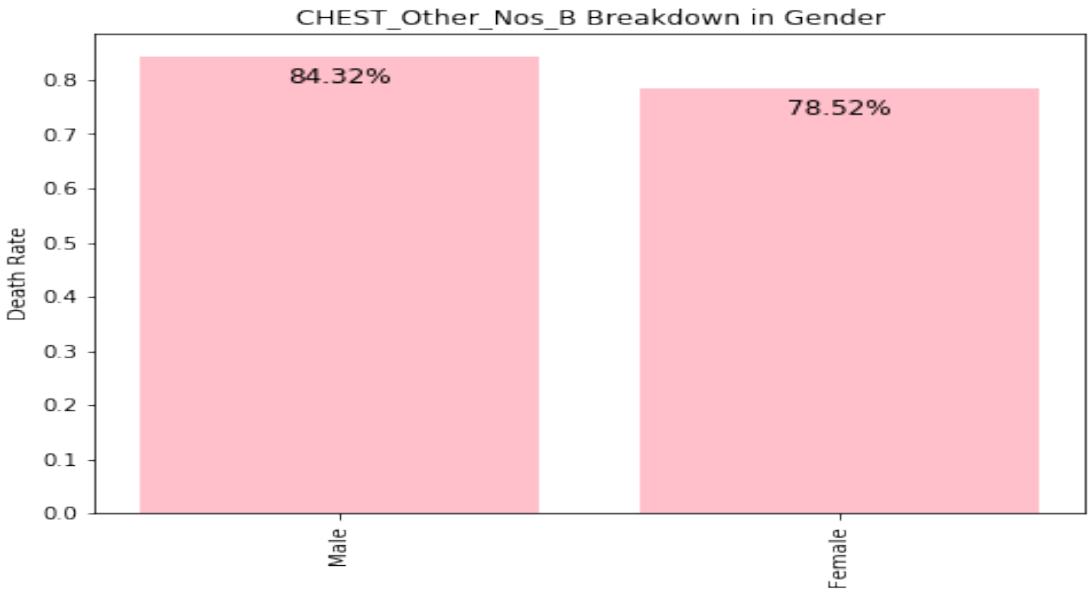
home to avoid being affected to COVID-19. We would also always recommend other groups of CHEST_Airway_Lungs_A patients above 49 year old to stay at home.

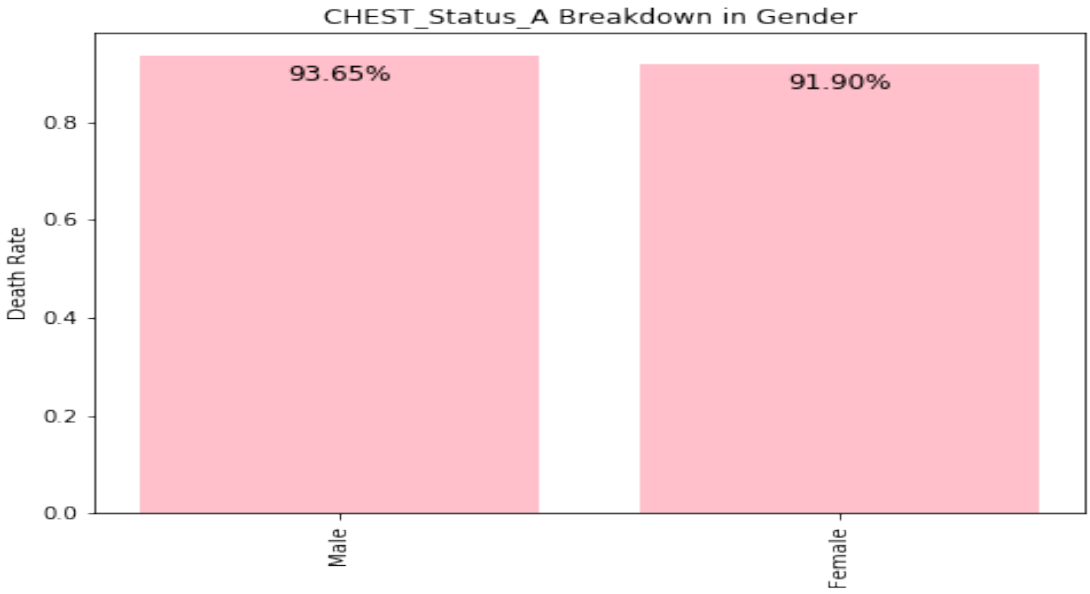


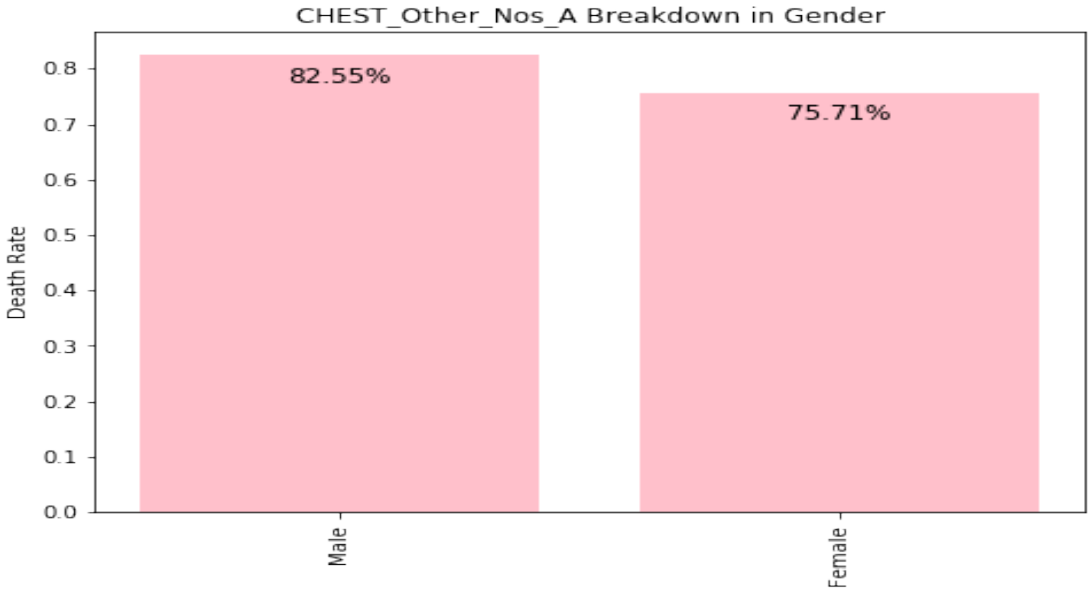




In another dimension, we break down this subcategory to see how this subcategory impacts the death rate of Male and Female differently. We can see female CHEST_Airway_Lungs_A patients have more COVID confirmed cases than male CHEST_Airway_Lungs_A patients. However, the death rate of female COVID-19 CHEST_Airway_Lungs_A patients is way lower than male

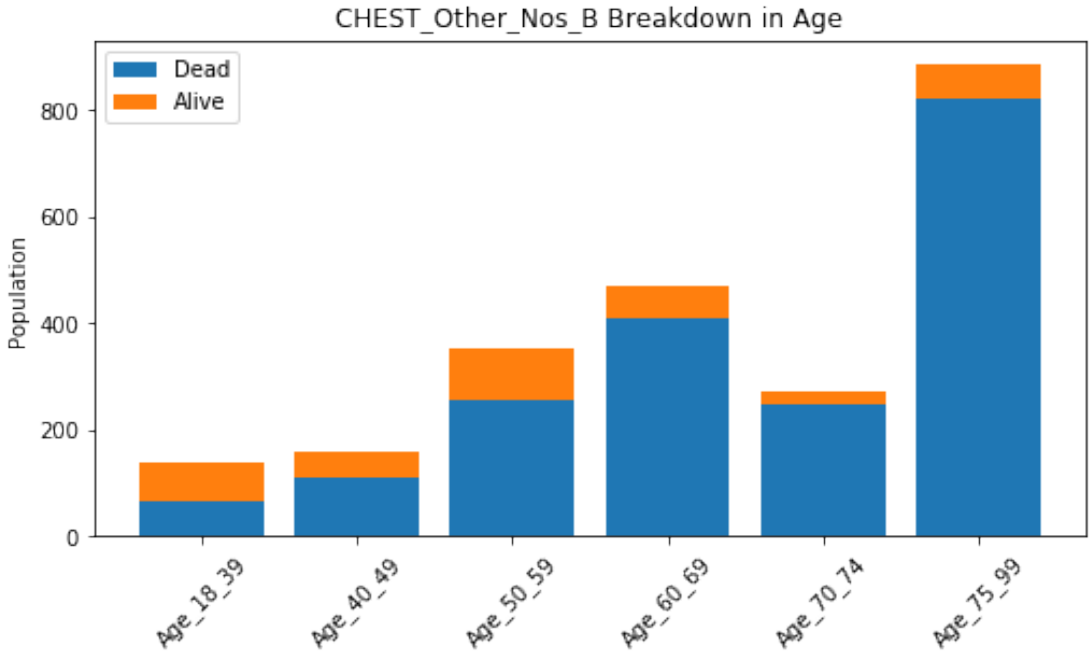
patients.  It might suggest male CHEST_Airway_Lungs_A patients are at higher risk than female patients at some level. However, the result might also be because, in general, men are at a higher risk than women responding to infection like COVID-19. Men and women are biologically different, and they differ in their sex chromosomes and the genes that lie on them. Women have two copies of a mid-sized chromosome (called the X). Men have only a single X chromosome and a small Y chromosome that contains few genes. Two X chromosomes are better than one. The X chromosome bears more than 1,000 genes with functions in all sorts of things including routine metabolism, blood clotting and brain development. The presence of two X chromosomes in XX females provides a buffer if a gene on one X is mutated. From the other major subcategories of CHEST (CHEST_Other_Nos_B, CHEST_Status_A,CHEST_Other_Nos_A), we also compared the death rate of female and male patients (as shown below), we have received similar results. Therefore, we can conclude that at least in CHEST category, male CHEST patients is at higher risk dying from COVID-19 than female CHEST patients.

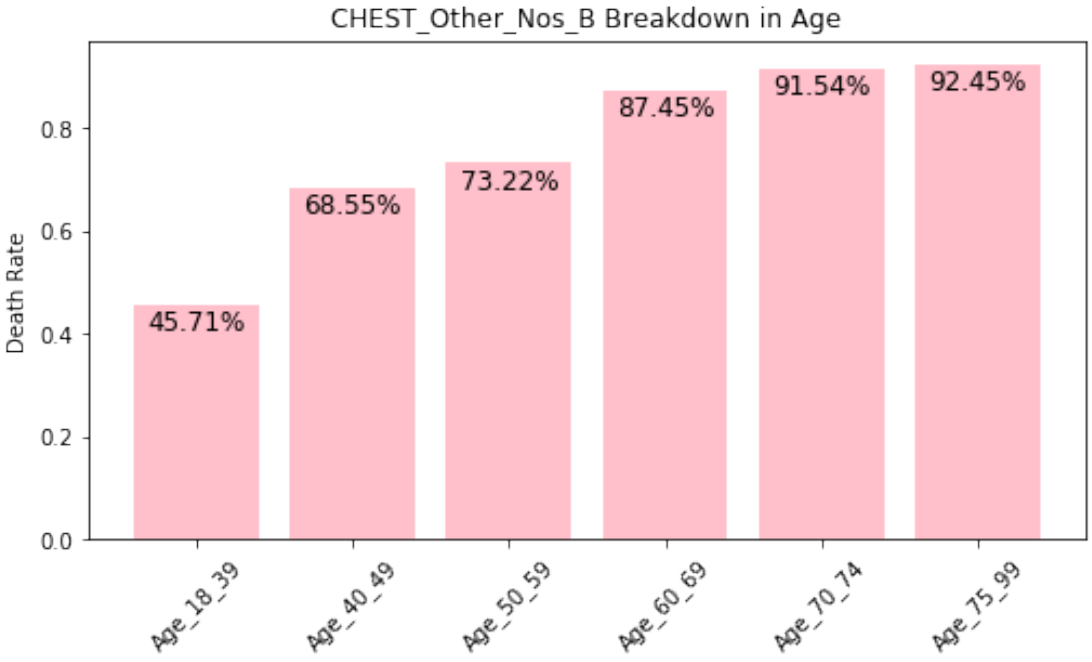



CHEST_Status_A Breakdown in Gender
93.65%
91.90%
Death Rate
0.0
0.2
0.4
0.6
0.8
Male
Female


CHEST_Other_Nos_A Breakdown in Gender
82.55%
75.71%
Death Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Male
Female

TOP 2 CHEST_Other_Nos_B




Based on feature importance from the result of our best model, CHEST_Other_Nos_B would be the second important feature in the CHEST category. Just as CHEST_Airway_Lungs_A

subcategory, we can see that there is a huge death rate gap (23%) between Group Age 18-49 and Group Age 40-49 in CHEST_Other_Nos_B subcategory. In the same way, the result might show the CHEST_Other_Nos_B subcategory increases the risk of death from COVID-19 in 40-49 year old adults. In the CHEST_Other_Nos_B subcategory, major diseases are Chronic respiratory failure, Congenital pneumonia, respiratory disorders etc. Therefore, we would highly recommend COVID-19 patients with these specific diseases at Age 40-49 and above keep social distance to protect themselves.

# III. Summary

Model part Logistics model was selected to conduct predictive analytics for risk profiling of Covid-19 patients in this study, with relatively high accuracy up to 0.937. 160 variables were picked by the machine learning model. By rolling up to the upper main categories, Age, ENDOC, CVASC, and CHEST are the four main categories making the largest contributions to higher mortality. Thus, we further explore the relationships and reasons behind these categories by applying horizontal joint analysis.

There are also some general patterns we could summarized from the analyses:

- Generally, people more than 70 years old are at the highest risk of dying when infected among the main DGL categories.
- When the infection rate is about the same in both groups, men who are infected are more likely to die among the main DGL categories.
- CVASC, which represents the cardiovascular disease, is the most important category in terms of subcategory. This group also has a positive relationship with age because the elderly tend to develop cardiovascular disease. And COVID-19 will cause severe complications or comorbidity to raise the death rate.

- One of the main factors accounting for death infected by COVID-19 is pulmonary infection, typical symptoms of this virus infection. In the CHEST category, we could also see a wider distribution of age groups with high mortality. We also notice that asthma and malignant neoplasm are keywords when carrying out text mining. Thus, people with lung disease are also the targeted high-risk population, considering a high percentage of people who suffered asthma in the United States.
- People who are already suffering from diseases related to the immune system are more likely to die because of potential complications and immune disorders brought from diseases they have. Common diseases mentioned here are like diabetes. These diseases originally accompany a higher possibility of death, not to mention the survival rate from the double whammy of age and complications.

# IV. Suggestions

Based on our model and findings, we would like to make the following recommendations of the at-risk populations and re-open efforts.

## At-risk populations

- People over 60 years old, especially over 70 years old
- People with lung disease, like asthma in all levels and malignant neoplasm
- People with preexisting heart disease and undiagnosed cardiovascular disease
- People with disease related to immune systems
- People with endocranium disease
- People with all kinds of chronic diseases

## Reopen Effort

- On the existing basis, strengthen the disinfection of nursing homes and other areas with large population density of the elderly. Social distancing should still be urged for this group of people
- Resumption of work should be  in batches and in different industries. For example, the government and health-care agencies could return to work first. In addition, at-risk population should be among the last group to return to onsite work for every age group. If they do want to work, working remotely and keeping social distancing are still necessary
- Limited social and medical resources could be more rationally distributed based on the population density and geographic distribution of at-risk populations
- At-risk population could be divided into five risk levels of death, from high to low, to allocate resources and give priority for vaccines
- The government should guarantee the living allowances for the high-risk population in case they cannot work due to social distancing