COVID-Net: A deep learning based and interpretable predication model for the county-wise trajectories of COVID-19 in the United States

Ting Tian[a,1], Yuknag Jiang[a,1], Yuting Zhang[a,1], Zhongfei Li[d], Xueqin Wang[b,2], and Heping Zhang[c,2]

[a]*School of Mathematics, Sun Yat-sen University*

[b]*School of Management,University of Science and Technology of China*

[c]*School of Public Health, Yale University*

[d]*School of Management,Sun Yat-sen University*

Abstract

The confirmed cases of novel coronavirus disease (COVID-19) have been reported in the United States since late January 2020. There were over 4.8 million confirmed cases and about 320,000 deaths as of May 19, 2020 in the world. We examined the characteristics of the confirmed cases and deaths of COVID-19 in all affected counties of the United States. We proposed a COVID-Net combining the architecture of both Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) by using the trajectories of COVID-19 during different periods until May 19, 2020, as the training data. The validation of the COVID-Net was performed by predicting the numbers of confirmed cases and deaths in subsequent 3-day, 5-day, and 7-day periods. The

[1]Ting Tian, Yukang Jiang and Yuting Zhang contributed equally to this article

[2]Xueqin Wang and Heping Zhang are corresponding authors

COVID-Net produced relatively smaller Mean Relative Errors (MREs) for the 10 counties with the most severe epidemic as of May 19, 2020. On average, MREs were 0.01 for the number of confirmed cases in all validation periods, and 0.01, 0.01, and 0.03 for the number of deaths in the 3-day, 5-day, and 7-day periods, respectively. The COVID-Net incorporated five risk factors of COVID-19 and was used to predict the trajectories of COVID-19 in Hudson County, New Jersey and New York County, New York until June 28, 2020. The risk factors include the percentage of the population with access to exercise opportunities, average daily $PM_{2.5}$, population size, preventable hospitalization rate, and violent crime rate. The expected number of cumulative confirmed cases and deaths depends on the dynamics of these five risk factors.

Significance Statement:

A COVID-Net model was built to predict the trajectories of COVID-19, based on the percentage of the population with access to exercise opportunities, average daily $PM_{2.5}$, population size, preventable hospitalization rate and violent crime rate in the metropolises areas. The increasing awareness of how these risk factors affect the pandemic helps policymakers develop plans that mitigate the spread of COVID-19.

Classification: Social Sciences, Statistics

Keywords: Deep learning; COVID-Net model; Access to exercise opportunities; Air pollution; Preventable hospital stays

1. INTRODUCTION

According to The New York Times (2020), the early confirmed cases of COVID-19 were reported on January 21, 2020 in the United States. In March, the outbreak of COVID-19 was proclaimed as a pandemic by the World Health Organization (2020). Since then, the United States has had the largest number of both confirmed cases and deaths in the world (National Health Commission of the People's Republic of China, 2020), where the confirmed cases and deaths were 1,536,447 and 91,936, respectively, as of May 19, 2020.

A vast majority of states in the United States had issued a stay at home order to reduce the transmission of COVID-19 since March 2020 (Governor's Press Office, 2020). As the states are in the process of reopening for normalcy, it is important to predict the trajectories of COVID-19 based on the risk factors to provide the decision-makers with a quantitative and dynamic assessment.

Besides personal factors, county-level factors are known to affect the spread of infectious diseases. For instance, the hospitalization rate of H1N1 2009 had a disproportionate impact on high-poverty areas in New York City (Balter et al., 2010) and on the small population of racial/ethnic groups in Wisconsin (Truelove et al., 2011). Consequently, we considered the data from county health ranking and roadmaps programs (A Robert Wood Johnson Foundation program, 2020). The details about the database are available from "https://www.countyhealthrankings .org/reports/county-health-rankings-reports." Specifically, we used five risk factors related to diet and exercise, air quality, population size, quality of care and community safety (See Materials and Methods section). Using deep learning, we included data from all counties in the United States to build a prediction model and selected the 10 most severe counties to validate the model. Then, we predicted future trajectories of COVID-19 in Hudson County, New Jersey and New York County, New York.

2. MATERIALS AND METHODS

2.1 Data Sources

We collected the daily numbers of cumulative confirmed cases and deaths from January 21 to May 19, 2020 for counties in the United States from the New York Times (New York Times, 2020). The daily cumulative confirmed cases and deaths were collected from health departments and U.S. Centers for Disease Control and Prevention (CDC), where patients were identified as "confirmed" based on the laboratory positive tests and clinical symptoms and exposure (New York Times, 2020). The five risk factors were compiled from 2020 annual data on the County Health Rankings and Roadmaps program official website (A Robert Wood Johnson Foundation program, 2020). The five risk factors including the percentage of the population with access to exercise opportunities, average daily $PM_{2.5}$, population size, preventable hospitalization rate and violent crime rate were listed in Table 1. Data analysis was conducted in version 3.7 Python with TensorFlow-GPU 1.14.0 and Keras 2.3.0.

Table 1: The definitions and sources of the selected five risk factors, and the literature supporting these factors.

Factors	**Meanings**	**Sources**	**Related topics**	**Previous studies**
Access to exercise opportunities	Percentage of population with access to locations for physical activity	Business Analyst, Delorme map data, ESRI, & US Census Tigerline Files	Social Distancing;Outdoor and indoor exercise	Novel
Average daily density of $PM_{2.5}$	Average daily density of fine particulate matter in micrograms per cubic meter ($PM_{2.5}$)	Environmental Public Health Tracking Network	Air pollution	Conticini, Frediani and Caro (2020), Contini and Costabile (2020), Martelletti and Martelletti (2020)
Population	Resident population	Census Population Estimates	Population density	Novel

Continued on next page

Table 1 – continued from previous page

Factors	Meanings	Sources	Related topics	Previous studies
Preventable Hospitalization Rate	Rate of hospital stays for ambulatory-care sensitive conditions per 100,000 Medicare enrollees	Mapping Medicare Disparities Tool	Poor health status; Healthcare providers	Novel
Violent crime rate	Number of reported violent crime offenses per 100,000 population	Uniform Crime Reporting - FBI	Community Safety; Psycholog ical stress	Novel

2.2 Statistical Analysis

We combined the architecture of LSTM and GRU (Hochreiter and Schmidhuber, 1997; Chung et al., 2014; Bandyopadhyay and Dutta, 2020) by using the five risk factors as auxiliary inputs. The COVID-Net was shown in Fig. 1. During the training process, the observed cumulative confirmed cases and deaths every past seven days in each county of the United States were used to predict the cumulative confirmed cases and deaths in the present day. All risk factors involved in the model were min-max normalized before used.

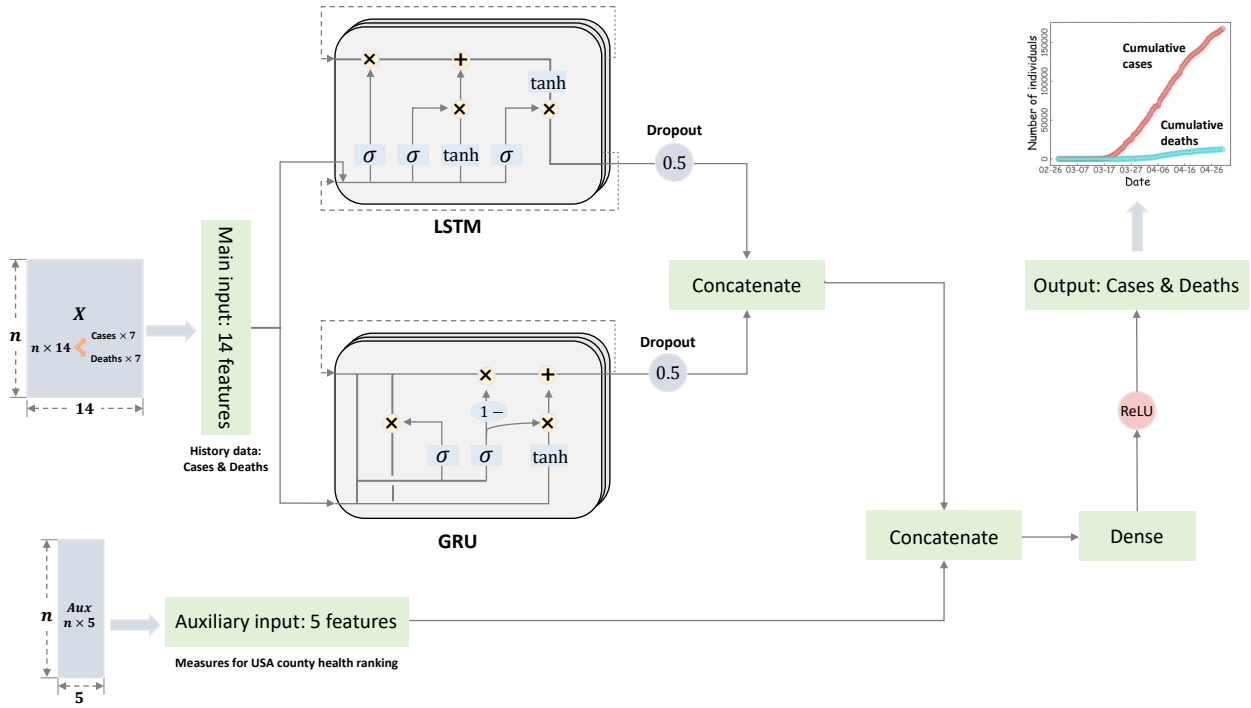

Figure 1: **The COVID-Net combining Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) using the five risk factors.**

Our models were divided into two parts. The first part was to learn the observed patterns of COVID-19 and then to validate the learned patterns, where the accuracy of the models was evaluated by Mean Relative Errors (MRE_t).

$$MRE_t = \frac{1}{t}\sum_{i=1}^{t}\frac{|Actual_i - Predicted_i|}{Actual_i}, t = 3, 5, 7. \tag{2.1}$$

In the first part, we conducted three experiments:

- (a) t=3: the training data were from January 21 to May 16, 2020, and the validation data from May 17 to May 19, 2020;
- (b) t=5: the training data were from January 21 to May 14, 2020, and the validation data from May 15 to May 19, 2020;

- (c) t=7: the training data were from January 21 to May 12, 2020, and the validation data from May13 to May 19, 2020.

The second part used the data after the training data period until June 28, 2020. We simulated the epidemic by considering different levels of the five covariates in the near future.

3. RESULTS

3.1 Model validation

Our COVID-Net was built by using cumulative counts of confirmed cases and deaths in three overlapping periods from January 21 to May 16, 2020, to May 14, 2020, and to May 12, 2020 as the training data. The mean relative errors (MREs) between the observed and projected counts from the day after the training period to May 19, 2020 were computed to assess the accuracy of the prediction. Depending on the training period, we obtained 3-day, 5-day, and 7-day MREs for the 10 most severe counties. Table 2 presented both individual and average MREs for those 10 counties. New York County, New York had the smallest MREs for the 3-day prediction of the confirmed cases and 5-day prediction of deaths. Also, Westchester County, New York had the smallest MREs for the 5-day and 7-day predictions of the confirmed cases. Nassau County, New York had the smallest MREs in the prediction of deaths for the other periods. Overall, all the averaged MREs were relatively small, assuring the accuracy of our COVID-Net model in predicting future trajectories of COVID-19 for the numbers of both confirmed cases and deaths. Additional data were presented in Figs. S1 and S2 of the Supporting Information (SI) Appendix.

Table 2: 10 most severe counties of COVID-19 and their MREs of cumulative confirmed cases and deaths.

State, County	Confirmed Cases			Deaths		
	MRE_3	MRE_5	MRE_7	MRE_3	MRE_5	MRE_7
New York, New York	0.003	0.021	0.025	0.006	0.001	0.010
Illinois, Cook	0.018	0.035	0.010	0.007	0.021	0.084
California, Los Angeles	0.016	0.008	0.030	0.020	0.012	0.089
New York, Nassau	0.021	0.014	0.003	0.002	0.001	0.005
New York, Suffolk	0.016	0.004	0.003	0.017	0.009	0.008
New York, Westchester	0.009	0.002	0.003	0.023	0.014	0.008
Pennsylvania, Philadelphia	0.018	0.009	0.018	0.016	0.013	0.017
Massachusetts, Middlesex	0.011	0.007	0.005	0.003	0.005	0.006
Michigan, Wayne	0.005	0.003	0.005	0.021	0.021	0.013
New Jersey, Hudson	0.007	0.002	0.008	0.010	0.007	0.007
Average	**0.012**	**0.010**	**0.011**	**0.013**	**0.010**	**0.025**

3.2 Prediction of future trajectories of COVID-19

Our COVID-Net model incorporated five risk factors to predict future trajectories of COVID-19 in both cumulative confirmed cases and deaths from May 20 to June 28, 2020. The results for the 10 most severe counties were shown in Fig. S3 in the SI Appendix. The weights of the five risk factors in the construction of the COVID-Net models were presented in Table S1 in the SI Appendix.

To offer insights into the prediction dynamics of the COVID-Net, we varied the levels of the five risk factors and presented the resulting trajectories of COVID-19 for New York County, New York was shown in Fig. 2. The impact of the five risk factors on COVID-19 in both cumulative confirmed cases and deaths was visible, depending on their weights of the risk factors (Table S1 in the SI Appendix). Specifically, increases in the percentage

of the population with access to exercise opportunities, average daily $PM_{2.5}$, population size, preventable hospitalization rate, and violent crime rate would increase in both the cumulative confirmed cases and deaths. For example, on June 28, 2020, if the percentage of the population with access to exercise opportunities was changed to 4 times versus 0.5 times of that on May 20, 2020, the cumulative confirmed cases would increase to 225,000 versus 204,000. If the violent crime rate was varied to 4 times versus 0.5 times of that on May 20, 2020, the cumulative confirmed cases would increase to 210,000 versus 206,000. Overall, the numbers of cumulative confirmed cases and deaths were projected to increase slowly in the coming month in New York County, New York. The results were similar for Hudson County, New Jersey as shown in Fig. S4 in the SI Appendix. Nonetheless, the trajectories were expected to have sustainable growth in the coming month.

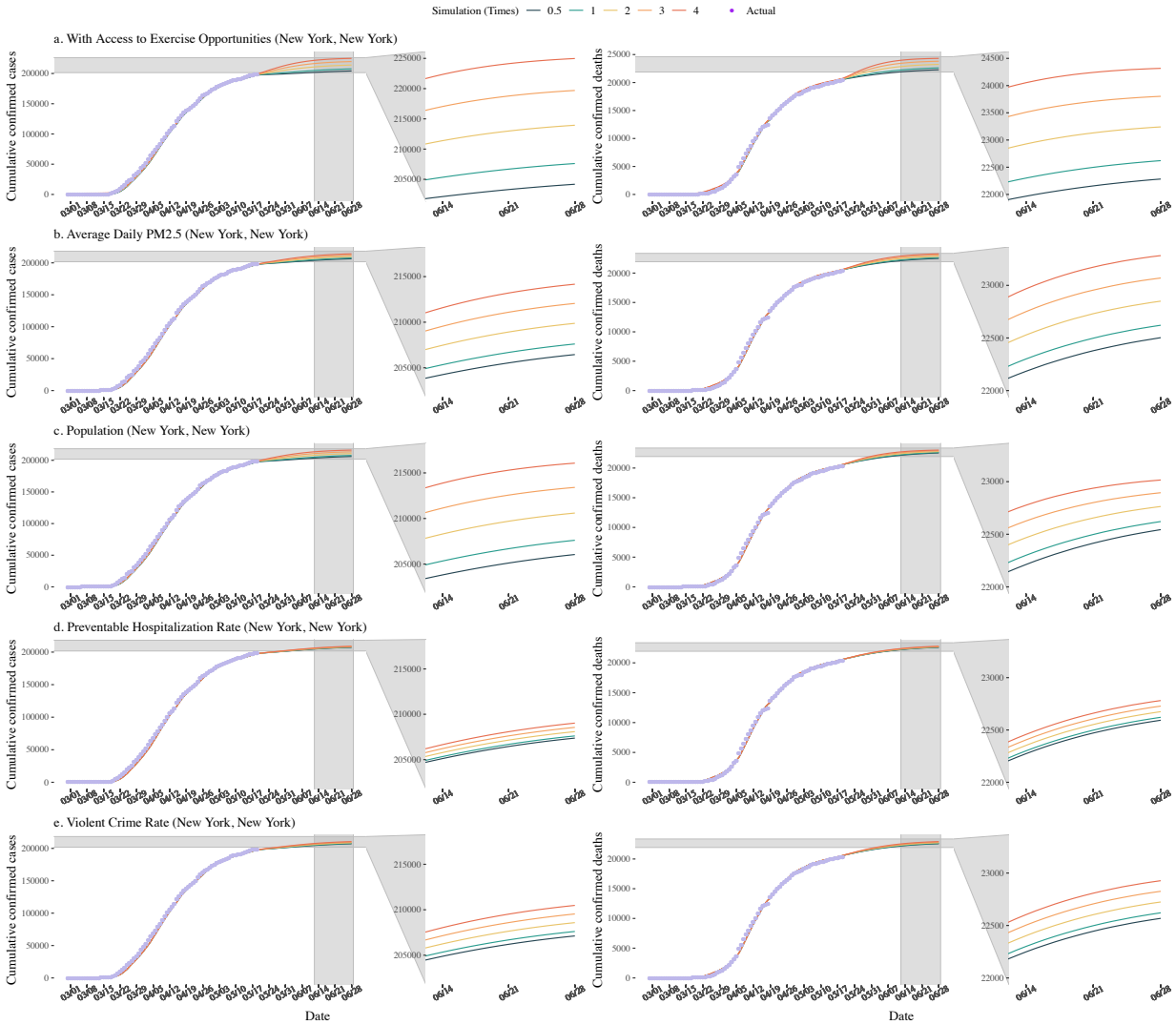

Figure 2: **The trajectories of COVID-19 for New York County, New York of cumulative confirmed cases and deaths from March 8 to June 28, 2020.**. The levels of the five risk factors were changed from 0.5 times to 4 times of those on May 20, 2020. The trajectories in the training period were observed.

4. DISCUSSION

Our COVID-Net was built by deep learning and was shown to effectively model and predict the cumulative confirmed cases and deaths in the counties of the United States. The risk factors that were used in the COVID-Net provided the visible evidence on actionable steps

that influenced the trajectories of COVID-19. Thus, COVID-Net took the advantages of both deep learning and interpretability of risk factors.

In our study, we found that the higher percentage of the population with access to locations for physical activity (e.g. indoor activities in the gym and outdoor activities in the squares), the higher risk of COVID-19 spread, likely by negating the effect the social distancing (McCloskey et al., 2020; Joscha Weber, 2020), which was identified to reduce transmission (Anderson et al., 2020; Greenstone and Nigam, 2020).

Several studies indicated that air pollution could influence the COVID-19 spread (Contini and Costabile, 2020; Martelletti and Martelletti, 2020; Conticini et al., 2020). We focused on the county-level average daily density (mg/m^3) of fine particulate matter ($PM_{2.5}$) because high levels of average daily $PM_{2.5}$ were found to increase in cumulative confirmed cases and deaths (Wu et al., 2020). Exposed to high levels of $PM_{2.5}$ led to systemic inflammation (Conticini et al., 2020; Yang et al., 2019), and chronic respiratory conditions (Cao et al., 2020). In the end, a weakened immune system may result in acute respiratory distress syndrome (ARDS) and even death.

High population density was a known infectious agent (Rocklöv and Sjödin, 2020). Metropolitan areas generally had larger populations and higher densities, and higher numbers of cumulative confirmed cases and deaths. For example, New York County, with an estimated population of 1,628,706 as of July 1, 2019 (United States Census Bureau, 2019) and under 23 square miles of area (GIS Geography, 2020), had 198,710 confirmed cases and 20,376 deaths as of May 19, 2020 (New York Times, 2020).

The higher number of Medicare enrollees who stayed for ambulatory-care sensitive conditions in hospitals, the more Medicare enrollees would have with poor health status, leading them to be vulnerable to novel diseases (Hatch et al., 2018) and become infected in the hospitals. Also, the more enrollees in the hospitals, the more healthcare providers were required, increasing the pressure of health care workers and exposing more people at risk. The higher violent crime rate may not be directly related to the COVID-19, but it could be a surrogate variable for community safety. If the community environment was not safe,

which might cause psychological distress (Wang et al., 2020; Ho et al., 2019) and trigger mental health illness (Xiang et al., 2020), especially under the "stay-at-home" order, it was reasonable to expect that the change in the violent crime rate would be relatively small in the near future. Therefore, other factors were expected to make greater differences in the COVID-19 trajectories.

There might be other factors that we could consider in building the COVID-Net. We chose to use the five risk factors because of the documented evidence in the literature, the interpretability of the results, and the accuracy of the prediction. Our goal was not to identify risk factors, but to build an accurate and interpretable prediction tool for the outbreak of COVID-19. Because COVID-Net used only five risk factors, the small number of variables made it easier to interpret and act for decision-makers, and with its relatively small MREs, the room for further prediction improvement was expected to be too small to make a practical difference.

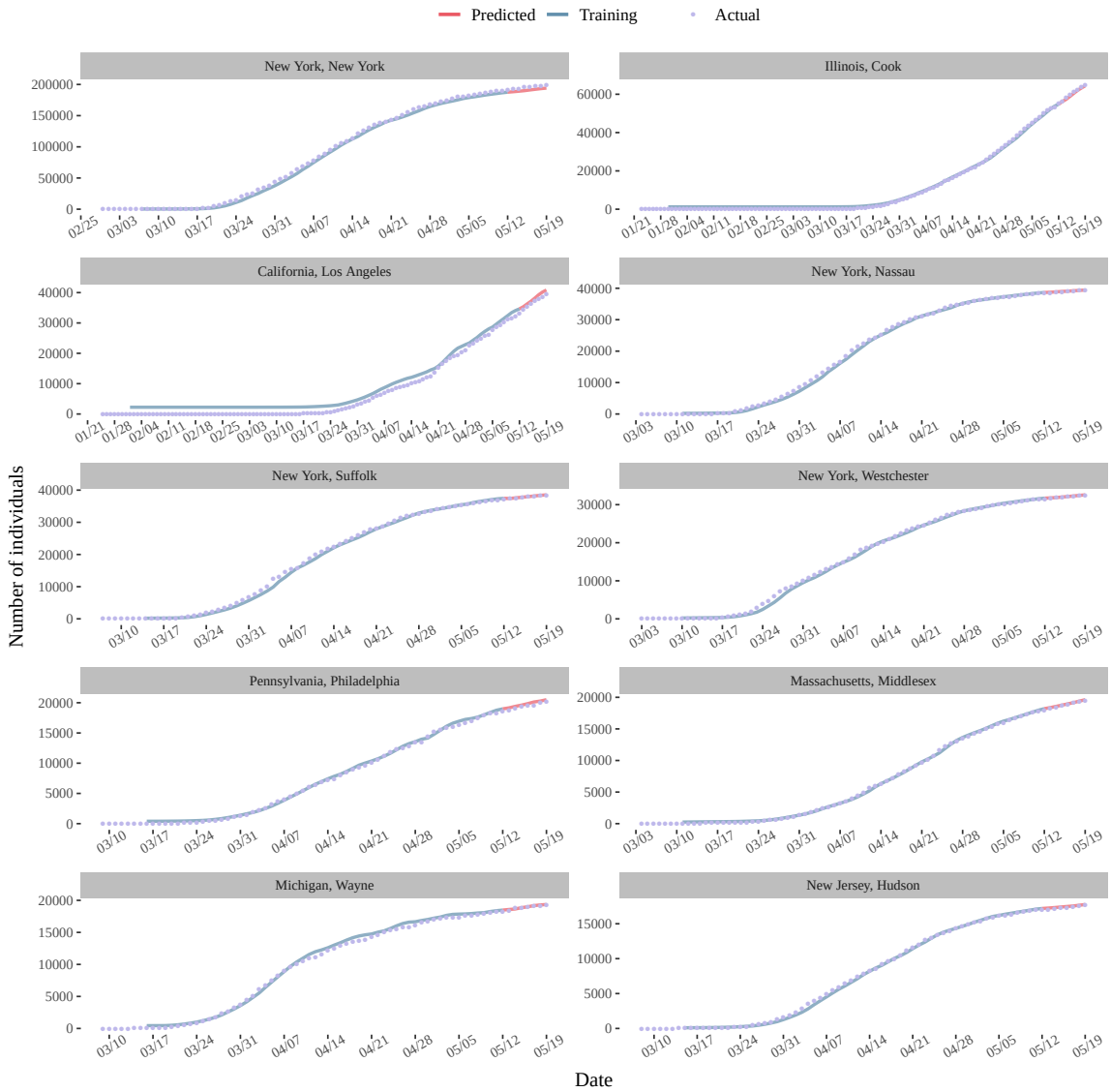

Figure S1: **The trajectories of COVID-19 for the 10 most severe counties as of May 19, 2020.** The dates of the initially confirmed cases were different for these 10 counties. Both observed confirmed cases and the projected ones were represented by purple points and red curves from May 13 to May 19, 2020.

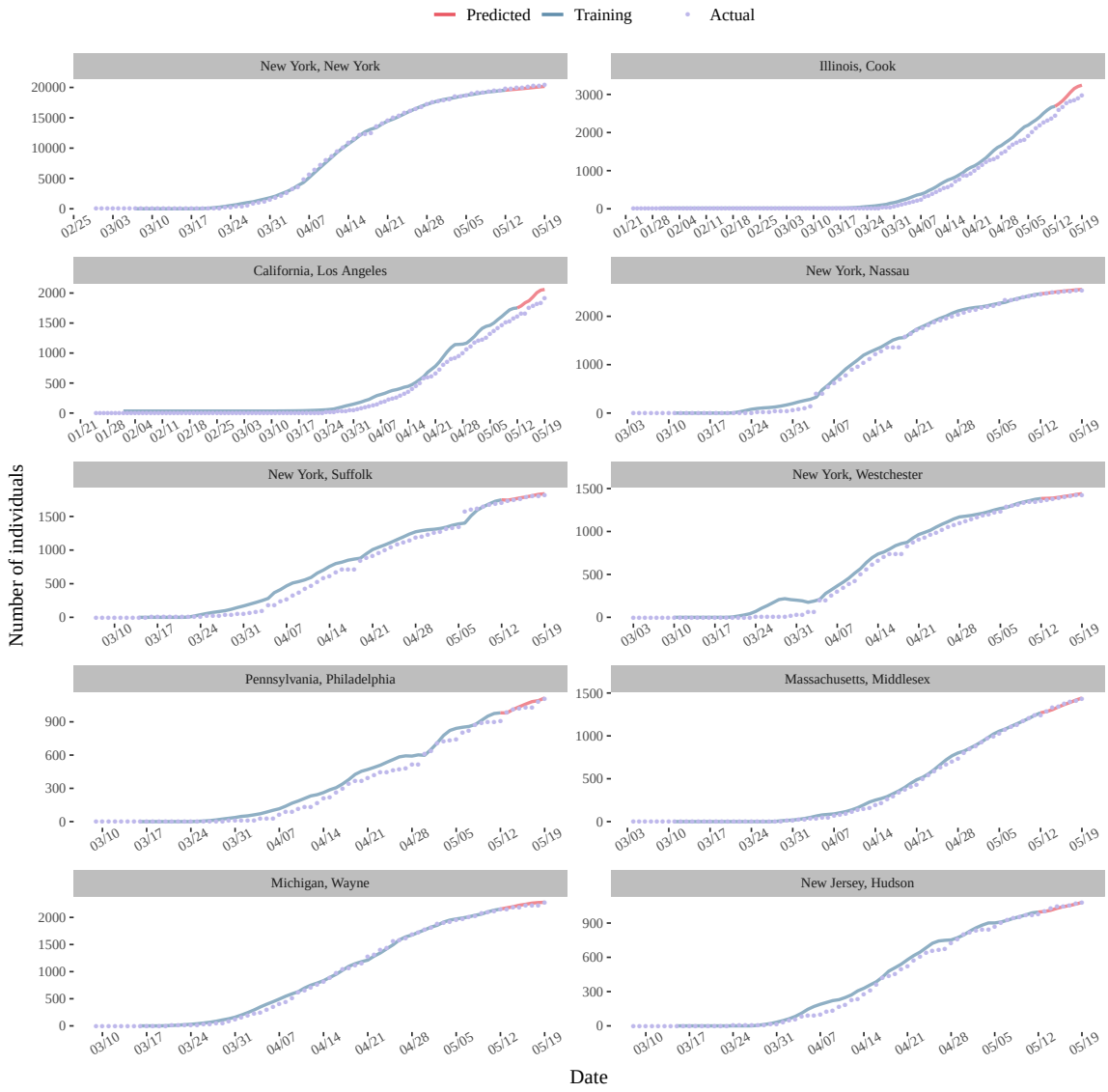

Figure S2: **The trajectories of COVID-19 for the 10 most severe counties as of May 19, 2020.**The dates of the initial deaths were different for these 10 counties. Both observed deaths and the projected ones were represented by purple points and red curves from May 13 to May 19, 2020.

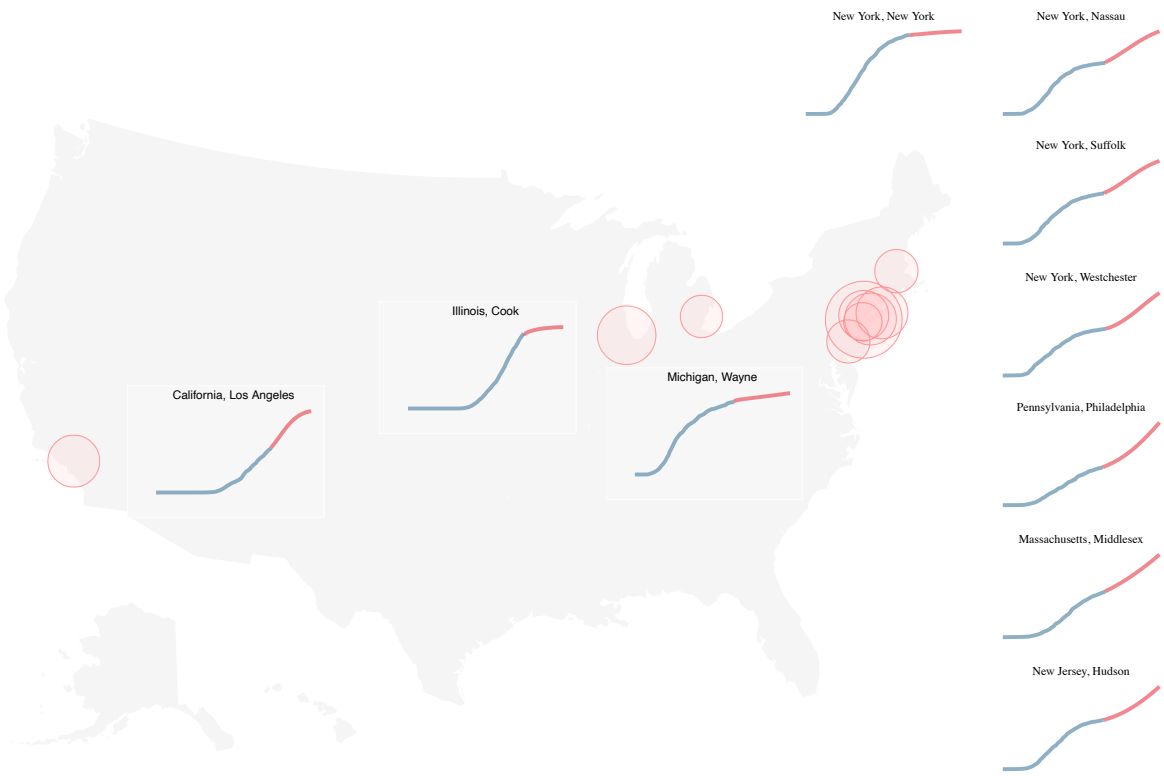

Figure S3: **The map of the 10 most severe counties.** The trajectories of COVID-19 for these counties until June 28, 2020 were displayed. The blue curves indicated the observed cumulative confirmed cases while the red curves indicated the projected ones from May 20 to June 28, 2020.

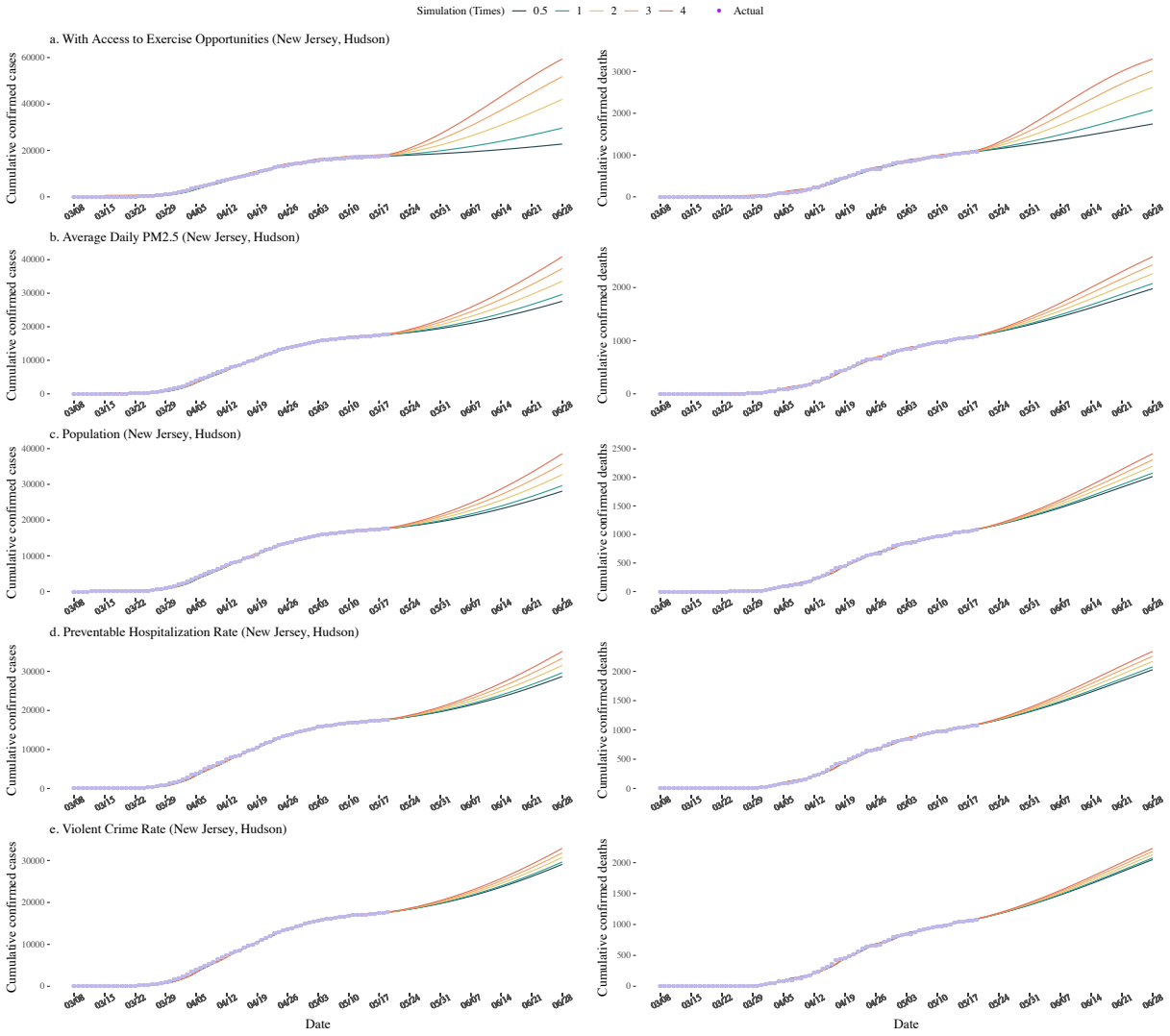

Figure S4: **The trajectories of COVID-19 for Hudson County, New Jersey in cumulative confirmed cases and deaths from March 1 to June 28, 2020.** The levels of the five risk factors were changed between 0.5 times and 4 times of that on May 20, 2020. The observed patterns of COVID-19 were from March 1 to May 19, 2020, and the projected ones from May 20 to June 28, 2020.

Table S1: The weights of the five risk factors in the COVID-Net.

Factors	**Parameters estimation for confirmed cases**	**Parameters estimation for deaths**
Access to exercise opportunities	0.003	0.003
Average daily density of $PM_{2.5}$	0.002	0.002
Population	0.012	0.002
Preventable Hospitalization Rate	0.001	0.002
Violent crime rate	0.001	0.001

ACKNOWLEDGMENTS

We would like to thank all individuals who collected epidemiological data of the COVID-19 outbreak, and the data in the county health ranking and roadmaps program.

REFERENCES

A Robert Wood Johnson Foundation program (2020), "County Health Rankings Reports,". **URL:** *https://www.countyhealthrankings.org/reports/county-health-rankings-reports*

Anderson, R. M., Heesterbeek, H., Klinkenberg, D., and Hollingsworth, T. D. (2020), "How will country-based mitigation measures influence the course of the COVID-19 epidemic?," *The Lancet*, 395(10228), 931–934.

Balter, S., Gupta, L. S., Lim, S., Fu, J., Perlman, S. E., Team, N. Y. C. . H. F. I. et al. (2010), "Pandemic (H1N1) 2009 surveillance for severe illness and response, New York, New York, USA, April–July 2009," *Emerging infectious diseases*, 16(8), 1259.

Bandyopadhyay, S. K., and Dutta, S. (2020), "Machine learning approach for confirmation of covid-19 cases: Positive, negative, death and release," *medRxiv*, .

Cao, Y., Chen, M., Dong, D., Xie, S., and Liu, M. (2020), "Environmental pollutants damage airway epithelial cell cilia: Implications for the prevention of obstructive lung diseases," *Thoracic Cancer*, 11(3), 505–510.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014), "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, .

Conticini, E., Frediani, B., and Caro, D. (2020), "Can atmospheric pollution be considered a co-factor in extremely high level of SARS-CoV-2 lethality in Northern Italy?," *Environmental pollution*, p. 114465.

Contini, D., and Costabile, F. (2020), "Does Air Pollution Influence COVID-19 Outbreaks?,".

GIS Geography (2020), "New York County Map,". **URL:** *https://gisgeography.com/new-york-county-map.html*

Governor's Press Office (2020), "Governor Cuomo Signs the 'New York State on PAUSE' Executive Order,".

URL: *https://www.governor.ny.gov/news/governor-cuomo-signs-new-york-state-pause-executive-order*

Greenstone, M., and Nigam, V. (2020), "Does social distancing matter?," *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2020-26).

Hatch, R., Young, D., Barber, V., Griffiths, J., Harrison, D. A., and Watkinson, P. (2018), "Anxiety, Depression and Post Traumatic Stress Disorder after critical illness: a UK-wide prospective cohort study," *Critical care*, 22(1), 310.

Ho, C. S., Tan, E. L., Ho, R., and Chiu, M. Y. (2019), "Relationship of anxiety and depression with respiratory symptoms: Comparison between depressed and non-depressed smokers in Singapore," *International journal of environmental research and public health*, 16(1), 163.

Hochreiter, S., and Schmidhuber, J. (1997), "Long short-term memory," *Neural computation*, 9(8), 1735–1780.

Joscha Weber (2020), "Coronavirus: Are outdoor sports healthy exercise or a dangerous risk?,".

URL: *https://www.dw.com/en/coronavirus-are-outdoor-sports-healthy-exercise-or-a-dangerous-risk/a-52971973*

Martelletti, L., and Martelletti, P. (2020), "Air pollution and the novel Covid-19 disease: a putative disease risk factor," *SN Comprehensive Clinical Medicine*, pp. 1–5.

McCloskey, B., Zumla, A., Ippolito, G., Blumberg, L., Arbon, P., Cicero, A., Endericks, T., Lim, P. L., and Borodina, M. (2020), "Mass gathering events and reducing further global spread of COVID-19: a political and public health dilemma," *The Lancet*, 395(10230), 1096–1099.

National Health Commission of the People's Republic of China (2020), "Distribution of COVID-19 cases in the world. [accessed 2020 April 26],".

URL: *http://2019ncov.chinacdc.cn/2019-nCoV/global.html*

New York Times (2020), "Coronavirus in the US: Latest Map and Case Count,".
URL: *https://github.com/nytimes/covid-19-data*

Rocklöv, J., and Sjödin, H. (2020), "High population densities catalyse the spread of COVID-19," *Journal of travel medicine*, 27(3), taaa038.

Truelove, S. A., Chitnis, A. S., Heffernan, R. T., Karon, A. E., Haupt, T. E., and Davis, J. P. (2011), "Comparison of patients hospitalized with pandemic 2009 influenza A (H1N1) virus infection during the first two pandemic waves in Wisconsin," *Journal of Infectious Diseases*, 203(6), 828–837.

United States Census Bureau (2019), "County Population Totals: 2010-2019,".
URL: *https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-total.html*

Wang, C., Pan, R., Wan, X., Tan, Y., Xu, L., Ho, C. S., and Ho, R. C. (2020), "Immediate psychological responses and associated factors during the initial stage of the 2019 coronavirus disease (COVID-19) epidemic among the general population in China," *International journal of environmental research and public health*, 17(5), 1729.

World Health Organization (2020), "WHO Director-General's opening remarks at the media briefing on COVID-19-11 March 2020,".
URL: *https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020*

Wu, X., Nethery, R. C., Sabath, B. M., Braun, D., and Dominici, F. (2020), "Exposure to air pollution and COVID-19 mortality in the United States," *medRxiv*, .

Xiang, Y.-T., Yang, Y., Li, W., Zhang, L., Zhang, Q., Cheung, T., and Ng, C. H. (2020), "Timely mental health care for the 2019 novel coronavirus outbreak is urgently needed," *The Lancet Psychiatry*, 7(3), 228–229.

Yang, J., Chen, Y., Yu, Z., Ding, H., and Ma, Z. (2019), "The influence of PM2. 5 on lung injury and cytokines in mice," *Experimental and therapeutic medicine*, 18(4), 2503–2511.