County-Level Longitudinal Clustering of COVID-19 Mortality to Incidence Ratio in the United States

Nasim Vahabi[1*], Masoud Salehi[2*], Julio D. Duarte[3], Abolfazl Mollalo[4], George Michailidis[1**]

1. Informatics Institute, University of Florida, Gainesville, FL, USA
2. Department of Biostatistics, College of Public Health, Iran University of Medical Sciences, Tehran, Iran
3. Center for Pharmacogenomics, Department of Pharmacotherapy and Translational Research, College of Pharmacy, University of Florida, Gainesville, FL, USA
4. Department of Public Health and Prevention Sciences, School of Health Sciences, Baldwin Wallace University, Berea, OH, USA

*: Co-first author
**: Correspondence author
gmichail@ufl.edu

ABSTRACT

Background: As of June 3, 2020, the mortality to incidence ratio (MIR) of COVID-19 was 5.8%. We utilized a longitudinal model-based clustering system based on the disease trajectories over time.

Methods: County-level COVID-19 cases and deaths (March-June 2020), and a set of potential risk factors were collected for 3050 U.S. counties. We used growth mixture models to identify clusters of counties exhibiting similar COVID-19 MIR growth trajectories and risk-factors over time.

Results: We identified clusters 1 (rural-areas of IA, NC, OK, VA, FL, GA, LA, OH states), and 7 (rural-areas of AR, CO, GA, KS, NE, TN, TX states) as the so-called "more vulnerable" clusters. Tuberculosis (OR=1.3), drug use disorder (OR=1.1), and particulate matter (OR=1.1) were significantly associated with increased odds of being in a more vulnerable cluster. Heart complications and cancer were the main risk factors increasing the COVID-19 MIR (range: 0.08%-0.72% MIR↑).

Conclusion: We identified two county-clusters exhibiting the highest COVID-19 MIR trajectories, indicating that enhancing the capacity and access to healthcare resources would be key to successfully manage COVID-19 in these "vulnerable" clusters. These findings provide insights for public health policymakers on the groups of people and locations they need to pay particular attention while managing the COVID-19 epidemic.

Keywords: COVID-19, Mortality to Incidence, County-level clustering, Longitudinal study, Latent Growth Model, Comorbidities

1 INTRODUCTION

As of June 3, 2020, the total number of confirmed COVID-19 (caused by the SARS-CoV-2 virus) cases and death in the United States were 1,840,631 and 106 554, respectively. Mortality rate (MR) was 32.5 per 100 000 population, and mortality to incidence ratio was 5.8%, i.e., 5.8% of the COVID-19 confirmed cases experienced death as the outcome (U.S. population on June 2020 was 328.2 million) (https://usafacts.org). Within the United States, according to the Center for Disease Control and Prevention (CDC) report, the maximum number of confirmed cases and death were reported in Queens County in New York state and King County in Washington state, respectively.

COVID-19 was first discovered in Wuhan, China on December 31, 2019, the outbreak of the disease was declared on January 30, 2020, and eventually was declared as a pandemic by World Health Organization (WHO) on March 11, 2020[1]. Shortly after, few countries, most notably Iran and Italy, experienced a significant increase in the number of confirmed cases and deaths[1]. The first COVID-19 case in the United States was confirmed on January 19, 2020, in Washington State. After that, New York City became one of the epicenters of the disease spread. On March 17, 2020, all fifty states across the United States had at least one confirmed case of COVID-19 and on March 26, 2020, the United States became the leading country in the number of COVID-19 cases worldwide, replacing Italy that was previously in the lead of COVID-19 cases (Center for Infectious Disease Research and Policy, 2020, https://www.cidrap.umn.edu/).

Studies have reported multiple risk factors mainly categorized in three groups: (1) comorbidities (including chronic lung disease, heart diseases, diabetes, cancer, and chronic liver disease), (2) demographics & social factors (including age, gender, ethnicity, and smoking status), and (3) environmental factors (including temperature, humidity, and air pollution).

We reviewed the primary published evidence (as of June 2020) reporting associations between the above-mentioned risk factors and COVID-19 incidence, mortality, and severity. We consider more severely impacted patients from COVID-19, those in need of requiring oxygen, hospitalization, or ventilation. Here we include part of the literature review. Full literature review is available in the Supplementary Materials.

Comorbidities:

Chronic lung diseases, CLD: COVID-19 is an acute respiratory disease that primarily affects the pulmonary alveolar epithelial cells, which can lead to respiratory failure and death[2]. There are different hypotheses about whether people with pre-existing CLD (especially chronic obstructive pulmonary disease, COPD) would be at higher risk of infection with the SARS-CoV-2 virus representing more severe symptoms than others.

Halpin et al. showed that the CLD prevalence among COVID-19 cases was less than the estimated prevalence in the general population[3]. In a study from Italy (March 23, 2020), COPD was not reported for any of the patients who died from COVID-19 (n=355, mean-age=79.5)[4]. Similarly, in data in the USA (March 31, 2020), chronic respiratory diseases were comorbidities in 8.5% of patients with COVID-19, compared to the GBD estimate of 11.3% for the same disease[5]. There are also several published studies showing the synergistic effect of CLD in worsening the severity of COVID-19[6-10]. Guan et al. reported more than 50% of chronic pulmonary disease for COVID-19 patients who were also admitted to ICU[11]. In a meta-analysis study on both Chinese- and English-language published articles, Zhao et al. showed that pre-existing COPD was significantly associated with a nearly 4-fold higher risk of developing severe COVID-19. The association remained significant in the subgroup of patients with death or ICU-required patients[7]. Moreover, in large case-series, they reported a higher prevalence of COPD in patients with severe presentation and worse outcomes[8]. In another meta-analysis (May 1, 2020), the reported prevalence of COPD patients was 2% in COVID-19 cases. They showed that although the COPD prevalence was low, it was significantly associated with a higher risk of more severe COVID-19 (63%), and higher mortality (60%)[12]. Brake et al. reported higher (upregulated) expression of the angiotensin-converting enzyme 2 (ACE2) in resected lung tissue from COPD patients compared to those with healthy lung function[9]. Some published evidence also indicates higher ACE2 expression in smokers compared to never smokers, which suggests that smokers can be more susceptible to infection by the SARS-CoV-2 virus[9, 10].

It is necessary to put all these findings into context and consider that people with CLD, especially past or current smokers are more likely to have immune dysregulation. Therefore, these groups of people can be at higher risk of developing more severe symptoms out of a simple upper respiratory infection (similar to Bhat et al. suggestion[13]).

Cardiovascular disease, CVD: In addition to respiratory complications, published studies are showing the impact of pre-exist CVDs on developing COVID-19 and on worsening its severity

and clinical outcomes. Hendren et al. showed that COVID-19 might cause myocarditis-like syndrome and acute myocardial injury associated with reduced left ventricular ejection fraction (LVEF), which can also be complicated by heart failure[14]. A different analysis in China showed that 8%-20% of the patients hospitalized with COVID-19 had abnormal cardiac troponin I (cTnI) who were also older and had more comorbid diseases[15, 16]. There are also published literature (not fully proven though) showing that SARS-CoV-2 can infect fibroblasts and cardiomyocytes via the ACE2-pathway causing myocardial injury[17-21]. Moreover, it is shown that patients with viral myocarditis, which commonly follows with chest pain can mimic a ventricular arrhythmia or coronary syndrome[22, 23]. Historically, researches have shown a significant increase in mortality for SARS patients with pre-existing CVD[24-29].

Demographic & Social Factors:

Age: People of 65 years of age and older are at significantly higher risk of experiencing COVID-19 or for hospitalization and death, especially if they have pre-existing comorbidities such as CVD, DM, CLD, Hypertensive heart disease, and obesity[30, 31]. Ferguson et al. reported that 27%-71% of patients older than 60 years needed especial care in an ICU with an infection fatality rate of about 2%-9.5%[32, 33]. Stang et al. discussed a probable bias in age-significance in COVID-19 patients due to overestimation caused by the limited testing capacity to more symptomatic patients. They showed that the fatality rate from COVID-19 started increasing after age 60 years in Italy, Spain, and the USA (as of April 20, 2020)[34, 35]. There is also a study on children with a median age of 7 years in China (April 1, 2020) in which most of the cases were male (not significant though) with mild symptoms[36].

Although, there is still not enough evidence and/or data to confirm whether this increase in mortality is directly related to age or other comorbidities that are not considered yet in the analysis.

Gender: Most evidence suggests that men are infected at a higher rate than women by COVID-19 and exhibit a higher mortality rate. However, most studies showed no significant differences in infection and mortality between men and women COVID-19 cases[2, 37]. Wenham et al. indicated that although an equal number of male and female COVID-19 cases was observed, there seem to be different MR by gender. They also suggested that women can be in high risk of

getting infected, since they have more front-line interaction with communities, and provide more informal care within families besides their physical and cultural differences[38, 39].

Further, selected studies report significantly different gender-distributions between male and female COVID-19 cases. For example, Zhao Y et al., using single-cell data, reported that ACE2 was upregulated in Asian males compared to women and other ethnicities which may lead to more severe incidents of COVID-19[10, 40, 41].

Environmental Factors:

Air pollution: Exposure to air pollution and particulate matter (PM) can have a positive association with increased risk of certain viral respiratory diseases such as influenza and SARS pandemic 2003. Studies are showing that exposure to PM increased the MR from 2009 H1N1 and Spanish influenza[42-45]. Air pollution is also linked to cellular damage, inflammation, CVD, and CLD, which are potential comorbidities associated with COVID-19 severity[42, 46-48]. Ye et al. showed that air pollution could also play a role in infectious disease transmission, although it has not been studied for COVID-19 as of May 15, 2020[49].

Wu et al. and Mollalo et al., in nationwide studies in the USA, showed that exposure to PM increased COVID-19 mortality and severity[42, 50, 51]. Setti et al. reported a significant relationship between PM and experiencing COVID-19 in Italy (January 1, 2020)[52].

A number of studies did not confirm the association between air pollution and COVID-19 severity, mortality, and transmission. However, they agreed that since exposure to air pollution and PM has a link with other complications, there can be a risk factor in increasing COVID-19 MR and disease severity[53-56].

Despite numerous claims in the literature of the significant role that pre-existing conditions play, the studies to date are not conclusive given the fast changing landscape of data and the current understanding of the disease. Moreover, to the best of our knowledge, longitudinal model-based clustering using the disease mortality pattern over time has not yet been considered in the published studies. Hence, this study using an appropriate modeling framework contributes to the literature by finding relevant clusters considering disease growth trajectories. To this end, we first, determined the county-level risk factors of COVID-19 MIR in the United State using a longitudinal generalized estimating equations (GEE) model. Next, we trained a latent growth mixture model (LGMM) to cluster the U.S. counties and to identify significant risk factors for

each cluster separately. This longitudinal model-based clustering approach enables us to incorporate the possible heterogeneity of COVID-19 MIR growth trajectories present, due to the previously mentioned factors. Note that such heterogeneity is not accounted in other simpler, but widely used models, such as the SIR (susceptible, infected, and recovered) model. Our methodology enables us to cluster different counties into distinctive subpopulations based on their similarities in COVID-19 patterns over time (March 25-June 3, 2020).

The proposed methodology aids in understanding the evolution of COVID-19 disease transmission and severity by examining MIR and developing a model-based clustering system that takes into consideration both the disease pattern over time and the pre-existing risk factors. It can aid in future healthcare planning, by identifying clusters of communities "more vulnerable" to the disease. Finally, the methodology is readily applicable to other countries, if similar granularity data are available.

2 RESULTS

Mean COVID-19 MIR in the contiguous United States significantly increased (P-value<0.001) from MIR=0.8% on March 25 to MIR=3.0% on April 22. Henceforth, the rate slightly increased (P-value=0.501) to MIR=3.2% on April 29 and remained at this level until June 3, 2020 (Table 1). About 93% (n=2830) of the U.S. counties had zero confirmed death (MIR=0%) on March 25, but this percentage decreased to 42.9% (n=1311) by June 3, 2020. On June 3, 2020, the median (Q1, Q3) population of the 3050 U.S. counties was 25884 (11 027, 67 644), with Loving county in Texas having the smallest population (n=169), and Los Angeles county in California the largest one (n=1 039 107). Queens County in New York state had the maximum number of confirmed cases at the beginning of the study on March 25 (n=6420), while Cook County in Illinois had the maximum confirmed cases (n=80 204) at the end of the study on June 3, 2020. Whereas, the maximum number of confirmed death was reported in King County in Washington state on March 25 (n=100) and in Kings County in New York state on June 3 (n=6774).

[Insert Table 1 here]

Based on the univariate variable selection method, some potential risk factors were excluded from the final analysis including, asthma (P-value=0.980), COPD (P-value=0.703), leukemia (P-

value=0.402), and rheumatic disease (P-value=0.774). The description table of the potential risk factors is provided in Table S1.

Results of the final multivariate GEE model (Table 2) showed significant **positive association** between COVID-19 MIR and CVD (0.21%±0.01%, P-value<0.001), cardiomyopathy and myocarditis (0.15%±0.04%, P-value=0.015), hypertensive heart disease (0.10%±0.03%, P-value=0.001), ischemia (0.08%±0.03%, P-value=0.024), mesothelioma (0.72%±0.27%, P-value=0.009), pancreatic cancer (0.51%±0.10, P-value<0.001), drug use disorder (0.08%±0.02, P-value<0.001), and age (5.36%±8.01%, P-value=0.032).

Whereas, there were **negative associations** between COVID-19 MIR and tracheal cancer (-0.03%±0.01%, P-value=0.002), and alcohol use disorder (-0.17%±0.05%, P-value=0.002). Moreover, the effect of time on the COVID-19 MIR was significant and positive (0.21%±0.01%, P-value<0.001) suggesting the use of longitudinal (repeated measures) approaches instead of cross-sectional studies to better evaluate the growth trajectory of COVID-19 MIR over time.

[Insert Table 2 here]

Table S2 shows the full result of the LGMs. Based on the information criteria, a non-linear LGM with a quadratic term exhibited a better fit than the linear LGM. Figure 1 shows the overall COVID-19 MIR growth trajectories based on both linear and non-linear models. The overall growth trajectory of the observed mean COVID-19 MIR for 1736 U.S. counties (with MIR>0) showed a sharp increase from MIR=1.9% on March 25 to MIR=5.6% on April 29. Henceforth, the rate slightly increased to MIR=5.9% on May 20 and then slightly decreased to MIR=5.7% till June 3, 2020 (Figure 1, Non-linear LGM).

[Insert Figure 1 here]

A clustered pattern of COVID-19 MIR across the U.S. is confirmed by *Moran's I* statistics (MIR-*Morans'I*=0.46, P-value<0.001). based on the LGMM results, an 8-cluster non-linear model was selected as the best model to find clusters of the U.S. counties. Detailed result of LGMM models is provided in Table S3. Table 3 and Figure 2 show the detailed MIR information over time (factor loadings are reported in Table S4).

[Insert Table 3 here]

Cluster 0 contains 1314 counties with zero confirmed death form COVID-19 during the study time (i.e., MIR=0).

Cluster 1 with 52 counties from 28 different states, had the highest MIR at the beginning of the study (intercept=12.9%±3.1%) compared to other clusters (Table 3). This cluster continued having the highest MIR at the end of the study, on June 3, 2020 (MIR=13.2%). IA (Audubon, Floyd, and Guthrie counties), IL (Carroll, Clinton, and Jasper counties), NC (McDowell, Moore, Orange, and Polk counties), OK (Cotton, Le Flore, Mayes counties), and VA (Northumberland, Page, and Scott counties) were the most frequent states present in this cluster. Within this cluster, McHenry (ND), Crowley (CO), Terrell (GA), and Shelby (KY) counties had the highest COVID-19 MIR. COVID-19 MIR growth trajectory for the counties in this cluster (Figure 3) showed a 5% decrease from March 25 (MIR=12.9%) to April 1 (MIR=7.9%) and stayed steady (flat) till April 8, 2020. From here, the rate slightly increased to MIR=9% and stayed at this level till May 6, and thereafter, had another increase to MIR=13.2% on June 3, 2020.

Cluster 6 with 64 counties (from 28 different states) had the second-highest MIR at the beginning of the study (intercept=9.8%±3.0%) compare to other clusters. However, on June 3, it had the third-lowest MIR compare to other clusters. GA (with 7 counties), KY (with 4 counties), MI (with 5 counties), OH (with 6 counties), and VA (with 6 counties) are the most frequent states in this cluster. COVID-19 MIR growth trajectory had a sharp increase from MIR=9.8% on March 25 to MIR=36.0% on April 1, 2020. Then. the rate had a sharp decrease to MIR=9.5% till April 22 and continued decreasing with a gentle slope till June 3, 2020 (MIR=7.7%). Iron (WI), Gallia (OH), Bourbon (KY), and Missaukee (MI) had the highest COVID-19 MIR trajectories within this cluster.

More information about the COVID-19 MIR estimation at both the beginning and the end of the study, amount of increase (or decrease) in this rate, and the rank of each cluster are presented in Table S5. One important point in Table S5 is that counties in **cluster 4** (MIR: 0.8%→10.5%) and **cluster 7** (MIR: 1.5%→11.6%) had the highest increase in COVID-19 MIR from March 25 to June 3, 2020.

Table 4 shows the significant risk factors in each cluster. To find the odds ratios (ORs) we used cluster 0 as the baseline one (with MIR=0) and compared all other clusters to it. The full results of the multinomial logit model for all risk factors are provided in Table S6 in the Supplement.

[Insert Table 4 here]

TB (OR=1.3) drug use disorder (OR=1.1), and PM (OR=1.1) are significantly associated exhibiting a 30%, 10% and 10% increase in the relative log-odds of being in “vulnerable cluster

7" vs. cluster 0, respectively. Table S6 contains the detailed output of the multinomial logit model for all potential risk factors in each cluster separately.

[Insert Figure 2 here]

Figure 3 shows the geographical distribution of the 8 clusters of the contiguous United States, based on the estimated COVID-19 MIR growth trajectory over time, from March 25 to June 3, 2020.

[Insert Figure 3 here]

3 DISCUSSION

This study investigated the county-level COVID-19 confirmed cases and death from March 25 to June 3, 2020 in a longitudinal fashion in the contiguous United States. We assessed the growth trajectories of COVID-19 MIR and found the county-level clusters of the contiguous United States with similarities in COVID-19 MIR growth trajectory over time. We also considered the effects of different county-level potential risk factors on MIR, including comorbidities & disorders, demographics & social factors, and environmental factors. We selected MIR as a measure of mortality since it also considers the number of confirmed cases to adjust the mortality rates. However, the estimates of all COVID-19 epidemiological-measures (i.e., incidence, prevalence, and mortality rates) are subject to bias due to the imprecise number of affected (confirmed) cases, especially those with mild or no disease symptoms. Moreover, there are yet not enough studies showing the association between different risk factors, especially pre-existing comorbidities, with COVID-19 incidence and mortality.

We found nine clusters of the U.S. counties (including a cluster of counties with zero MIR) based on the COVID-19 MIR pattern (growth trajectory) using a longitudinal LGMM. All counties in the same cluster have a similar COVID-19 MIR growth pattern over the study time. This approach considered both spatial and temporal heterogeneities in COVID-19 MIR due to pre-existing comorbidities, environmental factors, and demographics. We also identified significant risk factors associated with the identified clusters using a multinomial logit model. It is shown that different age and sex distributions in the U.S. counties impact differentially COVID-19 mortality and severity[57, 58]. Race is also a factor that leads to heterogeneity. For instance, number of findings reported African Americans having higher risk of getting infected, experiencing more severe COVID-19 and death[59]. In our study, about 43% of the northern and

central U.S. counties did not experience death from COVID-19 until June 3. Nearly 116 counties in clusters 1 and 6 had the highest mean COVID-19 MIR at the beginning of the study on March 25, 2020. On June 3, 2020, counties in cluster 1 still had the highest mean COVID-19 MIR (MIR=13.2%), while counties in cluster 6 improved to the third lowest (excluding the cluster with MIR=0). Counties in cluster 7 had a low level of COVID-19 MIR in the beginning of the study on March 25 (MIR=1.5%). However, they had a very dramatic increase (10.1%↑) in COVID-19 MIR till June 3, 2020 (MIR=11.6%). Cluster seven became the cluster with second highest COVID-19 MIR at the end of the study period on June 3, 2020. Based on the clustering result (as of June 2020), we considered clusters 1, and 7 as the so called "more vulnerable" clusters of counties which needs more attention to control disease mortality.

Cluster 7 includes of the following counties: Marion (KS), Seward (NE), Churchil (NV), Catron (MN), Crater (OK), Benton (TN), Gonzales (TX), Lavaca (TX), and Barbour (WV). Most frequent states in cluster 1 were IA (Audubon, Floyd, and Guthrie counties), IL (Carroll, Clinton, and Jasper counties), NC (McDowell, Moore, Orange, and Polk counties), OK (Cotton, Le Flore, Mayes counties), and VA (Northumberland, Page, and Scott counties). TB (OR=1.3) and drug use disorder (OR=1.1) are two significant factors that are respectively associated with a 30% and a 10% increase in the odds of being in cluster 7 vs. cluster 0. Among the demographic and environmental factors, male-WA% (OR=1.8) and PM (OR=1.1) are significantly associated with 80% and 10% increase in the relative log-odds of being in cluster 7 vs. cluster 0. Therefore, protecting subjects with TB and drug use disorder, and managing the $PM_{2.5}$ level of the air (mixture of solid particles and liquid droplets found in the air, such as, dust, dirt, or smoke) can help ameliorate the COVID-19 mortality in these counties.

Moreover, more than 80% of the counties in clusters 1 and 7 were rural areas based on the U.S. Census Bureau definition (https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural.html). Lack of access to health and critical care infrastructure, and more limited resources in general may be responsible for higher COVID-19 MIR. Therefore, addressing these factors would be beneficial in the long run for managing the epidemic.

We found a positive significant association between COVID-19 MIR and heart diseases including CVD (0.21% MIR↑), cardiomyopathy and myocarditis (0.15% MIR↑), and hypertensive heart disease (0.10% MIR↑). This finding is in accordance with recent studies on the topic, even though its etiology remains uncertain. This can be due to lipid and glucose

metabolism disorders caused by SARS-CoV2, antiviral drug-related heart damage, or poor diagnosis in patients with heart complication. Historically, it is shown that patients with pre-existing heart and lung diseases had a higher mortality rate from SARS[17, 24-29]. The same findings have been reported in China[15, 16, 60], and United Kingdom[61]. Lippi et al. showed that about 20% of hospitalized COVID-19 cases had heart complications as well[16]. A meta-analysis with 46248 confirmed COVID-19 cases showed that patients with severe disease were more likely to have CVD (odds ratio=3.4), and hypertensive heart disease (odds ratio= 2.4)[62]. Recent studies have reported ACE2 as the coreceptor for the coronavirus in people with heart and lung complications compared with healthy individuals[29, 63]. There is also evidence showing the critical role of the ACE2 and its peptides in the inflammatory[64, 65] and oxidative organ activities[66, 67] which are significant triggers in the initiation and progression of cardiovascular disease, cardiac hypertrophy, lung complications, and acute pancreatitis. However, ACE2 is also a component of the renin-angiotensin-aldosterone system (RAAS) may act as a competitive interceptor of SARS-CoV-2 and slow the virus and therefore protect from heart and lung complications[68].

We did not find a significant association between respiratory diseases (including COPD, Asthma, lower respiratory infection, and interstitial lung disease and pulmonary sarcoidosis) and COVID-19 MIR which is consistent with the Halpin et al. study[3], Onder et al. in Italy (March 2020)[4], and the *CDC* report of health conditions' prevalence in the USA (April 2020)[5]. One possible explanation might be that having CLD causes a different immune response, which eventually protects against infection form SARS-CoV2. Another possibility is that treatments and therapies used by patients with CLD can protect against COVID-19 as well (for instance, ECMO[69], topical intra-nasal sprays[70], and mPGES-1[71]), or that CLD treatments can reduce symptoms and hence affecting COVID-19 diagnosis[3]. However, the *Chinese CDC* (http://www.chinacdc.cn/en/) has reported a 6.3% COVID-19 case-fatality rate for cases with pre-existing chronic respiratory diseases.

Besides heart diseases, we found significant associations between COVID-19 MIR and cancer including mesothelioma (0.72% MIR↑) and pancreatic (0.51% MIR↑) in the United States. Typically, patients with cancer are known to be at higher risk for community respiratory viruses (such as influenza and coronaviruses) due to their suppressed immune system and poor physiological baseline[72-74]. Based on a descriptive study from Wuhan, China (March 2020), the

incidence of COVID-19 patients with pre-existing cancer was about 1%, which is five times higher than the general cancer incidence in China[56]. In a report of 72314 cases from the *Chinese CDC* (March 2020), the COVID-19 case fatality for patients with cancer was 3.5% higher than those without cancer[75]. In another report from Italy (April 2020), the prevalence of pre-existing cancer among COVID-19 death was 16.5%[4]. Du et al., in a multi-omics study, indicated an indirect connection between ACE2 pathway and cancer via Transforming Growth Factor Beta 1, *TGFB1*, association with colorectal cancer[76, 77].

Our findings also indicated that demographics and social factors at the county-level, such as mean age, and drug use disorders significantly increased COVID-19 MIR by 5.36% and 0.08%, respectively. Drug use disorders can result in increased inflammation of multiple organ systems, particularly lungs, which may lead to respiratory failure. In turn, it can directly contribute to the elevated mortality rate of COVID-19 among confirmed cases. Marsden et al. showed that people with opioid use disorder have a higher prevalence of co-occurrence of health problems, subsequently leading to an increased rate of COVID-19[78].

This study has several limitations. **First**, the mortality and MIR estimates from the current COVID-19 related data are biased since most of the individuals with mild or no symptoms have not been tested for COVID-19 in most of the counties. Moreover, the COVID-19 reporting system appears to differ regionally, which introduces further inaccuracies in the available data. For example, for a small number of counties we found MIR=100%, which is an unlikely event and can be due to an incomplete disease recording system. Timely sharing of information and collaboration between organizations and governors can partly solve this problem. To have higher quality data, there also needs to be additional testing and follow-ups, especially for younger individuals with mild symptoms. Recent data (CDC June 19, 2020[79]) showed that more young people are testing positive for COVID-19 in the United States. **Second**, the reporting of disease data are mostly based on ICD9/10 codes which can be fairly inaccurate[80]. **Third**, the analysis was based on county-level data. It would be beneficial to analyze individual-level and multi-countries data to gain deeper insights into the impact of risk factors into COVID-19 progression. **Forth**, some of the counties, especially in Maine, were excluded from the study because some of the environmental factors such as climate and air pollution were not directly available.

In summary, accounting for heterogeneity in both risk factors and COVID-19 mortality patterns over time leads to a more informative clustering system, which can then be leveraged in managing the epidemic by identifying and informing groups of people at higher risk and also in managing healthcare resources (access to facilities, ICUs, etc.) more judiciously. Findings of this study suggest that counties in clusters 1 and 7 experience higher COVID-19 MIR growth trajectories over time and are facing more challenges due to the prevalence of rural counties (more than 80%) in managing the disease. Further, heart complications and cancer were statistically significant pre-existing comorbidities related to COVID-19 MIR across the U.S. Drug use disorder, TB, and PM were specifically associated with an increased chance of being in a more "vulnerable" cluster.

4 METHODS

Data resources

We collected county-level cumulative COVID-19 confirmed cases and death from March 25 to June 3, 2020, across the contiguous United States from *USAFacts* (usafacts.org). MIR, as a proxy for survival rate, is calculated by dividing the number of confirmed deaths in each county by the confirmed cases in the same county at the same time-period multiplied by 100. MIR ranges from 0%-100%, 100% indicating the worst situation where all of the confirmed cases have died.

Thirty-four potential risk factors (covariates), including county-level MR of comorbidities & disorders, demographics & social factors, and environmental factors were retrieved from the *University of Washington Global Health Data Exchange* (http://ghdx.healthdata.org/us-data). Comorbidities and disorders includes CVD, cardiomyopathy and myocarditis and myocarditis, hypertensive heart disease, peripheral vascular disease, atrial fibrillation, cerebrovascular disease, hepatitis, HIV/AIDS, tuberculosis (TB), lower respiratory infection, interstitial lung disease and pulmonary sarcoidosis, asthma, COPD, ischemia, mesothelioma, tracheal cancer, leukemia, pancreatic cancer, rheumatic disease, drug use disorder, and alcohol use disorder. Demographics & social factors includes age, female African American%, female white American%, male African American%, male white American%, Asian%, smokers%, unemployed%, income rate, and uninsured%. Environmental factors include air quality index

(AQI), temperature, and PM. A descriptive table including all potential risk factors is provided in Table S1).

Analysis (descriptive methods and models)

We **first** provide summary statistics for COVID-19 data for the period under consideration. Full descriptive statistics for n=34 potential risk factors are provided in Table S1 in the Supplementary Materials.

Second, we applied GEE marginal approaches to model the COVID-19 MIR over time, and to find the significant risk factors. To this end, we first used the forward-selection method to select the most relevant risk factors (covariates) among the covariates using univariate GEE models[81], as follows:

$$\begin{cases} \mu_{ij}^{(1)} = \beta_0 + \beta_1 Time + \beta_{2(1)} X_{(1)} \\ \mu_{ij}^{(2)} = \beta_0 + \beta_1 Time + \beta_{2(2)}\, X_{(2)} \\ \qquad\qquad \vdots \\ \mu_{ij}^{(34)} = \beta_0 + \beta_1 Time + \beta_{2(34)}\, X_{(34)} \end{cases} , i = 1, \dots 3050\ (counties);\quad j = 1, \dots, 11\ (weeks). \qquad (1)$$

where μ_{ij} indicates the mean COVID-19 MIR for i^{th} county in week j^{th}, β_0 is the starting rate of MIR before considering the effect of any potential risk factor (intercept), β_1 and β_2s are the effects of time and risk factors X (such as Asthma) on the COVID-19 MIR. For variable selection purposes, we chose variables with (univariate) P-value<0.2 to be included in the final multivariate GEE model, as follows:

$$\mu_{ij} = \alpha_0 + \sum_{p=1}^{n_1} \alpha_p X_p, \qquad (2)$$

where μ_{ij} indicates the overall marginal mean MIR for the i^{th} county in the j^{th} week. α_0 is the intercept and α_p is the coefficient of the p^{th} potential risk factor (X_p), $p = 1,2, \dots, n_1$, where n_1 is the total number of the selected variables based on the univariate GEE model (Eq 1). Variables with (multivariate) P-value<0.05 will be selected as the potential risk factors. In each marginal model, an appropriate correlation structure (with the best goodness of fit index, QIC) was utilized. Statistical analysis and visualization for this step were performed using the *geepack* R-package (https://cran.r-project.org/web/packages/geepack/).

Third, we evaluated COVID-19 MIR growth trajectory over the study time using a latent growth model (LGM). An LGM approach considers both the mean MIR differences between counties at each time point (inter-subject) and MIR growth trajectories over time (intra-subject). Specifically, suppose y_{ti} is the COVID-19 MIR in the i^{th} county at time t; then, it can be modeled as follows[82]:

$$\begin{aligned} y_{ti} &= \eta_{0i} + \eta_{1i}\lambda_t + \varepsilon_{ti}, \\ \eta_{0i} &= \eta_0 + \varepsilon_{0i}, \\ \eta_{1i} &= \eta_1 + \varepsilon_{1i}, \end{aligned} \tag{3}$$

where η_{0i} and η_{1i} are two latent growth factors and λ_ts are time scores (factor loadings); ε_{ti} is a normally distributed error term for the i^{th} county at time t; η_0 and η_1 indicate the estimated overall mean COVID-19 MIR in each county and the average rate of MIR change, respectively. We also employed a number of non-linear (quadratic) LGMs, based on a polynomial time function (quadratic or higher-order) of time scores[83] to decrease estimation bias to account for the MIR trajectories exhibiting non-linear behavior over time. The non-linear LGM using a quadratic time function is given by:

$$\begin{aligned} y_{ti} &= \eta_{0i} + \eta_{1i}\lambda_t + \eta_{2i}\lambda_t^2 + \varepsilon_{ti}, \\ \eta_{0i} &= \eta_0 + \varepsilon_{0i}, \\ \eta_{1i} &= \eta_1 + \varepsilon_{1i}, \\ \eta_{2i} &= \eta_2 + \varepsilon_{2i}, \end{aligned} \tag{4}$$

where η_2 indicates the growth factor, which can be a concave or convex form of the COVID-19 MIR pattern over the study time, and λ_t^2 are the squared time scores. Both linear and non-linear LGMs were applied to 1736 U.S. counties with MIR>0, i.e., counties with at least one confirmed death between March 25 to June 3, 2020. We then used information criteria (AIC, BIC) to find the best model among linear and non-linear LGMs to determine the COVID-19 MIR changes and patterns over the study time. Smaller AIC and BIC values indicate better fit of the underlying models. We also calculated *Moran's I*[84] to evaluate the spatial autocorrelation of COVID-19 MIR across the U.S. counties.

Forth, we identified clusters of the U.S. counties based on the COVID-19 MIR growth trajectory over time using longitudinal LGMMs[82], as follows:

$$y_{it}^k = \eta_{i0}^k + \eta_{i1}^k \lambda_t^k + \varepsilon_{it}^k, \tag{5}$$

$$\eta_{i0}^{k} = \eta_{00}^{k} + \varepsilon_{i0}^{k},$$
$$\eta_{i1}^{k} = \eta_{10}^{k} + \varepsilon_{i1}^{k},$$

where k is the upper bound of the number of the clusters, η_{00}^{k} indicates the initial COVID-19 MIR at the beginning of the study, and η_{10}^{k} indicates the average rate of COVID-19 MIR change over time. To find the optimal number of clusters (k), we fit a series of LGMMs with different numbers of clusters of counties and conducted tests for the adequacy of the reduced models with respect to the number of clusters. Information criteria such as AIC, BIC, and a bootstrap likelihood ratio test (BLRT) were used to compare the k-cluster model to the $(k-1)$-clsuter model [85, 86]. Also, cluster sample sizes greater than 1% of the total sample size, and a relative entropy (REN) statistic greater than 0.8 were considered as the qualified latent class membership classification criteria[87]. The REN statistic for a k-class model is calculated as $REN(k) = 1 - \frac{-\sum_{i=1}^{N}\sum_{k=1}^{K} P_{ik} ln P_{ik}}{N - lnK}$, where k and i correspond to the number of clusters and counties, respectively, and P_{ik} indicates the posterior probability for the i^{th} county to be in cluster k. We then applied a multinomial logit model to find the significant risk factors in each cluster as follows:

$$ln\frac{p(y_i = k)}{p(y_i = 0)} = \alpha_k + \sum_{p=1}^{n_1} \beta_p X_p\,, \qquad k = 1, \dots, K\ (cluster) \tag{6}$$

where y_i is a categorical variable with K possible categories (indicating the cluster number), α_k is the intercept for cluster k, β_k is a vector of regression coefficients of the p^{th} potential risk factor (X_p), $p = 1,2, \dots, n_1$, where n_1 is the total number of the selected variables based on the univariate GEE model (Eq 1).

Statistical analysis for LGMMs and multinomial logit model were performed using *Mplus* v6.12 (Muthén & Muthén, CA, USA, www.statmodel.com) and the *nnet* R-package (https://cran.r-project.org/web/packages/nnet/index.html), respectively. Geographical distribution of the clusters was illustrated in a color-coded geographical map using *ArcGIS 10.7* (ESRI, Redland, CA).

5 REFERENCES

[1] Organization WH. Rolling updates on coronavirus disease (COVID-19): WHO characterizes COVID-19 as a pandemic. 2020.

[2] Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The lancet. 2020;395(10223):497-506.

[3] Halpin DM, Faner R, Sibila O, Badia JR, Agusti A. Do chronic respiratory diseases or their treatment affect the risk of SARS-CoV-2 infection? The Lancet Respiratory Medicine. 2020;8(5):436-8.

[4] Onder G, Rezza G, Brusaferro S. Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. Jama. 2020;323(18):1775-6.

[5] Covid C, COVID C, COVID C, Chow N, Fleming-Dutra K, Gierke R, et al. Preliminary estimates of the prevalence of selected underlying health conditions among patients with coronavirus disease 2019—United States, February 12–March 28, 2020. Morbidity and Mortality Weekly Report. 2020;69(13):382.

[6] Liu X, Zhou H, Zhou Y, Wu X, Zhao Y, Lu Y, et al. Temporal radiographic changes in COVID-19 patients: relationship to disease severity and viral clearance. 2020.

[7] Zhao Q, Meng M, Kumar R, Wu Y, Huang J, Lian N, et al. The impact of COPD and smoking history on the severity of Covid-19: A systemic review and meta-analysis. Journal of medical virology. 2020.

[8] Guan W-j, Ni Z-y, Hu Y, Liang W-h, Ou C-q, He J-x, et al. Clinical characteristics of coronavirus disease 2019 in China. New England journal of medicine. 2020;382(18):1708-20.

[9] Brake SJ, Barnsley K, Lu W, McAlinden KD, Eapen MS, Sohal SS. Smoking upregulates angiotensin-converting enzyme-2 receptor: a potential adhesion site for novel coronavirus SARS-CoV-2 (Covid-19). Multidisciplinary Digital Publishing Institute; 2020.

[10] Zhao Y, Zhao Z, Wang Y, Zhou Y, Ma Y, Zuo W. Single-cell RNA expression profiling of ACE2, the putative receptor of Wuhan 2019-nCov. BioRxiv. 2020.

[11] WHO. Coronavirus disease 2019 (COVID-19): situation report, 73. 2020.

[12] Alqahtani JS, Oyelade T, Aldhahir AM, Alghamdi SM, Almehmadi M, Alqahtani AS, et al. Prevalence, severity and mortality associated with COPD and smoking in patients with COVID-19: a rapid systematic review and meta-analysis. PloS one. 2020;15(5):e0233147.

[13] Bhat TA, Panzica L, Kalathil SG, Thanavala Y. Immune dysfunction in patients with chronic obstructive pulmonary disease. Annals of the American Thoracic Society. 2015;12(Supplement 2):S169-S75.

[14] Hendren NS, Drazner MH, Bozkurt B, Cooper J, Leslie T. Description and proposed management of the acute COVID-19 cardiovascular syndrome. Circulation. 2020.

[15] Shi S, Qin M, Shen B, Cai Y, Liu T, Yang F, et al. Association of cardiac injury with mortality in hospitalized patients with COVID-19 in Wuhan, China. JAMA cardiology. 2020.

[16] Lippi G, Lavie CJ, Sanchis-Gomar F. Cardiac troponin I in patients with coronavirus disease 2019 (COVID-19): Evidence from a meta-analysis. Progress in cardiovascular diseases. 2020.

[17] Hamming I, Timens W, Bulthuis M, Lely A, Navis G, van Goor H. Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis. The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland. 2004;203(2):631-7.

[18] Crackower MA, Sarao R, Oudit GY, Yagil C, Kozieradzki I, Scanga SE, et al. Angiotensin-converting enzyme 2 is an essential regulator of heart function. Nature. 2002;417(6891):822-8.
[19] Chen L, Li X, Chen M, Feng Y, Xiong C. The ACE2 expression in human heart indicates new potential mechanism of heart injury among patients infected with SARS-CoV-2. Cardiovascular research. 2020;116(6):1097-100.
[20] Tang N, Li D, Wang X, Sun Z. Abnormal coagulation parameters are associated with poor prognosis in patients with novel coronavirus pneumonia. Journal of Thrombosis and Haemostasis. 2020;18(4):844-7.
[21] Danzi G, Loffi M, Galeazzi G, Gherbesi E. Acute pulmonary embolism and COVID-19 pneumonia: a random association?[published online ahead of print March 30, 2020]. Eur Heart J.
[22] Ammirati E, Cipriani M, Moro C, Raineri C, Pini D, Sormani P, et al. Clinical Presentation and Outcome in a Contemporary Cohort of Patients With Acute Myocarditis: Multicenter Lombardy Registry. Circulation. 2018;138(11):1088-99.
[23] Hufnagel G, Pankuweit S, Richter A, Schönian U, Maisch B, Investigators E. The European Study of Epidemiology and Treatment of Cardiac Inflammatory Diseases (ESETCID) First Epidemiological Results. Herz. 2000;25(3):279-85.
[24] Li W, Moore MJ, Vasilieva N, Sui J, Wong SK, Berne MA, et al. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. Nature. 2003;426(6965):450-4.
[25] Zhou G, Zhao J, Wang S. Pathological study of impact of SARS coronavirus on heart and its conduction system in SARS patients. Medical Journal of Chinese People's Liberation Army. 1982(01).
[26] Oudit G, Kassiri Z, Jiang C, Liu P, Poutanen S, Penninger J, et al. SARS-coronavirus modulation of myocardial ACE2 expression and inflammation in patients with SARS. European journal of clinical investigation. 2009;39(7):618-25.
[27] Booth CM, Matukas LM, Tomlinson GA, Rachlis AR, Rose DB, Dwosh HA, et al. Clinical features and short-term outcomes of 144 patients with SARS in the greater Toronto area. Jama. 2003;289(21):2801-9.
[28] Li B, Yang J, Zhao F, Zhi L, Wang X, Liu L, et al. Prevalence and impact of cardiovascular metabolic diseases on COVID-19 in China. Clinical Research in Cardiology. 2020:1-8.
[29] South AM, Diz DI, Chappell MC. COVID-19, ACE2, and the cardiovascular consequences. American Journal of Physiology-Heart and Circulatory Physiology. 2020;318(5):H1084-H90.
[30] Muniyappa R, Gubbi S. COVID-19 pandemic, coronaviruses, and diabetes mellitus. American Journal of Physiology-Endocrinology and Metabolism. 2020;318(5):E736-E41.
[31] Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, et al. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. Jama. 2020.
[32] Ferguson N, Laydon D, Nedjati Gilani G, Imai N, Ainslie K, Baguelin M, et al. Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. 2020.
[33] Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. Jama. 2020;323(11):1061-9.
[34] Stang A, Standl F, Jöckel K-H. Characteristics of COVID-19 pandemic and public health consequences. Herz. 2020:1.

[35] Standl F, Jöckel K-H, Stang A. COVID-19 and the need of targeted inverse quarantine. European Journal of Epidemiology. 2020:1-2.
[36] Dong Y, Mo X, Hu Y, Qi X, Jiang F, Jiang Z, et al. Epidemiology of COVID-19 among children in China. Pediatrics. 2020;145(6).
[37] Emami A, Javanmardi F, Pirbonyeh N, Akbari A. Prevalence of underlying diseases in hospitalized patients with COVID-19: a systematic review and meta-analysis. Archives of academic emergency medicine. 2020;8(1).
[38] Wenham C, Smith J, Morgan R. COVID-19: the gendered impacts of the outbreak. The Lancet. 2020;395(10227):846-8.
[39] Feng Z. The Novel Coronavirus Pneumonia Emergency Response Epidemiology Team. China CDC Weekly. 2020;2(8):113-22.
[40] Cai H. Sex difference and smoking predisposition in patients with COVID-19. The Lancet Respiratory Medicine. 2020;8(4):e20.
[41] Atto B, Eapen MS, Sharma P, Frey U, Ammit AJ, Markos J, et al. New therapeutic targets for the prevention of infectious acute exacerbations of COPD: role of epithelial adhesion molecules and inflammatory pathways. Clinical Science. 2019;133(14):1663-703.
[42] Wu X, Nethery RC, Sabath BM, Braun D, Dominici F. Exposure to air pollution and COVID-19 mortality in the United States. medRxiv. 2020.
[43] Morales KF, Paget J, Spreeuwenberg P. Possible explanations for why some countries were harder hit by the pandemic influenza virus in 2009–a global mortality impact modeling study. BMC infectious diseases. 2017;17(1):642.
[44] Clay K, Lewis J, Severnini E. What explains cross-city variation in mortality during the 1918 influenza pandemic? Evidence from 438 US cities. Economics & Human Biology. 2019;35:42-50.
[45] Clay K, Lewis J, Severnini E. Pollution, infectious disease, and mortality: evidence from the 1918 Spanish influenza pandemic. The Journal of Economic History. 2018;78(4):1179-209.
[46] Kaan PM, Hegele RG. Interaction between respiratory syncytial virus and particulate matter in guinea pig alveolar macrophages. American journal of respiratory cell and molecular biology. 2003;28(6):697-704.
[47] Lambert AL, Trasti FS, Mangum JB, Everitt JI. Effect of preexposure to ultrafine carbon black on respiratory syncytial virus infection in mice. Toxicological sciences. 2003;72(2):331-8.
[48] Ciencewicki J, Jaspers I. Air pollution and respiratory viral infection. Inhalation toxicology. 2007;19(14):1135-46.
[49] Ye Q, Fu J-f, Mao J-h, Shang S-q. Haze is a risk factor contributing to the rapid spread of respiratory syncytial virus in children. Environmental Science and Pollution Research. 2016;23(20):20178-85.
[50] Mollalo A, Vahedi B, Rivera KM. GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. Science of The Total Environment. 2020:138884.
[51] Mollalo A, Rivera KM, Vahedi B. Artificial Neural Network Modeling of Novel Coronavirus (COVID-19) Incidence Rates across the Continental United States. International Journal of Environmental Research and Public Health. 2020;17(12):4204.
[52] Setti L, Passarini F, De Gennaro G, Barbieri P, Perrone MG, Piazzalunga A, et al. The potential role of particulate matter in the spreading of COVID-19 in Northern Italy: first evidence-based research hypotheses. medRxiv. 2020.
[53] Bontempi E. First data analysis about possible COVID-19 virus airborne diffusion due to air particulate matter (PM): the case of Lombardy (Italy). Environmental Research. 2020:109639.

[54] Fattorini D, Regoli F. Role of the chronic air pollution levels in the Covid-19 outbreak risk in Italy. Environmental Pollution. 2020:114732.
[55] Conticini E, Frediani B, Caro D. Can atmospheric pollution be considered a co-factor in extremely high level of SARS-CoV-2 lethality in Northern Italy? Environmental pollution. 2020:114465.
[56] Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. The Lancet. 2020;395(10223):507-13.
[57] Miller I, Becker A, Grenfell B, Metcalf C. Mapping the burden of COVID-19 in the United States. medRxiv. 2020. OnLine (0).1-21.
[58] Scully EP, Haverfield J, Ursin RL, Tannenbaum C, Klein SL. Considering how biological sex impacts immune responses and COVID-19 outcomes. Nature Reviews Immunology. 2020:1-6.
[59] Yancy CW. COVID-19 and African Americans. Jama. 2020.
[60] Nie Y, Li J, Huang X, Guo W, Zhang X, Ma Y, et al. Epidemiological and clinical characteristics of 671 COVID-19 patients in Henan Province, China. International Journal of Epidemiology. 2020.
[61] Lai AG, Pasea L, Banerjee A, Denaxas S, Katsoulis M, Chang WH, et al. Estimating excess mortality in people with cancer and multimorbidity in the COVID-19 emergency. medRxiv. 2020.
[62] Jordan RE, Adab P, Cheng K. Covid-19: risk factors for severe disease and death. British Medical Journal Publishing Group; 2020.
[63] Bavishi C, Maddox TM, Messerli FH. Coronavirus disease 2019 (COVID-19) infection and renin angiotensin system blockers. JAMA cardiology. 2020.
[64] Reddy Gaddam R, Chambers S, Bhatia M. ACE and ACE2 in inflammation: a tale of two enzymes. Inflammation & Allergy-Drug Targets (Formerly Current Drug Targets-Inflammation & Allergy). 2014;13(4):224-34.
[65] Yamamoto Y, Gaynor RB. Therapeutic potential of inhibition of the NF-κB pathway in the treatment of inflammation and cancer. The Journal of clinical investigation. 2001;107(2):135-42.
[66] Montezano AC, Cat AND, Rios FJ, Touyz RM. Angiotensin II and vascular injury. Current hypertension reports. 2014;16(6):431.
[67] Arwood MJ, Vahabi N, Lteif C, Sharma RK, Machado RF, Duarte JD. Transcriptome-wide analysis associates ID2 expression with combined pre-and post-capillary pulmonary hypertension. Scientific Reports. 2019;9(1):1-10.
[68] Batlle D, Wysocki J, Satchell K. Soluble angiotensin-converting enzyme 2: a potential approach for coronavirus infection therapy? Clinical science. 2020;134(5):543-5.
[69] Matthay MA, Aldrich JM, Gotts JE. Treatment for severe acute respiratory distress syndrome from COVID-19. The Lancet Respiratory Medicine. 2020;8(5):433-4.
[70] Basu S, Holbrook L, Kudlaty K, Mamdani M, Farzal Z, Fasanmade O, et al. Numerical evaluation of spray position for improved nasal drug delivery. Scientific Reports, 2019.
[71] Zhou S, Zhou Z, Ding K, Yuan Y, Loftin C, Zheng F, et al. DREAM-in-CDM Approach and Identification of a New Generation of Anti-inflammatory Drugs Targeting mPGES-1. Scientific Reports. 2020;10(1):1-9.
[72] Ganatra S, Hammond SP, Nohria A. The novel coronavirus disease (COVID-19) threat for patients with cardiovascular disease and cancer. ACC Journals; 2020.

[73] Rivera A, Ohri N, Thomas E, Miller R, Knoll MA. The Impact of COVID-19 on Radiation Oncology Clinics and Cancer Patients in the US. Advances in Radiation Oncology. 2020.
[74] Thom KA, Kleinberg M, Roghmann M-C. Infection prevention in the cancer center. Clinical infectious diseases. 2013;57(4):579-85.
[75] Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. Jama. 2020;323(13):1239-42.
[76] Du M, Cai G, Chen F, Christiani DC, Zhang Z, Wang M. Multi-omics evaluation of gastrointestinal and other clinical characteristics of SARS-CoV-2 and COVID-19. Gastroenterology. 2020.
[77] Chou C-H, Chuang L-Y, Lu C-Y, Guh J-Y. Interaction between TGF-β and ACE2-Ang-(1–7)-Mas pathway in high glucose-cultured NRK-52E cells. Molecular and cellular endocrinology. 2013;366(1):21-30.
[78] Marsden J, Darke S, Hall W, Hickman M, Holmes J, Humphreys K, et al. Mitigating and learning from the impact of COVID-19 infection on addictive disorders. Addiction. 2020.
[79] Stokes EK, Zambrano LD, Anderson KN, Marder EP, Raz KM, Felix SEB, et al. Coronavirus Disease 2019 Case Surveillance—United States, January 22–May 30, 2020. Morbidity and Mortality Weekly Report. 2020;69(24):759.
[80] O'malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. Health services research. 2005;40(5p2):1620-39.
[81] Diggle P, Diggle PJ, Heagerty P, Liang K-Y, Heagerty PJ, Zeger S. Analysis of longitudinal data: Oxford University Press; 2002.
[82] Muthén L, Muthén B. Mplus. The comprehensive modelling program for applied researchers: user's guide. 2016;5.
[83] Wang J, Wang X. Structural equation modeling: Applications using Mplus: John Wiley & Sons; 2019.
[84] Moran PA. Notes on continuous stochastic phenomena. Biometrika. 1950;37(1/2):17-23.
[85] McLachlan GJ. On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. Applied statistics. 1987:318-24.
[86] Nylund KL, Asparouhov T, Muthén BO. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Structural equation modeling. 2007;14(4):535-69.
[87] Clark SL. Mixture modeling with behavioral data: University of California, Los Angeles; 2010.

6 AUTHOR CONTRIBUTION

NV, and M.S. conceived and designed research; AM, and N.V. extracted and cleaned data; M.S., and N.V. analyzed data, and interpreted the results; N.V., and M.S. drafted manuscript; G.M., and J.D. edited and revised manuscript; All authors reviewed and approved final version of the manuscript.

7 COMPETING INTEREST

The authors declare no competing interests.

8 DATA AVAILEBILITY

All datasets used in the current study are publicly available (sources are mentioned in Table S1). Datasets generated during the study are available from the corresponding author.

9 ABBREVIATION

ACE2: Angiotensin-converting enzyme 2
AIC: Akaike Information Criteria
AQI: Air quality index
ARDS: Acute respiratory distress syndrome
BIC: Bayesian information criteria
BLRT: Bootstrap Likelihood Ratio Test
CLD: Chronic Lung Disease
COPD: Chronic obstructive pulmonary disease
COVID-19: Coronavirus disease 2019
cTnI: Cardiac troponin I
CVD: Cardiovascular disease
DM: Diabetes mellitus
DPP4: Dipeptidyl peptidase 4
GEE: Generalized Estimating Equations
H1N1: Swine flu
H5N1: Avian influenza
IL: Interleukin
LF: Liver function
LGM: Latent growth model
LGMM: Latent growth mixture model
LVEF: Left ventricular ejection fraction
MIR: Mortality to incidence ratio
MR: Mortality rate

PM: Particulate matter
QIC: Quasi-likelihood information criterion
OR: Odds Ratio
REN: Relative entropy
SIR: Susceptible, infected, and recovered
TB: Tuberculosis
TGFB1: Transforming Growth Factor Beta 1
TNF: Tumor necrosis factor
WA: White American

10 FIGURES

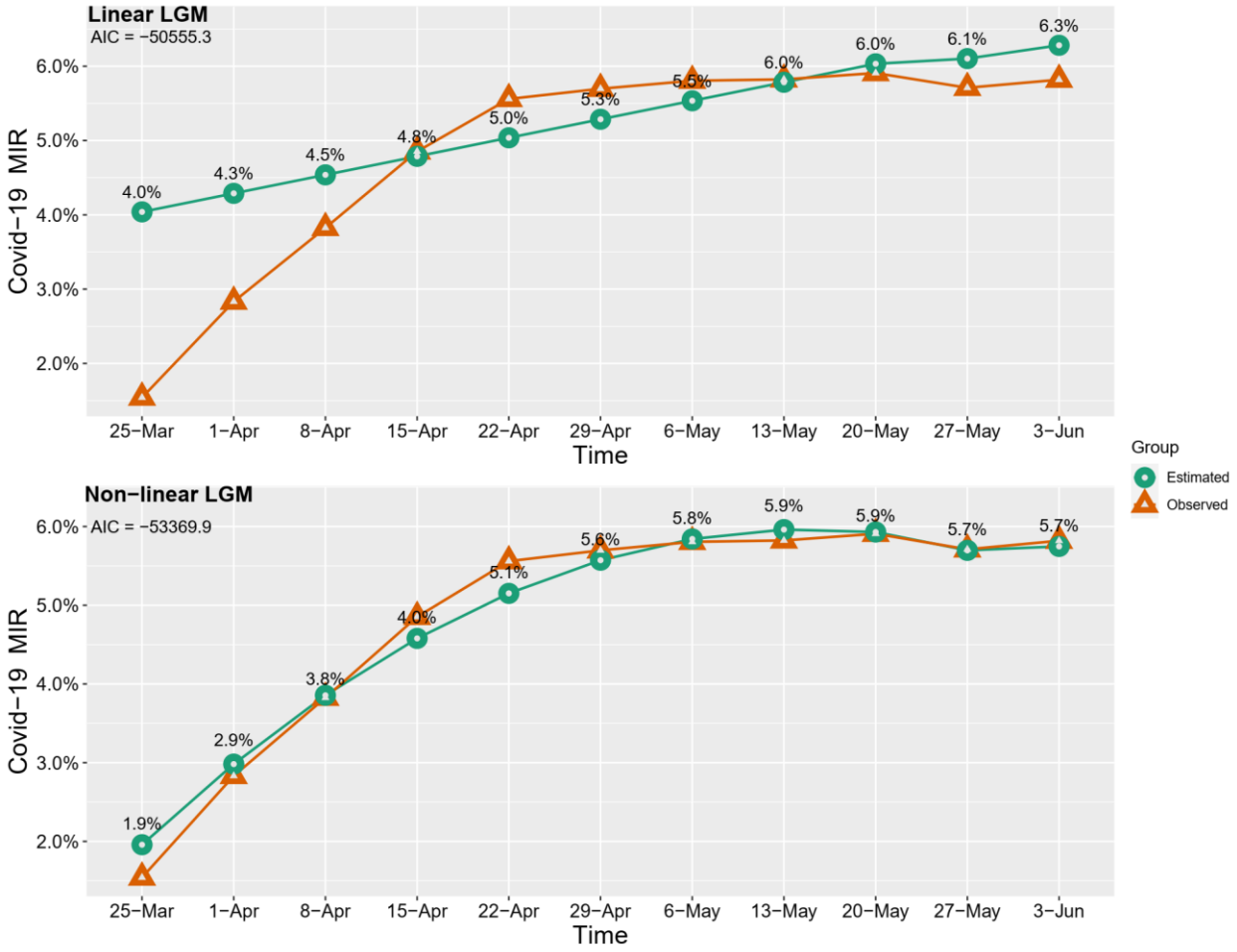

Figure 1. Overall growth trajectories of observed and estimated weekly COVID-19 MIR from March 25 to June 3, 2020, USA. Green lines indicate the estimated MIR trajectories using an LGM model (linear and non-linear). Orange lines indicate the observed mean MIR.

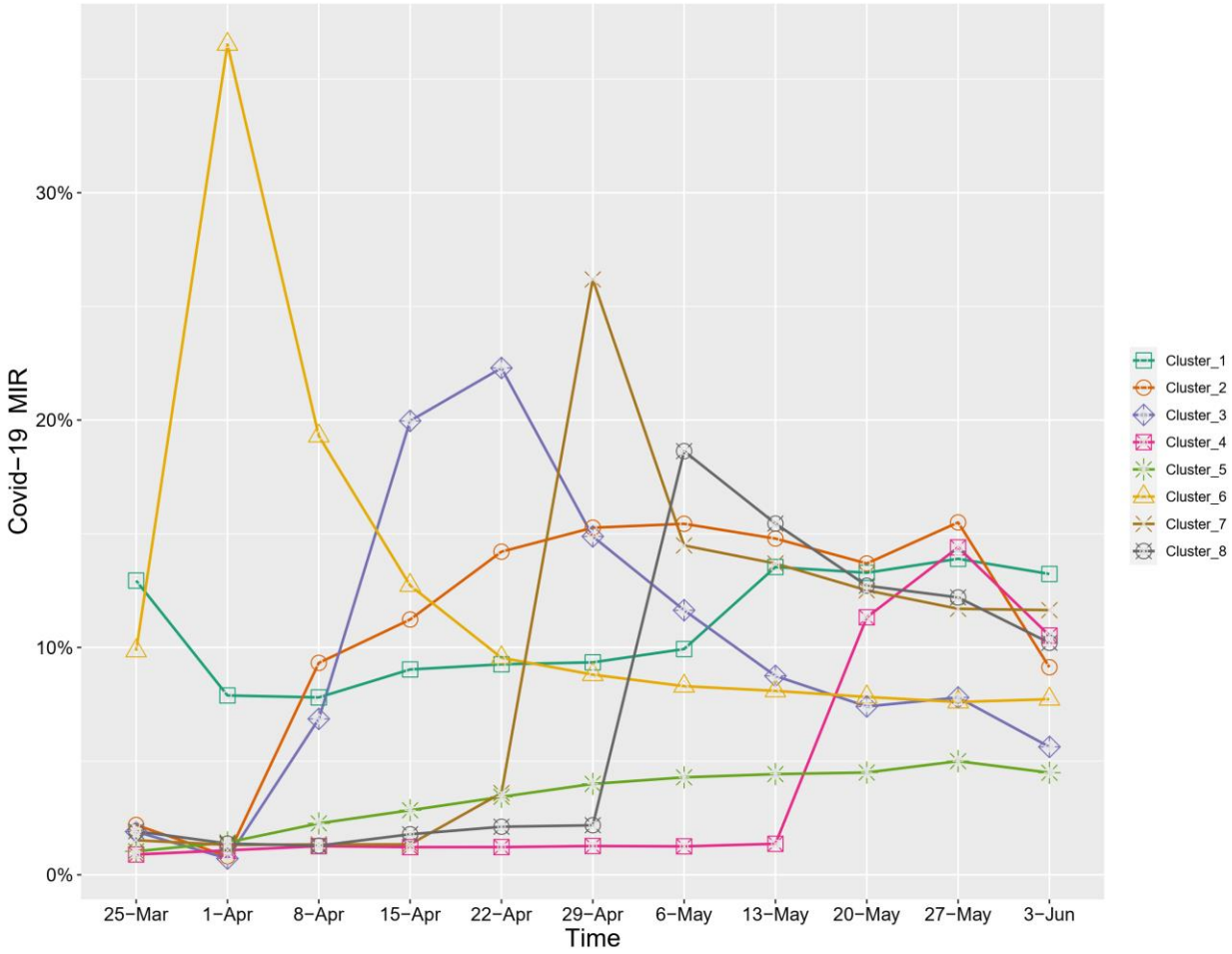

Figure 2. Estimated mean COVID-19 MIR growth trajectories from March 25 to June 3, 2020, for 8 clusters of the U.S. counties.

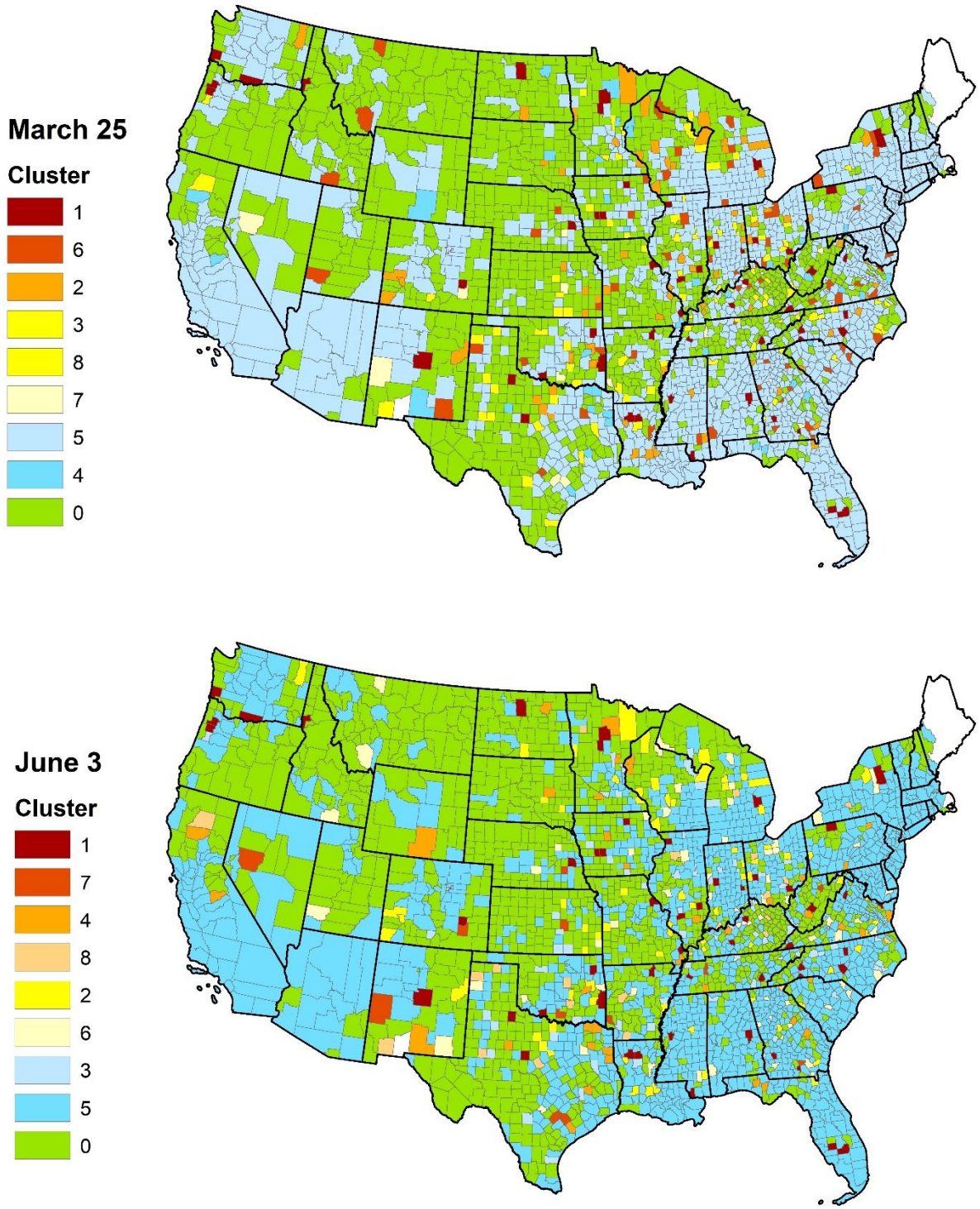

Figure 3. Geographical distribution of the 9 clusters of U.S. counties based on the estimated COVID-19 MIR growth trajectories over time (March 25, June 3, 2020), USA. Red color indicates the highest MIR, blue color indicates the lowest MIR, and green color shows the counties with MIR=0.

11 TABLES

Table 1. Descriptive statistics of COVID-19 MIR in the United States from March 25 to June 3, 2020 (n=3050 counties)

Time*	**COVID-19 MIR**					**P-value*****
	Minimum (N, %)	Maximum (N, %)	Mean	SD	Mean Difference**	
March 25	0.0 (2830, 92.8%)	1.0 (9, 0.3%)	0.8%	6.5%	*NA*	*NA*
April 1	0.0 (2507, 82.2%)	1.0 (11, 0.4%)	1.6%	7.5%	0.7%	**<0.001**
April 8	0.0 (2185, 71.6%)	1.0 (10, 0.3%)	2.1%	7.9%	0.5%	**0.004**
April 15	0.0 (1936, 63.5%)	1.0 (7, 0.2%)	2.6%	6.4%	0.5%	**0.002**
April 22	0.0 (1763, 57.8%)	1.0 (8, 0.3%)	3.0%	6.4%	0.4%	**0.020**
April 29	0.0 (1643, 53.9%)	1.0 (4, 0.1%)	3.2%	5.4%	0.1%	0.501
May 6	0.0 (1553, 50.9%)	0.55 (9, 0.3%)	3.2%	5.1%	0.08%	0.600
May 13	0.0 (1487, 48.8%)	0.50 (3, 0.1%)	3.2%	5.2%	0.02%	0.900
May 20	0.0 (1417, 46.4%)	1.0 (1, 0.0%)	3.2%	5.1%	0.02%	0.900
May 27	0.0 (1376, 45.1%)	1.0 (1, 0.0%)	3.2%	5.2%	-0.00%	0.989
June 3	0.0 (1311, 42.9%)	1.0 (1, 0.0%)	3.2%	5.0%	-0.01%	0.900

* Year of 2020

** Mean difference between mean COVID-19 MIR at each time and the previous time.

** P-values from the t-test comparing mean COVID-19 MIR in each time with the previous time

Table 2. COVID-19 MIR risk factors based on a longitudinal GEE model. Table includes both the results from the forward selection (univariate GEE model), and the final model (multivariate GEE model) from March 25 to June 3, 2020, USA.

	Variable	Variable Selection (Univariate GEE)		Final Model (Multivariate GEE)		
		Est.	P-value (univariate)*	Est.	SD	P-value (multivariate)**
	Time	0.33%	<0.001	0.21%	0.01%	<0.001
Comorbidities & Disorders	**CVD**	0.003%	0.009	0.07%	0.032%	**0.015**
	Cardiomyopathy & myocarditis	0.21%	<0.001	0.15%	0.047%	**<0.001**
	Hypertensive heart disease	0.06%	<0.001	0.10%	0.035%	**0.001**
	Peripheral vascular disease	0.42%	0.021	0.28%	0.152%	0.060
	Atrial fibrillation	-0.12%	0.006	-0.01%	0.061%	0.881
	Cerebrovascular disease	0.01%	0.071	0.06%	0.035%	0.054
	Hepatitis	0.65%	0.189	-0.62%	0.572%	0.278
	HIV/AIDS	0.21%	0.026	0.07%	0.063%	0.240
	TB	2.01%	0.001	-0.65%	0.685%	0.339
	Lower respiratory infection	0.02%	0.031	0.002%	0.011%	0.915
	Interstitial lung disease & pulmonary sarcoidosis	0.26%	<0.001	0.06%	0.089%	0.498
	Asthma	0.04%	0.980	-	-	-
	COPD	0.002%	0.703	-	-	-
	Ischemia	0.002%	0.178	0.07%	0.032%	**0.024**
	Mesothelioma	0.67%	0.013	0.72%	0.279%	**0.009**
	Tracheal cancer	0.02%	0.001	-0.03%	0.011%	**0.002**
	Leukemia	0.08%	0.402	-	-	-
	Pancreatic cancer	0.45%	<0.001	0.51%	0.106%	**<0.001**
	Rheumatic disease	0.02%	0.774	-	-	-
	Drug use disorder	0.06%	0.001	0.08%	0.025%	**<0.001**
	Alcohol use disorder	-0.07%	0.001	-0.17%	0.055%	**0.002**
Demographics & Social	**Age**	0.44%	<0.001	5.36%	8.01%	**0.032**
	Female-AA%	3.63%	<0.001	11.53%	10.56%	0.275
	Female-WA%	-3.03%	<0.001	6.18%	98.89%	0.487
	Male-AA%	3.72%	<0.001	-15.41%	11.27%	0.172
	Male-WA%	-3.10%	<0.001	-10.30%	11.27%	0.172
	Asian%	0.12%	<0.001	0.02%	0.038%	0.565
	Smokers%	0.07%	0.001	0.05%	0.053%	0.281
	Unemployed%	0.15%	0.002	-0.02%	0.063%	0.881
	Income Rate	0.35%	0.002	-0.03%	0.014%	0.802
	Uninsured%	-0.03%	0.037	0.02%	0.021%	0.233
Environme	AQI	0.09%	<0.001	0.06%	0.152%	0.706
	Temperature	0.04%	<0.001	0.01%	0.017%	0.628

PM	0.35%	<0.001	0.03%	0.561%	0.962

* Univariate P-value<0.2 is considered as significant.

** Multivariate P-value<0.05 is considered as significant.

Table 3. Results of the 8-class GLMM clustering of the 1736 counties (with MIR>0) based on COVID-19 MIR form March 25 to June 3, 2020, USA

Cluster	Cluster Size	Intercept*		Slope**	
	N (%)	Mean (SE)	P-value	Mean (SE)	P-value
0	1314 (43.1%)	0% (0%)	*NA*	0% (0%)	*NA*
1	52 (1.7%)	12.9% (3.1%)	<0.001	-1.0% (0.6%)	0.122
2	74 (2.4%)	2.2% (0.8%)	0.010	3.5% (1.0%)	<0.001
3	66 (2.1%)	1.9% (0.9%)	0.027	2.8% (0.4%)	<0.001
4	39 (1.3%)	0.9% (0.5%)	0.089	2.0% (0.4%)	<0.001
5	1406 (46.1%)	1.0% (0.3%)	<0.001	-3.0% (0.5%)	<0.001
6	64 (2.1%)	9.8% (3.0%)	0.001	3.4% (0.7%)	<0.001
7	12 (0.4%)	1.5% (1.3%)	0.236	-3.1% (0.5%)	<0.001
8	23 (0.8%)	1.9% (1.3%)	0.127	-4.2% (0.0%)	*NA*

* Intercept indicates the estimated mean MIR of COVID-19 at the beginning of the study on March 25, 2020, for each cluster.

** Slope indicates the overall change of MIR of COVID-19 over the study time, for each cluster.

Table 4. Significant risk factors and their odds ratios in each cluster compare to cluster 0 (MIR=0). Blank spots indicate the insignificant risk factors

	Variable	Cluster							
		1	2	3	4	5	6	7	8
Comorbidities & Disorders	CVD								
	Cardiomyopathy & myocarditis								1.3*
	Hypertensive heart disease								
	Peripheral vascular disease					0.5			
	Atrial fibrillation					0.8			
	Cerebrovascular disease					0.9			
	Hepatitis	0.5	2.0		5.1		0.1	0.7	1.8
	HIV/AIDS					1.2			
	TB	0.7	0.5	1.9	0.4		0.7	1.3	0.6
	Lower respiratory infection	0.9	0.9			0.6			
	Interstitial lung disease & pulmonary sarcoidosis					1.3	1.3		
	Ischemia								
	Mesothelioma		3.8		2.8	2.4			6.0
	Tracheal cancer		0.9		1.0	0.9			
	Pancreatic cancer		1.6			1.2			
	Drug use disorder		1.1	1.1		1.1		1.1	
	Alcohol use disorder			0.8		0.9			
Demographics & Social	Age		0.9	0.8		0.8	0.8	0.9	
	Female-AA%				2.7	19.5		0.9	0.7
	Female-WA%	0.3	0.2		0.2	0.3		0.6	
	Male-AA%		2.0		2.2			0.7	0.6
	Male-WA%	0.3	0.1		0.2	3.8		0.7	1.3
	Asian%	1.6	0.6		1.8	1.8		0.1	1.8
	Smokers%				0.8				
	Unemployed%				1.2	0.1			
	Income Rate					0.8			
	Uninsured%					0.9			
Environmental	AQI	1.2		1.4	1.3	1.2	1.2		1.5
	Temperature	1.1				1.1			
	PM	0.6		0.4	0.4		0.6	1.1	0.3

* for instance, OR=1.3 means that 1% increase in cardiomyopathy & myocarditis MR is associated with 30% increase in the relative log odds of being in cluster 8 vs. cluster 0 (MIR=0).