**Outline**

Introduction

* COVID + Significance
* COVID Pneumonia + Significance

Exploratory Data Analysis (EDA)

* Dataset
* Univariate/bivariate analysis
* Correlation analysis

Methodology

* Data preparation – feature engineering & pre-processing
* Logistic regression (LR)
* Support vector classifier (SVC)

Results: Active features

Conclusion

Evaluation

* Limitations
* Applications
* Future research
  + Comorbidities that may increase the risk of death from CCOVID-19 pneumonia
    - <https://www.nature.com/articles/s41598-021-82862-5>
    - <https://pubmed.ncbi.nlm.nih.gov/33966261/>

**Draft 1**

**Title**: Predictors of COVID-19 Pneumonia

**Introduction**

In December 2019, the Coronavirus Disease 2019 (COVID-19) began its rapid spread worldwide, becoming the fifth documented pandemic throughout history (Moore, 2021). Despite many skepticisms, its asymptomatic and contagious incubation period allowed it to spread effectively in a series of surges (Achenbach, Cha, & Sellers, 2021; Maragakis, 2021). By the end of 2020, there were almost 2 million COVID-19 deaths reported, with an estimate of at least 3 million excess mortalities (World Health Organization, 2021).

Being a significantly pro-inflammatory condition, one of the common mechanisms that COVID-19 kills is through instigating pneumonia as a lung complication (Galiatsatos, 2022). In 2022, about 15% of people with COVID-19 develop serious complications, with 75% being pneumonia (Wiersinga, Rhodes, Cheng, Peacock, & Prescott, 2020; Cleveland Clinic, 2022).

Pneumonia is a respiratory infection that causes the air sacs in one or both lungs to fill with pus and other liquids (Johns Hopkins, n.d.). Although most people respond well to pneumonia treatment, it has the potential to be life-threatening. In particular, for people with impaired immune systems such as elderlies, children, or patients with pre-existing conditions, pneumonia tends to lead to complications (Johns Hopkins, n.d.).

Unlike the standard pneumonia, pneumonia caused by COVID-19 are often severe which are often associated with very high mortality. This is because, it uses the individuals’ immune system to spread, allowing it to last longer and cause widespread damage in multiple areas of the body (Cleveland Clinic, 2022). It also often impairs both lungs, limiting one’s ability to take in oxygen causing agonizing symptoms, such as shortness of breath and coughing, alongside lasting lung injuries that persists months after its recovery (Galiatsatos, 2022).

Unfortunately, a significant proportion of COVID-19 pneumonia patients are presented with a lack of certain symptoms. For example, a significant proportion of COVID-19 patients do not exhibit dyspnoea, despite it being a common predictor of regular pneumonia. This insidious nature of COVID-19 pneumonia worsened prognosis by increasing the difficulty of disease diagnosis, resulting in a sudden increase in need for high-intensity care (Goyal, et al., 2021). This is a problem as a substantial increase in severe cases can overwhelm less prepared hospitals, which doubles the mortality rate of COVID-19 pneumonia from 20% to 40% (Cleveland Clinic, 2022). To exacerbate the issue is the fact that resource-constrained environments are now common due to sudden surges of cases and poor economics. All of these factors associate COVID-19 pneumonia with severe disease development and high mortality (Mahendra, Nuchin, Kumar, Shreedhar, & Mahesh, 2021).

To further complicate the issue, patients without dyspnoea may instead present with other symptoms of hypoxia, such as severe fatigue, exertional fatigue and/or altered mental status (Goyal, et al., 2021). Ultimately, the great difficulty of diagnosis identifies a necessity in discerning symptoms specific to COVID-19 pneumonia to help with diagnosis and reduce its mortality rate.

Logically, to help with diagnosis, we need to identify explainable risk factors associated with COVID-19 pneumonia, especially latent ones. By identifying risk factors, healthcare workers can better predict and prevent the development of pneumonia by employing early interventions. Doing so is more effective than treating severe COVID-19 pneumonia after it arises and is advantageous in conserving resources. Being aware of more risk factors can also further inform the general public on how to keep themselves safe through the understanding of to adopt and avoid. For example, a known risk factor of COVID-19 pneumonia is smoking because it damages the lungs and by quitting smoking the risk of an infection that leads to pneumonia decreases (Cleveland Clinic, 2022). Accordingly, the differing and deadly nature of COVID-19 pneumonia stresses the importance of exploring risk factors to reduce its mortality rate.

**Exploratory Data Analysis (EDA)**

*Dataset*

The dataset used was from the ‘COVID-19 Case Surveillance Restricted Use Detailed Data’. It is a database of patient-level data recorded in the United States (US) and uploaded on a monthly basis from April of 2020 till the current date November 2022. Despite an incomplete download the entire database, 158 of the files were downloaded and unzipped. The downloaded files contained 1,433,415,248 observations in total. The dataset contained at most 33 features which can be found in Appendix A, with some of the earlier files missing one or two features.

The incompleteness of the dataset is due to the ‘Unknown’, ‘Missing’, and ‘NA’ values which varies by jurisdiction and time period. The value “Unknown” represents when jurisdictions specify that the value is unknown, the value “Missing” represents when jurisdictions do not provide a value, and the value “NA” represents when the value is suppressed as part of privacy protections. For the sake of simplicity, these three values are put under one class ‘Unknown’ for all predictors.

As ‘pna\_ya’ is the target variable, the ‘Unknown’, ‘Missing’ and ‘NA’ values in this feature are removed as they do not provide any useful information. Since there is an ample amount of data, it was assumed that dropping these observations would not lead to any significant impact. Accordingly, out of 1,433,415,248 observations, 1,337,516,466 observations were dropped, which leaves 95,898,782 (6.69% data remaining).

Due to computational power and time, only 3 files with removed observations were used for modelling. The three files were:

1. COVID\_Cases\_Restricted\_Details\_03312021\_Part\_1.parquet
2. COVID\_Cases\_Restricted\_Details\_03312021\_Part\_2.parquet
3. COVID\_Cases\_Restricted\_Details\_03312021\_Part\_3.parquet

These three files represented the data collected during March of 2021. It is important to understand the context around this time in the US to better handle the dataset. Firstly, earlier in January 2021, vaccines were made available for first responders and individuals 65 and older (macmillan learning, 2022).

These files were combined into one data frame containing 1515261. The dataset also contained all 33 features, however ‘race’ and ‘ethnicity’ was merged into ‘race\_ethnicity\_combined’ which resulted in 32 features. Further feature reduction was conducted by removing all 4 datetime data: ‘cdc\_report\_dt’, ‘cdc\_case\_earliest\_dt’, ‘onset\_dt’, and ‘pos\_spec\_dt’. This is because the features lack value as they only indicate the initial date, but not the duration.

A number of predictors was also removed due to target leakage, which included: hosp\_yn, death\_yn icu\_yn, mechvent\_yn. This is because the pattern within these predictors likely have a causal relationship with the target variable, however with the target variable causing the results.

This leaves .. relevant variables

*Pneumonia*

The target variable ('pna\_yn') is heavily imbalanced, with 'No' making up 93% of the training observations Figure 1. Accordingly, of 1015224 observations, only 6.7% of patients presented with pneumonia. Need to take into consideration the different COVID-19 variants when considering the risk of pneumonia, especially for America where the data is collected from. Also need to consider COVID-19 vaccine dose administered as it leads to less pneumonia 1. For example, the Omicron variant, which spread since December 2021, rarely causes pneumonia 2.

1. https://pubs.rsna.org/doi/10.1148/radiol.220129#:~:text=The%20answer%3A%20pneumonia%20was%20present,for%20the%20difference%20between%20groups).
2. https://www.google.com/search?q=risk+of+pneumonia+for+omnicrom&oq=risk+of+pneumonia+for+omnicrom&aqs=chrome.

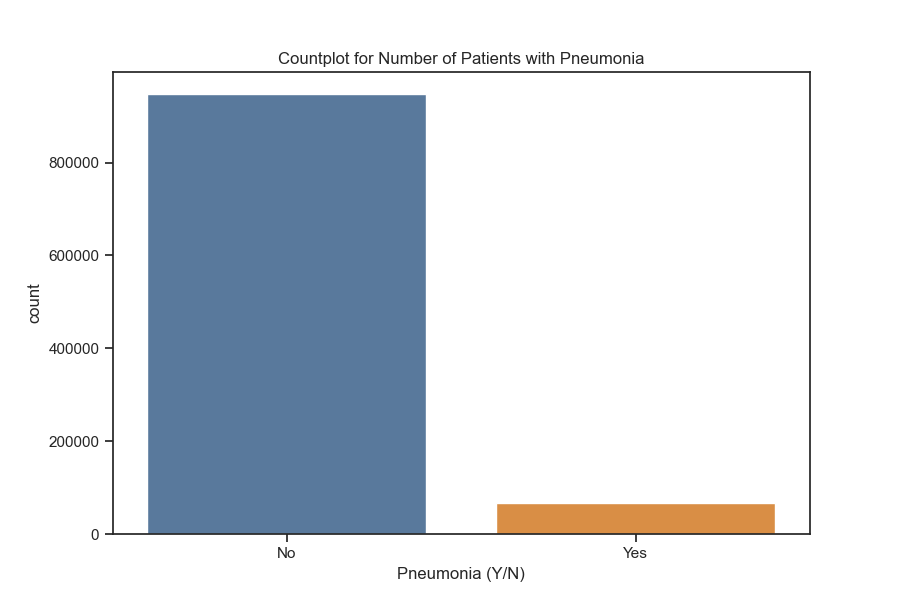


Figure : Pneumonia Countplot

Of the 33 predictors, only … was evaluated to be important.

*Race & ethnicity combined*

**Data Pre-Processing**

*Categorical encoding*

*‘Unknown’ class*

To deal with the missing values,

*Feature scaling*

Feature scaling is important for logistic regression as it allows it to quickly optimize the model.

*Train Test Split*

The dataset was split into a training set with the test set making up 33% of the entire dataset. There was no split for validation set as cross-validation (cv) will be conducted to tune the hyperparameter.

**Methodology**

Two models were selected for the current dataset: Logistic Regression (LR) and Linear Support Vector Classifier (LinearSVC).

*Logistic regression*

# 

# References

Achenbach, J., Cha, A. E., & Sellers, F. S. (2021, March 9). *A viral tsunami: How the underestimated coronavirus took over the world*. Retrieved from The Washington Post: https://www.washingtonpost.com/health/2021/03/09/coronavirus-spread-world/

Cleveland Clinic. (2022, October 8). *COVID Pneumonia*. Retrieved from Cleveland Clinic: https://my.clevelandclinic.org/health/diseases/24002-covid-pneumonia#:~:text=How%20many%20people%20with%20COVID,severe%20cases%20of%20COVID%2D19.

Galiatsatos, P. (2022, February 28). *COVID-19 Lung Damage*. Retrieved from Johns Hopkins: https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/what-coronavirus-does-to-the-lungs#:~:text=The%20pneumonia%20that%20COVID%2D19,breath%2C%20cough%20and%20other%20symptoms.

Goyal, D., Inada-Kim, M., Mansab, F., Iqbal, A., McKinstry, B., Naasan, A. P., . . . Burke, D. (2021, October 19). Improving the early identification of COVID-19 pneumonia: a narrative review. *BMJ Open Respiratory Research, 8*(1). doi:10.1136/bmjresp-2021-000911

ILO, FAO, IFAD & WHO. (2020, October 13). *Impact of COVID-19 on people's livelihoods, their health and our food systems*. Retrieved from World Health Organization: https://www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people's-livelihoods-their-health-and-our-food-systems#:~:text=The%20economic%20and%20social%20disruption,the%20end%20of%20the%20year.

Johns Hopkins. (n.d.). *Pneumonia*. Retrieved from Johns Hopkins: https://www.hopkinsmedicine.org/health/conditions-and-diseases/pneumonia

Mahendra, M., Nuchin, A., Kumar, R., Shreedhar, S., & Mahesh, P. A. (2021). Predictors of mortality in patients with severe COVID-19 pneumonia - a retrospective study. *Adv Respir Med, 89*(2), 135-144. doi:10.5603/ARM.a2021.0036.

Maragakis, L. (2021, October 21). *Coronavirus Second Wave, Third Wave and Beyond: What Causes a COVID Surge*. Retrieved from Johns Hopkins Medicine: https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/first-and-second-waves-of-coronavirus

Moore, S. (2021, September 28). *History of COVID-19*. Retrieved from News Medical: https://www.news-medical.net/health/History-of-COVID-19.aspx

Park, J. Y., Freer, R., Stevens, R., Soneji, N., & Jones, N. (2021, March 4). *The accuracy of chest CT in the diagnosis of COVID-19: An umbrella review*. Retrieved from THe Centre for Evidence-Based Medicine: https://www.cebm.net/covid-19/the-accuracy-of-chest-ct-in-the-diagnosis-of-covid-19-an-umbrella-review/

Pawar, M. (2021, November 2). *More than 144 million jobs lost, COVID caused a global employment shift*. Retrieved from Charles Sturt University: https://news.csu.edu.au/opinion/more-than-144-million-jobs-lost,-covid-caused-a-global-employment-shift

Wiersinga, W., Rhodes, A., Cheng, A., Peacock, S., & Prescott, H. (2020, July 10). Pathophysiology, Transmission, Diagnosis, and Treatment of Coronavirus Disease 2019 (COVID-19): A Review. *JAMA, 324*(8), 782-293. doi:10.1001/jama.2020.12839

World Health Organization. (2021). *The true death toll of COVID-19: estimating global excess mortality*. Retrieved from World Health Organization: https://www.who.int/data/stories/the-true-death-toll-of-covid-19-estimating-global-excess-mortality

Zheng, Z., Yao, Z., Wu, K., & Zheng, J. (2020, April 29). The diagnosis of pandemic coronavirus pneumonia: A review of radiology examination and laboratory test. *J Clin Virol, 128*(104396). doi:10.1016/j.jcv.2020.104396

# Appendix A

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Description** | **Source** | **Values** | **Type** |
| current\_status | Current status of the persons’ COVID-19 tests? | Case Report Form | Laboratory-confirmed case Probable case | String |
| cdc\_report\_dt | Date case was first reported to the CDC | Calculated | YYYY-MM-DD | Date |
| cdc\_case\_earliest\_dt | Date related to illness/ Date received by CDC | Calculated | YYYY-MM-DD | Date |
| sex | Sex | Case Report Form | Male Female Unknown Other Missing | String |
| age\_group | Age group | Calculated | 0 - 9 Years 10 - 19 Years 20 - 39 Years 40 - 49 Years 50 - 59 Years 60 - 69 Years 70 - 79 Years 80 + Years Missing NA | String |
| race | Race | Calculated | American Indian/Alaska Native Asian Black Multiple/Other Native Hawaiian/Other Pacific Islander White Unknown Missing NA | String |
| ethnicity | Ethnicity | Case Report Form | Hispanic Non-Hispanic Unknown Missing NA | String |
| county\_fips\_code | County FIPS Code | Calculated | NA if suppressed null if could not be matched | String |
| res\_county | County of residence | Case Report Form | NA if suppressed Missing if not available | String |
| res\_state | State of residence | Case Report Form | NA if suppressed Missing if not available | String |
| onset\_dt | Date of symptom onset | Case Report Form | YYYY-MM-DD | Date |
| pos\_spec\_dt | Date of first positive specimen collection | Case Report Form | YYYY-MM-DD | Date |
| hosp\_yn | Was the patient hospitalized? | Case Report Form | Yes No Unknown Missing | String |
| icu\_yn | Was the patient admitted to an intensive care unit (ICU)? | Case Report Form | Yes No Unknown Missing | String |
| death\_yn | Did the patient die as a result of this illness? | Case Report Form | Yes No Unknown Missing | String |
| hc\_work\_yn | Is the patient a health care worker in the United States? | Case Report Form | Yes No Unknown Missing | String |
| pna\_yn | Did the patient develop pneumonia? | Case Report Form | Yes No Unknown Missing | String |
| abxchest\_yn | Did the patient have an abnormal chest X-ray? | Case Report Form | Yes No Unknown Missing | String |
| acuterespdistress\_yn | Did the patient have acute respiratory distress syndrome? | Case Report Form | Yes No Unknown Missing | String |
| mechvent\_yn | Did the patient receive mechanical ventilation (MV)/intubation? | Case Report Form | Yes No Unknown Missing | String |
| fever\_yn | Fever >100.4F (38C) | Case Report Form | Yes No Unknown Missing | String |
| sfever\_yn | Subjective fever (felt feverish) | Case Report Form | Yes No Unknown Missing | String |
| chills\_yn | Chills | Case Report Form | Yes No Unknown Missing | String |
| myalgia\_yn | Muscle aches (myalgia) | Case Report Form | Yes No Unknown Missing | String |
| runnose\_yn | Runny nose (rhinorrhea) | Case Report Form | Yes No Unknown Missing | String |
| sthroat\_yn | Sore throat | Case Report Form | Yes No Unknown Missing | String |
| cough\_yn | Cough (new onset or worsening of chronic cough) | Case Report Form | Yes No Unknown Missing | String |
| sob\_yn | Shortness of breath (dyspnea) | Case Report Form | Yes No Unknown Missing | String |
| nauseavomit\_yn | Nausea or Vomiting | Case Report Form | Yes No Unknown Missing | String |
| headache\_yn | Headache | Case Report Form | Yes No Unknown Missing | String |
| abdom\_yn | Abdominal pain | Case Report Form | Yes No Unknown Missing | String |
| diarrhea\_yn | Diarrhea (≥3 loose/looser than normal stools/24hr period) | Case Report Form | Yes No Unknown Missing | String |
| medcond\_yn | Pre-existing medical conditions? | Case Report Form | Yes No Unknown Missing | String |