Racial Composition as a Key Factor in Predicting COVID-19 Impacts in U.S. Counties

Agrawal, S., Hooper-Antunez, Y., Lian, C.

Department of Statistics, Grinnell College

Professor Shonda Kuiper

Abstract The COVID-19 pandemic has had a severe impact across the United States, with a death toll surpassing that of the 1918 influenza pandemic. The unequal distribution and slow deployment of healthcare resources was found to disproportionately affected minority populations primarily, resulting in higher case fatality rates (CFR) among these groups. This study used data collected during COVID-19 as an example to examine the impact of demographic, socio-demographic, and health-related factors on the healthcare system during crises. The study selected effective predictors and developed a comprehensive model to assess pandemic severity and communal healthcare capacities. The model is anticipated to become a reliable guide for crisis responses in the future. Statistical techniques such as best subsets regression and extra-sum-of-squares (ESS) test were implemented to validate the model's effectiveness. The results found race as an important predictor of CFR. This highlights the need for equitable resource allocation and customized healthcare policies accommodating needs for communities with distinct racial compositions.

Ethnic Composition as a Key Factor in Predicting COVID-19 Impacts in U.S. Counties

The COVID-19 pandemic has been the most devastating health crisis in recent American history, surpassing the death toll of the 1918 influenza (Lovelace, 2021). The Center of Disease Control and Prevention (CDC, 2020) reported that from 2020 to 2022, COVID-19 resulted in over 100 million cases and approximately one million deaths in the United States. Despite early warnings, the U.S. was unprepared for the pandemic, and the unequal distribution and slow deployment of healthcare resources disadvantaged minority populations even further, considering their peak case fatality rate (CFR, Equation 1) than the general public (CDC, 2020).

Kadri et al. (2021) reported that nearly 25% of COVID-19 deaths may have been caused by a lack of immediate medical treatment. The fragility of our healthcare system during surges in pandemic caseloads has raised concerns and highlights the need for improvements in crisis response and resource allocation mechanism. Additionally, serving as the core component of COVID-19 treatments for severely infected patients, more than one third of intensive care units (ICUs) across the nation have already been occupied before the abrupt increase of more than 15,000 patients per county on average (CDC, 2020).

While the most acute workforce challenges have eased with decreasing case and hospitalization rates as of March 1, 2022, understanding the key contributing factors behind this unmanageable pandemic is still essential for developing effective prevention and intervention strategies.

This report examined the combination of demographic (i.e., communal composition by age and ethnicity), socioeconomic (i.e., housing estimates and economic characteristics), and healthcare-related factors (i.e., chronic diseases and health insurance coverage) impact the communal burden on the healthcare system during prolonged healthcare crises instantiated by COVID-19. This study uses 2021 data compiled from CDC and the U.S. Census Bureau to generate a comprehensive model for assessing communal healthcare capacities and instructing inter-community resource-sharing plans to balance healthcare availability.

1 Materials and Methods

1.1 Study Population

The COVID-19 data for all U.S. counties, including cases/deaths, hospital capacity, and vaccination, was collected by the U.S. Department of Health and Human Services (HHS, 2020; HHS, 2021) and New York Times (2020) on a weekly basis. To better reflect the severeness of the pandemic, case fatality rate (CFR) was calculated via Equation 1 (Rajgor et al., 2020). For the purpose of this study, county-level demographic and health-related information was obtained from the U.S. Census Bureau via American Community Survey (Bureau, 2021a, 2021b) and CDC (2021). We only chose 2021 to reduce noises induced by high proportions of missing, inconsistent data from 2020, 2022, and 2023.

R (version 4.3.0) was used to process (**Error! Reference source not found.**) and analyze the collected data (**Error! Reference source not found.**).

1.2 Methods

1.2.1 Data Cleaning

The original datasets (Table 1) were merged by FIPS code and week number of the collection date to remove observations that are lack of cases from COVID-19 Reported Patient Impact and Hospital Capacity by Facility (HHS, 2020). Missing or inconsistently outstanding cases (which have been primarily prevalent at the beginning of 2021) were replaced by *NA*.

Preliminary data analysis was performed via correlation matrix and scatterplot to better understand the study population, identify inconsistent or collinear variables, and pinpoint potentially biased ones. Herein, the data was considered cleaned, and the remaining variables were considered as the initial models.

To test the effects of race in the final prediction model, we shortlisted all race-related variables (i.e., "white", "black", "asian", "hispanic", and "native") for the following steps. Best subsets regression was carried out on the initial model subsequently to identify the effective predictors of CFR. After shortlisting the important explanatory variables from two iterations (Figure 2, Figure 3), a collective best subset regression was done to give a collection of variables with the greatest explanatory power of CFR. Thereafter, necessary variable transformations were carried out based on residual plots (). Additionally, terms with a notably high VIF value, such as "chd" and "csmoking" (Appendix 3.2.2), were neutralized by introducing interaction terms to reduce collinearity (Appendix 3.3.2). Herein, the final reduced model was confirmed (Appendix 3.3.2), and the full model was generated by recalling the race-related terms (Appendix 3.3.1).

And extra-sum-of-squares (ESS) test (α-level = 0.5) was conducted between the reduced and full models to assess the impact of race on CFR prediction (Appendix 3.3.3).

2 Results & Discussion

With a p-value < 2.2×10^{-16} (F-value = 185.16) from the ESS test, we found the introduction of race-related terms improved CFR prediction for the full model. This was as well reflected by the improved $R\text{-}sq$ ($R\text{-}sq_{\text{reduced}}$ = 0.4525; $R\text{-}sq_{\text{full}}$ = 0.4722). Therefore, we may conclude that the county-level CFR is better predicted by the full model (Appendix 3.3.4).

Our finding, as per the prediction model, suggested that the severity of COVID-19 in U.S. counties was influenced by their racial composition, which was consistent with previous studies (CDC, 2023; Dollemore, 2020; Ndugga et al., 2022). The inclusion of race-related terms (i.e., "white", "black", "asian", "hispanic", and "native") in the model provided a notable improvement to the prediction outcome.

Counties with higher proportions of Asian and Native Americans had notably lower COVID-19 severity than counties dominated by White, American, and Hispanic Americans (Figure 7). Surprisingly, although Black-dominated counties had the most staffed ICU beds than other racial communities (Figure 8), they still suffered significantly more than other racial communities in the pandemic (Figure 7), which raises the question: "If the lack of healthcare resources is not the issue, what other factors could contribute to the greater severity of COVID-19 that were exclusive to Black communities?"

Despite the high effectiveness of COVID-19 reported in studies (Ferdinands et al., 2022), both Black and White-dominated counties showed high hesitancy to vaccinations (Figure 9). As Black-dominated counties also exhibited a notably high unemployment rate (Figure 10), leaving them unprotected by employee health insurance plans. This adverse combination physically exposed them to the infectious disease, causing much severe symptoms than which experienced by other ethnical communities.

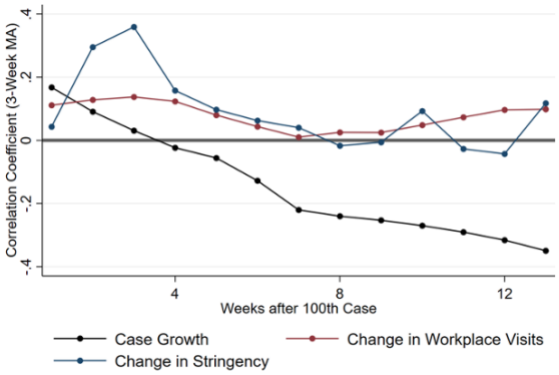

Figure 1. Correlation of approval rates over time: health vs the economy (Trebesch et al., 2020). A negative correlation between case growth and change in stringency was shown, whereas there was not a notable response to the change in workspace visits from policy-makers during the pandemic.

Additionally, the study found that counties with different racial compositions exhibit distinct correlations with CFR, as shown in Figure 6. This suggests that suggesting

may need to consider different approaches to accommodate the needs and cultural disparities across communities.

2.1 Future Directions

Our data provided strong support for the hypothesis that the racial composition of a county is a significant factor in determining the severity of a pandemic in that county. However, future research could benefit from expanding the selection of datasets to further validate and refine these findings. For example, non-pharmaceutical interventions (NPIs) such as social distancing could be a key predictor for pandemic progression and mortality, as previous studies have shown a strong correlation between case growth rate and the extent of prioritizing healthcare over economics from the administration (Correia et al., 2022). In future crises, this comprehensive prediction model may be aided by advanced monitoring technologies from government agencies such as Bureau of Economic Analysis (BEA) (Correia et al., 2022; Trebesch et al., 2020; *U.S. Bureau of Economic Analysis (BEA)*, 2023).

Previous news reports have highlighted concerns among experts about the potential for increased social contacts and outdoor population density during weekends and holidays (Malani & Inskeep, 2023; Yan & Elamroussi, 2021), which would indirectly promote pathogen transmissions. Yet, Figure 11 did not find a similar correlation between CFR and the use of public transportation, which involves frequent close social contacts as well. Considering the possibility that the impact of social interactions was clouded by the work-from-home policy executed in 2021, we should expand the examination on NPIs' impact on case growth rates. To further increase the flexibility of the predictive model at different times of the year, date factors such as holiday breaks could be considered. Additionally, given the varying transmission capacity of pathogens in different climatic environments or seasons (Chen et al., 2021), the data could be broken into quarters to provide season or date-specific predictions. We anticipate that this approach would improve the accuracy of the model's predictions.

References

Bureau, U. S. C. (2021a). *ACS Demographic and Housing Estimates: 2021: ACS 1-Year Estimates Data Profiles* (ACSDP1Y2021.DP05. https://data.census.gov/table?q=county+DP05&tid=ACSDP1Y2021.DP05

Bureau, U. S. C. (2021b). *Selected Economic Characteristics: 2021: ACS 1-Year Estimates Data Profiles* (ACSDP1Y2021.DP03. https://data.census.gov/table?tid=ACSDP1Y2021.DP03

CDC, C. f. D. C. a. P. (2020, 2020-03-28). *COVID Data Tracker*. @CDCgov. Retrieved May 11 from https://covid.cdc.gov/covid-data-tracker

CDC, C. f. D. C. a. P. (2023, 2023-04-24T08:58:08Z). *Risk for COVID-19 Infection, Hospitalization, and Death By Race/Ethnicity | CDC*. @CDCgov. Retrieved 05-17 from https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-race-ethnicity.html

CDC, C. f. D. C. a. P., National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health. (2021). *PLACES: County Data (GIS Friendly Format), 2021 release* (009:020. https://chronicdata.cdc.gov/500-Cities-Places/PLACES-County-Data-GIS-Friendly-Format-2021-releas/kmvs-jkvx

Correia, S., Luck, S., & Verner, E. (2022). Pandemics Depress the Economy, Public Health Interventions Do Not: Evidence from the 1918 Flu. *The Journal of Economic History*, *82*(4), 917-957. https://doi.org/10.1017/s0022050722000407

Dollemore, D. (2020). Why is the COVID-19 mortality rate highest for Black Americans? | COVID-19 CENTRAL @THEU. *U OF U HEALTH*. https://coronavirus.utah.edu/research-news/why-is-the-covid-19-mortality-rate-highest-for-black-americans/

Ferdinands, J. M., Rao, S., Dixon, B. E., Mitchell, P. K., DeSilva, M. B., Irving, S. A., Lewis, N., Natarajan, K., Stenehjem, E., Grannis, S. J., Han, J., McEvoy, C., Ong, T. C., Naleway, A. L., Reese, S. E., Embi, P. J., Dascomb, K., Klein, N. P., Griggs, E. P., . . . Fireman, B. (2022). Waning 2-Dose and 3-Dose Effectiveness of mRNA Vaccines Against COVID-19-Associated Emergency Department and Urgent Care Encounters and Hospitalizations Among Adults During Periods of Delta and Omicron Variant Predominance - VISION Network, 10 States, August 2021-January 2022. *MMWR Morb Mortal Wkly Rep*, *71*(7), 255-263. https://doi.org/10.15585/mmwr.mm7107e2

HHS, U. S. D. o. H. H. S. (2020). *COVID-19 Reported Patient Impact and Hospital Capacity by Facility*. https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/anag-cw7u

HHS, U. S. D. o. H. H. S. (2021). *COVID-19 Vaccinations in the United States, County*. https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh

Kadri, S. S., Sun, J., Lawandi, A., Strich, J. R., Busch, L. M., Keller, M., Babiker, A., Yek, C., Malik, S., Krack, J., Dekker, J. P., Spaulding, A. B., Ricotta, E., Powers, J. H., 3rd, Rhee, C., Klompas, M., Athale, J., Boehmer, T. K., Gundlapalli, A. V., . . . Warner, S. (2021). Association Between Caseload Surge and COVID-19 Survival in 558 U.S. Hospitals, March to August 2020. *Ann Intern Med*, *174*(9), 1240-1251. https://doi.org/10.7326/M21-1213

Lovelace, B., Jr. (2021, 2021-09-20). Covid is officially America's deadliest pandemic as U.S. fatalities surpass 1918 flu estimates. *CNBC*. https://www.cnbc.com/2021/09/20/covid-is-americas-deadliest-pandemic-as-us-fatalities-near-1918-flu-estimates.html

Ndugga, N., Hill, L., & Artiga, S. (2022, 2022-11-17). COVID-19 Cases and Deaths, Vaccinations, and Treatments by Race/Ethnicity as of Fall 2022. *KFF*. https://www.kff.org/racial-equity-and-health-policy/issue-brief/covid-19-cases-and-deaths-vaccinations-and-treatments-by-race-ethnicity-as-of-fall-2022/

New York Times, N. Y. T. (2020). *Coronavirus (COVID-19) Data in the United States*. https://github.com/nytimes/covid-19-data

Rajgor, D. D., Lee, M. H., Archuleta, S., Bagdasarian, N., & Quek, S. C. (2020). The many estimates of the COVID-19 case fatality rate. *Lancet Infect Dis*, *20*(7), 776-777. https://doi.org/10.1016/S1473-3099(20)30244-9

Trebesch, C., Konradt, M., Ordoñez, G., & Herrera, H. (2020, 6 Nov 2020). The political consequences of the Covid pandemic: Lessons from cross-country polling data. *VOXEU*. https://cepr.org/voxeu/columns/political-consequences-covid-pandemic-lessons-cross-country-polling-data

U.S. Bureau of Economic Analysis (BEA). (2023). Retrieved May 17 from https://www.bea.gov/

3 Appendix

3.1 Data Cleaning & Processing

$$\text{Case fatality rate (CFR)} = \frac{\text{Number of deaths}}{\text{Number of cases}} \times 100\% \qquad \text{(Equation 1)}$$

Name	Source
ACS Demographic and Housing Estimates: ACS 1-Year Estimates Data Profiles	Bureau (2021a)
Selected Economic Characteristics: ACS 1-Year Estimates Data Profiles	Bureau (2021b)
PLACES: County Data (GIS Friendly Format), 2021 release	CDC (2021)
COVID-19 Reported Patient Impact and Hospital Capacity by Facility	HHS (2020)
COVID-19 Vaccinations in the United States, County	HHS (2021)
Coronavirus (COVID-19) Data in the United States	New York Times (2020)

Table 1. Original datasets and corresponding sources.

3.2 Data Processing

3.2.1 Best Subsets Regressions

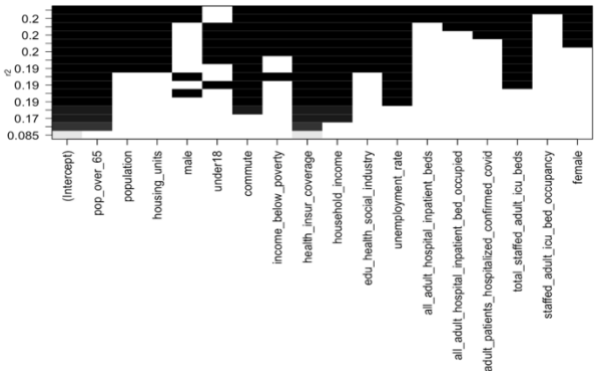

Figure 2. Results from the first best subsets iteration of the cleaned dataset (first half).

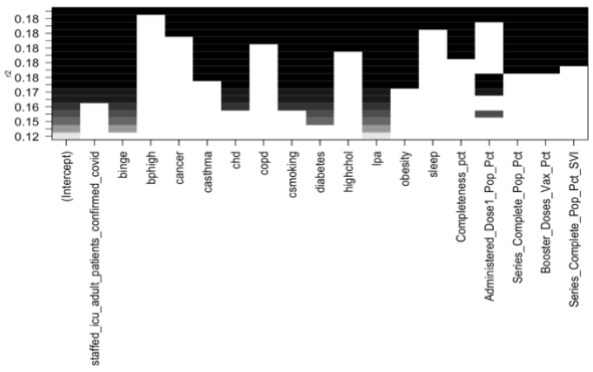

Figure 3. Results from the first best subsets iteration of the cleaned dataset (second half).

Variables selected: First iteration

```
"pop_over_65", "under18", "commute", "health_insur_coverage", "household_income", "male",
"income_below_poverty", "total_staffed_adult_icu_beds_7_day_avg"
```

Variables selected: Second iteration

```
"staffed_icu_adult_patients_confirmed_covid_7_day_avg", "Binge", "casthma", "chd", "csmoking",
"diabetes", "lpa", "obesity", "Administered_Dose1_Pop_Pct"
```

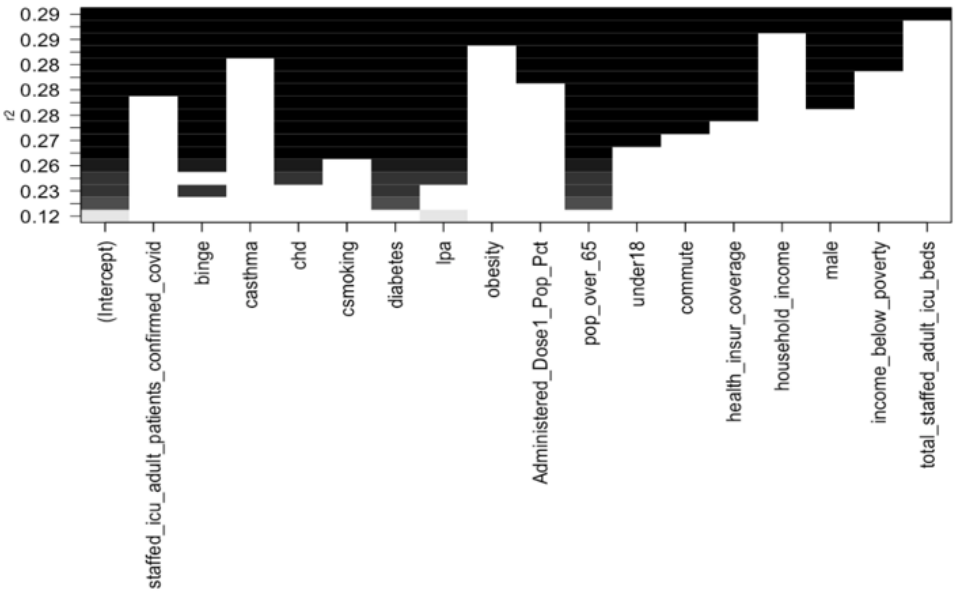

Figure 4. Results from the collective final best subsets iteration of selected variables from the first and second iterations.

Variables selected: Final iteration

```
"staffed_icu_adult_patients_confirmed_covid_7_day_avg", "binge", "chd", "csmoking", "diabetes",
"lpa", "pop_over_65", "under18", "commute", "health_insur_coverage", "male"
```

3.2.2 Variance inflation factor (VIF)

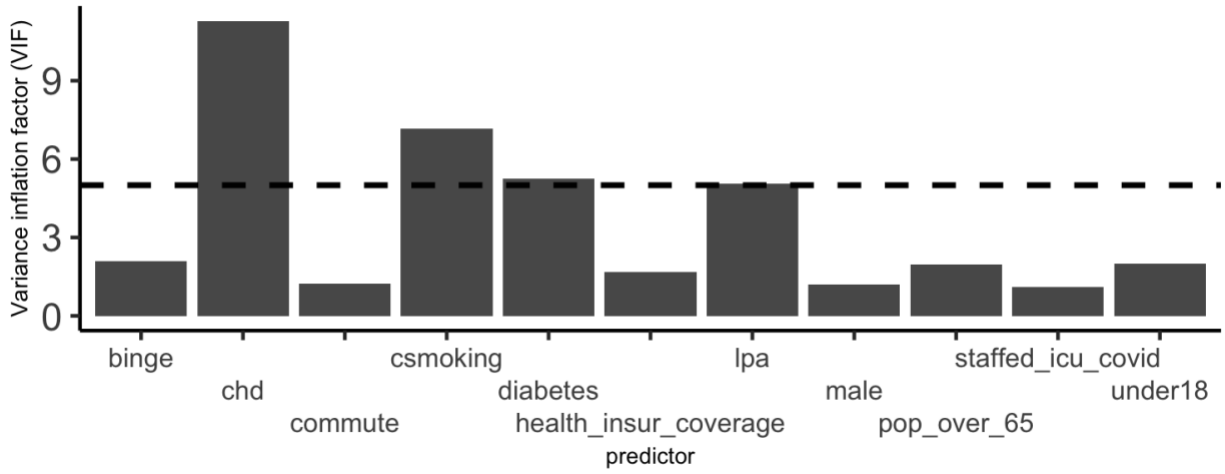

Figure 5. Variance inflation factor (VIF) results for variables of the reduced model.

3.3 Full & Reduced Models for CFR

3.3.1 Full Model (with race-related variables)

```
lm(formula = CFR ~ white + black + hispanic + native + asian + binge + chd + csmoking + diabetes +
lpa + log_male + log_pop_over_65 + under18 + log_health_insur_coverage + log_commute + Log_staffed_icu
+ csmoking * log_commute + diabetes * under18 + lpa * log_pop_over_65 + under18 * log_commute +
log_health_insur_coverage * log_commute + csmoking * binge + binge * diabetes + chd * diabetes + chd
* lpa, data = regfin3)

Residual standard error: 0.4574 on 24874 degrees of freedom
```

```
Multiple R-squared: 0.4722, Adjusted R-squared: 0.4716
F-statistic: 890.1 on 25 and 24874 DF, p-value: < 2.2e-16
```

3.3.2 *Reduced Model (without race-related variables)*

```
lm(formula = CFR ~ binge + chd + csmoking + diabetes + lpa +  log_male + log_pop_over_65 + under18 +
log_health_insur_coverage  +  log_commute  +  log_staffed_icu  +  csmoking  *  log_commute  +  diabetes  *
under18 + lpa * log_pop_over_65 + under18 * log_commute +  log_health_insur_coverage * log_commute +
csmoking * binge + binge * diabetes + chd * diabetes + chd * lpa, data = regfin2)

Residual standard error: 0.4658 on 24879 degrees of freedom
Multiple R-squared:0.4525, Adjusted R-squared:  0.4521
F-statistic: 1028 on 20 and 24879 DF, p-value: < 2.2e-16
```

3.3.3 *ESS test between the reduced and full models*

```
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1  24879 5397.6
2  24874 5204.0  5    193.69 185.16 < 2.2e-16 ***
```

3.3.4 *Linear Regression Equation: Full Model*

```
County-level CFR = 7.48 + 2.17 * white + 3.43 * black + 1.85 * hispanic + 3.28 * native + 3.15 * asian
+ -0.21 * binge + -0.42 * chd + -0.03 * csmoking + -1.13 * diabetes + 0.04 * lpa + -1.45 * log_male
+ -0.01 * log_pop_over_65 + -14.11 * under18 + 8.62 * log_health_insur_coverage + 0.42 * log_commute
+  -0.09  *  log_staffed_icu  +  -0.01  *  csmoking:log_commute  +  1.17  *  diabetes:under18  +  0.05  *
lpa:log_pop_over_65 + -0.68 * under18:log_commute + 1.4 * log_health_insur_coverage:log_commute + 0
* binge:csmoking + 0.02 * binge:diabetes + 0.1 * chd:diabetes + -0.03 * chd:lpa + ε
```

3.4 Discussion

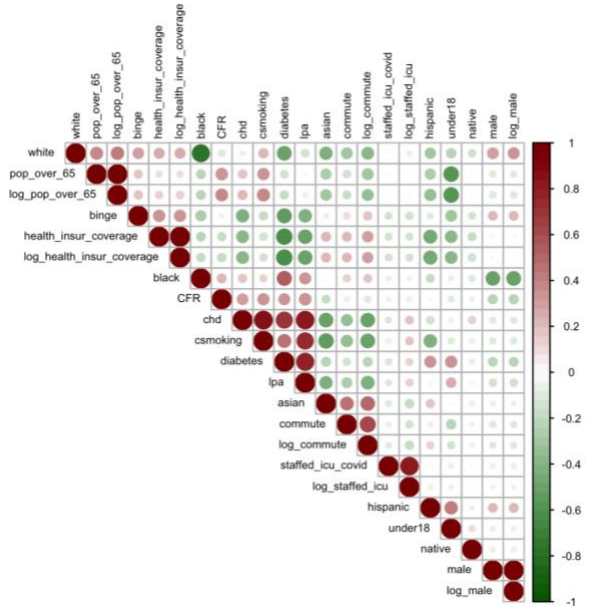

Figure 6. Correlation plot for the full model.

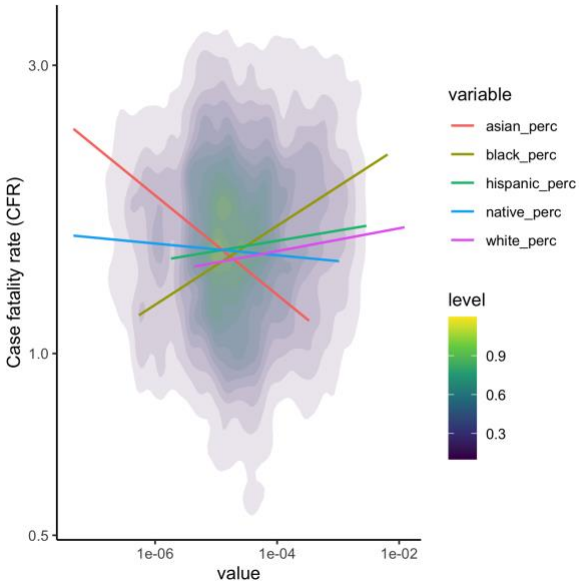

Figure 7. Two-dimensional density plot for the relationship between CFR and ethnic composition.

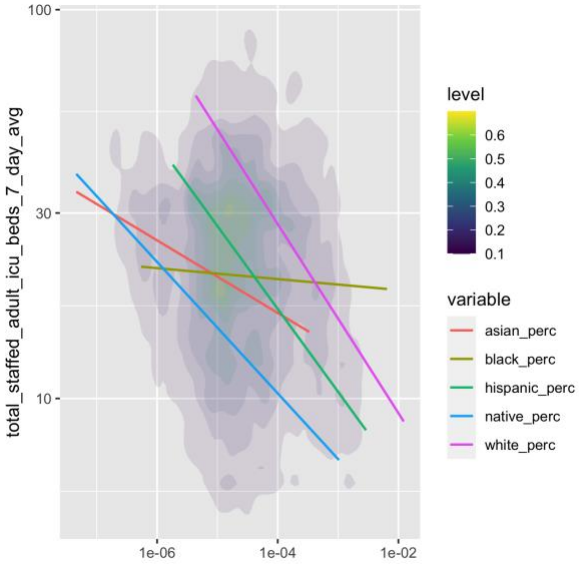

Figure 8. Two-dimensional density plot for the relationship between the 7-day average total number of staffed adult ICU beds and racial composition.

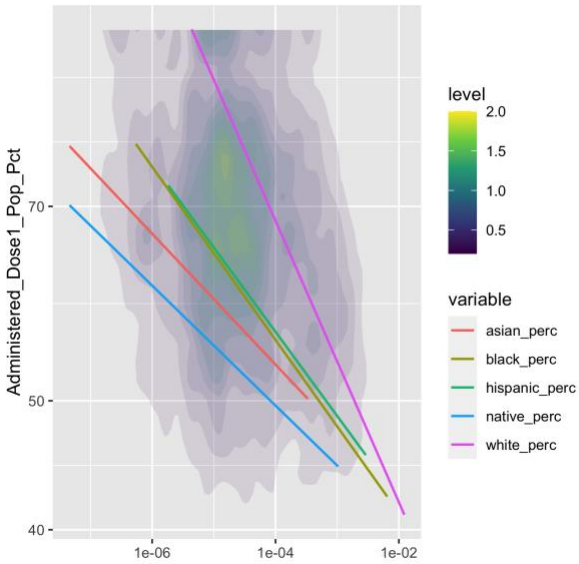

Figure 9. Two-dimensional density plot for the relationship between percent of population administrated first dose of COVID-19 vaccines and racial composition.

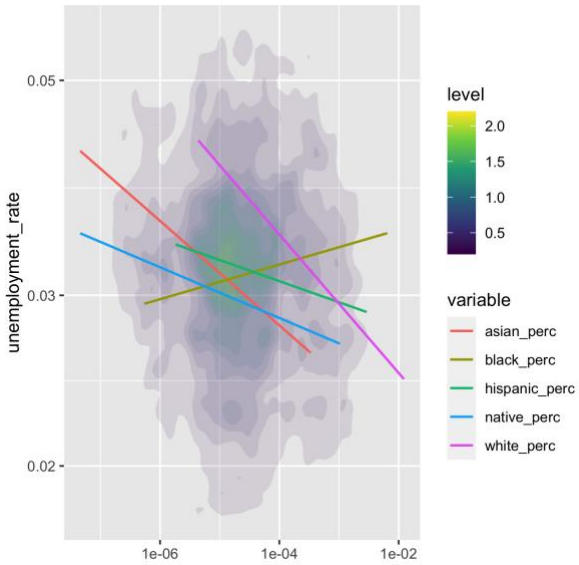

Figure 10. Two-dimensional density plot for the relationship between unemployment rate and racial composition.

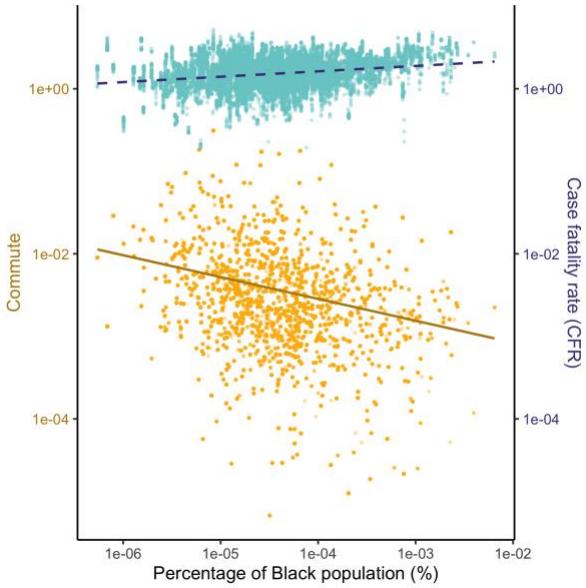

Figure 11. Scatterplot for the relationship of percentage of Black population with both commute (yellow dots and solid regression line) and CFR (blue dots and dashed regression line).

3.5 Variable Dictionary

Model	Name	Description	Source
NA	fips	FIPS code by county	New York Times (2020)
NA	week	The ith week of the year	New York Times (2020)
Response	CFR	Case fatality ratio (Equation 1)	New York Times (2020)
Reduced	pop_over_65	Percentage of population ≥ 65 yrs	Bureau (2021a)
Reduced	population	Total Population	Bureau (2021a)
Full	white	Percentage of only-white population	Bureau (2021a)
Full	black	Percentage of only-black population	Bureau (2021a)
Full	hispanic	Percentage of only-hispanic population	Bureau (2021a)
Full	native	Percentage of only-native population	Bureau (2021a)
Full	asian	Percentage of only-asian population	Bureau (2021a)
Reduced	male	Percentage of male population	Bureau (2021a)
Reduced	under18	Percentage of population < 18 yrs	Bureau (2021a)
Reduced	commute	Percentage of workers (≥ 16 yrs) commuting via public transportation	Bureau (2021b)
Reduced	health_insur_coverage	Percentage of civilian noninstitutionalized population with health insurance coverage	Bureau (2021a)
NA	household_income	Median household income in 2021 inflation-adjusted dollars	Bureau (2021b)
NA	edu_health_social_industry	Percentage of employed population (≥ 16 yrs) in educational services, and health care and social assistance	Bureau (2021b)
Reduced	unemployment_rate	Percentage of unemployed civilian labor force	Bureau (2021b)
Reduced	total_staffed_adult_icu_ beds_7_day_avg	7-day average number of staffed inpatient adult ICU beds reported	HHS (2020)
Reduced	staffed_icu_adult_patients_ confirmed_covid_7_day_avg	7-day average number of adult ICU patients have laboratory-confirmed COVID-19	HHS (2020)
Reduced	binge	Prevalence of binge drinking among adults aged >=18 years	CDC (2021)
Reduced	chd	Prevalence of coronary heart disease among adults aged ≥ 18 years	CDC (2021)
Reduced	csmoking	Prevalence of current smoking among adults aged ≥ 18 years	CDC (2021)
Reduced	diabetes	Prevalence of diagnosed diabetes among adults aged ≥ 18 yrs	CDC (2021)
Reduced	lpa	Prevalence of no leisure-time physical activity among adults aged ≥ 18 yrs	CDC (2021)
NA	Administered_Dose1_ Pop_Pct	Percent of total population with at least one dose of COVID-19 vaccines	HHS (2021)

Table 2. Data dictionary for the full model. If a variable is included in the reduced model, it is default that it will be included in the full model. Only race-related variables were removed from the full model for the purpose of this study. Variables designated *NA* in the Model column were not included in either reduced or full model, yet was mentioned in this report.