COVID

MDPI

Article

Changes over Time in Association Patterns between Estimated COVID-19 Case Fatality Rates and Demographic, Socioeconomic and Health Factors in the US States of Florida and New York

Mansi Joshi [1], Yanming Di [1,*], Sharmodeep Bhattacharyya [1] and Shirshendu Chatterjee [2]

[1] Department of Statistics, Oregon State University, Corvallis, OR 97331, USA
[2] Department of Mathematics, The City University of New York, New York, NY 10031, USA
* Correspondence: diy@oregonstate.edu

Abstract: The United States struggled exceptionally during the COVID-19 pandemic. For researchers and policymakers, it is of great interest to understand the risk factors associated with COVID-19 when examining data aggregated at a regional level. We examined the county-level association between the reported COVID-19 case fatality rate (CFR) and various demographic, socioeconomic and health factors in two hard-hit US states: New York and Florida. In particular, we examined the changes over time in the association patterns. For each state, we divided the data into three seasonal phases based on observed waves of the COVID-19 outbreak. For each phase, we used tests of correlations to explore the marginal association between each potential covariate and the reported CFR. We used graphical models to further clarify direct or indirect associations in a multivariate setting. We found that during the early phase of the pandemic, the association patterns were complex: the reported CFRs were high, with great variation among counties. As pandemics progressed, especially during the winter phase, socioeconomic factors such as median household income and health-related factors such as the prevalence of adult smokers and mortality rate of respiratory diseases became more significantly associated with the CFR. It is remarkable that common risk factors were identified for both states.

Keywords: COVID-19; association; graphical model

Citation: Joshi, M.; Di, Y.; Bhattacharyya, S.; Chatterjee, S. Changes over Time in Association Patterns between Estimated COVID-19 Case Fatality Rates and Demographic, Socioeconomic and Health Factors in the US States of Florida and New York. *COVID* **2022**, 2, 1417–1434. https://doi.org/10.3390/covid2100102

Academic Editors: Guglielmo Campus and Dora Marinova

Received: 23 August 2022
Accepted: 28 September 2022
Published: 6 October 2022

1. Introduction

COVID-19 is a global challenge that demands researchers, policymakers, and governments to address multiple dimensions related to public health, socioeconomic impact, psychological impact, educational gap and many other issues [1]. The United States had high numbers of coronavirus cases and deaths, with high variability in cases and mortality among communities across the nation [2]. Researchers have identified risk factors for COVID-19 mortality, such as age and comorbid diseases [3]. Based on the CDC data as of 10 February 2021, 81.2% (359,956) of deaths were reported in the older population of age 65 and up in the United States [4]. There were also data suggesting that the social determinants of health such as poverty, physical environment (e.g., smoke exposure, homelessness) and race or ethnicity can also have a considerable effect on COVID-19 outcome [3,5,6]. COVID-19 has also disproportionately affected racial and ethnic minority groups, with high rates of death in African American, Native American and Latin populations [7]. At the regional level, it is important for policymakers to understand how the average COVID-19 risks are associated with the region's demographic, socioeconomic and health factors and how the association patterns are changing over time.

We are interested in, at the county level, how associations between risk factors and COVID-19 case fatality rate (CFR) have changed over the course of the pandemic in two

states, New York and Florida, that were hit hard by the pandemic. For each state, we divided the COVID-19 pandemic period into three seasonal phases based on the observed major peaks of COVID-19 outbreaks. We looked at county-level associations between demographic, socioeconomic and health factors and the reported COVID-19 CFR in the three phases from the beginning of the pandemic till mid-January 2021. This was the time period before COVID-19 vaccines became widely available. We used correlation tests to identify significant associations and used graphical models to untangle direct and indirect associations with the multivariate data comprising all the relevant covariates.

In this analysis, we observed that, during the early phases of the pandemic, the reported CFR was much higher than in later phases, and association patterns were complex and suggest that risk factors were multifaceted. The percentage of the population aged 65 and older was associated with reported CFR in Florida but not in New York. As the pandemic progressed, especially during the winter phase, socioeconomic factors such as median household income and health factors such as the prevalence of adult smoking and fatality rate due to respiratory diseases became more significantly associated with the reported CFR. We used causal graphical models [8,9] to sort the associated factors according to which ones were more likely to be directly associated with the COVID-19 case fatality rate. The limitation of observational studies does not allow us to make causal conclusions, but trends and changes in association patterns furnish important data-supported hypotheses for understanding the pandemic in depth. It was remarkable that there was as much commonality in association patterns during the third phase between the two states, even though the two states implemented different mitigation measures and were commonly contrasted in news reports.

2. Materials and Methods

2.1. Data Sources and Preparation

County-level, daily reported COVID-19 cases and deaths numbers were obtained from USAFacts [10]. For each county, we computed the reported CFR as the ratio of the reported confirmed deaths to the reported confirmed cases. As discussed in Angelopoulos et al. [11], this simple and "naïve" estimation of CFR has limitations since, for example, both cases and deaths can be under-reported, but compensating for such biases would be difficult without collecting substantial additional data. This naïve estimation is still an informative and practical measure of the severity of an ongoing pandemic.

For New York and Florida, we divided the pandemic period up until 15 January 2021 into three seasonal phases based on the observed major peaks in each state (see Figures 1 and 2). We computed the reported CFR during each phase for the counties of New York and Florida. The first phase in New York included the days from 1 March to 30 June 2020, and the first phase in Florida included the days from March 1 to 31 May 2020. The second phase in New York was the period between 1 July and 30 September. The second phase in Florida included the days from 1 June to 30 September 2020. The third phases in both states were the days between 1 October 2020 and 15 January 2021, which was the period when two major holidays, Thanksgiving and Christmas, likely increased indoor gatherings. Since both deaths and cases were aggregated over each phase, the estimated CFR was less affected by the reporting lag.

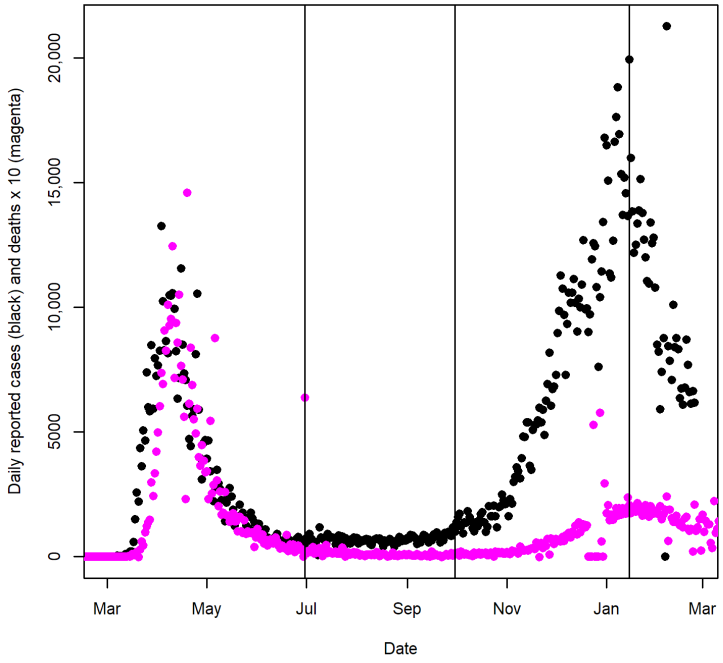

Figure 1. Total numbers of daily reported COVID-19 cases and deaths (multiplied by 10) in New York State. The three vertical lines mark the dates: 30 June 2020, 30 September 2020 and 15 January 2021.

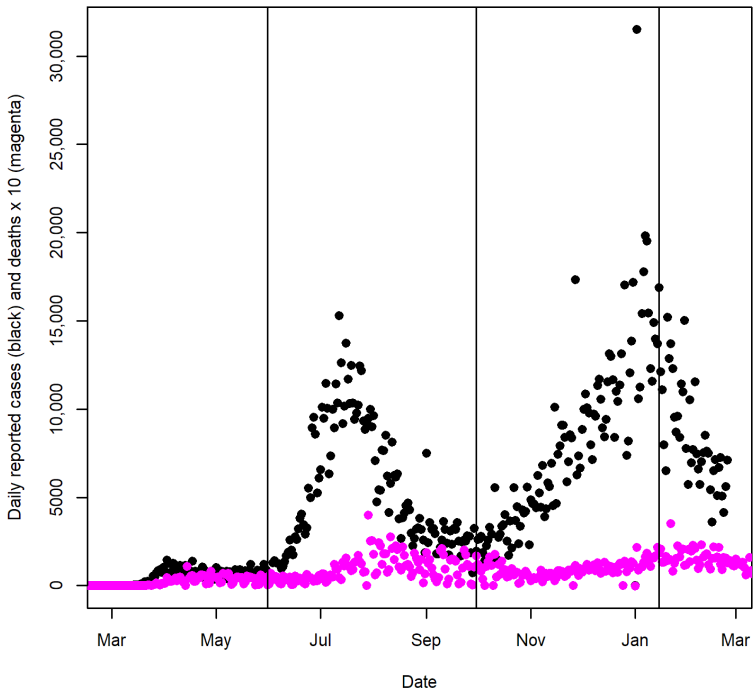

Figure 2. Total numbers of daily reported COVID-19 cases and deaths (multiplied by 10) in Florida State. The three vertical lines mark the dates: 31 May 2020, 30 September 2020 and 15 January 2021.

We obtained covariate data reflecting various potential demographic, socioeconomic and health risk factors from the County Health Rankings 2021 and 2022 Analytics Datasets [12,13] and an abridged dataset curated by the Yu group [14]. Table 1 summarizes the 21 covariates used in our analyses in detail. The 2022 County Health Rankings data updated the health-related covariates in our list to reflect 2019's measurements.

The datasets were merged by the Federal Information Processing Standards (FIPS) code for each county to create a final dataset for the statistical analysis in the R software (R Core Team, Vienna, Austria) [15].

Table 1. Description of all covariates used in the analyses. The first column lists the short variable names used in our analyses and summaries (e.g., tables and figures). The second column gives a detailed description of each variable. The third column lists the years of the data.

Variable	Description	Years of Data
age65+	% of population 65 or older	2019
age18−	% of population below 18 years of age	2019
Black	% Non-Hispanic Black	2019
AIAN	% American Indian and Alaska Native	2019
Asian	% Asian	2019
NHPI	% Native Hawaiian/Other Pacific Islander	2019
Hispanic	% Hispanic	2019
White	% Non-Hispanic White	2019
income	Median household income	2019
housing	Severe housing problems: % of households with at least 1 of 4 housing problems: overcrowding, high housing costs, lack of kitchen facilities or lack of plumbing facilities	2015–2019
unemployment	% of population aged 16 and older unemployed but seeking work	2019
uninsured	% of adults under age 65 without health insurance	2019
SVI	Social vulnerability index	2018
obesity	% of the adult population (age 18 and older) that reports a body mass index (BMI) greater than or equal to 30 kg/m^2 (age-adjusted)	2019
smoking	% of adults who are current smokers (age-adjusted)	2019
drinking	% of adults reporting binge or heavy drinking (age-adjusted)	2019
diabetes	% of adults aged 20 and above with diagnosed diabetes (age-adjusted)	2019
HIV	Number of people aged 13 years and older living with a diagnosis of human immunodeficiency virus (HIV) infection per 100,000 population	2019
heart	Heart disease mortality rate	2014–2016
stroke	Stroke mortality rate	2014–2016
respiratory	Chronic respiratory disease mortality rate	2014

2.2. Statistical Analysis

For univariate analysis, we conducted tests of correlations between the 21 covariates and the reported CFR for each COVID-19 phase among counties in the states of New York and Florida. We explored both the Pearson correlation and Spearman's rank correlation. The test of correlation using the Pearson correlation coefficient is equivalent to the test of the regression coefficient in a univariate simple linear regression. Spearman's rank correlation was also considered since it is less sensitive to the influence of outlying data points. Exploratory visualization was used to examine the correlations between the reported CFR and the covariates.

Tetrad software (Ramsey et al, Pittsburgh, USA) 7.1.0-0 [16] was used to build graphical models to examine associations between the covariates and the reported CFR during the third COVID-19 phase in each state. The Greedy Fast Casual Inference Algorithm for continuous random variables (GFCIc) [17] was used to build the graphical models. GFCIc takes a dataset of continuous variables as input and outputs a graphical model called a partial ancestral graph (PAG), which represents a set of causal Bayesian networks that cannot be distinguished by the algorithm. The GFCIc assumes that cases in the data are independent and identically distributed. The formal interpretation of nodes and edges in PAG is listed in Table 2 [18]. At a more intuitive level, in the large sample limit, the PAG [19] output by GFCIc will entail the set of conditional independence relationships judged to hold in the population represented by its input dataset. With a finite sample size, the resulting graphical model may not be completely accurate, but one can use the PAG that GFCIc

returns as a data-supported hypothesis about causal relationships that exist among the variables in the dataset.

Table 2. Interpretation of nodes and edges in the graphical models (based on Tetrad's manual). If an edge is green, it means there is no latent confounder. Otherwise, there is possibly a latent confounder. If an edge is bold (thickened), it means it is definitely direct. Otherwise, it is possibly direct.

Edge Type	Relationships That Are Present	Relationships That Are Absent
A —> B	A is a cause of B. It may be a direct or indirect cause that may include other measured variables. Further, there may be an unmeasured confounder of A and B.	A is a cause of B.
A <—> B	There is an unmeasured confounder (call it L) of A and B. There may be measured variables along the causal pathway from L to A or from L to B.	A is not a cause of B. B is not a cause of A.
A o—>B	Either A is the cause of B (i.e., A —>B) or there is an unmeasured confounder of A and B (i.e., A<—>B) or both.	B is not a cause of A.
A o—o B	Exactly one of the following holds: 1. Ais a cause of B. 2. B is a cause of A. 3. there is an unmeasured confounder of A and B. 4. both 1 and 3. 5. both 2 and 3.	

3. Results

3.1. New York State

In order to explore potential county-level demographic, socioeconomic and health risk factors of COVID-19 during each of the three seasonal phases (see Figure 1), we performed tests of correlation between the reported CFR in each phase and the county-level covariates listed in Table 1. We will focus our discussion on test results based on Spearman's rank correlation coefficients since the rank correlation is less sensitive to the influence of a small number of outlying data points. Results based on the Pearson correlation coefficient—which is equivalent to the test of the regression coefficient in a univariate simple linear regression—are listed in Appendix A, Table A1.

According to Spearman's rank correlation test results (listed in Table 3), during the first phase (1 March to 30 June of 2020) of COVID-19, among counties in New York State, the reported CFR was significantly correlated with the prevalence of an obese population, and also moderately correlated with stroke and respiratory diseases mortality rates. However, the signs of the correlation coefficients are all counter-intuitive: for example, counties with higher stroke mortality rates actually tended to have lower reported CFRs for COVID-19 during this period. This is likely due to confounding factors: for example, one contributing factor is that the virus hit major cities first, at a time when the county was not yet prepared to respond. To visually inspect the trends, we plotted the reported CFR against the covariates (Figures 3–5). By examining the scatter plots, we see that the reported CFRs were much higher than in later phases. As we know now, COVID-19 tests were not readily available during this period, and it was likely patients with severe symptoms were more likely to be diagnosed and recorded. The reported CFRs also showed great variation among counties, and for many covariates, the scatter plots display nonlinear and nonmonotonic trends that cannot be captured by Spearman's rank correlation.

During the second phase (1 July to 30 September 2020), increased COVID-19 CFR at the county level was associated with an increased percentage of the black population and decreased percentage of the white population, increased SVI, increased prevalence of diabetes, increased HIV prevalence and increased heart mortality rate. Note that New York State did not experience a high peak during this phase.

During the third phase (1 October 2020 to 15 January 2021), increased COVID-19 CFR at the county level was associated with many measurements of racial compositions, decreased median household income, increased prevalence of adult smokers, increased HIV prevalence and increased stroke and respiratory disease mortality rates. The negative

association with housing problems was likely due to confounding. The noise level was much lower during this phase. From the scatter plots, we see that overall, the reported CFRs had greatly decreased in phase 3. When it comes to median household income, the reduction in CFR among the high-income counties was greater than among the low-income counties.

Table 3. Test for correlation between the reported CFR and covariates among New York counties for the three COVID-19 phases. The columns are Spearman rank correlations and corresponding test *p*-values.

	Phase 1		**Phase 2**		**Phase 3**	
	r_s	p_s	r_s	p_s	r_s	p_s
age65+	0.087	0.500	−0.161	0.210	0.227	0.077
age18−	0.065	0.616	0.209	0.103	0.091	0.479
Black	0.134	0.300	0.282	**0.026**	−0.275	**0.031**
AIAN	−0.165	0.199	0.210	0.101	−0.091	0.481
Asian	0.241	0.059	0.245	0.055	−0.261	**0.041**
NHPI	0.214	0.095	0.225	0.079	−0.430	**0.001**
Hispanic	0.101	0.436	0.226	0.077	−0.377	**0.003**
White	−0.096	0.457	−0.259	**0.042**	0.316	**0.013**
income	0.223	0.082	−0.214	0.095	−0.387	**0.002**
housing	0.151	0.241	0.119	0.356	−0.323	**0.011**
unemployment	−0.116	0.369	0.201	0.117	0.218	0.088
uninsured	0.154	0.233	0.225	0.079	−0.158	0.220
SVI	−0.081	0.532	0.436	**0.000**	0.012	0.924
obesity	−0.331	**0.009**	0.053	0.682	0.235	0.066
smoking	−0.186	0.148	0.037	0.775	0.421	**0.001**
drinking	−0.133	0.302	−0.194	0.131	0.126	0.327
diabetes	−0.035	0.789	0.402	**0.001**	−0.015	0.910
HIV	0.159	0.226	0.269	**0.038**	−0.292	**0.024**
heart	0.151	0.242	0.338	**0.007**	0.149	0.248
stroke	−0.254	**0.047**	−0.177	0.168	0.503	**0.000**
respiratory	−0.276	**0.030**	−0.139	0.282	0.316	**0.013**

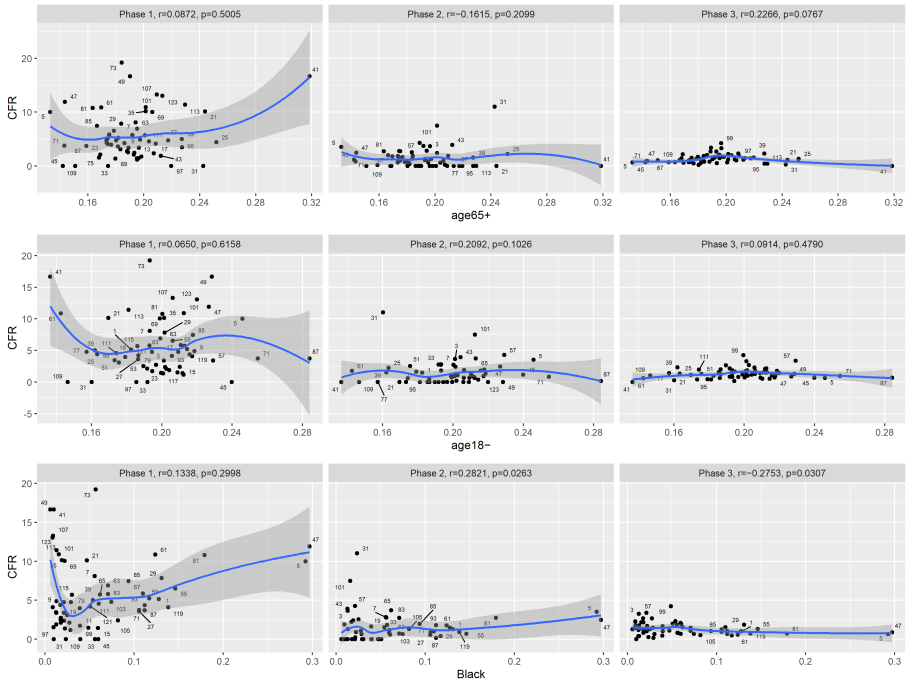

Figure 3. *Cont.*

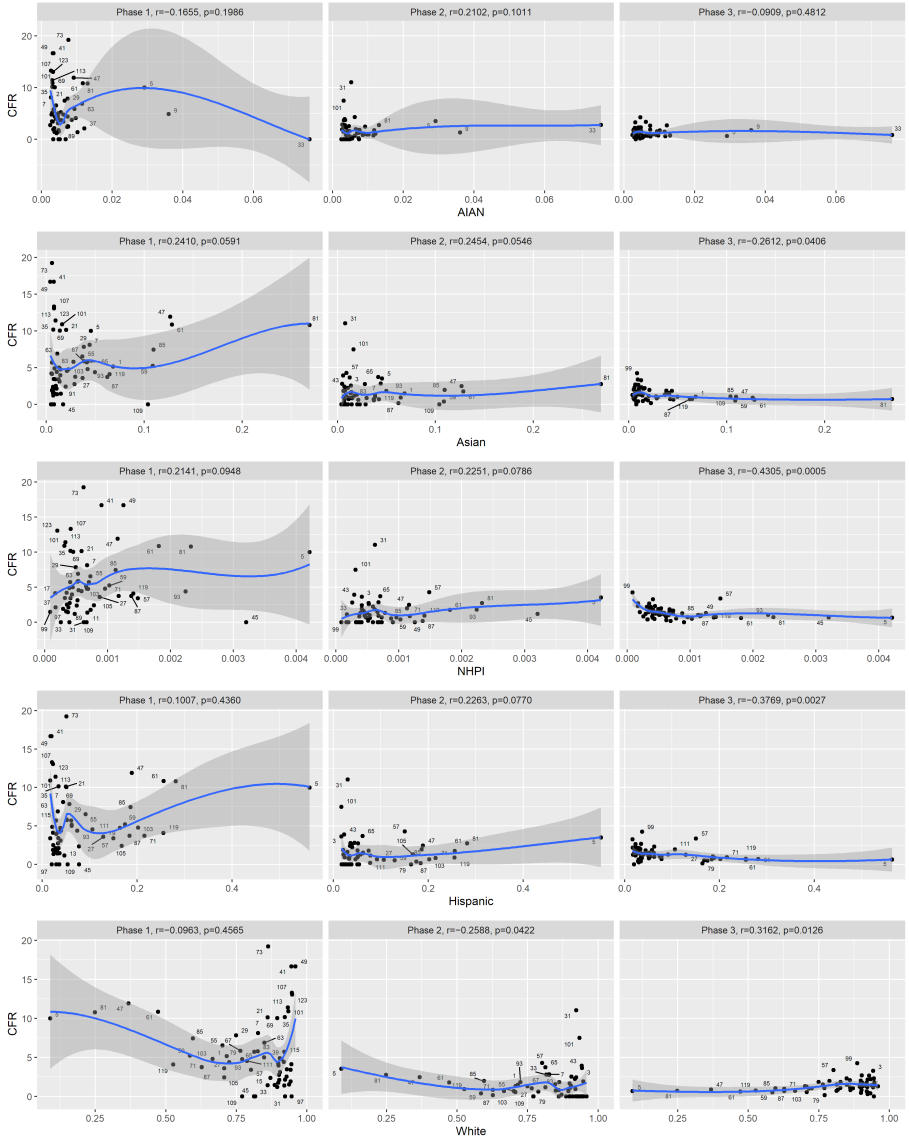

Figure 3. Scatter plots of the reported CFR versus demographic covariates among New York counties during the three COVID-19 phases. The *y*-axes are the reported CFR. The *x*-axes in the plots (from top to bottom) are age65+, age18−, Black, AIAN, Asian, NHPI, Hispanic and White (See Table 1 for detailed descriptions of the variables).

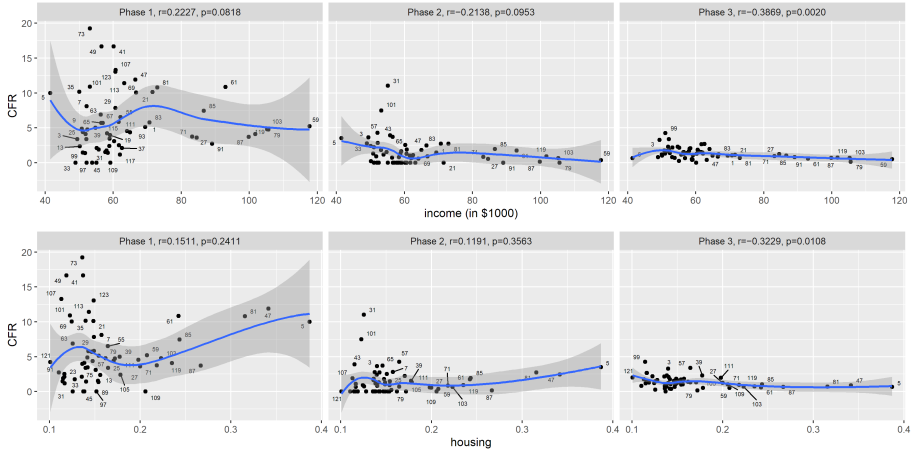

Figure 4. *Cont.*

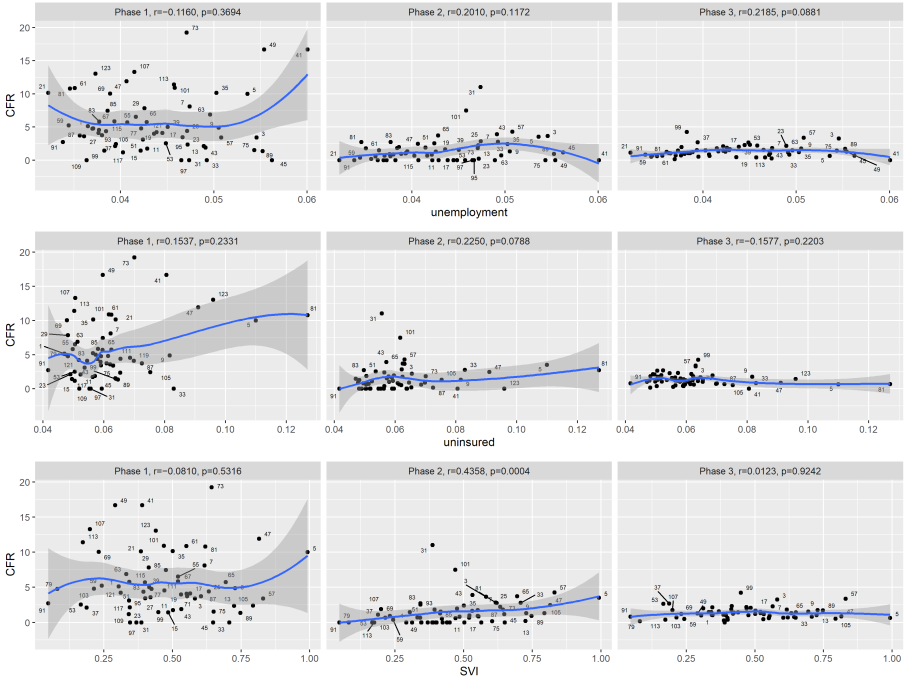

Figure 4. Scatter plots of the reported CFR versus socioeconomic covariates among New York counties during the three COVID-19 phases. The *y*-axes are the reported CFR. The *x*-axes in the plots (from top to bottom) are income, housing, unemployment, uninsured and SVI (See Table 1 for detailed descriptions of the variables).

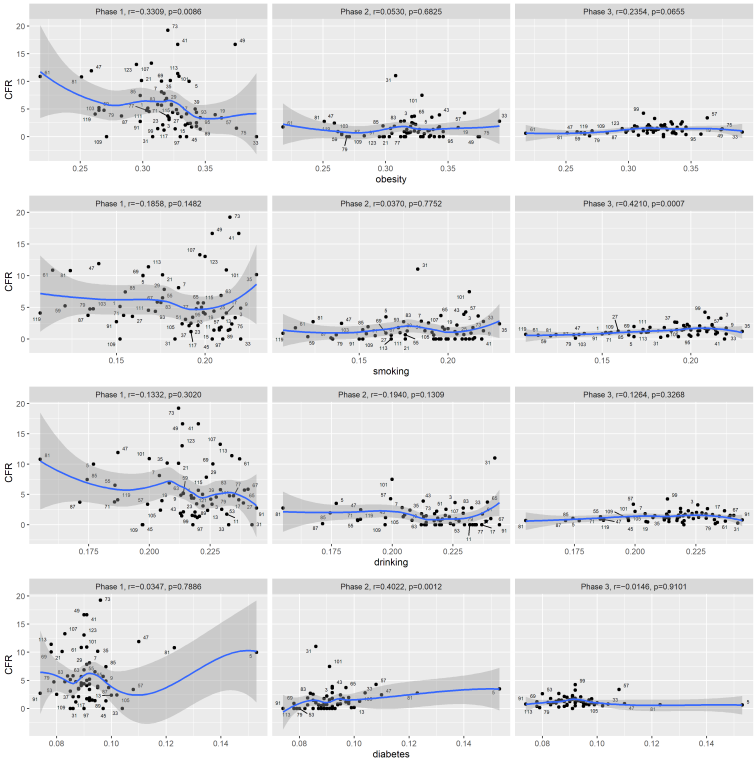

Figure 5. *Cont.*

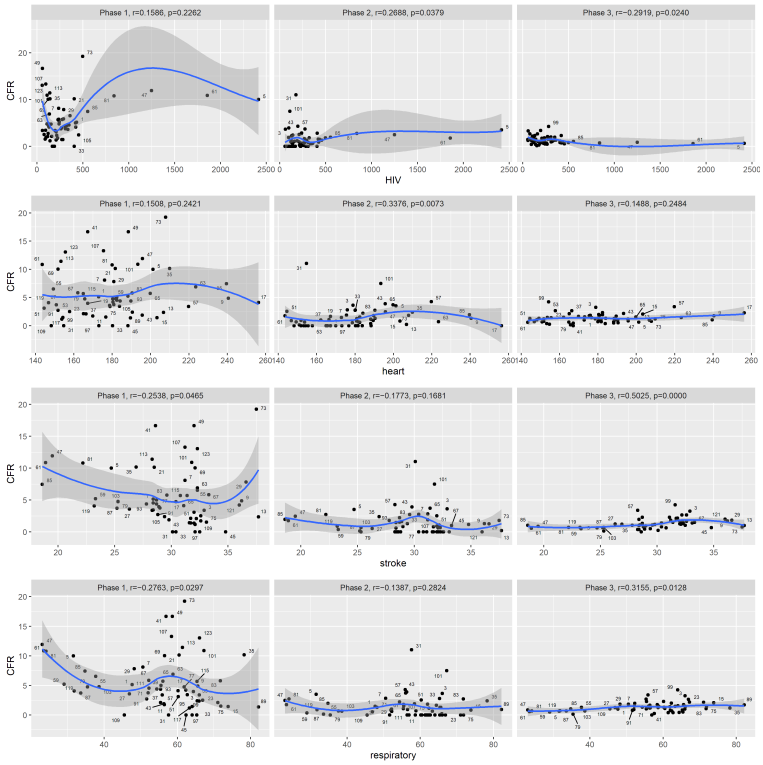

Figure 5. Scatter plots of the reported CFR versus health covariates among New York counties during the three COVID-19 phases. The y-axes are the reported CFR. The x-axes in the plots (from top to bottom) are obesity, smoking, drinking, diabetes, HIV, heart, stroke and respiratory (See Table 1 for detailed descriptions of the variables).

We fit graphical models to the CFR and covariates data from the third phase using the software Tetrad. With a finite sample size, the resulting graphical model may not be completely accurate and should be viewed as a data-supported hypothesis about causal relationships that exist among the variables. Intuitively speaking, the factors that are closer to the CFR in the resulting graphical model are more (likely to be) directly associated with the CFR (see Table 2 for technical interpretation of the graph). The graphical model (Figure 6) shows that the percentage of adult smokers is the only covariate that is directly associated with the reported CFR. In other words, the graphical model indicates that once we apply the percentage of adult smokers, other covariates are no longer significantly correlated with the reported CFR. This suggests that data from New York counties are most compatible with the interpretation that the other significant covariates reported from the correlation analysis are indirectly associated with CFR through the effect of smoking.

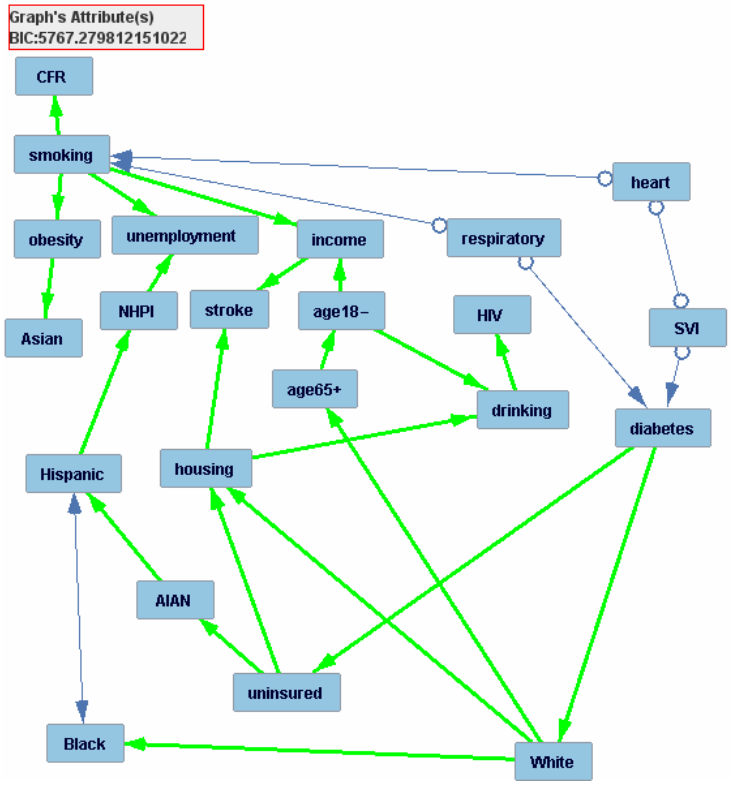

Figure 6. Graphical model for phase 3 of New York State. Nodes closer to CFR are more likely to be directly associated with the risk of CFR. See Table 2 for technical interpretation of the edges.

3.2. Florida State

The results from the test of Spearman correlations are summarized in Table 4 for the three phases of Florida (The results from the test of Pearson correlations are summarized in Appendix A Table A2). Among Florida counties, during the first phase of COVID-19 (1 March to 31 May 2020), the reported CFR was significantly correlated with the percentage of the population aged 65 and older and the percentage of the AIAN population. The reported CFR was also moderately associated with the prevalence of obesity and the fatality rate of respiratory diseases, but similar to New York State's case, the signs of correlations were counter-intuitive and were likely due to confounding and also the nonmonotonic nature of the curves, which Spearman's correlation can not handle. Note that Florida only experienced a minor peak during this phase.

During the second phase (1 June to 30 September 2020), the reported CFR was significantly correlated with the percentage of the population aged 65 or older, the percentage of the population aged 18 or younger, the percentage of the black population and the prevalence of obesity. During the third phase (1 October 2020 to 15 January 2021), the percentage of the population aged 65 or older was still significantly correlated with the reported CFR, but the correlation was less significant compared to the first two phases. The reported CFR was significantly correlated with median household income, unemployment rate, the prevalence of smokers and the fatality rate of respiratory diseases.

In Figures 7–9, we plotted the reported CFR in Florida counties against demographic, socioeconomic and health-related covariates.

Table 4. Test for correlation between the reported CFR and covariates among Florida counties for the three COVID-19 phases. Listed are Spearman rank correlations and the corresponding test *p*-values.

	Phase 1		**Phase 2**		**Phase 3**	
	r_s	p_s	r_s	p_s	r_s	p_s
age65+	0.340	**0.005**	0.477	**0.000**	0.286	**0.019**
age18−	−0.210	0.088	−0.280	**0.022**	−0.189	0.125
Black	−0.128	0.302	−0.248	**0.044**	−0.063	0.611
AIAN	−0.347	**0.004**	−0.221	0.072	0.235	0.056
Asian	0.146	0.237	0.181	0.142	−0.335	**0.006**
NHPI	−0.083	0.504	0.082	0.510	−0.112	0.365
Hispanic	0.029	0.815	0.179	0.147	−0.195	0.115
White	0.112	0.366	0.096	0.439	0.237	0.054
income	0.235	0.055	0.170	0.167	−0.304	**0.013**
housing	0.130	0.295	−0.063	0.611	−0.240	0.051
unemployment	0.010	0.937	0.241	0.050	0.358	**0.003**
uninsured	0.090	0.469	0.044	0.720	0.039	0.756
SVI	−0.220	0.074	−0.217	0.078	0.220	0.073
obesity	−0.267	**0.029**	−0.365	**0.002**	0.151	0.222
smoking	−0.203	0.099	−0.218	0.076	0.412	**0.001**
drinking	0.201	0.103	0.233	0.058	0.060	0.632
diabetes	−0.205	0.097	−0.236	0.054	0.100	0.421
HIV	0.056	0.653	−0.120	0.332	−0.136	0.271
heart	−0.217	0.077	−0.116	0.349	0.198	0.108
stroke	−0.216	0.079	−0.182	0.141	−0.032	0.799
respiratory	−0.252	**0.039**	−0.221	0.072	0.328	**0.007**

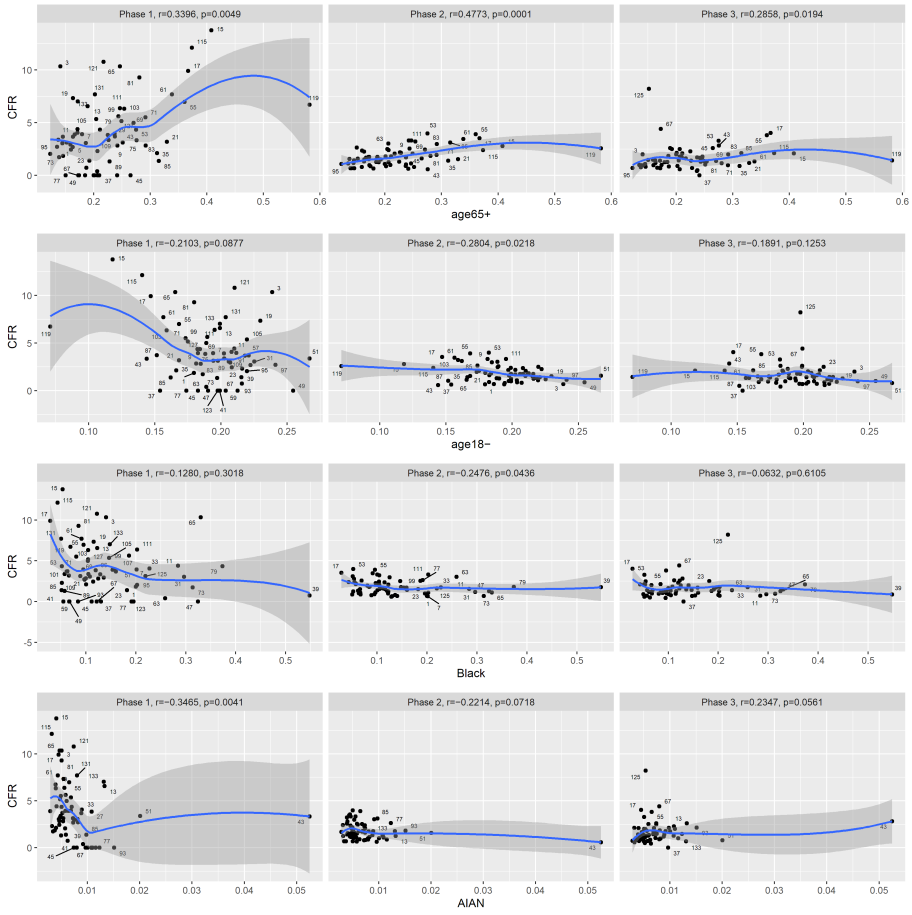

Figure 7. *Cont.*

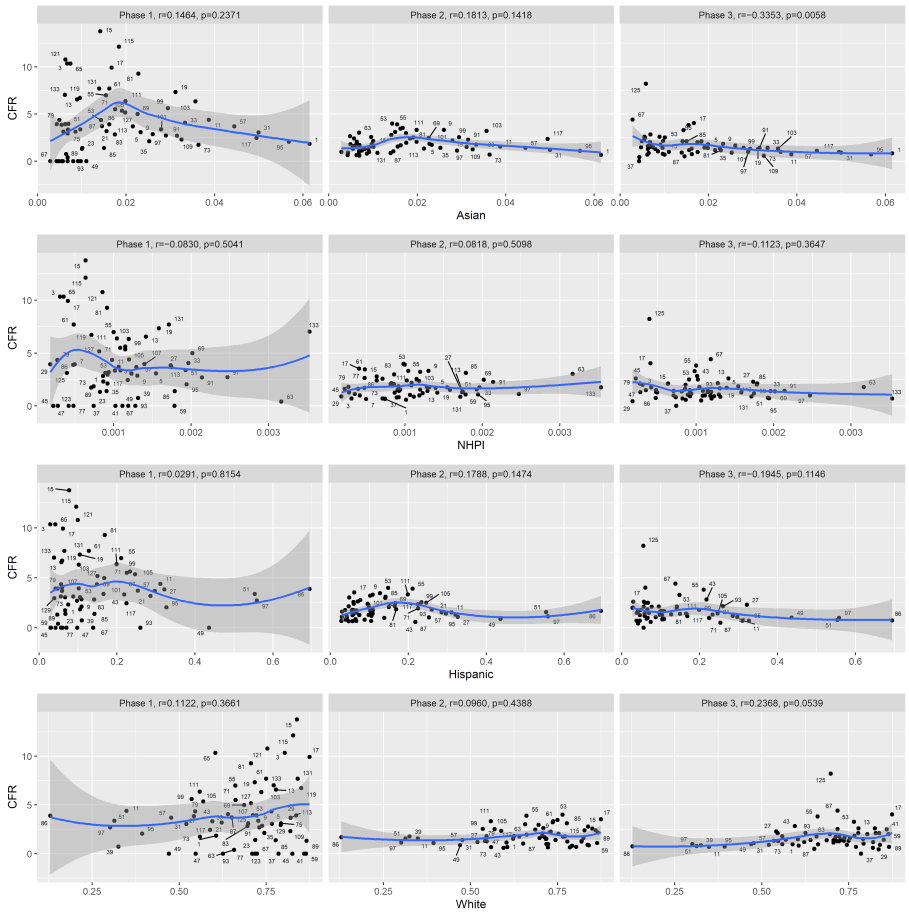

Figure 7. Scatter plots of the reported CFR versus demographic covariates among Florida counties during the three COVID-19 phases. The *y*-axes are the reported CFR. The *x*-axes in the plots (from top to bottom) are age65+, age18−, Black, AIAN, Asian, NHPI, Hispanic and White (See Table 1 for detailed descriptions of the variables).

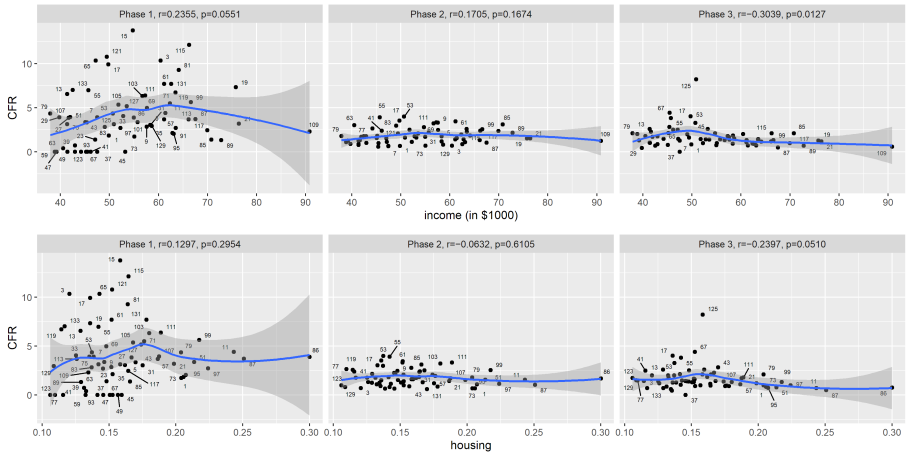

Figure 8. *Cont*.

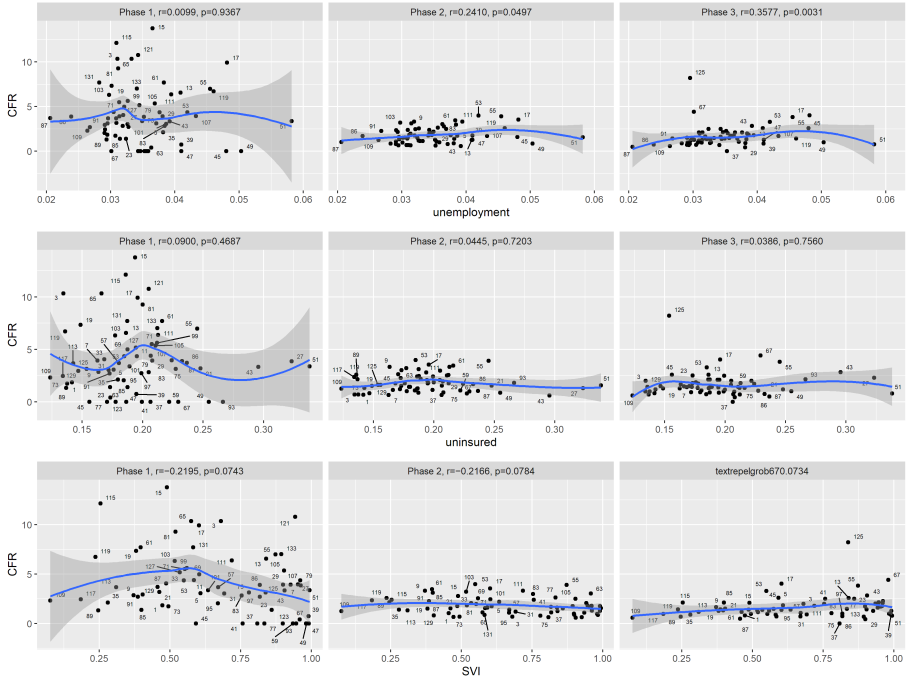

Figure 8. Scatter plots of the reported CFR versus socioeconomic covariates among Florida counties during the three COVID-19 phases. The *y*-axes are the reported CFR. The *x*-axes in the plots (from top to bottom) are income, housing, unemployment, uninsured and SVI (See Table 1 for detailed descriptions of the variables).

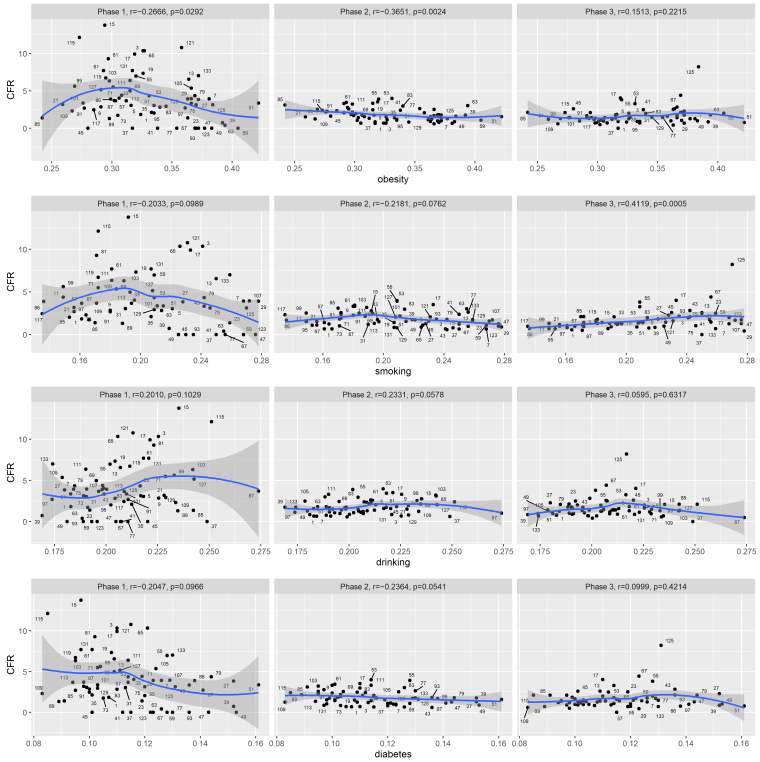

Figure 9. *Cont.*

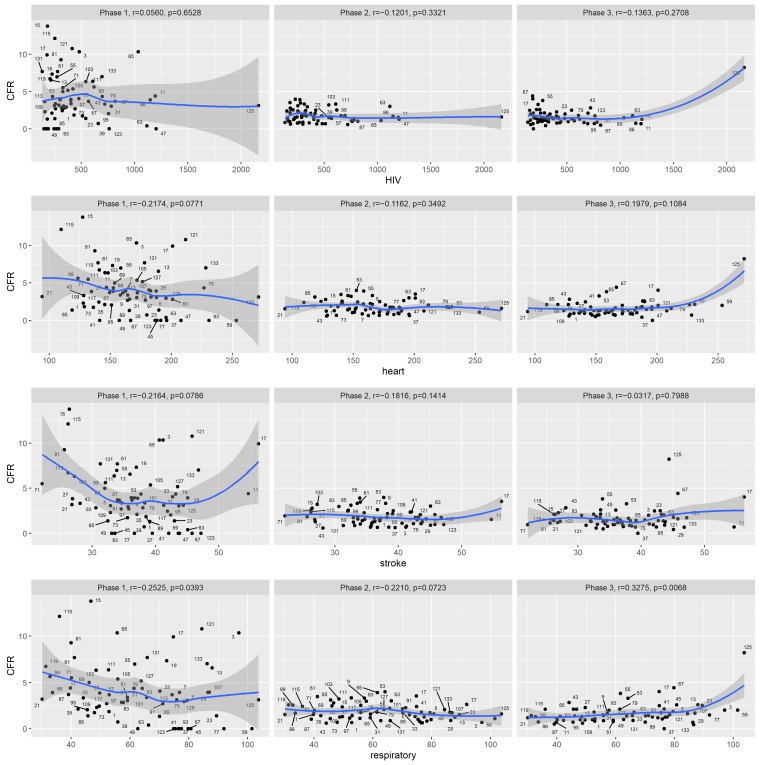

Figure 9. Scatter . plots of the reported CFR versus health covariates among Florida counties during the three COVID-19 phases. The *y*-axes are the reported CFR. The *x*-axes in the plots (from top to bottom) are obesity, smoking, drinking, diabetes, HIV, heart, stroke and respiratory (See Table 1 for detailed descriptions of the variables).

The graphical model (Figure 10) estimated using Tetrad shows that for the third phase, the prevalence of adult smokers is the only covariate that is directly associated with the reported CFR. In other words, once we adjust for the effect of this covariate, other covariates are no longer significant. This agrees with New York State's graphical model.

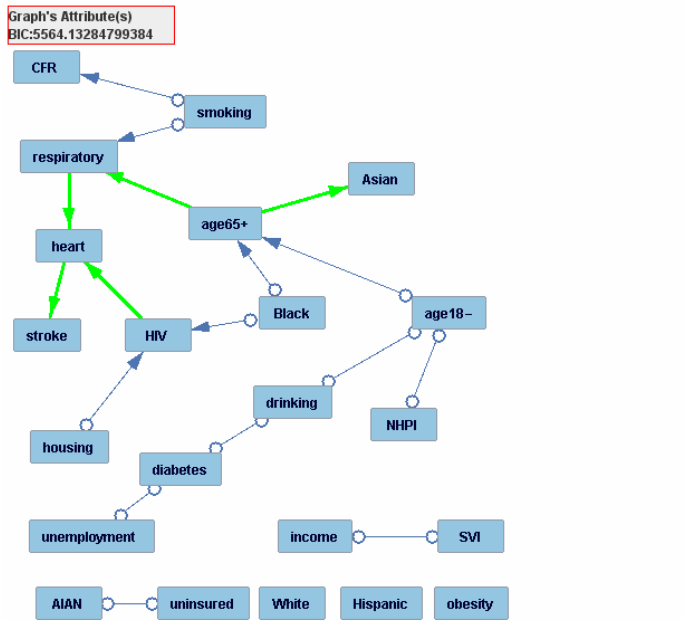

Figure 10. Graphical model for phase 3 of Florida State. Nodes closer to CFR tend to be more directly associated with the risk of CFR. See Table 2 for technical interpretation of the edges.

4. Discussion

In this study, we examined how the association patterns between demographic, socioeconomic and health factors and the reported CFR changed over time in the states of New York and Florida during the COVID-19 pandemic up until mid-January 2021. In each state, we divided the pandemic period into three phases. It is understandable that data from the early phase of the COVID-19 pandemic were noisy and did not display clear association patterns. Many factors contributed to the uncertainty in the observed CFR: diagnostic tests were not widely available, knowledge on effective treatments was limited, steps had not yet been taken to protect the vulnerable groups, and state-wide lockdown protocols were still being discussed. In the early phases, there was evidence in Florida State that the age distribution of the county residents was associated with the reported CFR early in the pandemic. As the pandemic progressed, the association with age distribution became weaker. However, the association with age distribution was not significant in New York.

As the pandemic progressed, the data became much less noisy, and interesting association patterns started to emerge. In the third phase of both states, the reported CFR was more associated with socioeconomic factors and health factors. The graphical model suggests that in both states, the prevalence of adult smokers was most directly associated with reported CFR. It is not surprising that the median household income was associated with the CFR during the third phase in both states. A cnbc.com article [20] pointed out that low-income workers tended to work in sectors that were considered "essential" during the pandemic, and their jobs often required them to be on-site. These factors increased their risk of being infected by COVID-19. In the third phase, lockdown restrictions were lighter compared to the earlier two phases, but increased economic activities also meant an increased possibility of those essential workers being infected with COVID-19. People living in more socio-economically disadvantaged neighborhoods and minority ethnic groups have higher rates of almost all the known underlying clinical risk factors associated with the severity and mortality of COVID-19, including hypertension, diabetes, asthma, chronic obstructive pulmonary disease (COPD), heart disease, liver disease, renal disease, cancer and cardiovascular disease [6], so an exact causal relationship is difficult to infer from observational data.

We considered different multiple regression models, but we decided to use graphical models for explicability and simplicity. The graphical model for the CFR and the covariates provides suggestive evidence regarding direct or indirect associations. For spatial considerations, we explored clustering of the counties based on the observed death counts from the first wave. There were a few challenges. For example, the cases and death numbers were highly uneven across counties. The pandemic reached different regions at different times. The confounding factors affecting the association analysis would also impact the clustering analysis. Further, while clustering allows us to group the counties, it does not directly reveal the factors underlying the groups. For temporal considerations, by aggregating counts over time in our association analysis, we alleviated the impact of some of the potential confounding factors.

The associations we discovered are not causation: they should be viewed as data-supported hypotheses. The verification of the causal hypotheses will be more challenging, as is usually the case: one can eliminate some unlikely causal links by testing conditional independences on new datasets, but ultimately randomized experiments will be needed to confirm a causal relation.

There is no simple unit measure for the severity of COVID-19. For example, for our purpose, the population fatality rate would not be a good measure since when the pandemic first hit, it only affected a few counties. For this reason, we chose the reported CFR in this study. In fact, the impact of COVID-19 on society is multifaceted. Our study reflects one aspect of the pandemic.

5. Conclusions

As pointed out by many, during the COVID-19 pandemic, vulnerable groups are restricted to not only elderly people but also people with ill health and/or comorbidities. Socioeconomic conditions can also have a considerable effect on COVID-19 outcomes. In this work, we highlighted how the associations between demographic, health and socioeconomic factors and the reported CFR had changed over time as the pandemic progressed. We examined data up to January of 2021, the period before vaccination became widely available. We saw that during the early phase of the pandemic, the reported CFR was high, and the data were highly noisy. It is likely that many factors contributed to the high reported CFR: the lack of tests, the lack of effective treatment and the lack of mitigation strategies. During this phase, it is hard to clearly identify simple risk factors and simple risk groups. As the pandemic progressed, the usual suspect factors started to emerge as significant risk factors for the report CFR. It is quite remarkable that during the third phase of the pandemic, New York and Florida actually shared common risk factors for the CFR: the two states implemented quite different mitigation measures. In future work, we would like to examine the impact of vaccination on COVID-19 and the demographic, health and socioeconomic variables.

Author Contributions: Conceptualization, M.J., Y.D., S.B. and S.C.; methodology, M.J., Y.D., S.B. and S.C.; software, M.J. and Y.D.; formal analysis, M.J. and Y.D.; visualization, M.J. and Y.D.; investigation, M.J., Y.D., S.B. and S.C.; data curation, Y.D.; writing—original draft preparation, M.J. and Y.D.; writing—review and editing, M.J., Y.D., S.B. and S.C.; supervision, Y.D., S.B. and S.C. All authors have read and agreed to the published version of the manuscript.

Funding: S.C. was funded by NSF DMS Award #2154564.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and R code used for the correlation tests and the scatter plots are available on the GitHub page: https://github.com/diystat/covid-association-study (accessed on 4 August 2022).

Acknowledgments: We would like to thank participants of the COVID-19 discussion group at Oregon State University. We thank the two reviewers for their constructive comments.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CFR	case fatality rate
AIAN	American Indian and Alaska Native
PAG	partial ancestral graph
SVI	Social vulnerability index

Appendix A. Test of Pearson correlations

Table A1. Test for correlation between the reported CFR and covariates among New York counties for the three COVID phases. Listed are Pearson correlations and corresponding test *p*-values.

	Phase 1		**Phase 2**		**Phase 3**	
	r	***p***	***r***	***p***	***r***	***p***
age65+	0.199	0.120	0.036	0.782	0.019	0.885
age18−	−0.017	0.893	0.007	0.959	0.087	0.500
Black	0.185	0.149	0.091	0.482	−0.255	**0.045**
AIAN	−0.116	0.370	0.121	0.350	−0.091	0.484
Asian	0.182	0.156	0.068	0.601	−0.277	**0.029**
NHPI	0.144	0.265	0.136	0.292	−0.316	**0.012**
Hispanic	0.133	0.303	0.087	0.500	−0.311	**0.014**
White	−0.164	0.202	−0.096	0.459	0.321	**0.011**
income	0.023	0.861	−0.226	0.078	−0.402	**0.001**
housing	0.169	0.189	0.056	0.667	−0.328	**0.009**
unemployment	0.054	0.674	0.200	0.120	0.129	0.318
uninsured	0.319	**0.012**	0.148	0.251	−0.184	0.153
SVI	−0.022	0.864	0.281	**0.027**	0.037	0.773
obesity	−0.208	0.105	0.046	0.720	0.271	**0.033**
smoking	−0.074	0.568	0.088	0.497	0.427	**0.001**
drinking	−0.162	0.207	−0.057	0.661	0.123	0.342
diabetes	0.115	0.374	0.253	**0.048**	−0.057	0.659
HIV	0.297	**0.021**	0.145	0.269	−0.250	0.054
heart	0.101	0.435	0.127	0.326	0.149	0.247
stroke	−0.166	0.198	−0.094	0.469	0.376	**0.003**
respiratory	−0.211	0.100	−0.022	0.864	0.348	**0.006**

Table A2. Test for correlation between the reported CFR and covariates among Florida counties for the three COVID phases. Listed are Pearson correlations and corresponding test *p*-values.

	Phase 1		**Phase 2**		**Phase 3**	
	r	***p***	***r***	***p***	***r***	***p***
age65+	0.424	**0.000**	0.501	**0.000**	0.122	0.327
age18−	−0.314	**0.010**	−0.337	**0.005**	−0.127	0.305
Black	−0.188	0.127	−0.209	0.089	−0.051	0.681
AIAN	−0.186	0.133	−0.237	0.054	0.110	0.376
Asian	0.003	0.979	0.021	0.864	−0.310	**0.011**
NHPI	−0.126	0.311	0.061	0.623	−0.182	0.140
Hispanic	−0.090	0.470	−0.028	0.822	−0.195	0.114
White	0.196	0.111	0.159	0.198	0.223	0.069
income	0.201	0.103	0.108	0.386	−0.247	**0.044**
housing	0.031	0.805	−0.117	0.348	−0.207	0.093
unemployment	−0.012	0.925	0.230	0.061	0.162	0.190
uninsured	−0.033	0.790	−0.016	0.896	0.033	0.790
SVI	−0.240	0.050	−0.206	0.094	0.223	0.070
obesity	−0.274	**0.025**	−0.330	**0.006**	0.195	0.114
smoking	−0.202	0.101	−0.218	0.077	0.372	**0.002**
drinking	0.264	**0.031**	0.200	0.104	0.028	0.825
diabetes	−0.285	**0.020**	−0.233	0.058	0.125	0.313
HIV	−0.066	0.594	−0.135	0.277	0.305	**0.012**
heart	−0.210	0.088	−0.135	0.275	0.339	**0.005**
stroke	−0.189	0.125	−0.117	0.345	0.148	0.233
respiratory	−0.211	0.086	−0.219	0.075	0.369	**0.002**

References

1. Lambert, H.; Gupte, J.; Fletcher, H.; Hammond, L.; Lowe, N.; Pelling, M.; Raina, N.; Shahid, T.; Shanks, K. COVID-19 as a global challenge: Towards an inclusive and sustainable future. *Lancet Planet. Health* **2020**, *4*, e312–e314. [CrossRef]
2. Dong, E.; Du, H.; Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **2020**, *20*, 533–534. [CrossRef]
3. Hawkins, R.B.; Charles, E.J.; Mehaffey, J.H. Socio-economic status and COVID-19–related cases and fatalities. *Public Health* **2020**, *189*, 129–134. [CrossRef] [PubMed]
4. CDC. Provisional Death Counts for Coronavirus Disease 2019 (COVID-19). 2021. Available online: https://www.cdc.gov/nchs/covid19/mortality-overview.htm (accessed on 1 July 2022).
5. Abrams, E.M.; Szefler, S.J. COVID-19 and the impact of social determinants of health. *Lancet Respir. Med.* **2020**, *8*, 659–661. [CrossRef]
6. Bambra, C.; Riordan, R.; Ford, J.; Matthews, F. The COVID-19 pandemic and health inequalities. *J. Epidemiol. Community Health* **2020**, *74*, 964–968. [CrossRef] [PubMed]
7. Tai, D.B.G.; Shah, A.; Doubeni, C.A.; Sia, I.G.; Wieland, M.L. The disproportionate impact of COVID-19 on racial and ethnic minorities in the United States. *Clin. Infect. Dis.* **2021**, *72*, 703–706. [CrossRef] [PubMed]
8. Glymour, C.; Zhang, K.; Spirtes, P. Review of causal discovery methods based on graphical models. *Front. Genet.* **2019**, *10*, 524. [CrossRef] [PubMed]
9. Pearl, J. *Causality*; Cambridge University Press: Cambridge, UK, 2009.
10. USAFacts. Detailed Methodology and Sources: COVID-19 Data. 2022. Available online: https://usafacts.org/articles/detailed-methodology-covid-19-data (accessed on 1 July 2022).
11. Angelopoulos, A.N.; Pathak, R.; Varma, R.; Jordan, M.I. On identifying and mitigating bias in the estimation of the COVID-19 case fatality rate. *Harv. Data Sci. Rev.* **2020**, *Special Issue 1*. [CrossRef]
12. University of Wisconsin Population Health Institute. County Health Rankings & Roadmaps 2021. Available online: www.countyhealthrankings.org (accessed on 1 July 2022).
13. University of Wisconsin Population Health Institute. County Health Rankings & Roadmaps 2022. Available online: www.countyhealthrankings.org (accessed on 1 July 2022).
14. Altieri, N.; Barter, R.L.; Duncan, J.; Dwivedi, R.; Kumbier, K.; Li, X.; Netzorg, R.; Park, B.; Singh, C.; Tan, Y.S.; et al. Curating a COVID-19 Data Repository and Forecasting County-Level Death Counts in the United States. *Harv. Data Sci. Rev.* **2021**, *Special Issue 1*. [CrossRef]
15. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.
16. Ramsey, J.D.; Zhang, K.; Glymour, M.; Romero, R.S.; Huang, B.; Ebert-Uphoff, I.; Samarasinghe, S.; Barnes, E.A.; Glymour, C. TETRAD—A toolbox for causal discovery. In Proceedings of the 8th International Workshop on Climate Informatics, Boulder, CO, USA, 19–21 September 2018.
17. Ogarrio, J.M.; Spirtes, P.; Ramsey, J. A hybrid causal search algorithm for latent variable models. In Proceedings of the Conference on Probabilistic Graphical Models (PMLR), Lugano, Switzerland, 6–9 September 2016; pp. 368–379.
18. TETRAD. Tetrad Manual. 2020. Available online: https://cmu-phil.github.io/tetrad/manual/ (accessed on 1 July 2022).
19. Zhang, J. Causal reasoning with ancestral graphs. *J. Mach. Learn. Res.* **2008**, *9*, 1437–1474.
20. Connley, C. How COVID-19 Exacerbated America's Racial Health Disparities. 2020. Available online: https://www.cnbc.com/2020/05/14/how-covid-19-exacerbated-americas-racial-health-disparities.html (accessed on 1 July 2022).