Modeling Socioeconomic, Demographic, and Chronic Medical Conditions Correlated with COVID-19 Incidence

Jenna Chuan

Memorial High School 935 Echo Lane, Houston, Texas 77024, U.S.A.; jenna.ctx4@gmail.com

ABSTRACT: Since its initial discovery in November 2019, COVID-19 has caused a worldwide pandemic with substantial medical, economic, and social repercussions, with over 4.22 million deaths.[1] Although pandemics are inevitable, consequences from them can be minimized. Non-pharmaceutical interventions have been helpful in decreasing COVID-19 transmission, but they are broad and have severe educational and economic consequences. Understanding the factors that affect disease transmission is pivotal for non-pharmaceutical interventions to maximize their effectiveness while minimizing their repercussions. This study examined COVID-19 cases in each U.S. county to create a multiple linear regression model predicting incidence based on socioeconomic, demographic, and chronic disease factors. The model accuracy had a coefficient of determination of 71%, and the most impactful factors were population density, rural/urban status, social vulnerability, and multi-unit housing. Results from this study are necessary to help stakeholders decide how to allocate resources and target interventions to minimize the spread of the virus.

KEYWORDS: Translational Medical Sciences; Disease Prevention; virology; Computational Epidemiology; COVID-19.

Introduction

COVID-19, the coronavirus disease caused by the SARS-CoV-2 virus, has caused a widespread pandemic resulting in severe economic, social, and chronic medical consequences. After first being detected in Wuhan, China, in November 2019, it has spread to over 197 million people worldwide.[1] Within the United States, COVID-19 cases first began in January 2020, and by June 2021, there were over 33 million cases and nearly 600 thousand deaths.[2] Nearly 1 in every 10 Americans has received a positive COVID-19 test[3] and the number of Americans that have had COVID-19 is likely 2-3 times that number.[4]

Within the US, there have been striking differences in incidences across geographic areas. According to data collected by the New York Times (NYT) on June 29, 2021, of over 3,000 U.S. state counties, there have been 18 counties in which over 20% of the population has had COVID-19 and ten counties in which less than 2% of the population has had COVID-19.[2] This is a tenfold difference in COVID-19 incidence between these two different groups of counties. A reasonable follow-up question to this interesting find would be, what factors lead to differences in COVID-19 incidences across different areas within the US?

The COVID-19 pandemic has caused significant disruption to daily lives. Consequently, non-pharmaceutical interventions have been a primary means of reducing the spread of the virus. Many sectors have moved to remote workplaces to decrease workspace transmission rates,[5] and consumers have changed their shopping habits from in-person to online platforms.[6] Many schools moved to remote learning or hybrid (combination of both virtual and in-person) plans to minimize classroom crowding.[7] The pandemic has even caused trends of migration away from metropolitan areas. A recent study showed a 3.92% increase in the movement of people across the United States, most moving away from large cities such as New York, Chicago, San Francisco, and Los Angeles.[8] Many of these behavioral changes are based on assumptions and extrapolation from previous knowledge of virus transmission. However, these have adverse effects, such as learning loss for children not in school,[9] and economic consequences for closed businesses.[10] More research is needed so that the impact of these non-pharmaceutical interventions (on reducing the spread of the virus) can be effectively maximized while minimizing the pandemic's economic, educational, and social costs.

Although previous papers have been published regarding demographic, socioeconomic, and spatial factors that correlate with COVID-19 incidence, the great majority of these papers used data from the first six months of the pandemic and thus under-represent the current situation. This paper updates these studies by analyzing 16 months of data to determine which demographic, socioeconomic, and medical factors continue to correlate with COVID-19 incidence strongly. The aspects of focus include demographic characteristics (age, ethnicity), socioeconomic factors (the Social Vulnerability Index, poverty, multi-unit or single-unit homes, rural versus urban populations, population density, average household size), percent of the population staying at home (from pre-pandemic compared to the pandemic time period), and chronic health conditions (obesity and diabetes).

Furthermore, this study combines the factors that have shown a correlation in previous papers into a model that explains the varying incidence of cumulative COVID-19 incidence within U.S. counties. Understanding this model and the components that lead to increased COVID-19 incidence are critical for di-

DOI: 10.36838/v5i1.12

recting policies and decisions for the future of this pandemic. It can also help institutions decide how to allocate resources and target interventions. Post-pandemic, these results are still crucial in planning for future pandemics.

Literature Review:

Previous research has been done on demographic factors that significantly impact COVID-19 incidences, such as age, sex, and ethnicity. Multiple studies have found that underrepresented racial and ethnic groups were two to three times more likely to contract this virus.[11,12] In August 2020, Moore *et al*. examined U.S. counties with sufficient data on COVID-19 infection and race, finding that Hispanic, Black, American Indian, Asian, and Pacific Islander races were two to four times more disproportionately affected by COVID disease.[11] Mahajan and Larkins-Pettigrew also found a positive correlation between counties with increased percentages of African Americans and Asian Americans and increased COVID-19 incidence and mortality rates in May 2020.[12]

Many studies combined demographic and socioeconomic factors in their COVID-19 risk factor analysis. Rozenfeld *et al*. studied over 34,000 patients who took COVID-19 tests in the Providence Health System in April 2020 and found that the COVID-19 positive patients were more likely to be male, of older age, of Asian or Black race, Latino ethnicity, non-English language, or residing in a neighborhood with financial insecurity, low air quality, housing insecurity, transportation insecurity, or in a senior living community.[11] Gunness and Milheiser compared communities in Queens, NY, in April 2020 and found overcrowded communities with lower education levels, less healthcare access, and more chronic disease rates were more severely affected by the COVID-19 outbreak.[13]

Poverty, income disparity, and social vulnerability, therefore, seem to play a vital role in COVID-19 incidence in communities. The Centers for Disease Control and Prevention (CDC) Social Vulnerability Index (SVI) measures a community's ability to prevent human suffering and financial loss in disasters. This index is subdivided into socioeconomic status, household composition and disability, minority status and language, and housing type and transportation in communities and ranges from 0-1 (1 representing an extremely vulnerable community).[14] In May 2020, Karaye and Horney used the SVI in U.S. counties and found that a percentile increase in SVI resulted in a 16% increase in COVID-19 case counts.[15] In July 2020, Karmakar *et al.* also studied SVI and found a 0.1 point increase (or 10% increase on the 0-1 scale) in SVI resulted in a 14.3% increase in the COVID-19 incidence rate.[16] Long-standing inequality in healthcare access for vulnerable communities contributes to increased incidence.

Several studies demonstrated increased COVID-19 transmission in urban settings. Karim and Chen divided all U.S. counties into metropolitan, micropolitan, and rural. They collected data up to June 2020, showing that most COVID-19 cases and deaths were in urban areas.[17] Hamidi *et al*. also noted in May 2020 that metropolitan areas correlated with increased COVID-19 incidence, but population density did not correlate with increased COVID-19 incidence after controlling for urban populations.[18] Similarly, in June 2020, Cromer *et al*. noted that patients who tested positive for COVID-19 were more likely to live in census tracts with higher rates of household crowding, a higher percentage of multifamily homes, and lower rates of high school completion.[19]

Only a few studies focused on movement and COVID-19 incidence. Badr *et al*. examined anonymous mobile phone data before April 2020 and determined that decreased movement resulted in decreased COVID-19 transmission 9-12 days later.[20] Gatalo *et al*. continued this analysis with data into July 2020. Although early data showed a correlation between mobility data and transmission, later time frames showed weaker correlations, which suggested other non-pharmaceutical measures had a more important role as time passed.[21]

As noted, many of these studies were conducted with data from the first six months of the pandemic; however, dynamics may have changed since then because the pandemic has persisted for about 16 months at the time of this study. Academic literature has identified many factors that may correlate with increased COVID-19 incidences across counties. All of these studied factors are combined and analyzed with 16 months of data to build a machine learning model (multiple linear regression) for predicting COVID-19 incidence.

Methods

Data Collection:

Data for each of the following factors was collected from publicly available sources.

COVID-19 Incidence per U.S. County:

COVID-19 cumulative U.S. county incidence data was gathered from the NYT GitHub website, a publicly available repository of U.S. COVID-19 data.[2] Each county was identified by unique Federal Information Processing Standards (FIPS) codes. Data related to the U.S. territories or data not associated with a FIPS code (for example, cases associated with a state but unknown county) comprised 3.28% of the data set and were eliminated. Cumulative COVID-19 cases from January 21, 2020, to June 29, 2021, were used in the study.

Percentage of Multi-unit and Single-unit Housing:

Data was collected from the 2019 American Community Survey (ACS) on the United States Census Bureau website.[22] To determine percentages from the data, single-family homes (detached and attached) per county were summed and divided by the total housing units in that county. Likewise, to determine the percentage of multifamily dwellings, structures with greater than 20 units were also added per county and divided by the total housing units in that county. FIPS codes were used to identify each county.

Population Density, Average Household Size, and Poverty:

U.S. Census Bureau published data on April 7, 2020, that provided average household size per county and population density based on people per square kilometer.[23] The percentage of residents in poverty was available through the 2019 US Census Small Area Income and Poverty Estimates (SAIPE) data.[24] The SAIPE data is based on income tax returns, U.S. census estimates, and Supplemental Nutrition Assistance Program (SNAP) benefits.[25]

DOI: 10.36838/v5i1.12

Rural-Urban Continuum Codes:

The Rural-Urban Continuum Codes (RUCC) were collected from the U.S. Department of Agriculture Economic Research Service (ERS) website.[26] The ERS divided counties into nine classifications, with lower numbers signifying a more urban area and higher numbers corresponding to a more rural area. The first three RUCC codes were divided by population size into metropolitan areas, and the last six codes were non-metropolitan areas classified by proximity to a metro area and population.

Population Mobility:

The U.S. Department of Transportation Bureau of Transportation Statistics.[27] provides mobility statistics for each U.S. county. The data is produced from a survey of an anonymous national panel of mobile devices.[28] The data also defines trips as movements that include a stay of longer than ten minutes at an anonymized location away from home. Every day, the county population is divided into two parts: the percentage staying at home and the percentage not staying at home.

The data can be averaged over a period of time. This study collected data from April 7, 2020, to January 10, 2021. The data start date was chosen because most mandatory stay-at-home orders were initiated between mid-March and April 7, 2020,[29] while the end date is representative of the peak of the pandemic (i.e., the date with the most daily cases).[1] Daily cases declined after the peak of January 10, 2021, so it was inferred that before this date would be when the importance of social distancing would be most crucial. The percentage of the population that stayed at home over the pandemic time period in each county was calculated as the average number of people that stayed at home each day from April 7, 2020, to January 10, 2021, divided by the total population monitored in that county.

In addition, the pandemic time period (April 7, 2020, to January 10, 2021) was compared to the same time period one year prior (April 7, 2019, to January 10, 2020) to establish a baseline of comparison for the movement of individuals on a county level. The change in movement was calculated as the number of people that stayed home during the pandemic period, subtracted by the number of people that stayed home during the baseline period divided by the baseline period again.

Obesity and Diabetes Rates:

The original data used to estimate obesity and diabetes rates in each county was derived from the Behavioral Risk Factor Surveillance System (BRFSS) through the CDC.[30] The most recent BRFSS survey data was from 2019 and reported the percentage of the population diagnosed with diabetes and the percentage of the population with obesity (defined as a body mass index greater than 30).

Social Vulnerability Index (SVI), Percentage of Population Greater than 65 years of age, and Percentage of Minority Population:

The Social Vulnerability Index (SVI), as mentioned previously, is a measure of a community's susceptibility to loss in a disaster or anthropogenic event. Each county is given a score between 0 and 1, with higher scores indicating higher levels of vulnerability. SVI is calculated from four key factors: socioeconomic status, household composition and disability, minority status and language, and housing type and transportation. The data is available from 2018 on the CDC website.[31] This SVI database also included estimations for the number of people in each county over 65 years old, minorities, and the total population.

Data Preprocessing:

All analyses in this study were done with Python 3.7.11.[32] A Jupyter notebook[33] using Google Colab workspaces[34] was utilized as well. Data preprocessing, descriptive statistics, data visualizations, data transformations, linear models, and multiple linear regression were all performed for the data set used in this study.

Before fitting the linear regression model, data were preprocessed in two steps: removing outliers and feature scaling (standardizing normally distributed variables and normalizing variables with non-normal distributions). Initially, outliers were considered as data points greater than three standard deviations away from the mean. Still, parameters were modified to 2.5 standard deviations to account for the large number of outliers. The removed points consisted of 1.24% of the data. After adjusting for outliers, the dataset used in this study had 2,662 data points and 12 features.

For feature scaling, the data was divided into normally distributed variables and non-normally distributed variables. The normally distributed predictor variables were standardized using StandardScaler from the sci-kit-learn library in Python 3.7.11. This shifted the mean to zero and scaled the variables to unit variance. The non-normally distributed predictor variables were normalized using MinMaxScaler from the sci-kit-learn library in Python 3.7.11. This adjusts all values in the non-normal features to a 0-1 range. Both of these two feature scaling techniques were necessary to have a standard range for all the variables, as their values encompassed a wide range (for instance, percentages range from 0 - 100, SVI ranges from 0 - 1, rural/urban status ranges from 1 – 9, etc.).

Linear Regression:

To create the multiple linear regression model, the data was divided into 80% for training and 20% for testing, also known as the 80-20 split. These ratios were based upon the Pareto Principle, which states that 20% of causes form 80% of effects.[35] The linear model was then fit to the training data and analyzed using sklearn's ordinary least squares linear regression summary. Statistically insignificant features ($p<0.05$) were removed iteratively until a parsimonious linear model was fit. The outcome (COVID cases per county) was noted to have a non-normal distribution, so it was transformed into a normal distribution.

For the results of a regression model to be accurate, the following five assumptions must be met: linearity, normal distribution of error terms, no multicollinearity among predictors, homoscedasticity, and no relationship between residuals and variables. Linearity is the ability to draw a linear relationship between the predictor and outcome variables. Next, error terms must be normally distributed, which can be checked using an Anderson-Darling test or quantile-quantile plot. Multicol linearity among predictor variables cannot be present; if there is, it is challenging to make interpretations from the model. Homoscedasticity is associated with no relationship or patterns between residuals and variables. Violation of this implies the lack of a linear relationship. After trying out different transformation techniques, including log transforming the outcome variable (elaborated in the results section), the Box-Cox method was the most optimal for this study.

Results

In the preliminary analysis, the model with COVID-19 cases as the outcome variable yielded higher accuracy than that with COVID-19 deaths as the outcome. For this reason, data was focused on COVID-19 cases. Three metrics were utilized to evaluate the performance of the model: the coefficient of determination measured how much of the variance in outcome was explained by the predictors, the mean squared error measured the average magnitude of errors, and the root mean squared error also measured magnitude, but was more tolerant of outliers. The metrics after each model iteration is summarized in Table 1. After data preprocessing, the first model fit did not have high accuracy because the data had not yet been narrowed down or transformed. It concluded in a coefficient of determination of 52%, mean absolute error of 4916.28, and mean squared error of 9536.76. After fitting the ordinary least squares regression model, features with a p-value <0.05 were considered statistically insignificant to the model. As a result, two components (average household size and percentage of single-family homes) were removed from the model one at a time to achieve a parsimonious model. This new model had a coefficient of determination of 58%, a mean absolute error of 5398.93, and a mean squared error of 10344.53. However, the linear model assumptions were violated because the outcome variable (COVID-19 cases) was not normally distributed.

Table 1: Coefficient of determination, mean absolute error and mean squared error of the experimental models.

	Coefficient of determination	Mean absolute error	Mean squared error
First model after data preprocessing	52%	4916.28	9536.76
Model after fitting OLS regression	58%	5398.93	10344.53
Model after log transformation	58%	5391.53	10382.79
Final model after Box Cox method	71%	0.46	0.60

To rectify this, data transformation was executed on the outcome by log transforming it. After fitting the model to the dataset with the log-transformed outcome, the coefficient of determination was 58%, the mean absolute error was 5391.53, and the mean squared error was 10382.79. Although the model's accuracy improved, it still needed modification because it did not pass all of the assumptions of a linear model. The model violated linearity and homoscedasticity (Figures 1 and 2). Because there are numerous ways to transform an outcome from a non-normal variable to a normal variable (log transformation, square root transformation, inverse transformation, etc.), there were no guarantees as to which exact transformation would work for the outcome. Consequently, a Box-Cox method was chosen to transform the outcome variable. The Box-Cox method does this by finding the best lambda value to transform the outcome variable to follow a normal distribution optimally. This function returns a lambda value (For instance, lambda = 2 for a square root transformation; lambda = -1 for an inverse transformation; etc.).

Following the Box-Cox method, the optimal lambda value for COVID-19 cases was -0.02187. Figure 3 shows the distribution of the outcome variable before and after the box cox transformation. Fitting the model with this newly transformed outcome, all five linear model assumptions were met. Figure 4 displays a plot of predicted and actual values, following the assumption of linearity. Figure 5 is a quantile-quantile (QQ) plot, which tests for the normality of residuals. Figure 6 is a residual graph with equal variance throughout, following homoscedasticity of error terms. In Table 2, there are no significant variance inflation values, proving no multicollinearity was present in my variables. Finally, the model also passed the Durbin-Watson test, which checked for independent errors assumption. The test resulted in 2.11, which is in the normal range of 1.5 - 2.5. Thus, there was no multicollinearity present in my variables.

After ensuring the validity of all five tests, the final model had a coefficient of determination of 71%, a mean absolute error of 0.46, and a mean squared error of 0.60. The details for the final model are available in Table 3. Now that all five assumptions have been met, results could be drawn from the model.

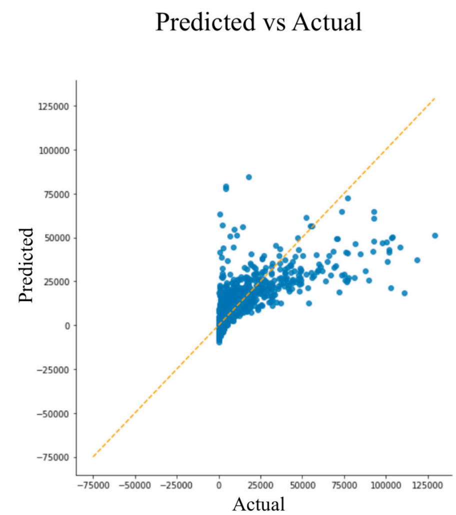

Figure 1: A residual graph of predicted values and actual values. The log-transformed model violates linearity, as the points are patterned and not clustered evenly along the orange line.

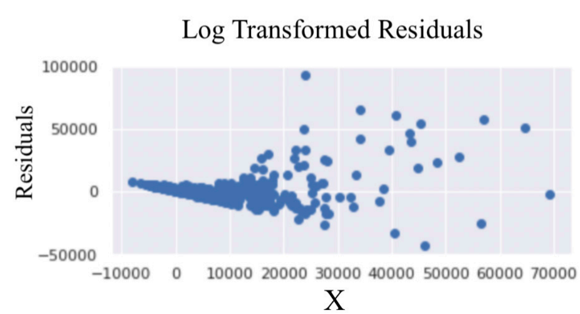

Figure 2: A graph of the log-transformed residuals, which violates homoscedasticity. As X increases, the variance of the residual increases, creating a patterned funnel shape that is inconsistent.

DOI: 10.36838/v5i1.12

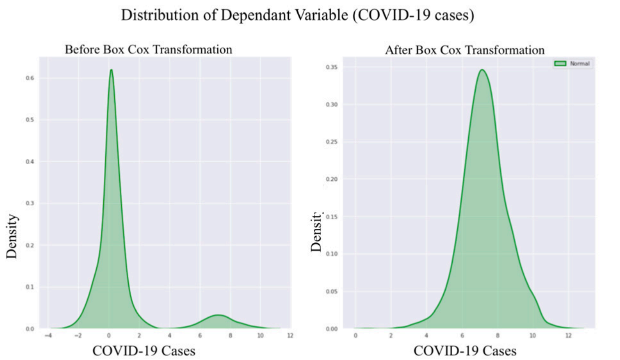

Figure 3: Distribution of dependent variable (COVID-19 cases) before and after box cox transformation. The image on the right shows the COVID-19 cases transformed into a normal distribution.

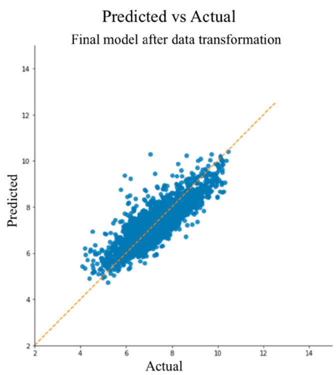

Figure 4: Residual graph (plot of predicted values vs. actual values). The figure shows that the linearity assumption is well satisfied with a strong linear relationship.

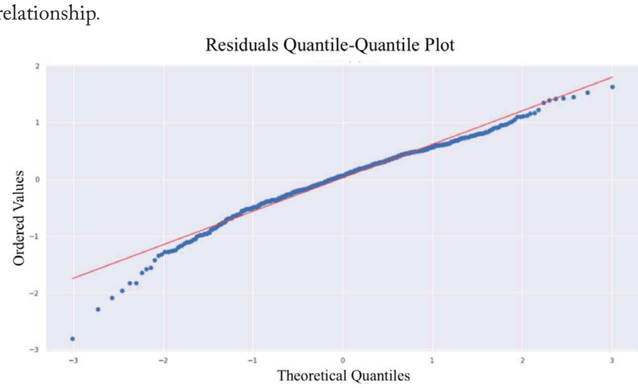

Figure 5: This quantile-quantile (QQ) plot indicates the normality of the residuals, as the points closely lie along the orange line.

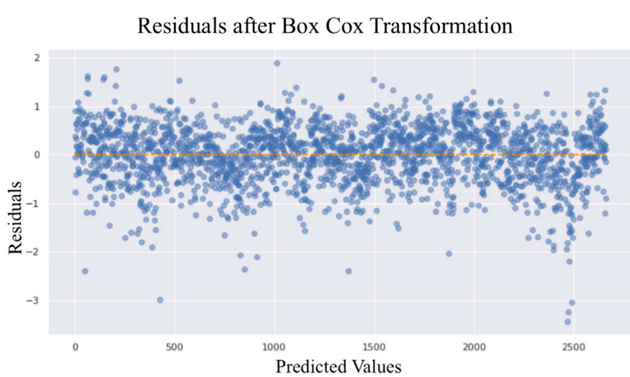

Figure 6: The model qualifies as Homoscedasticity of Error Terms. This model graphs variance in residuals, and there is no pattern.

Discussion

Variance Inflation Factors of Model Variables

	Model Variables	VIF
1	Population Density	1.89
2	% of Housing Units with > 20 Units	3.20
3	% Diabetes	2.76
4	% Obesity	2.01
5	% Poverty	2.12
6	% of Population Over 65	1.85
7	% Minority	5.08
8	Social Vulnerability Index	9.28
9	Rural Urban Classification Status	4.31
10	Change in % Population at Home	1.70

Table 3: The greater absolute value of the model coefficient determines the weight the factor is given in the model. This chart includes factors that had a significant p-value (p<0.05), the remaining two factors not included (percentage of single-family homes and average household size, with p-values of 0.356 and 0.232, respectively) were eliminated because their p-values were greater than 0.05.

Factors that affect number of COVID-19 cases	*Model Coefficient*	*Standard Error of Model Coefficient*	*P-Value <0.05 considered significant*
Percentage of population over 65 years of age	-0.1977	0.016	0.000
Change in percentage of population staying at home each day during pandemic time (4/7/20 to 1/10/21) compared to prepandemic time (4/7/19 to 1/10/20)	0.2261	0.015	0.000
Social Vulnerability Index	1.0200	0.082	0.000
Percentage of homes in multiunit buildings with greater than 20 units	2.0076	0.085	0.000
Rural versus urban code (urban codes lower)	-1.1441	0.051	0.000
Population Density	1.3430	0.174	0.000
Percentage of population with diabetes	0.0895	0.019	0.000
Percentage of population in poverty	-0.1167	0.021	0.000
Percentage of population with a BMI>30	0.0427	0.016	0.009
Percentage of population that is of minority ethnicity	-0.5654	0.073	0.000

The multi-linear regression model was able to predict COVID-19 cases successfully using the ten factors listed in Table 3. Two variables (average household size and percent of housing units that were single-family homes) were statistically insignificant (p>0.05) and subsequently removed from the model. The model had a coefficient of determination of 0.71, indicating reasonable accuracy. The four factors that had the most influence on the model included the percentage of multi-unit housing, population density, rural-urban classification status, and Social Vulnerability Index (SVI).

Percentage of Multi-Unit Housing:

Multi-unit housing is the percentage of housing units in buildings with more than twenty units per building, correlated with increased COVID-19 incidence. These results concurred with Cromer *et al*., who found that high household crowding and occupancy contributed to higher COVID-19 incidence.[19]

Population Density:

The population density was calculated by dividing the total county population by the county's area in square kilometers. In the model coefficients, population density contributed positively, showing increased COVID-19 cases in denser counties. Results from some previous studies are consistent with this, while other studies did not find any significant contribution of population density to COVID-19 incidences.[18,36] For instance, Hamidi *et al*. showed that population density did not correlate positively with increased COVID-19 incidence after controlling for metropolitan populations.[18] In contrast, Wong *et al*. showed that population density did correlate positively.[36] These results concur with Wong and show that population density positively correlates with COVID-19 incidence.

Rural Urban Classification:

According to the model, urban areas had increased incidences of COVID-19. A 0.1 unit increase in rural status would correspond to a decrease of 18.69% in COVID-19 cases. Calculated from the model coefficients, this statistic supported the finding of Karim *et al*. that metropolitan areas continue to have higher rates of disease compared to micropolitan and rural areas, and the majority of COVID-19 cases and deaths in 2020 were in urban areas.[17] This data is also consistent with Hamidi *et al*., that metropolitan areas correlated with increased COVID-19 incidence.[18]

Social Vulnerability Index:

Social vulnerability plays a vital role in COVID-19 incidence in communities. According to the model coefficients, a percentile increase in SVI leads to 4.79% more COVID-19 cases. This is consistent with other studies of COVID-19 and SVI, although the magnitude of contribution varies in each study. Karaye and Horney found that a percentile increase in SVI resulted in 16% more COVID-19 cases.[15] Karmakar *et al*. found a 0.1 point increase in SVI increased the incidence of COVID-19 by 14.3%.[16] Long-standing inequities in healthcare access for vulnerable communities appear to contribute to increased incidence; however, variations in the effect of SVI changes on COVID-19 cases in these different studies could be due to the various time periods examined.

Obesity and Diabetes:

The remaining two factors, diabetes and obesity did not add much significance to the model. Nonetheless, the model showed a positive correlation between increased incidence of obesity and diabetes with increased incidence of COVID-19, which supported findings in other studies. A one-unit increase in the percentage of the population with obesity matched with a 4.16% increase in COVID-19 cases. This parallels with Jayawardena *et al*., who also found that increased rates of obesity correlated with increased incidence of COVID-19.[37] In addition, Chen *et al*. found that COVID-19 hospitalized patients with diabetes had two to three times higher mortality rates.[38]

Other Factors:

Some factors had minimal effects in the model that seemed to contradict previous studies. First, there was a positive correlation between an increase in the percentage of the population that stayed at home during the pandemic and COVID-19 incidence. However, this might reflect stricter stay-at-home orders in metropolitan areas with higher incidences of COVID-19. Next, the model indicated that higher poverty rates corresponded with lower rates of COVID-19. Again, this could be attributed to metropolitan areas having higher rates of COVID-19 and increased average income. Metropolitan poverty rates in 2019 were only 11.9% compared to 15.4% in non-metropolitan areas.[39]

An increase in the percentage of the population that were of minority ethnicity also unexpectedly had a negative correlation with COVID-19 cases. This contradicted previous studies that concluded that minority populations were positively correlated with COVID-19.[11] However, this played only a minor role in the model and could be because the data targeted the number of COVID-19 cases as opposed to the percentage of the population that had been diagnosed with COVID-19.

Finally, the model showed that counties with higher percentages of people over 65 had lower cumulative incidences of COVID-19. This seemed to challenge earlier reports that indicated elderly patients were more at risk of contracting COVID-19.[39] However, reviewing CDC weekly cases by age 40, early data from the first three months showed older age groups having a higher incidence of COVID-19. However, after June 2020, the group with the highest incidence of COVID-19 transitioned to ages 18-29. Perhaps elderly patients can socially distance themselves more due to their retired status, so as the pandemic progressed, the incidence of elderly patients declined.

Pearson's Correlation Coefficient:

In addition to the multiple linear regression model, the strength of correlation and direction between the predicting factors and COVID-19 incidence was tested. Correlation between the independent variables and the normally-distributed dependent variable (with all outliers removed) was measured. As shown in Table 4, the Pearson's correlation coefficient was calculated for each of the ten statistically significant variables from the initial model ($p<0.05$). The factors that correlated best with higher COVID-19 cases included greater population density, more multi-unit housing, more urban locations, and an increased percentage of the population that stayed at home. These results concur with the results seen both in this model and in other studies as well.[18,19,21,36]

DOI: 10.36838/v5i1.12

Table 4: Pearson correlation coefficient – correlation of various factors studied with cumulative COVID-19 incidence in U.S. countries.

Factor affecting COVID-19 incidence	*r* Pearsons correlation coefficient
Percentage of population over 65 years of age	-0.188
Change in percentage of population staying at home each day during pandemic time (4/7/20 to 1/10/21) compared to prepandemic time (4/7/19 to 1/10/20)	0.279
Population Density	0.524
Rural versus urban code (urban codes lower)	-0.287
Precentage of homes are in multiunit buildings with greater than 20 units	0.514
Percentage of population in poverty	-0.084
Percentage of population with diabetes	-0.148
Percentage of population is minority ethnicity	0.213
Social Vulnerability Index	0.035
Percentage of population obese (BMI>30)	-0.196

In conclusion, this study emphasizes that COVID-19 transmission is complicated, with numerous aspects to consider. Upon examining this model, the data suggests that the most effective non-pharmaceutical measures would include reducing shared spaces in multi-unit housing, enforcing social distancing, especially in urban areas with increased population density, and investing more resources into medically underserved communities to compensate for persistent healthcare inequities.

Limitations and Future Work:

There were several significant limitations to this study. First, COVID-19 incidence relies on COVID-19 testing being accessible. At the beginning of the pandemic, there were severe supply shortages for testing, and many people with suspected symptoms were never able to receive a confirmatory test. Therefore, the data may be skewed towards more severe cases of COVID-19, as milder cases are less likely to be tested. In addition, some areas had more severe test shortages than others, highlighting the effect of healthcare iniquities. Secondly, demographic information was gathered from surveys between 2018-2020, so some of these variables may have changed compared to 2020. Thirdly, this study focused on the number of COVID-19 cases in each county, in contrast to most studies that focused on the incidence of COVID-19 per 100,000 population. Consequently, the data was skewed towards larger, more populated counties. Attempts to minimize this bias were made with feature scaling and pre-processing, but the bias still may affect the data.

There are several directions in which this research can progress. Although this study was based on 16 months of pandemic data, the pandemic continues to evolve with time, and factors that affect incidence will continue to change. In addition, effective immunizations against COVID-19 will also change the transmission dynamics of COVID-19. Even more importantly, variants such as the new delta variant may increase transmission as initial studies have already found it to be 50% more contagious and can be transmitted through brief contact.[41]

Conclusion

This paper compares COVID-19 cases at a county level to multiple factors, including age, obesity, diabetes, poverty, minorities, urban populations, population density, multi-unit housing, SVI, and population mobility. Furthermore, a reliable model was created to accurately predict the incidence of COVID-19 and showed that multi-unit housing, population density, urban population, and SVI were the most significant factors. The model also provided a measure of how each factor changed COVID-19 incidence; for example, a 0.1 unit move towards more rural status led to 18.69% fewer COVID-19 cases. In addition, a 0.01 unit increase in SVI (meaning more social vulnerability) increased COVID-19 cases by 4.79%. This study highlights the importance of continuing demographic studies over time as the pandemic evolves. Analyzing the factors that lead to higher COVID-19 rates can determine which non-pharmaceutical interventions are most effective. This knowledge is critical to crafting policies that reduce the consequences of pandemics and to improving health worldwide.

Acknowledgements

I would like to thank the staff of the Summer Stem Institute for their guidance and the opportunity to further my interests in epidemiology research. I am especially grateful to my research mentor, Marian Obuseh, for her incredible dedication and guidance throughout the program.

References

1. "COVID Live Update: 198,031,336 Cases and 4,224,346 Deaths from the Coronavirus - Worldometer," July 2021.
2. "COVID-19 Map - Johns Hopkins Coronavirus Resource Center," July 2021.
3. "covid-19-data/us-counties.csvat1eea97d0652b04ceda2d9b88c3f21db5ef643b36 · nytimes/covid-19-data · GitHub," June.
4. "COVID Vaccine Updates: 1 in 10 people in US has tested positive

for virus - ABC7 New York," Apr. 2021.
5. "Estimated COVID-19 Burden | CDC," July 2021.
6. A. Bartik, Z. Cullen, E. Glaeser, M. Luca, and C. Stanton, "What Jobs are Being Done at Home During the Covid-19 Crisis? Evidence from Firm- Level Surveys by Alexander Bartik, Zoe Cullen, Edward L. Glaeser, Michael Luca, Christopher Stanton: SSRN," June 2021.
7. "We're all shopping more online as consumer behaviour shifts | World Economic Forum," July 2021.
8. "Map: Coronavirus and School Closures in 2019-2020," June 2021.
9. C. Bowman, "Coronavirus Moving Study Shows More Than 15.9 Million People Moved During COVID-19 | MYMOVE," June 2021.
10. "New Data from Curriculum Associates Quantifies Impact of COVID Learning Loss; Raises Questions about At-Home Testing | Curriculum Associates," Oct. 2020.
11. A. Klein and E. Smith, "Explaining the economic impact of COVID-19: Core industries and the Hispanic workforce."
12. "A model of disparities: risk factors associated with COVID-19 infection | International Journal for Equity in Health | Full Text."
13. U. Mahajan, "Racial demographics and COVID-19 confirmed cases and deaths: a correlational analysis of 2886 US counties," 2020.
14. G. Harlem and M. Lynn, "Descriptive analysis of social determinant factors in urban communities affected by COVID-19," June 2020.
15.] "At A Glance: CDC/ATSDR Social Vulnerability Index | Place and Health | ATSDR," Apr. 2020.
16. I. Karaye and J. Horney, "The Impact of Social Vulnerability on COVID-19 in the U.S.: An Analysis of Spatially Varying Relationships," Sept. 2020.
17. M. Karmaker, P. Lantz, and R. Tipirneni, "Association of Social and Demographic Factors With COVID-19 Incidence and Death Rates in the US | Health Disparities | JAMA Network Open | JA MA Network," Jan. 2021.
18. S. Karim and H.-F. Chen, "Deaths From COVID-19 in Rural, Micropolitan, and Metropolitan Areas: A County-Level Comparison - Karim - 2021 - The Journal of Rural Health - Wiley Online Library," Nov. 2020.
19. S. Hamidi, S. Sabouri, and R. Ewing, "Full article: Does Density Aggravate the COVID-19 Pandemic?," 2021.
20. S. Cromer, L.Chirag, D. Wexkler, and M. U. Patel, "Geospatial Analysis of Individual and Community-Level Socioeconomic Factors Impacting SARS- CoV-2 Prevalence and Outcomes | medRxiv," 2020.
21. H. Badr, H. Du, M. Marshall, and E. Dong, "Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study | Request PDF," July 2020.
22. H. Badr, M. Marshall, E. Dong, and M. Squire, "Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study - The Lancet Infectious Diseases, " July 2020.
23. "Census - Table Results," 2019.
24. "Census COVID-19 Data Hub: Average Household Size and Population Density - County: Average Household Size and Population Density - County," June 2020.
25. "Small Area Income and Poverty Estimates (SAIPE) Program," 2019.
26. "SAIPE," 2019.
27. "USDA ERS - Rural-Urban Continuum Codes," Dec. 2020.
28. "Bureau of Transportation Statistics," Jan. 2021.
29. "Trips by Distance | Open Data | Socrata," 2020.
30. "Timing of State and Territorial COVID-19 Stay-at-Home Orders and Changes in Population Movement — United States, March 1–May 31, 2020, | MMWR," Sept. 2020.
31. L. Berry, "Living Atlas: Diabetes, Obesity, and Inactivity by US County: Diabetes, Obesity, and Inactivity by US County," June 2021.
32. "CDC/ATSDR SVI Data and Documentation Download | Place and Health | ATSDR," 2018.
33. "Python Release Python 3.7.11 | Python.org," June 2021.
34. "Project Jupyter | Home."
35. "Welcome To Colaboratory - Colaboratory."
36. T. D. D., "The 80/20 Split Intuition and an Alternative Split Method | by The Data Detective | Towards Data Science," Jan. 2020.
37. D. Wong and Y. Li, "Spreading of COVID-19: Density matters," Dec. 2020.
38. R. Jayawardena, D. Jeyakumar, A. Misra, H. Andrew, and P. Ranasinghe, "Obesity: A potential risk factor for infection and mortality in the current COVID-19 epidemic."
39. X. Chen, D. Yang, B. Cheng, J. Chen, A. Peng, C. Yang, C. Liu, M. Xiong, A. Deng, Y. Zheng, and L. Zheng, "Clinical Characteristics and Outcomes of Patients With Diabetes and COVID-19 in Association With Glucose-Lowering Medication," July 2020.
40. "USDA ERS - Rural Poverty & Well-Being," June 2021.
41. "CDC COVID Data Tracker," July 2021.
42. K. Katella, "5 Things To Know About the Delta Variant > News > Yale Medicine," July 2021

Author

Jenna Chuan is a sophomore at Memorial High School in Houston, Texas. She is passionate about understanding physiology and disease and aspires to study medicine in the future. She pursues Tae Kwon Do, piano, and loves to read in her free time.

DOI: 10.36838/v5i1.12