AI-Opening Case Study

A Case Study Using Predictive Analytics for Risk Profiling of Covid-19 Patients

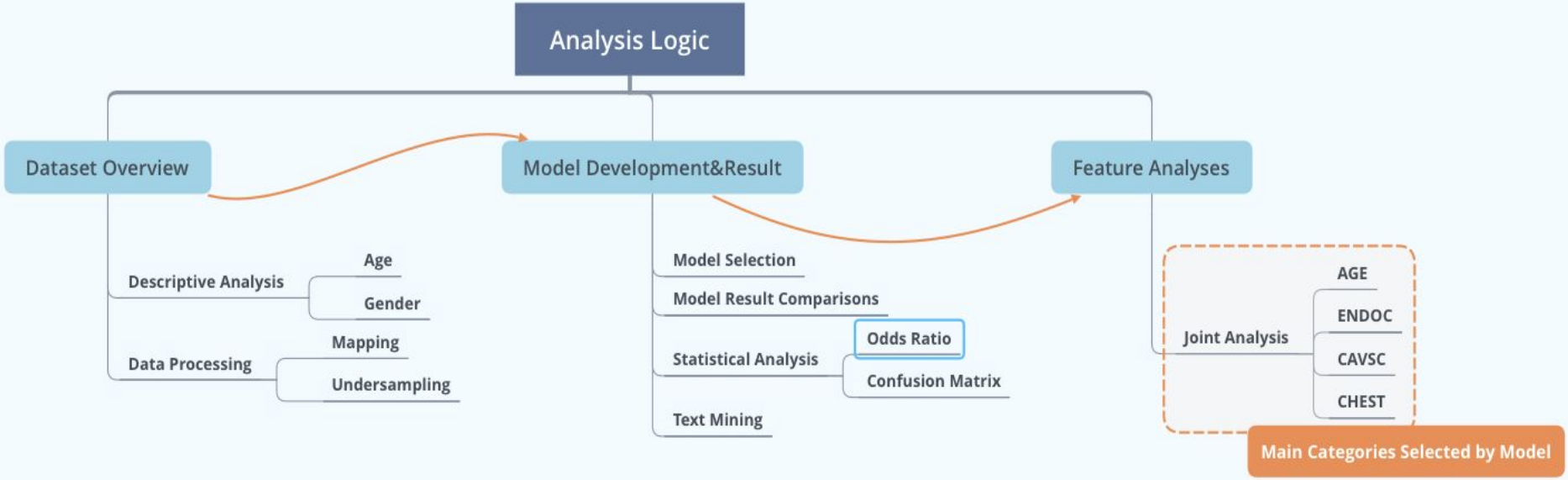
Group 3
Team Member：Xuhui Bai, Jiajie Yuan, Ying Tung Lau, Yunyi Wang, Zoey Yang

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….

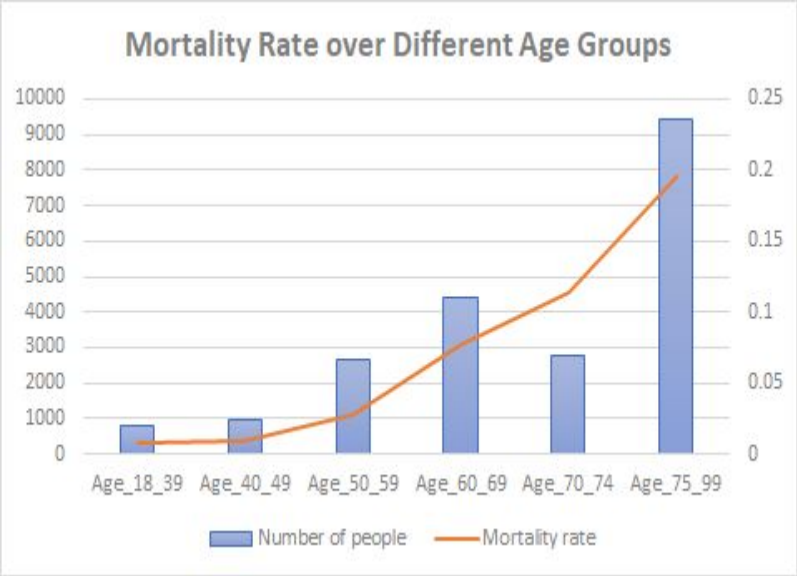
As the White House has decided to reopen the market at the beginning of May in sake of the economy, models project a sharp rise in deaths as states reopen. At this time, forecasting COVID-19 Deaths in the US is critical. Forecasts of deaths will help inform public health decision-making by projecting 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.

Analysis Logic



Dataset Overview

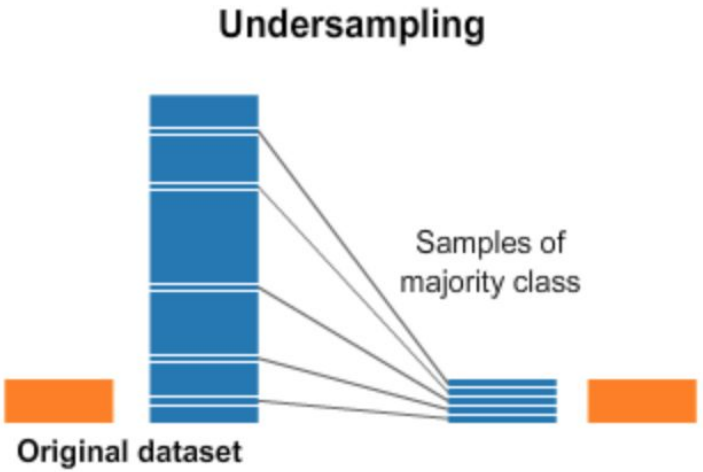
Age & Sex Analysis

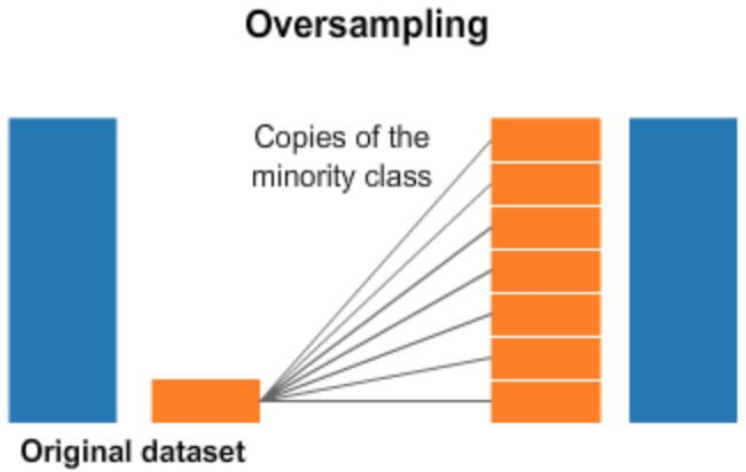


- Age increases, mortality rate increases
- Males have higher mortality rate

Data Procession & Model Development

Sampling Method

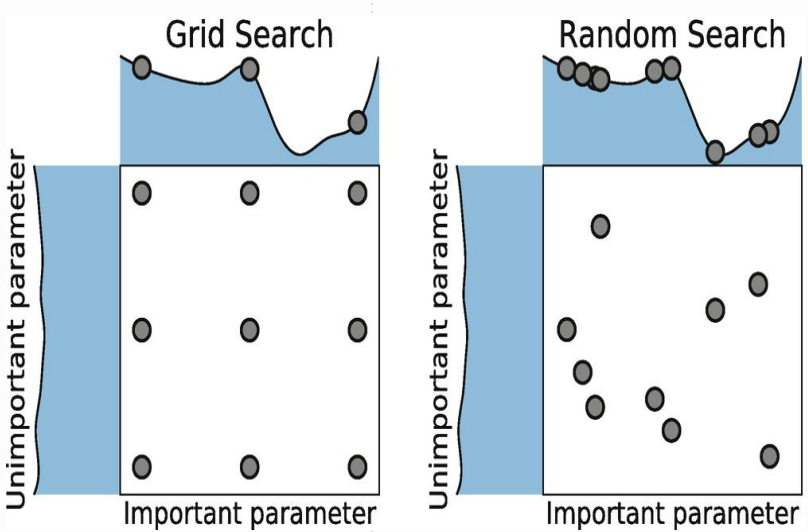




Random under-sampling is very simple and intuitive under- sampling technique .
Method works by randomly choosing the samples from dominant class.

Data Procession & Model Development

Hyper-parameter Optimization



- Choosing the best model and hyperparameters are challenges that must be solved for improvements in predictions.
- Grid Search looks through each combination of hyperparameters. This means that every combination of specified hyperparameter values will be tried.

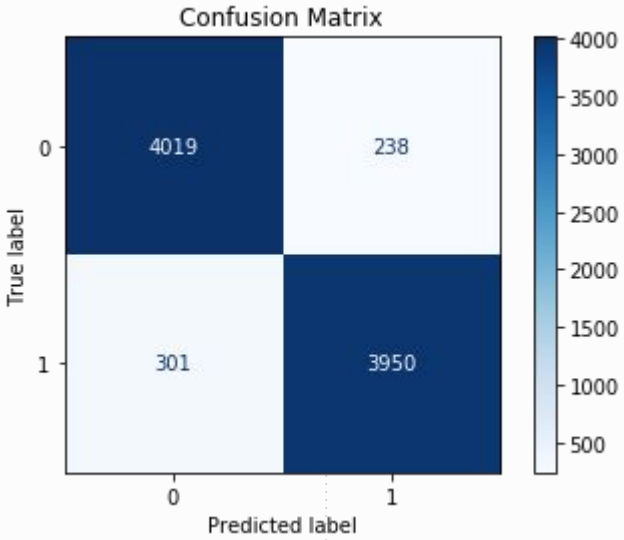
Data Procession & Model Development

Model Results Table

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

“ We will use Logistic Regression in our case as the score on testing set is the best. 160 features were selected. ”

Confusion Matrix



Accuracy	0.937
Recall	0.929
Precision	0.943
F1 Score	0.936

“ The indicators are all pretty good, meaning a relatively high accuracy of our predictive model. ”

Odds Ratio

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

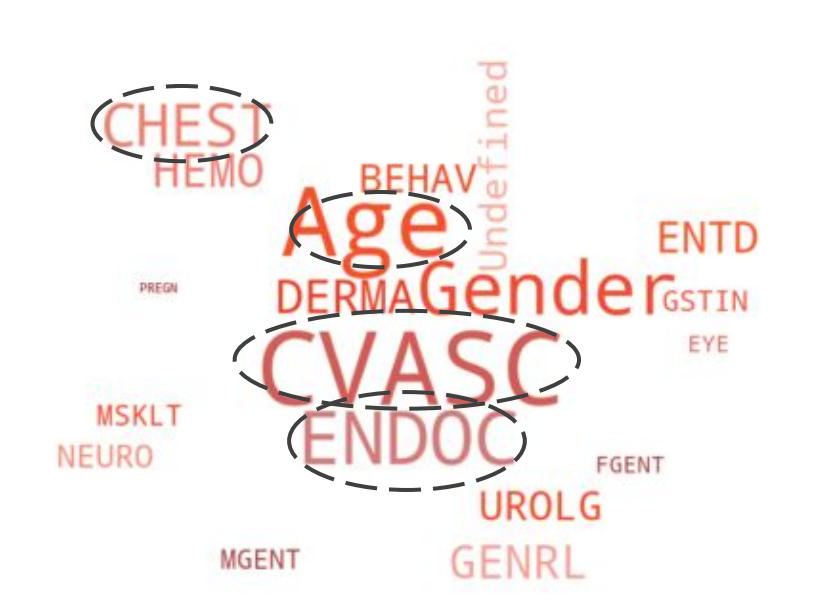
Clinical Chapter

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 are mostly associated with heart, lung, or vascular diseases.

Top Risk Factors Revealed by Model

Categorical Analysis on Pre-existing Health Condition



Features with the same initial capital letters are regarded and groupby the same main category.

The weight of each word was the averaged coefficient for features within the category.

AGE, ENDOC, CVASC, CHEST are the four main categories.

CVASC
Cardiovascular Diseases

ENDOC
Diabetes

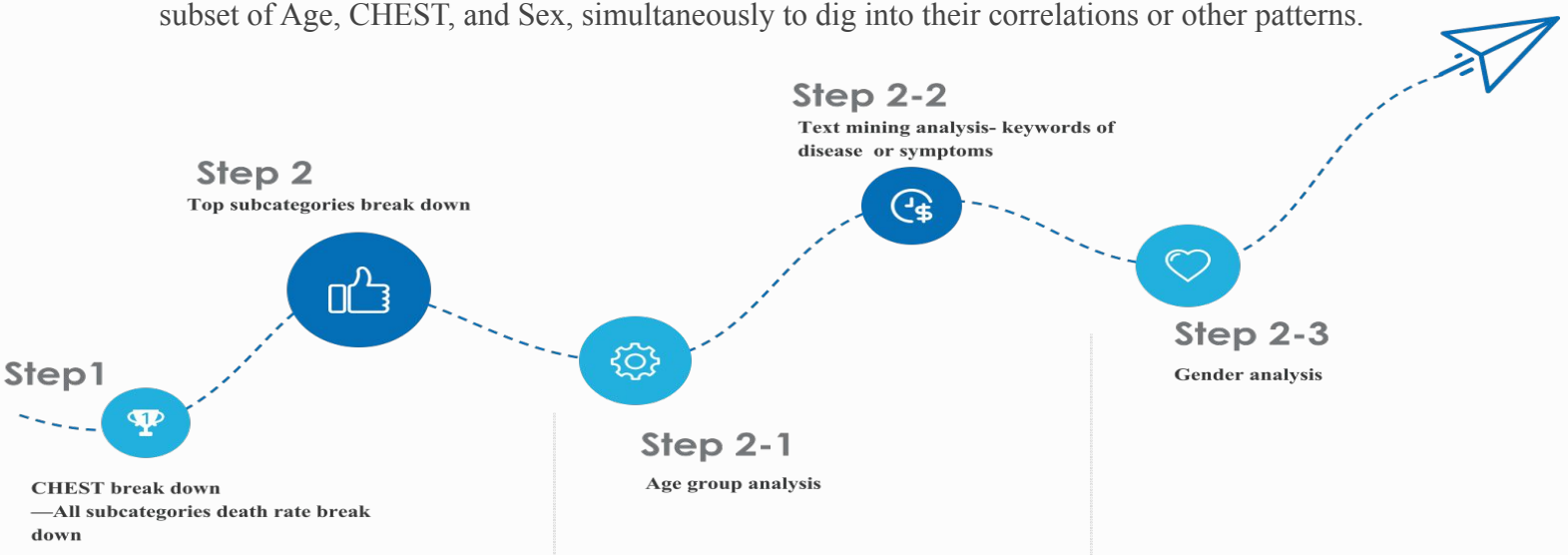
CHEST
Respiratory Diseases

Feature Analyses

Taking CHEST As An Example

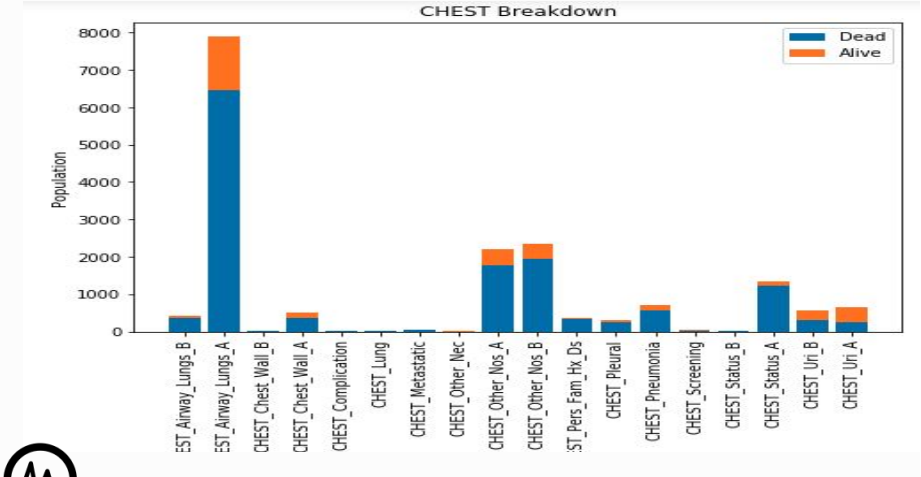
Highlight of Analyses Process

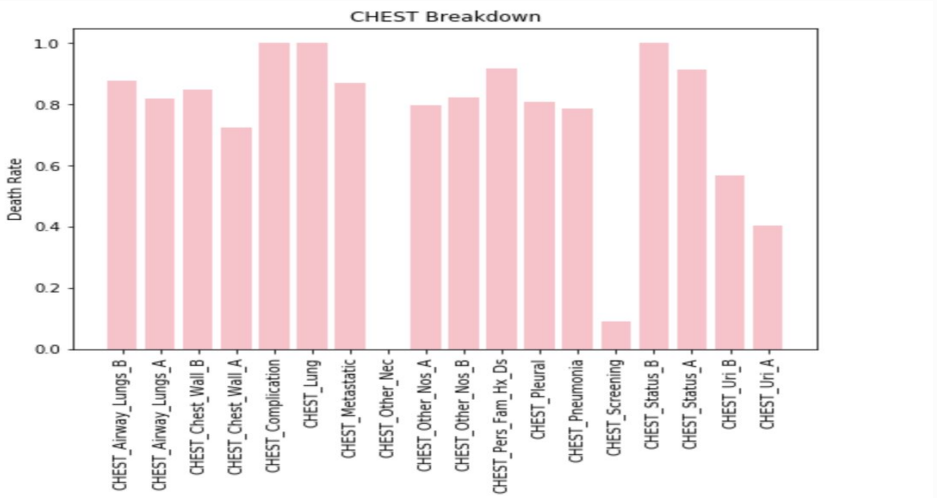
AGE, ENDOC, CVASC, and CHEST are the five main categories which could be the determinants leading to a higher risk of mortality. While previous points started from looking at effecting factors separately, in the following part, we will conduct horizontal joint analysis, meaning that considering several factors, like a subset of Age, CHEST, and Sex, simultaneously to dig into their correlations or other patterns.



Feature Analyses -- CHEST Subcategories Break Down

Taking CHEST As An Example

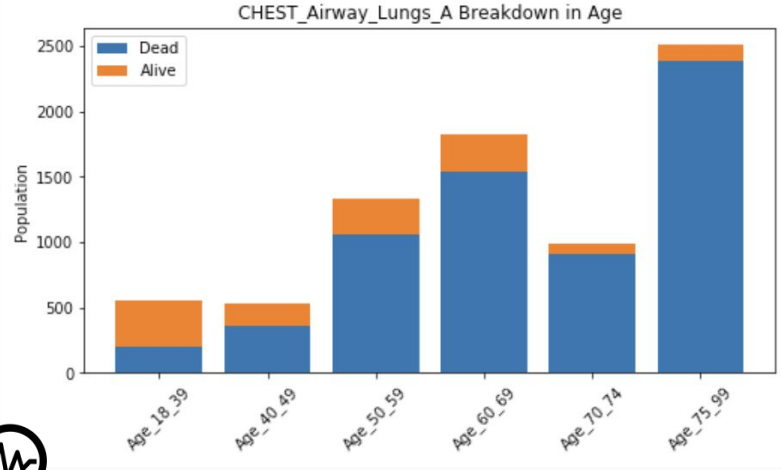


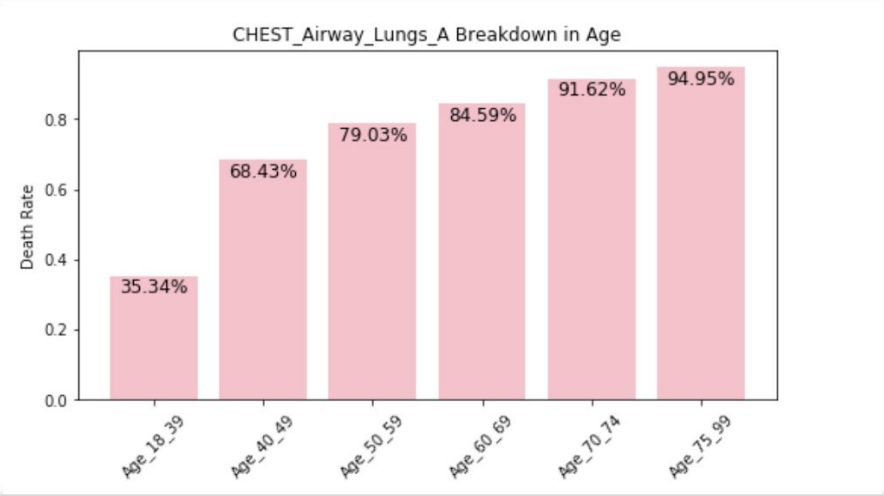


The charts above shows the comparison of death and alive population for each CHEST subcategory, the top 4 subcategories based on total population perfectly match the feature importance based on the result of our model. Based on the feature importance, the top 4 important subcategories of CHEST is CHEST_Airway_Lungs_A (coef = 1.01), CHEST_Other_Nos_B (coef=0.88), CHEST_Status_A(coef = 0.72), CHEST_Other_Nos_A(coef = 0.66).

Feature Analyses -- CHEST Top Subcategories Analysis

TOP 1: CHEST_Airway_Lungs_A



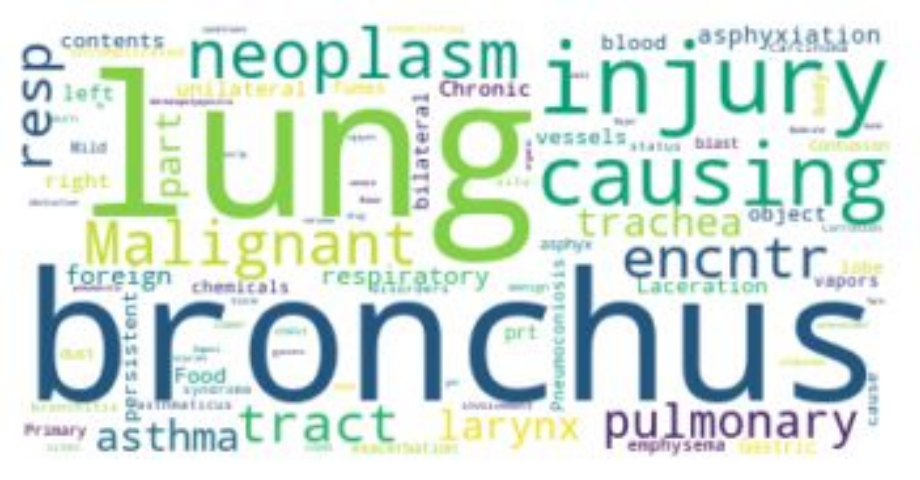


Age 18-39 group have slightly greater confirmed cases than Age 40-49. However, the dead population of Age 18-49 is much lower than the dead population from Age 40-49. The death rate slightly grows as the age group changes to older as elder people have less robust immune systems to confront COVID-19 plus 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.

Feature Analyses -- CHEST Top Subcategories Analysis

TOP 1: CHEST_Airway_Lungs_A

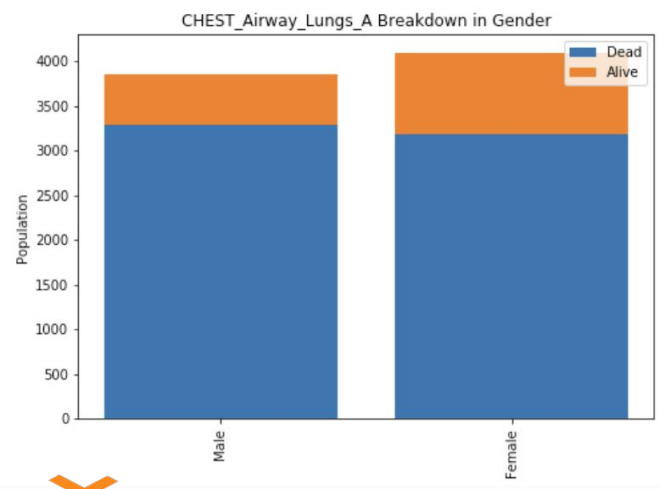


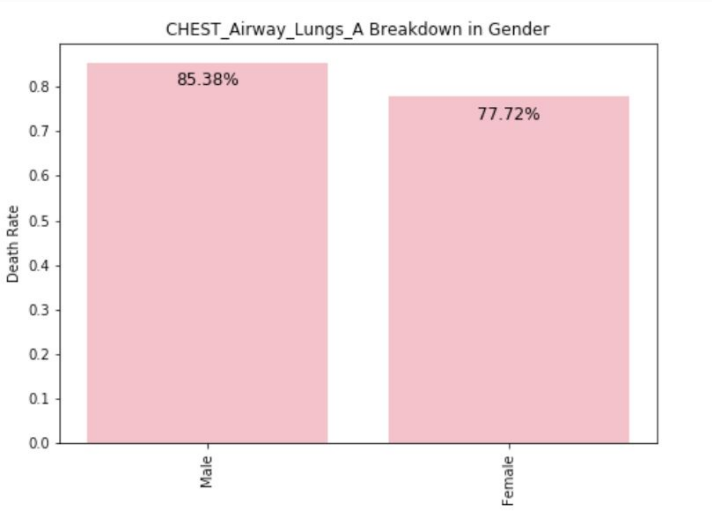
- Major diseases are asthma in all levels and malignant neoplasm of lung and etc. Lung and asthma are the most frequently appeared words in CHEST_Airway_Lungs_A subcategory's ICD10 code description.

- The 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.

Feature Analyses -- CHEST Top Subcategories Analysis

TOP 1: CHEST_Airway_Lungs_A

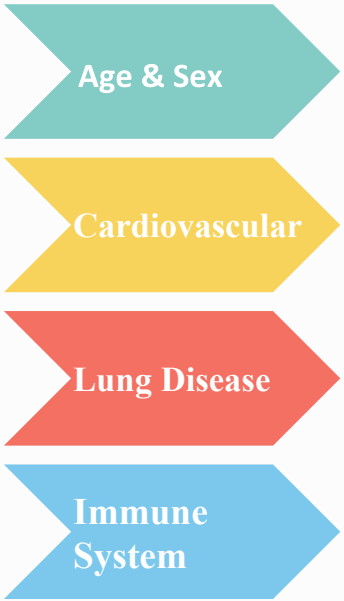




Male CHEST_Airway_Lungs_A patients are at higher risk than female patients at some level when infected with COVID-19. There are major gender difference in this case.

Summary

Some general patterns we could summarized from the analyses



People more than 70 years old and Males are at the highest risk of dying when infected among the main DGL categories.

The group with cardiovascular disease 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.

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 complications and immune disorders brought from diseases they have.

At-risk populations

Elderly People
People over 60 years old, especially over 70 years old.

Pulmonary Patient Group
People with lung disease, like asthma in all levels and malignant neoplasm.

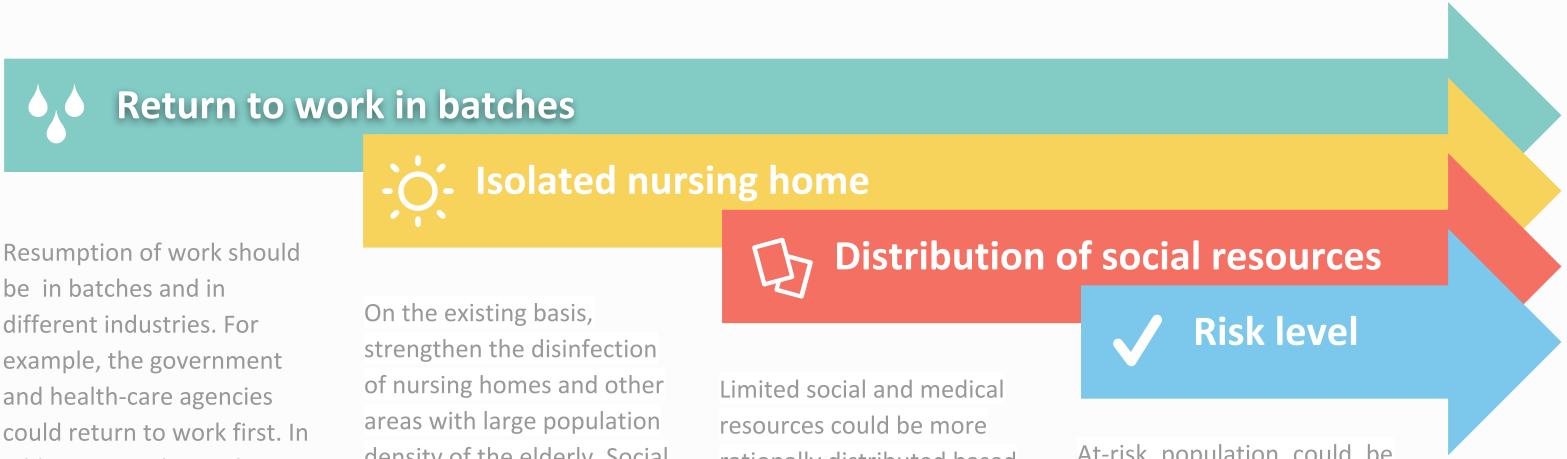
Cardiovascular Disease Group
People with preexisting heart disease and undiagnosed cardiovascular disease.

Immuse System Disease Group
People with disease related to immune systems.

Endocrinium Disease Group
People with endocrinium disease.

Chronic Disease Group
People with all kinds of chronic diseases.

Reopen Effort



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

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.

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.