HIGH SCHOOL EDITION
Journal of Student Research
Volume 10 Issue 3 (2021)

Predicting the Severity of Coronavirus Cases Given Demographics and Pre-existing Conditions

Mahi Ravi[1], Jackie Li[2], Vineet Burugu[3], Sarvagya Goyal[4], Sireesh Pedapenki[5], Aditya Goel[6], Nandana Nambiar[7], Aadhya Subhash[8], Larry McMahan[#]

1Saratoga High School, Saratoga, CA, USA
2BASIS High School, San Jose, CA, USA
3Westwood High School, Austin, TX, USA
4Dougherty Valley High School, San Ramon, CA, USA
5Washington High School, Fremont, CA, USA
6 Quarry Lane High School, Dublin, CA, USA
7Mission San Jose High School, Fremont, CA, USA
8Marquette High School, Chesterfield, MO, USA
#Advisor

ABSTRACT

Beginning in early 2020, coronavirus disease (COVID-19) has rapidly spread all over the world. As of now there have been over 102.52 million confirmed cases along with 2.21 million deaths worldwide. Our objective is to create an algorithm that will predict the severity of a COVID-19 case for an individual based on demographic data such as race, age, gender, and location. Using international, national and local datasets, we collected the demographic data and organized them into their respective categories, namely age, race, gender, and location of origin. We then inputted this data into an algorithm that works around the principle of probability. Our algorithm uses such trends to develop a risk assessment and create a model. While compiling that data we noted common trends within the three demographics. Specifically, around the age thirty, cases were higher compared to other age ranges. The data collected and trends noted can be used to prioritize and prepare for patients that may be in critical danger, providing a chance for hospitals and vaccine distribution centers to preemptively address higher risk cases early.

Introduction

COVID-19, the disease caused by the SARS-CoV-2 virus, is a pandemic infecting millions globally. Since the first suspected case in Wuhan, China, there have been 98.55 million confirmed cases in the United States along with 2.11 million deaths worldwide, as of January 23, 2021. [1] Throughout the ongoing effort to combat COVID-19, Artificial intelligence (AI) has been implemented to provide effective solutions for a plethora of issues, such as creating accurate data models depicting the whereabouts of the viral spread, effectively diagnosing the coronavirus, as well as even aiding in the creation of a potential vaccine for the coronavirus. Additionally, AI has been used to predict the severity of COVID-19 cases. In one scenario, New York University researchers utilized AI and machine learning to determine the likelihood of a patient developing acute respiratory distress syndrome (ARDS) as a result of contracting COVID-19. Their research looked at biomarkers in the blood of patients and identified hospitalized individuals who may potentially suffer from longer lasting effects of COVID-19. Considering factors other than age and pre-existing conditions, prior studies have already achieved a seventy to eighty percent accuracy. With this information, hospital staff can effectively determine precedents to set for their patients, i.e. which patients should take a priority in treatment for the disease. AI research aims to create a “smart” algorithm that can determine the severity of COVID-19 in infected

individuals given certain pre-existing conditions, mainly demographic data. The factors that we consider analysing include age, gender, existence of pre-existing conditions or underlying diseases, and, if possible, biomarkers inside an individual's bloodstream.

Review of Literature

Severity Detection for the Coronavirus Disease 2019 Patients Using a Machine Learning Model Based on the Blood and Urine Tests Haochen Yao et al, 31 July, 2020 [2]:

This research focused on the detection of severity of COVID-19 in an individual by analysing 32 characteristics which were "significantly associated" with severity with the support of a vector machine (a model which finds a hyperdimensional plane in order to classify data), an ML model.

Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT:

Evaluation of the Diagnostic Accuracy Lin Li et al, 19 March 2020 [3]:

This research created a model which accurately found the occurrence of coronavirus independent from other lung diseases. The researchers used 4563 CT scans to create a deep learning model. They then took the CT scans and used RestNet50 to generate features from corresponding scans and combined the extracted features in a pool to present three probability scores for COVID-19, CAP, and non-pneumonia. However, this research article did not detect the exact areas in which the model uses to differentiate between viral pneumonia and COVID-19, because researchers were not able to supply community-acquired pneumonia into the model.

Artificial Intelligence Tool Predicts Which Patients with Pandemic Virus Will Develop Serious Respiratory Disease [4]:

This research study was one on biological predictors which determined which patients were more likely to develop ARDS (Acute Respiratory Distress Syndrome), a severe side effect of COVID-19. This study collected data on 53 patients and used data including demographic, laboratory, and radiological findings. The data was then implemented in an AI program which made decisions based on the input data, with the program becoming "smarter" as more data was included in its algorithms and models. Researchers found that levels of hemoglobin, ALT, and myalgia could determine the likelihood of a patient developing ARDS with an accuracy of 80%. However, some limitations of the study included a limited set of data to work with and the average patient being younger or middle-aged.

Methods

Data

In order to develop our algorithm we referenced twenty datasets on various demographic features. These datasets are located in the references section of this publication. We then used dataset data to generate histograms and analyze common trends based on the independent variable being tested (age, race, location), with mortality rates being plotted as the dependent variable.

There are various ways to measure the data. With the tables that we have, concrete number ranges are all different. For age data, CDC reports it with a 10 year age range (10-20, 20-30, etc.) while Alameda County reports it with a 9 year age range (31-40, 41-50, etc.). In that case, we went through our data tables and generated histograms of them.

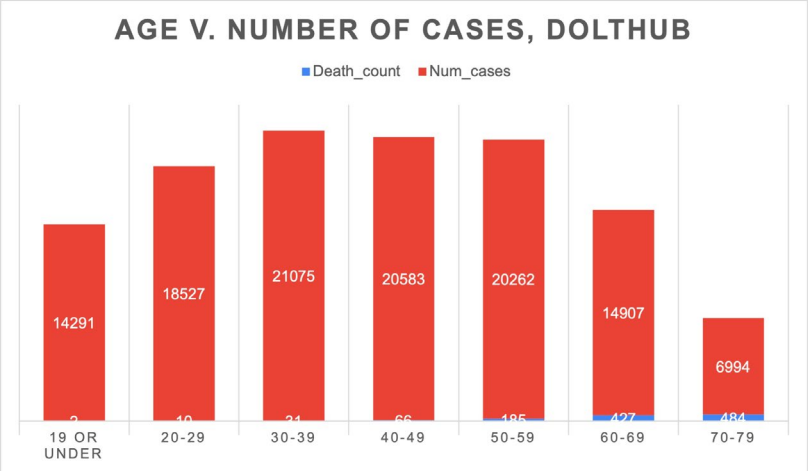

Figure 1.1

Figure 1.1 is for Age vs. Number of Cases for Dolthub's dataset combining data from Johns Hopkins University, China's Center for Disease Control, and the Singapore government [5]. The histogram above compares ages (in nine-year age ranges) to the number of cases and deaths contracted for each age range. The peak of this histogram is around the 30-39 year age range.

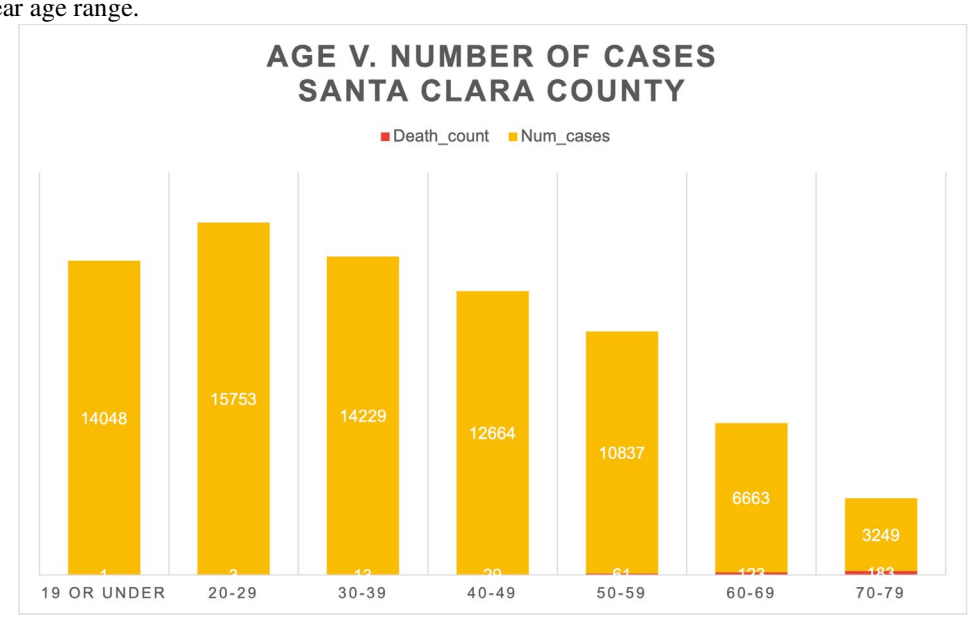

Figure 1.2

Figure 1.2 of Age vs Number of cases above is derived from the County of Santa Clara, California. Looking at the number of cases based on age (compared in nine- year ranges), the histogram displays a peak around the age range 18-30 [6]. There seems to be an increase in the number of cases within this age range among the data set set above.

As shown by the histograms for two different datasets above, even though the age ranges are different, certain ranges of ages have a peak or a decline in cases. In Figure 1.1, there is a peak at ages 30-39, whereas in Figure 1.2, the peak is seen at the 18-30 years range. From comparing the two peaks, we conclude that an individual is most likely to contract COVID-19 at age 30.

Algorithm

Our algorithm works around the principle of possibility. First, we take in risk factors, including demographic data such as race, gender, and location. With each type of factor we have a base risk approximate. We will also produce a

death risk based on the data. Thus the input to the algorithm was a person's own age, gender, and demographics, and the output was a person's risk of dying from SARS-CoV-2. Data is taken and each subsection of a factor, for example, if cases in California were ½ million and the US has 22 million cases, Californians would have a .5/22 (1/44) chance of dying. With the basic fractions for each subsection of risks set into place, basic dictionaries and if statements are created for each factor. Each factor has its own function. Each of the risks based on individual factors are listed, and then the factor risk fractions are averaged to provide a total risk of dying from COVID-19. To generate risk approximates, a basic standard deviation algorithm as shown below was used. In order to utilize standard deviation, numpy, an advanced Python mathematical library, had to be imported. An importance is given to certain factors by multiplying them by a value higher than one.

Results

Using our generated histograms we compare peaks to find an over-represented age range, gender, or race/ethnicity. So far, our results indicate that individuals ages 29-30 are most likely to contract COVID-19, and individuals ages 79-81 are most likely to die from catching COVID-19.

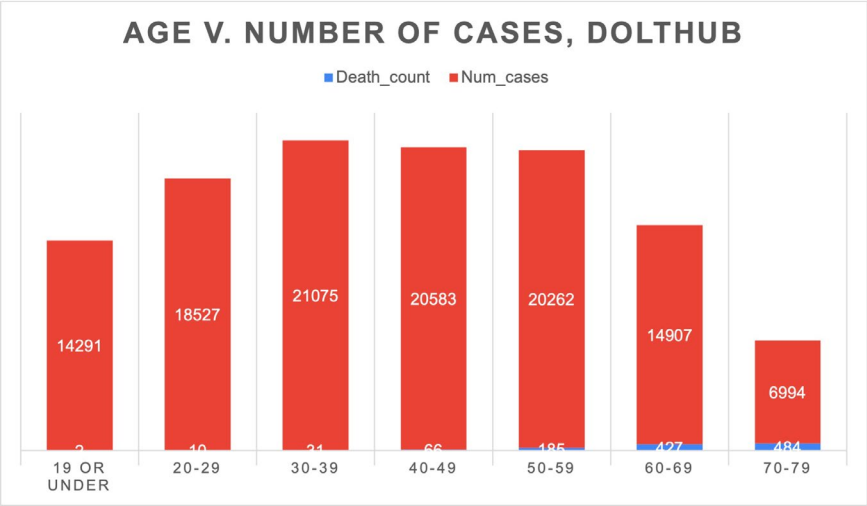

Figure 1.1

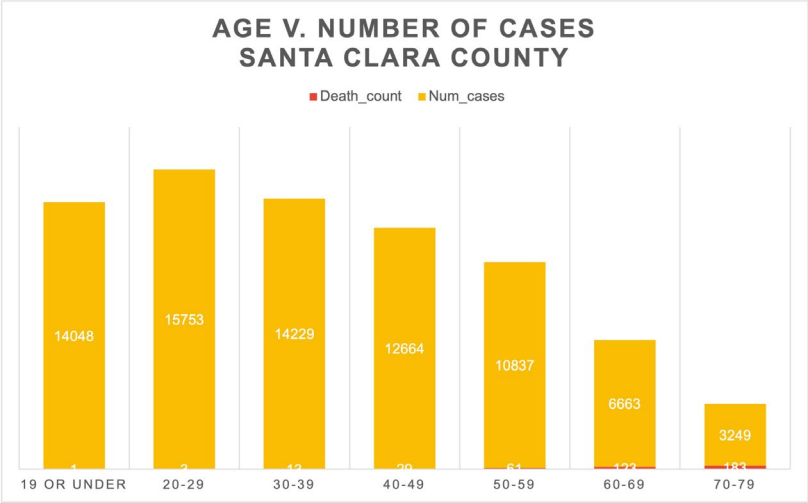

Figure 1.2

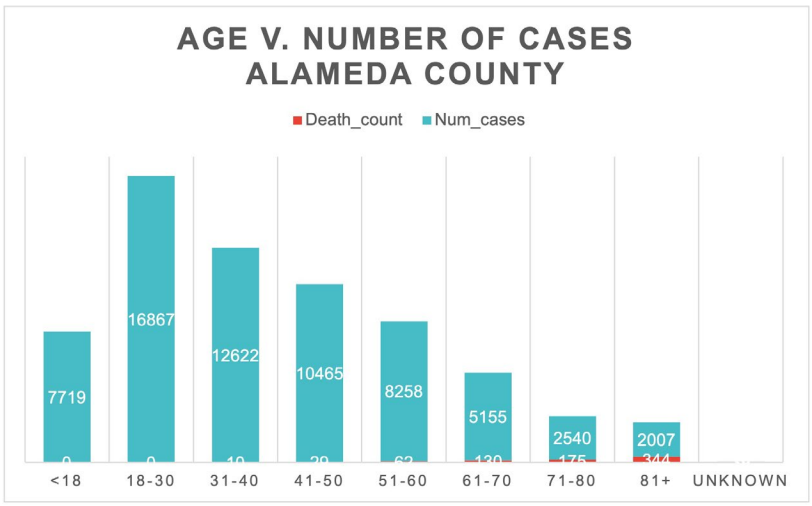

Figure 1.3

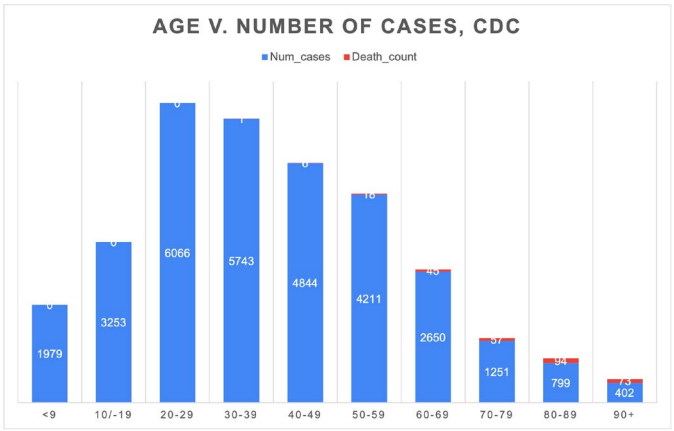

Figure 1.4 A comparison of Age v. Number of cases graphs from four data sets. As the graphs show, the peaks occur at around the 20-29 and the 30-39 age ranges.

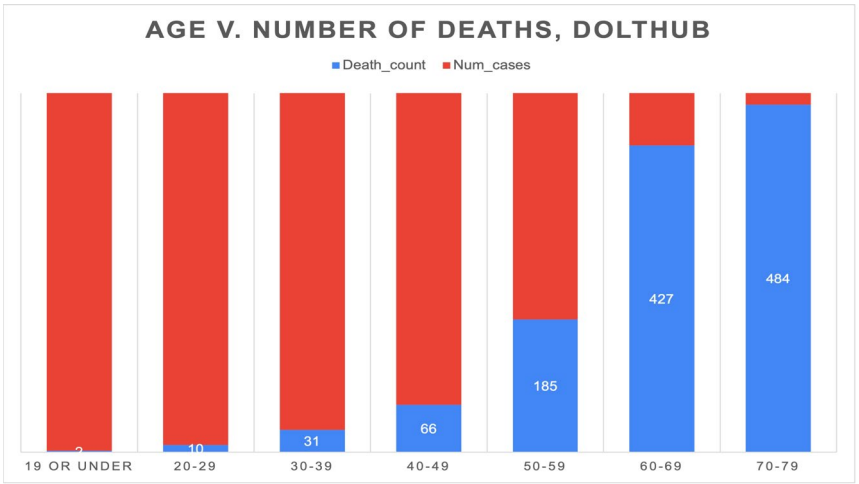

Figure 1.5

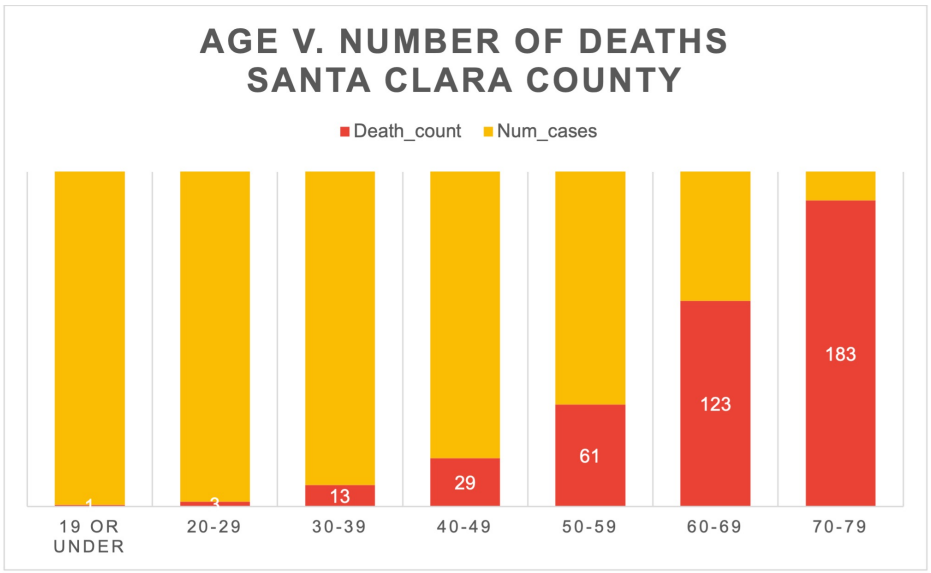

Figure 1.6

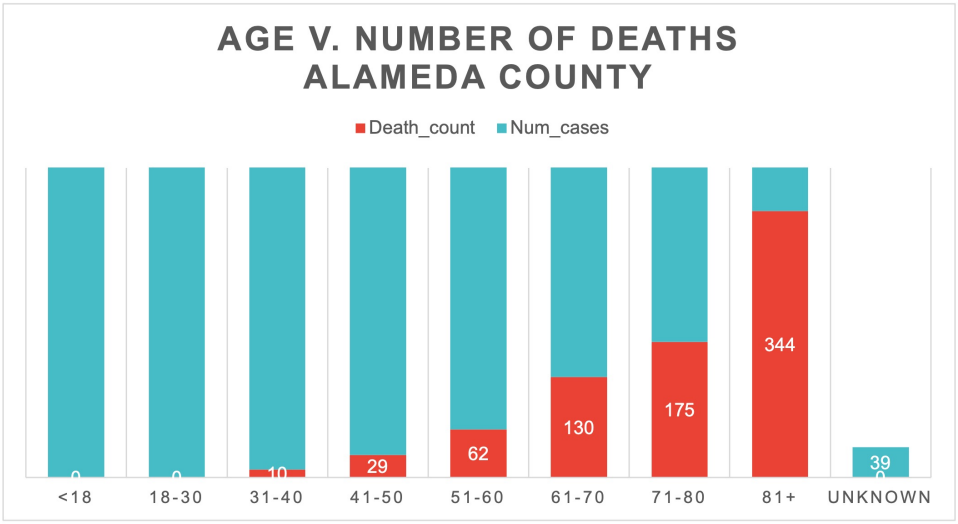

Figure 1.7

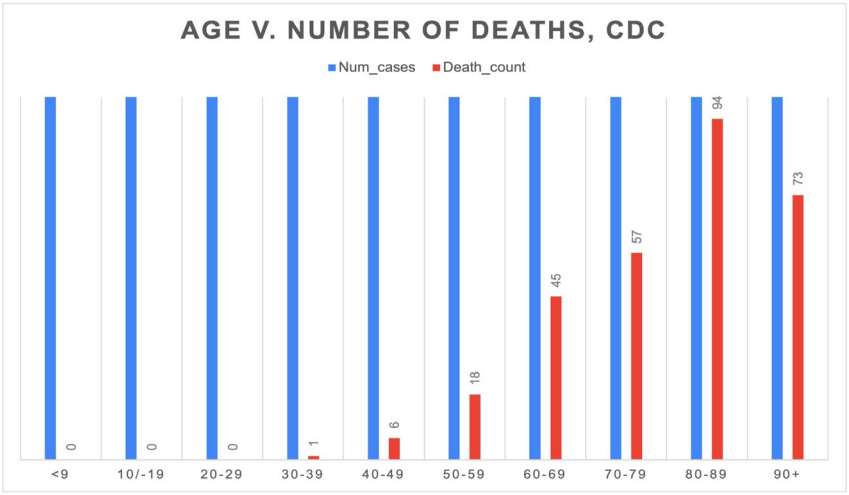

Figure 1.8

A comparison of Age v. Number of deaths graphs from four data sets. As the skewed graphs show, the peaks generally occur around the 70+ age ranges for all of the data sets.

In gender's case, the results which we got were more ambiguous - while in most of the datasets it was seen that females had a slightly higher contraction rate than males, the numbers were usually only off by around a few thousand. The death rate for males was also higher than those for females, again not by a substantial amount. For instance, in the Santa Clara County COVID-19 dataset, the amount of females who contracted COVID-19 was 40,438 people compared to 39,053 males, while the female death count was 380 compared to 445 males. [6]

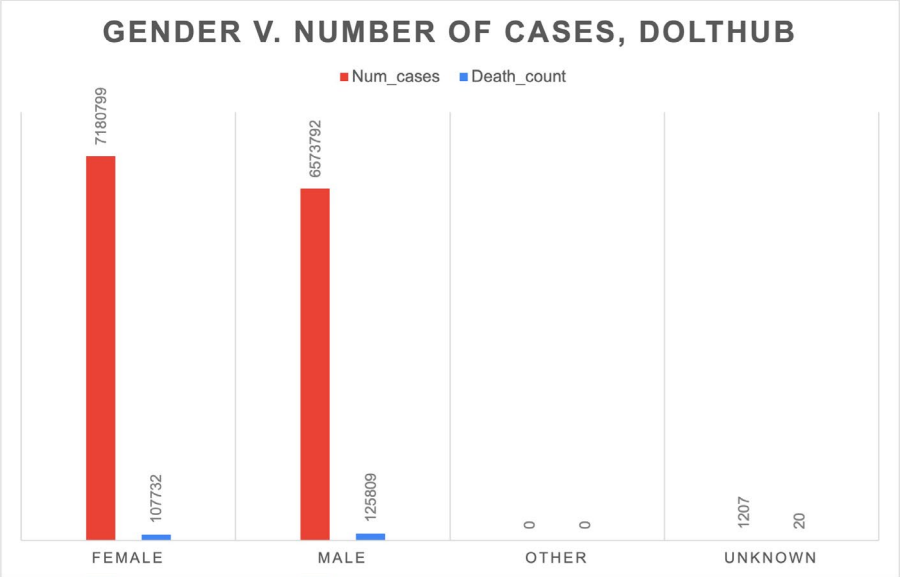

Figure 2.1

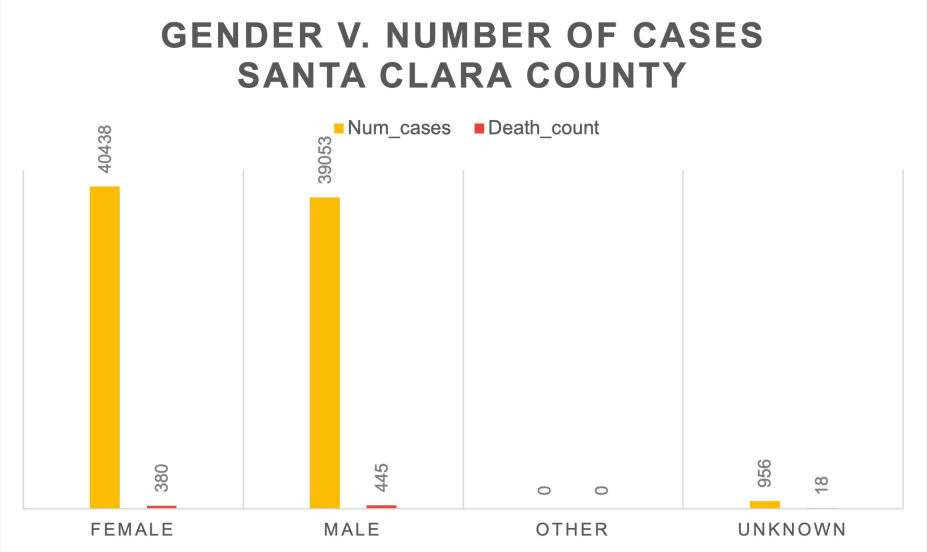

Figure 2.2

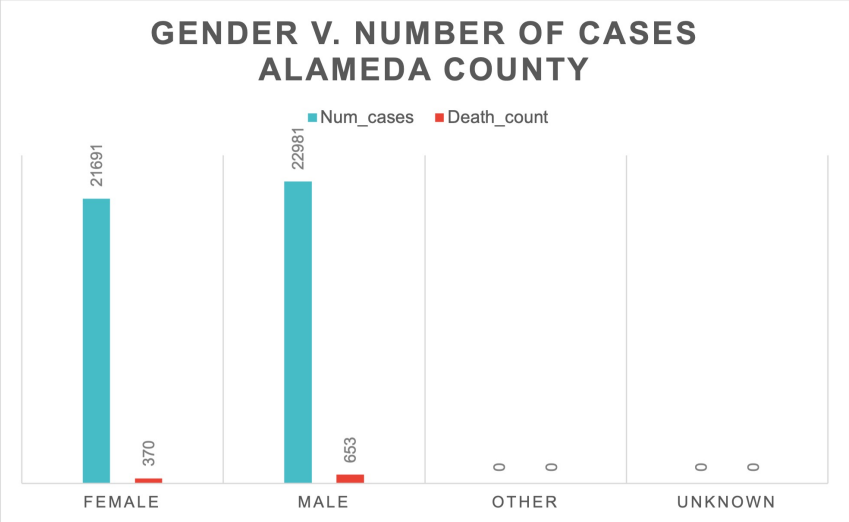

Figure 2.3

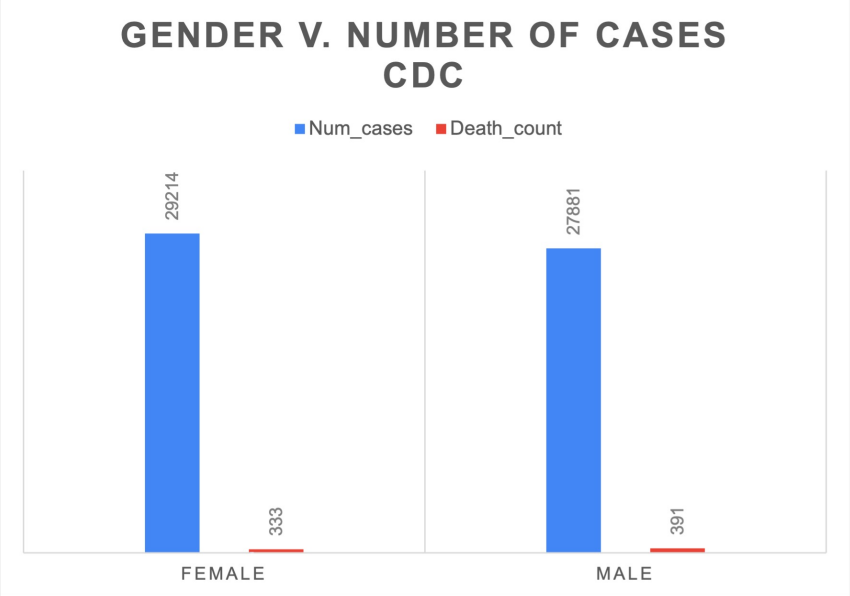

Figure 2.4 A comparison of Gender v Number of cases graphs. In three out of four of the graphs, females had slightly higher case numbers than males, although not enough to be a significant difference.

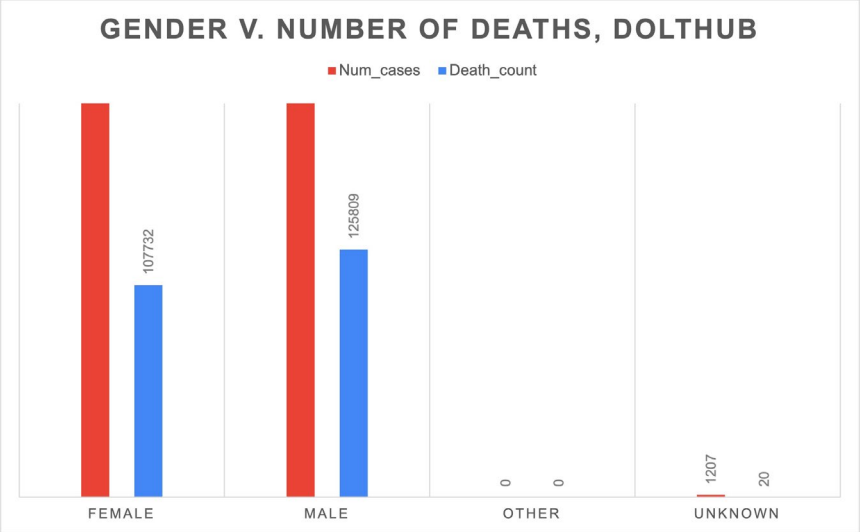

Figure 2.5

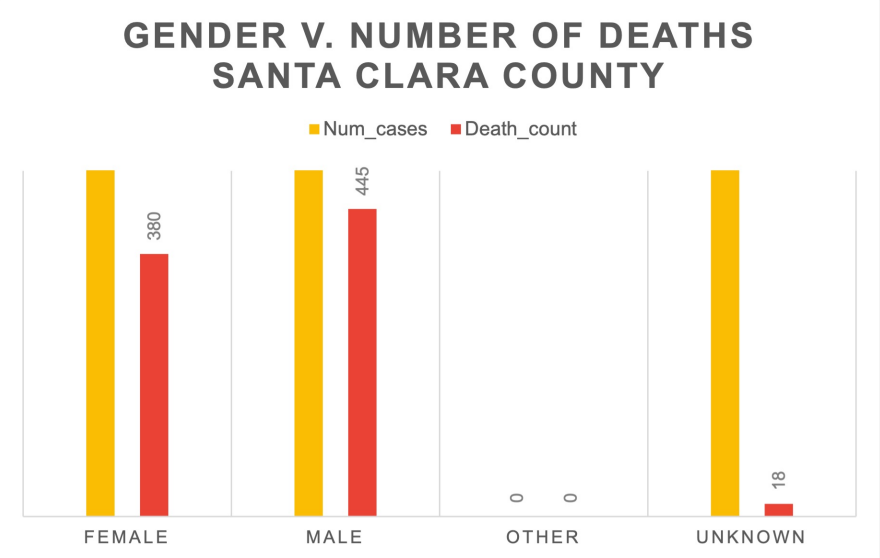

Figure 2.6

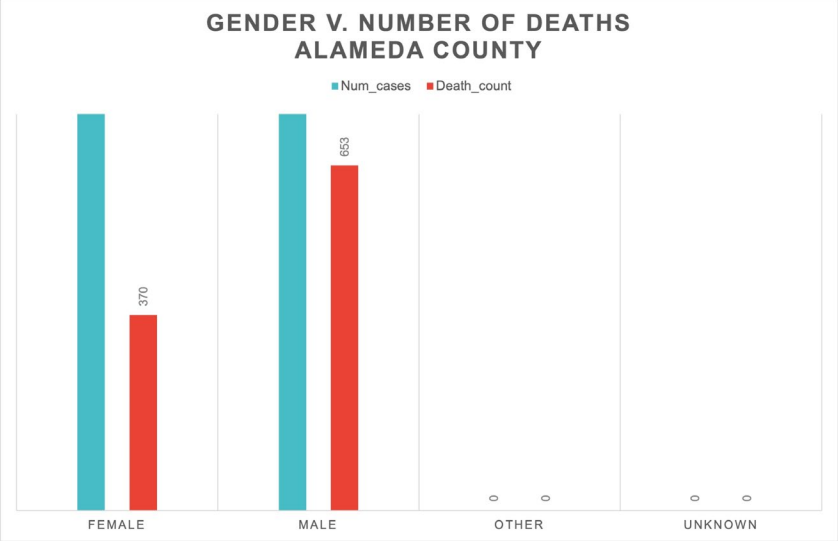

Figure 2.7

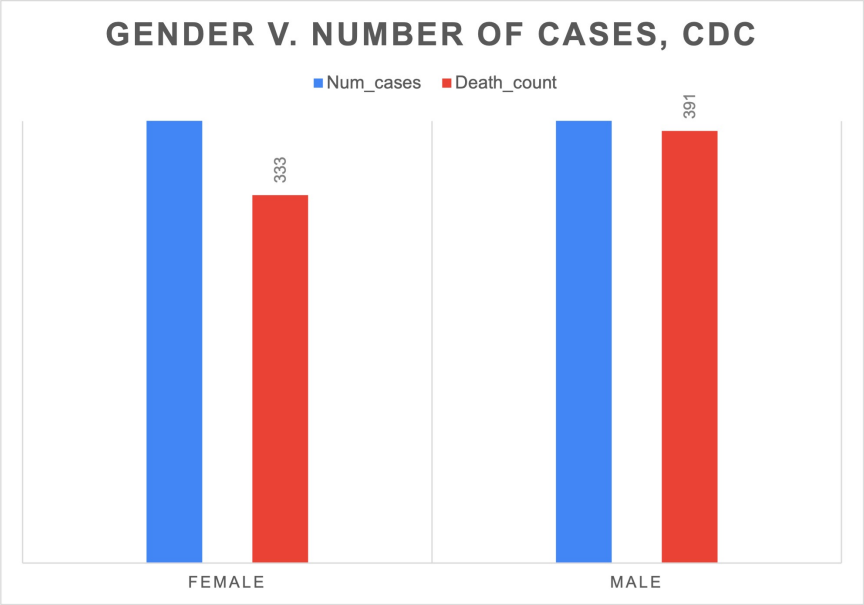

Figure 2.8 A comparison of Gender v. Number of deaths graphs. In all four data sets, males had higher death counts than females. In three out of four of the data sets, that difference was insignificant, but in the third data set (yellow), the number of deaths for males almost doubled the number of deaths for females.

In terms of race/ethnicity, a slight majority of the datasets analysed showed that people of Hispanic origin had the highest chance of contracting COVID-19, while the minority projected that for people of caucasian descent. A slight majority of the datasets predicted that white people were more susceptible to death after contracting COVID-19 against people of other racial origin.

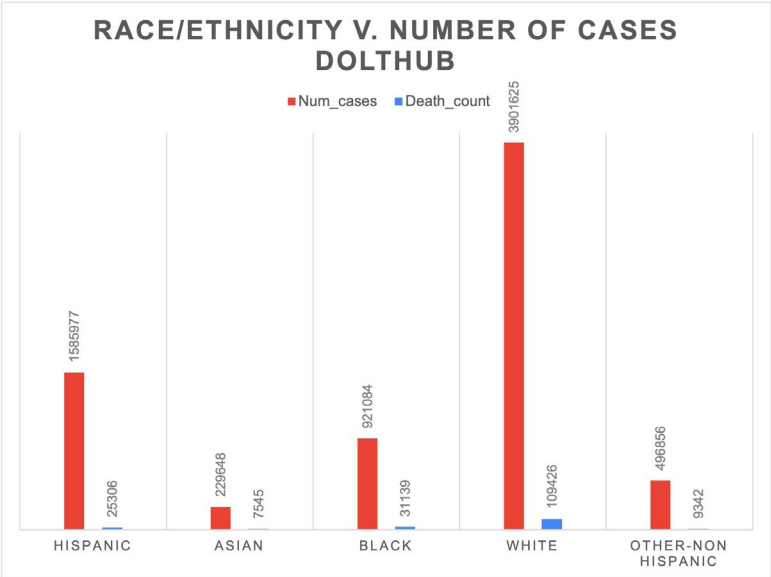

Figure 3.1

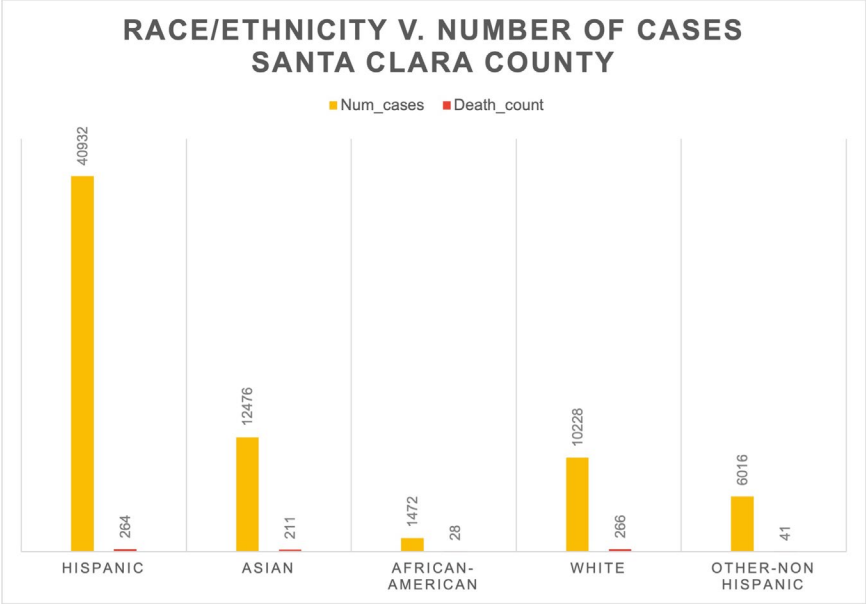

Figure 3.2

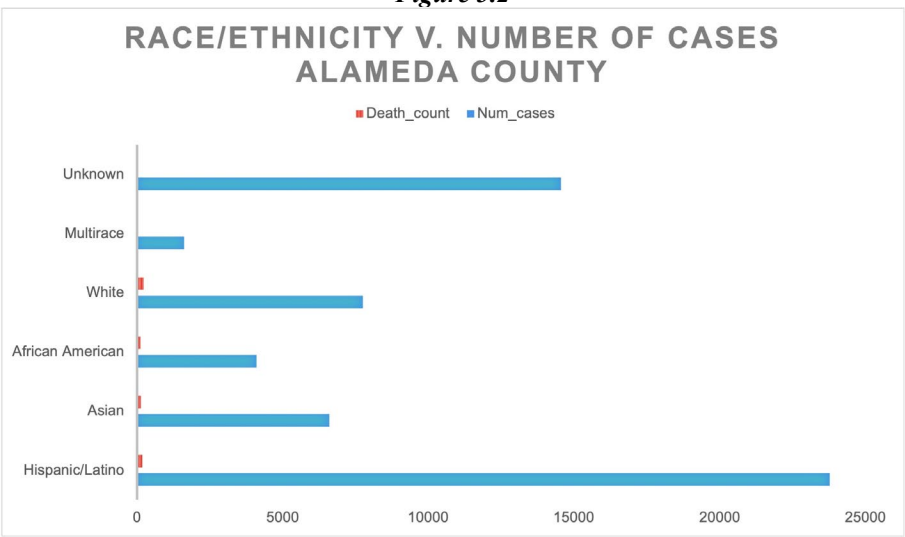

Figure 3.3 A comparison of three Race/ethnicity v. Number of cases graphs. Two out of three of the above data sets show that the Hispanic population has the highest number of cases, while one (dark blue) shows that the white population has the highest rate of contraction.

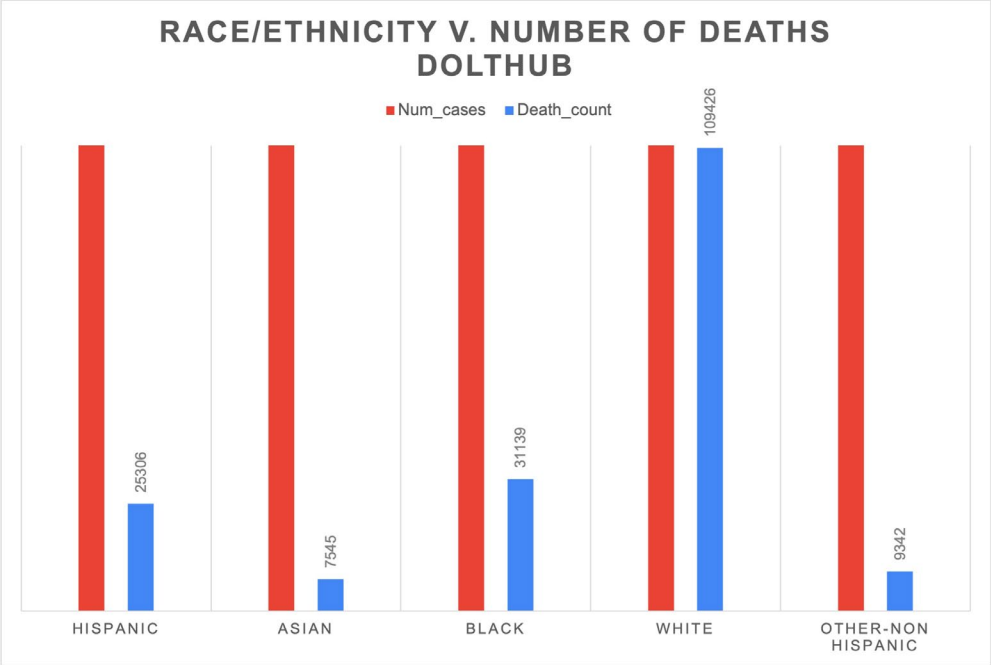

Figure 3.4

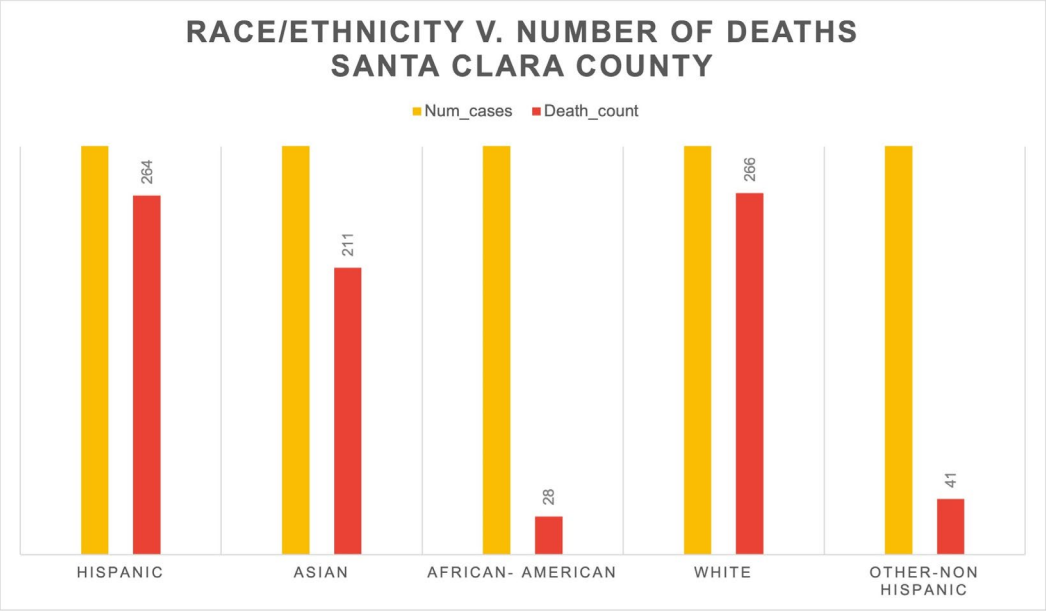

Figure 3.5

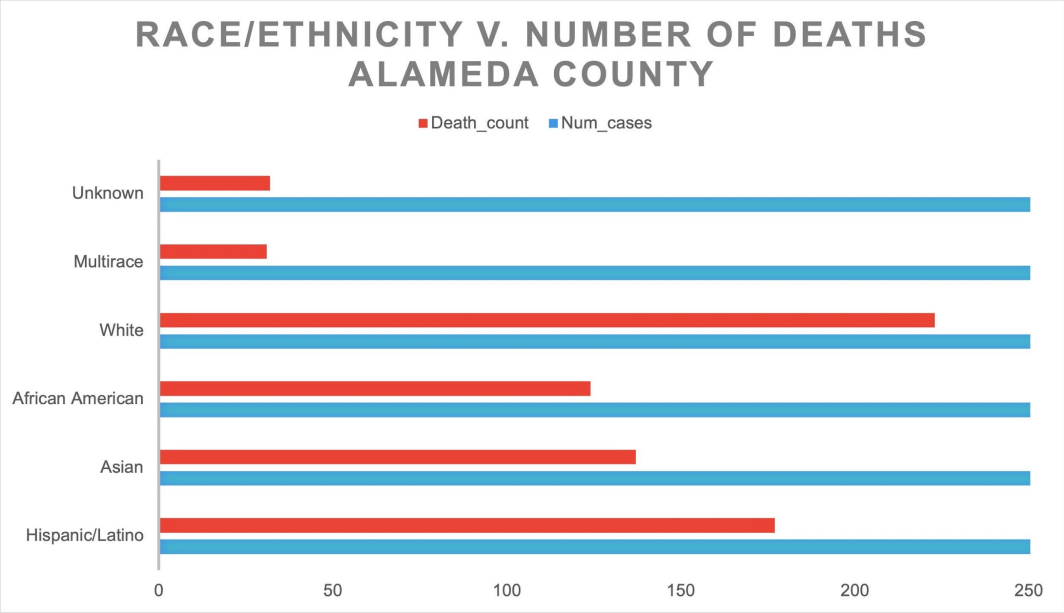

Figure 3.6 A comparison of the Race/ethnicity v. number of death graphs for three data sets. In all three data sets, the white population was shown to have tallied the most deaths; however, in the second data set (turquoise), the difference between the number of deaths of Hispanics and whites was only two.

Discussions and Conclusions

From the above analysis, we predict that age is the largest factor among the three factors analysed for the likelihood of catching and dying from COVID-19. Out of all three of the demographic features which we looked at, almost all of the datasets we analysed had conclusive results about the peak age for people to contract COVID-19, and even more so when it came to death. We were unable to find any association for race and COVID-19. For gender there is even less of a correlation for which gender is more likely to contract and perish from COVID-19. For race, there was a small correlation between a specific race being the most likely to contract and perish from COVID-19. As aforementioned, some of our reported data was only off by a few thousand people out of a 100,000 plus sample size. Therefore, the demographic which had the most affirmed results was age.

Future Research

For future improvements, one could consider specific biomarker data into our algorithm, in order to train it to consider more factors. With an algorithm that is expected to predict the severity of COVID-19 in an individual, a good follow up for that work would be an analysis of aftereffects of people who have contracted the virus.

Limitations

We do not have substantial data for the biological factors portion of our study. The only data which we have on individual biological factors is a dataset from China's National Center for Bioinformation. In the future, we can consider pursuing similar international datasets for similar data [9]. Receiving more data would require us to have the proper permissions to obtain due to issues of anonymity. There is also the limitation of where our datasets come from. As different places in the nation hold different demographics of people, some of the data (especially the race/ethnicity

data) were skewed towards diversity. For instance, two of the datasets that we used in comparison to the national datasets gathering information from all parts of the country were local county datasets in California's Bay Area region, where the demographics there differ from that of the rest of the country. This may have possibly caused some skew towards Hispanics being more prone to contracting COVID-19.

Acknowledgements

We would like to thank our advisor Dr. Larry McMahan for meaningful discussions.

References

Johns Hopkins University. "Johns Hopkins Coronavirus Resource Center." *Johns Hopkins Coronavirus Resource Center*, Johns Hopkins University & Medicine, 2020, coronavirus.jhu.edu/map.html.

Yao, Haochen, et al. "Severity Detection for the Coronavirus Disease 2019 (COVID-19) Patients Using a Machine Learning Model Based on the Blood and Urine Tests." Frontiers, Frontiers, 6 July 2020, www.frontiersin.org/articles/10.3389/fcell.2020.00683/full.

Li, Lin, et al. "Using Artificial Intelligence to Detect COVID-19 and Community-Acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy." Radiology, Radiological Society of North America, 19 Mar. 2020, pubs.rsna.org/doi/10.1148/radiol.2020200905.

Jiang, X., Coffee, M., Bari, A., Wang, J., Jiang, X. et al. (2020). Towards an Artificial Intelligence Framework for Data-Driven Prediction of Coronavirus Clinical Severity. CMC-Computers, Materials & Continua, 63(1), 537–551.

Dolthub.com, 2021, www.dolthub.com/repositories/dolthub/corona-virus. Accessed 17 Jan. 2021.

"Coronavirus (COVID-19) Data Dashboard - Novel Coronavirus (COVID-19) - County of Santa Clara." *Www.sccgov.org*, www.sccgov.org/sites/covid19/Pages/dashboard.aspx.

"Data | COVID-19 | Alameda County Public Health." *Covid-19.Acgov.org*, covid-19.acgov.org/data. Accessed 17 Jan. 2021.

CDC. "COVID-19 Cases, Deaths, and Trends in the US | CDC COVID Data Tracker." *Centers for Disease Control and Prevention*, 28 Mar. 2020, covid.cdc.gov/covid-data-tracker/#underlying-med-conditions.

"Clinical Records, China National Center for Bioinformation." *Clinical Records, CNCB* , 2019 Novel Coronavirus Resource (2019nCoVR), 20 Jan. 2021, bigd.big.ac.cn/ncov/clinic?lang=en.