Oregon Health & Science University
School of Medicine

Scholarly Projects Final Report

Title *(Must match poster title; include key words in the title to improve electronic search capabilities.)*

Effects of Social Determinants of Health and Political Leaning on COVID19 Severity in the United States

Student Investigator's Name

Samuel Clarin

Date of Submission *(mm/dd/yyyy)*

3/14/2024

Graduation Year

2024

Project Course *(Indicate whether the project was conducted in the Scholarly Projects Curriculum; Physician Scientist Experience; Combined Degree Program [MD/MPH, MD/PhD]; or other course.)*

Scholarly Project Curriculum

Co-Investigators *(Names, departments; institution if not OHSU)*

Nicole Weiskopf, Ph.D.. OHSU Department of Medical Informatics and Clinical Epidemiology

Mentor's Name

Nicole Weiskopf

Mentor's Department

Department of Medical Informatics and Clinical Epidemiology

Scholarly Project Final Report

Concentration Lead's Name

Mark Baskerville, MD JD MBA

Project/Research Question

Does education impact COVID19 severity?

Type of Project *(Best description of your project; e.g., research study, quality improvement project, engineering project, etc.)*

Descriptive Epidemiology

Key words *(4-10 words describing key aspects of your project)*

Social Determinants of Health, Statistical Analysis

Meeting Presentations

If your project was presented at a meeting besides the OHSU Capstone, please provide the meeting(s) name, location, date, and presentation format below (poster vs. podium presentation or other).

N/A

Publications *(Abstract, article, other)*

If your project was published, please provide reference(s) below in JAMA style.

N/A

Submission to Archive

Final reports will be archived in a central library to benefit other students and colleagues. Describe any restrictions below (e.g., hold until publication of article on a specific date).

Possible Presentation at AMIA fall of 2024. Hold publication until Jan. 1, 2025

Scholarly Project Final Report

Next Steps

What are possible next steps that would build upon the results of this project? Could any data or tools resulting from the project have the potential to be used to answer new research questions by future medical students?

Analyze social media usage and COVID misinformation, surveys in the Oregon community about COVID19 vaccine hesitancy and social determinants and/or social media and/or political affiliation/beliefs.

Please follow the link below and complete the archival process for your Project in addition to submitting your final report.

https://ohsu.ca1.qualtrics.com/jfe/form/SV_3ls2z8V0goKiHZP

Student's Signature/Date *(Electronic signatures on this form are acceptable.)*
This report describes work that I conducted in the Scholarly Projects Curriculum or alternative academic program at the OHSU School of Medicine. By typing my signature below, I attest to its authenticity and originality and agree to submit it to the Archive.

X Samuel Clarin

Student's full name

Mentor's Approval *(Signature/date)*

Scholarly Project Final Report

Report*: Information in the report should be consistent with the poster, but could include additional material. Insert text in the following sections targeting 1500-3000 words overall; include key figures and tables. Use Calibri 11-point font, single spaced and 1-inch margin; follow JAMA style conventions as detailed in the full instructions.*

Introduction *(≥250 words)*

The COVID-19 pandemic has underscored the significant impact of social determinants on health outcomes. Social determinants encompass factors such as economic stability, education, and healthcare access. Vulnerable populations, including racial minorities and those with lower socioeconomic status, experience disproportionately negative health outcomes. Disparities in access to healthcare services, coupled with occupational risks in essential industries, are suggested to contribute to differential outcomes.
Much of this communication occurred via social media, across which scientific communication and misinformation are sometimes difficult to distinguish. Scientific communication has traditionally been done via academic journals and conferences, although this has not translated as effectively to social media. The context a person is in can impact the content they see and how they integrate it into their lives.
During a pandemic, clear communication about scientific recommendations is important. Despite progress, data and research gaps persist regarding how a person's context impacts their decisions, highlighting the need for comprehensive studies. Understanding these intricacies is crucial for formulating effective public health strategies and interventions.
Aim: Determine potential barriers to effective scientific communication during the COVID19 pandemic.
Hypothesis: Educational level would be negatively correlated to COVID19 severity.
Predictions: The social determinants of health would be highly correlated and it would be challenging to determine the underlying driving factor. Also, political leaning would have an impact on COVID19 severity.

Methods *(≥250 words)*

Case, death and vaccine data was downloaded from the CovidActNow.org on 8/6/2022.
Race, income, education, gender and age by county was downloaded from the 2020 Census 5 year American Community survey. 2020 election data by county was taken from https://electionlab.mit.edu/data. Data cleaning utilized "dplyr" and "tidyverse" packages.
Initial features were selected based on literature review and one feature from each "set" (ie income) was used for analysis. Challenges with collinearity, execution and scope of analysis with multiple features of each "set." For race, "Black" was chosen due to Black patients being reported to have increased mortality from COVID19 and was found to be collinear with "White." For income, mean income was chosen over median income or any specific income bracket. For education, "less than high school" was chosen over "high school graduate +/- some college" and "college graduate or higher" with focus on answering the primary question of if lack of education impaired communication on COVID-19. For election data, candidate votes were condensed into "Trump", "Biden" and "Other." Votes for each candidate was then divided by total votes for each county and created a percentage vote. "Percent vote for Trump" was selected as "Other" candidates did not garner a significant portion of the vote and is the reciprocal of "percent vote for Biden".
Mortality rate was calculated using deaths from CDC data divided by population per county from the Census' race and age survey.
Date selected for pre vs post vaccination was 2/1/2021, which was the date first dose of vaccination administered outnumbered cases.

Values above 100k for Total Cases per 100k and Vaccinations per 100k on 8/1/2022 were removed, as well as counties with missing values for any of the social determinants or election data, which reduced N=3111 to N=3069.

To create a proxy of COVID19 maximum impact, the highest 14 day case rate was determined by iterating over all dates for pre and post vaccination periods for the maximum mortality and case rates. This was used to create "PreCases" and "PreMort" data sets for the maximum case incidence rate and maximum mortality rate, respectively, prior to 2/1/2021. "PostCases" and "PostMort" were maximum case incidence rate and maximum mortality rate over a 14 day stretch between 2/1/2021 and 8/5/2022. "TotCases" represents the per 100k numbers of cases on 8/5/2022 to examine total incidence of COVID19 over this period.

The collinearity matrix with independent and dependent variables was generated using "cor" package with p-values and heatmap for collinearity and p-values were created in Microsoft Excel.

Multiple linear regression across each dependent variable using "lm". Table of coefficients and p-value heatmap were generated with Microsoft Excel.

K-Nearest Neighbors analysis was done with the "caret" package and examined K of 5, 7 and 9.

Regression Tree analysis was done with "rpart" package using leave one out cross validation and then averaged to create the final model. The average Regression Tree visualized with "rpart.plot" and "barplot."

Results *(≥500 words)*

Collinearity

Evaluation of collinearity between selected features did not uncover significant collinearity between any parameters. There was no significant collinearity between any of the predictors. Significant collinearity used 0.7-1 as cutoffs, while moderate used $0.3 < x < 0.7$ and 0.3 or below was little to no collinearity. The majority of these interactions were statistically significant, with a minority of interactions of being statistically insignificant and therefore difficult to assess.

Multiple Linear Regression

Generally speaking, there were few statistically and functionally significant interactions for the mortality prior to vaccination or post vaccination. Vaccinations per 100k people on 8/1/2022 was several orders of magnitude below the most important factors for mortality in pre and post vaccination periods. There were no factors that were statistically significant for pre vaccination mortality outside of the intercept. For post vaccination mortality, the most important two features that were statistically significant were "%male" (p= 0.04) and "%>65" (p= 1.50E-19). Cases had more statistically and functionally significant coefficients. For the pre vaccination period, the "%VoteTrump" (p= 1.22E-12) was an order of magnitude more important than the next closest coefficient "%male" (p= 1.06E-93), though both were smaller in size than intercept. For case incidence rate in the post vaccination period, "%VoteTrump" (p=5.08E-19) was the most most important factor by an order of magnitude, although all chosen features were statistically significant outside of "%male" and the intercept. Total cases were all statically significant outside of "%Black" (p= 0.12). The coefficient for "%VoteTrump" was an order of magnitude larger than the next largest coefficients of "% >65" and "%male", although it was an order of magnitude smaller than the intercept.

K-Nearest Neighbors

K-Nearest Neighbors analysis was performed using leave one out cross validation. For total cases, K=7 produced R^2=0.25. All other iterations used K=9. Pre vaccination cases had an R^2 of 0.27 and pre vaccination mortality had an R^2 of 0.01. Post vaccination cases and post vaccination mortality had R^2 of 0.19 and 0.13 respectively.

Regression tree

The regression tree analysis was performed and found weak correlation for mortality data, with R^2 of 0.0003 for pre vaccination period and 0.11 for post vaccination period. Analysis for cases had somewhat

better correlations, with pre vaccination R^2 of 0.28, post vaccination of 0.25 and total cases of 0.18. In descending order, the three most important for pre vaccination cases were "% male", "%VoteTrump" and "%Black", with the first major decision point utilizing "% male <59" as seen in figure 4.1. The three most important features for post vaccination cases were "%VoteTrump", "%>65" and per 100,000 vaccination rate, in descending order while the first decision point was percent vote for Trump >71 as seen in figure 5.2 and 5.1 respectively.

Pre Vaccination Cases

Figure 4.1

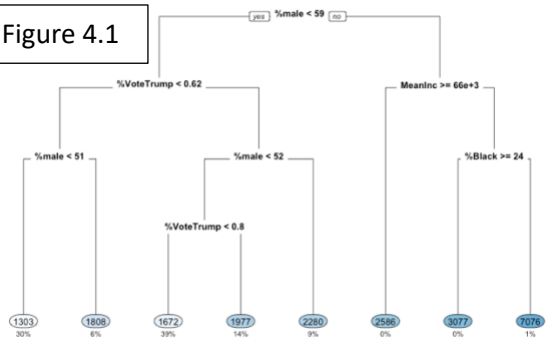


Figure 4.2

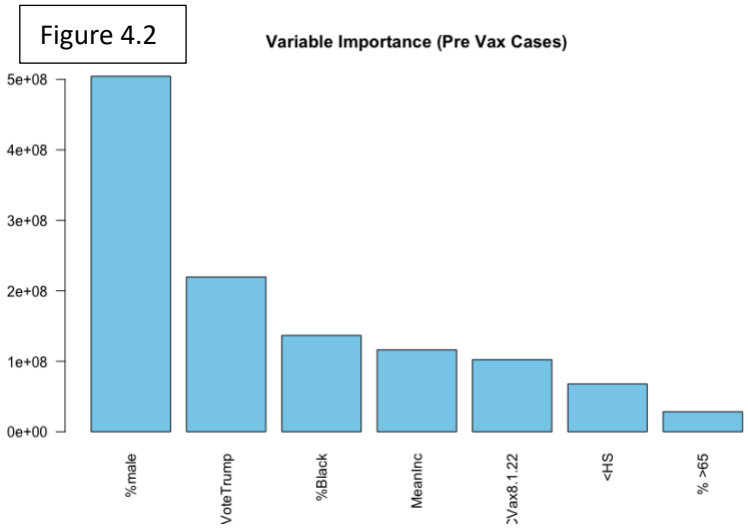
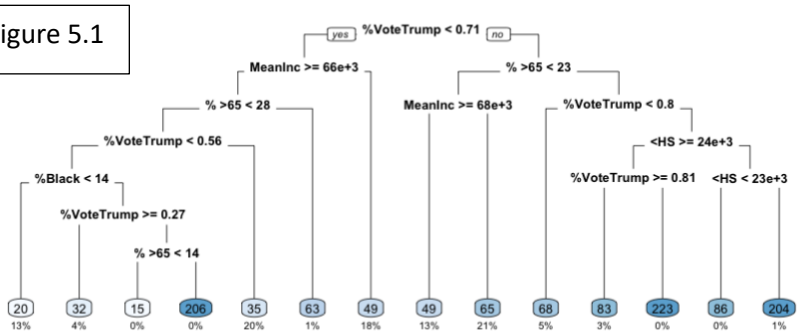


Figure 4.1 Decision Tree for maximum 14 day stretch Pre Vaccination Cases per 100k: Graphical representation of the decision tree for the same factors as above multiple linear regression. Nodes are organized by order of reducing variance of the remainder of the dataset but not in order of importance. Order of importance is demonstrated in figure 4.2 for the overall impact these features have for the post vaccination period. R^2= 0.27

Figure 5.1



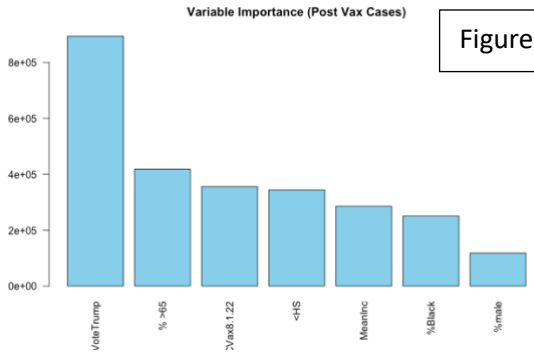
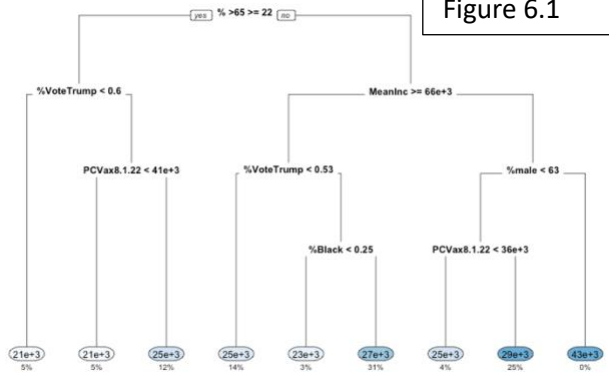


Figure 5.2

Figure 5.1 Decision Tree for maximum 14 day stretch Post Vaccination Cases per 100k: Graphical representation of the decision tree for the same factors as above multiple linear regression. Nodes are organized by order of reducing variance of the remainder of the dataset but not in order of importance. Order of importance is demonstrated in figure 5.2 for the overall impact these features have for the post vaccination period. R^2= 0.26

Figure 6.1



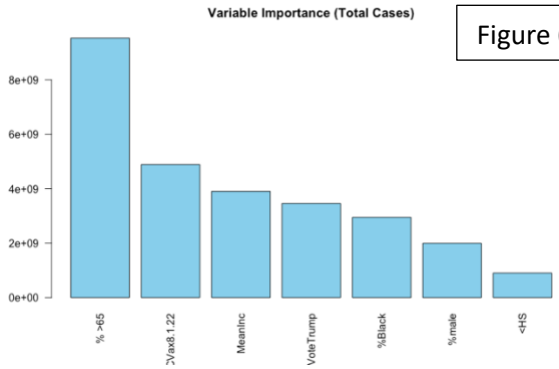


Figure 6.2

Figure 6.1 Decision Tree for Total Cases per 100k: Graphical representation of the decision tree for the same factors as above multiple linear regression. Nodes are organized by order of reducing variance of the remainder of the dataset but not in order of importance. Order of importance is demonstrated in figure 6.2 for the overall impact these features have for the post vaccination period. R^2= 0.18
Not featured: Pre and post Vaccination Mortality rates as the R^2 was 0.0003 and 0.11 respectively.

Discussion *(≥500 words)*

Null hypothesis was accepted regarding education and that the social determinants would be highly correlated. It is difficult to determine any underlying driving factors and political leaning did have an impact on COVID19 severity. While there has been much discussion of social determinants impacting political leaning, there was not high collinearity between any of the predictors and each other, or any of the predictors and any dependent variable. Interestingly, the selected features were not linearly interrelated to a large degree, and also did not explain the dependent variables in a linear fashion. This could be due to the selection of features that had more than two options (ie race), as selecting one of two features would be either positive or negative and would not affect the analysis. It would be interesting to see how addition of different races, income levels, education status or other factors not mentioned in this paper could contribute to COVID19 information spread.

The linear regression analysis noted significant correlation with percent vote for Trump and Cases at all three time points, though it was smaller than the intercept for Total and Pre-Vaccination slices. There were several other statistically significant correlations that are functionally zero, which suggests minimal linear correlation with outcomes despite statistical significance.

Neither KNN or decision tree analysis were able to achieve an R^2 of greater than 0.3, which suggests that there is weak correlation between the predictor and dependent variables. Of the features presented, percent of a county that voted for Donald Trump in 2020 appears to be the most important factor in predicting COVID19 severity, though the correlation is weak. This tracks with the linear regression findings, and prior research noting the political aspect to COVID19 spread and vaccination.

Curiously, vaccination rate was not affiliated with lower mortality. This is likely due to choosing to use the highest 14 day stretch as the maximum value, which might not correlate with herd immunity in each county. I selected the date where more first doses of vaccines had been distributed than cases as the vaccination turn point, though a more complete picture of each county's vaccination status would likely be gained through analysis of when each county reached herd immunity, which was outside of the scope of this project. There is significant data that vaccinations reduce mortality, and are correlated with political affiliation though this analysis found that cases were correlated with political affiliation. Further work will need to be done to determine the lag time between vaccination with mRNA vaccines and full effect, and also the durability of this immunity. The immunity durability work can utilize these county level analyses as well as serological studies to determine antibody response in the long term.

Further research on spread of information on a more granular level could help elucidate differences that are not as apparent on a county-wide basis to further improve scientific communication to the public. Having access to social media data about misinformation spread and demographic information could provide more specific information about who is at risk for misinformation, and who might not have access to government sponsored information at all based on algorithmic recommendations.

Conclusions *(2-3 summary sentences)*

Political affiliation was more important than education or any of the non-modifiable social determinants that were analyzed, which might be the most challenging of the factors to adjust via communication with the proliferation of highly polarized networks on social media. Efforts to increase bipartisan support for national scientific communication are vital to improve future outcomes, and further focus on this area of research could yield significant benefits.

References *(JAMA style format)*

1. Cortez MF, Court E. More Americans have received at least 1 Covid vaccine dose than tested positive. Bloomberg.com. February 1, 2021. Accessed March 10, 2024. https://www.bloomberg.com/news/articles/2021-02-01/u-s-hits-milestone-in-pandemic-with-more-vaccinated-than-cases.
2. Social Determinants of Health at CDC. Centers for Disease Control and Prevention. December 8, 2022. Accessed March 10, 2024. https://www.cdc.gov/about/sdoh/index.html#:~:text=Social%20determinants%20of%20health%20(SDOH,the%20conditions%20of%20daily%20life.
3. Why is addressing social determinants of health important for CDC and Public Health? Centers for Disease Control and Prevention. December 8, 2022. Accessed March 10, 2024. https://www.cdc.gov/about/sdoh/addressing-sdoh.html.
4. Muhammed T S, Mathew SK. The disaster of misinformation: a review of research in social media. *Int J Data Sci Anal*. 2022;13(4):271-285. doi:10.1007/s41060-022-00311-6
5. Johnson SS and S. Males and the Hispanic, American Indian and Alaska native populations experienced disproportionate increases in deaths during pandemic. Census.gov. June 29, 2023. Accessed March 10, 2024. https://www.census.gov/library/stories/2023/06/covid-19-impacts-on-mortality-by-race-ethnicity-and-sex.html.