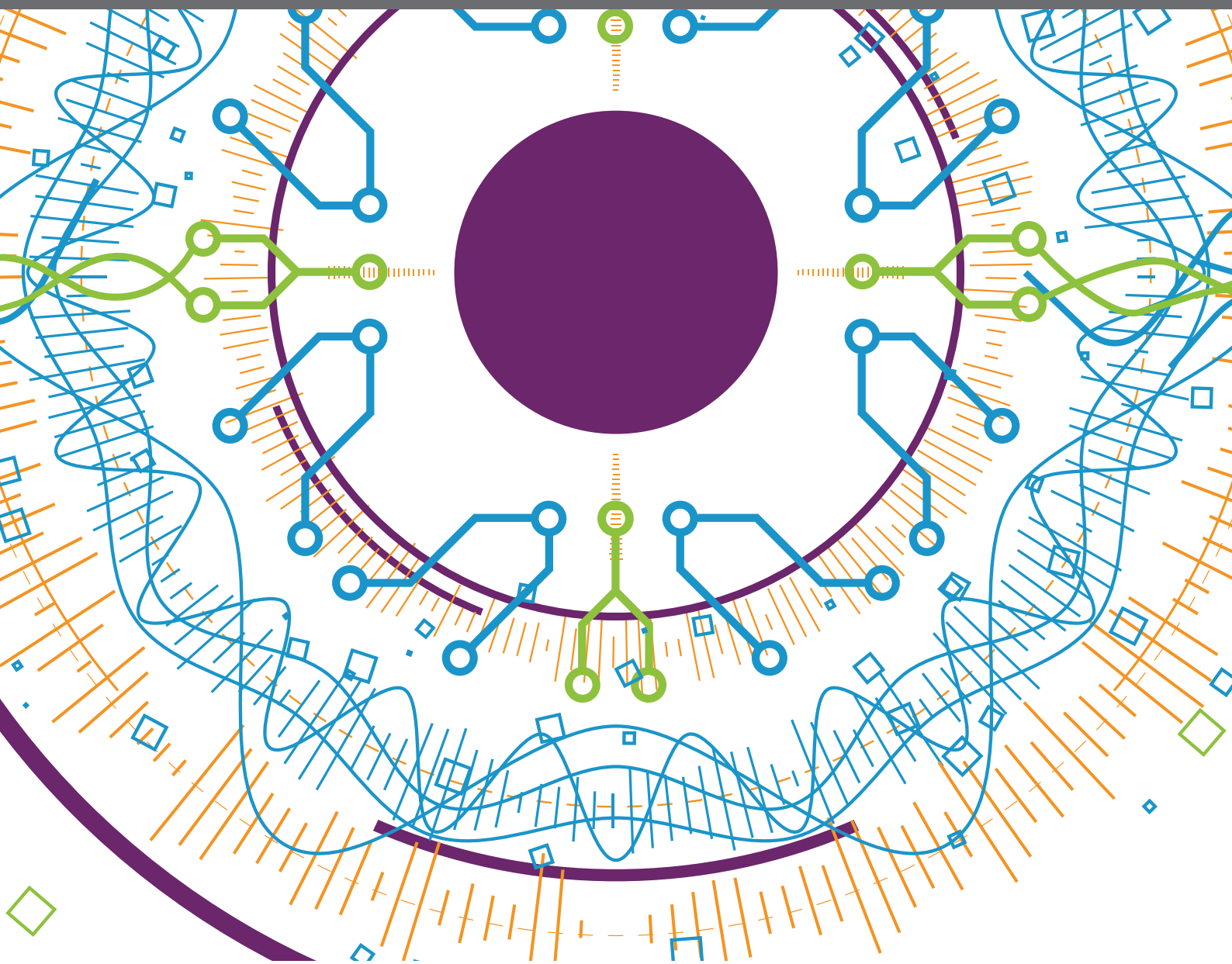
AI-ENABLED DATA SCIENCE FOR COVID-19

EDITED BY: Da Yan, Hong Qin, Hsiang-Yun Wu and Jake Y. Chen
PUBLISHED IN: Frontiers in Artificial Intelligence and Frontiers in Big Data



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88974-050-5
DOI 10.3389/978-2-88974-050-5

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

AI-ENABLED DATA SCIENCE FOR COVID-19

Topic Editors:
Da Yan, University of Alabama at Birmingham, United States
Hong Qin, University of Tennessee at Chattanooga, United States
Hsiang-Yun Wu, Vienna University of Technology, Austria
Jake Y. Chen, University of Alabama at Birmingham, United States

Citation: Yan, D., Qin, H., Wu, H.-Y., Chen, J. Y., eds. (2022). AI-enabled Data Science for COVID-19. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-050-5

Table of Contents

frontiers in Big Data

EDITORIAL
published: 24 November 2021
doi: 10.3389/fdata.2021.802452

Editorial: AI-Enabled Data Science for COVID-19

Da Yan[1]*, *Hong Qin*[2], *Hsiang-Yun Wu*[3] and *Jake Y. Chen*[1,4]

[1]Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, United States, [2]Department of Computer Science and Engineering, University of Tennessee at Chattanooga, Chattanooga, TN, United States, [3]Research Unit of Computer Graphics, Institute of Visual Computing and Human-Centered Technology, TU Wien, Austria, [4]Informatics Institute, School of Medicine, University of Alabama at Birmingham, Birmingham, AL, United States

Keywords: COVID-19, artificial intelligence, AI, pandemic, data mining

Editorial on the Research Topic

AI-Enabled Data Science for COVID-19

COVID-19 is a pandemic that has swept all over the world. As of this writing, the New York Times reported that the United States has over 45.4 million cases and 736,000 deaths, and the worldwide numbers are over 240 million cases and 4.9 million deaths. New variants of SARS-CoV-2 continue to emerge and can be more infectious, as we witnessed new surges of the Delta variant worldwide in 2021. Therefore, fighting against COVID-19 is a public health topic of paramount importance.

Many COVID-19 related datasets have already been collected, and the rapid advancement of AI and Data Science has created new software tools for researchers to characterize epidemiological and biological characteristics of COVID-19. In this Research Topic that started in mid-2020, we have openly solicited and collected eight articles in this research direction. This Research Topic represents recent advances in computational approaches to epidemiological modeling, risk analysis, precision diagnosis, and disease progression of COVID-19.

Two papers studied the spread of COVID-19, and such epidemiological models are useful to help the authority decide the proper preventive measures such as stay-at-home orders, travel restrictions, school closure, mask-wearing mandate, and so forth. Li et al. proposed a time-dependent SEIR model that considers the incubation period to mathematically describe the dynamic of the COVID-19 pandemic. The model takes immunity, reinfection, and vaccination into account and can monitor the trajectories of changing parameters, such as transmission rate, recovery rate, and the basic reproduction number. Potgieter et al. emphasized the use of mobility data in modeling the COVID-19 spread through the population. Different mobility data sources were compared to provide insight on which data provides what type of information and in what situations a particular data source is the most useful.

Some COVID-19 patients may develop severe pneumonia in both lungs. COVID-19 pneumonia is a serious illness that can be deadly, so a lot of works have merged that conduct computer-aided detection of such patients from chest CT or X-ray images, using the deep learning technology for computer vision. Nguyen et al. raised the concern about the generalizability of such models, given the heterogeneous factors in training datasets. Their study examined the severity of this problem by evaluating deep learning classification models trained to identify COVID-19 positive patients on 3D CT datasets from different countries. The study confirmed that such models cannot easily generalize to an entirely new dataset distribution never seen before due to factors including patient demographics and differences in image acquisition or reconstruction; and the best-performing model for a particular dataset tends to be a model trained on multiple datasets.

Four works studied how to train robust models and/or interpretable models with the electronic health records (EHRs) of COVID-19 patients to predict symptoms, mortality, and other risk factors. Such prediction models would help the planning of medical resources to individuals most at-risk

OPEN ACCESS

Edited and reviewed by:
Thomas Hartung,
Johns Hopkins University,
United States

***Correspondence:**
Da Yan
yanda@uab.edu

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Big Data

Received: *26 October 2021*
Accepted: *04 November 2021*
Published: *24 November 2021*

Citation:
Yan D, Qin H, Wu H-Y and Chen JY (2021) Editorial: AI-Enabled Data Science for COVID-19. Front. Big Data 4:802452. doi: 10.3389/fdata.2021.802452

when healthcare services are under high pressure and would help improve the healthcare outcomes of COVID-19 patients in time. To build a cohort-independent robust mortality prediction model, Bai et al. conducted an international, bi-institutional study from China and Germany. A random forest model was applied to 1,352 patients from the Wuhan cohort, which identified five effective clinical features at admission, including lymphocyte, neutrophil count, C-reactive protein, lactate dehydrogenase, and α-hydroxybutyrate dehydrogenase. These features were also found to be robust over time when patients are in the hospital, and the model was found to generalize well on the independent Würzburg cohort. Mamidi et al. developed an interpretable COVID-19 risk calculator for individuals by utilizing de-identified electronic health records (EHR) from UAB-i2b2 COVID-19 repository under the U-BRITE framework. The generated risk scores are analogous to commonly used credit scores where higher scores indicate higher risks for COVID-19 infection. The authors found that within the 2 weeks before a COVID-19 diagnosis, the most predictive features were respiratory symptoms and other chronic conditions; when extending the timeframe to include all medical conditions across all time, their models also uncovered several chronic conditions impacting the respiratory, cardiovascular, central nervous and urinary organ systems. Snider et al. used SHAP (SHapley Additive exPlanations) to study the impacts of various attributes of the COVID-19 patients in an XGBoost model, which was applied to a dataset containing 57,390 individual COVID-19 cases and 2,822 patient deaths in Ontario, Canada. The most important attributes were found to be age, date of the positive test, sex, income, dementia and some others. Jamshidi et al. conducted a comprehensive evaluation of existing machine learning methods, and created two models based solely on the age, gender, and medical histories of 23,749 hospital-confirmed COVID-19 patients from February to September 2020: a symptom prediction model (SPM) and a mortality prediction model (MPM).

Finally, this Research Topic also included a survey paper by Abdulkareem and Petersen, who carefully summarized recent technological tools, artificial intelligence (AI) tools in particular, that have been used in the detection, diagnosis and epidemiological predictions, forecasting and social control for combating COVID-19. The work highlighted areas of successful applications and underscored issues that need to be addressed to achieve significant progress in battling COVID-19 and future pandemics.

AUTHOR CONTRIBUTIONS

DY led the writing of this manuscript. HQ, H-YW, and JC contributed to the writing and review of the manuscript.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 18 March 2021
doi: 10.3389/frai.2021.648579

Toward the Impact of Non-pharmaceutical Interventions and Vaccination on the COVID-19 Pandemic With Time-Dependent SEIR Model

Yuexin Li[1], *Linqiang Ge*[2], *Yang Zhou*[3], *Xuan Cao*[4] *and Jingyi Zheng*[1]*

[1] Department of Mathematics and Statistics, Auburn University, Auburn, AL, United States, [2] TSYS School of Computer Science, Columbus State University, Columbus, GA, United States, [3] Department of Computer Science and Software Engineering, Auburn University, Auburn, AL, United States, [4] Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH, United States

OPEN ACCESS

Edited by:
Hong Qin,
University of Tennessee at Chattanooga, United States

Reviewed by:
Ramaraju Rudraraju,
University of Alabama at Birmingham, United States
Zongliang Yue,
University of Alabama at Birmingham, United States

***Correspondence:**
Jingyi Zheng
jingyi.zheng@auburn.edu

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: 31 December 2020
Accepted: 22 February 2021
Published: 18 March 2021

Citation:
Li Y, Ge L, Zhou Y, Cao X and Zheng J (2021) Toward the Impact of Non-pharmaceutical Interventions and Vaccination on the COVID-19 Pandemic With Time-Dependent SEIR Model. Front. Artif. Intell. 4:648579. doi: 10.3389/frai.2021.648579

The outbreak of COVID-19, caused by the SARS-CoV-2 coronavirus, has been declared a pandemic by the World Health Organization (WHO) in March, 2020 and rapidly spread to over 210 countries and territories around the world. By December 24, there are over 77*M* cumulative confirmed cases with more than 1.72*M* deaths worldwide. To mathematically describe the dynamic of the COVID-19 pandemic, we propose a time-dependent SEIR model considering the incubation period. Furthermore, we take immunity, reinfection, and vaccination into account and propose the SEVIS model. Unlike the classic SIR based models with constant parameters, our dynamic models not only predicts the number of cases, but also monitors the trajectories of changing parameters, such as transmission rate, recovery rate, and the basic reproduction number. Tracking these parameters, we observe the significant decrease in the transmission rate in the U.S. after the authority announced a series of orders aiming to prevent the spread of the virus, such as closing non-essential businesses and lockdown restrictions. Months later, as restrictions being gradually lifted, we notice a new surge of infection emerges as the transmission rates show increasing trends in some states. Using our epidemiology models, people can track, timely monitor, and predict the COVID-19 pandemic with precision. To illustrate and validate our model, we use the national level data (the U.S.) and the state level data (New York and North Dakota), and the resulting relative prediction errors for the infected group and recovered group are mostly lower than 0.5%. We also simulate the long-term development of the pandemic based on our proposed models to explore when the crisis will end under certain conditions.

Keywords: COVID-19, epidemiology, dynamic modeling, reinfection, vaccination, time-dependent SEIR model

1. INTRODUCTION

On March 11, 2020, the World Health Organization (WHO) declared that the outbreak of the novel coronavirus (COVID-19) can be characterized as a pandemic. The COVID-19 outbreak started in Wuhan, China in December, 2019. By the end of January, 2020, the confirmed cases in China went up to 11,791. Only 1 month later, the number increased nearly seven-fold to 80,134 and the COVID-19 cases gradually showed up in other countries. Starting from March, 2020, the outbreak spread to more than 100 countries. By the end of 2020, the pandemic has led to $77.5M$ confirmed cases and more than $1.72M$ fatalities worldwide. **Figure 1** summarizes the percentage of global confirmed cases contributed by each country. As of December 24, the United States, India, and Brazil are the three countries most impacted by the COVID-19 pandemic. The trajectories of the confirmed cases in the three countries are also displayed.

The COVID-19 virus has caused a great disruption to the human health, social life, developments, and economics. To stop the spread of COVID-19 virus, governments have carried out numerous preventive measures such as stay-at-home orders, travel restrictions, school closure, mask-wearing mandate, and so forth. The impact on the society came later in all aspects, including rising unemployment, protests against restrictions, and psychological anxiety and stress brought to the public. However, a significant decrease in the transmission rate occurred, which proved that these mitigation measures were effective. Months later, many states in the U.S. have loosened their restrictions and lifted orders to allow businesses to reopen to the public. Consequently, the diagnoses of daily confirmed cases have displayed a consequential increasing trend after the reopen in some states such as Alabama. By looking at the numbers only, it is difficult to assess what stage we are at in the COVID-19 pandemic and when it is going to end. Hence, mathematical models considering the epidemiological characteristics of COVID-19 become crucial and significant to track and forecast the trend of the spread.

The classic epidemiology model exhibits compelling results, especially during the early period of the pandemic. The compartmental models, which are the simplified versions of mathematical models for infectious diseases, divide the population into different compartments between which people may progress. Different diseases are represented by different compartmental models (Schmidt, 1981; Sharomi and Gumel, 2011; Gao et al., 2016). The Susceptible-Infectious-Recovered (SIR) model, as one of the simplest and most classic compartmental models, characterizes the dynamic changes in each compartment using ordinary differential equations. There are three compartments in this model: susceptible (S), infectious (I), and recovered/deceased (R). The number of individuals in each compartment varies over time. The deterministic SIR and its derivatives are widely used to predict infectious deceases like COVID-19 (Chen et al., 2020; Katul et al., 2020; Toda, 2020). Besides compartmental models, statistical learning techniques are also widely used in biomedical fields (Zheng et al., 2018, 2019; Hsieh and Zheng, 2019; Ganyani et al., 2020; Murray, 2020; You et al., 2020). For example, IHME team (Murray, 2020) employed a statistical model to predict the number of deaths, the demand of hospital beds, ICU beds and ventilators in a few months.

In this paper, we develop a time-dependent Susceptible-Exposed-Infectious-Recovered (SEIR) model with coefficients estimated by Least Absolute Shrinkage and Selection Operator (LASSO) regression. This model is inspired by the SIR model and takes the existence of incubation period (the time from exposure to development of symptoms) into consideration. The individuals who have been infected but are not yet infectious are labeled as exposed (E). Instead of the constant parameters used in traditional SIR based models, we propose to model the dynamic with time-dependent parameters. Additionally, we extend our SEIR model to accommodate other crucial factors such as immunity, reinfection, and vaccination cases into account. With the epidemiology models, we aim at answering the following questions:

- What is the trajectory of transmission rate, incubation rate, and recovery rate?
- Has the inflection point been reached. If so, when?
- How does the reopen order affect the spread of the pandemic?
- How do reinfection and vaccination affect the pandemic?
- When will the mortality reach the peak?
- How many cases do we expect to have when the pandemic is over?

The remainder of the paper is organized as follows: we build the time-dependent SEIR model in section 2. Then we extend the model to include the vaccinated group as well as analyze the asymptotic stability of its disease-free equilibrium in section 3. To validate our model, we perform numerical analysis, prediction, and model simulation using national level data of the United States, and the state level data of two selected states, New York and North Dakota. The results are presented in section 4. Lastly, we conclude this paper in section 5.

2. THE TIME-DEPENDENT SEIR MODEL

Our proposed SEIR model with time-dependent parameters describes the transmission dynamic of an epidemic. It is assumed that there are totally four states in which an individual would experience: susceptible, exposed, infected, and recovered. In the susceptible state, the individual does not have the disease but can be infected by someone infectious through an effective contact. Once being infected, the individual moves to the exposed state. The exposed individual is not able to infect others until the incubation period is over. Eventually, the infected individual recovers from the disease. Altogether the four groups of individuals at different states compose the entire population and we denote the number of individuals in each group at time t by $S(t), E(t), I(t)$, and $R(t)$. In this model, a person is assumed to be immune to the virus after recovery and will not return to the susceptible state. Accordingly, the number of deaths caused by the disease is also counted in the recovered group $R(t)$ since

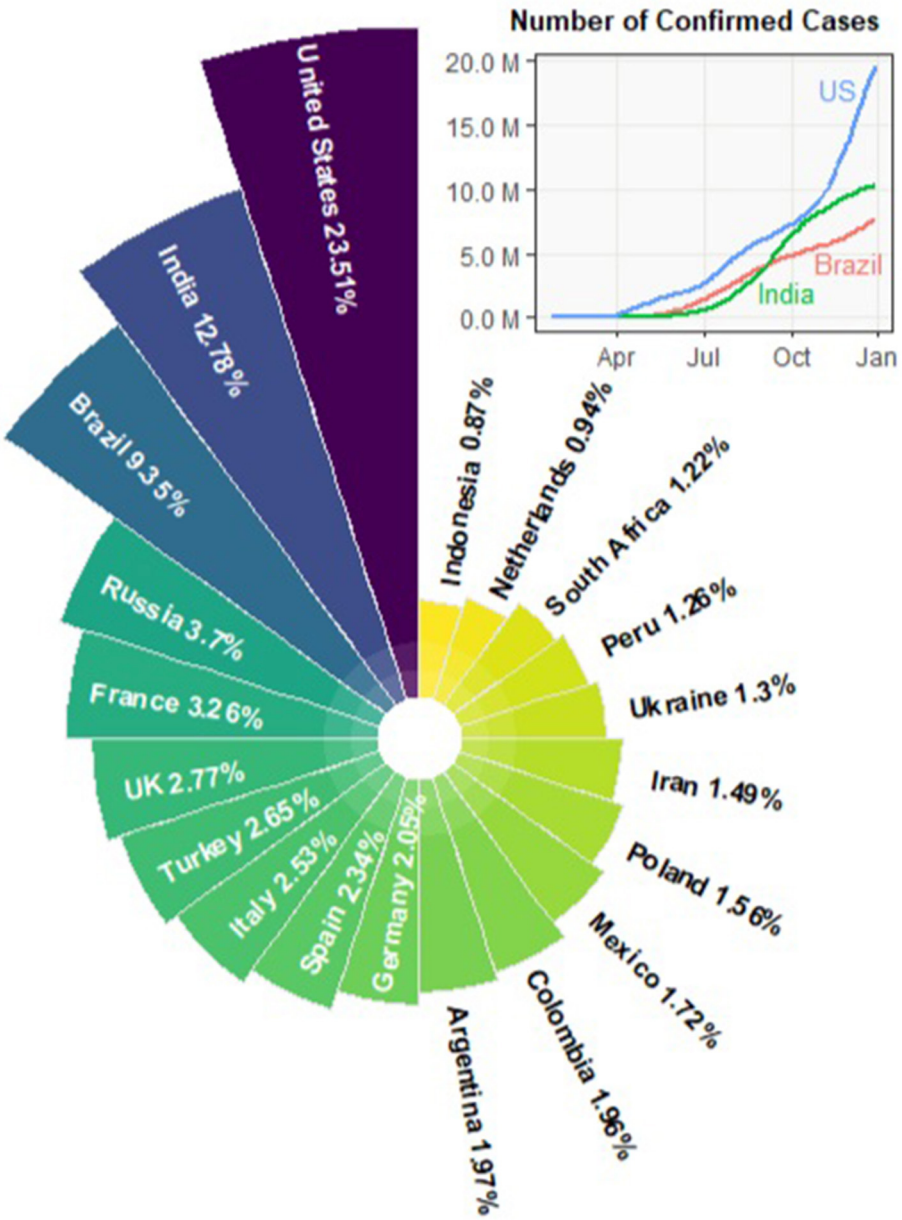


FIGURE 1 | Countries most impacted by COVID-19, updated by 2020-12-24.

neither of the recovered and dead has any more impact on the spread of the virus.

The differential equations that govern the trajectories of the four compartments are formulated as:

$$\frac{dS}{dt} = -\frac{\beta_t S(t) I(t)}{N}, \tag{1}$$

$$\frac{dE}{dt} = \frac{\beta_t S(t) I(t)}{N} - \sigma_t E(t), \tag{2}$$

$$\frac{dI}{dt} = \sigma_t E(t) - \gamma_t I(t), \tag{3}$$

$$\frac{dR}{dt} = \gamma_t I(t), \tag{4}$$

with a constant total population N,

$$N = S(t) + E(t) + I(t) + R(t), \tag{5}$$

and therefore, we have:

$$\frac{dS}{dt} + \frac{dE}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0. \tag{6}$$

Three time-dependent parameters, the transmission rate β_t, the transition rate σ_t, and the recovery rate γ_t are introduced in this model, which are all assumed to vary with respect to time. The descriptions and empirical ranges are listed in **Table 1**.

The proportion of susceptible and infected individuals in the population at time t are $\frac{S(t)}{N}$ and $\frac{I(t)}{N}$, respectively. Given the transmission rate β_t, which describes the flow of susceptible becoming exposed to the virus, and the total population N, the number of newly exposed people is $\frac{\beta_t S(t) I(t)}{N}$. Later, the exposed individuals make the transition to the infected state at the transition rate σ_t, which is the inverse of the incubation period. The number of exposed individuals who complete the transition at time t is $\sigma_t E(t)$. Similarly, people recovered at time t is $\gamma_t I(t)$, given the recovery rate γ_t, which is the number of individuals recover from the infected state per person per time.

2.1. Discrete Time-Dependent SEIR Model

Since the COVID-19 case report is updated daily, we revise the differential Equations (1)–(4) into discrete time difference equations as follows:

$$S(t+1) - S(t) = -\frac{\beta_t S(t) I(t)}{N}, \tag{7}$$

$$E(t+1) - E(t) = \frac{\beta_t S(t) I(t)}{N} - \sigma_t E(t), \tag{8}$$

$$I(t+1) - I(t) = \sigma_t E(t) - \gamma_t I(t), \tag{9}$$

$$R(t+1) - R(t) = \gamma_t I(t), \tag{10}$$

with the four variables satisfying (5) and

$$\begin{aligned} &S(t+1) - S(t) + E(t+1) - E(t) + \\ &I(t+1) - I(t) + R(t+1) - R(t) = 0. \end{aligned} \tag{11}$$

Assuming historical data for a certain time period $0 \le t \le T$ is available, i.e., we have $\{S(t), E(t), I(t), R(t) | 0 \le t \le T\}$. By deduction from (7) to (10), we can compute historical values of the parameter series $\{\beta_t, \sigma_t, \gamma_t | 0 \le t \le T-1\}$ using the following formulas:

$$\beta_t = \frac{N(E(t+1) - E(t) + I(t+1) - I(t) + R(t+1) - R(t))}{S(t) I(t)}, \tag{12}$$

$$\sigma_t = \frac{I(t+1) - I(t) + R(t+1) - R(t)}{E(t)}, \tag{13}$$

$$\gamma_t = \frac{R(t+1) - R(t)}{I(t)}. \tag{14}$$

Now predicting future values of the parameters $\{\beta_t, \sigma_t, \gamma_t | t \ge T\}$ given historical values can be converted to a regression problem.

2.2. Tracking the Transmission Rate β_t, Transition Rate σ_t, and Recovery Rate γ_t

There are several approaches predicting future values of the time-dependent parameters. For instance, we can use linear models (e.g., linear regression), nonlinear methods (e.g., spline), or time series models (e.g., autoregressive model), etc. In this subsection, we fit the following LASSO regression models:

$$\hat{\beta}_{t+1} = a_0 + \sum_{i=1}^{I} a_i \beta_{t-i}, \tag{15}$$

$$\hat{\sigma}_{t+1} = b_0 + \sum_{j=1}^{J} a_j \sigma_{t-j}, \tag{16}$$

$$\hat{\gamma}_{t+1} = c_0 + \sum_{k=1}^{K} a_k \gamma_{t-k}, \tag{17}$$

where I, J, and K are the orders of the autoregressive process, and $\{a_i | 0 \le i \le I\}$, $\{b_j | 0 \le j \le J\}$ and $\{c_k | 0 \le k \le K\}$ are the regression coefficients.

These coefficients are determined by minimizing the following loss functions, which are composed of the residual sums of squares (RSS) and regularization terms:

$$L(\beta) = \sum_{t=I+1}^{T-1} (\beta_t - a_0 - \sum_{i=1}^{I} a_i \beta_{t-i})^2 + \lambda_\beta \sum_{i=0}^{I} |a_i^2|, \tag{18}$$

$$L(\sigma) = \sum_{t=J+1}^{T-1} (\sigma_t - b_0 - \sum_{j=1}^{J} b_j \sigma_{t-j})^2 + \lambda_\sigma \sum_{j=0}^{J} |b_j^2|, \tag{19}$$

$$L(\gamma) = \sum_{t=K+1}^{T-1} (\gamma_t - c_0 - \sum_{k=1}^{K} c_k \gamma_{t-k})^2 + \lambda_\gamma \sum_{k=0}^{K} |c_k^2|, \tag{20}$$

λ_β, λ_σ, and λ_γ are the regularization parameters deciding the penalty to the flexibility of model, and all regularization parameters can be optimized by cross-validation.

2.3. Estimating the Exposed $\hat{E}(t)$, Infections $\hat{I}(t)$, and Recovered $\hat{R}(t)$ Groups

Given the historical data $\{S(t), E(t), I(t), R(t), 0 \le t \le T\}$, we first compute the time-dependent parameter series $\{\beta_t, \sigma_t, \gamma_t, 0 \le t \le T-1\}$ introduced in section 2.1. Then we predict future values $\{\hat{\beta}_t, \hat{\sigma}_t, \hat{\gamma}_t, t \ge T\}$ using the model built in section 2.2. According to (8), (9), (10), and (5), we can further predict the number of cases for the future as follows:

$$\hat{E}(t+1) = \hat{E}(t) + \frac{\hat{\beta}_t \hat{S}(t) \hat{I}(t)}{N} - \hat{\sigma}_t \hat{E}(t), \quad t \ge T+1, \tag{21}$$

TABLE 1 | Model parameters.

Parameter	Description	Empirical range	References
β_t	Transmission rate (effective contact rate) at a given time	0.5–1.5 day^{-1}	Ngonghala et al., 2020; Read et al., 2020; Shen et al., 2020
σ_t	Transition rate from exposed to infections at a given time	$\frac{1}{5.1}$	Fairoza Amira et al., 2020; Ngonghala et al., 2020
γ_t	Recovery rate at a given time	$\frac{1}{10}$	Fairoza Amira et al., 2020; Ngonghala et al., 2020
v_t	Fraction of susceptible individuals vaccinated at a given time		
w	Fraction of infections gain immunity after recovery		

$$\hat{I}(t+1) = \hat{I}(t) + \hat{\sigma}_t\hat{E}(t) - \hat{\gamma}_t\hat{I}(t), \quad t \geq T+1, \tag{22}$$

$$\hat{R}(t+1) = \hat{R}(t) + \hat{\gamma}_t\hat{I}(t), \quad t \geq T+1, \tag{23}$$

$$\hat{S}(t+1) = N - \hat{E}(t+1) - \hat{I}(t+1) - \hat{R}(t+1), \quad t \geq T+1, \tag{24}$$

Note that for the special case when estimating $\{\hat{S}(t), \hat{E}(t), \hat{I}(t), \hat{R}(t)|t = T+1\}$, i.e., the numbers of cases at $t = T+1$, we use the true values of $\{S(t), E(t), I(t), R(t)|t = T\}$ instead of using the estimated values $\{\hat{S}(t), \hat{E}(t), \hat{I}(t), \hat{R}(t)|t = T\}$ as in the formulas (21), (22), (23), and (24). The detailed steps of the entire procedure are summarized in Algorithm 1.

Algorithm 1: Tracking discrete time time-dependent SEIR model

Input: $\{E(t), I(t), R(t)|0 \leq t \leq T\}$, regularization parameters λ_β, λ_σ and λ_γ, orders of autoregressive process I, J, K, prediction window t_w.

Output: $\{\beta_t, \sigma_t, \gamma_t|0 \leq t \leq T-1\}$, $\{\hat{\beta}_t, \hat{\sigma}_t, \hat{\gamma}_t|T \leq t \leq T+t_w-1\}$, $\{\hat{E}(t), \hat{I}(t), \hat{R}(t)|T+1 \leq t \leq T+t_w\}$.

Compute $\{\beta_t, \sigma_t, \gamma_t|0 \leq t \leq T-1\}$ using (12), (13), and (14);

Train the LASSO regression models using $\{\beta_t, \sigma_t, \gamma_t|0 \leq t \leq T-2\}$ as the predictors and $\{\beta_{t_1}, \sigma_{t_2}, \gamma_{t_3}|I+1 \leq t_1 \leq T-1, J+1 \leq t_2 \leq T-1, K+1 \leq t_3 \leq T-1\}$ as the response;

while $T \leq t \leq T+t_w-1$ **do**

  Predict $\hat{\beta}_t$, $\hat{\sigma}_t$ and $\hat{\gamma}_t$ using (15), (16), and (17);

  Estimate $\hat{E}(t+1)$, $\hat{I}(t+1)$ and $\hat{R}(t+1)$ using (21), (22), and (23), respectively;

3. SEIR VARIATION CONSIDERING IMMUNITY, REINFECTION, AND VACCINATION

The human immune system protects the body against diseases with two parts. The first part, known as the innate immune response, includes the release of chemicals that cause inflammation and white blood cells that can destroy infected cells. It is always ready to take actions as soon as any foreign invader is detected inside the body. However, this part is not specific to coronavirus. It will not learn and develop immunity to the virus. Instead, the second part: the adaptive immune response produces targeted antibodies that can stick to the virus and stop the spread to the body. The T cells[1] would attack the cells infected by the virus.

Existing research shows that most COVID-19 patients had an antibody response at 10 days or later after onset of symptoms (To et al., 2020). If the adaptive immune response is powerful enough, it could leave a lasting memory of the infection that will provide protection in the future. Other findings also suggest that strong responders (with higher antibody level) are significantly higher in severe patients, while it is unclear whether the asymptomatic or mildly symptomatic patients will develop sufficient adaptive immune response and gain immunity to the disease after recovery (Tan et al., 2020). In fact, there have been several reported cases of COVID reinfection in China, Hong kong, Belgium, the Netherlands, and the U.S. (Tan et al., 2020), and the reinfection case are indeed increasing. This implies the necessity of taking reinfection into consideration.

On the other hand, the worldwide endeavor to create a safe and effective COVID-19 vaccine is beginning to bear fruit. A wide variety of vaccines has already been authorized around the globe while many more remain in development. According to the U.S. CDC, as of December 13, 2020, the Pfizer-BioNTech COVID-19 vaccine has been authorized and large-scale (Phase 3) clinical trials are in progress or being planned for three other vaccines in the United States. Currently the supply of COVID-19 vaccine in the U.S. is limited, but it will increase in the upcoming weeks and months. Once large quantities are available, the increasingly large-scale vaccination will have a substantial impact on the pandemic.

3.1. The Time-Dependent SEVIS Model

To take the factors of immunity, reinfection, and vaccination into account, we modify the proposed SEIR model by removing the recovered group $R(t)$ and adding a vaccinated group $V(t)$, which represents the vaccinated individuals. In this susceptible, exposed, vaccinated, and infected modeling framework, the previous assumption for the SEIR model that an infected

[1]T cells are one of the important white blood cells of the immune system, and play a central role in the adaptive immune response.

individual will not become susceptible again after recovery is no longer employed. Instead, we assume that a fraction of the infected individuals gain immunity after recovery through producing antibodies while the rest return to the susceptible state. The former is counted in the $V(t)$ group along with the vaccinated individuals since, epidemiologically speaking, both are immune to the virus and can no longer be infected. The new SEVIS model is governed by the following differential equations:

$$\frac{dS}{dt} = -\frac{\beta_t S(t)I(t)}{N} - v_t S + (1-w)\gamma_t I(t), \tag{25}$$

$$\frac{dE}{dt} = \frac{\beta_t S(t)I(t)}{N} - \sigma_t E(t), \tag{26}$$

$$\frac{dV}{dt} = v_t S + w\gamma_t I(t), \tag{27}$$

$$\frac{dI}{dt} = \sigma_t E(t) - \gamma_t I(t), \tag{28}$$

with a constant total population N,

$$N = S(t) + E(t) + V(t) + I(t), \tag{29}$$

and therefore, we have:

$$\frac{dS}{dt} + \frac{dE}{dt} + \frac{dV}{dt} + \frac{dI}{dt} = 0. \tag{30}$$

The parameter settings of the transmission rate β_t, the transition rate σ_t, and the recovery rate γ_t remain the same as in the SEIR model. The vaccination rate v_t is low at the beginning of vaccine administration and gradually increasing as supply is growing. $w \in [0, 1]$ is the fraction of infected cases that become immune after recovery. In addition, we assume it to be constant in this model. Hence, the number of infected individuals recover at time t is $\gamma_t I(t)$, and $w\gamma_t I(t)$ join the $V(t)$ group while $(1-w)\gamma_t I(t)$ fail to gain immunity and return to the susceptible state $S(t)$.

3.2. Baseline Epidemiological Parameters

In previous studies, the transmission rate, β (as a constant), ranges from around 0.5 to 1.5 per person per day (Ngonghala et al., 2020; Read et al., 2020; Shen et al., 2020) and decreases as time goes. Based on existing literature, the incubation period (the time from exposure to development of symptoms) of COVID-19 and other coronaviruses ranges from 2 to 14 days. On average, symptoms show up in the newly exposed person about 5.1 days after contact (Fairoza Amira et al., 2020; Ngonghala et al., 2020). Thus, the transition rate, which is the inverse of the incubation period, is estimated to be $\frac{1}{5.1}$.

3.3. Basic Reproduction Number and Asymptotic Stability of Disease-Free Equilibrium

In this subsection we give the closed-form expression for the time-dependent basic reproduction number of the SEVIS model using the next generation operator method (Diekmann et al., 1990; van den Driessche and Watmough, 2002). The basic reproduction number $\mathcal{R}_0$ is defined as the average number of secondary infections caused by a single infectious individual who enters an entirely susceptible population. That actually is the special case where all parameters and compartments are at their initial state at time $t = 0$. Since we propose the parameters to be time-dependent in our model, we revise the basic reproduction number to a time-dependent version $\mathcal{R}_t$ as well. When $\mathcal{R}_t > 1$, the infection will be able to start spreading in the population and develop into an epidemic. Generally speaking, it is more difficult to control the epidemic with the larger the value of the basic reproduction number.

Let X be the vector of infected classes and Y be the vector of uninfected classes. For the SEVIS model (25)–(28), we have:

$$X = \begin{bmatrix} E \\ I \end{bmatrix}, Y = \begin{bmatrix} S \\ V \end{bmatrix}.$$

Next we define the matrix of new infection terms $\mathcal{F}$, which only includes the flow from X to Y, and matrix of all other terms $\mathcal{V}$, which includes flows within X and flows leaving the system. For each compartment, in-flow in $\mathcal{V}$ is negative and out-flow in $\mathcal{V}$ is positive.

$$\mathcal{F} = \begin{bmatrix} \frac{\beta_t SI}{N} \\ 0 \end{bmatrix}, \mathcal{V} = \begin{bmatrix} \sigma_t E \\ -\sigma_t E + \gamma_t I \end{bmatrix}.$$

The next generation matrix is defined as FV^{-1} where:

$$F = \left.\frac{\partial \mathcal{F}}{\partial X}\right|_{DFE}, V = \left.\frac{\partial \mathcal{V}}{\partial X}\right|_{DFE}.$$

The disease-free equilibrium (DFE) of the SEVIS model is given by: $(S^*, E^*, V^*, I^*) = (N, 0, 0, 0)$, and we have

$$F = \begin{bmatrix} 0 & \beta_t \\ 0 & 0 \end{bmatrix}, V = \begin{bmatrix} \sigma_t & 0 \\ -\sigma_t & \gamma_t \end{bmatrix}.$$

Therefore, the next generation matrix is:

$$FV^{-1} = \begin{bmatrix} \beta_t/\gamma_t & \beta_t/\gamma_t \\ 0 & 0 \end{bmatrix}.$$

$\mathcal{R}_t$, the basic reproduction number at time t, is given by the dominant eigenvalue of FV^{-1}:

$$\mathcal{R}_t = \frac{1}{2}\left(\frac{\beta_t}{\gamma_t} + \sqrt{\frac{\beta_t}{\gamma_t}\left(\frac{\beta_t}{\gamma_t} + 4\right)}\right). \tag{31}$$

Similarly, we can obtain the same basic reproduction number for the time-dependent SEIR model. The DFE is locally asymptotically stable if $\mathcal{R}_t < 1$, and unstable if $\mathcal{R}_t > 1$.

4. NUMERICAL RESULTS, PREDICTIONS, AND SIMULATIONS

In this section, we will give the numeric results obtained by implementing Algorithm 1 on the national level data of the

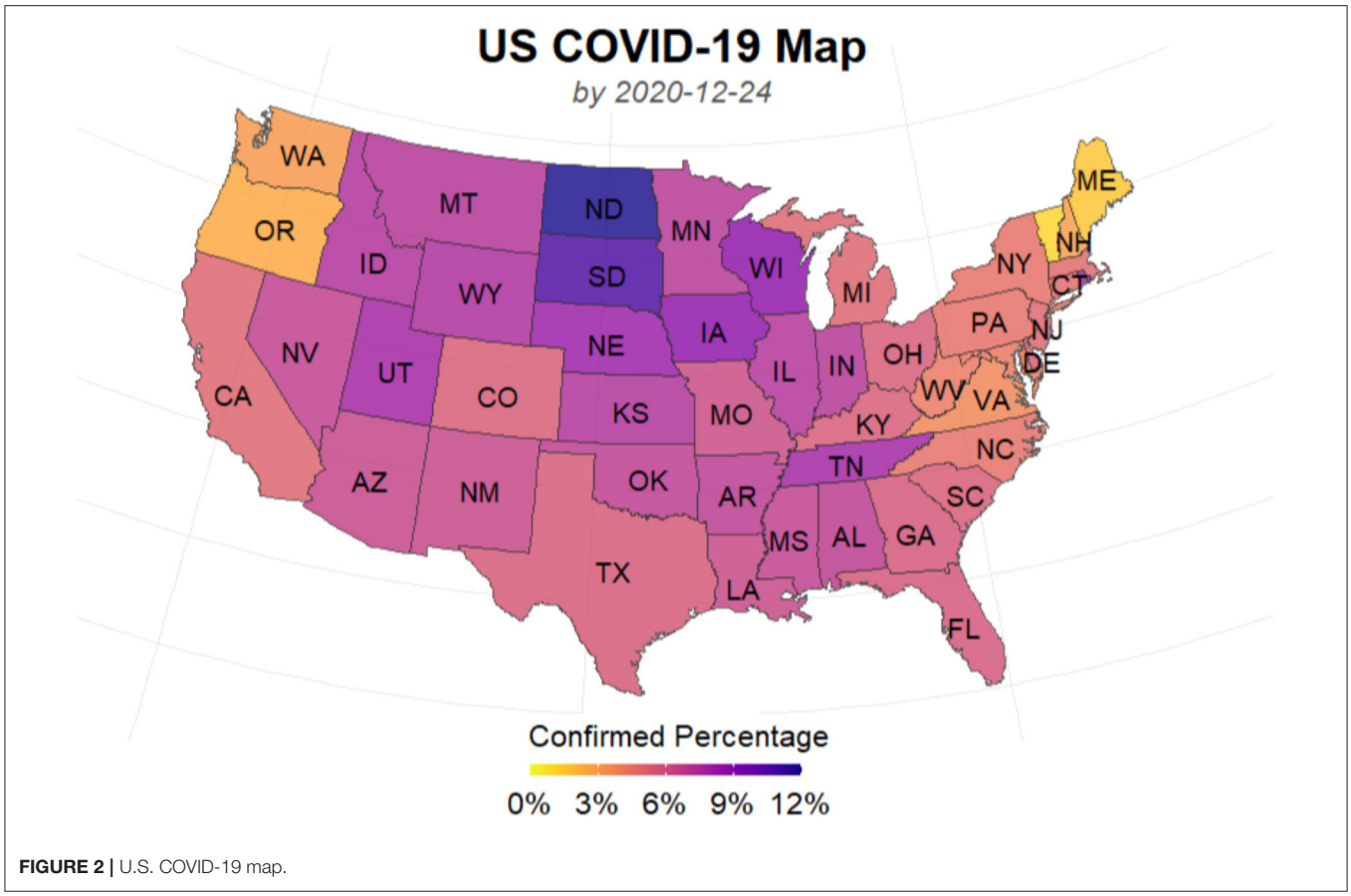


FIGURE 2 | U.S. COVID-19 map.

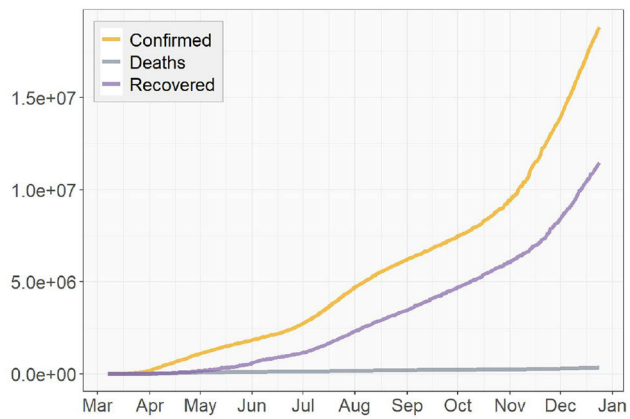


FIGURE 3 | Cumulative numbers of COVID-19 confirmed cases, recoveries, and deaths in the United States.

United States (US) as well as the state level data of a few representative states.

In spring 2020, the New York Metropolitan Area experienced the largest COVID-19 outbreaks. As thousands of cases were being confirmed daily in New York, the state was the epicenter of the nation's crisis and on a different scale than the rest of the country. Though some new batches of hotspots have emerged across the country during the past months, the state of New York (NY) is still a region worth studying. On the other hand, as of December 24, a pack of northern states close to the Canada-US border have the highest percentages of cumulative confirmed cases in their populations as shown in **Figure 2**. The top one, North Dakota, has 11.94% of its population infected cumulatively, followed by South Dakota (10.69%), Wisconsin (8.61%), and some other nearby states. In this case, as a representative of this particular area, we take North Dakota (ND) as another example to illustrate our algorithm. We used the dataset that was collected from the COVID-19 data repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (Dong et al., 2020) and the **nCov2019** R package (Wu et al., 2020). The dataset contains time series of the numbers of confirmed cases, recovered cases and deaths up to December 24, 2020. The starting date of the training set used for model training varies according to the actual spread of the pandemic in each of the three regions: US, NY, and ND. For each region, a different start date of training set is chosen for model fitting according to the time when a relatively clear trend emerges.

Figures 3, **4** presents the cumulative numbers of COVID-19 confirmed cases, recoveries and deaths reported in US, NY, and ND. The data starts at the beginning of the pandemic for US and ND, but it starts a while after the initial point for NY. The

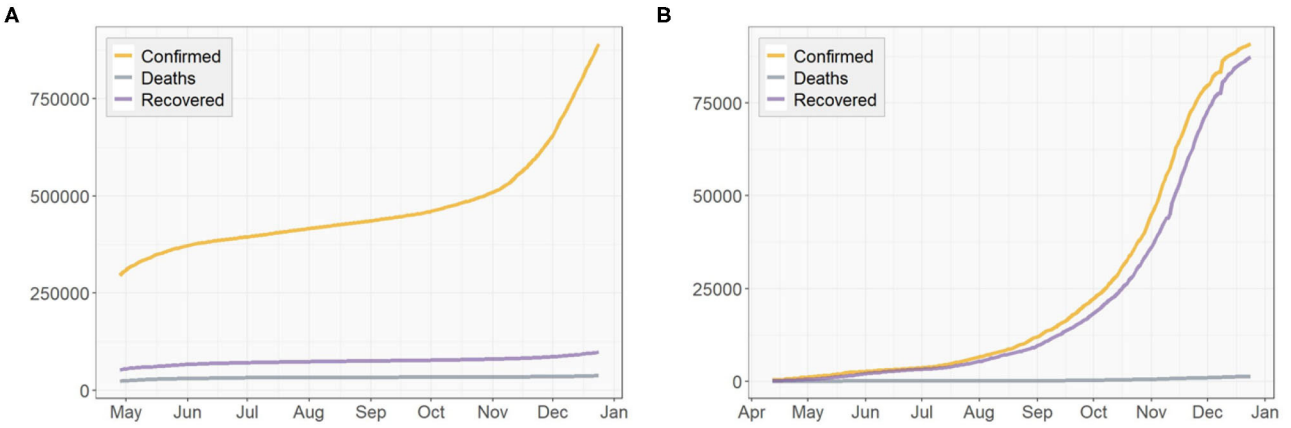


FIGURE 4 | Cumulative numbers of COVID-19 confirmed cases, recoveries, and deaths in **(A)** New York, **(B)** North Dakota.

reason is that, back when the pandemic first started, a series of well-recorded numbers of recoveries were not available for many states, including NY. To obtain complete data on the three type of cases for computation, a cut-off is made. Therefore, the starting point of the data we collected for NY is about 2 months later than the actual date when the first case of COVID-19 was confirmed.

Due to the unavailability of the numbers of the exposed individuals $E(t)$ in any of these regions, we substitute our model in section 2.1 with a simplified version as in Chen et al. (2020) that only includes the other three compartments $S(t)$, $I(t)$, and $R(t)$. To validate our algorithm, we compare the prediction results with known data to see how well it performs, or how large the prediction errors are. Then we implement the algorithm again to predict how the COVID-19 pandemic will spread in the future.

At the end of this section, we simulate the long-term development of the pandemic based on the epidemiology models proposed in sections 2, 3 by constructing certain conditions and assigning assumed values to the parameters listed in **Table 1**. Based on the results, we discuss what they indicate as well as what differences we expect to see in reality compared to the simulation.

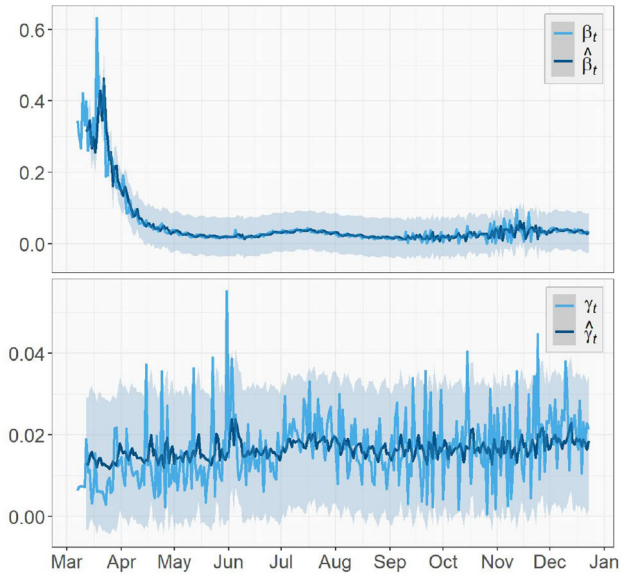


FIGURE 5 | Parameter tracking and prediction for the United States.

4.1. Parameter Tracking and Prediction

First we compute the true values of the transmission rate β_t and the recovery rate γ_t using (12), (13), and (14). Then starting from the sixth day in the parameter series, we take the value of a time-dependent parameter for each day as a subject for testing and a 5-day window before it as a corresponding observation used for training, i.e., $I, K = 5$ in section 2.2. By doing this, we construct the training and testing sets for model fitting. The R package **glmnet** is used to fit the LASSO regression models and choose the optimal values of λ_β and λ_γ that yield the minimum mean cross-validated errors.

Figures 5, **6** depict the true values $\{\beta_t, \gamma_t | 0 \leq t \leq T-1\}$ and predicted values $\{\hat{\beta}_{t_1}, \hat{\gamma}_{t_2} | I+1 \leq t_1 \leq T-1, K+1 \leq t_2 \leq T-1\}$ of both the transmission rate and the recovery rate of US, NY, and ND, respectively. The 95% prediction intervals are shown as the gray bands above and below the curves.

For the U.S. case, there was a sharp decrease in the transmission rate from mid-March to May, just about 1 week after the spread of the virus started. This was an evidence that the social distancing measures and community lockdowns implemented across the country have effectively and significantly slowed down the spread of the pandemic. It kept decreasing for about a month before a surge appeared in July, which is possibly caused by the nationwide celebration of Independence Day. In the fall, starting from early September, the transmission rate slowly rose again with increasingly larger oscillations, which showed consistency with the surge in the fall that

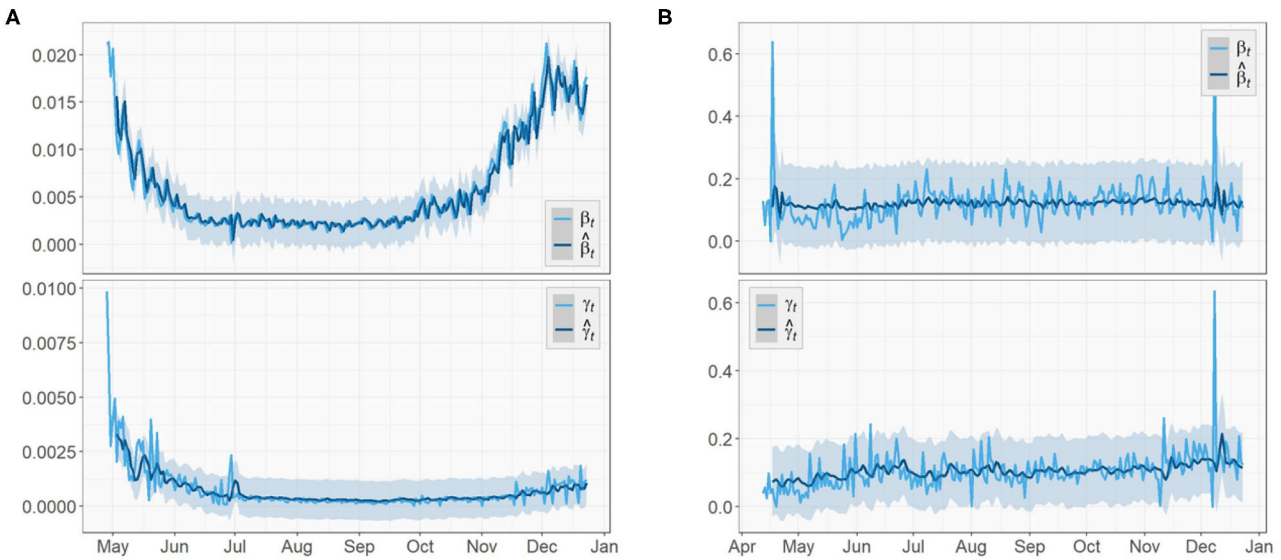


FIGURE 6 | Parameters tracking and prediction for **(A)** New York **(B)** North Dakota.

pushed the total number of confirmed cases in US past 11*M*. This could be a result of a series of events prior to that (e.g., school opening, Halloween), and a prelude to the upcoming large gathering (e.g., Thanksgiving, Black Friday, Christmas). We expect this increase in the transmission rate to continue toward early 2021 and start to gradually decrease after the vaccination is administrated at a large scale in U.S. The recovery rate also had an slight increase around the same time in July but not as large as the one in the transmission rate. Overall, the recovery rate of U.S. is relatively steady and does not show any significant increasing or decreasing trend.

Similar to the US case, the transmission rate of NY started high and then reduced rapidly in the next few weeks. The trend maintained stationary for about 3 months until a rise appeared in late September and kept increasing toward the end. By December, the transmission rate is nearly as high as when it first started. The recovery rate of NY also had a large initial value followed by a 2-month-long decrease, but no clear trend was shown after a small spike at the beginning of July.

As for the ND case, the recovery rate started with a mild increase in the first 2 month. Later on, it remained steady just like the previous two regions. For the transmission rate, the overall trend is much more stationary compared to the results of US and NY and no significant change could be observed. However, the true values of the two parameters of ND have the greatest oscillations, i.e., the largest ranges of oscillations, among the three regions. Note the two unusually acute spikes in the transmission rate respectively in May and December and one in the recovery rate in December that deviate from the entire curves. In the absence of any pre or post trend, we consider these points as outliers in this paper and exclude them in model training.

4.2. Algorithm Validation and Relative Percentage Errors

In this section, we use the computed values of the parameters to estimate the three variables $S(t)$, $I(t)$, and $R(t)$ as in section 2.3. Instead of directly predicting future values for $t > T$, we use the historical data $\{I(t), R(t)|T - t_w \leq t \leq T - 1\}$ and the predicted parameter series $\{\hat{\beta}_t, \hat{\gamma}_t|T - t_w \leq t \leq T - 1\}$ to estimate the last t_w days of the entire period of time by which the data is covered, i.e., predict $\{\hat{I}(t), \hat{R}(t)|T - t_w + 1 \leq t \leq T\}$. Moreover, we also compare the proposed model with the classic SIR model with constant parameters by replacing the time-dependent parameter series with their means.

We evaluate the model performance using the relative percentage errors (*RPE*) of the prediction for the infected group $I(t)$ and the recovered group $R(t)$ as follows:

$$RPE_I = \frac{|I(t) - \hat{I}(t)|}{I(t)}, \quad T - t_w + 1 \leq t \leq T, \tag{32}$$

$$RPE_R = \frac{|R(t) - \hat{R}(t)|}{R(t)}, \quad T - t_w + 1 \leq t \leq T. \tag{33}$$

To assess the predictions of the proposed method and compare with the classic SIR model, we compute the *RPE* series for the past week (i.e., $t_w = 7$) for the two models. The *RPE* series for US, NY, and ND are displayed in **Figures 7**, **8** respectively, with their means summarized in the top-left corner of each figure. Using the proposed model with time-dependent parameters, the mean relative percentage errors for $I(t)$ and $R(t)$, i.e., RPE_I and RPE_R, are 2.35 and 0.39% for US, 0.2 and 0.2% for NY, and 4.67 and 0.09% for ND, respectively. Using the classic SIR model with

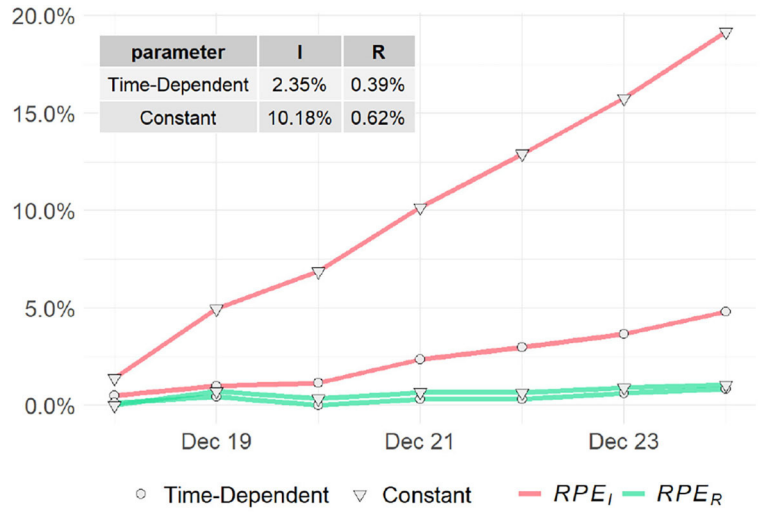


FIGURE 7 | Relative prediction errors for the United States.

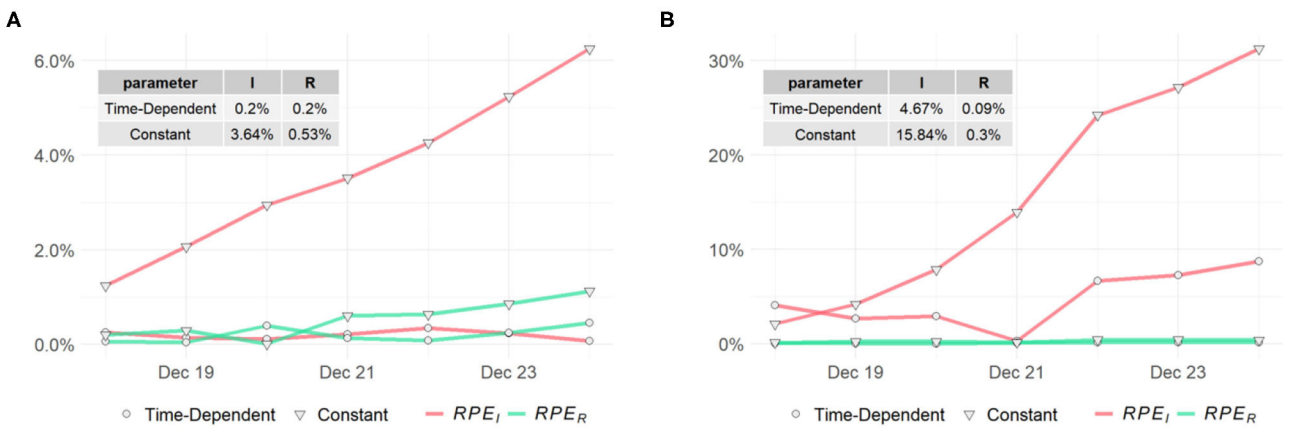


FIGURE 8 | Relative prediction errors for **(A)** New York **(B)** North Dakota.

constant parameters, RPE_I and RPE_R are 10.18 and 0.62% for US, 3.64 and 0.53% for NY, and 15.84 and 0.3% for ND, respectively. All errors are significantly larger than the former, which clearly shows the proposed time-dependent model yields better results in predicting the spread of the pandemic than the traditional SIR model with fixed parameters. Details of the model training and validation process are summarized in **Table 2**.

4.3. One-Day Prediction for *I*(*t*), *R*(*t*), and Basic Reproduction Numbers

Next we implement Algorithm 1 to predict the number of infected $I(t)$ and recovered individuals $R(t)$ for the future $\{\hat{I}(t), \hat{R}(t) | T+1 \leq t \leq T+t_w\}$. We reset the prediction window t_w to be 30, as we are to predict the spread of COVID-19 pandemic in the next 30 days after December 24, 2020. The results of 1-day prediction for US, NY, and ND are shown in **Figures 9**, **10**, respectively. For NY, the sharp increase in the infected group since November is predicted to continue toward the next year, due to the oscillatory rise in the transmission rate shown in **Figure 6**. On the other hand, the growth of the recovered group remains slow. For ND, the number of infected will stay low after the small surge was contained in November, while the rapid growth in the recovered group is expected to be continuous but might slow down. For US, the prediction shows that both curves will keep climbing at a high rate, which indicates that there will still be a long way to go before the pandemic finally ends. The prediction results are summarized in **Table 3**.

To assess the spread of COVID-19, we also obtain the 1-day prediction for the time-dependent basic reproduction number $\mathcal{R}_t$ using (31). The results for the three regions

TABLE 2 | Modeling training and validation.

Region	Start date of training data	Size of training set	Order	Prediction window t_w	Mean RPE_I (%)	Mean RPE_R (%)
United States	2020 – 03 – 07	287	5	7	2.35	0.39
New York	2020 – 04 – 28	235	5	7	0.2	0.2
North Dakota	2020 – 04 – 12	251	5	7	4.67	0.09

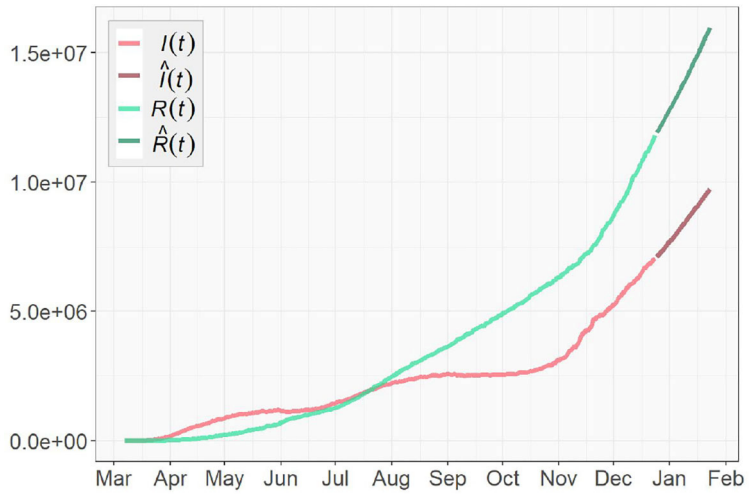


FIGURE 9 | One-day prediction of 30 days for the United States.

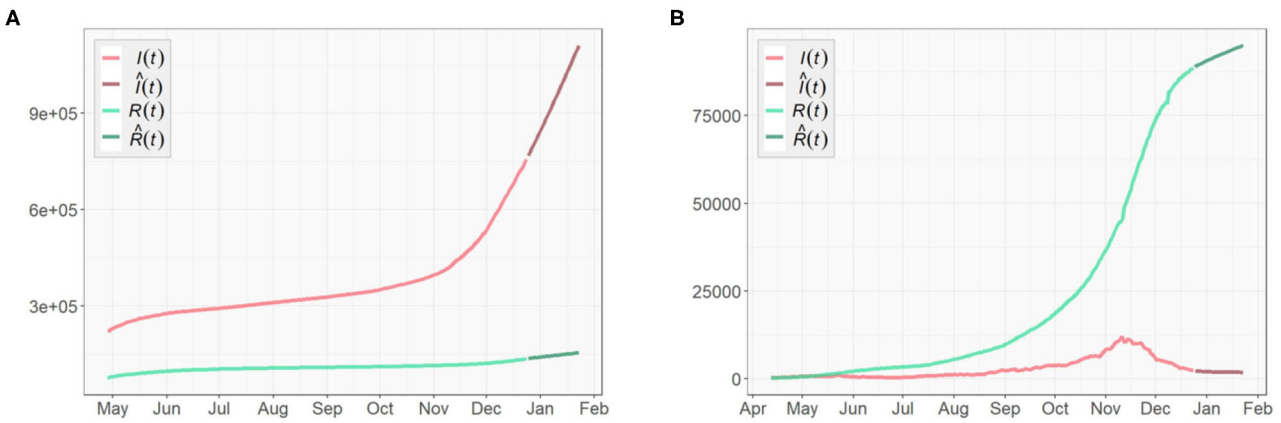


FIGURE 10 | One-day prediction of 30 days for **(A)** New York **(B)** North Dakota.

are presented in **Figures 11**, **12**, with horizontal lines representing $\mathcal{R}_t = 1$. As discussed in section 3.3, the virus will decline and gradually die out when $\mathcal{R}_t < 1$. Otherwise, it will continue to spread. According to the results shown in **Figures 11**, **12**, only very few points fall below the horizontal line, while the majority lies above it. For NY, the surge in fall, 2020 and some scattered large values agree with the increasing trends in both the confirmed cases and the transmission rate we see in **Figures 4**, **6**, respectively.

The basic reproduction numbers $\mathcal{R}_t$ for each of the next 30 days are estimated to be >1 for all three regions. The means of predicted values are found to be 2.48 for US, 22.28 for NY and 1.68 for ND, which suggests the inflection point, where $\mathcal{R}_t$ stabilizes below 1 afterwards, has not been reached yet, especially for the NY case, where instead of having a decreasing trend, an increasing $\mathcal{R}_t$ actually emerges over time. For US and ND, the curves gradually approaching the horizontal lines of $\mathcal{R}_t < 1$ indicates that the measures taken to tackle the pandemic are taking effect, but at this point it is sill too early to relax them.

TABLE 3 | Prediction results.

Region	Total confirmed cases on last day	Prediction window t_w	$\hat{I}(t)$	$\hat{R}(t)$	Predicted total confirmed cases
United States	18,829,816	30	9,723,682	15,971,038	25,694,720
New York	891,270	30	1,111,117	153,277	1,264,394
North Dakota	90,947	30	1,760	95,016	96,776

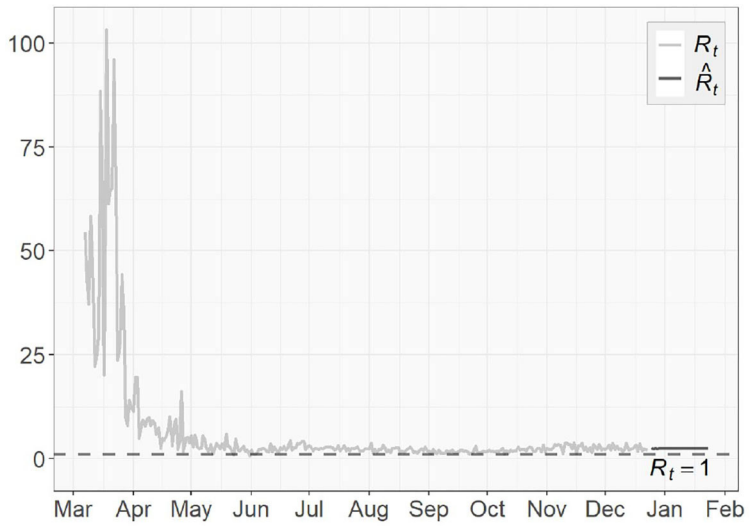


FIGURE 11 | Time-dependent basic reproduction number for the United States.

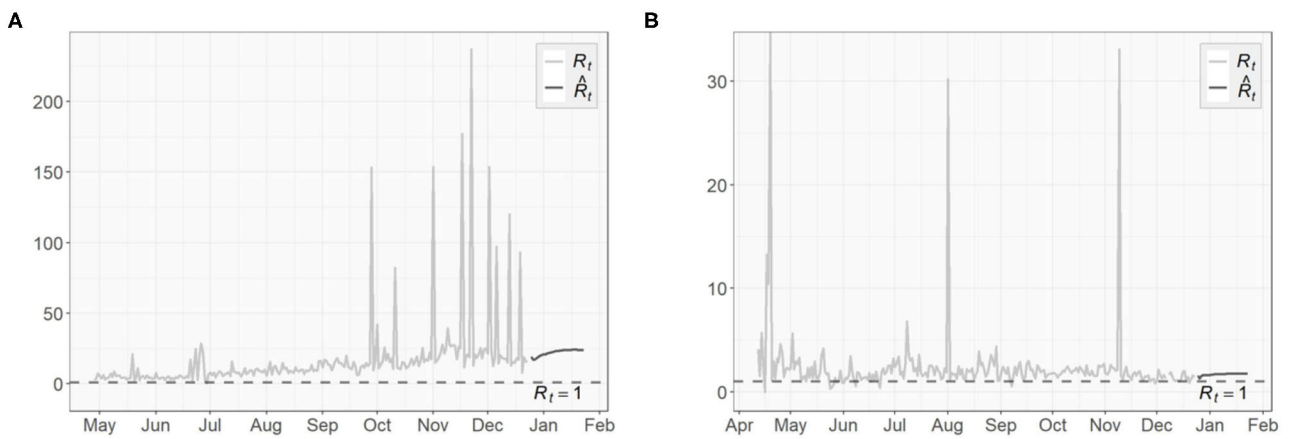


FIGURE 12 | Time-dependent basic reproduction number for **(A)** New York **(B)** North Dakota.

4.4. Simulation Results for the SEIR and SEVIS Models

We also simulate the long-term development of the COVID-19 pandemic based on SEIR and SEVIS models. March 17, 2020, the first day in our US data, is chosen as the starting date of the pandemic in the simulations.

For the SEIR model, we set the transition rate to $\sigma_t = \frac{1}{5.1}$ according to **Table 1**. To simulate as close to the reality as possible, we set the transmission rate β_t and the recovery rate γ_t to the means of their true value series obtained in section 4.1. To construct the initial conditions of the system, we use the initial values $I(0) = 311$ and $R(0) = 27$ obtained from the data as well. In previous studies, the average Infected-Suspected ratio in China, one of the earliest hot spots of the global COVID-19 outbreak, was found to be 2.399 (e.g., Fairoza Amira et al., 2020). In this simulation, due to the lack of data

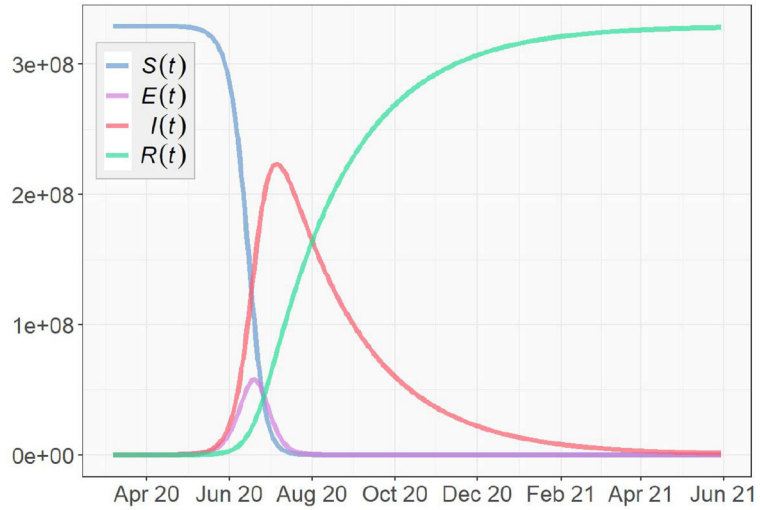


FIGURE 13 | Simulation based on the SEIR Model for the U.S.

of the exposed group, we use the same ratio to initialize $E(t)$, i.e., $E(0) = \frac{1}{2.399} I(0) \approx 130$. According to the U.S. and World Population Clock (United States Census Bureau, 2020), the U.S. population is $N = 329,227,746$. Using (5), we have: $S(0) = N - E(0) - I(0) - R(0) \approx 329,227,278$.

With the aforementioned parameter settings and initial conditions, we simulate the COVID-19 pandemic for the US. As shown in **Figure 13**, the number of infected people reaches a peak in early July, 2020, and the pandemic gradually dies out in summer 2021. It is important to note that the simulation is only theoretical and restricted by given conditions. These conditions can be dramatically different in realty. Moreover, no mitigation measure of any kind that can possibly prevent or limit the spread of the virus is considered in the simulation, such as wearing facial coverings, social distancing, community lockdowns, and work-from-home policies. Being free of the influences of such factors indicates that the pandemic might develop slower in the simulation than in reality. Since many states of the U.S. are following the strict guidelines set by CDC, the pandemic is highly likely to end earlier than the simulation result.

Next, we take immunity, reinfection and vaccination into account, and simulate the pandemic according to the SEVIS model proposed in section 3.1. The parameter settings of β_t, σ_t, and γ_t remain the same as in the SEIR simulation. For the vaccination rate ν_t, we clarify a starting date of vaccination t_v. Before the vaccination starts, i.e., for $t < t_v$, $\nu_t = 0$. When $t \geq t_v$, ν_t becomes positive and based on the discussion in section 3, we assume ν_t to start at a low value in realty and exponentially increase as time goes on. Here, we simplify this process by assuming the mean of $\{\nu_t | t \geq t_v\}$ to be 1% per day and assigning it to ν_t, and let the vaccination start on January 1, 2021. As for the last parameter w in **Table 1**, the fraction of infected cases that become immune after recovery is currently unknown. In this simulation, we assume w to be 0.5.

Figure 14 shows the simulation result with the vertical dashed line representing $t = t_v$, (i.e., the first day of 2021). We notice that the trajectories obtained from the SEVIS model before the vaccination are nearly identical to the previous SEIR simulation. Once vaccination begins, the growth of the immunity group $V(t)$ and the decrease of the infected group $I(t)$ clearly accelerate. However, different from SEIR model which assumes no reinfection, the SEVIS model does allow reinfection, which leads to a longer time for the virus to die out. To speed up the process, we can employ a larger value for w, i.e., increased flows from $I(t)$ to $V(t)$ and reduced flows from $I(t)$ to $S(t)$.

5. CONCLUSION

Considering the incubation period of COVID-19, we first proposed a time-dependent SEIR model with the time-dependent parameters estimated by LASSO regression. The proposed model is validated using the national level data (the United States) and state level data (New York and North Dakota). Overall, our proposed model outperforms the SIR model with smaller prediction errors. Furthermore, by taking immunity, reinfection, and vaccination into account, we proposed a time-dependent SEVIS model without assuming guaranteed immunity after recovery as in the SEIR model. Simulations are performed using the proposed two models to predict the spread of COVID-19 pandemic for the United States.

With the daily recorded data in the U.S., our algorithm predicts that the numbers of the infected and recovered individuals will keep increasing at a high rate in the short future. The total number of confirmed cases in the U.S. is estimated to reach close to $25.7M$ by late January, 2021, while North Dakota and New York will face 1.26 and $0.96M$ total confirmed cases, respectively. Given the historical transmission and recovery rate of the COVID-19, the simulation of SEVIS model predicts that

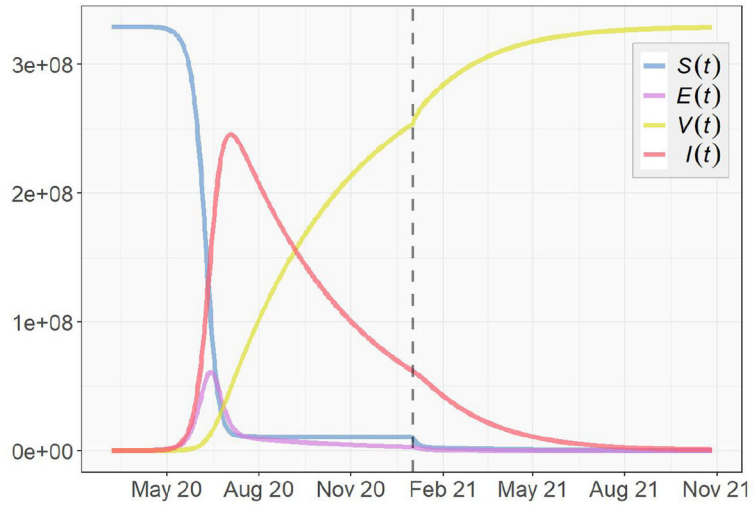


FIGURE 14 | Simulation based on the SEVIS Model for the U.S.

the pandemic will die down in fall 2021, assuming the mean vaccination rate to be 1% per day and the probability of gaining immunity after recovery to be 50%. Note that this prediction is subject to change with more accurate parameters chosen according to the real data once vaccination starts.

In addition, it is crucial to understand that neither of the prediction and simulation takes any mitigation measures that can prevent or limit the growth of the pandemic into consideration, such as social distancing, facial covering, lockdown restrictions, and closing non-essential businesses. As a result, the end of the pandemic in reality is highly likely to come earlier than the numeric outcome. However, at this point the spread of the pandemic is still ongoing and has not been contained yet, as the time-dependent basic reproduction number for US is still steadily positive. Also, in some particular parts of US (e.g., New York), a new surge in the transmission rate was detected as the end of the year 2020 approaches. These all could serve as an alert that it is too early to relax the measures already implemented to tackle the pandemic. Fortunately, these measures have been proven effective by evidences. We expect them to continue taking effect over time and suggest the necessity of bring in more. Hopefully, with effort made by people around the world and the upcoming release of vaccine, we will be able to conquer this global crisis in no time.

Another limitation of the proposed time-dependent SEVIS model is that, it assume absolute immunity to the virus after vaccination, while in reality, the effectiveness of the vaccine is not 100% guaranteed. For example, as reported by the BBC news, a single dose of the Moderna vaccine can provide 80.2% protection. When a second dose is injected after a period of time, the effectiveness rise to 95.6%. In the future, we would like to extend the model by factoring in changing effectiveness at different stage of the vaccination.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

JZ designed the study. YL collected data for analysis and interpreted the results and drafted the manuscript. JZ and YL analyzed the data and developed the models. LG, YZ, XC, and JZ revised the manuscript. All authors gave final approval for publication.

REFERENCES

Chen, Y.-C., Lu, P.-E., Chang, C.-S., and Liu, T.-H. (2020). A time-dependent sir model for covid-19 with undetectable infected persons. *IEEE Trans. Netw. Sci. Eng.* 7, 3279–3294. doi: 10.1109/TNSE.2020.3024723

Diekmann, O., Heesterbeek, J. A. P., and Metz, J. A. J. (1990). On the definition and the computation of the basic reproduction ratio r0 in models for infectious diseases in heterogeneous populations. *J. Math. Biol.* 28, 365–382. doi: 10.1007/BF00178324

Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track covid-19 in real time. *Lancet Infect. Dis.* 20, 533–534. doi: 10.1016/S1473-3099(20)30120-1

Fairoza Amira, B. H., Cher, H., Hafeez, N., Dominic, L., Guanhua, L., Mohammad, S., et al. (2020). *Coronatracker: World-Wide Covid-19 Outbreak Data Analysis and Prediction*. Bulletin of the World Health Organization.

Ganyani, T., Kremer, C., Chen, D., Torneri, A., Faes, C., Wallinga, J., et al. (2020). Estimating the generation interval for covid-19 based on symptom onset data, March 2020. *Eurosurveilance* 25:2000257. doi: 10.2807/1560-7917.ES.2020.25.17.2000257

Gao, D., Porco, T. C., and Ruan, S. (2016). Coinfection dynamics of two diseases in a single host population. *J. Math. Anal. Appl.* 442, 171–188. doi: 10.1016/j.jmaa.2016.04.039

Hsieh, F., and Zheng, J. (2019). Unraveling pattern-based mechanics defining self-organized recurrent behaviors in a complex system: a zebrafish's calcium brain-wide imaging example. *Front. Appl. Math. Stat.* 5:13. doi: 10.3389/fams.2019.00013

Katul, G. G., Mrad, A., Bonetti, S., Manoli, G., and Parolari, A. J. (2020). Global convergence of COVID-19 basic reproduction number and estimation from early-time sir dynamics. *PLoS ONE* 15:e239800. doi: 10.1371/journal.pone.0239800

Murray, C. J. (2020). Forecasting COVID-19 impact on hospital bed-days, icu-days, ventilator-days and deaths by us state in the next 4 months. *medRxiv [Preprint]*. doi: 10.1101/2020.03.27.20043752

Ngonghala, C. N., Iboi, E., Eikenberry, S., Scotch, M., MacIntyre, C. R., Bonds, M. H., et al. (2020). Mathematical assessment of the impact of non-pharmaceutical interventions on curtailing the 2019 novel coronavirus. *Bellman Prize Math. Biosci.* 325:108364. doi: 10.1016/j.mbs.2020.108364

Read, J. M., Bridgen, J. R., Cummings, D. A., Ho, A., and Jewell, C. P. (2020). Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. *medRxiv [Preprint]*. doi: 10.1101/2020.01.23.20018549

Schmidt, W. (1981). Eisen, M.: Mathematical models in cell biology and cancer chemotherapy. Lecture notes in biomathematics, vol. 30. Springer-Verlag, Berlin-Heidelberg-New York 1979. IX, 431 s., 70 abb., 17 tab., DM 39,-. *Biometr. J.* 23, 519–520. doi: 10.1002/bimj.4710230517

Sharomi, O., and Gumel, A. (2011). Dynamical analysis of a sex-structured chlamydia trachomatis transmission model with time delay. *Nonlin. Anal. Real World Appl.* 12, 837–866. doi: 10.1016/j.nonrwa.2010.08.010

Shen, M., Peng, Z., Xiao, Y., and Zhang, L. (2020). Modelling the epidemic trend of the 2019 novel coronavirus outbreak in china. *Innovation* 1:100048. doi: 10.1016/j.xinn.2020.100048

Tan, W., Lu, Y., Zhang, J., Wang, J., Dan, Y., Tan, Z., et al. (2020). Viral kinetics and antibody responses in patients with covid-19. *medRxiv [Preprint]*. doi: 10.1101/2020.03.24.20042382

To, K. K. W., Tsang, O. T. Y., Leung, W. S., Tam, A. R., Wu, T. C., Lung, D. C., et al. (2020). Temporal profiles of viral load in posterior oropharyngeal saliva samples and serum antibody responses during infection by sars-CoV-2: an observational cohort study. *Lancet Infect. Dis.* 20, 565–574. doi: 10.1016/S1473-3099(20)30196-1

Toda, A. A. (2020). Susceptible-infected-recovered (sir) dynamics of COVID-19 and economic impact. *arXiv [Preprint] arXiv*:2003.11221.

United States Census Bureau (2020). *U.S. and World Population Clock*. United States Census Bureau. Available online at: https://www.census.gov/ (accessed December 24, 2020).

van den Driessche, P., and Watmough, J. (2002). Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Bellman Prize Math. Biosci.* 180, 29–48. doi: 10.1016/S0025-5564(02)00108-6

Wu, T., Ge, X., Yu, G., and Hu, E. (2020). Open-source analytics tools for studying the COVID-19 coronavirus outbreak. *medRxiv [Preprint]*. doi: 10.1101/2020.02.25.20027433

You, C., Deng, Y., Hu, W., Sun, J., Lin, Q., Zhou, F., et al. (2020). Estimation of the time-varying reproduction number of COVID-19 outbreak in china. *Int. J. Hyg. Environ. Health* 228:113555. doi: 10.1016/j.ijheh.2020.113555

Zheng, J., Fushing, H., and Ge, L. (2019). A data-driven approach to predict and classify epileptic seizures from brain-wide calcium imaging video data. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 17, 1858–1870. doi: 10.1109/TCBB.2019.2895077

Zheng, J., Liang, M., Ekstrom, A. D., Ge, L., Yu, W., and Hsieh, F. (2018). "On association study of scalp EEG data channels under different circumstances," in *International Conference on Wireless Algorithms, Systems, and Applications* (New York, NY), 683–695. doi: 10.1007/978-3-319-94268-1_56

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

REVIEW
published: 14 May 2021
doi: 10.3389/frai.2021.652669

The Promise of AI in Detection, Diagnosis, and Epidemiology for Combating COVID-19: Beyond the Hype

Musa Abdulkareem [1,2,3]* *and Steffen E. Petersen* [1,2,3,4]

[1] *Barts Heart Centre, Barts Health National Health Service (NHS) Trust, London, United Kingdom,* [2] *National Institute for Health Research (NIHR) Barts Biomedical Research Centre, William Harvey Research Institute, Queen Mary University of London, London, United Kingdom,* [3] *Health Data Research UK, London, United Kingdom,* [4] *The Alan Turing Institute, London, United Kingdom*

COVID-19 has created enormous suffering, affecting lives, and causing deaths. The ease with which this type of coronavirus can spread has exposed weaknesses of many healthcare systems around the world. Since its emergence, many governments, research communities, commercial enterprises, and other institutions and stakeholders around the world have been fighting in various ways to curb the spread of the disease. Science and technology have helped in the implementation of policies of many governments that are directed toward mitigating the impacts of the pandemic and in diagnosing and providing care for the disease. Recent technological tools, artificial intelligence (AI) tools in particular, have also been explored to track the spread of the coronavirus, identify patients with high mortality risk and diagnose patients for the disease. In this paper, areas where AI techniques are being used in the detection, diagnosis and epidemiological predictions, forecasting and social control for combating COVID-19 are discussed, highlighting areas of successful applications and underscoring issues that need to be addressed to achieve significant progress in battling COVID-19 and future pandemics. Several AI systems have been developed for diagnosing COVID-19 using medical imaging modalities such as chest CT and X-ray images. These AI systems mainly differ in their choices of the algorithms for image segmentation, classification and disease diagnosis. Other AI-based systems have focused on predicting mortality rate, long-term patient hospitalization and patient outcomes for COVID-19. AI has huge potential in the battle against the COVID-19 pandemic but successful practical deployments of these AI-based tools have so far been limited due to challenges such as limited data accessibility, the need for external evaluation of AI models, the lack of awareness of AI experts of the regulatory landscape governing the deployment of AI tools in healthcare, the need for clinicians and other experts to work with AI experts in a multidisciplinary context and the need to address public concerns over data collection, privacy, and protection. Having a dedicated team with expertise in medical data collection, privacy, access and sharing, using federated learning whereby AI scientists hand over training algorithms to the healthcare institutions to train models locally, and taking full advantage of biomedical data stored in biobanks can alleviate some of problems posed by these challenges. Addressing these challenges

OPEN ACCESS

Edited by:
Jake Y. Chen,
University of Alabama at Birmingham, United States

Reviewed by:
Maria F. Chan,
Memorial Sloan Kettering Cancer Center, United States
Akram Mohammed,
University of Tennessee Health Science Center (UTHSC), United States

****Correspondence:***
Musa Abdulkareem
musa.abdulkareem@nhs.net

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *12 January 2021*
Accepted: *13 April 2021*
Published: *14 May 2021*

Citation:
Abdulkareem M and Petersen SE (2021) The Promise of AI in Detection, Diagnosis, and Epidemiology for Combating COVID-19: Beyond the Hype. Front. Artif. Intell. 4:652669. doi: 10.3389/frai.2021.652669

will ultimately accelerate the translation of AI research into practical and useful solutions for combating pandemics.

Keywords: artificial intelligence, COVID-19, detection, diagnosis, epidemiology, social control, contact tracing, medical imaging

INTRODUCTION

COVID-19, a type of coronavirus disease caused by Severe Acute Respiratory Syndrome Corona-Virus 2 (SARS-CoV-2), has created enormous suffering, affecting lives and causing deaths. The novel nature of the virus means that humans are only newly exposed to the virus (Brüssow, 2020; Wan et al., 2020). First reported in China in December 2019, it was declared by The World Health Organization (WHO) to be a Public Health Emergency of International Concern (PHEIC) on January 30, 2020 and a pandemic on March 11, 2020 (Team, 2020; WHO, 2020). It is an infectious disease that spreads in humans mainly through respiratory droplets produced by an already infected person through sneezing or talking, or airborne transmission (Moriyama et al., 2020). The early symptoms of the disease include persistent high temperature, dry continuous coughing, loss of taste or smell, and difficulty in breathing (Kooraki et al., 2020; Wang et al., 2020a). Severe cases of the disease cause death (Rothan and Byrareddy, 2020; Zhou et al., 2020a).

Due to the ease with which the coronavirus can spread and grow exponentially within the human population, healthcare resources and manpower to rapidly control it is limited as the number of doctors, nurses, and other healthcare workers and resources that could help control it is finite. Moreover, the disease has exposed weaknesses of many healthcare systems around the world. Indeed, the lack of affordable, quick and accurate means of detecting the disease is one of the most important reasons it has rapidly spread (Ai et al., 2020).

Since the emergence of COVID-19, many governments, research communities, commercial enterprises and other institutions and stakeholders around the world have been fighting in various ways to curb the spread of the disease (Chen et al., 2020a; Dong et al., 2020). Science and technology have helped in the implementation of policies of many governments that are directed toward mitigating the impacts of the pandemic and in developing cures and vaccines for the disease. They also offer unique opportunity to support healthcare workers by providing them with tools that would save them time, improve their ability to carry out their job and enhance the management of healthcare systems developed to combat the pandemic, and much more. Many resources have been made available to support the battle against COVID-19, such as datasets (Cheng et al., 2020; Cohen et al., 2020; Zhao et al., 2020a), computing resource (Hack and Papka, 2020), and research funding (Casigliani et al., 2020; Glasziou et al., 2020; Janiaud et al., 2020; Patel et al., 2020; Prudêncio and Costa, 2020; UKCDR, 2020).

The scope of combating COVID-19 using technology is very broad and it includes understanding the socio-economic and medical impacts of the pandemic. From a healthcare perspective, it includes disease detection, diagnosis, and monitoring (Huang et al., 2020a; Kong et al., 2020; Thevarajan et al., 2020; Xu et al., 2020a), epidemiology (Chan et al., 2020; Jin et al., 2020a; Li et al., 2020a), social control (Jin et al., 2020a; Kandel et al., 2020; Qian et al., 2020), virology and pathogenesis (Andersen et al., 2020; Jin et al., 2020a; Lu et al., 2020b; Walls et al., 2020), and drug discovery (Chen et al., 2020b; Phua et al., 2020). For example, during the early phase of the outbreak of the pandemic, China used facial recognition cameras to track infected patients and drones to disinfect public places and broadcast audio messages to the public asking them to stay at home (Ruiz Estrada, 2020). As another example, Taiwan linked its national medical insurance database with the immigration and custom database in order to inform the healthcare practitioners of the travel history of patients (Wang et al., 2020b).

The term artificial intelligence (AI) refers to the study of developing computer algorithms with human-like intelligence to accomplish specific tasks. Machine learning (ML) methods are a set of techniques in AI and includes supervised (Kotsiantis et al., 2007), unsupervised (Barlow, 1989), semi-supervised (Zhu, 2005; Chapelle et al., 2009), and reinforcement learning (Sutton and Barto, 1998). Some of these methods and other terms often encountered in the AI literature are briefly described in **Table 1**.

The applications of AI can be found in many disciplines and industries in modern society, and healthcare is not an exemption. The rapid growth of AI-based techniques and tools in healthcare are addressing complex problems such as identifying previously undiscovered relationships in patient phenotypes (Shivade et al., 2014), optimizing healthcare pathways (Lu and Wang, 2019; Blasiak et al., 2020), and improving accuracy of medical decision making (Bennett and Hauser, 2013; Shortliffe and Sepúlveda, 2018).

Advances and accessibility to high-performance scalable computing equipment have driven the recent popularity of the use of AI in many real-world applications. This development has also prompted an expansion of research into novel AI techniques and algorithms. AI algorithms have the potential to interpret biomedical and healthcare data particularly for tasks where conventional statistical methods are less efficient. The algorithms are even more suitable for datasets of large scale and high dimensions. These algorithms can therefore be used to solve problems such as optimizing care pathways, standardizing clinical diagnosis, identifying relationships in patient phenotypes and developing predictive models (Johnson et al., 2017). While AI-based methods can be used to solve many problems in medicine and healthcare, the success of AI projects, in many cases, depends on the choice of the AI technique, the quality of the dataset to be used and the context associated with the way the dataset is used. For instance, deep learning (DL) algorithms such as convolutional neural networks (CNN) are particularly suitable for computer vision problems such as image segmentation (Shen

TABLE 1 | Some terms and methods commonly used in AI.

General terms	
Artificial Intelligence (AI)	The concept of developing computer algorithms with human-like intelligence to solve specific tasks.
Deep Learning (or Deep Neural Network)	A set of ML algorithms that are based on neural network (NN) that are used for feature learning. The term "deep" refers to the fact that they have multiple layers between the input and the output layers.
Machine Learning (ML)	A subset of AI and consists of a collection of techniques to achieve AI.
Reinforcement Learning	A set of ML algorithms that is based on the interaction between an agent and its environment. In general, the agent seeks to take actions in the environment by maximizing a cumulative reward.
Supervised Learning	A set of ML algorithms for developing mathematical models using data that consists of both the input and the desired output data.
Unsupervised Learning	A set of ML algorithms for finding underlying structures or patterns in datasets using only the input data.
Convolutional Neural Network (CNN)	A set of DL algorithms that are particularly efficient in developing AI-based applications involving images. CNN acts as the backbone of many well-known neural network architectures (such as U-net) used in image processing.
Random Forests (RF) Method	A set of learning algorithms involving several decision trees and whose output is the class that is the statistical mode (in classification tasks) or statistical mean (in regression tasks) of each of the decision trees. These algorithms are often used for classification tasks and regression analysis problems.
Support Vector Machines (SVM)	A set of supervised learning algorithms that constructs hyperplanes in a high-dimensional space. These algorithms are often used for classification tasks, regression analysis, and other problems. In a classification problem, for instance, out of the many hyperplanes, the one that has the largest distance to the data point of any class is considered the 'optimal' classifier.

Reference List of AI Algorithms Mentioned in this Paper

- AlexNet (Russakovsky et al., 2015; Krizhevsky et al., 2017)
- Artificial Neural Networks (ANN) (Hopfield, 1988; Jain et al., 1996)
- Adaptive-Network-based Fuzzy Inference System (ANFIS) (Jang, 1993)
- CNN (LeCun et al., 2015)
- CNN segmentation model (Region Proposal Network structure) (Ren et al., 2016)
- CNN model with Inception (Szegedy et al., 2016)
- Decision Tree (DT) (Breiman et al., 1984)
- Extreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016)
- Generative Adversarial Networks (GANs) (Goodfellow et al., 2014)
- Gated Recurrent Unit (GRU) recurrent neural network (Cho et al., 2014; Chung et al., 2014)
- k-mean clustering (Kanungo et al., 2002)
- k-nearest neighbor (Cover and Hart, 1967)
- Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression (Tibshirani, 1996)
- Logistic regression (Hosmer Jr et al., 2013)
- LSTM (Hochreiter and Schmidhuber, 1997)
- RF (Breiman, 2001; Liaw and Wiener, 2002)
- ResNet (He et al., 2016)
- SVM (Cortes and Vapnik, 1995)
- U-Net (Ronneberger et al., 2015)

et al., 2017). Recent advancement in AI research has led to the development of tools in medicine and healthcare that are useful in combating global pandemics. Researchers across several areas of expertise and industries have therefore explored and exploited the use of AI in the battle against COVID-19.

There are many ways in which AI can help in the fight against the COVID-19 pandemic. For example, AI could be used to track the spread of the virus (Al-Qaness et al., 2020; Bandyopadhyay and Dutta, 2020; Carrillo-Larco and Castillo-Cara, 2020; Hu et al., 2020; Jana and Bhaumik, 2020; Huang et al., 2020b; Kavadi et al., 2020; Sameni, 2020), identify patients with high mortality risk (Jiang et al., 2020a; Qi et al., 2020; Xu et al., 2020b; Yan et al., 2020a), diagnose and screen a population for COVID-19 (Ghoshal and Tucker, 2020; Hassanien et al., 2020; Hemdan et al., 2020; Jin et al., 2020b; Maghdid et al., 2020a; Narin et al., 2020; Wang et al., 2020c,e; Wu et al., 2020a; Zhang et al., 2021; Xu et al., 2020c), or reduce the time for diagnosis (Vaishya et al., 2020a). Many of the AI techniques currently being deployed in the battle already existed prior to the pandemic. These techniques include those that can process and understand medical imaging

data from modalities such as computed tomography (CT) and X-ray that are being used for detection and diagnosis (Li et al., 2020b; Wang et al., 2020e; Wynants et al., 2020) and those involving non-imaging data that are being used for mortality rate and outcome prediction, prognosis, outbreak prediction, contact tracing and social control of COVID-19 (John and Shaiba, 2019; Bandyopadhyay and Dutta, 2020; Chen et al., 2020e; Goh et al., 2020; Pourhomayoun and Shakibi, 2020; Xu et al., 2020b). Other AI techniques have also found new application areas due to COVID-19. For example, in Shi et al. (2020a), argued for the development of AI-based tools for automated acquisition of medical images in order to optimize the imaging workflow and reduce healthcare practitioners' risk of exposure to the virus by minimizing or eliminating contact with COVID-19 patients.

Several reviews, such as Albahri et al. (2020), Bansal et al. (2020), Bragazzi et al. (2020), Bullock et al. (2020), Jamshidi et al. (2020), Kricka et al. (2020), Kumar et al. (2020), Lalmuanawma et al. (2020), Martin et al. (2020), Naudé (2020), Nguyen (2020), Rasheed et al. (2020), Suri et al. (2020), Shi et al. (2020a), Vaishya et al. (2020b), Zhou et al. (2020b), and Chen et al. (2020c), have been published to showcase the opportunities AI presents in the current effort to fight against COVID-19. In this paper, areas where AI techniques are being used in the detection, diagnosis and epidemiological predictions, forecasting and social control for combating COVID-19 are discussed, highlighting areas of successful applications and underscoring issues that need to be addressed to achieve significant progress in battling COVID-19 and future pandemics. The paper assumes a basic background knowledge of AI techniques, the reader is invited to consult (Raghu and Schmidt, 2020) for further information of these AI methods. Useful introduction to the epidemiology and clinical features of COVID-19 can be found in, for example, C Disease Control (2020).

AI IN COVID-19 DETECTION AND DIAGNOSIS

Detection and diagnosis of COVID-19 is an important part in the fight against the virus. Current diagnostic testing methods are mostly non-invasive methods and they include chest CT and chest X-ray medical imaging, nucleic acid, serologic, and viral throat swab testing methods (Fang et al., 2020; Li et al., 2020c; Lu et al., 2020a; Ozturk et al., 2020; Schwartz, 2020; Zeng et al., 2020). In order to contain the spread of the pandemic and isolate the virus, fast and early detection and tracking of infected patients is crucial and there is clearly the need of innovation in this area (Ai et al., 2020; Fang et al., 2020). In the subsections that follow, AI tools that have been developed for the detection and diagnosis of SARS-CoV-2 and COVID-19 are presented.

Nucleic Acid Amplification-Based Diagnostics

A type of nucleic acid amplification test (Udugama et al., 2020), the Reverse Transcription-Polymerase Chain Reaction (RT-PCR) test, is one of the most widely used standard testing methods for detecting whether patients have COVID-19 (Ai et al., 2020). The RT-PCR however suffers from inadequate sensitivity, as low as 71% as reported in Fang et al. (2020), as a result of many factors such as low detection efficiency and complicated sample preparation (Lu et al., 2020a; Wu et al., 2020a). This low sensitivity issue results in multiple testing of many patients usually over several days apart in order to obtain a reliable conclusion.

A ML model was reported in Wu et al. (2020a) that uses 11 key blood indices to distinguish between patients with and without COVID-19. The model was developed using the random forest (RF) ML technique and 49 clinical available blood test parameters (consisting of 24 routine hematological and 25 biochemical parameters) from 169 patients with a total number of 253 data samples of which 105 samples are from patients confirmed to have the COVID-19 disease using the RT-PCR test. The remaining samples consisted of 98 samples from patients with common pneumonia and 25 samples each from patients with tuberculosis and lung cancer. The data was divided into 149 training, 33 testing, and 74 validating datasets. The model achieved accuracy of 96.97% on the testing set and 97.95% for the cross-validation set. While this model (which could be further investigated for reliability and also improved further) offers a promising tool for preliminary assessment of suspected patients with COVID-19, it so far has not made it to front-line in the fight against the coronavirus.

Medical Imaging Diagnostics

Medical imaging is one of the main areas in which AI has found practical applications in medicine and healthcare. Imaging data obtained using different modalities, such as computed tomography (CT), magnetic resonance imaging (MRI) and X-ray, are of high dimension. They contain very rich information that can be used to develop AI applications. Imaging data can be used to generate many useful image-derived phenotypes that are obtained *via* qualitative and quantitative assessment of structural changes (that often characterize the structural and functional properties of an organ), significantly shortening the time for radiologists to accomplish these tasks (Petersen et al., 2017; Suinesiaputra et al., 2018; Mauger et al., 2019). Imaging data can also be combined with non-imaging data from Electronic Health Record (EHR) or elsewhere for identifying biomarkers and predicting disease risk factors (Alaa et al., 2019). Faster and automated reading and interpretation of image workflow can be achieved using AI-based tools (Petersen et al., 2019; Robinson et al., 2019; Bai et al., 2020).

Furthermore, segmentation of medical images is useful as these images are often affected by noise, artifacts, and other uncertainties associated with imaging. Image segmentation involves contouring a medical image into biologically relevant structures, helping to quantify those structures and their functions and to produce measurements that act as biomarkers (such as quantities that can be used to diagnose, monitor, or prognosticate diseases). In particular, AI-based automatic image segmentation tools are beneficial as they help in eliminating variability that would have been introduced if segmentation were manually done. Consequently, the use of these AI-based technologies has contributed in the fight against COVID-19

(Bullock et al., 2020). Medical imaging modalities, such as chest CT and X-ray imaging, have provided significant support to clinicians in diagnosing COVID-19 (Apostolopoulos and Mpesiana, 2020; Bernheim et al., 2020; Kanne, 2020). A typical workflow for diagnosing COVID-19 with medical imaging modalities involves the following three phases: (i) pre-scan preparation according to a given protocol; (ii) image acquisition; and (iii) diagnosis.

AI tools for COVID-19 diagnosis with medical images often consist of one or a combination of several AI models (or networks) involving the following two main components: (i) image segmentation models, and (ii) image classification. Image segmentation is used to mark and identify the region of interest (ROI) while an image classification task extracts features from the ROI and uses those features as a basis for classifying (diagnosing) the images.

CT Medical Imaging

Chest CT images are being used for early diagnosis of COVID-19 by identifying ground-glass opacity (GGO) around the subpleural region (Ai et al., 2020; Chung et al., 2020; Fang et al., 2020; Kanne, 2020; Wong et al., 2020). In Pan et al. (2020), the dynamic radiological patterns in chest CT images of COVID-19 patients was reported with the following four stages identified: (i) 0–4 days: early stage; (ii) 5–8 days: progressive stage; (iii) 9–13 days: peak stage; and (iv) 14 days and beyond: absorption stage. These distinct manifestations of COVID-19 in CT images provide evidence and severity of the disease that are exploited using AI systems for diagnosing the disease. Generally, the process of COVID-19 diagnosis with CT images involves the following steps: (i) image pre-processing, (ii) image segmentation, (iii) classification, and (iv) model evaluation.

AI tools for COVID-19 diagnosis with CT images involving lung tissue segmentation are reported in Jin et al. (2020b), Li et al. (2020b), and Xu et al. (2020c). As an example, the AI system presented in Jin et al. (2020b) classifies chest CT input image slices into the following four categories: non-pneumonia, non-viral community acquired pneumonia (CAP), influenza-A/B and COVID-19. The predicted class of an image is that which has the highest probability among the four classes. This AI system was developed using two main DL algorithms: the U-Net for performing the lung segmentation task and the ResNet for performing the classification (diagnosis) task. The ROI in the image include the lung, lung lobes, bronchopulmonary segments, and infected lesions. The dataset consists of 10,250 CT scans from three centers in China and three publicly available external databases. This multi-center dataset was from 7,917 subjects consisting of 3,686 scans of COVID-19, 2,886 scans of CAP, 132 scans of influenza-A/B and 3,546 scans of non-pneumonia subjects. The COVID-19 subjects were all confirmed using the RT-PCR diagnostic test. The imaging dataset of 10,250 was divided into a total training dataset of 5,104 and a total testing dataset of 5,146. As a measure of accuracy, on internal testing dataset of 3,203 images (out of the 5,146) the AI system achieved an AUC of 97.17%, a sensitivity of 90.19% and a specificity of 95.76%. It achieved an AUC of 97.77% on the remaining (external) dataset of 1,943 images.

As another example, the AI system presented in Xu et al. (2020c) classifies chest CT input image slices into the following three categories: influenza-A viral pneumonia (IAVP), COVID-19 and irrelevant to infection (i.e., cases that do not belong to the other two categories) cases. The predicted class of an image is that which has the highest probability that the image belongs to it. This AI system consists of two main DL algorithms: a three-dimensional (3D) CNN segmentation model (Region Proposal Network structure) for performing a lung segmentation task and a ResNet-based model for performing the image classification task. The dataset consists of 618 CT scans from three hospitals in China's Zhejiang Province of which 110 subjects (219 scans) were confirmed of COVID-19 using the RT-PCR diagnostic test; 224 subjects (224 scans) had IAVP, and the remaining 175 scans are healthy subjects. The imaging dataset of 528 (189 COVID-19 cases plus 194 IAVP cases plus 145 healthy cases) were used for training and validation, and the remaining 90 scans (30 COVID-19 cases plus 30 IAVP cases plus 30 healthy cases) were used as a testing dataset. On the testing dataset, the AI system achieved an f_1-score of 83.9% for COVID-19 cases, 84.7% for IAVP cases, 91.5% for healthy cases and an overall accuracy rate of 86.7%.

Several AI systems, such as Ardakani et al. (2020), Chen et al. (2020d), Gozes et al. (2020), Kang et al. (2020), Li et al. (2020b), Shi et al. (2020b), Song et al. (2020), Tang et al. (2020) and Wang et al. (2020c,d), have been developed for diagnosing COVID-19. Compared to the examples in the two preceding paragraphs, these AI systems mainly differ in their choices of the algorithm for image segmentation of the ROI and the algorithm used for classification or diagnosis. The image segmentation algorithms used include U-Net, U-Net++, V-Net, and others, and the image classification algorithms include ResNet and CNN model with Inception.

In order to address the problem of lack of large datasets of COVID-19 patients for developing AI-based models, researchers, such as in Jin et al. (2020b) and Zhao et al. (2020a), have used different techniques such as data augmentation and transfer learning, to solve the CT image classification problems for COVID-19 diagnosis. In Qian et al. (2020), the classification task was to classify COVID-19 patients into those that will have short-term and long-term hospital stay. Some AI models, such as Shi et al. (2020b,c), went further after the image segmentation task to predict the severity of COVID-19 in patients using algorithms such as least absolute shrinkage and selection operator (LASSO) logistic regression model and RF.

X-Ray Medical Imaging

The X-ray technology is a very popular imaging modality in medical imaging (Wang et al., 2017). The CT and X-ray medical imaging modalities have been more widely accessible and used to provide evidence and for COVID-19 diagnosis compared to other imaging modalities due to their fast acquisition. In fact, in many healthcare centers and hospitals, X-ray imaging, due to its accessibility and quickness to obtain, is often the first-line imaging modality for suspected COVID-19 patients (Bullock et al., 2020; Shi et al., 2020a). Although the chest X-ray images are less informative compared to CT images for diagnosing COVID-19 due to lower sensitivity of chest X-ray images, the popularity

and availability of X-ray imaging facilities means that it is widely used for the diagnosis of the disease. As with chest CT imaging, chest X-ray imaging is being used for diagnosis of COVID-19 by identifying ground-glass opacity (GGO) around the subpleural region, and these manifestations of COVID-19 in chest X-ray images provide evidence and classification of severity of the disease that are being exploited using AI systems for diagnosing the disease.

In general, the process of COVID-19 diagnosis with chest X-ray images using AI tools involves the following steps: (i) image pre-processing, (ii) image classification, and (iii) model evaluation. In other words, compared to the AI tool for CT images, the image segmentation process is absent although some researchers, such as Hassanien et al. (2020), included classical computer vision methods (i.e., not AI-based methods, such as image thresholding) for carrying out the image segmentation step as well. The AI-based image segmentation part of the process is particularly difficult in the case of chest X-ray images given that the ribs are projected onto other tissues on these images (Chen et al., 2020c) so researchers often skip that step completely. Classification tasks were binary, multi-class, multi-labeled or hierarchical classifications (Albahri et al., 2020).

Several AI systems, such as Ghoshal and Tucker (2020), Hassanien et al. (2020), Hemdan et al. (2020), Maghdid et al. (2020a), Narin et al. (2020), Wang et al. (2019, 2020d), Zhang et al. (2021), have been developed for diagnosing COVID-19 using chest X-ray images. These AI systems mainly differ in their choice of the algorithms used for the image classification task and often combine several algorithms (often, to achieve a feature extraction step before a classification process). The image classification algorithms that are being used include Support Vector Machines (SVM), CNN, AlexNet, ResNet, and CNN model with Inception.

The large number of AI techniques available for diagnosing and classifying a disease means that it can be daunting to select the most appropriate technique (in terms of accuracy and computation efficiency) for a given problem given that many of the researchers have used different (and sometimes conflicting) evaluation criteria for their adopted techniques (Alsalem et al., 2018, 2019; Zaidan et al., 2020). In Albahri et al. (2020), carried out a literature review of AI techniques involving medical images that are being used for diagnosing COVID-19 in an attempt to evaluate and establish benchmarking procedures for these techniques. A detailed description of the proposed methodology for the evaluation and benchmarking of these AI techniques is beyond the scope of this paper and the reader is invited to consult (Albahri et al., 2020) for further information.

Other Tools for Diagnostics

In Schuller et al. (2020), presented a potential computer audition tool that uses AI-based speech and sound analysis to COVID-19 diagnosis. The authors surveyed automatic recognition and monitoring of contextually significant phenomena from speech or sound, such as dry and wet coughing or sneezing sounds, pain, speech under cold, and breathing for diagnostic exploitation using AI techniques such as Generative Adversarial Networks (GANs) (Pascual et al., 2017).

In Wang et al. (2020f), an AI-based classification model was proposed that is able to distinguish respiratory pattern from six other viral infection respiratory patterns using the Gated Recurrent Unit (GRU) recurrent neural network algorithm with bi-directional attention mechanism. As measures of accuracy, the reported precision, recall, f_1-score, and accuracy of the model were 94.4, 95.1, 94.8, and 94.5%, respectively. Other models that use respiratory or coughing data for COVID-19 diagnosis can be found in Brown et al. (2020), Imran et al. (2020), and Jiang et al. (2020b).

Researchers, such as in Maghdid et al. (2020b), have also proposed frameworks for using in-built mobile phone sensors including cameras (to scan CT images, for example), temperature sensors, and so on, for COVID-19 diagnosis. The computer audition tools for diagnosing COVID-19, models that use respiratory or coughing data for COVID-19 diagnosis and other AI-based computational frameworks that use speech and sound analysis and in-built mobile sensors, such as Iqbal and Faiz (2020) have not yet gone beyond the conceptual phase.

AI IN EPIDEMIOLOGY

In the subsections that follow, AI tools that have been developed for epidemiological predictions, forecasting and social control for combating COVID-19 are presented.

AI for Prognosis

The ability to forecast possible patient outcomes is vital in the planning and management of a pandemic such as COVID-19. In order to improve prognosis and not to overwhelm healthcare systems, the ability to predict number of patients at risk of developing acute respiratory distress syndrome and patients at risk of hospitalization or death can be very important (Bullock et al., 2020). In the fight against MERS Co-V, for example, AI-based models have been used to predict prognosis in patients' infection (in particular, patients' recovery) using patients' profession (e.g., whether healthcare workers or not), age, pre-existing healthcare conditions, and disease severity as model input parameters (John and Shaiba, 2019). Similar AI-based applications and methods have been developed for Ebola patients (Colubri et al., 2016; Riad et al., 2019). These and other similar tools can help, for example, to assess healthcare preparedness for a pandemic and to determine treatment methods and resource allocation during a pandemic, and some of the these algorithms could be adapted for decision making in the management of COVID-19 (Bansal et al., 2020).

Epidemiological research is a vast area, and a huge amount of publications on epidemiological modeling of COVID-19 using well-established classical methods have surfaced since the beginning of the pandemic (Cooper et al., 2020; Jewell et al., 2020; Ndairou et al., 2020). Recently, researchers have proposed several AI-based techniques for predicting mortality rate, long-term patient hospitalization (Qi et al., 2020) and patient outcomes for COVID-19 (Jiang et al., 2020a; Yan et al., 2020a). AI-based techniques that have been used to accomplish the prediction tasks include artificial neural networks (ANN), SVM, and XGBoost. For example, in Pourhomayoun and Shakibi

(2020), using dataset of more than 117,000 confirmed COVID-19 patients from 76 countries described in Xu et al. (2020b), the authors used several AI-methods (including SVM, ANN, RF, Decision Tree (DT), logistic regression, and k-nearest neighbor) for the prediction of mortality rate of COVID-19 patients using 112 features consisting of 80 features from patients' doctors notes and health status and 32 features from patients' demographic and physiological data.

AI for Outbreak Forecasting and Control

The development of forecasting models in order to help policy makers and other stakeholders understand the progression of the pandemic is one of the first areas where mathematical methods were applied to tackle the COVID-19 pandemic. It is therefore not surprising that outbreak forecasting is also one of the first areas in which AI methods have been applied in the fight against the COVID-19 pandemic (Rasheed et al., 2020). There are many existing statistical and dynamic methods for modeling the spread of infectious diseases and understand the impact of interventions to curb these diseases, such as mass vaccination or social distancing, in any given population (Anderson and May, 1979; May and Anderson, 1979; Mena-Lorcat and Hethcote, 1992; Isham and Medley, 1996; Vynnycky and White, 2010; Siettos and Russo, 2013; Pastor-Satorras et al., 2015). Several of these methods have been used to understand and forecast the spread of COVID-19 from available data (Karako et al., 2020; Sameni, 2020; Wu et al., 2020b; Zhao et al., 2020b). These methods can be used to determine transmission factors in order to establish preventive and control measures for the pandemic.

The majority of AI applications developed in the fight against COVID-19 have focused on predicting national and local statistics such as the number of confirmed cases, deaths, and people recovered from COVID-19 (Bullock et al., 2020). AI models that have been developed for outbreak predictions include (Al-Qaness et al., 2020; Bandyopadhyay and Dutta, 2020; Carrillo-Larco and Castillo-Cara, 2020; Hu et al., 2020; Jana and Bhaumik, 2020; Huang et al., 2020b; Kavadi et al., 2020; Sameni, 2020), and the modeling techniques used for these models include CNN, long short-term memory (LSTM), adaptive-network-based fuzzy inference system (ANFIS), partial derivative regression and non-linear machine learning (PDR-NML) (Kavadi et al., 2020), SVM and k-mean clustering.

For example, in Carrillo-Larco and Castillo-Cara (2020), a model based on the k-means clustering algorithm was developed and used to categorize countries based on the number of confirmed COVID-19 cases using a dataset that contains features such as the prevalence of HIV/AIDS, diabetes, and tuberculosis in 156 countries in addition to data on the number of COVID-19 related deaths, confirmed cases and recovered cases. In Al-Qaness et al. (2020), an ANFIS-based model was developed to estimate and forecast the number of confirmed cases of COVID-19 10 days ahead using data of previously confirmed cases. And in Ribeiro et al. (2020), for 10 Brazilian states with a high daily COVID-19 incidence, a stacked ensemble of learning algorithms [autoregressive integrated moving average (ARIMA), cubist regression (CUBIST), RF, ridge regression (RIDGE), SVM] with a Gaussian process (GP) meta-learner was used to conduct 1, 3, and 6-days ahead time series forecasting of the COVID-19 cumulative confirmed cases, achieving errors in a range of 0.87–3.51%, 1.02–5.63%, and 0.95–6.90% in 1, 3, and 6-days-ahead predictions, respectively.

In addition, some of these AI-based models, such as in Kavadi et al. (2020), have reported prediction accuracies that are superior to traditional linear regression-based methods. Researchers, such as in Fong et al. (2020), have also proposed techniques for comparing these different models that have mostly been developed using different architectures and trained with non-identical hyperparameters.

AI for Contact Tracing and Social Control

The implementation of indiscriminate lockdowns in several countries in an attempt to control the COVID-19 pandemic have had severe social and economic consequences. Despite the physical distancing measures in-place when some of the lockdown restrictions where gradually relaxed, other public health measures were necessary in order to control the pandemic (Hellewell et al., 2020; Hope et al., 2020; Park et al., 2020; Salathé et al., 2020; Kretzschmar et al., 2020a,b), and contact tracing (whether conventional methods that rely on interviewing COVID-19 patients or mobile phone application technology) has been one of the methods that have been adopted in many parts of the world for this purpose. Contact tracing involves contact identification, contact listing and contact follow-up (Kricka et al., 2020).

For contact tracing purposes, mobile applications that have been deployed to notify every participating user that a person with COVID-19 was within a certain distance of the user for more than a specific amount of time include COVIDSafe (Australia), Ketju (Finland), CoronaApp (Germany), StopCovid (France), NZ COVID Tracer (New Zealand), TraceTogether (Singapore), NHS Covid-19 App (United Kingdom), to mention a few (Lalmuanawma et al., 2020). As far as we know, none of these digital technologies have been confirmed to use AI-based models as tools, for example, in identifying those in contact with a COVID-19 patient [in Lalmuanawma et al. (2020) though, there is a report that AI tools are being used but this could not be confirmed in the references provided by the authors]. There are however promises [see Kricka et al. (2020), for example] that data gathered through these applications could be exploited for developing AI-based tools in the future.

In addition, AI techniques have been used to develop applications for managing and control the spread of the COVID-19 pandemic. Technologies, such as drones and surveillance cameras equipped with AI-based models for enforcing social isolation (Ahmed et al., 2021), have been reported. As for the impacts of the various social control strategies, the reader is invited to consult (Chang et al., 2020; Hellewell et al., 2020; Kissler et al., 2020; Koo et al., 2020) for further information.

DISCUSSION

Promising and encouraging AI-based techniques and frameworks for the detection, diagnosis, and epidemiological predictions, forecasting and social control of COVID-19 have

been proposed in the fight against the disease. For these AI techniques to gain wide acceptance and use in practical clinical settings however, there would need to be a framework on how these models would be incorporated into clinical practice systems. Importantly, the models, which have been developed with mostly limited amount of data using different algorithms and architectures, would need to be trained and validated with larger amount of data and issues such as overfitting and biasness should be appropriately addressed. Evaluating and comparing the performance of AI models is crucial but challenging. This is partly due to complex relationships amongst the choice of algorithms, architectures, hyperparameters, and the quality and amount of data used for these models.

In addition, many (if not the majority) of the proposed or developed AI-based techniques and models for COVID-19 diagnosis and epidemiological forecasting have not been externally evaluated and caution must be exercised in the interpretation of these results. Indeed, despite the urgency for the publication of research results during the COVID-19 pandemic, these models cannot be used in clinical practice in their current form as critical review and external assessment of the techniques and models with multi-center datasets should be carried out.

To illustrate the scale of the lack of external evaluation problem with an example, consider a recent study presented in Yan et al. (2020b) where the authors have used blood samples from 485 infected patients in the region of Wuhan, China, to identify crucial predictive biomarkers of disease mortality using AI-based tools. In this relatively simple severity and outcome prediction task, and with a small validation sample size and no external model evaluation, the authors have used the XGBoost classifier method to identify three biomarkers [namely, lactic dehydrogenase (LDH), lymphocyte count and high-sensitivity C-reactive protein (hs-CRP)] that will allow the prediction of the mortality of COVID-19 patients more than 10 days in advance with reportedly more than 90% accuracy. External evaluation of this result by several other researchers, such as in Barish et al. (2020), Giacobbe (2020), Quanjel et al. (2020), and Dupuis et al. (2021), has shown that the results of Yan et al. (2020b) have limited clinical utility as it was impossible to replicate the findings and arrive at the same conclusion. If a huge external evaluation problem exists even for simpler problems (such as prediction and forecasting problems), one can only imagine the scale of the problem when using AI-based model for more complicated problems such as those involving images (computer vision-related problems).

AI has huge potential in the battle against the COVID-19 pandemic. Despite several AI approaches and techniques proposed for the detection, diagnosis and epidemiological predictions, forecasting, and social control in the combat against the pandemic, successful practical deployments of these AI-based tools have so far been limited. There are challenges that have led to the limited applicability of these AI-based tools. In the following paragraphs, some of these challenges are discussed with some suggestions of how some of these obstacles may be tackled in order to achieve significant progress in battling COVID-19 and future pandemics using AI techniques.

Data Accessibility

One of the key challenges that AI experts have faced during the COVID-19 pandemic is the lack of access to sufficiently large datasets for training and external validation of AI models upon which deployable and successful applications depend. In order to tackle this problem for COVID-19 and future pandemics, healthcare centers would need a dedicated team with expertise in medical data collection, privacy, access, and sharing. In short, data governance frameworks and protocols for pandemics and other emergency times will need to be designed and put in place.

One of the sources of data that has not been taken full advantage of so far for developing AI-based applications and solutions during the COVID-19 pandemic are data from biobanks. Biobanks provide infrastructure for the collection and storage of biomedical data, including data related to health records and lifestyle of participants, with the aim of advancing scientific research and improving healthcare. They are often large databases that can store imaging data, text data from electronic health record (EHR) and lifestyle information, and numerical data obtained by physical measurements of consented participants. Several types of biobanks exist around the world with different population sizes, including genetic banks, blood banks, and tissue banks. These biobanks contain valuable data that can provide insights into how the health of a population develops over years and provide a rich source of data that can be harnessed to unveil complex relationships amongst variables [such as environmental (Wright et al., 2002; Hall et al., 2014), lifestyle choice (Rutten-Jacobs et al., 2018; Said et al., 2018), and genetics (Arnau-Soler et al., 2019; Wang et al., 2019)] that are associated with COVID-19. Biobanking is particularly useful in that it provides a unified data repository with mostly standardized data collecting protocols. In contrast, the hospital data are "messy" due to the nature of data collection and storage across multiple repositories. Examples of biobanks include the Kaiser Permanente's Research Program on Genes, Environment and Health (RPGEH) with 200,000 participants (Kaiser Permanente, 2020), the UK Biobank with 500,000 participants (Biobank, 2014), China Kadoorie Biobank with 500,000 participants (Chen et al., 2005a, 2011), India's Chennai biobank with 500,000 participants (Gajalakshmi et al., 2007), and Biobank of Vanderbilt University Medical Center (BioVU) with over 1.4 million participants (Roden et al., 2008). Not all these biobanks have data of COVID-19 patients. The UK Biobank, one of the largest biobanks in the world in terms of data volume and depth including multi-organ imaging, is an example of one that has been integrated with pre-existing data of COVID-19 patients. UK Biobank's data has been used for research related to COVID-19 [for example, see Armstrong et al. (2020), Atkins et al. (2020), Grant and McDonnell (2020), Hastie et al. (2020), Jimenez-Solem et al. (2020), Kenneth and So (2020), Pereira et al. (2020), Sattar et al. (2020), Toh and Brody (2020), and Zimmerman and Kalra (2020)]. Few AI-based applications, such as in Jimenez-Solem et al. (2020), Kenneth and So (2020), Pereira et al. (2020), Toh and Brody (2020), and Zimmerman and Kalra (2020), exist that have used biobanks' data for their development, and it is likely

that the use of biobanks' data for the development of AI solutions will increase in the near future.

In addition, researchers, such as in Brisimi et al. (2018), Lee et al. (2018), Rieke et al. (2020), Li et al. (2020d), and Xu et al. (2020d), have proposed the use of federated learning (FL) whereby, rather than participating healthcare institutions hand over healthcare data to AI experts to develop AI models, AI experts will handover training algorithms to the healthcare institutions to train their models locally. The AI experts only get the model or model parameters in return—thus, eliminating some of the problems of data governance and privacy associated with data transfer between different parties while giving access to large amount of data. FL is not without its challenges, such as lesser accuracy of the final model (Li et al., 2020d), and the reader is invited to consult (Brisimi et al., 2018; Lee et al., 2018; Li et al., 2020d; Xu et al., 2020d) for further information of this approach.

External Evaluation

Many of the developed AI-based techniques and models for COVID-19 diagnosis and epidemiological forecasting have not been externally evaluated. External model evaluation helps in assessing the generalisability of the predictions on independent datasets and ensures that the model has learnt the underlying features of the process that produces the data rather than "memorized" the features of a particular set of data. For illustration, **Figure 1** shows the steps in developing models using AI algorithms, highlighting the model evaluation stage of the development process.

Many publicly available datasets for COVID-19 diagnosis do not necessarily generalize to the whole population (i.e., they are usually for a specific country or regions of a country or a specific number of hospitals). The implication is that most ML models based on them will be biased (He and Garcia, 2009), which can reduce the performance of the models in practical settings (Chawla et al., 2002) and can promote healthcare inequalities (Petersen et al., 2019). Many mainstream ML algorithms for classification problems, including SVM, decision trees, and nearest neighbor, were developed based on the assumption that the dataset has balanced class distribution (Chen et al., 2005b; Almogahed and Kakadiaris, 2015), resulting in significant error when classifying the minority class. Algorithms that have been developed to overcome this problem algorithmically or at data level can be found in Hart (1968), Kubat and Matwin (1997), Laurikkala (2001), Barandela et al. (2003), Oh (2011), and Almogahed and Kakadiaris (2015). In addition, it is important for published research to report the pre-processing, the cleaning and the feature engineering steps applied to the data used for developing AI-based solutions.

AI Regulatory Landscape

Recently, frameworks for strong regulatory and ethical requirements of AI-based clinical utility tools are being developed but significant hurdles still persist (Petersen et al., 2019). Many AI experts are unaware of the regulatory landscape governing the development of AI tools in healthcare and have not considered this matter in their development. Proof of model performance is not sufficient. Issues, such as model biasness, safety, effectiveness, and benefit-versus-harm analysis have mostly been ignored by developers of many critical AI-based healthcare technologies.

Collaboration Between AI Experts and Clinicians

While AI provides the opportunity to reduce the time for disease diagnosis and improve accuracy, the workload of healthcare professionals is very high during a pandemic. The impact of this includes the difficulty for healthcare professionals to be up to date with the progress being made in areas relevant to their work. It has also hindered their contribution toward that progress. The absence and lack of engagement of clinicians to contribute and review research results during the COVID-19 pandemic has contributed to the limited impact, reliability and clinical utility of many of these research findings. The COVID-19 pandemic has highlighted the importance of domain specific knowledge in AI. It is not sufficient for clinicians to handover data to AI experts who understand how to develop and use classical AI algorithms. Rather, it is important for the clinicians to work with AI experts to help them understand the context of the solutions being developed, to help them interpret the results from those solutions, and to guide them on how those solutions could be used and integrated into existing clinical healthcare pathways or workflows. Thus, multidisciplinary research collaborations will no doubt accelerate the translation of AI research into practical solutions in healthcare and funding bodies could help in this by ensuring multidisciplinary collaborations as a condition for funding. An important lesson that should be learnt in using AI techniques for combating COVID-19 and future pandemics is that the applicability of these techniques is limited if AI experts work in isolation. Important progress in healthcare using AI technologies can be achieved only in a multi-disciplinary setting where clinicians, epidemiologists, computer scientists, software developers, AI experts and others work together to achieve the common goal of improving healthcare services through innovative technologies.

Public Engagements Over Privacy Concerns

While AI-based technologies embedded in digital systems have played a role in controlling the spread of COVID-19 and the general management of the disease by many governments across the world, the concerns of the general public over privacy have had an impact on the acceptance of many of these technologies and even other potential applications. Consider, for instance, the contract tracing mobile applications that many governments have deployed as a tool for controlling COVID-19, concerns over the possibility that data gathered through these applications could be exploited for other purposes has meant that the general public have been very reluctant in using them (Clark et al., 2020; Lewis, 2020). It has also meant that tools applicable to one country (such as China's use of facial recognition cameras to track infected patients or the linking of the national medical insurance database with the immigration and custom database in order to

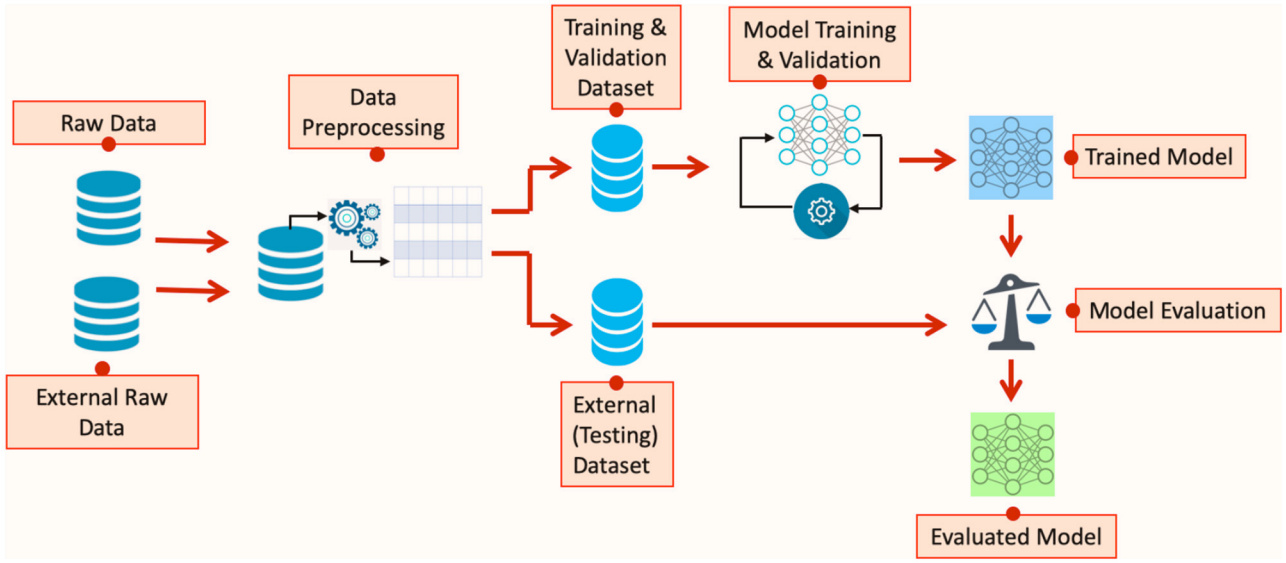


FIGURE 1 | Steps in developing models using AI algorithms.

inform the healthcare practitioners in Taiwan of the travel history of patients) may not be applicable to others.

A framework that will ensure transparency over the legal basis of data use, that data collection is safe and that there are controls and mechanisms to protect misuse of data is critical now and in future. Thus, while it is essential to gather data to address the challenges posed by a pandemic, the authorities would need to do work on gaining the trust of the population through effective engagements with all stakeholders on the mechanisms that would be in place in order to protect privacy and data misuse.

Potential Misuse of AI Applications

One of the dangers of reliance on AI applications during a pandemic is the potential for misuse. Medical imaging involves several stages including image acquisition, reconstruction, and transmission for storage using Digital Imaging and Communications in Medicine (DICOM) protocol. A cyber-attack could disrupt the use of the devices such as CT devices that can be critical for disease diagnosis during a pandemic (Mahler et al., 2018). With the advent of advanced AI techniques such as generative adversarial network in medical imaging (Yi et al., 2019), one can envisage sophisticated scenarios where AI technologies are used for cyber-attacks that can alter the output of imaging modalities (for instance, by removing or adding a tissue to medical images) altering the results of medical examination, which could lead to fatal consequences. With increasing cyber-attack activities during COVID-19 (Lallie et al., 2021; Muthuppalaniappan and Stevenson, 2021), healthcare providers must be prepared for preventing the occurrence and also detecting and mitigating the impacts that these type of AI attacks will cause when they occur.

In addition, while FL can resolve data governance issues, it does not necessarily guarantee data security on its own as it may be possible to reconstruct parts of the training dataset from the weights on decentralized computer nodes (Kaissis et al., 2020). This possibility can allow attackers to steal sensitive personal information in the training datasets from the nodes and even reconstruct medical images with high degree of accuracy (Fredrikson et al., 2015; Hitaj et al., 2017), leading to patient confidentiality violations.

Problems associated with data imbalance, variability and incompleteness resulting from the use of datasets that are not accurate representation of the population on which AI models was built for can lead to biased treatment of certain ethnic, sex, age, and other groups. In many cases, these data biases are often introduced inadvertently by AI algorithm developers but unscrupulous individuals can take advantage of this to exacerbate bias from cultural prejudices and increase disparities in delivering healthcare services. Moreover, misuse of AI models can also result when the datasets used for model training do not take into account future use-case conditions; for example, radiologists can easily adapt to change in MRI field strength and breathing motion artifacts but these changes will affect the performance of AI models unless they have been specifically allowed for during the training of the models (Brady and Neri, 2020).

These issues of misuse of AI as highlighted here show that it is important to provide safeguards to ensure that new AI solutions during a pandemic are assessed before being deployed at scale. It is important to emphasize that these challenges posed by AI are not necessarily associated with the limitations of AI *per se* (Rodriguez et al., 2018). Rather, they apply to particular use-cases and emphasize the importance of understanding the relationships that AI models use in arriving at their predictions. As such, guards against spurious predictions must be put in place in order to limit data misuse.

We finish by noting that, recently, there have been several promising initiatives from key players (e.g., government bodies,

commercial institutions, and policy makers) to collect and manage data in order to address or alleviate some of the problems highlighted in this paper. We mention a few of them in the following paragraphs.

In the United Kingdom, NHSX, the government's unit responsible for developing and setting national policy on digital, data and technology for National Health Service (NHS), has developed the National COVID-19 Chest Imaging Database (NCCID) in order to collect patient data and facilitate research and the development and validation of technologies that are promising for improving COVID-19 care (NHSX, 2021). The categories of collected data include chest X-ray, CT, and MR images including those performed in the 3 years preceding the first COVID-19-related imaging study, routine demographic data, biochemical and hematological data, and outcome data.

In the European Union (EU), SoBigData is a research initiative under the EU's Horizon 2020 programme (Grant No. 654024 and 871042) which provides an integrated ecosystem of "big data" for ethnic-sensitive scientific discoveries in multiple fields including mathematics, ICT, and human, social, and economic sciences (SoBigData.eu, 2021). The idea is to promote repeatable and open science by meeting the data and infrastructural needs of researchers while also ensuring that users' data are gathered for specific application and timebound (e.g., relates to dealing with the COVID-19 pandemic only and data will be deleted afterwards), the data cannot be shared without consent, and the data must be of direct benefit to the users whose data were gathered. Another initiative in the EU, The Confederation of Laboratories for Artificial Intelligence (CLAIRE) (CLAIRE, 2021), has warned that it is very likely that our societies will be confronted with other crises at a scale similar to COVID-19 in the not-so-far future and have outlined an European approach with the recommendation that standards and frameworks that would facilitate the development of efficient management of medical data that will not erode human dignity must be developed (Ishmaev et al., 2021).

In the United States, three national institutions namely, National Center for Advancing Translational Sciences (NCATS), Clinical and Translational Science Awards (CTSA) Program and Institutional Development Award Networks for Clinical and Translational Research (IDeA-CTR), have partnered to form the National COVID Cohort Collaborative (N3C) in an attempt to enable collaborators to contribute and use COVID-19 clinical data for scientific research that will have impact in the battle against the pandemic (NCATS-US, 2021). As at the time of writing this paper, the data of more than 950,000 COVID-19 positive patients are available from N3C for researchers to examine associations between COVID-19 patient outcomes and other determinants of health and, at least, 144 projects are already on-going for this purpose. Interestingly, in addition to patient data being de-identified for privacy reasons, this cloud-based data repository consists of synthetic (that is, computationally derived) data that statistically resemble original patient information but are not the actual data of the patients, adding another layer of privacy protection for patients.

The summary of the key messages and the main lessons learnt on the application of AI-based techniques and frameworks for the detection, diagnosis and epidemiological predictions, forecasting and social control of COVID-19 is as follows:

- We recommend that healthcare centers set up dedicated teams with expertise in medical data collection, privacy, access and sharing, and data governance frameworks and protocols for pandemics and other emergency times.
- External model evaluation is important to avoid the problems associated with model overfitting and biasness, such as arriving at clinically unusable solutions or introducing inequalities in health and healthcare. We recommend the establishment of independent units at national level or through international collaboration with the goal of assessing and validating AI applications developed for healthcare during pandemics before such applications are adopted and scaled up.
- The regulatory landscape (covering issues such as safety, effectiveness and benefit-versus-harm analysis) governing the development of AI tools in healthcare need to be accessible and understandable to AI experts. We recommend that professional bodies that will oversee certification programmes for AI experts working in healthcare be introduced to ensure that, through continuing professional development, these professionals adhere to common ethical standards and are aware of the current ethical and social issues related to their work.
- The COVID-19 pandemic has highlighted the importance of domain specific knowledge in AI, and multidisciplinary research collaborations will only accelerate the translation of AI research into practical and useful solutions in healthcare. In funding AI projects, we recommend that research fund awarding bodies should make the collaboration between AI scientists and domain specific experts a condition for grant awards.
- In order to gain the trust of the population in terms of data collection, privacy and protection, we recommend that all stakeholders work together in the development of a data use and sharing framework that will ensure effective data management is in place for the development and advancement of AI applications in healthcare.

CONCLUSION

In this paper, AI techniques that are being used in the detection, diagnosis and epidemiological predictions, forecasting and social control for combating COVID-19 have been discussed. While AI has huge potential in the battle against COVID-19, the successful practical deployments of these AI-based tools have so far been limited due to challenges such as limited data accessibility, need for external evaluation of AI models, lack of awareness of AI experts of the regulatory landscape governing the deployment of AI tools in healthcare, the need for clinicians and other experts to work with AI experts in a multidisciplinary context and the need to address public concerns over data collection, privacy and protection. Overcoming these challenges will lead to significant progress in battling COVID-19 and future pandemics using AI techniques.

AUTHOR CONTRIBUTIONS

MA drafted the first version of the manuscript. MA and SP contributed to the content and writing of the final version. Both authors contributed to the article and approved the submitted version.

FUNDING

MA and SP acknowledge support from the CAP-AI programme (led by Capital Enterprise in partnership with Barts Health NHS Trust and Digital Catapult and funded by the European Regional Development Fund and Barts Charity) and Health Data Research UK (HDR UK—an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities; www.hdruk.ac.uk). SP acknowledges support from the National Institute for Health Research (NIHR) Biomedical Research Center at Barts and from the SmartHeart EPSRC programme grant (www.nihr.ac.uk; EP/P001009/1). SP has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825903 (euCanSHare project).

REFERENCES

Ahmed, I., Ahmad, M., Rodrigues, J., Jeon, G., and Din, S. (2021). A deep learning-based social distance monitoring framework for COVID-19. *Sustain. Cities Soc.* 65:102571. doi: 10.1016/j.scs.2020.102571

Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., et al. (2020). Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1,014 cases. *Radiology* 2020:200642. doi: 10.1148/radiol.2020200642

Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H. F., and van der Schaar, M. (2019). Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. *PLoS ONE* 14:e0213653. doi: 10.1371/journal.pone.0213653

Albahri, O. S., Zaidan, A. A., Albahri, A. S., Zaidan, B. B., Abdulkareem, K. H., Al-Qaysi, Z. T., et al. (2020). Systematic review of artificial intelligence techniques in the detection and classification of COVID-19 medical images in terms of evaluation and benchmarking: taxonomy analysis, challenges, future solutions and methodological aspects. *J. Infect. Public Health* 13, 1381–1396. doi: 10.1016/j.jiph.2020.06.028

Almogahed, B. A., and Kakadiaris, I. A. (2015). NEATER: filtering of over-sampled data using non-cooperative game theory. *Soft Comput*. 19, 3301–3322. doi: 10.1007/s00500-014-1484-5

Al-Qaness, M. A. A., Ewees, A. A., Fan, H., and Abd El Aziz, M. (2020). Optimization method for forecasting confirmed cases of COVID-19 in China. *J. Clin. Med*. 9:E674. doi: 10.3390/jcm9030674

Alsalem, M. A., Zaidan, A. A., Zaidan, B. B., Albahri, O. S., Alamoodi, A. H., Albahri, A. S., et al. (2019). Multiclass benchmarking framework for automated acute Leukaemia detection and classification based on BWM and group-VIKOR. *J. Med. Syst.* 43:212. doi: 10.1007/s10916-019-1338-x

Alsalem, M. A., Zaidan, A. A., Zaidan, B. B., Hashim, M., Albahri, O. S., Albahri, A. S., et al. (2018). Systematic review of an automated multiclass detection and classification system for acute Leukaemia in terms of evaluation and benchmarking, open challenges, issues and methodological aspects. *J. Med. Syst.* 42:204. doi: 10.1007/s10916-018-1064-9

Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., and Garry, R. F. (2020). The proximal origin of SARS-CoV-2. *Nat. Med.* 26, 450–452. doi: 10.1038/s41591-020-0820-9

Anderson, R. M., and May, R. M. (1979). Population biology of infectious diseases: Part, I. *Nature* 280, 361–367. doi: 10.1038/280361a0

Apostolopoulos, I. D., and Mpesiana, T. A. (2020). Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. *Phys. Eng. Sci. Med*. 2020:1. doi: 10.1007/s13246-020-00865-4

Ardakani, A. A., Kanafi, A. R., Acharya, U. R., Khadem, N., and Mohammadi, A. (2020). Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: results of 10 convolutional neural networks. *Comput. Biol. Med.* 2020:103795. doi: 10.1016/j.compbiomed.2020.103795

Armstrong, J., Rudkin, J. K., Allen, N., Crook, D. W., Wilson, D. J., Wyllie, D. H., et al. (2020). Dynamic linkage of covid-19 test results between public health england's second generation surveillance system and uk biobank. *Microb. Genom.* 6:397. doi: 10.1099/mgen.0.000397

Arnau-Soler, A., Macdonald-Dunlop, E., Adams, M. J., Clarke, T.-K., MacIntyre, D. J., Milburn, K., et al. (2019). Genome-wide by environment interaction studies of depressive symptoms and psychosocial stress in UK Biobank and Generation Scotland. *Transl. Psychiatry* 9, 1–13. doi: 10.1038/s41398-018-0360-y

Atkins, J. L., Masoli, J. A. H., Delgado, J., Pilling, L. C., Kuo, C.-L., Kuchel, G. A., et al. (2020). Preexisting comorbidities predicting COVID-19 and mortality in the UK Biobank Community Cohort. *J. Gerontol. Ser. A* 75, 2224–2230. doi: 10.1093/gerona/glaa183

Bai, W., Suzuki, H., Huang, J., Francis, C., Wang, S., Tarroni, G., et al. (2020). A population-based phenome-wide association study of cardiac and aortic structure and function. *Nat. Med.* 2020, 1–9. doi: 10.1038/s41591-020-1009-y

Bandyopadhyay, S. K., and Dutta, S. (2020). Machine learning approach for confirmation of covid-19 cases: positive, negative, death and release. *MedRxiv*. doi: 10.2196/preprints.19526

Bansal, A., Padappayil, R. P., Garg, C., Singal, A., Gupta, M., and Klein, A. (2020). Utility of artificial intelligence amidst the COVID 19 pandemic: a review. *J. Med. Syst*. 44, 1–6. doi: 10.1007/s10916-020-01617-3

Barandela, R., Sánchez, J. S., Garca, V., and Rangel, E. (2003). Strategies for learning in class imbalance problems. *Pattern Recognit*. 36, 849–851. doi: 10.1016/S0031-3203(02)00257-1

Barish, M., Bolourani, S., Lau, L. F., Shah, S., and Zanos, T. P. (2020). External validation demonstrates limited clinical utility of the interpretable mortality prediction model for patients with COVID-19. *Nat. Mach. Intell*. 2020, 1–3. doi: 10.1038/s42256-020-00254-2

Barlow, H. B. (1989). Unsupervised learning. *Neural Comput*. 1, 295–311. doi: 10.1162/neco.1989.1.3.295

Bennett, C. C., and Hauser, K. (2013). Artificial intelligence framework for simulating clinical decision-making: a Markov decision process approach. *Artif. Intell. Med*. 57, 9–19. doi: 10.1016/j.artmed.2012.12.003

Bernheim, A., Mei, X., Huang, M., Yang, Y., Fayad, Z. A., Zhang, N., et al. (2020). Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection. *Radiology* 2020:200463. doi: 10.1148/radiol.2020200463

Biobank, U. K. (2014). *About UK Biobank*. Available online at: https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us (accessed January 12, 2021).

Blasiak, A., Khong, J., and Kee, T. (2020). CURATE. AI: optimizing personalized medicine with artificial intelligence. *SLAS Technol. Transl. Life Sci. Innov*. 25, 95–105. doi: 10.1177/2472630319890316

Brady, A. P., and Neri, E. (2020). Artificial intelligence in radiology—ethical considerations. *Diagnostics* 10:231. doi: 10.3390/diagnostics10040231

Bragazzi, N. L., Dai, H., Damiani, G., Behzadifar, M., Martini, M., and Wu, J. (2020). How big data and artificial intelligence can help better manage the COVID-19 pandemic. *Int. J. Environ. Res. Public Health* 17:3176. doi: 10.3390/ijerph17093176

Breiman, L. (2001). Random forests. *Mach Learn* 45, 5–32. doi: 10.1023/A:1010933404324

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. Boca Raton, FL: CRC Press.

Brisimi, T. S., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I. C., and Shi, W. (2018). Federated learning of predictive models from federated electronic health records. *Int. J. Med. Inform.* 112, 59–67. doi: 10.1016/j.ijmedinf.2018.01.007

Brown, C., Chauhan, J., Grammenos, A., Han, J., Hasthanasombat, A., Spathis, D., et al. (2020). Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data. *ArXiv Prepr ArXiv200605919*. doi: 10.1145/3394486.3412865

Brüssow, H. (2020). The novel coronavirus – a snapshot of current knowledge. *Microb. Biotechnol.* 13, 607–612. doi: 10.1111/1751-7915.13557

Bullock, J., Pham, K. H., Lam, C. S. N., Luengo-Oroz, M. (2020). Mapping the landscape of artificial intelligence applications against COVID-19. *ArXiv Prepr ArXiv200311336*. doi: 10.1613/jair.1.12162

C Disease Control (2020). *Coronavirus Disease 2019 (COVID-19): Frequently Asked Questions*. Webpage.

Carrillo-Larco, R. M., and Castillo-Cara, M. (2020). Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: an unsupervised machine learning approach. *Wellcome Open Res.* 5:56. doi: 10.12688/wellcomeopenres.15819.1

Casigliani, V., De Nard, F., De Vita, E., Arzilli, G., Grosso, F. M., Quattrone, F., et al. (2020). Too much information, too little evidence: is waste in research fuelling the covid-19 infodemic. *BMJ* 370:m2672. doi: 10.1136/bmj.m2672

Chan, J. F.-W., Yuan, S., Kok, K.-H., To, K. K.-W., Chu, H., Yang, J., et al. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet* 395, 514–523. doi: 10.1016/S0140-6736(20)30154-9

Chang, S. L., Harding, N., Zachreson, C., Cliff, O. M., and Prokopenko, M. (2020). Modelling transmission and control of the COVID-19 pandemic in Australia. *ArXiv Prepr ArXiv200310218*. doi: 10.1038/s41467-020-19393-6

Chapelle, O., Scholkopf, B., and Zien, A. (2009). Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Trans. Neural Netw.* 20:542. doi: 10.1109/TNN.2009.2015974

Chawla, N., V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Chen, J., Li, K., Zhang, Z., Li, K., and Yu, P. S. (2020c). A survey on applications of artificial intelligence in fighting against covid-19. *ArXiv [Preprint]. ArXiv: 200702202*. Available online at: https://arxiv.org/abs/2007.02202

Chen, J., Wu, L., Zhang, J., Zhang, L., Gong, D., Zhao, Y., et al. (2020d). Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study. *MedRxiv*. doi: 10.1101/2020.02.25.20021568

Chen, J. J., Tsai, C. A., Young, J. F., and Kodell, R. L. (2005b). Classification ensembles for unbalanced class sizes in predictive toxicology. *SAR QSAR Environ. Res.* 16, 517–529. doi: 10.1080/10659360500468468

Chen, Q., Allot, A., and Lu, Z. (2020a). Keep up with the latest coronavirus research. *Nature* 2020:193. doi: 10.1038/d41586-020-00694-1

Chen, S., Yang, J., Yang, W., Wang, C., and Bärnighausen, T. (2020e). COVID-19 control in China during mass population movements at New Year. *Lancet* 395, 764–766. doi: 10.1016/S0140-6736(20)30421-9

Chen, T., and Guestrin, C. (2016). "Xgboost: a scalable tree boosting system," in *Proc. 22nd acm sigkdd Int. Conf. Knowl. Discov. data Min.* (San Francisco, CA), 785–794. doi: 10.1145/2939672.2939785

Chen, W.-H., Strych, U., Hotez, P. J., and Bottazzi, M. E. (2020b). The SARS-CoV-2 vaccine pipeline: an overview. *Curr. Trop. Med. Rep.* 6, 1–4. doi: 10.1007/s40475-020-00201-6

Chen, Z., Chen, J., Collins, R., Guo, Y., Peto, R., Wu, F., et al. (2011). China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int. J. Epidemiol.* 40, 1652–1666. doi: 10.1093/ije/dyr120

Chen, Z., Lee, L., Chen, J., Collins, R., Wu, F., Guo, Y., et al. (2005a). Cohort profile: the Kadoorie study of chronic disease in China (KSCDC). *Int. J. Epidemiol.* 34, 1243–1249. doi: 10.1093/ije/dyi174

Cheng, C., Barceló J., Hartnett, A. S., Kubinec, R., and Messerschmidt, L. (2020). COVID-19 government response event dataset (CoronaNet v. 1.0). *Nat. Hum. Behav.* 4, 756–68. doi: 10.1038/s41562-020-0909-7

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: encoder-decoder approaches. *ArXiv Prepr ArXiv14091259*. doi: 10.3115/v1/W14-4012

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *ArXiv [Preprint]. ArXiv: 14123555*. Available online at: https://arxiv.org/abs/1412.3555

Chung, M., Bernheim, A., Mei, X., Zhang, N., Huang, M., Zeng, X., et al. (2020). CT imaging features of 2019 novel coronavirus (2019-nCoV). *Radiology* 2020:200230. doi: 10.1148/radiol.2020200230

CLAIRE (2021). *The Confederation of Laboratories for Artificial Intelligence (CLAIRE) in Europe*. Available online at: https://claire-ai.org/ (accessed March 24, 2021).

Clark, E., Chiao, E. Y., and Amirian, E. S. (2020). Why contact tracing efforts have failed to curb coronavirus disease 2019 (covid-19) transmission in much of the united states. *Clin Infect Dis.* 2020:ciaa1155. doi: 10.1093/cid/ciaa1155

Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T. Q., and Ghassemi, M. (2020). Covid-19 image data collection: prospective predictions are the future. *ArXiv [Preprint]. ArXiv: 200611988*. Available online at: https://arxiv.org/abs/2006.11988

Colubri, A., Silver, T., Fradet, T., Retzepi, K., Fry, B., and Sabeti, P. (2016). Transforming clinical data into actionable prognosis models: machine-learning framework and field-deployable app to predict outcome of Ebola patients. *PLoS Negl. Trop. Dis.* 10:e0004549. doi: 10.1371/journal.pntd.0004549

Cooper, I., Mondal, A., and Antonopoulos, C. G. (2020). A SIR model assumption for the spread of COVID-19 in different communities. *Chaos Solitons Fractals* 139:110057. doi: 10.1016/j.chaos.2020.110057

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach Learn.* 20, 273–297. doi: 10.1007/BF00994018

Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* 13, 21–27. doi: 10.1109/TIT.1967.1053964

Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis.* 20, 533–534. doi: 10.1016/S1473-3099(20)30120-1

Dupuis, C., De Montmollin, E., Neuville, M., Mourvillier, B., Ruckly, S., and Timsit, J. F. (2021). Limited applicability of a COVID-19 specific mortality prediction rule to the intensive care setting. *Nat. Mach. Intell.* 2020, 1–3. doi: 10.1038/s42256-020-00252-4

Fang, Y., Zhang, H., Xie, J., Lin, M., Ying, L., Pang, P., et al. (2020). Sensitivity of chest CT for COVID-19: comparison to RT-PCR. *Radiology* 2020:200432. doi: 10.1148/radiol.2020200432

Fong, S. J., Li, G., Dey, N., Gonzalez-Crespo, R., and Herrera-Viedma, E. (2020). Finding an accurate early forecasting model from small dataset: a case of 2019-nCoV novel coronavirus outbreak. *Int. J. Interact. Multimed. Artif. Intell.* 6, 132–140. doi: 10.9781/ijimai.2020.02.002

Fredrikson, M., Jha, S., and Ristenpart, T. (2015). "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur* (Denver, CO), 1322–1333. doi: 10.1145/2810103.2813677

Gajalakshmi, V., Peto, R., Kanimozhi, V. C., Whitlock, G., and Veeramani, D. (2007). Cohort profile: the Chennai prospective study of mortality among 500,000 adults in Tamil Nadu, South India. *Int. J. Epidemiol.* 36, 1190–1195. doi: 10.1093/ije/dym091

Ghoshal, B., and Tucker, A. (2020). Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. *ArXiv [Preprint]. ArXiv: 200310769*. Available online at: https://arxiv.org/abs/2003.10769

Giacobbe, D. R. (2020). Clinical interpretation of an interpretable prognostic model for patients with COVID-19. *Nat. Mach. Intell.* 2020:1. doi: 10.1038/s42256-020-0207-0

Glasziou, P. P., Sanders, S., and Hoffmann, T. (2020). Waste in covid-19 research. *BMJ* 369:m1847. doi: 10.1136/bmj.m1847

Goh, K. J., Kalimuddin, S., and Chan, K. S. (2020). Rapid progression to acute respiratory distress syndrome: review of current understanding of critical illness from coronavirus disease 2019 (COVID-19) infection. *Ann. Acad. Med. Singapore* 49, 108–118. doi: 10.47102/annals-acadmedsg.202057

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Adv. Neural Inf. Process Syst.* 27, 2672–2680. Available online at: https://arxiv.org/abs/1406.2661

Gozes, O., Frid-Adar, M., Greenspan, H., Browning, P. D., Zhang, H., Ji, W., et al. (2020). Rapid ai development cycle for the coronavirus (covid-19) pandemic: initial results for automated detection & patient monitoring using deep learning ct image analysis. *ArXiv [Preprint]. ArXiv: 200305037*. Available online at: https://arxiv.org/abs/2003.05037

Grant, W. B., and McDonnell, S. L. (2020). Letter in response to the article: vitamin D concentrations and COVID-19 infection in UK biobank (Hastie et al.). *Diabetes Metab. Syndr*. 15, 643–644. doi: 10.1016/j.dsx.2020.05.046

Hack, J. J., and Papka, M. E. (2020). The US high-performance computing consortium in the fight against COVID-19. *Comput. Sci. Eng*. 22, 75–80. doi: 10.1109/MCSE.2020.3019744

Hall, M. A., Dudek, S. M., Goodloe, R., Crawford, D. C., Pendergrass, S. A., Peissig, P., et al. (2014). Environment-wide association study (EWAS) for type 2 diabetes in the Marshfield Personalized Medicine Research Project Biobank. Biocomput. *World Sci.* 2014, 200–211. doi: 10.1142/9789814583220_0020

Hart, P. (1968). The condensed nearest neighbor rule (Corresp). *IEEE Trans. Inf. Theory* 14, 515–516. doi: 10.1109/TIT.1968.1054155

Hassanien, A. E., Mahdy, L. N., Ezzat, K. A., Elmousalami, H. H., and Ella, H. A. (2020). Automatic x-ray covid-19 lung image classification system based on multi-level thresholding and support vector machine. *MedRxiv*. doi: 10.1101/2020.03.30.20047787

Hastie, C. E., Mackay, D. F., Ho, F., Celis-Morales, C. A., Katikireddi, S. V., Niedzwiedz, C. L., et al. (2020). Vitamin D concentrations and COVID-19 infection in UK Biobank. *Diabetes Metab. Syndr. Clin. Res. Rev*. 14, 561–565. doi: 10.1016/j.dsx.2020.04.050

He, H., and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng*. 21, 1263–1284. doi: 10.1109/TKDE.2008.239

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit* (Las Vegas, NV), 770–778. doi: 10.1109/CVPR.2016.90

Hellewell, J., Abbott, S., Gimma, A., Bosse, N. I., Jarvis, C. I., Russell, T. W., et al. (2020). Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *Lancet Glob. Heal*. 8, 488–496. doi: 10.1101/2020.02.08.20021162

Hemdan, E. E.-D., Shouman, M. A., and Karar, M. E. (2020). Covidx-net: a framework of deep learning classifiers to diagnose covid-19 in x-ray images. *ArXiv [Preprint]. ArXiv: 200311055*. Available online at: https://arxiv.org/abs/2003.11055

Hitaj, B., Ateniese, G., and Perez-Cruz, F. (2017). "Deep models under the GAN: information leakage from collaborative deep learning," in *Proc. 2017 ACM SIGSAC Conf. Comput. Commun. Secur*. (Dallas, TX), 603–618. doi: 10.1145/3133956.3134012

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput*. 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Hope, T., Portenoy, J., Vasan, K., Borchardt, J., Horvitz, E., Weld, D. S., et al. (2020). SciSight: combining faceted navigation and research group detection for COVID-19 exploratory scientific search. *ArXiv Prepr ArXiv200512668*. doi: 10.1101/2020.05.23.112284

Hopfield, J. J. (1988). Artificial neural networks. *IEEE Circuits Devices Mag*. 4, 3–10. doi: 10.1109/101.8118

Hosmer, D. W. Jr., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression. Vol. 398*. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9781118548387

Hu, Z., Ge, Q., Li, S., Boerwinkle, E., Jin, L., and Xiong, M. (2020). Forecasting and evaluating multiple interventions for COVID-19 worldwide. *Front. Artif. Intell*. 3:41. doi: 10.3389/frai.2020.00041

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., et al. (2020a). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 395, 497–506. doi: 10.1016/S0140-6736(20)30183-5

Huang, C.-J., Chen, Y.-H., Ma, Y., and Kuo, P.-H. (2020b). Multiple-input deep convolutional neural network model for covid-19 forecasting in china. *MedRxiv*. doi: 10.1101/2020.03.23.20041608

Imran, A., Posokhova, I., Qureshi, H. N., Masood, U., Riaz, S., Ali, K., et al. (2020). AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples *via* an app. *ArXiv Prepr ArXiv200401275*. doi: 10.1016/j.imu.2020.100378

Iqbal, M. Z., and Faiz, M. F. I. (2020). Active Surveillance for COVID-19 through artificial in-telligence using concept of real-time speech-recognition mobile application to analyse cough sound. *arXiv*. doi: 10.31219/osf.io/cev6x

Isham, V., and Medley, G. (1996). *Models for Infectious Human Diseases: Their Structure and Relation to Data. Vol. 6*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511662935

Ishmaev, G., Dennis, M., and van den Hoven, M. J. (2021). Ethics in the COVID-19 pandemic: myths, false dilemmas, and moral overload. *Ethics Inf. Technol.* 2021, 1–16. doi: 10.1007/s10676-020-09568-6

Jain, A. K., Mao, J., and Mohiuddin, K. M. (1996). Artificial neural networks: a tutorial. *Computer* 29, 31–44. doi: 10.1109/2.485891

Jamshidi, M., Lalbakhsh, A., Talla, J., Peroutka, Z., Hadjilooei, F., Lalbakhsh, P., et al. (2020). Artificial intelligence and COVID-19: deep learning approaches for diagnosis and treatment. *IEEE Access* 8, 109581–109595. doi: 10.1109/ACCESS.2020.3001973

Jana, S., and Bhaumik, P. (2020). A multivariate spatiotemporal spread model of COVID-19 using ensemble of ConvLSTM networks. *MedRxiv* (2020).

Jang, J.-S. (1993). ANFIS: adaptive-network-based fuzzy inference system. *IEEE Trans. Syst. Man. Cybern*. 23, 665–685. doi: 10.1109/21.256541

Janiaud, P., Axfors, C., Van't Hooft, J., Saccilotto, R., Agarwal, A., Appenzeller-Herzog, C., et al. (2020). The worldwide clinical trial research response to the COVID-19 pandemic-the first 100 days. *F1000Research* 9:2. doi: 10.12688/f1000research.26707.2

Jewell, N. P., Lewnard, J. A., and Jewell, B. L. (2020). Predictive mathematical models of the COVID-19 pandemic: underlying principles and value of projections. *JAMA* 323, 1893–1894. doi: 10.1001/jama.2020.6585

Jiang, X., Coffee, M., Bari, A., Wang, J., Jiang, X., Huang, J., et al. (2020a). Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. *C Comput. Mater. Contin.* 63, 537–551. doi: 10.32604/cmc.2020.010691

Jiang, Z., Hu, M., Gao, Z., Fan, L., Dai, R., Pan, Y., et al. (2020b). Detection of respiratory infections using RGB-infrared sensors on portable device. *IEEE Sens. J.* 20, 13674–13681. doi: 10.1109/JSEN.2020.3004568

Jimenez-Solem, E., Petersen, T. S., Hansen, C., Hansen, C., Lioma, C., Igel, C., et al. (2020). Developing and validating COVID-19 adverse outcome risk prediction models from a bi-national European Cohort of 5,594 patients. *MedRxiv*. doi: 10.1101/2020.10.06.20207209

Jin, C., Chen, W., Cao, Y., Xu, Z., Zhang, X., Deng, L., et al. (2020b). Development and evaluation of an AI system for COVID-19 diagnosis. *Nat. Commun.* 11:20039834. doi: 10.1101/2020.03.20.20039834

Jin, Y., Yang, H., Ji, W., Wu, W., Chen, S., Zhang, W., et al. (2020a). Virology, epidemiology, pathogenesis, and control of COVID-19. *Viruses* 12:372. doi: 10.3390/v12040372

John, M., and Shaiba, H. (2019). Main factors influencing recovery in MERS Co-V patients using machine learning. *J. Infect. Public Health* 12, 700–704. doi: 10.1016/j.jiph.2019.03.020

Johnson, K. W., Shameer, K., Glicksberg, B. S., Readhead, B., Sengupta, P. P., Björkegren, J. L. M., et al. (2017). Enabling precision cardiology through multiscale biology and systems medicine. *JACC Basic Transl. Sci.* 2, 311–327. doi: 10.1016/j.jacbts.2016.11.010

Kaiser Permanente (2020). *Kaiser Permanente: Research Program on Genes, Environment and Health*. Available online at: https://divisionofresearch.kaiserpermanente.org/genetics/rpgeh/rpgehabout (accessed November 1, 2020).

Kaissis, G. A., Makowski, M. R., Rückert, D., and Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach Intell.* 2, 305–311. doi: 10.1038/s42256-020-0186-1

Kandel, N., Chungong, S., Omaar, A., Xing, J. (2020). Health security capacities in the context of COVID-19 outbreak: an analysis of International Health Regulations annual report data from 182 countries. *Lancet* 395, 1047–1053. doi: 10.1016/S0140-6736(20)30553-5

Kang, H., Xia, L., Yan, F., Wan, Z., Shi, F., Yuan, H., et al. (2020). Diagnosis of coronavirus disease 2019 (covid-19) with structured latent multi-view representation learning. *IEEE Trans. Med. Imaging*. 39, 2606–2614. doi: 10.1109/TMI.2020.2992546

Kanne, J. P. (2020). Chest CT findings in 2019 novel coronavirus (2019-nCoV) infections from Wuhan, China: key points for the radiologist. *Radiology* 2020:200241. doi: 10.1148/radiol.2020200241

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., and Wu, A. Y. (2002). An efficient k-means clustering algorithm: analysis

and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 881–892. doi: 10.1109/TPAMI.2002.1017616

Karako, K., Song, P., Chen, Y., and Tang, W. (2020). Analysis of COVID-19 infection spread in Japan based on stochastic transition model. *Biosci. Trends*. 14, 134–138. doi: 10.5582/bst.2020.01482

Kavadi, D. P., Patan, R., Ramachandran, M., and Gandomi, A. H. (2020). Partial derivative non-linear global pandemic machine learning prediction of covid 19. *Chaos Solitons Fractals* 139:110056. doi: 10.1016/j.chaos.2020.110056

Kenneth, C. Y., and So, H.-C. (2020). Uncovering clinical risk factors and prediction of severe COVID-19: a machine learning approach based on UK Biobank data. *MedRxiv*. doi: 10.1101/2020.09.18.20197319

Kissler, S. M., Tedijanto, C., Lipsitch, M., and Grad, Y. (2020). Social distancing strategies for curbing the COVID-19 epidemic. *MedRxiv*. doi: 10.1101/2020.03.22.20041079

Kong, W.-H., Li, Y., Peng, M.-W., Kong, D.-G., Yang, X.-B., Wang, L., et al. (2020). SARS-CoV-2 detection in patients with influenza-like illness. *Nat. Microbiol.* 5, 675–678. doi: 10.1038/s41564-020-0713-1

Koo, J. R., Cook, A. R., Park, M., Sun, Y., Sun, H., Lim, J. T., et al. (2020). Interventions to mitigate early spread of SARS-CoV-2 in Singapore: a modelling study. *Lancet Infect. Dis.* 20, 678–688. doi: 10.1016/S1473-3099(20)30162-6

Kooraki, S., Hosseiny, M., Myers, L., and Gholamrezanezhad, A. (2020). Coronavirus (COVID-19) outbreak: what the department of radiology should know. *J. Am. Coll. Radiol.* 17, 447–451. doi: 10.1016/j.jacr.2020.02.008

Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: a review of classification techniques. *Emerg. Artif. Intell. Appl. Comput. Eng.* 160, 3–24. doi: 10.1007/s10462-007-9052-3

Kretzschmar, M., Rozhnova, G., and van Boven, M. (2020b). Isolation and contact tracing can tip the scale to containment of COVID-19 in populations with social distancing. *SSRN 3562458*. doi: 10.2139/ssrn.3562458

Kretzschmar, M. E., Rozhnova, G., Bootsma, M. C. J., van Boven, M., van de Wijgert, J. H. H. M., and Bonten, M. J. M. (2020a). Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. *Lancet Public Heal.* 5, e452–e459. doi: 10.1016/S2468-2667(20)30157-2

Kricka, L. J., Polevikov, S., Park, J. Y., Fortina, P., Bernardini, S., Satchkov, D., et al. (2020). Artificial intelligence-powered search tools and resources in the fight against COVID-19. *Ejifcc* 31:106.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks mark. *Commun. ACM* 60, 84–90. doi: 10.1145/3065386

Kubat, M., and Matwin, S. (1997). Addressing the curse of imbalanced training sets: one-sided selection. *Icml* 97, 179–186.

Kumar, A., Gupta, P. K., and Srivastava, A. (2020). A review of modern technologies for tackling COVID-19 pandemic. *Diabetes Metab Syndr Clin Res Rev*. 14, 569–573. doi: 10.1016/j.dsx.2020.05.008

Lallie, H. S., Shepherd, L. A., Nurse, J. R. C., Erola, A., Epiphaniou, G., Maple, C., et al. (2021). Cyber security in the age of covid-19: a timeline and analysis of cyber-crime and cyber-attacks during the pandemic. *Comput. Secur.* 2021:102248. doi: 10.1016/j.cose.2021.102248

Lalmuanawma, S., Hussain, J., and Chhakchhuak, L. (2020). Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: a review. *Chaos Solitons Fractals* 2020:110059. doi: 10.1016/j.chaos.2020.110059

Laurikkala, J. (2001). "Improving identification of difficult small classes by balancing class distribution," in *Conf. Artif. Intell. Med. Eur* (Berlin), 63–66. doi: 10.1007/3-540-48229-6_9

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lee, J., Sun, J., Wang, F., Wang, S., Jun, C.-H., and Jiang, X. (2018). Privacy-preserving patient similarity learning in a federated environment: development and analysis. *JMIR Med. Informatics* 6:e20. doi: 10.2196/medinform.7744

Lewis, D. (2020). Why many countries failed at COVID contact-tracing-but some got it right. *Nature* 588, 384–387. doi: 10.1038/d41586-020-03518-4

Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., et al. (2020b). Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. *Radiology*. 2020:200905. doi: 10.1148/radiol.2020200905

Li, L., Zhang, Q., Wang, X., Zhang, J., Wang, T., Gao, T.-L., et al. (2020a). Characterizing the propagation of situational information in social media during covid-19 epidemic: a case study on weibo. *IEEE Trans. Comput. Soc. Syst.* 7, 556–562. doi: 10.1109/TCSS.2020.2980007

Li, M., Lei, P., Zeng, B., Li, Z., Yu, P., Fan, B., et al. (2020c). Coronavirus disease (COVID-19): spectrum of CT findings and temporal progression of the disease. *Acad. Radiol.* 27, 603–608. doi: 10.1016/j.acra.2020.03.003

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. (2020d). Federated learning: challenges, methods, and future directions. *IEEE Signal Process Mag.* 37, 50–60. doi: 10.1109/MSP.2020.2975749

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. *R News* 2, 18–22. Available online at: https://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf

Lu, H., and Wang, M. (2019). RL4health: crowdsourcing reinforcement learning for knee replacement pathway optimization. *ArXiv [Preprint]. ArXiv: 190601407*. Available online at: https://arxiv.org/abs/1906.01407

Lu, R., Wu, X., Wan, Z., Li, Y., Zuo, L., Qin, J., et al. (2020a). Development of a novel reverse transcription loop-mediated isothermal amplification method for rapid detection of SARS-CoV-2. *Virol Sin.* 2020:1. doi: 10.3390/ijms21082826

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., et al. (2020b). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 395, 565–574. doi: 10.1016/S0140-6736(20)30251-8

Maghdid, H. S., Asaad, A. T., Ghafoor, K. Z., Sadiq, A. S., and Khan, M. K. (2020a). Diagnosing COVID-19 pneumonia from X-ray and CT images using deep learning and transfer learning algorithms. *ArXiv Prepr ArXiv200400038*. doi: 10.1117/12.2588672

Maghdid, H. S., Ghafoor, K. Z., Sadiq, A. S., Curran, K., and Rabie, K. (2020b). A novel ai-enabled framework to diagnose coronavirus covid 19 using smartphone embedded sensors: design study. *ArXiv Prepr ArXiv200307434*. doi: 10.1109/IRI49571.2020.00033

Mahler, T., Nissim, N., Shalom, E., Goldenberg, I., Hassman, G., Makori, A., et al. (2018). Know your enemy: characteristics of cyber-attacks on medical imaging devices. *ArXiv [Preprint]. ArXiv: 180105583*. Available online at: https://arxiv.org/abs/1801.05583

Martin, A., Nateqi, J., Gruarin, S., Munsch, N., Abdarahmane, I., Zobel, M., et al. (2020). An artificial intelligence-based first-line defence against COVID-19: digitally screening citizens for risks *via* a chatbot. *Sci. Rep.* 10, 1–7. doi: 10.1038/s41598-020-75912-x

Mauger, C., Gilbert, K., Lee, A. M., Sanghvi, M. M., Aung, N., Fung, K., et al. (2019). Right ventricular shape and function: cardiovascular magnetic resonance reference morphology and biventricular risk factor morphometrics in UK Biobank. *J. Cardiovasc. Magn. Reson.* 21:41. doi: 10.1186/s12968-019-0551-6

May, R. M., and Anderson, R. M. (1979). Population biology of infectious diseases: part I. *Nature* 280, 455–461. doi: 10.1038/280455a0

Mena-Lorcat, J., and Hethcote, H. W. (1992). Dynamic models of infectious diseases as regulators of population sizes. *J. Math. Biol.* 30, 693–716. doi: 10.1007/BF00173264

Moriyama, M., Hugentobler, W. J., and Iwasaki, A. (2020). Seasonality of respiratory viral infections. *Annu. Rev. Virol.* 7:22445. doi: 10.1146/annurev-virology-012420-022445

Muthuppalaniappan, M., and Stevenson, K. (2021). Healthcare cyber-attacks and the COVID-19 pandemic: an urgent threat to global health. *Int. J. Qual. Heal Care* 33:mzaa117. doi: 10.1093/intqhc/mzaa117

Narin, A., Kaya, C., and Pamuk, Z. (2020). Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. *ArXiv [Preprint]. ArXiv: 200310849*. Available online at: https://arxiv.org/abs/2003.10849

Naudé, W. (2020). Artificial intelligence vs. COVID-19: limitations, constraints and pitfalls. *Ai Soc.* 2020:1. doi: 10.1007/s00146-020-00978-0

NCATS-US (2021). *National COVID Cohort Collaborative (N3C)*. Available online at: https://ncats.nih.gov/n3c (accessed March 24, 2021).

Ndairou, F., Area, I., Nieto, J. J., and Torres, D. F. M. (2020). Mathematical modeling of COVID-19 transmission dynamics with a case study of Wuhan. *Chaos Solitons Fractals* 2020:109846. doi: 10.1016/j.chaos.2020.109846

Nguyen, T. T. (2020). Artificial intelligence in the battle against coronavirus (COVID-19): a survey and future research directions. *ArXiv Prepr ArXiv*. doi: 10.36227/techrxiv.12743933

NHSX. (2021). *National COVID-19 Chest Image Database (NCCID)*. Available online at: https://nhsx.github.io/covid-chest-imaging-database/index.html (accessed March 24, 2021).

Oh, S.-H. (2011). Error back-propagation algorithm for classification of imbalanced data. *Neurocomputing* 74, 1058–1061. doi: 10.1016/j.neucom.2010.11.024

Ozturk, T., Talo, M., Yildirim, E. A., Baloglu, U. B., Yildirim, O., and Acharya, U. R. (2020). Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Comput. Biol. Med.* 2020:103792. doi: 10.1016/j.compbiomed.2020.103792

Pan, F., Ye, T., Sun, P., Gui, S., Liang, B., Li, L., et al. (2020). Time course of lung changes on chest CT during recovery from 2019 novel coronavirus (COVID-19) pneumonia. *Radiology* 2020:200370. doi: 10.1148/radiol.2020200370

Park, Y. J., Choe, Y. J., Park, O., Park, S. Y., Kim, Y.-M., Kim, J., et al. (2020). Contact tracing during coronavirus disease outbreak, South Korea, 2020. *Emerg. Infect. Dis.* 26, 2465–2468. doi: 10.3201/eid2610.201315

Pascual, S., Bonafonte, A., and Serra, J. (2017). SEGAN: speech enhancement generative adversarial network. *ArXiv Prepr ArXiv170309452*. doi: 10.21437/Interspeech.2017-1428

Pastor-Satorras, R., Castellano, C., Van Mieghem, P., and Vespignani, A. (2015). Epidemic processes in complex networks. *Rev. Mod. Phys.* 87:925. doi: 10.1103/RevModPhys.87.925

Patel, S. S., Webster, R. K., Greenberg, N., Weston, D., and Brooks, S. K. (2020). Research fatigue in COVID-19 pandemic and post-disaster research: causes, consequences and recommendations. *Disaster Prev. Manag. An Int. J.* 29, 445–455. doi: 10.1108/DPM-05-2020-0164

Pereira, N. L., Ahmad, F., Cummins, N. W., Byku, M., Morris, A. A., Owens, A., et al. (2020). COVID-19: understanding inter-individual variability and implications for precision medicine. *Mayo Clin. Proc.* 96, 446–463. doi: 10.1016/j.mayocp.2020.11.024

Petersen, S. E., Abdulkareem, M., and Leiner, T. (2019). Artificial intelligence will transform cardiac imaging–opportunities and challenges. *Front. Cardiovasc. Med.* 6:133. doi: 10.3389/fcvm.2019.00133

Petersen, S. E., Aung, N., Sanghvi, M. M., Zemrak, F., Fung, K., Paiva, J. M., et al. (2017). Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort. *J. Cardiovasc. Magn. Reson.* 19:18. doi: 10.1186/s12968-017-0327-9

Phua, J., Weng, L., Ling, L., Egi, M., Lim, C.-M., Divatia, J. V., et al. (2020). Intensive care management of coronavirus disease 2019 (COVID-19): challenges and recommendations. *Lancet Respir. Med.* 8, 506–517. doi: 10.1016/S2213-2600(20)30161-2

Pourhomayoun, M., and Shakibi, M. (2020). Predicting mortality risk in patients with COVID-19 using artificial intelligence to help medical decision-making. *MedRxiv*. doi: 10.1101/2020.03.30.20047308

Prudêncio, M., and Costa, J. C. (2020). Research funding after COVID-19. *Nat. Microbiol.* 5:986. doi: 10.1038/s41564-020-0768-z

Qi, X., Jiang, Z., Yu, Q., Shao, C., Zhang, H., Yue, H., et al. (2020). Machine learning-based CT radiomics model for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: a multicenter study. *MedRxiv*. doi: 10.1101/2020.02.29.20029603

Qian, X., Ren, R., Wang, Y., Guo, Y., Fang, J., Wu, Z.-D., et al. (2020). Fighting against the common enemy of COVID-19: a practice of building a community with a shared future for mankind. *Infect Dis. Poverty* 9, 1–6. doi: 10.1186/s40249-020-00650-1

Quanjel, M. J. R., van Holten, T. C., der Vliet, P. C., Wielaard, J., Karakaya, B., Söhne, M., et al. (2020). Replication of a mortality prediction model in Dutch patients with COVID-19. *Nat. Mach. Intell.* 2020, 1–2. doi: 10.1038/s42256-020-00253-3

Raghu, M., and Schmidt, E. (2020). A survey of deep learning for scientific discovery. *ArXiv [Preprint]. ArXiv: 200311755*. Available online at: https://arxiv.org/abs/2003.11755

Rasheed, J., Jamil, A., Hameed, A. A., Aftab, U., Aftab, J., Shah, S. A., et al. (2020). A survey on artificial intelligence approaches in supporting frontline workers and decision makers for COVID-19 pandemic. *Chaos Solitons Fractals* 2020:110337. doi: 10.1016/j.chaos.2020.110337

Ren, S., He, K., Girshick, R., and Sun, J. (2016). Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1137–1149. doi: 10.1109/TPAMI.2016.2577031

Riad, M. H., Sekamatte, M., Ocom, F., Makumbi, I., and Scoglio, C. M. (2019). Risk assessment of Ebola virus disease spreading in Uganda using a two-layer temporal network. *Sci. Rep.* 9, 1–17. doi: 10.1038/s41598-019-52501-1

Ribeiro, M. H. D. M., da Silva, R. G., Mariani, V. C., and dos Santos Coelho, L. (2020). Short-term forecasting COVID-19 cumulative confirmed cases: perspectives for Brazil. *Chaos Solitons Fractals* 2020:109853. doi: 10.1016/j.chaos.2020.109853

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H., Albarqouni, S., et al. (2020). The future of digital health with federated learning. *ArXiv Prepr ArXiv200308119*. doi: 10.1038/s41746-020-00323-1

Robinson, R., Valindria, V. V., Bai, W., Oktay, O., Kainz, B., Suzuki, H., et al. (2019). Automated quality control in image segmentation: application to the UK Biobank cardiovascular magnetic resonance imaging study. *J. Cardiovasc. Magn. Reson.* 21, 1–14. doi: 10.1186/s12968-019-0523-x

Roden, D. M., Pulley, J. M., Basford, M. A., Bernard, G. R., Clayton, E. W., Balser, J. R., et al. (2008). Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin. Pharmacol. Ther.* 84, 362–369. doi: 10.1038/clpt.2008.89

Rodriguez, F., Scheinker, D., and Harrington, R. A. (2018). Promise and perils of big data and artificial intelligence in clinical medicine and biomedical research. *Circ. Res.* 123, 1282–1284. doi: 10.1161/CIRCRESAHA.118.314119

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. *Lect. Notes Comput. Sci.* 9351, 234–241. Available online at: https://arxiv.org/abs/1505.04597

Rothan, H. A., and Byrareddy, S. N. (2020). The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. *J. Autoimmun.* 2020:102433. doi: 10.1016/j.jaut.2020.102433

Ruiz Estrada, M. A. (2020). The uses of drones in case of massive epidemics contagious diseases relief humanitarian aid: Wuhan-COVID-19 crisis. *SSRN 3546547*. doi: 10.2139/ssrn.3546547

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Rutten-Jacobs, L. C. A., Larsson, S. C., Malik, R., Rannikmäe, K., Sudlow, C. L., Dichgans, M., et al. (2018). Genetic risk, incident stroke, and the benefits of adhering to a healthy lifestyle: cohort study of 306 473 UK Biobank participants. *BMJ* 363:k4168. doi: 10.1136/bmj.k4168

Said, M. A., Verweij, N., and van der Harst, P. (2018). Associations of combined genetic and lifestyle risks with incident cardiovascular disease and diabetes in the UK Biobank Study. *JAMA Cardiol.* 3, 693–702. doi: 10.1001/jamacardio.2018.1717

Salathé M., Althaus, C. L., Neher, R., Stringhini, S., Hodcroft, E., Fellay, J., et al. (2020). COVID-19 epidemic in Switzerland: on the importance of testing, contact tracing and isolation. *Swiss Med. Wkly* 150:w20225. doi: 10.4414/smw.2020.20225

Sameni, R. (2020). Mathematical modeling of epidemic diseases; a case study of the COVID-19 coronavirus. *ArXiv [Preprint]. ArXiv: 200311371*. Available online at: https://arxiv.org/abs/2003.11371

Sattar, N., Ho, F. K., Gill, J. M. R., Ghouri, N., Gray, S. R., Celis-Morales, C. A., et al. (2020). BMI and future risk for COVID-19 infection and death across sex, age and ethnicity: preliminary findings from UK biobank. *Diabetes Metab. Syndr. Clin. Res. Rev.* 14, 1149–1151. doi: 10.1016/j.dsx.2020.06.060

Schuller, B. W., Schuller, D. M., Qian, K., Liu, J., Zheng, H., and Li, X. (2020). Covid-19 and computer audition: an overview on what speech and sound analysis could contribute in the SARS-CoV-2 Corona crisis. *ArXiv Prepr ArXiv200311117*. doi: 10.3389/fdgth.2021.564906

Schwartz, D. A. (2020). An analysis of 38 pregnant women with COVID-19, their newborn infants, and maternal-fetal transmission of SARS-CoV-2: maternal coronavirus infections and pregnancy outcomes. *Arch. Pathol. Lab. Med.* 144, 799–805. doi: 10.5858/arpa.2020-0901-SA

Shen, D., Wu, G., and Suk, H.-I. (2017). Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* 19, 221–248. doi: 10.1146/annurev-bioeng-071516-044442

Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., et al. (2020a). Review of artificial intelligence techniques in imaging data acquisition,

segmentation and diagnosis for covid-19. *IEEE Rev. Biomed. Eng.* 14:2987975. doi: 10.1109/RBME.2020.2987975

Shi, F., Xia, L., Shan, F., Wu, D., Wei, Y., Yuan, H., et al. (2020b). Large-scale screening of covid-19 from community acquired pneumonia using infection size-aware classification. *ArXiv Prepr ArXiv200309860*. doi: 10.1088/1361-6560/abe838

Shi, W., Peng, X., Liu, T., Cheng, Z., Lu, H., Yang, S., et al. (2020c). Deep learning-based quantitative computed tomography model in predicting the severity of COVID-19: a retrospective study in 196 patients. *Ann. Transl. Med.* 9:216. doi: 10.2139/ssrn.3546089

Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P. J., Elhadad, N., Johnson, S. B., et al. (2014). A review of approaches to identifying patient phenotype cohorts using electronic health records. *J. Am. Med. Informat. Assoc.* 21, 221–230. doi: 10.1136/amiajnl-2013-001935

Shortliffe, E. H., and Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *JAMA* 320, 2199–2200. doi: 10.1001/jama.2018.17163

Siettos, C. I., and Russo, L. (2013). Mathematical modeling of infectious disease dynamics. *Virulence* 4, 295–306. doi: 10.4161/viru.24041

SoBigData.eu (2021). *The Project SoBigData*. Available online at: http://project.sobigdata.eu/ (accessed March 24, 2021).

Song, Y., Zheng, S., Li, L., Zhang, X., Zhang, X., Huang, Z., et al. (2020). Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images. *MedRxiv*. doi: 10.1109/TCBB.2021.3065361

Suinesiaputra, A., Sanghvi, M. M., Aung, N., Paiva, J. M., Zemrak, F., Fung, K., et al. (2018). Fully-automated left ventricular mass and volume MRI analysis in the UK Biobank population cohort: evaluation of initial results. *Int. J. Cardiovasc. Imaging* 34, 281–291. doi: 10.1007/s10554-017-1225-9

Suri, J. S., Puvvula, A., Biswas, M., Majhail, M., Saba, L., Faa, G., et al. (2020). COVID-19 pathways for brain and heart injury in comorbidity patients: a role of medical imaging and artificial intelligence-based COVID severity classification: a review. *Comput. Biol. Med.* 2020:103960. doi: 10.1016/j.compbiomed.2020.103960

Sutton, R. S., and Barto, A. G. (1998). *Introduction to Reinforcement Learning. Vol. 135*. Cambridge: MIT press.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit,* 2818–2826. doi: 10.1109/CVPR.2016.308

Tang, Z., Zhao, W., Xie, X., Zhong, Z., Shi, F., Liu, J., et al. (2020). Severity assessment of coronavirus disease 2019 (COVID-19) using quantitative features from chest CT images. *ArXiv Prepr ArXiv200311988*. doi: 10.1088/1361-6560/abbf9e

Team, E. E. (2020). Note from the editors: World Health Organization declares novel coronavirus (2019-nCoV) sixth public health emergency of international concern. *Eurosurveillance* 25:200131e. doi: 10.2807/1560-7917.ES.2020.25.5.200131e

Thevarajan, I., Nguyen, T. H. O., Koutsakos, M., Druce, J., Caly, L., van de Sandt, C. E., et al. (2020). Breadth of concomitant immune responses prior to patient recovery: a case report of non-severe COVID-19. *Nat. Med.* 26, 453–455. doi: 10.1038/s41591-020-0819-2

Tibshirani, R. (1996). Regression shrinkage and selection *via* the lasso. *J. R Stat. Soc. Ser. B* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Toh, C., and Brody, J. P. (2020). Evaluation of a genetic risk score for severity of COVID-19 using human chromosomal-scale length variation. *Hum. Genom.* 14, 1–5. doi: 10.1186/s40246-020-00288-y

Udugama, B., Kadhiresan, P., Kozlowski, H. N., Malekjahani, A., Osborne, M., Li, V. Y. C., et al. (2020). Diagnosing COVID-19: the disease and tools for detection. *ACS Nano* 14, 3822–3835. doi: 10.1021/acsnano.0c02624

UKCDR (2020). *COVID-19 Research Project Tracker by UKCDR & GloPID-R. UK Collab Dev Res*. Available online at: https://www.ukcdr.org.uk/covid-circle/covid-19-research-project-tracker/ (accessed January 1, 2021).

Vaishya, R., Haleem, A., Vaish, A., and Javaid, M. (2020a). Emerging technologies to combat the COVID-19 pandemic. *J Clin Exp Hepatol.* 10, 409–411. doi: 10.1016/j.jceh.2020.04.019

Vaishya, R., Javaid, M., Khan, I. H., and Haleem, A. (2020b). Artificial Intelligence (AI) applications for COVID-19 pandemic. *Diabetes Metab. Syndr. Clin. Res. Rev.* 14, 337–339. doi: 10.1016/j.dsx.2020.04.012

Vynnycky, E., and White, R. (2010). *An Introduction to Infectious Disease Modelling*. Oxford: OUP.

Walls, A. C., Park, Y.-J., Tortorici, M. A., Wall, A., McGuire, A. T., and Veesler, D. (2020). Structure function and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* 181, 281–292. doi: 10.1016/j.cell.2020.02.058

Wan, Y., Shang, J., Graham, R., Baric, R. S., and Li, F. (2020). Receptor recognition by novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS. *J. Virol.* 2020, JVI00127–00120. doi: 10.1128/JVI.00127-20

Wang, C. J., Ng, C. Y., and Brook, R. H. (2020b). Response to COVID-19 in Taiwan: big data analytics, new technology, and proactive testing. *JAMA*. 2020:10.1001/jama.2020.3151. doi: 10.1001/jama.2020.3151

Wang, H., Zhang, F., Zeng, J., Wu, Y., Kemper, K. E., Xue, A., et al. (2019). Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. *Sci. Adv.* 5:eaaw3538. doi: 10.1126/sciadv.aaw3538

Wang, L., Lin, Z. Q., and Wong, A. (2020d). Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Sci. Rep.* 10, 1–12. doi: 10.1038/s41598-020-76550-z

Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., et al. (2020c). A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). *MedRxiv*. doi: 10.1101/2020.02.14.20023028

Wang, X., Deng, X., Fu, Q., Zhou, Q., Feng, J., Ma, H., et al. (2020e). A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT. *IEEE Trans. Med. Imaging.* 39:2995963. doi: 10.1109/TMI.2020.2995965

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R. M., et al. (2017). "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.,* 2097–2106. doi: 10.1109/CVPR.2017.369

Wang, Y., Hu, M., Li, Q., Zhang, X.-P., Zhai, G., and Yao, N. (2020f). Abnormal respiratory patterns classifier may contribute to large-scale screening of people infected with COVID-19 in an accurate and unobtrusive manner. *ArXiv [Preprint]. ArXiv: 200205534*. Available online at: https://arxiv.org/abs/2002.05534

Wang, Y., Wang, Y., Chen, Y., and Qin, Q. (2020a). Unique epidemiological and clinical features of the emerging 2019 novel coronavirus pneumonia (COVID-19) implicate special control measures. *J. Med. Virol.* 92, 568–576. doi: 10.1002/jmv.25748

WHO (2020). *World Health Organization Director-General's opening remarks at the media briefing on COVID-19*. Geneva: WHO.

Wong, H. Y. F., Lam, H. Y. S., Fong, A. H.-T., Leung, S. T., Chin, T. W.-Y., Lo, C. S. Y., et al. (2020). Frequency and distribution of chest radiographic findings in COVID-19 positive patients. *Radiology* 2020:201160. doi: 10.1148/radiol.2020201160

Wright, A. F., Carothers, A. D., and Campbell, H. (2002). Gene–environment interactions—the BioBank UK study. *Pharmacogenom. J.* 2, 75–82. doi: 10.1038/sj.tpj.6500085

Wu, J., Zhang, P., Zhang, L., Meng, W., Li, J., Tong, C., et al. (2020a). Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results. *MedRxiv*. doi: 10.1101/2020.04.02.20051136

Wu, J. T., Leung, K., and Leung, G. M. (2020b). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet.* 395, 689–697. doi: 10.1016/S0140-6736(20)30260-9

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., et al. (2020). Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 369:m1328. doi: 10.1136/bmj.m1328

Xu, B., Gutierrez, B., Mekaru, S., Sewalk, K., Goodwin, L., Loskill, A., et al. (2020b). Epidemiological data from the COVID-19 outbreak, real-time case information. *Sci. Data* 7, 1–6. doi: 10.1038/s41597-020-0448-0

Xu, J., Glicksberg, B. S., Su, C., Walker, P., Bian, J., and Wang, F. (2020d). Federated learning for healthcare informatics. *J. Healthc. Informat. Res.* 2020, 1–19. doi: 10.1007/s41666-020-00082-4

Xu, X., Jiang, X., Ma, C., Du, P., Li, X., Lv, S., et al. (2020c). A deep learning system to screen novel coronavirus disease 2019 pneumonia. *Engineering* 6, 1122–1129. doi: 10.1016/j.eng.2020.04.010

Xu, Y., Li, X., Zhu, B., Liang, H., Fang, C., Gong, Y., et al. (2020a). Characteristics of pediatric SARS-CoV-2 infection and potential evidence for

persistent fecal viral shedding. *Nat. Med.* 26, 502–505. doi: 10.1038/s41591-020-0817-4

Yan, L., Zhang H-T, Goncalves, J., Xiao, Y., Wang, M., Guo, Y., et al. (2020b). An interpretable mortality prediction model for COVID-19 patients. *Nat. Mach. Intell.* 2020, 1–6. doi: 10.1038/s42256-020-0180-7

Yan, L., Zhang, H.-T., Xiao, Y., Wang, M., Sun, C., Liang, J., et al. (2020a). Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. *MedRxiv*. doi: 10.1101/2020.02.27.20028027

Yi, X., Walia, E., and Babyn, P. (2019). Generative adversarial network in medical imaging: a review. *Med. Image Anal.* 58:101552. doi: 10.1016/j.media.2019.101552

Zaidan, A. A., Zaidan, B. B., Alsalem, M. A., Albahri, O. S., Albahri, A. S., and Qahtan, M. Y. (2020). Multi-agent learning neural network and Bayesian model for real-time IoT skin detectors: a new evaluation and benchmarking methodology. *Neural Comput. Appl.* 32, 8315–8366. doi: 10.1007/s00521-019-04325-3

Zeng, H., Xu, C., Fan, J., Tang, Y., Deng, Q., Zhang, W., et al. (2020). Antibodies in infants born to mothers with COVID-19 pneumonia. *JAMA* 323, 1848–1849. doi: 10.1001/jama.2020.4861

Zhang, J., Xie, Y., Pang, G., Liao, Z., Verjans, J., Li, W., et al. (2021). Viral pneumonia screening on chest X-rays using confidence-aware anomaly detection. *IEEE Trans Med Imaging* 40, 879–890, doi: 10.1109/TMI.2020.3040950

Zhao, J., Zhang, Y., He, X., and Xie, P. (2020a). COVID-CT-Dataset: a CT scan dataset about COVID-19. *ArXiv [Preprint]. ArXiv: 200313865*. Available online at: https://arxiv.org/abs/2003.13865v1

Zhao, S., Musa, S. S., Lin, Q., Ran, J., Yang, G., Wang, W., et al. (2020b). Estimating the unreported number of novel coronavirus (2019-nCoV) cases in China in the First Half of January 2020: a data-driven modelling analysis of the early outbreak. *J. Clin. Med.* 9:20388. doi: 10.3390/jcm902 0388

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., et al. (2020a). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* 395, 1054–1062. doi: 10.1016/S0140-6736(20) 30566-3

Zhou, Y., Wang, F., Tang, J., Nussinov, R., and Cheng, F. (2020b). Artificial intelligence in COVID-19 drug repurposing. *Lancet Digit. Heal.* 2, 667–676. doi: 10.1016/S2589-7500(20) 30192-8

Zhu, X. J. (2005). *Semi-Supervised Learning Literature Survey*. Technical Report, University of Wisconsin, Madison, WI, United States.

Zimmerman, A., and Kalra, D. (2020). Usefulness of machine learning in COVID-19 for the detection and prognosis of cardiovascular complications. *Rev. Cardiovasc. Med*. 21, 345–352. doi: 10.31083/j.rcm.2020.03.120

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Big Data

ORIGINAL RESEARCH
published: 04 June 2021
doi: 10.3389/fdata.2021.675882

Development of An Individualized Risk Prediction Model for COVID-19 Using Electronic Health Record Data

Tarun Karthik Kumar Mamidi[1†], *Thi K. Tran-Nguyen*[2†], *Ryan L. Melvin*[3] *and Elizabeth A. Worthey*[1,2*]

[1]*Center for Computational Genomics and Data Science, Departments of Pediatrics and Pathology, University of Alabama at Birmingham School of Medicine, Birmingham, AL, United States,* [2]*Hugh Kaul Precision Medicine Institute, University of Alabama at Birmingham, Birmingham, AL, United States,* [3]*Department of Anesthesiology and Perioperative Medicine, University of Alabama at Birmingham, Birmingham, AL, United States*

OPEN ACCESS

Edited by:
Hong Qin,
University of Tennessee at Chattanooga, United States

Reviewed by:
Yaojiang Huang,
Minzu University of China, China
Minh Pham,
University of South Florida, United States

***Correspondence:**
Elizabeth A. Worthey
eworthey@peds.uab.edu

[†]*These authors have contributed equally to this work and share first authorship.*

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Big Data

Received: *04 March 2021*
Accepted: *19 May 2021*
Published: *04 June 2021*

Citation:
Mamidi TKK, Tran-Nguyen TK, Melvin RL and Worthey EA (2021) Development of An Individualized Risk Prediction Model for COVID-19 Using Electronic Health Record Data. Front. Big Data 4:675882. doi: 10.3389/fdata.2021.675882

Developing an accurate and interpretable model to predict an individual's risk for Coronavirus Disease 2019 (COVID-19) is a critical step to efficiently triage testing and other scarce preventative resources. To aid in this effort, we have developed an interpretable risk calculator that utilized de-identified electronic health records (EHR) from the University of Alabama at Birmingham Informatics for Integrating Biology and the Bedside (UAB-i2b2) COVID-19 repository under the U-BRITE framework. The generated risk scores are analogous to commonly used credit scores where higher scores indicate higher risks for COVID-19 infection. By design, these risk scores can easily be calculated in spreadsheets or even with pen and paper. To predict risk, we implemented a Credit Scorecard modeling approach on longitudinal EHR data from 7,262 patients enrolled in the UAB Health System who were evaluated and/or tested for COVID-19 between January and June 2020. In this cohort, 912 patients were positive for COVID-19. Our workflow considered the timing of symptoms and medical conditions and tested the effects by applying different variable selection techniques such as LASSO and Elastic-Net. Within the two weeks before a COVID-19 diagnosis, the most predictive features were respiratory symptoms such as cough, abnormalities of breathing, pain in the throat and chest as well as other chronic conditions including nicotine dependence and major depressive disorder. When extending the timeframe to include all medical conditions across all time, our models also uncovered several chronic conditions impacting the respiratory, cardiovascular, central nervous and urinary organ systems. The whole pipeline of data processing, risk modeling and web-based risk calculator can be applied to any EHR data following the OMOP common data format. The results can be employed to generate questionnaires to estimate COVID-19 risk for screening in building entries or to optimize hospital resources.

Keywords: COVID-19, electronic health record, risk prediction, ICD-10, credit scorecard model

INTRODUCTION

Despite recent progress in the Coronavirus Disease 2019 (COVID-19) vaccines approval and distribution, the pandemic continues to pose a tremendous burden to our healthcare system. Global resources to manage this current crisis continued to be in short supply. It remains critical to quickly and efficiently identify, screen and monitor individuals with the highest risks for COVID-19 so that distribution of therapeutics can be based on individual risks. Many factors including pre-existing chronic conditions (Liu et al., 2020), age, sex, ethnicity and racial background, access to health care, and other social-economic components (Rashedi et al., 2020) have been shown to affect an individual's risk for this disease.

Accordingly, several predictive models that seek to optimize hospital resource management and clinical decisions have been developed (Jehi et al., 2020a; Jehi et al., 2020b; Gong et al., 2020; Liang et al., 2020; Wynants et al., 2020; Zhao et al., 2020). To a large degree, these informatic tools leverage the vast and rich health information available from Electronic Health Record (EHR) data (Jehi et al., 2020b; Oetjens et al., 2020; Osborne et al., 2020; Vaid et al., 2020; Wang et al., 2021a; Wang et al., 2021b; Estiri et al., 2021; Halalau et al., 2021; Schwab et al., 2021). EHR systems contain longitudinal data about patients' demographics, health history, current and past medications, hospital admissions, procedures, current and past symptoms and conditions. Although the primary purpose of EHRs is clinical, over the last decade researchers have used them to conduct clinical and epidemiological research. This has been notable especially during the COVID-19 pandemic where such research that generated invaluable data about COVID-19 risks, comorbidities, transmission and outcomes was quickly adapted for clinical decision making (Daglia et al., 2021). To ensure interoperability across multiple hospital systems, EHR data incorporate standard reference terminology and standard classification systems such as the International Classification of Diseases (ICD) that organize and classify diseases and procedures for facile information retrieval (Bowman, 2005). Incorporated into the Medical Outcomes Partnership (OMOP) common data model (Blacketer, 2021), these ICD9/ICD10 codes facilitate systemic analyses of disparate EHR datasets across different healthcare organizations.

Many of these insights were generated using machine learning methods, based on multi-dimensional data (Mitchell, 1997). Studies have employed a variety of classification and/or regression methods including Naive Bayes, Support Vector Machine, Decision Tree, Random Forest, AdaBoost, K-nearest-neighbor, Gradient-boosted DT, Logistic Regression, Artificial Neural Network, and Extremely Randomized Trees (Alballa and Al-Turaiki, 2021). Among these, the most popular methods applied to COVID-19 have been linear regression, XGBoost, and Support Vector Machine (Alballa and Al-Turaiki, 2021).

To develop a COVD-19 risk model, we chose a Logistic Regression based Credit Scorecard modeling approach to estimate the probability of COVID-19 diagnosis given an individual's ICD9/ICD10 encoded symptoms and conditions. Credit Scorecard is a powerful predictive modeling technique widely adopted by the financial industry to manage risks and control losses when lending to individuals or businesses by predicting the probability of default (Bailey, 2006). The Credit Scorecard model is most frequently used by scorecard developers not only due to its high prediction accuracy, but also due to its interpretability, transparency and ease of implementation. This method has been implemented previously for EHR data based COVID-19 risk prediction (Jehi et al., 2020a; Jehi et al., 2020b).

Application of feature selection methods that attempt to retain the subset of features that are most applicable for classification has been applied to increase interpretability, enhance speed, reduce data dimensionality and prevent overfitting (Alballa and Al-Turaiki, 2021). While there are many feature selection methods, sparse feature selection methods such as LASSO (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996) and Elastic-Net (Zou and Hastie, 2005) provide advantages. LASSO places an upper bound constraint on the sum of the absolute values of the model parameters by penalizing the regression coefficients based on their size and forcing certain coefficients to zero and eventually excluding them to retain the most useful features (Tibshirani, 1996). Expanded from LASSO, Elastic-Net adds a quadratic penalty term to the calculation of coefficients to prevent the "saturation" problem encountered when a limited number of variables are selected (Zou and Hastie, 2005). Several COVID-19 risk prediction models employed LASSO (Gong et al., 2020; Liang et al., 2020; Feng et al., 2021) and Elastic-Net (Heldt et al., 2021; Hu et al., 2021; Huang et al., 2021).

The major goals for this analysis were to determine whether we could: 1) leverage the existing hierarchical structure of the ICD9/ICD-10 classification system, in an unbiased approach, to capture patients' symptoms and conditions and estimate their possibilities of having a COVID-19 diagnosis, 2) examine the temporal aspect of EHR (i.e., within a timeframe, for example, symptoms within 2-weeks of infection/diagnosis). to evaluate what current symptoms and/or pre-existing conditions affect COVID-19 risks, 3) apply a Credit Scorecard modeling approach to develop and validate a predictive model for COVID-19 risk from retrospective EHR data, and 4) develop a pipeline requiring minimal manual curation capable of generating COVID-19 risk models from any EHR data using the OMOP common data model (Blacketer, 2021). To demonstrate the latter goal a web application was created to take answers from individuals and produces a COVID-19 risk score. We have made the code freely available for anyone wishing to reproduce and deploy such a model at gitlab.rc.uab.edu/center-for-computational-genomics-and-data-science/public/covid-19_risk_predictor.

MATERIALS AND METHODS

Dataset

The UAB Informatics Institute Integrating Biology and the Bedside (i2b2) COVID-19 Limited Datasets (LDS) contain de-identified EHR data that are also part of the NIH COVID-19 Data Warehouse

(NCATS, 2020). Data was made available through the UAB Biomedical Research Information Technology Enhancement (U-BRITE) framework. Access to the level-2 i2b2 data was granted upon self-service pursuant to an IRB exemption. Our dataset contains longitudinal data of patients in the UAB Health System who had COVID-19 testing and/or diagnosis from January to June 2020. Aggregated from six different databases, our dataset was transformed to adhere to the OMOP Common Data Model Version 5.3.1 (Blacketer, 2021) to enable systemic analyses of EHR data from disparate sources.

The UAB i2b2 COVID-19 LDS is comprised of 14 tables corresponding to different domains: PERSON, OBSERVATION_PERIOD, SPECIMEN, DEATH, VISIT_OCCURRENCE, PROCEDURE_OCCURENCE, DRUG_EXPOSURE, DEVICE_EXPOSURE, CONDITION_OCCURENCE, MEASUREMENT, OBSERVATION, LOCATION, CARE_SITE and PROVIDER. For the purpose of this study, we limit assessment to previous reported conditions (from CONDITION_OCCURENCE) and lifestyle/habits (from OBSERVATION).

Data Processing

Data wrangling was performed using Python 3.8.5 with the Pandas package 1.2.1 and Numpy package 1.19.5. Code for recreating our process is freely available (see code availability statement below). The following subsections detail the information retrieved from the database tables mentioned above.

Person Table

Demographic information (i.e., age, gender, race, and ethnicity) for each de-identified individual was extracted from the PERSON table. Ages were extracted using the "year of birth" values.

Measurement Table

Information about COVID-19 testing was stored in the Measurement table. We extracted the date, test type and test result for each person.

COVID-19 positivity was determined by the presence of either one of the three criteria: positive COVID-19 antibody test, positive COVID-19 Polymerase Chain Reaction (PCR) test, or the presence of ICD-10 U07.1 code in the EHR record. COVID-19 negativity was assigned if the person were tested for COVID-19 but has never had a positive test nor an ICD-10 U07.1 code.

Condition_Occurence Table

We extracted medical conditions (such as signs and symptoms, injury, abnormal findings and diagnosis) for each patient from this table by leveraging the inherent hierarchical structure of the ICD-10 classification system.

Observation Table

Lifestyle and habits (i.e., BMI, smoking, alcohol and substance use) were extracted from this table. This table also includes the current status (i.e., current, former, never or unknown) of habits for each patient.

Feature Filtering and Extraction

Demographics, lifestyle/habits and conditions (encoded by ICD-9/ICD-10) are obtained as features in our model. For the purpose of using the updated version of ICD codes as features, we converted all ICD-9 codes to ICD-10 codes using a publicly available converter script (Hanratty, 2019). We used these converted codes along with the original ICD-10 codes to map and extract conditions reported in the EHR for each patient.

Before feature extraction, we filtered out all COVID-19 related ICD-10 codes such as U07.1 (COVID-19, virus identified), Z86.16 (personal history of COVID-19), J12.82 (pneumonia due to coronavirus disease 2019), B94.8 (sequelae of COVID-19), B34.2 (Coronavirus infection, unspecified), and B97.2 (Coronavirus as the cause of diseases classified elsewhere). Discarding COVID-19-related codes is imperative to prevent data leakage in our predictive model. Data leakage refers to the inclusion of information about the target of the prediction in the features used for making the prediction that should not be (legitimately) available at the time a prediction is made (Huang et al., 2000; Nisbet et al., 2009; Kaufman et al., 2012; Filho et al., 2021).

Temporal Filter for Medical Condition data

For the positive cohort, we used the date of patients' first COVID-19 testing or their first assignment of the COVID-19-related ICD-10 codes (U07.1, U07.2, Z86.16, J12.82, B94.8, B34.2, or B97.29) as the timestamp to apply a temporal filter for feature selections. For the negative cohort, we also used the date of their first COVID-19 testing as the timestamp. We define temporal filter as a restricted timeframe to study the effect of conditions for infection (i.e., to assess risk using medical conditions occurred within 2 weeks before an infection). This temporal filter is crucial to once again avoid data leakage by excluding features that may emerge as a result of a COVID-19 infection or diagnosis.

To investigate how the timing of medical events and conditions may affect the risk for COVID-19, we extracted the condition data over two distinct time intervals. The first timeframe only considers the conditions within the 2-week window prior to the date of diagnosis whereas the second timeframe retains all condition data before a given patient's first COVID-19 test or diagnosis.

Credit Scorecard Model

Variable (Feature) Selections

After extracting patients' demographic information, lifestyle, habits and ICD-10 condition codes, we converted them to features using one-hot encoding. Features with more than 95% missing data or 95% identical values across all observations were removed. The remaining variables underwent weight-of-evidence (WoE) transformation, which standardizes the scale of features and establishes a monotonic relationship with the outcome variable (Zdravevski et al., 2011). WoE transformation also handles missing and extreme outliers while supporting interpretability through enforcing strict linear relationships (Zdravevski et al., 2011). WoE transformations

TABLE 1 | Demographics and Clinical Characteristics of the UAB LDS N3C Cohort.

UAB LDS N3C cohort (*n* = 7,262)		
COVID-19 testing:		
COVID-19 results	Positive (*n* = 912)	Negative (*n* = 6,350)
Total COVID tests	1,328	7,596
COVID Tests/Person	1.46	1.20
All medical tests:		
All tests	1,951,404	17,395,613
All tests/person	2,139	2,739
Age	mean = 52 (10–119)	mean = 52 (<1–119)
Gender:		
Male (%)	394 (43%)	3,035 (48%)
Female (%)	516 (57%)	3,314 (52%)
Unknown (%)	2 (0%)	1 (0%)
Race:		
White (%)	337 (37%)	3,441 (54%)
Black (%)	416 (46%)	2,497 (39%)
Asian (%)	27 (3%)	70 (1%)
Hispanic (%)	28 (3%)	174 (3%)
Others (%)	104 (11%)	168 (3%)
Conditions:		
Total conditions	129,091	1,133,396
Unique conditions	9,224	24,101
#Conditions/Person	142	178
#Unique conditions/Person	10	4
Smoking:		
Current smoker	81 (9%)	1,602 (25%)
Former smoker	196 (21.5%)	1,625 (26%)
Never smoker	368 (40%)	2,589 (41%)
Unknown	13 (1%)	64 (1%)
Substance use:		
Current substance abuse	27 (3%)	895 (14%)
No substance abuse	632 (69%)	4,716 (74%)
Former substance abuse	32 (3.5%)	402 (6%)
Unknown	15 (1.6%)	74 (1%)
Alcohol use:		
Current alcohol	273 (30%)	1954 (31%)
Former alcohol	58 (6%)	652 (10%)
No alcohol	379 (41.5%)	3,459 (54.5%)
Unknown	12 (1.3%)	80 (1%)
Weight:		
Underweight (BMI < 19)	20 (2%)	271 (4%)
Normal weight (BMI = 20–25)	49 (5%)	563 (9%)
Overweight (BMI = 25–40)	320 (35%)	2,439 (38%)
Obese (BMI > 40)	120 (13%)	773 (12%)

require all continuous or discrete variables to be binned. This binning process is carried out programmatically based on conditional inference trees (Hothorn et al., 2006). Missing values for each feature are placed in their own bin and eventually assigned their own WoE values. Each level (x) of the binned values for each feature is then assigned a WoE value via $WoE(x) = \ln\left(\frac{P(x|y=1)}{P(x|y=0)}\right)$ where $P(x/y)$ is the conditional probability of x given y, and y is the binary response variable. All values of the independent variables, including missing values, are then replaced with their corresponding WoE value (Zdravevski et al., 2011; Szepannek, 2020). These transformed variables were then used in logistic regression to assign weights for the Scorecard.

For feature selection and regression on these transformed variables, we tested two regularization approaches, LASSO (Tibshirani, 1996) and Elastic-Net (Zou and Hastie, 2005), using a cross-validation-based logistic regression method from the Python package *Scikit-Learn* (version 0.23.2). This method incorporates the use of stratified cross-validation to determine optimal parameters for LASSO and Elastic-Net. LASSO is a modification to typical generalized linear modeling techniques such as logistic regression. Under the constraint the sum of the absolute value of the model coefficients are less than a constant, the residual sum of square errors is minimized (Tibshirani, 1996). The application of this constraint results in some coefficients being 0, making LASSO a simultaneous variable selection and model fitting technique. Building on LASSO, Elastic-Net adds a quadratic penalty term to the calculation of coefficients. Practically, this additional term prevents the "saturation" (Zou and Hastie, 2005) problem sometimes experienced with LASSO where an artificially limited number of variables are selected. Both techniques employ penalty terms to shrink variable coefficients to eliminate uninformative features and avoid collinearity.

Collinearity is a major problem in extracting features from ICD codes since some codes are frequently reported together, or different providers may use inconsistent and incomplete codes. Between the two approaches, LASSO is a more stringent variable selector. For example, in the case of two highly similar features, LASSO tends to eliminate one of them while Elastic-Net will shrink the corresponding coefficients and keep both features (Hastie et al., 2001).

The regularization strength (for both LASSO and Elastic-Net) parameter and mixing parameter (for Elastic-Net) were selected using 10-fold stratified cross-validation (CV). This method creates 10 versions of the model using a fixed set of parameters, each trained on 90% of the training data with 10% held out in each "fold" for scoring that particular instance of the model. The stratified variant of CV ensures that the distribution of classes (here COVID-positive patients and COVID-negative patients) is identical across the 90%/10% split of each fold. This process enables the model developer to assess the predictive capability of the model given the specific set of parameters being tested. The scores over all folds are averaged to assign an overall score for the given set of parameters. This process is repeated for all candidate sets of parameters being tested. Cross-validation aids in preventing overfitting, i.e., failing to generalize the pattern from the data, because the model is judged based on its predictions on hold-out data, which are not used for training the model.

For scoring candidate sets of parameters, we chose negative log loss, a probability-based scoring metric, because a Scorecard model is based on probabilities rather than strict binary predictions. In particular, negative log loss penalizes predictions based on how far their probability is from the correct response (Bishop, 2016). For example, consider a patient who is in truth COVID-negative. A forecast that a COVID-positive diagnosis is 51% likely will be penalized less harshly than a forecast that COVID-positive is 99% likely.

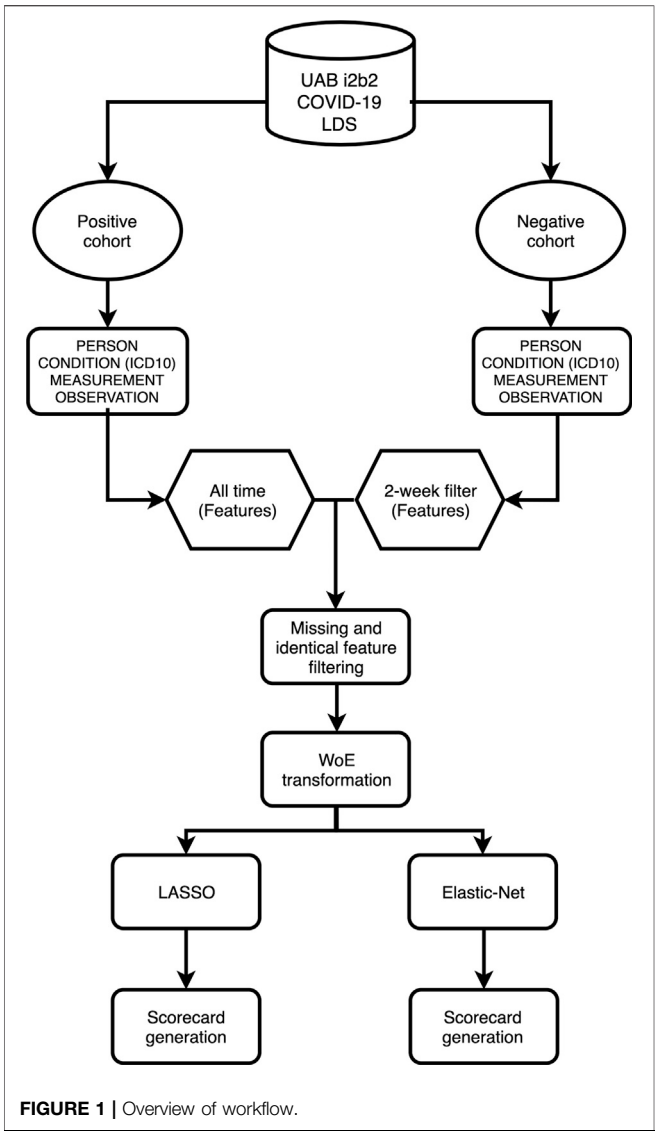


FIGURE 1 | Overview of workflow.

Conversely, a forecast that a positive diagnosis is 49% likely will be rewarded less than one that such a diagnosis is 1% likely.

The hyperparameters evaluated for candidate LASSO models was regularization strength, or the inverse of lambda referred to in (Tibshirani, 1996). One-hundred candidate values on a log scale between $1e^{-4}$ and $1e^{4}$ were considered. The model with the best score from the technique described above was considered to have the optimal hyperparameters. For Elastic-Net, the same set of regularization strength parameters was considered. Additionally, Elastic-Net has a mixing parameter that controls the relative strength of the LASSO-like penalty and the additional Elastic-Net penalty term. Ten evenly spaced values between 0 and 1 were considered for this hyperparameter.

To address the class imbalance between COVID-19 positive and negative group in the training data, we weighted each observation inversely proportional to the size of its class. Likewise, the use of a stratified cross-validation method reduces the risk of inflating some scoring metrics by the model preferring to simply predict the dominant class. Using the above methods, we wanted to compare and contrast four models to predict the risk for infection. Below are the four models:

1. LASSO with all conditions/features reported before the infection/diagnosis
2. Elastic-Net with all conditions/features reported before the infection/diagnosis
3. LASSO with only conditions/features reported within 2 weeks of infection/diagnosis
4. Elastic-Net with only conditions/features reported within 2 weeks of infection/diagnosis

Model Evaluations

Data were randomly split into 80% for the train set and 20% for the test set. The quality of the four models built from two different time-filtered datasets and two different regularization techniques were evaluated by plotting the Receiving Operating Characteristic (ROC) curve and measuring the corresponding Area Under the ROC Curve (AUC). We also considered other model quality metrics such as Accuracy (ACC)—the percent of correct responses—and F-score—the harmonic mean of precision and recall. We also used the confusion matrices to judge the quality of our candidate models. Considering that these models are built to recommend COVID-19 testing, we sought to avoid False Negative predictions while being more lenient towards False Positive errors.

Risk Score Scaling Using the Scorecard Method

Coefficients from the resulting logistic regression models were then combined with the WoE-transformed variables to establish scores for each feature in the Scorecard. This scorecard generation was performed using the Scorecard method implemented in the *scorecardpy* python package (version 0.1.9.2). As opposed to pure logistic regression models, scorecard models allow a strictly linear combination of scores that can be calculated even on a piece of paper, without the aid of any technology. Calculating the probabilities from a logistic regression model would require inverse transformations of log odds. We chose the scorecard model for the strict linear interpretation and corresponding ease of deployment anywhere.

This method requires users to select target odds and target points (a baseline number of points corresponding to a baseline score) along with the points required to double the odds. As these choices are arbitrary, we used the package defaults, which set the target odds to 1/19, the corresponding target points to 600, and the default points required to double the odds to 50. **Supplemental Figure S1** shows an example of a Scorecard distribution calculated in this manner. Since the final Scorecard model is a linear function of the predictors (i.e., higher scores indicate higher COVID-19 risks), using scorecards has many benefits such as transparency, interpretability and facile implementation.

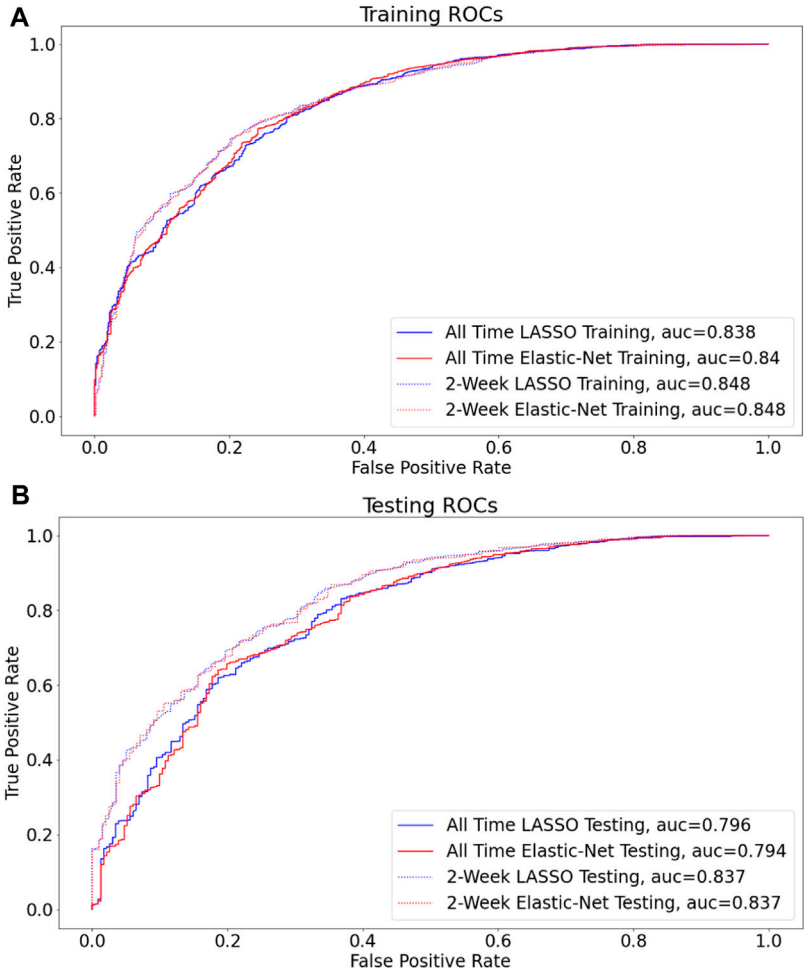


FIGURE 2 | LASSO vs Elastic-Net model performance on two sets of data Receiver operating characteristic (ROC) curves are shown for the final model for each of the four assessed techniques **(A,B)**, and the corresponding areas under curves (AUC) are presented in the figure legend. By AUC on hold out data (0.815), the models built on data filtered by two-week before COVID (non)diagnosis perform the best (B).

Building a Web Application to Predict COVID-19 Risks

Based on the final Scorecard model results, we used the *streamlit* package (version 0.77.0) in Python to build an interface and used interactive indicator plot from *plotly* to visualize the risk score. The Python code to build this application can be found in our gitlab repository at gitlab.rc.uab.edu/center-for-computational-genomics-and-data-science/public/covid-19_risk_predictor.

RESULTS

Our dataset was composed of 7,262 patients from within the UAB Health System who received COVID-19 testing or diagnosis from January to June 2020. The demographic information of this study population is shown in **Table 1**. Among them, 912 patients were diagnosed with COVID-19 and the remaining 6,350 patients, were not. On average, patients in the positive group received 13% more COVID-19 tests (1.45 vs. 1.19 tests/person). While there is no statistically significant difference in age and gender between the two groups, African American (46 vs. 39%), Asian (3 vs. 1%) and Others (11 vs. 3%) ethnicity were overrepresented in the positive group, a finding which is concordant with other reports about the racial disparity in COVID-19 (Kullar et al., 2020). In this UAB Health System dataset, a greater number of patients in the negative group reported substance abuse (14 vs. 3%) and current smoking (25 vs. 9%). There was no difference in Body Mass Index (BMI) between the two groups. Although the COVID-19 negative group had more reported medical conditions (178 vs. 142 medical conditions/person), they had fewer unique medical conditions (4 vs. 10 unique conditions/person).

The workflow to build the predictive model for COVID-19 diagnosis based on EHR data is summarized in **Figure 1**. We used condition data extracted from ICD-9/ICD-10 codes from two different timeframes to assess how the timing of medical symptoms and conditions may affect our COVID-19 risk

TABLE 2 | Model metrics Evaluation of four models (LASSO and Elastic-Net with patient's conditions information from two timeframes) while training and testing (i.e., holdout) data set. For each model, the accuracy, F-Score, and AUC with 95% CI using DeLong's method (DeLong et al., 1988) are shown. The accuracy metric indicates the percent of correct predictions. F-score is the harmonic mean of precision and recall. Area under receiver operating curve (AUC) is the area under the curve resulting from plotting the true positive against the false positive rate.

Training metrics			
All-Time + LASSO		**All-Time + Elastic-Net**	
Accuracy	0.746	Accuracy	0.755
F-Score	0.834	F-Score	0.840
AUC	0.838	AUC	0.840
95% AUC CI	[0.82 0.86]	95% AUC CI	[0.82 0.86]
2-Week + LASSO		**2-Week + Elastic-Net**	
Accuracy	0.774	Accuracy	0.775
F-Score	0.847	F-Score	0.848
AUC	0.848	AUC	0.848
95% AUC CI	[0.83 0.87]	95% AUC CI	[0.83 0.87]
Testing Metrics			
All-time + LASSO		**All-time + Elastic-Net**	
Accuracy	0.741	Accuracy	0.744
F-Score	0.832	F-Score	0.834
AUC	0.796	AUC	0.794
95% AUC CI	[0.76 0.83]	95% AUC CI	[0.76 0.83]
2-Week + LASSO		**2-Week + Elastic-Net**	
Accuracy	0.753	Accuracy	0.755
F-Score	0.833	F-Score	0.835
AUC	0.837	AUC	0.837
95% AUC CI	[0.81 0.87]	95% AUC CI	[0.81 0.87]

predictions. The first timeframe considers the data reported within a 2-week window of testing/diagnosis while the second timeframe retains all condition data prior to a COVID-19 test or diagnosis. Such condition data suffer from collinearity issues in that a group of medical conditions tends to be reported together, and different providers may use inconsistent codes for the same conditions. To address these collinearity issues, we utilized two different regularized regression techniques, LASSO and Elastic-Net. Applying these two methods on the two data timeframes yielded four different models with reasonable discriminatory power, as judged by performance metrics on testing data. With LASSO, we achieved 0.75 accuracy and 0.84 [CI: 0.81–0.87] AUC for the 2-week data and 0.74 accuracy and 0.80 [CI: 076–0.83] AUC for all-time data (**Figure 2**; **Table 2**). Elastic-Net models also performed with a similar accuracy of 0.76 and AUC of 0.84 [CI: 0.81–0.87] for the 2-week data and an accuracy of 0.74 and AUC of 0.79 [CI: 0.76–0.83] for the all-time data (**Figure 2**; **Table 2**).

Using LASSO, a more stringent regularization method where many variables are eliminated through shrinkage, after filtering, 30 out of the 58 features were retained (**Supplemental Table S1**) in the 2-week data, and 93 out of 212 features were retained in the all-time data (**Supplemental Table S2**). Within two weeks before a COVID-19 diagnosis, features that predict higher risks for this disease were cough (R05), abnormalities of breathing (R06), pain in throat and chest (R07), abnormal findings on diagnostic imaging of lung (R91), respiratory disorder (J98), disorders of fluid, electrolyte and acid-base balance (E87), nicotine dependence (F17), major depressive disorder (F32) and overweight and obesity (E66) (**Supplemental Table S1**). The LASSO model on all-time data identified similar features from the 2-week data such as cough (R05), but it also delineated other important features related to acute respiratory infections such as fever (R50), pain (R52), acute upper respiratory infections (J06), respiratory failure (J96), respiratory disorder (J98), pneumonia (J18), vasomotor and allergic rhinitis (J30), and other disorders of nose and nasal sinuses (J34). Most notably, the all-time model uncovered several chronic conditions in other organ systems besides the respiratory system including neurological disorders e.g. postviral fatigue syndrome (G93, R41), kidney diseases (I12, I13, N17), diseases of the heart and circulation including hypertension and kidney failure (I49, I51, J95) and fibrosis/cirrhosis of the liver (K74), suggesting that long-term chronic conditions in other organ systems may increase the risks for contracting an acute respiratory illness such as COVID-19.

Even though LASSO is an effective method to handle collinearity issues, it may not work well with multicollinearity where several features are correlated among each other, as observed in our condition data. Considering that LASSO may eliminate important features through the stringent shrinkage process, we also implemented the Elastic-Net regularization method as a less stringent variable selector. This approach retained more features than the LASSO with 43 features remained for the 2-week data and 179 features for the all-time data. All features selected from the LASSO method also remained in the Elastic-Net method. Several new predictive features emerged from the 2-week data including primary hypertension (I10) and gastro-esophageal reflux disease (K21). In the all-time data, many distinct yet similar conditions from the LASSO model also appeared such as acute myocardial infarction (I21), cardiomyopathy (I42), other cardiac arrhythmias (I49), cerebral infarction (I63), complications and ill-defined descriptions of heart disease (I51), peripheral vascular diseases (I73), and other cerebrovascular diseases (I67), pointing to vascular disorders. Other medical conditions also emerged including viral hepatitis (B19), bacterial infection (B96), thrombocytopenia (D69), epilepsy and recurrent seizures (G40), although the predictive powers of these variables were low.

Among the four candidate models we generated based on the UAB-i2b2 data, the LASSO method on the 2-week filtered data retained the fewest variables while achieving similar performance with other more complex models (**Figures 2**, **3**; **Table 2**; **Supplemental Tables S1–S4**). For this reason, we believed this is a superior model and selected it as the model for our web application. This interactive web application (**Figure 4**) gathers participant questionnaire inputs and generates a risk prediction score of having COVID-19. The Scorecard distribution based on the logistic regression model can be found in **Supplemental Figure S1**. This tool can be used for individuals to check their risks based on their symptoms or conditions, or for organizations

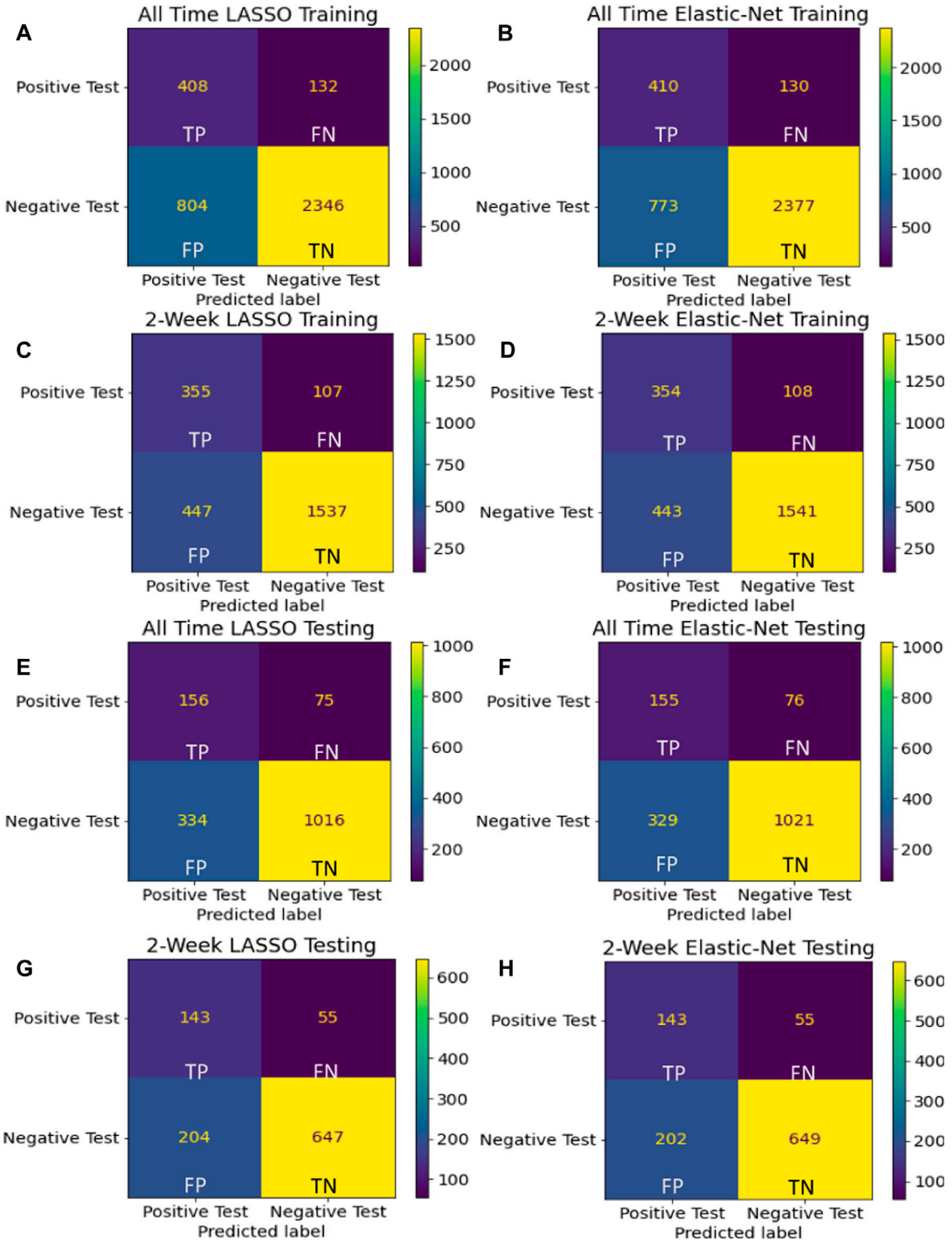


FIGURE 3 | Confusion matrices Confusion matrices using training **(A–D)** and holdout **(E–H)** data are shown for the final model for each of the four assessed techniques. Considering that these models are built to recommend COVID-19 testing, we sought to avoid False Negative predictions while being more lenient towards False Positive errors.

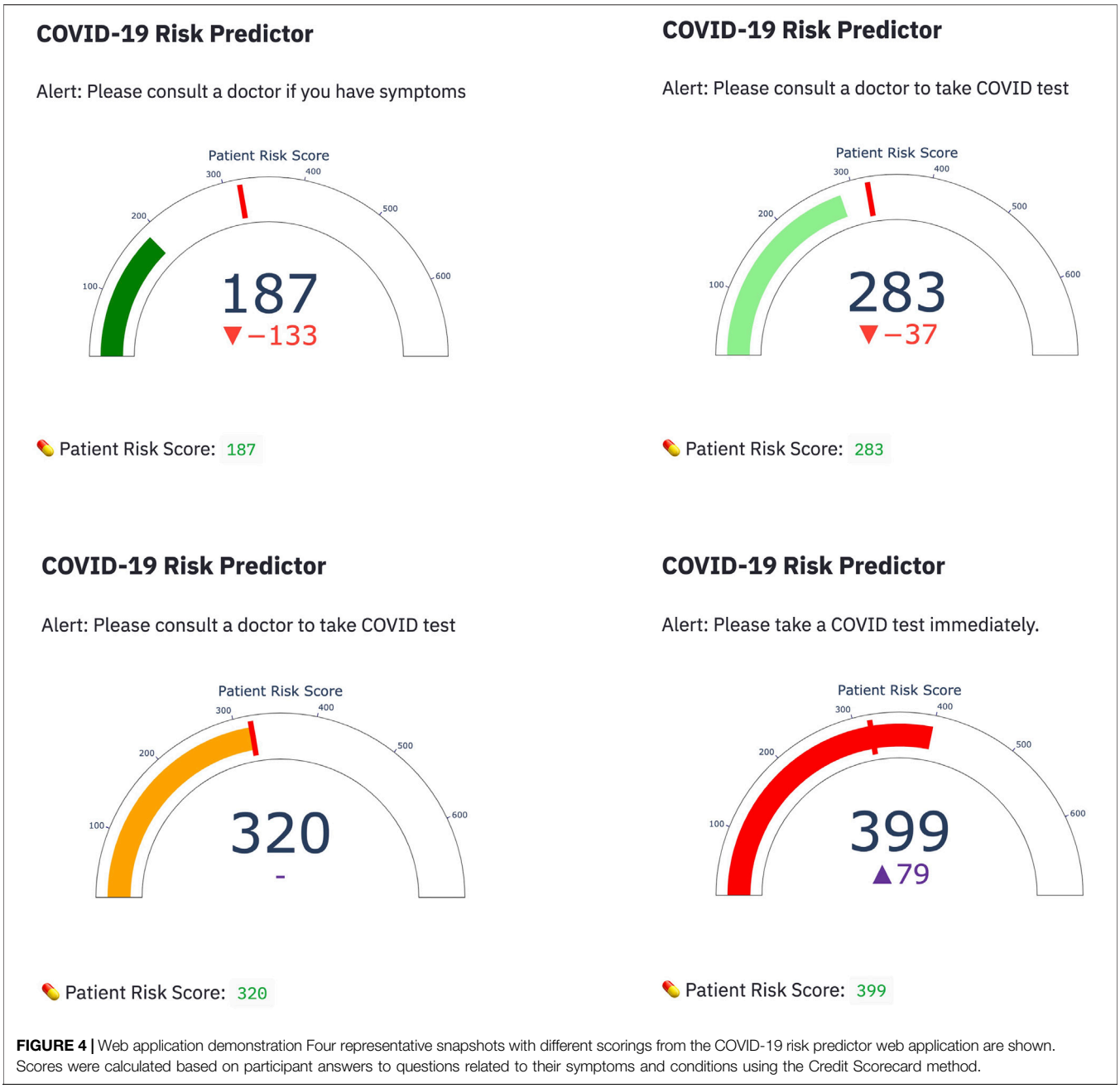


FIGURE 4 | Web application demonstration Four representative snapshots with different scorings from the COVID-19 risk predictor web application are shown. Scores were calculated based on participant answers to questions related to their symptoms and conditions using the Credit Scorecard method.

to build questionnaires to perform COVID-19 screening for building entries. An example questionnaire from our final model is provided in **Table 3**.

DISCUSSION

In this project, we built a data processing and predictive analytics workflow to predict the risks for COVID-19 diagnosis using patients' longitudinal medical conditions encoded by the ICD-9/ICD-10 classification system. We tested the implications of applying different time windows and alternative variable regularization methods to extract the most predictive features from the condition data.

Although the all-time data model selected more features with implications about pre-existing chronic medical conditions increasing the risk of contracting COVID-19, we determined that it was prone to capturing spurious correlations with distant historical data and had weaker performance than the 2-week models (**Figures 2**, **3**; **Table 2**; **Supplemental Tables S1–S4**). With regards to modeling techniques, we found that a more stringent regularized regression approach such as LASSO resulted in simpler models and still achieved high performance as compared to more complex models built

TABLE 3 | Example questionnaire Example questionnaire built using our selected model using the UAB-i2b2 data—the LASSO method on the 2-week filtered data. Base score is 320 and the risk increases/decreases based on the answers in the questionnaire. Any score between 450 and 696 is considered high risk for infection. Disclaimer: This questionnaire is intended only as an example output from a model built using our pipeline. It is not itself a diagnostic tool.

Questions	Yes	No
Do you have chronic kidney disease?	36	−6
Do you have cough?	36	−44
Have you delivered a baby?	35	−2
Are you having acute upper respiratory infections?	30	−6
Do you have fever?	24	−5
Are you having depression, anxiety, problems with cognitive functions or other brain disorders?	17	−4
Are you having pneumonia?	17	−3
Are you having respiratory failure?	16	−3
Are you dependent on nicotine?	14	−4
Do you have allergic rhinitis?	14	−2
Do you have retention of urine?	14	−1
Do you have pain?	14	−1
Do you have hernia?	13	−1
Do you have liver fibrosis/cirrhosis?	13	−1
Do you have disturbances of skin sensation?	12	−2
Are you having anemia?	10	−1
Are you having bacterial infection?	9	−1
Do you have complications from heart disease?	8	−2
Do you have hypotension?	8	−1
Do you have complications of cardiac and vascular prosthetic devices, implants and grafts?	6	0
Are you vitamin D deficient?	2	0
Do you have cardiac arrhythmias?	2	0

from the Elastic-Net method (**Figures 2**, **3**; **Table 2**; **Supplemental Tables S1–S4**). As simpler models tend to be more generalizable, more interpretable, and less likely to be overfit, we consider the LASSO model using the 2-week data filter the superior model for its parsimony without sacrificing performance. Many COVID-19 risk prediction studies also employed LASSO (Alballa and Al-Turaiki, 2021) with a few other studies used Elastic-Net (Heldt et al., 2021; Hu et al., 2021; Huang et al., 2021) as feature selection methods. A COVID-19 diagnostic prediction study by (Feng et al., 2021) compared the performance of four different feature selection methods including LASSO, Ridge, Decision Tree and AdaBoost also found LASSO produced the best performance in both the testing and the validation set.

While our workflow focuses on automatically extracting predictive features from ICD9/10 codes, the majority of COVID-19 prediction studies selected features from a wide-range of additional clinical data components such as chest computed tomography (CT) scan results, laboratory blood tests, which includes complete blood count (e.g., leukocyte, erythrocyte, platelet count, and hematocrit), metabolic factors (e.g., glucose, sodium, potassium, creatinine, urea, albumin, and bilirubin), clotting factors (e.g., prothrombin and fibrinogen), inflammation markers such as C-reactive protein and interleukin 6 (IL-6) (Alballa and Al-Turaiki, 2021). Furthermore, whereas some studies selected the initial sets of features from EHR data based on expert opinions (Estiri et al., 2021; Feng et al., 2021; Schwab et al., 2021) and/or literature review (Joshi et al., 2020; Schwab et al., 2021), we took an unbiased approach to use ICD9/10 codes along with demographic information as the initial set of features. Our data wrangling workflow is limited to the data available in the OMOP common data model, which facilitates scaling up the analyses when we have access to more data of the same format in the future.

Our results showed several COVID-19 predictive features that overlapped with existing published findings. For example, several respiratory symptoms such as cough, abnormalities of breath, and chest pain prioritized by our models—particularly within the 2-week timeframe—are well-known symptoms of COVID-19 (Fu et al., 2020; Huang et al., 2020). Other chronic conditions selected from our models have also been reported to increase COVID-19 risks such as obesity (Popkin et al., 2020), allergic rhinitis (Yang et al., 2020), cardiovascular diseases (Nishiga et al., 2020) and kidney diseases (Adapa et al., 2020) while there are still on-going debates about the role of nicotine and smoking in COVID-19 risks (Polosa and Caci, 2020). Similar to other studies, we found that major depressive disorder is associated with COVID-19 diagnoses. However, it is unclear whether severe mental health problems are the cause, the effect, or the confounding factors with other features associated with COVID-19 (Ettman et al., 2020; Nami et al., 2020; Skoda et al., 2020).

A major limitation in our predictive modeling pipeline relates to the fact that our model is based entirely on correlations between medical conditions and COVID-19 testing/diagnosis. Therefore, by design, this workflow cannot establish causal relationships. As examples, there are several medical conditions associated with lower risks for COVID-19 (**Supplemental Tables S1–S4**) which may highlight distinct features in our negative cohort but may not directly affect COVID-19 risks. This problem, however, is inevitable in predictive analytic workflows that derive inferences from retrospective data. Similar to all studies that apply machine learning methods to model COVID-19 diagnosis, our classifier is prone to imbalanced class distribution where there the positive

COVID-19 instances are underrepresented in the training data (Alballa and Al-Turaiki, 2021). However, we addressed this class imbalance issue by weighing each observation inversely proportional to the size of its class (see the Methods *Variable (Feature) Selections*). Finally; we choose a generalized linear model approach where we assume linear relationships on a logistic scale between medical conditions and COVID-19 risks, and consequently, potential non-linear relationships are not considered.

Although our workflow is straightforward to implement, there are substantial trade-offs by using the ICD-9/ICD-10 standard vocabulary system as opposed to alternative text mining approaches to extract medical conditions from EHR data. ICD code accuracy is a major problem in some cases with classification error rates as high as 80% (O'Malley et al., 2005). The sources of these errors are wide-ranging including poor communication between patients and providers, clinician' mistakes or biases, transcription/scanning errors, coders' experience, and intentional or unintentional biases (e.g., upcoding and unbundling for higher billing/reimbursement value) (O'Malley et al., 2005). Inconsistent, incomplete, systemic and random errors in ICD coding (Cox et al., 2009) introduce noise in the dataset, which is another limitation of our workflow.

Despite these inherent limitations, our study shows the promising utility of incorporating the ICD-10 system in an unbiased manner for novel inferences of EHR data, particularly to study medical symptoms and conditions that influence the risks for COVID-19. Future studies can consider incorporating other standard vocabularies available in EHR data such as Systemized Nomenclature of Medicine (SNOMED), Current Procedural Terminology (CPT), Logical Observation Identifiers Names and Codes (LOINC) as well as adding additional datasets such as patient' medication uses to further understand the risks and the long-term consequences of COVID-19.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: All restrictions of the Limited Data Set (LDS) from the UAB i2b2 system apply to this dataset. Requests to access these datasets should be directed to https://www.uab.edu/ccts/research-commons/berd/55-research-commons/informatics/325-i2b2.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

All authors listed have made direct and substantial contribution to the article and approved the submission of this article.

ACKNOWLEDGMENTS

This COVID-19 risk prediction project was initiated and partially executed during a 3-day Hackathon event at UAB. We thank UAB Informatics Institute for providing data management support with U-BRITE (http://www.ubrite.org/), which made this work possible. We also thank the UAB IT Research Computing who maintains the Cheaha Supercomputer resources, which was supported in part by the National Science Foundation under Grants Nos. OAC-1541310, the University of Alabama at Birmingham, and the Alabama Innovation Fund. The authors also sincerely thank Jelai Wang and Matt Wyatt for providing access to UAB N3C dataset and Ryan C. Godwin for his encouragement on the viability of our pipeline and models.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2021.675882/full#supplementary-material

REFERENCES

Adapa, S., Chenna, A., Balla, M., Merugu, G. P., Koduri, N. M., Daggubati, S. R., et al. (2020). COVID-19 Pandemic Causing Acute Kidney Injury and Impact on Patients with Chronic Kidney Disease and Renal Transplantation. *J. Clin. Med. Res.* 12 (6), 352–361. doi:10.14740/jocmr4200

Alballa, N., and Al-Turaiki, I. (2021). Machine Learning Approaches in COVID-19 diagnosis, Mortality, and Severity Risk Prediction: A Review. *Inform. Med.* 24, 100564. doi:10.1016/j.imu.2021.100564

Bailey, M. (2006). *Practical Credit Scoring: Issues and Techniques*. Bristol, United Kingdom: White Box Publishing.

Bishop, C. M. (2016). *Pattern Recognition and Machine Learning*. Springer.

Blacketer, C. (2021). Chapter 4. The Common Data Model [Online]. Available at: https://ohdsi.github.io/TheBookOfOhdsi/CommonDataModel.html

Bowman, S. E. (2005). Coordination of SNOMED-CT and ICD-10: Getting the Most out of Electronic Health Record Systems. *Perspectives in Health Information Management* [Online]. http://library.ahima.org/doc?oid=106578#.YDXOMGNMEXx

Cox, E., Martin, B. C., Van Staa, T., Garbe, E., Siebert, U., and Johnson, M. L. (2009). Good Research Practices for Comparative Effectiveness Research: Approaches to Mitigate Bias and Confounding in the Design of Nonrandomized Studies of Treatment Effects Using Secondary Data Sources: The International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report-Part II. *Value in Health*. 12 (8), 1053–1061. doi:10.1111/j.1524-4733.2009.00601.x

Dagliati, A., Malovini, A., Tibollo, V., and Bellazzi, R. (2021). Health Informatics and EHR to Support Clinical Research in the COVID-19 PANDEMIc: An Overview. *Brief Bioinform* 22 (2), 812–822. doi:10.1093/bib/bbaa418

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: a Nonparametric Approach. *Biometrics* 44 (3), 837–845. doi:10.2307/2531595

Estiri, H., Strasser, Z. H., Klann, J. G., Naseri, P., Wagholikar, K. B., and Murphy, S. N. (2021). Predicting COVID-19 Mortality with Electronic Medical Records. *Npj Digit. Med.* 4 (1), 15. doi:10.1038/s41746-021-00383-x

Filho, A. C., de Moraes Batista, A. F., and dos Santos, H. G. (2021). Data Leakage in Health Outcomes Prediction With Machine Learning. Comment on "Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning". *J. Med. Internet. Res.* 23, 1–3. doi:10.2196/10969

Ettman, C. K., Abdalla, S. M., Cohen, G. H., Sampson, L., Vivier, P. M., and Galea, S. (2020). Prevalence of Depression Symptoms in US Adults before and during the COVID-19 Pandemic. *JAMA Netw. Open* 3 (9), e2019686. doi:10.1001/jamanetworkopen.2020.19686

Feng, C., Wang, L., Chen, X., Zhai, Y., Zhu, F., Chen, H., et al. (2021). A Novel Artificial Intelligence-Assisted Triage Tool to aid in the Diagnosis of Suspected COVID-19 Pneumonia Cases in Fever Clinics. *Ann. Transl. Med.* 9 (3), 201.

Fu, L., Wang, B., Yuan, T., Chen, X., Ao, Y., Fitzpatrick, T., et al. (2020). Clinical Characteristics of Coronavirus Disease 2019 (COVID-19) in China: A Systematic Review and Meta-Analysis. *J. Infect.* 80 (6), 656–665. doi:10.1016/j.jinf.2020.03.041

Gong, J., Ou, J., Qiu, X., Jie, Y., Chen, Y., Yuan, L., et al. (2020). A Tool for Early Prediction of Severe Coronavirus Disease 2019 (COVID-19): A Multicenter Study Using the Risk Nomogram in Wuhan and Guangdong, China. *Clin. Infect. Dis.* 71 (15), 833–840. doi:10.1093/cid/ciaa443

Halalau, A., Imam, Z., Karabon, P., Mankuzhy, N., Shaheen, A., Tu, J., et al. (2021). External Validation of a Clinical Risk Score to Predict Hospital Admission and In-Hospital Mortality in COVID-19 Patients. *Ann. Med.* 53 (1), 78–86. doi:10.1080/07853890.2020.1828616

Hanratty, B. (2019). *ICD9CMtoICD10CM [Online]* (Accessed March, 2, 2021)

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer

Heldt, F. S., Vizcaychipi, M. P., Peacock, S., Cinelli, M., McLachlan, L., Andreotti, F., et al. (2021). Early Risk Assessment for COVID-19 Patients From Emergency Department Data Using Machine Learning. *Sci. Rep.* 11 (1), 4200. doi:10.1038/s41598-021-83784-y

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *J. Comput. Graphic. Stat.* 15 (3), 651–674. doi:10.1198/106186006X133933

Hu, C., Liu, Z., Jiang, Y., Shi, O., Zhang, X., Xu, K., et al. (2021). Early Prediction of Mortality Risk Among Patients With Severe COVID-19, Using Machine Learning. *Int. J. Epidemiol.* 49 (6), 1918–1929. doi:10.1093/ije/dyaa171

Huang, Y., Radenkovic, D., Perez, K., Nadeau, K., Verdin, E., Furman, D., et al. (2021). Modeling Predictive Age-Dependent and Age-Independent Symptoms and Comorbidities of Patients Seeking Treatment for COVID-19: Model Development and Validation Study. *J. Med. Internet Res.* 23 (3), e25696. doi:10.2196/25696

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., et al. (2020). Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China. *Lancet* 395 (10223), 497–506. doi:10.1016/S0140-6736(20)30183-5

Jehi, L., Ji, X., Milinovich, A., Erzurum, S., Merlino, A., Gordon, S., et al. (2020a). Development and Validation of a Model for Individualized Prediction of Hospitalization Risk in 4,536 Patients with COVID-19. *PLoS One* 15 (8), e0237419. doi:10.1371/journal.pone.0237419

Jehi, L., Ji, X., Milinovich, A., Erzurum, S., Rubin, B. P., Gordon, S., et al. (2020b). Individualizing Risk Prediction for Positive Coronavirus Disease 2019 Testing. *Chest* 158 (4), 1364–1375. doi:10.1016/j.chest.2020.05.580

Joshi, R. P., Pejaver, V., Hammarlund, N. E., Sung, H., Lee, S. K., Furmanchuk, A., et al. (2020). A predictive Tool for Identification of SARS-CoV-2 PCR-Negative Emergency Department Patients Using Routine Test Results. *J Clin. Virol.* 129, 104502. doi:10.1016/j.jcv.2020.104502

Kohavi, R., Brodley, C., and Frasca, B. (2000). KDD-Cup 2000 Organizers' Report: Peeling the Onion. *ACM SIGKDD Explorations Newsletter* 2 (2), 86–98. doi:10.1145/380995.381033

Kaufman, S., Rosset, S., Perlich, C., and Stitelman, O. (2012). Leakage in Data Mining: Formulation, Detection, and Avoidance. *ACM Trans Knowl Discov Data* 6, 563–556. doi:10.1145/2382577.2382579

Kullar, R., Marcelin, J. R., Swartz, T. H., Piggott, D. A., Macias Gil, R., Mathew, T. A., et al. (2020). Racial Disparity of Coronavirus Disease 2019 in African American Communities. *J. Infect. Dis.* 222 (6), 890–893. doi:10.1093/infdis/jiaa372

Liang, W., Liang, H., Ou, L., Chen, B., Chen, A., Li, C., et al. (2020). Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients with COVID-19. *JAMA Intern. Med.* 180 (8), 1081–1089. doi:10.1001/jamainternmed.2020.2033

Liu, H., Chen, S., Liu, M., Nie, H., and Lu, H. (2020). Comorbid Chronic Diseases Are Strongly Correlated with Disease Severity Among COVID-19 Patients: A Systematic Review and Meta-Analysis. *Aging Dis.* 11 (3), 668–678. doi:10.14336/AD.2020.0502

Mitchell, T. (1997). *Machine Learning*. New York: McGraw Hill.

Nami, M., Gadad, B. S., Chong, L., Ghumman, U., Misra, A., Gadad, S. S., et al. (2020). The Interrelation of Neurological and Psychological Symptoms of COVID-19: Risks and Remedies. *J Clin Med.* 9 (8), 2624. doi:10.3390/jcm9082624

NCATS (2020). *COVID-19 Clinical Data Warehouse Data Dictionary Based on OMOP Common Data Model Specifications*. Version 5.3.1

Nisbet, R., Elder, J., and Miner, J. (2009). *Handbook of Statistical Analysis and Data Mining Applications*. Academic Press.

Nishiga, M., Wang, D. W., Han, Y., Lewis, D. B., and Wu, J. C. (2020). COVID-19 and Cardiovascular Disease: from Basic Mechanisms to Clinical Perspectives. *Nat. Rev. Cardiol.* 17 (9), 543–558. doi:10.1038/s41569-020-0413-9

O'Malley, K. J., Cook, K. F., Price, M. D., Wildes, K. R., Hurdle, J. F., and Ashton, C. M. (2005). Measuring Diagnoses: ICD Code Accuracy. *Health Serv. Res.* 40 (5 Pt 2), 1620–1639. doi:10.1111/j.1475-6773.2005.00444.x

Oetjens, M. T., Luo, J. Z., Chang, A., Leader, J. B., Hartzel, D. N., Moore, B. S., et al. (2020). Electronic Health Record Analysis Identifies Kidney Disease as the Leading Risk Factor for Hospitalization in Confirmed COVID-19 Patients. *PLoS One* 15 (11), e0242182. doi:10.1371/journal.pone.0242182

Osborne, T. F., Veigulis, Z. P., Arreola, D. M., Röösli, E., and Curtin, C. M. (2020). Automated EHR Score to Predict COVID-19 Outcomes at US Department of Veterans Affairs. *PLoS One* 15 (7), e0236554. doi:10.1371/journal.pone.0236554

Polosa, R., and Caci, G. (2020). COVID-19: Counter-intuitive Data on Smoking Prevalence and Therapeutic Implications for Nicotine. *Intern. Emerg. Med.* 15 (5), 853–856. doi:10.1007/s11739-020-02361-9

Popkin, B. M., Du, S., Green, W. D., Beck, M. A., Algaith, T., Herbst, C. H., et al. (2020). Individuals with Obesity and COVID-19: A Global Perspective on the Epidemiology and Biological Relationships. *Obes. Rev.* 21 (11), e13128. doi:10.1111/obr.13128

Rashedi, J., Mahdavi Poor, B., Asgharzadeh, V., Pourostadi, M., Samadi Kafil, H., Vegari, A., et al. (2020). Risk Factors for COVID-19. *Infez Med.* 28 (4), 469–474.

Schwab, P., Mehrjou, A., Parbhoo, S., Celi, L. A., Hetzel, J., Hofer, M., et al. (2021). Real-time Prediction of COVID-19 Related Mortality Using Electronic Health Records. *Nat. Commun.* 12 (1), 1058. doi:10.1038/s41467-020-20816-7

Skoda, E.-M., Bäuerle, A., Schweda, A., Dörrie, N., Musche, V., Hetkamp, M., et al. (2020). Severely Increased Generalized Anxiety, but Not COVID-19-Related Fear in Individuals with Mental Illnesses: A Population Based Cross-Sectional Study in Germany. *Int. J. Soc. Psychiatry.*, 20764020960773. doi:10.1177/0020764020960773

Szepannek, G. (2020). An Overview on the Landscape of R Packages for Credit Scoring. *arXiv XX*, 1–25.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodological)* 58 (1), 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Vaid, A., Somani, S., Russak, A. J., De Freitas, J. K., Chaudhry, F. F., Paranjpe, I., et al. (2020). Machine Learning to Predict Mortality and Critical Events in a Cohort of Patients with COVID-19 in New York City: Model Development and Validation. *J. Med. Internet Res.* 22 (11), e24018. doi:10.2196/24018

Wang, Q., Davis, P. B., Gurney, M. E., and Xu, R. (2021a). COVID-19 and Dementia: Analyses of Risk, Disparity, and Outcomes from Electronic Health Records in the US. *Alzheimer's Demen.* doi:10.1002/alz.12296

Wang, Q., Davis, P. B., and Xu, R. (2021b). COVID-19 Risk, Disparities and Outcomes in Patients with Chronic Liver Disease in the United States. *EClinicalMedicine* 31, 100688. doi:10.1016/j.eclinm.2020.100688

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., et al. (2020). Prediction Models for Diagnosis and Prognosis of Covid-19: Systematic Review and Critical Appraisal. *BMJ* 369, m1328. doi:10.1136/bmj.m1328

Yang, J. M., Koh, H. Y., Moon, S. Y., Yoo, I. K., Ha, E. K., You, S., et al. (2020). Allergic Disorders and Susceptibility to and Severity of COVID-19: A Nationwide Cohort Study. *J. Allergy Clin. Immunol.* 146 (4), 790–798. doi:10.1016/j.jaci.2020.08.008

Zdravevski, E., Lameski, P., and Kulakov, A. (2011). Weight of Evidence as a tool for Attribute Transformation in the Preprocessing Stage of Supervised Learning Algorithms in: *The 2011 International Joint Conference on Neural Networks*, 181–188.

Zhao, Z., Chen, A., Hou, W., Graham, J. M., Li, H., Richman, P. S., et al. (2020). Prediction Model and Risk Scores of ICU Admission and Mortality in COVID-19. *PLoS One* 15 (7), e0236618. doi:10.1371/journal.pone.0236618

Zou, H., and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc B* 67 (2), 301–320. doi:10.1111/j.1467-9868.2005.00503.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 22 June 2021
doi: 10.3389/frai.2021.673527

Symptom Prediction and Mortality Risk Calculation for COVID-19 Using Machine Learning

Elham Jamshidi[1†], Amirhossein Asgary[2†], Nader Tavakoli[3†], Alireza Zali[1], Farzaneh Dastan[4], Amir Daaee[5], Mohammadtaghi Badakhshan[6], Hadi Esmaily[4], Seyed Hamid Jamaldini[7], Saeid Safari[1], Ehsan Bastanhagh[8], Ali Maher[9], Amirhesam Babajani[10], Maryam Mehrazi[3], Mohammad Ali Sendani Kashi[11], Masoud Jamshidi[12], Mohammad Hassan Sendani[13], Sahand Jamal Rahi[14] and Nahal Mansouri[15,16,17]**

[1]Functional Neurosurgery Research Center, Shohada Tajrish Comprehensive Neurosurgical Center of Excellence, Shahid Beheshti University of Medical Sciences, Tehran, Iran, [2]Department of Biotechnology, College of Sciences, University of Tehran, Tehran, Iran, [3]Trauma and Injury Research Center, Iran University of Medical Sciences, Tehran, Iran, [4]Department of Clinical Pharmacy, School of Pharmacy, Shahid Beheshti University of Medical Sciences, Tehran, Iran, [5]School of Mechanical Engineering, Sharif University of Technology, Tehran, Iran, [6]School of Electrical and Computer Engineering, Engineering Faculty, University of Tehran, Tehran, Iran, [7]Department of Genetic, Faculty of Advanced Science and Technology, Tehran Medical Sciences, Islamic Azad University, Tehran, Iran, [8]Department of Anesthesiology, Tehran University of Medical Sciences, Tehran, Iran, [9]School of Management and Medical Education, Shahid Beheshti University of Medical Sciences, Tehran, Iran, [10]Department of Pharmacology, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran, [11]Department of Business Management, Faculty of Management, University of Tehran, Tehran, Iran, [12]Department of Exercise Physiology, Tehran University, Tehran, Iran, [13]Department of Computer Engineering, Sharif University of Technology, Tehran, Iran, [14]Laboratory of the Physics of Biological Systems, Institute of Physics, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, [15]Division of Pulmonary Medicine, Department of Medicine, Lausanne University Hospital (CHUV), University of Lausanne (UNIL), Lausanne, Switzerland, [16]Swiss Institute for Experimental Cancer Research (ISREC), School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, [17]Research Group on Artificial Intelligence in Pulmonary Medicine, Division of Pulmonary Medicine, Lausanne University Hospital (CHUV), Lausanne, Switzerland*

OPEN ACCESS

Edited by:
Hong Qin, University of Tennessee at Chattanooga, United States

Reviewed by:
Ziwei Ma, University of Tennessee at Chattanooga, United States
Feng Liu, Stevens Institute of Technology, United States

***Correspondence:**
Nahal Mansouri
nahal.mansouri@chuv.ch
Sahand Jamal Rahi
sahand.rahi@epfl.ch

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *27 February 2021*
Accepted: *14 May 2021*
Published: *22 June 2021*

Citation:
Jamshidi E, Asgary A, Tavakoli N, Zali A, Dastan F, Daaee A, Badakhshan M, Esmaily H, Jamaldini SH, Safari S, Bastanhagh E, Maher A, Babajani A, Mehrazi M, Sendani Kashi MA, Jamshidi M, Sendani MH, Rahi SJ and Mansouri N (2021) Symptom Prediction and Mortality Risk Calculation for COVID-19 Using Machine Learning. Front. Artif. Intell. 4:673527. doi: 10.3389/frai.2021.673527

Background: Early prediction of symptoms and mortality risks for COVID-19 patients would improve healthcare outcomes, allow for the appropriate distribution of healthcare resources, reduce healthcare costs, aid in vaccine prioritization and self-isolation strategies, and thus reduce the prevalence of the disease. Such publicly accessible prediction models are lacking, however.

Methods: Based on a comprehensive evaluation of existing machine learning (ML) methods, we created two models based solely on the age, gender, and medical histories of 23,749 hospital-confirmed COVID-19 patients from February to September 2020: a symptom prediction model (SPM) and a mortality prediction model (MPM). The SPM predicts 12 symptom groups for each patient: respiratory distress, consciousness disorders, chest pain, paresis or paralysis, cough, fever or chill, gastrointestinal symptoms, sore throat, headache, vertigo, loss of smell or taste, and muscular pain or fatigue. The MPM predicts the death of COVID-19-positive individuals.

Results: The SPM yielded ROC-AUCs of 0.53–0.78 for symptoms. The most accurate prediction was for consciousness disorders at a sensitivity of 74% and a specificity of 70%.

Abbreviations: AI, artificial intelligence; COVID-19, coronavirus disease of 2019; intensive care unit, IQR; interquartile range, KS; Kolmogorov-Smirnov, LR; logistic regression, ML; machine learning, RF; random forest, RDW; red blood cell, distribution width; ROC, receiver operating characteristic; HIS, hospital information system.

2,440 deaths were observed in the study population. MPM had a ROC-AUC of 0.79 and could predict mortality with a sensitivity of 75% and a specificity of 70%. About 90% of deaths occurred in the top 21 percentile of risk groups. To allow patients and clinicians to use these models easily, we created a freely accessible online interface at www.aicovid.net.

Conclusion: The ML models predict COVID-19-related symptoms and mortality using information that is readily available to patients as well as clinicians. Thus, both can rapidly estimate the severity of the disease, allowing shared and better healthcare decisions with regard to hospitalization, self-isolation strategy, and COVID-19 vaccine prioritization in the coming months.

Keywords: COVID-19, artificial intelligence, machine learning, symptom, mortality

INTRODUCTION

The COVID-19 pandemic of the 2019 novel coronavirus (SARS-CoV-2) started in December 2019 and is spreading rapidly, with approximately 62.5 million confirmed cases and 1.5 million deaths by the end of November 2020 (WHO, 2020).

The severity of the disease varies widely between different patients, ranging from no symptoms to a mild flu-like illness, to severe respiratory symptoms, and to multi-organ failure leading to death. Among the symptoms, fever, cough, and respiratory distress are more prevalent than symptoms such as consciousness disorders and loss of smell and taste (Tabata et al., 2020; Jamshidi et al., 2021b). In general, complications are common among elderly patients and those with pre-existing conditions. The intensive care unit (ICU) admission rate is substantially higher for these groups (Abate et al., 2020; Jamshidi et al., 2021a).

The Center for Disease Control (CDC) and the World Health Organization (WHO) consider the identification of individuals at higher risk a top priority. This identification could be used for numerous solutions to moderate the consequences of the pandemic for the most vulnerable (CDC COVID-19 Response Team, 2020) as well as minimize the presence of actively ill patients in society.

This requires the prediction of the symptoms and mortality risk for infected individuals. While symptom prediction models exist for cancer, no such models have been designed for COVID-19 (Levitsky et al., 2019; Goecks et al., 2020). To make rapid, evidence-based decisions possible, they will ideally be based on readily available patient information, i.e., demographic attributes and past medical history (PMH) as opposed to costly laboratory tests. Early decision-making is critical for timely triage and clinical management of patients. For instance, clinical and laboratory data can only be assessed after presenting the individual to a health care center, increasing the risk of unnecessary exposures to the virus and increasing costs (Sun et al., 2020). These parameters are not available immediately and are partly subject to human error. Also, factors like genetic predisposition may increase the models' accuracy but are not broadly available.

With the growth of big data in healthcare and the introduction of electronic health records, artificial intelligence (AI) algorithms can be integrated into hospital IT systems and have shown promise as computer-aided diagnosis and prognostic tools. In the era of COVID-19, AI has played an essential role in the early diagnosis of infection, the prognosis of hospitalized patients, contact tracing for spread control, and drug discovery (Lalmuanawma et al., 2020). AI methods can have a higher accuracy over classical statistical analyses.

In contrast to the few previously available COVID-19 risk scales, our mortality prediction model uses a selection of variables that are in principle accessible to all patients and thus can be used immediately after diagnosis (Assaf et al., 2020; Pan et al., 2020). This model not only has a significant benefit in early decision making in the hospital setting, but because it does not require clinicians or laboratories, it can serve as a triage tool for patients in an outpatient setting, in telemedicine, or as a self-assessment tool. For example, decisions on outpatient vs. inpatient care can be made remotely by estimating the most probable symptoms and severity risks. This lessens the strain on health care resources, unnecessary costs, and unwanted exposures to infected patients.

Here, we implemented 2 ML methods to predict the symptoms and the mortality of patients with COVID-19. Overall, 23,749 patients were included in the study. The predictors used for the models were age, sex, and PMH of the patients. Both of these models achieved predictions with high accuracy. To our knowledge, this is one of the largest datasets of COVID-19 cases and is the only study that uses patient-available data for the prediction of COVID-19 symptoms and mortality. Furthermore, this study is the most extensive study for mortality prediction for COVID-19 using ML-based on any set of predictors (An et al., 2020; Gao et al., 2020; Vaid et al., 2020; Yadaw et al., 2020).

We also created an online calculator where each individual can predict their COVID-19 related symptoms and risk (www.aicovid.net).

For a standardized representation of the methodology and results of this analysis an adapted version of the Transparent Reporting of Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guideline was followed (Collins et al., 2015).

TABLE 1 | List of predictors**.** Predictor variables for mortality risk and symptom prediction of COVID-19.

Category	Variable	Description
Demographic	Age	In years
	Sex	Male or female
Past/Current Medical Conditions	*Cancer*	Current chemotherapy, radiotherapy, immunotherapy, bone marrow or stem cell transplantation
	Liver disorders	Chronic hepatitis (type B or C), alcohol-related liver disease, primary biliary cirrhosis, primary sclerosing cholangitis, hemochromatosis, cirrhosis
	Blood disorders	Anemia (iron deficiency, thalassemia minor and major, sickle cell disease), coagulopathies (hemophilia and platelet disorders)
	Immune disorders	Immune deficiency (acquired immunodeficiency syndrome, treatment with steroids and immune suppressors), autoimmune disease (rheumatoid arthritis, systemic lupus erythematosus, ankylosing spondylitis, vasculitis).
	Cardiovascular disease	Congestive heart failure, cardiovascular events (myocardial infarction, stroke, angina), valvular heart disease, arrhythmia (e.g. atrial fibrillation)
	Kidney disorders	Chronic kidney disease (stage 3, 4, and end-stage renal disease)
	Respiratory disorders	Asthma, chronic obstructive pulmonary disease (emphysema and chronic bronchitis), extrinsic allergic alveolitis, cystic fibrosis, interstitial lung disease, sarcoidosis, bronchiectasis, pulmonary hypertension
	Neurological disorders	Epilepsy, Parkinson's disease, motor neuron disease, cerebral palsy, dementia, multiple sclerosis
	Endocrine disorders	Hyperthyroidism, hypothyroidism, cushing syndrome, pheochromocytoma, thyroiditis, hyperaldosteronism
	Diabetes mellitus	Type 1 and type 2 diabetes, maturity onset diabetes of the young, insipidus, gestational diabetes
	Hypertension	Primary and secondary
	Psychiatric disorders (removed due to low prevalence)	Bipolar disorder, psychosis, schizophrenia, major depression disorder
	Thrombosis (removed due to low prevalence)	Venous thromboembolism, pulmonary thromboembolism

METHODS

Source of Data and Participants

In this cohort study, we used the Hospital Information System (HIS) of 74 secondary and tertiary care hospitals across Tehran, Iran. The eligibility criteria were defined as confirmed or suspected SARS-CoV-2 infections of people aged 18–100 years registered in the referred HIS. The final database used to design the models was obtained by aggregating the 74 hospitals' HIS. The study included patients referred to any of the hospitals between February 1, 2020, and September 30, 2020. Patients were followed up through October 2020 until all the registered patients had the specific death or survival outcome needed for the mortality prediction model (MPM). This study was approved by the Iran University of Medical Sciences Ethics Committee.

Outcome

Symptom Prediction Model

The patients' symptoms at the time of admission, as recorded in the HIS, were considered as the outputs of the Symptom Prediction Model (SPM). All stated symptoms were clustered in 12 categories to be predicted by the model. The groups are cough, loss of smell or taste, respiratory distress, vertigo, muscular pain or fatigue, sore throat, fever or chill, paresis or paralysis, gastrointestinal problems, headache, chest pain, and consciousness disorders.

Mortality Prediction Model

Death or survival as per the HIS records was defined as the output of the mortality prediction model (MPM).

Predictors

The patients' age, sex, and past medical history (PMH), as detailed in **Table 1**, were used as predictors for both models. The selection of variables as predictors was based on the available recorded data. All these predictors were recorded in the HIS at the time of admission.

Missing Data

We only included patients with the required data. Due to the absence of missing data, there was no imputation of missing values.

Pre-Processing

Symptoms and predictor variables from the medical histories with an incidence of less than 0.2% were removed to reduce noise. This removed past COVID-19 infections, thrombosis, psychiatric disorders, and organ or bone marrow transplantation from the set of predictor variables. The removed symptoms were tachycardia, seizure, nasal congestion, and skin problems.

Sex, PMH, and symptoms were encoded as binary variables. In training and test sets, the only continuous predictor, age, was standardized to zero mean and unit standard deviation.

TABLE 2 | Patient characteristics and symptoms. Baseline characteristics, symptoms, and death outcomes for COVID-19 patients.

Continuous variables	
Variable	**Median (±IQR)**
Age	52 (±29)
Categorical/Binary variables	
Variable	Count (percent)
Sex	
Male	12,597 (53.04%)
Female	11,152 (46.96%)
Cardiovascular disease	2,471 (10.4%)
Diabetes	2,068 (8.71%)
Hypertension	2,004 (8.44%)
Respiratory diseases	546 (2.3%)
Cancer	477 (2.01%)
Kidney disorders	416 (1.75%)
Neurological disorders	264 (1.11%)
Immune disorders	178 (0.75%)
Blood disorders	152 (0.64%)
Current pregnancy	139 (0.59%)
Liver disorders	119 (0.5%)
Endocrine disorders	97 (0.41%)
Organ or bone marrow transplant	29 (0.12%)
Mental illnesses	19 (0.08%)
Thrombosis	15 (0.06%)
Past COVID-19 infection	10 (0.04%)
Outcomes	
Survived	21,309 (89.73%)
Dead	2,440 (10.27%)
Symptoms	
Cough	11,995 (50.51%)
Respiratory distress	10,342 (43.55%)
Muscular pain or fatigue	9,249 (38.94%)
Fever or chill	8,553 (36.01%)
Gastrointestinal problems	2,469 (10.4%)
Headache	1,120 (4.72%)
Chest pain	745 (3.14%)
Consciousness disorders	698 (2.94%)
Loss of smell or taste	659 (2.77%)
Vertigo	501 (2.11%)
Sore throat	157 (0.66%)
Paresis or paralysis	121 (0.51%)

Machine Learning Methods

To ensure generalizability, a 5-fold cross-validation algorithm was employed [Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation, (Wong, 2015). All records were randomly separated into five independent subsets. Four subsets were used as training data, and one subset was retained as a validation set for model testing. The cross-validation process was then iterated four more times, with each of the five subsets being used as validation data exactly once. Subsequently, model performance metrics were evaluated for training and validation groups separately in each model iteration.

By separating deceased and surviving patients separately into five mortality-stratified subsets first and then combining these into the final five subsets, we maintained the same proportion of deceased and surviving patients in each of the final five subsets.

We evaluated several machine learning techniques for both models: Logistic Regression, Random Forest, Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), and Naive Bayes.

We took advantage of the Scikit-learn machine learning library to implement both preprocessing algorithms and models (Garreta and Moncecchi, 2013).

Symptom Prediction Model

The SPM output predicts symptoms for SARS-CoV-2 positive patients. Since there are 12 symptom groups, we judged the models' overall performance by a single metric, the prevalence-weighted mean of the twelve ROC-AUCs (Mandrekar, 2010), in which the ROC-AUCs were weighted by symptom prevalence.

Mortality Prediction Model

The MPM calculates the probability of death for SARS-CoV-2 positive patients. Each model's performance was measured in terms of a ROC-AUC.

RESULTS

Participants

Baseline characteristics of patients and their symptoms are shown in **Table 2**. Of all 23,749 confirmed or suspected COVID-19 patients, 2,440 (10.27%) passed away at the end of the study (see *Discussion*). A comparison of the characteristics of survived and deceased patients is shown in **Table 3**. A comparison of the characteristics of patients with and without each symptom is shown in **Supplementary Tables S1–S16**.

We used statistical hypothesis tests to demonstrate each predictor variable's significance to the model outputs. We employed the F-test (Snedecor, 1957) technique for age, a continuous variable, and the Chi-square (Snedecor, 1957) technique for other categorical variables such as sex and PMH.

Model Specification

We evaluated six machine learning methods for both the SPM and MPM, which are listed, together with the hyperparameters used in **Table 4**.

Model Performance

Symptom Prediction Model

The SPM can be considered as 12 separated classifiers; each predicts the occurrence of a specific symptom. While the performance of each sub-classifier can be evaluated separately, the overall performance can be assessed using the prevalence-weighted mean of the ROC-AUCs, since the symptoms have different prevalence. The prevalence-weighted mean ROC-AUC for each method is illustrated in **Figure 1**. Although the KNN method provided the highest weighted mean ROC-AUC for the test data, it was the least robust method since its performance varied considerably for different validation folds (note standard deviation bars). The Random Forest method achieved better overall performance and robustness. The weighted mean ROC-AUC value of this method was 0.582 for the test data.

TABLE 3 | Comparison between survived and deceased patient groups. Comparative evaluation of the characteristics of survived and deceased COVID-19 patients.

Continuous variables				
Variable	**Median in survivors (±IQR)**	**Median in deceased (±IQR)**	**F-test statistics**	**F-test *p*-value**
Age	49 (±27)	70 (±21)	2,039.47	<0.001
Categorical/Binary variables				
Variable	Count in survivors (percent in survivors)	Count in deceased (percent in deceased)	Chi2 statistics	Chi2 *p*-value
Sex				
Male	11,163 (52.39%)	1,434 (58.77%)	16.82	<0.001
Female	10,146 (47.61%)	1,006 (41.23%)	19	<0.001
Cardiovascular disease	2,039 (9.57%)	432 (17.7%)	139.29	<0.001
Diabetes	1,693 (7.94%)	375 (15.37%)	138.57	<0.001
Hypertension	1,676 (7.87%)	328 (13.44%)	80.71	<0.001
Respiratory disorder**s**	462 (2.17%)	84 (3.44%)	15.47	<0.001
Cancer	343 (1.61%)	134 (5.49%)	164.28	<0.001
Kidney disorders	317 (1.49%)	99 (4.06%)	82.54	<0.001
Neurological disorders	207 (0.97%)	57 (2.34%)	36.68	<0.001
Immune disorders	152 (0.71%)	26 (1.07%)	3.62	0.057
Blood disorders	112 (0.53%)	40 (1.64%)	42.43	<0.001
Current pregnancy	133 (0.62%)	6 (0.25%)	5.35	0.021
Liver disorders	101 (0.47%)	18 (0.74%)	3.04	0.081
Endocrine disorders	88 (0.41%)	9 (0.37%)	0.1	0.747
Organ or bone marrow transplant	25 (0.12%)	4 (0.16%)	0.39	0.533
Psychiatric disorders	16 (0.08%)	3 (0.12%)	0.63	0.428
Thrombosis	13 (0.06%)	2 (0.08%)	0.15	0.696
Past COVID-19 infection	10 (0.05%)	0 (0.0%)	1.15	0.285

TABLE 4 | Machine learning methods and hyperparameters used.

The Mortality Prediction Model		
Method	**Parameter**	**Value(s)**
Logistic Regression	C	1.0
Random Forest	Number of trees	500
	Min. Number of samples at a leaf node	%0.1 of all samples
	Criterion	Gini
Artificial Neural Networks	Number of layers	3
	Output space dimensionality for each layer	32, 16, 1
	Activation function for each layer	Tanh, tanh, sigmoid
K-Nearest Neighbors	K	10
	Weight function	Distance
Linear Discriminant Analysis	Solver	SVD
Naive Bayes	Interval size of age categories	0.1
The Symptom Prediction Model		
Method	Parameter	Value
Logistic Regression	C	1.0
Random Forest	Number of trees	200
	Min. Number of samples at a leaf node	%0.1 of all samples
	Criterion	Gini
Artificial Neural Network	Number of layers	4
	Output space's dimensionality for each layer	32, 32, 32, 12
	Activation function for each layer	Tanh, tanh, tanh, tanh, sigmoid
K-Nearest Neighbors	K	5
	Weight function	Distance
Linear Discriminant Analysis	Solver	SVD
Naive Bayes	Interval size of age categories	0.1

Moreover, the performance of the SPM can be evaluated for each symptom separately. The ROC-AUC values for predicting consciousness disorder, paresis or paralysis, and chest pain were 0.785, 0.729, and 0.686, respectively. Also, at a specificity of 70%, the sensitivities were 73%, 50%, and 53%, respectively.

As shown in **Figure 1**, the random forest model with a mean ROC-AUC of 0.8 and 0.79 has the highest efficiency in the

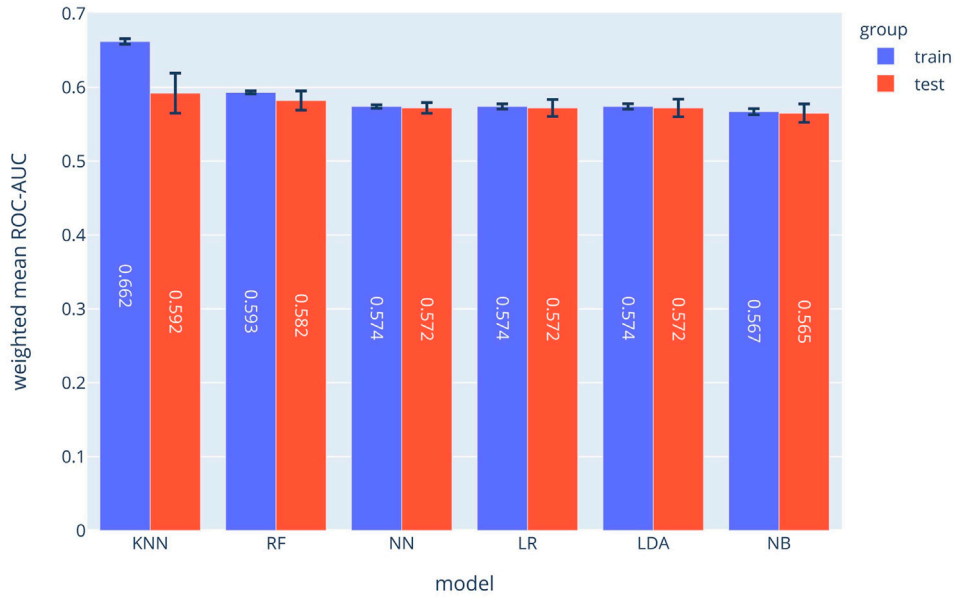


FIGURE 1 | Prevalence-weighted means ROC-AUCs for different ML models. The models were used to implement the Symptom Prediction Model (SPM). Error bars denote the standard deviation over different cross-validation folds.

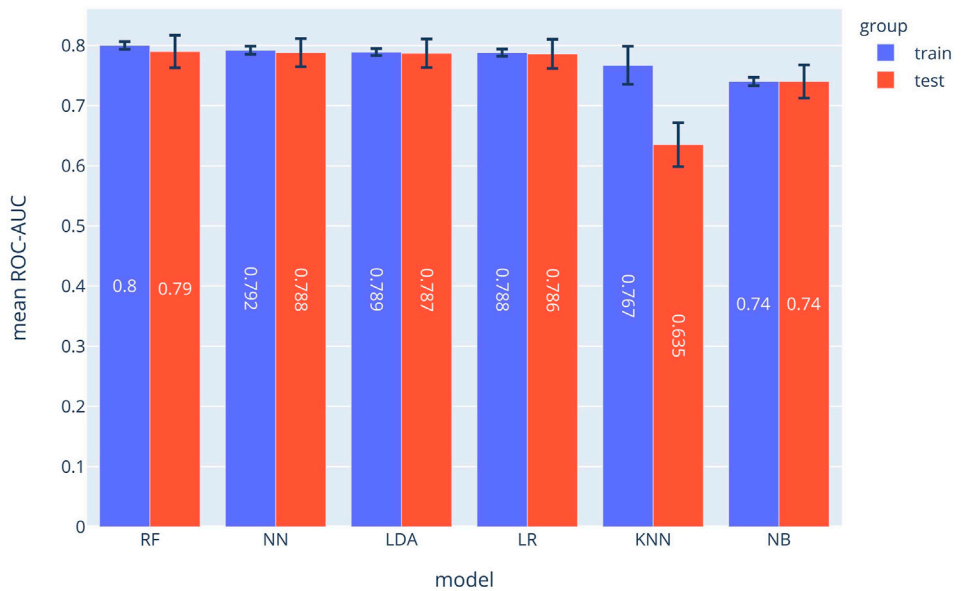


FIGURE 2 | ROC-AUCs of different ML models which were used to implement the MPM. The Random Forest (RF) model outperformed the other approaches.

training and the validation groups, respectively, followed by the Neural Network and LDA. In the symptom prediction model, the ROC-AUC values of all models in addition to the weighted average of ROC-AUC of different ML methods for each symptom are shown in **Supplementary Figure S1**. **Supplementary Figure S2** delineates each method's performance for all symptoms as a Radar chart.

Based on the ROC diagram and the information from the database, the other performance metrics of the other models were identified. In addition to the ROC-AUC of the risk prediction model, we calculated the sensitivity and the negative and positive predictive value (NPV and PPV respectively) for each model. The detailed results of all six algorithms for both MPM and SPM are shown in **Supplementary Tables S17 and S18**.

The calibration plot of the RF implementation for each symptom predictor (sub-classifier) is depicted in **Supplementary Figure S3** which shows the calibration plot of the RF implementation of each symptom.

Mortality Prediction Model

The ROC-AUC values for each method are depicted in **Figure 2**. In the MPM classifier, the Random Forest method outperformed the other methods just as for the SPM. The achieved ROC-AUC value was 0.79 for the test data.

Supplementary Figure S4 shows ROC diagrams representing the true-positive rates vs. false-positive rates for each method used to implement the MPM. The calibration plot of the RF model is depicted in **Supplementary Figure S5**. Calibration indications such as Mean Calibration Error are also shown in the **Supplementary Figure S5** for different methods.

Model Input-Output Correlations

We used the Chi-square test and the F-test to evaluate the extent to which PMH, sex, or age predict the outputs of the SPM and MPM. The larger the values of these test values are for each predictor variable, the more the predictor variable is predictive of the output of the models. For categorical predictor variables (i.e., PMH and sex), the Chi-square hypothesis test was used. To evaluate the predictive value of a categorical variable, we examined whether it was more common in patients who died (MPM) or in patients with a particular symptom (SPM). For the only continuous variable (age), we used the F-test. To find the impact of age, we examined if the age median was higher in dead patients (MPM) or patients with a particular symptom (SPM).

For the SPM model, **Supplementary Figure S6** shows how each factor in the PMH was correlated with each symptom using the Chi-square test. For example, patients with diabetes or cardiovascular disease were more likely to have consciousness disorders and chest pain in case of infection with COVID-19. The effect of age on each symptom is shown in **Supplementary Figure S7** using the F-test. Older patients were more likely to develop symptoms such as respiratory distress and consciousness disorder but also less likely to develop symptoms such as muscular pain or fatigue.

In addition, for the MPM, the impact of each PMH on death is shown in **Supplementary Figure S8**. In our analysis, cancer, cardiovascular disease, and diabetes have the greatest effects on the risk of death in patients with Covid 19; on the other hand, pregnancy or being female decreased the chances of death. The F-test statistic of age in the MPM model is 2,039.47, which explains the increase in mortality risk from aging.

DeLong's test shows the statistically significant difference between AUCs of models. The DeLong tests for the MPM and SPM predictions are shown through **Supplementary Figures S9–S21**.

Validation of the Model for Each Mortality Peak

For additional validation of our model, we evaluated the performance of the final random forest for MPM during the periods with the highest rate of mortality. The data corresponding to each available mortality peak (april, February, and September 2020) was selected from the validation dataset of each model, and the outcome (recovery or death of the patient) during each period was predicted by the model and shown as a ROC diagram (**Figure 3**). Despite the variation of the AUC in the mortality peaks, the weighted average of the AUC values corresponding to each period was approximately equal to the average model yield for the entire data. We can conclude that the model continues to perform equally well during each mortality peak. The cause of the high yield in april could be explained by the large number of available samples which would allow the algorithm to learn more accurately.

DISCUSSION

Our objective in this study was to develop 2 ML models to predict the mortality and symptoms of COVID-19-positive patients among the general population using age, gender, and comorbidities alone. These models can guide the design of measures to combat the COVID-19 pandemic. The prediction of vulnerability using the models allows people in different risk groups to take appropriate actions if they contract COVID-19. For example, people who fall into the low-risk group can start isolation sooner when the predicted symptoms appear and refer to a hospital only if the symptoms persist. As a result, the risk of disease spread and the pressure on the health care system from unnecessary hospital visits, costs, and psychological and physical stress to the medical staff could be reduced (Emanuel et al., 2020). In contrast, people who are predicted to be at higher risk are recommended to seek medical care immediately. Predictions can speed up the treatment process and ultimately decrease mortality.

Our study has shown that multiple symptoms have strong correlations with different medical history factors. Symptoms can be either amplified or attenuated by health backgrounds; for instance, hypertension, diabetes, and respiratory and neurological disorders increased the chances of loss of smell or taste; however, pregnancy, cancer, higher age, cardiovascular disease, and liver, immune system, blood, and kidney disorders have attenuated the appearance of this symptom.

Due to the complexity of the COVID-19 pathogenesis, many clinical studies revealed contradictory results, for example, the effectiveness or ineffectiveness of remdisivir (Beigel et al., 2020; Goldman et al., 2020; Wang et al., 2020). We hypothesize that the imbalance of mortality risks between the intervention and control groups could have been a problem in these studies. With the help of our model, such problems could be partially solved by equalizing the mortality baseline in different clinical groups.

Our AI models can also be beneficial for COVID-19 vaccine testing and prioritization strategies. The limited number of approved vaccines in the first months of the vaccination process and the potential shortages make vaccine prioritization inevitable. This prioritization would be more important for developing countries that do not have the resources to pre-

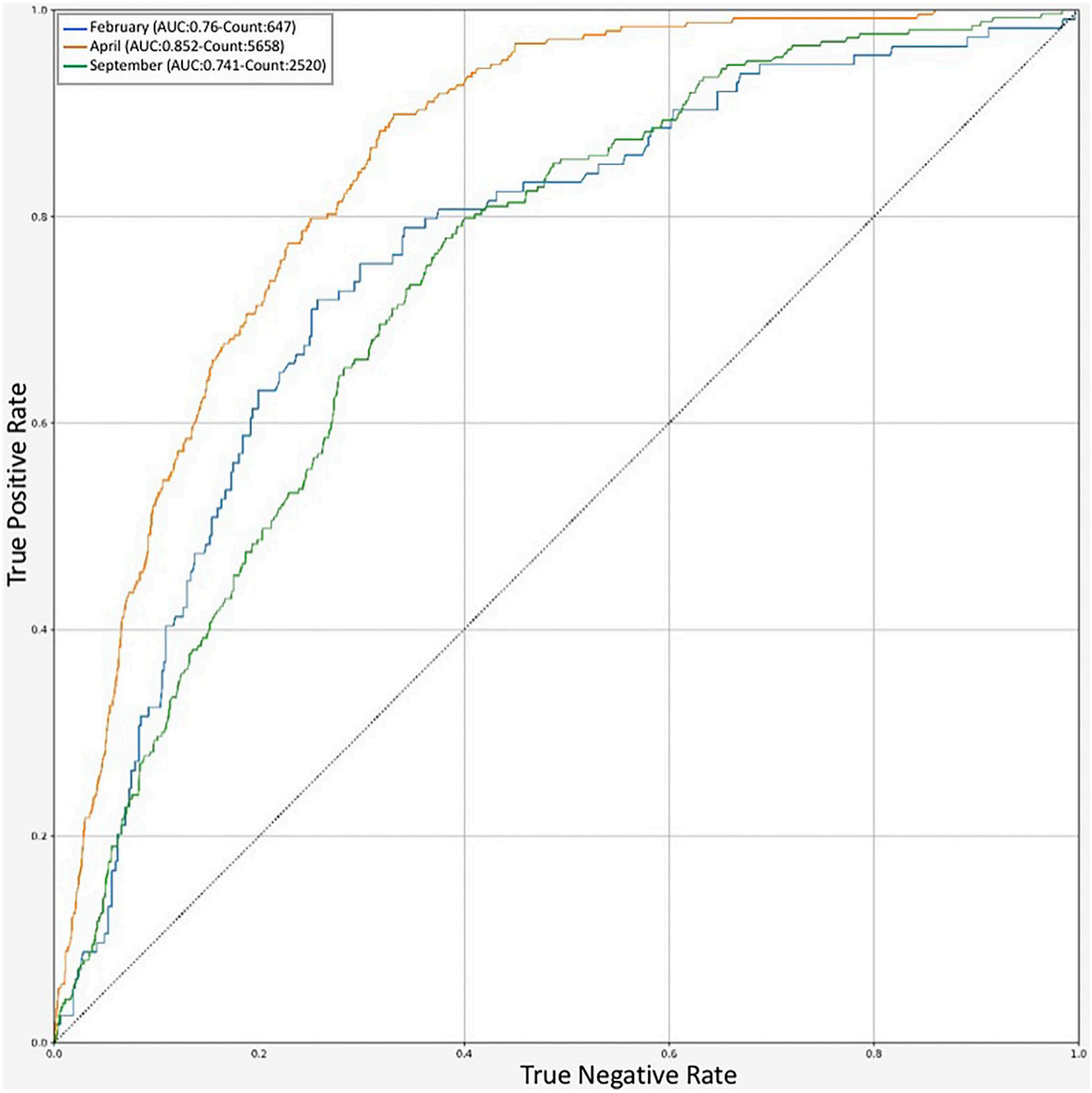


FIGURE 3 | ROC chart for Prediction of RF models in different timelines**.** An indicator of model performance on validation dataset in the peak months of COVID-19 outbreak in Iran.

order vaccines from multiple companies (Persad et al., 2020). Having a mortality prediction tool for each individual could be a valuable tool for governments to decide on vaccines' allocation.

Limitations

Since our dataset was collected by the HIS, it did not contain COVID-19 patients that did not refer to a hospital or had no major symptoms to be identified as infected. This could explain the high mortality rate in our and other studies (Fumagalli et al., 2020; Gue et al., 2020; Yadaw et al., 2020). However, for a systematic study with few confounding variables, uniform data collection is essential, which can only be realistically ensured with hospital data.

Also, other variables such as the viral load may be important but are difficult to measure and are not readily available. We opted for easily accessible predictor variables to allow the widespread use of the models.

One way to improve the models is to subgroup-specific factors in the medical history or specific symptoms further. The main

reason for grouping factors and symptoms was the low prevalence of certain subsets in the dataset.

In conclusion, we evaluated 15 parameters (**Table 1**) for predicting the symptoms and the mortality risk of COVID-19 patients. The ML models trained in this study could help people quickly determine their mortality risk and the probable symptoms of the infection. These tools could aid patients, physicians, and governments with informed and shared decision-making.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Iran National Committee for Ethics in Biomedical Research. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

EJ: Conceptualization-Equal, Methodology-Equal, Project administration-Equal, Writing-original draft-Equal, Writing-review and editing-Equal; AA: Conceptualization-Equal, Methodology-Equal, Project administration-Equal, Writing-original draft-Equal, Writing-review and editing-Equal; NT: Conceptualization-Equal, Methodology-Equal, Project administration-Equal, Writing-original draft-Equal, Writing-review and editing-Equal; AZ: Data curation-Equal, Investigation administration-Equal, Writing-original draft-Equal, Writing-review and editing-Equal; FD: Data curation-Equal, Investigation-Equal, Writing-original draft-Equal, Writing-review and editing-Equal; AD: Data curation-Equal, Investigation Equal, Writing-original draft-Equal, Writing-review and editing-Equal; MB: Data curation-Equal, Investigation-Equal, Writing-original draft-Equal, Writing-review and editing-Equal; HE: Data curation-Equal, Investigation-Equal, Writing-original draft-Equal, Writing-review and editing-Equal; SJ: Data curation-Equal, Investigation-Equal, Writing-original draft-Equal, Writing-review and editing-Equal; SS: Data curation-Equal, Investigation-Equal, Writing-original draft-Equal, Writing-review and editing-Equal; EB: Data curation-Equal, Investigation-Equal, Writing-original draft-Equal, Writing-review and editing-Equal; AM: Data curation-Equal, Investigation-Equal, Writing-original draft-Equal, Writing-review and editing-Equal; AB: Data curation-Equal, Investigation-Equal, Writing-original draft-Equal, Writing-review and editing-Equal; MM: Data curation-Equal, Investigation-Equal, Writing-original draft-Equal, Writing-review and editing-Equal; MS: Data curation-Equal, Investigation-Equal and Writing-original draft-Equal, Writing-review and editing-Equal; MJ: Data curation-Equal, Investigation-Equal and Writing-original draft-Equal; SJ: Conceptualization-Equal, Methodology-Equal, Project administration-Equal, Supervision-Equal, Writing-original draft-Equal, Writing-review and editing-Equal; NM: Conceptualization-Equal, Methodology-Equal, Project administration-Equal, Supervision-Equal, Writing-original draft-Equal, Writing-review and editing-Equal.

FUNDING

SJ thanks the École polytechnique fédérale de Lausanne for generous support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2021.673527/full#supplementary-material

REFERENCES

Abate, S. M., Ali, S. A., Mantfardo, B., and Basu, B. (2020). Rate of Intensive Care Unit Admission and Outcomes Among Patients with Coronavirus: A Systematic Review and Meta-Analysis. *PLoS One.* 15, e0235653. doi:10.1371/journal.pone.0235653

An, C., Lim, H., Kim, D.-W., Chang, J. H., Choi, Y. J., and Kim, S. W. (2020). Machine Learning Prediction for Mortality of Patients Diagnosed with COVID-19: a Nationwide Korean Cohort Study. *Sci. Rep.* 10, 1–11. doi:10.1038/s41598-020-75767-2

Assaf, D., Gutman, Y. a., Neuman, Y., Segal, G., Amit, S., Gefen-Halevi, S., et al. (2020). Utilization of Machine-Learning Models to Accurately Predict the Risk for Critical COVID-19. *Intern. Emerg. Med.* 15, 1435–1443. doi:10.1007/s11739-020-02475-0

Beigel, J. H., Tomashek, K. M., Dodd, L. E., Mehta, A. K., Zingman, B. S., Kalil, A. C., et al. (2020). Remdesivir for the Treatment of Covid-19 - Final Report. *N. Engl. J. Med.* 383, 1813–1826. doi:10.1056/nejmoa2007764

CDC COVID-19 Response Team (2020). Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) - United States, February 12-March 16, 2020. *MMWR. Morb. Mortal. Wkly. Rep.* 69(12), 343–346. doi:10.15585/mmwr.mm6912e2

Collins, G. S., Reitsma, J. B., Altman, D. G., and Moons, K. G. (2015). Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD Statement. *Br. J. Surg.* 102, 148–158. doi:10.1002/bjs.9736

Emanuel, E. J., Persad, G., Upshur, R., Thome, B., Parker, M., Glickman, A., et al. (2020). Fair Allocation of Scarce Medical Resources in the Time of Covid-19. *N. Engl. J. Med.* 382, 2049–2055. doi:10.1056/nejmsb2005114

Fumagalli, C., Rozzini, R., Vannini, M., Coccia, F., Cesaroni, G., Mazzeo, F., et al. (2020). Clinical Risk Score to Predict In-Hospital Mortality in COVID-19 Patients: a Retrospective Cohort Study. *BMJ Open.* 10, e040729. doi:10.1136/bmjopen-2020-040729

Gao, Y., Cai, G.-Y., Fang, W., Li, H.-Y., Wang, S.-Y., Chen, L., et al. (2020). Machine Learning Based Early Warning System Enables Accurate Mortality Risk

Prediction for COVID-19. *Nat. Commun.* 11, 1–10. doi:10.1038/s41467-020-18684-2

Garreta, R., and Moncecchi, G. (2013). *Learning Scikit-Learn: Machine Learning in Python*. Birmingham United Kingdom: Packt Publishing Ltd

Goecks, J., Jalili, V., Heiser, L. M., and Gray, J. W. (2020). How Machine Learning Will Transform Biomedicine. *Cell* 181, 92–101. doi:10.1016/j.cell.2020.03.022

Goldman, J. D., Lye, D. C. B., Hui, D. S., Marks, K. M., Bruno, R., Montejano, R., et al. (2020). Remdesivir for 5 or 10 Days in Patients with Severe Covid-19. *N. Engl. J. Med.* 383, 1827–1837. doi:10.1056/nejmoa2015301

Gue, Y. X., Tennyson, M., Gao, J., Ren, S., Kanji, R., and Gorog, D. A. (2020). Development of a Novel Risk Score to Predict Mortality in Patients Admitted to Hospital with COVID-19. *Sci. Rep.* 10, 1–8. doi:10.1038/s41598-020-78505-w

Jamshidi, E., Asgary, A., Tavakoli, N., Zali, A., Esmaily, H., Jamaldini, S. H., et al. (2021a). Using Machine Learning to Predict Mortality for COVID-19 Patients on Day Zero in the ICU. *bioRxiv*. doi:10.1101/2021.02.04.21251131

Jamshidi, E., Babajani, A., Soltani, P., and Niknejad, H. (2021b). Proposed Mechanisms of Targeting COVID-19 by Delivering Mesenchymal Stem Cells and Their Exosomes to Damaged Organs. *Stem Cel Rev Rep* 17, 176–192. doi:10.1007/s12015-020-10109-3

Lalmuanawma, S., Hussain, J., and Chhakchhuak, L. (2020). Applications of Machine Learning and Artificial Intelligence for Covid-19 (SARS-CoV-2) Pandemic: A Review. *Chaos, Solitons & Fractals*. 139, 110059. doi:10.1016/j.chaos.2020.110059

Levitsky, A., Pernemalm, M., Bernhardson, B.-M., Forshed, J., Kölbeck, K., Olin, M., et al. (2019). Early Symptoms and Sensations as Predictors of Lung Cancer: a Machine Learning Multivariate Model. *Sci. Rep.* 9, 16504. doi:10.1038/s41598-019-52915-x

Mandrekar, J. N. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *J. Thorac. Oncol.* 5, 1315–1316. doi:10.1097/jto.0b013e3181ec173d

Pan, P., Pan, Y., Xiao, Y., Han, B., Su, L., Su, M., et al. (2020). Prognostic Assessment of COVID-19 in the Intensive Care Unit by Machine Learning Methods: Model Development and Validation. *J. Med. Internet Res.* 22, e23128. doi:10.2196/23128

Persad, G., Peek, M. E., and Emanuel, E. J. (2020). Fairly Prioritizing Groups for Access to COVID-19 Vaccines. *JAMA* 324, 1601. doi:10.1001/jama.2020.18513

Snedecor, G. W. (1957). Statistical Methods. *Soil Sci.* 83, 163. doi:10.1097/00010694-195702000-00023

Sun, Q., Qiu, H., Huang, M., and Yang, Y. (2020). Lower Mortality of COVID-19 by Early Recognition and Intervention: Experience from Jiangsu Province. *Ann. Intensive Care*. 10, 33. doi:10.1186/s13613-020-00650-2

Tabata, S., Imai, K., Kawano, S., Ikeda, M., Kodama, T., Miyoshi, K., et al. (2020). Clinical Characteristics of COVID-19 in 104 People with SARS-CoV-2 Infection on the Diamond Princess Cruise Ship: a Retrospective Analysis. *Lancet Infect. Dis.* 20, 1043–1050. doi:10.1016/s1473-3099(20)30482-5

Vaid, A., Somani, S., Russak, A. J., De Freitas, J. K., Chaudhry, F. F., Paranjpe, I., et al. (2020). Machine Learning to Predict Mortality and Critical Events in a Cohort of Patients with COVID-19 in New York City: Model Development and Validation. *J. Med. Internet Res.* 22, e24018. doi:10.2196/24018

Wang, Y., Zhang, D., Du, G., Du, R., Zhao, J., Jin, Y., et al. (2020). Remdesivir in Adults with Severe COVID-19: a Randomised, Double-Blind, Placebo-Controlled, Multicentre Trial. *The Lancet*. 395, 1569–1578. doi:10.1016/s0140-6736(20)31022-9

WHO (2020). WHO Coronavirus Disease (COVID-19) Dashboard. Available at: https://covid19.who.int (Accessed December 2, 2020)

Wong, T-T. (2015). Performance Evaluation of Classification Algorithms by K-fold and Leave-One-Out Cross Validation. *Pattern Recognit*. 48, 2839–2846. doi:10.1016/j.patcog.2015.03.009

Yadaw, A. S., Li, Y. C., Bose, S., Iyengar, R., Bunyavanich, S., and Pandey, G. (2020). Clinical Features of COVID-19 Mortality: Development and Validation of a Clinical Prediction Model. *Lancet Digital Health*. 2, e516-e525. doi:10.1016/S2589-7500(20)30217-X

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 29 June 2021
doi: 10.3389/frai.2021.694875

Deep Learning–Based COVID-19 Pneumonia Classification Using Chest CT Images: Model Generalizability

Dan Nguyen[1,2], Fernando Kay[3], Jun Tan[2], Yulong Yan[2], Yee Seng Ng[3], Puneeth Iyengar[2], Ron Peshock[3] and Steve Jiang[1,2]**

[1]*Medical Artificial Intelligence and Automation (MAIA) Laboratory, University of Texas Southwestern Medical Center, Dallas, TX, United States,* [2]*Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX, United States,* [3]*Department of Radiology, University of Texas Southwestern Medical Center, Dallas, TX, United States*

Since the outbreak of the COVID-19 pandemic, worldwide research efforts have focused on using artificial intelligence (AI) technologies on various medical data of COVID-19–positive patients in order to identify or classify various aspects of the disease, with promising reported results. However, concerns have been raised over their generalizability, given the heterogeneous factors in training datasets. This study aims to examine the severity of this problem by evaluating deep learning (DL) classification models trained to identify COVID-19–positive patients on 3D computed tomography (CT) datasets from different countries. We collected one dataset at UT Southwestern (UTSW) and three external datasets from different countries: CC-CCII Dataset (China), COVID-CTset (Iran), and MosMedData (Russia). We divided the data into two classes: COVID-19–positive and COVID-19–negative patients. We trained nine identical DL-based classification models by using combinations of datasets with a 72% train, 8% validation, and 20% test data split. The models trained on a single dataset achieved accuracy/area under the receiver operating characteristic curve (AUC) values of 0.87/0.826 (UTSW), 0.97/0.988 (CC-CCCI), and 0.86/0.873 (COVID-CTset) when evaluated on their own dataset. The models trained on multiple datasets and evaluated on a test set from one of the datasets used for training performed better. However, the performance dropped close to an AUC of 0.5 (random guess) for all models when evaluated on a different dataset outside of its training datasets. Including MosMedData, which only contained positive labels, into the training datasets did not necessarily help the performance of other datasets. Multiple factors likely contributed to these results, such as patient demographics and differences in image acquisition or reconstruction, causing a data shift among different study cohorts.

Keywords: deep learning, generalizability, convolutional neural network, classification, computed tomography, COVID-19, SARS-CoV-2

OPEN ACCESS

Edited by:
Da Yan,
University of Alabama at Birmingham, United States

Reviewed by:
Zhao Wang,
Zhejiang University, China
Ke Li,
University of Wisconsin-Madison, United States

****Correspondence:***
Dan Nguyen
Dan.Nguyen@UTSouthwestern.edu
Steve Jiang
steve.jiang@utsouthwestern.edu

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *13 April 2021*
Accepted: *02 June 2021*
Published: *29 June 2021*

Citation:
Nguyen D, Kay F, Tan J, Yan Y, Ng YS, Iyengar P, Peshock R and Jiang S (2021) Deep Learning–Based COVID-19 Pneumonia Classification Using Chest CT Images: Model Generalizability. Front. Artif. Intell. 4:694875. doi: 10.3389/frai.2021.694875

INTRODUCTION

Since the outbreak of the 2019 coronavirus disease (COVID-19) in December 2019, the total worldwide death count due to COVID-19 has exceeded a million (Pérez-Peña, 2020). COVID-19 can affect multiple organ systems and cause fever, flu-like symptoms, cardiovascular damage, and pulmonary injury. The most common manifestation of COVID-19 upon initial presentation is pneumonia. While some patients are asymptomatic or have mild symptoms, a small percentage of

patients may develop severe acute respiratory distress syndrome (ARDS) that requires intubation in the intensive care unit and is associated with poor prognosis. The mortality rate is over 60% once they progress to the severe illness stage (Guan et al., 2020). Since chest CTs are performed for reasons other than pulmonary symptoms as well, an automated tool that can opportunistically screen chest CTs for the disease can potentially be used to identify patients with COVID-19. First, it has been suggested that patients with COVID-19 when identified in the early stage can be treated to prevent progression to the later stage of the disease (McCullough, et al., 2020a; McCullough, et al., 2020b; FLARE, 2020). Second, identification of asymptomatic patients in the early stage using CT (Ali and Ghonimy, 2020) provides a time window during which they can isolate themselves to prevent the spread to others.

Several efforts around the world have been focused on the identification or categorization of COVID-19–positive patients according to their various types of medical data. As part of the effort to understand and control this disease, large COVID-19 datasets of different formats have been curated and publicly released around the world. One group of studies focuses on using artificial intelligence (AI) technologies, in particular deep learning (DL)–based models, to detect COVID-19 through chest radiography and computed tomography (CT). These studies found high accuracy rates ranging from 82 to 98% (Wang L. et al., 2020; Sethy et al., 2020; Narin et al., 2021; Apostolopoulos and Mpesiana, 2020; Hemdan et al., 2020; Wang S. et al., 2020; Xu et al., 2020; Ozturk et al., 2020; Shibly et al., 2020; Oh et al., 2020; Jin et al., 2020). The high accuracy rates are promising and encourage the use of this technology in the clinical setting.

However, the generalizability of these models to other clinical settings around the world is not clear. The data usually found in clinical practice are often incomplete and noisy, and they may have high intra- and inter-study variability among different environments. This scenario often makes it difficult from a research perspective to develop algorithms and implement them in the clinic. Due to various restrictions on sharing patient data, many algorithms are developed with limited data that are specific to a clinic or a region. However, differences in several demographic factors, such as population distribution of race, ethnicity, and geography, can greatly impact the overall accuracy and performance of an algorithm in a different clinical setting (Topol, 2020). In addition, different methods of data collection by hospitals around the world may also impact an algorithm's performance. Because the boom of DL technologies has happened only within the last several years, the number of studies testing the robustness and performance of AI algorithms across various clinical settings is extremely limited (Topol, 2020). Therefore, there is very little knowledge about how well a model will perform in a realistic clinical environment over time.

For example, Barish et al. (2021) demonstrated a particular public model developed by Yan (2020) that predicted mortality from COVID-19–positive patients—which performed well on an internal dataset with an accuracy of 0.878—failed to accurately predict the mortality on an external dataset, with an accuracy of only around 0.5. Another similar negative study applied Yan et al.'s model on an external dataset and drew similar conclusions about the accuracy of its mortality prediction (Quanjel et al., 2021). A systematic review of 107 studies with 145 prediction models was conducted, and the studies reported that all models had a high bias, due to nonrepresentative control datasets and overly optimistic reported performance (Wynants et al., 2020), which can additionally lead to unrealistic expectations among clinicians, policy makers, and patients (Laghi, 2020). Bachtiger et al. concluded that this boom of DL models for COVID-19 focused far too much on developing novel prediction models without a comprehensive understanding of its practical application and biases from the dataset (Bachtiger et al., 2020). Others have similarly concluded that AI has yet to have any impact on the prevailing pandemic and that extensive and comprehensive gathering of diagnostic COVID-19–related data will be essential to develop useful AI models (Naudé, 2020).

As part of the efforts to collect data, large datasets of 3D computed tomography (CT) scans with COVID-19–related labels have been publicly released. This provides an opportunity to study the generalizability of DL algorithms developed using these volumetric datasets. In this study, we collected and studied one internal dataset collected at UT Southwestern (UTSW) and three large external datasets from around the world: 1) China Consortium of Chest CT Image Investigation (CC-CCII) Dataset (China) (Zhang et al., 2020), 2) COVID-CTset (Iran) (Rahimzadeh et al., 2021), and 3) MosMedData (Russia) (Morozov et al., 2020). We trained DL-based classification models on various combinations of datasets and evaluated the model performance on the held-out test data from each of the datasets.

METHODS

Data

We collected one internal dataset at UTSW and three large datasets from around the world that are publicly available—1) China Consortium of Chest CT Image Investigation (CC-CCII) Dataset (China), 2) COVID-CTset (Iran), and 3) MosMedDat (Russia)—which is summarized in **Table 1**. The UTSW dataset is composed of three subsets of anonymized imaging data obtained retrospectively. The study protocol was approved by the institutional review board and the requirement for informed consent was waived. The first subset includes patients who tested positive for severe acute respiratory syndrome coronavirus 2 on real-time polymerase chain reaction between March and November 2020 and who had a chest CT scan performed within the first 7 days of diagnosis. All chest CT scans were obtained according to the standard clinical care—common clinical indications were to assess the worsening respiratory status and to rule out pulmonary thromboembolism. Chest CT is not obtained as a first-line modality to diagnose or screen for COVID-19 at UTSW. As such, the collected dataset had a mixture of contrast-enhanced CTs and non-contrast CTs. The second and third subsets include patients who had a chest CT scan obtained as part of the standard clinical care between March and May 2019, that is, the pre–COVID-19 pandemic phase. The radiologic reports of

TABLE 1 | Summary of data used in the study. These datasets include full volumetric CT scans of the patients.

Dataset	Origin	Description				Available at:
		Details	# Patients	# 3D scans	Label	
UTSW	UT Southwestern Medical Center	CT vendors: Phillips, Toshiba, GE Medical Systems Image resolution: 512 × 512 Pixel size range: 0.45 mm to 0.83 mm Slice thickness range: 0.9–3 mm Format: DICOM	101 118 118	101 118 118	COVID-19 positive Infection (negative) Findings Unrelated to Infection (negative)	*See footnote[1]
China Consortium of Chest CT Image Investigation (CC-CCII) Dataset	Sun Yat-sen Memorial Hospital and Third Affiliated Hospital of Sun Yat-sen University, Guangzhou, China The First Affiliated Hospital of Anhui Medical University, Anhui, China West China Hospital, Sichuan, China Nanjing Renmin Hospital, Nanjing, China Yichang Central People's Hospital, Hubei, China Renmin Hospital of Wuhan University, Wuhan, China	CT vendor: unreported Image resolution: mostly 512 × 512 (a few were 128 × 128) Pixel size range: unreported Slice thickness range: 1–5 mm	929 964 849	1544 1556 1078	COVID-19 positive Common Pneumonia (negative) Normal Lung (negative)	http://ncov-ai.big.ac.cn/download
COVID-CTset	Negin Medical Center, Sari, Iran	CT vendor: Siemens Image resolution: 512 × 512 Pixel size range: unreported Slice thickness range: unreported	95 282	281 1068	COVID-19 positive Normal lung (negative)	https://github.com/mr7495/COVID-CTset
MosMedData	Municipal hospitals in Moscow, Russia	CT vendor: Toshiba Image resolution: 512 × 512 Pixel size range: unreported Slice thickness: 1 mm	254 684 125 45 2	254 684 125 45 2	CT-0—not consistent with pneumonia (can include both COVID-19 positive and negative) CT-1—Mild (COVID-19 positive) CT-2—Moderate (COVID-19 positive) CT-3—Severe (COVID-19 positive) CT-4—Critical (COVID-19 positive)	https://mosmed.ai/

these studies were screened by a cardiothoracic radiologist with 12 years of clinical experience. The reports were labeled as having radiologic findings suggestive of infection or not. The adjudication was based on the presence of radiologic patterns usually associated with infection, including ground-glass opacities, consolidation, and nodular pattern, if such findings were described as fitting a differential diagnosis of infectious process based on the impression by the primary interpreting radiologists. These studies were consecutively selected to match the sex and age distribution of the COVID-19–positive subset and to represent two control groups with a balanced representation of chest CT showing findings suggestive of the infection (118) and findings not related to infection (118). The CC-CCII dataset was obtained from six different hospitals: 1) Sun Yat-sen Memorial Hospital and Third Affiliated Hospital of Sun Yat-sen University, 2) The first Affiliated Hospital of Anhui Medical University, 3) West China Hospital, 4) Nanjing Renmin Hospital, 5) Yichang Central People's Hospital, and 6) Renmin Hospital of Wuhan University. The COVID-CTset dataset was from the Negin Medical Center, and the MosMedData dataset was from municipal hospitals in Moscow, Russia.

For consistency in training and testing the models in our study, we divided all the data into two classes: 1) COVID-19–positive and 2) COVID-19–negative patients. Note that MosMedData does not

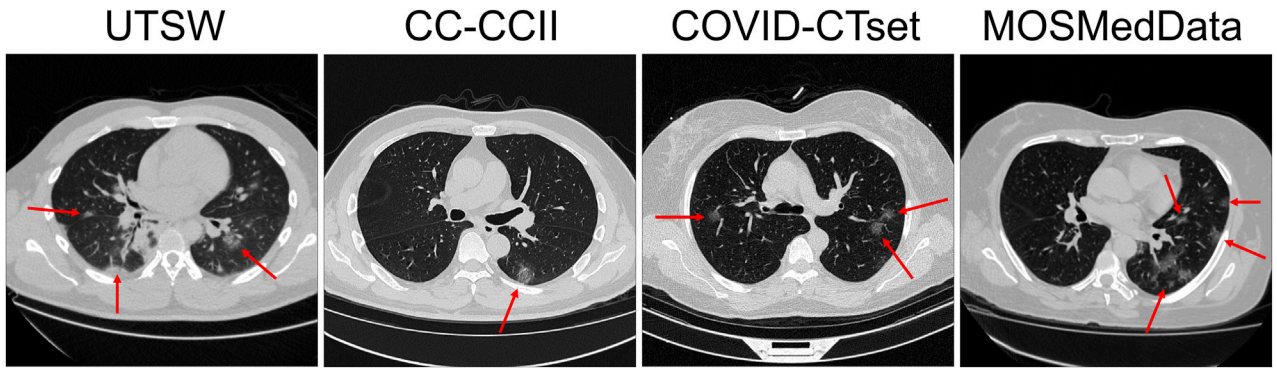


FIGURE 1 | Slice view of example CTs from each dataset. Red arrows show patchy ground-glass opacities with round morphology, which are typical findings in COVID-19 pneumonia.

have conclusive negative–label patients, as CT-0 might contain both positive and negative patients. Accordingly, we omitted the CT-0 category from this study. Most scans in this study had a matrix size of 512 × 512 × n, where n is the variable number of slices. For the small number of scans that had a reduced matrix size, the images were linearly interpolated to match the 512 × 512 × n resolution.

Most of the data were available in Hounsfield units (HU) or CT number (e.g. 0–4095). Some of the data in the CC-CCII dataset were provided in relative intensity values (e.g., 0–255). Because the data formatting varied across datasets, we performed clipping and normalization operations. First, if the data were displayed in HU, we clipped the minimum number to be −1,000 HU. For evaluation, the data were normalized from 0 to 1 prior to evaluation by the DL model. For training, multiple normalization methods were used as part of a data augmentation technique. The complete data augmentation is further described in the section *Training and Data Augmentation*. **Figure 1** shows example CTs of COVID-19–positive patients from each dataset.

For training, validating, and testing the model, the positive labels of the UTSW dataset were randomly split into 73 train, 8 validation, and 20 test patients and scans (one 3D scan per patient). The positive labels of the CC-CCII dataset were randomly split into 669 train, 74 validation, and 186 test patients, or 1,110 train, 122 validation, 312 test scans. The positive labels of the COVID-CTset were randomly split into 68 train, 8 validation, and 19 test patients, or 201 train, 23 validation, and 57 test scans. The positive labels of MosMedData were randomly split into 616 train, 69 validation, and 171 test patients and scans (one 3D scan per patient; CT-0 category was omitted).

For the negative labels, the UTSW dataset was randomly split into 170 training, 18 validation, and 48 testing patients and scans (one 3D scan per patient). The CC-CCII dataset was randomly split into 1,305 train, 145 validation, and 363 test patients, or 1,891 train, 203 validation, and 540 test scans. The COVID-CTset was randomly split into 259 train, 29 validation, and 72 test patients, or 770 train, 84 validation, and 214 test scans.

Model Architecture

The model used in this study was a classification style convolutional neural network (CNN) model (LeCun et al., 1989; LeCun and Bengio, 1995; LeCun et al., 1998; LeCun et al., 1999), with specifics shown in **Figure 2**. The input shape was set to 512 × 512 × 128. There are five resolution levels of convolutions and four downsampling operations prior to the flattening operation. The downsampling size also varied each time and was set as (4,4,4), (4,4,2), (4,4,2), and (2,2,2), respectively. This converts the data shape from 512 × 512 × 128 to 4 × 4 × 4. At each resolution level, a series of operations consisting of convolution, Rectified Linear Unit activation (ReLU), Group Normalization (Wu and He, 2018), and DropBlock (Ghiasi et al., 2018) is applied twice, consecutively. The convolution kernel size varied at each resolution level: (3,3,3), (5,5,5), (5,5,3), (5,5,3), and (3,3,3). The number of filters, indicated by red numbers in **Figure 2**, at each convolution started at eight and doubled after each downsampling operation. After these operations, the feature data are flattened into a single vector of length 8,192. Then, a series of operations consisting of fully connected calculations, ReLU, Group Normalization, and Dropout (Srivastava et al., 2014) follows. This is performed a total of four times, calculating 1,024 features each time. Then, one more full connection is applied to reduce the data into two outputs, and a softmax operation is applied.

Training and Data Augmentation

In total, nine models were trained in this study using the training and validation data outlined in *Data* and were split into two categories: 1) single dataset training and 2) multiple dataset training. We trained three models on a single dataset, one each on the UTSW, CC-CCII, and the COVID-CTset datasets. No model was trained on MosMedData by itself, since this dataset does not have any negative labels. For multiple dataset training, we trained six models with different combinations of datasets: 1) UTSW + CC-CCII, 2) UTSW + COVID-CTset, 3) CC-CCII + COVID-CTset, 4) UTSW + CC-CCII + COVID-CTset, 5) CC-CCII + COVID-CTset + MosMedData, and 6) UTSW + CC-CCII + COVID-CTset + MosMedData.

Some additional operations were applied to format and augment the CT data for model training. For CT data with less than 128 slices, slices of zeros were padded onto the CT slices until the total data volume had 128 slices. The number of slices superior and inferior to the CT data was uniformly and

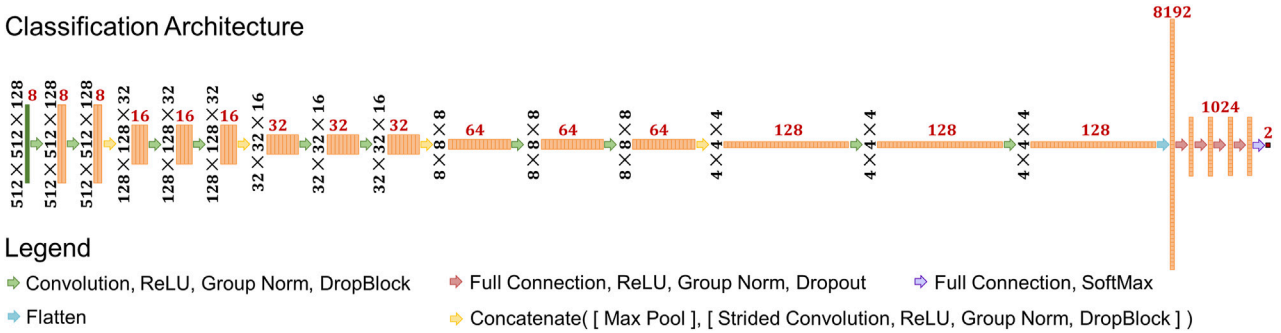


FIGURE 2 | Schematic of deep learning architecture used in the study. Black numbers represent the feature shape of each layer prior to the flattening operation. Red numbers represent the number of features at each layer.

randomly decided at each iteration. For data with more than 128 slices, a random continuous volume of 128 slices was selected. The data were then normalized in one of two ways: 1) from 0 to 1, or 2) from 0 to $\frac{max\ (data)}{2^n}$, where n is the smallest integer possible while keeping 2^n larger than the maximum value in the CT volume. The normalization method was randomly chosen with a 50% chance during each training iteration. An additional step was applied to decide, at a 50% chance, whether this data would be fed into the model for training or if additional data augmentation would be applied. If yes to additional data augmentation, then the function randomly flipped, transposed, rotated, or scaled the data. For the flip augmentation, there was a 50% chance that it would individually apply a flip to each axis (row, column, and slice). For the transpose augmentation, there was a 50% chance that it would transpose the row and column of the data (no transpose operation was ever applied using the slice dimension). For the rotate augmentation, a random integer, {0,1,2,3}, was generated and multiplied against 90° to determine the rotation angle, then applied only on the row and column dimensions. For the scale augmentation, there was a 50% chance that a scaling factor was applied, and the scale was a uniform random number from 0 to 1.

Each model was trained for a total of 2,50,000 iterations—which is about 1,029, 83, 544, and 406 epochs for the UTSW, CC-CCII, COVID-CTSet, and MosMedData, respectively—using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 1×10^{-5}. To prevent overfitting on the training data, the accuracy was evaluated on the validation data for every 500 iterations, and the instance of the model with the highest validation accuracy was saved as the final model for evaluation. The models were trained using NVIDIA V100 GPUs with 24 GB of memory.

Evaluation

All nine of the trained models were evaluated on the test data of each dataset. For volumes with less than 128 slices, zero padding on the slices was evenly applied in the superior and inferior directions, to keep the data centered. For volumes greater than 128 slices, a sliding window technique was applied across the volume, and the model made multiple predictions. The number of slices in a patch was 128, and the stride size was 32 slices. The prediction with the highest COVID-19 probability was taken as the model's final prediction.

A threshold was selected based on maximizing the prediction accuracy on the validation data and applied to the testing set. In the cases where the "optimal" threshold was a trivial value (e.g., threshold = 0 for MosMedData, which only has positive labels), we took the argmax of the output as the prediction instead. The true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) were counted, and a normalized confusion matrix was generated for each dataset. Averaged confusion matrices were calculated with and without MosMedData. An evenly weighted average was chosen.

Receiver operating characteristic (ROC) curves were calculated on the test data by varying the positive predictive threshold from 0 to 1, at 0.01 intervals. The area under the curve (AUC) was calculated to determine the overall performance of each model on each dataset. We additionally used the Bayesian approximate technique called Monte Carlo dropout (Gal and Ghahramani, 2016) to additionally estimate the uncertainty on the AUC. MosMedData was excluded from the ROC and AUC analyses, since it was missing negative labels.

RESULTS

Each model took about 5 days on average to train on a GPU. For nine models, this is equivalent to 45 GPU-days of training. Each model prediction takes an average of 0.53 s, which makes it very useful for near real-time applications.

The single dataset models' predictive accuracy $\left(\frac{TP+TN}{TP+TN+FP+FN}\right)$ on the test dataset is displayed in **Figure 3**. Overall, each model performed best on the dataset that it was trained on, with an accuracy as high as 0.97 for the CC-CCII model evaluated on the CC-CCII data. The model that performed the worst on its own dataset was COVID-CTset, with an accuracy of 0.86. The UTSW model had an accuracy of 0.87 on its own dataset. Since the test data were held out of the training and validation phase, it is a strong indicator that the model did not overfit to its specific

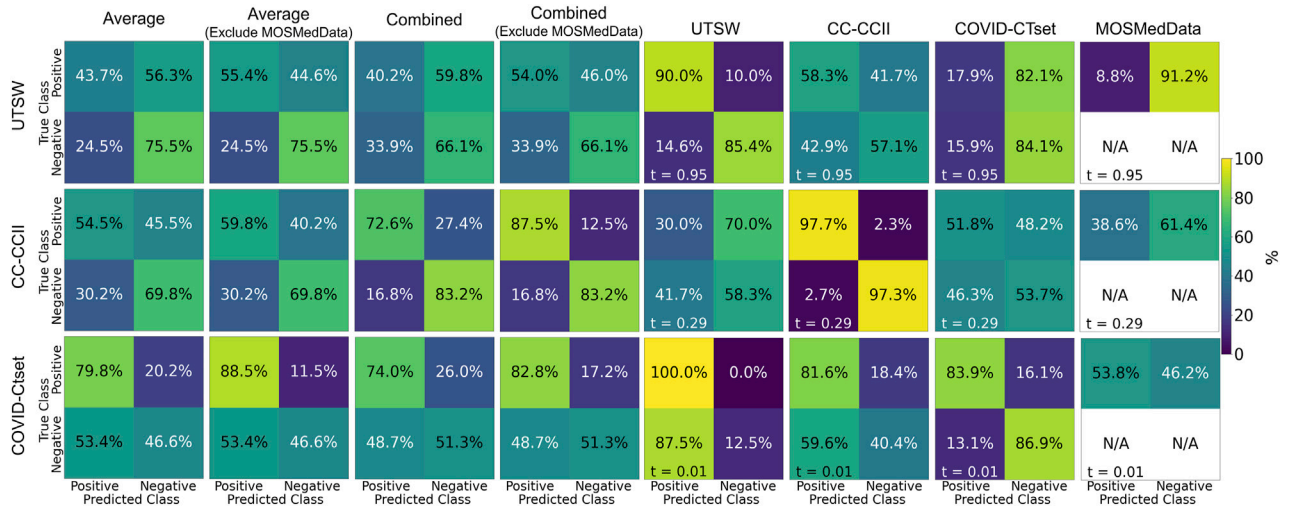


FIGURE 3 | Confusion matrices on the test data for each of the models trained on a single dataset. Each row represents the datasets that the model was trained on, and each column represents the datasets that the model was evaluated on. Note that MosMedData does not have any negative label data. The labeling threshold used for each model is indicated on the lower left of each confusion matrix (t = #). The "Average" confusion matrix is an equally weighted average among each dataset, while the "Combined" confusion matrix is calculated from all samples from the datasets.

training data. However, the models performed much more poorly when evaluated on a dataset they had not seen before, which signifies that the model did not generalize well to the new dataset type. The worst performance was the COVID-CTset model evaluated on the UTSW dataset, which had an accuracy of 0.38. All three models had poor performance on the MosMedData dataset.

Figure 4 shows the confusion matrices of the performance of models trained on multiple datasets against the test data. The multiple dataset model that had the best accuracy when evaluated on the UTSW test set was the UTSW + CC-CCII model, with 0.93 accuracy. When evaluating the CC-CCII test set, the model with the best accuracy of 0.96 was the UTSW + CC-CCII model. When evaluating the COVID-CTset, the UTSW + COVID-CTset performed best, with an accuracy of 0.94. The best multiple dataset models outperformed their single dataset counterparts with regards to accuracy. However, these models still had poor accuracy when evaluated on a test dataset they have not seen before. For example, the model trained with the UTSW and COVID-CTset together had improved accuracies to 0.90 and 0.94 when evaluated on the test sets of the UTSW and COVID-CTset datasets, respectively. However, when evaluated on the CC-CCII dataset, the accuracy was 0.53. Including MosMedData in the model training improved the total average performance but did not improve the performance when evaluating models on the individual UTSW, CC-CCII, and COVID-CTset datasets.

Figure 5 shows the ROC curves of the single dataset models. The models, when evaluated on the same dataset that they were trained on, showed good AUCs (mean ± standard deviation) of 0.826 ± 0.024 (UTSW), 0.988 ± 0.002 (CC-CCII), and 0.873 ± 0.012 (COVID-CTset). The models performed considerably worse when evaluated on different datasets, with AUCs ranging from 0.405 to 0.570, which is close to just random guessing (i.e., AUC = 0.5). The ROC curves of the multiple dataset models are shown in **Figure 6**. For each dataset—UTSW, CC-CCII, and COVID-CTset—the best performing models were the UTSW + COVID-CTset (AUC = 0.937 ± 0.018), the UTSW + CC-CCII + COVID-CTset (AUC = 0.989 ± 0.002), and the UTSW + COVID-CTset (AUC = 0.926 ± 0.010) models, respectively. Since the test data were held entirely separate from the model development process, and used only for evaluation, this shows once again that the models did not overfit their own training data. Similar to the single dataset models, the multiple dataset models also performed poorly when predicting on datasets they had never seen before, with AUCs ranging from 0.380 to 0.540.

DISCUSSION

In this study, we demonstrate that our DL models can correctly identify patients that are COVID-19–positive with high accuracy, but only when the model was trained on the same datasets that it was tested on. Otherwise, the performance is poor—close to random guessing—which indicates that the model cannot easily generalize to an entirely new dataset distribution that it has never seen before for COVID-19 classification. Several data augmentation techniques were applied during training to prevent overfitting on the test set. In addition, the weights of the model that performed the best on the validation data with regards to accuracy were used as the final model. Dropout and DropBlock regularization were added to further prevent the model from overfitting.

We additionally observed that certain combined dataset models performed best for particular datasets in detecting patients who are positive for COVID-19. For example, we

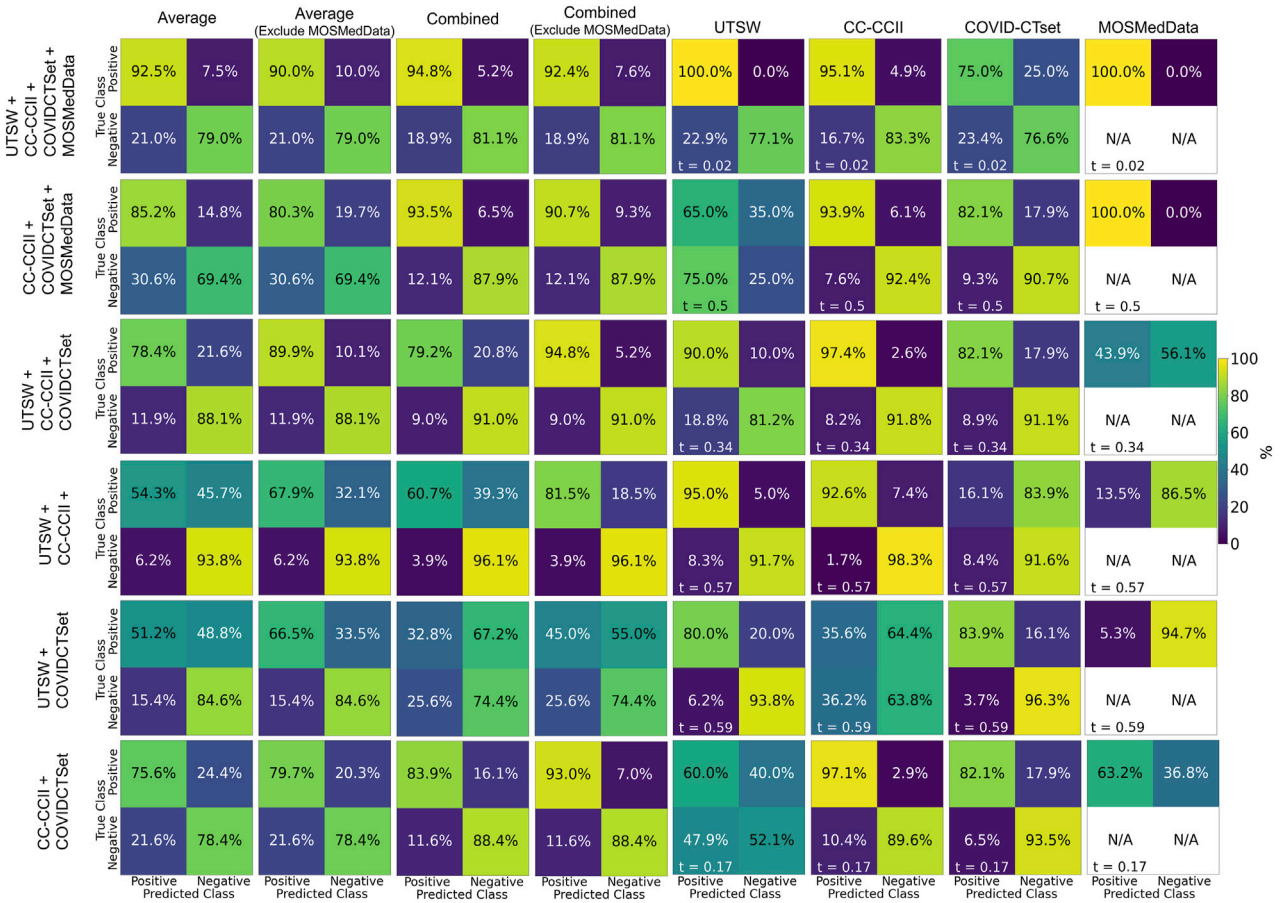


FIGURE 4 | Confusion matrices on the test data for each of the models trained on multiple datasets. Each row represents the datasets that the model was trained on, and each column represents the datasets that the model was evaluated on. The labeling threshold used for each model is indicated on the lower left of each confusion matrix (t = #). The "Average" confusion matrix is an equally weighted average among each dataset, while the "Combined" confusion matrix is calculated from all samples from the datasets.

found that the highest performing model in the dataset from the UTSW dataset was obtained when the training step combined UTSW and CC-CCII datasets. This may have occurred due to the relatively low sample count in the UTSW dataset (73 positive, 170 negative patients for training); therefore, adding data samples from COVID-CTset improved with DL-model's AUC from 0.826 to 0.937 on the UTSW dataset. Overall, the best-performing model for a particular dataset tended to be a multiple dataset model that included that same dataset in the training. When used properly, training on multiple datasets allows for having more training examples for the model to improve its overall feature extraction capabilities. There are many similarities between images, such as the texture and edges, which the model can learn from all the images. For example, it has been shown that models that pretrain on ImageNet (millions of images) can perform better on other classification tasks (Xie and Richmond, 2018). However, adding more data from different distributions into the training did not always monotonically improve the model's performance. For example, adding the CC-CCII data for training did not improve the model performance, with the AUC of 0.920 for the UTSW dataset. Adding MosMedData into the training lowered the performance of the model on the other three datasets. This is likely because the original intent of MosMedData was to train a model to categorize the severity of COVID-19 into five classes and, therefore, lacked negative labels. Without definitive negative labels, our models likely learned simply to identify the data source as MosMedData and compromised some of their learning capacity and performance to use the relevant imaging features for the predictions. This does serve as an important lesson in data collection: datasets from a particular healthcare center or region should be fully representative of the task at hand to be used in training. Simply collecting COVID-19–positive patients from one source and negative patients from a different source is likely to introduce an uncorrectable bias during training that led to a poor model performance.

We did include some state-of-the-art modules in our model, such as Group Normalization (Wu and He, 2018) and DropBlock (Ghiasi et al., 2018), that allowed for a high performance similar to other COVID-19 classification studies (Wang Z. et al., 2020;

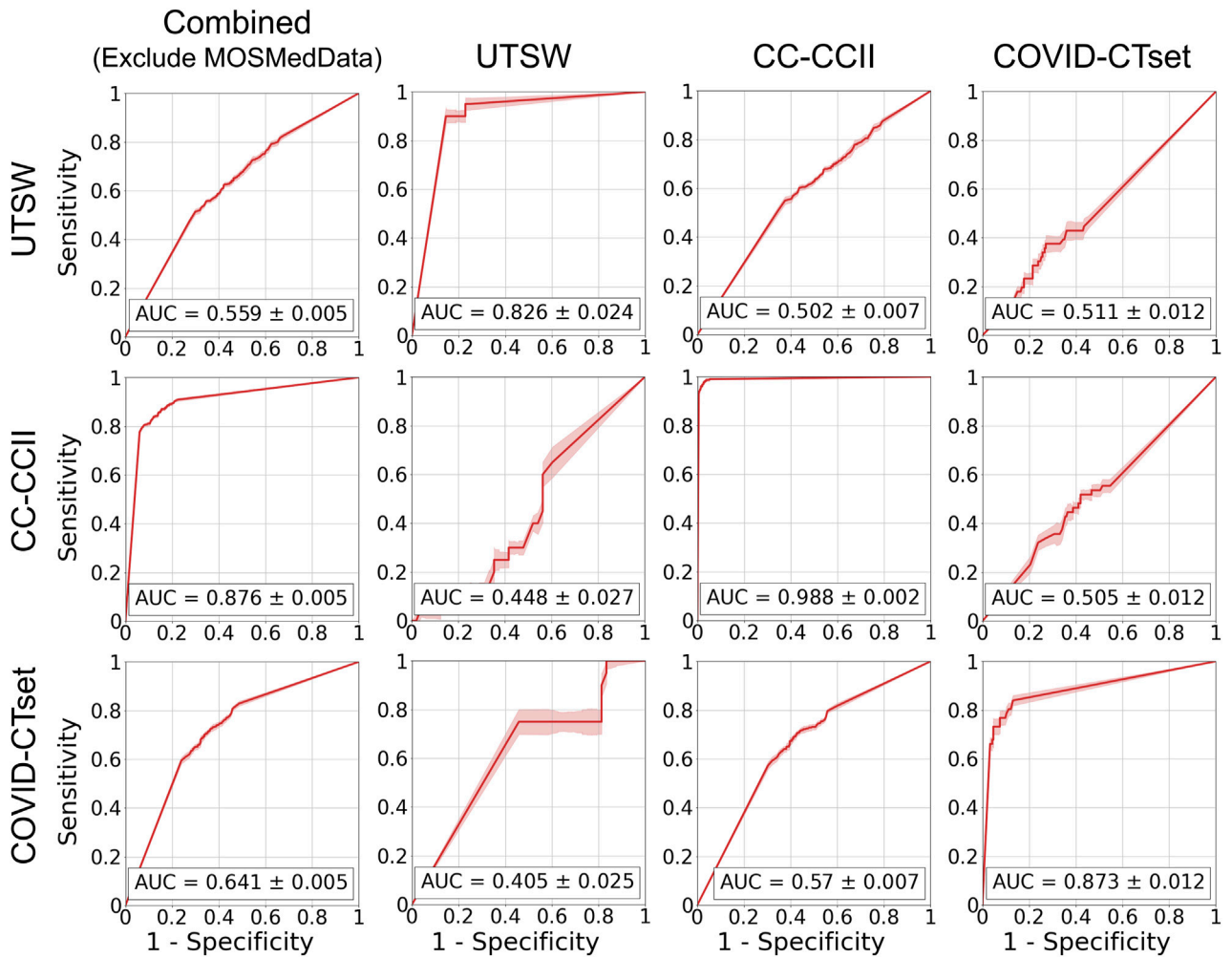


FIGURE 5 | ROC curves on the test data for the models trained on single datasets. Each row represents the datasets that the model was trained on, and each column represents the datasets that the model was evaluated on. The error band and the error value in the reported AUC represent 1 standard deviation.

Ali et al., 2021; Song, 2021). Zech et al. investigated model generalizability in CT scans and found a similar conclusion, but a better one than a random guess on the unseen dataset (Zech et al., 2018). The major difference between this study and our study, where we only found a performance of around a random guess on an unseen test dataset, is that we investigated the generalizability of datasets across different countries around the world. The other study by Zech investigated datasets only from the United States, so it is likely that the differences in protocol, standards, and demographics between the datasets are much smaller than the dataset that we used. We intend to further investigate these differences and their impact across both intranation and internation datasets in a future study.

A potential source of bias may come from the discretization of data. While CT is typically stored in a 12-bit format, having 4,096 levels of discretization, some of the data in the CC-CCII dataset were stored in relative intensity values from 0 to 255. While we were careful with our normalization and data augmentation techniques, the more inherent coarseness in some of the data may have affected the model's generalizability between datasets. When sharing or collecting datasets, it is of utmost importance to disclose the data's exact format, as these can add more variability outside of the scanning protocol, quality, and demographics of a particular institution or region.

Between the UTSW dataset, CC-CCII dataset, and the COVID-CTset dataset, the CC-CCII dataset consistently yielded models that had the highest accuracy and AUC when evaluated on its own dataset. The exact reason for this is unknown, but it may be possible that there was an implicit bias within the dataset. For example, if one of the participating hospitals had a very different distribution of image quality, but also were a large provider of the data, then the model may have learned to simply distinguish that hospital specifically instead of the disease. However, the exact breakdown of where each individual scan originated from is not available. We will continue to investigate such cases and determine whether there was some sort of bias that allowed the CC-CCII dataset to yield models that gave high-accuracy values.

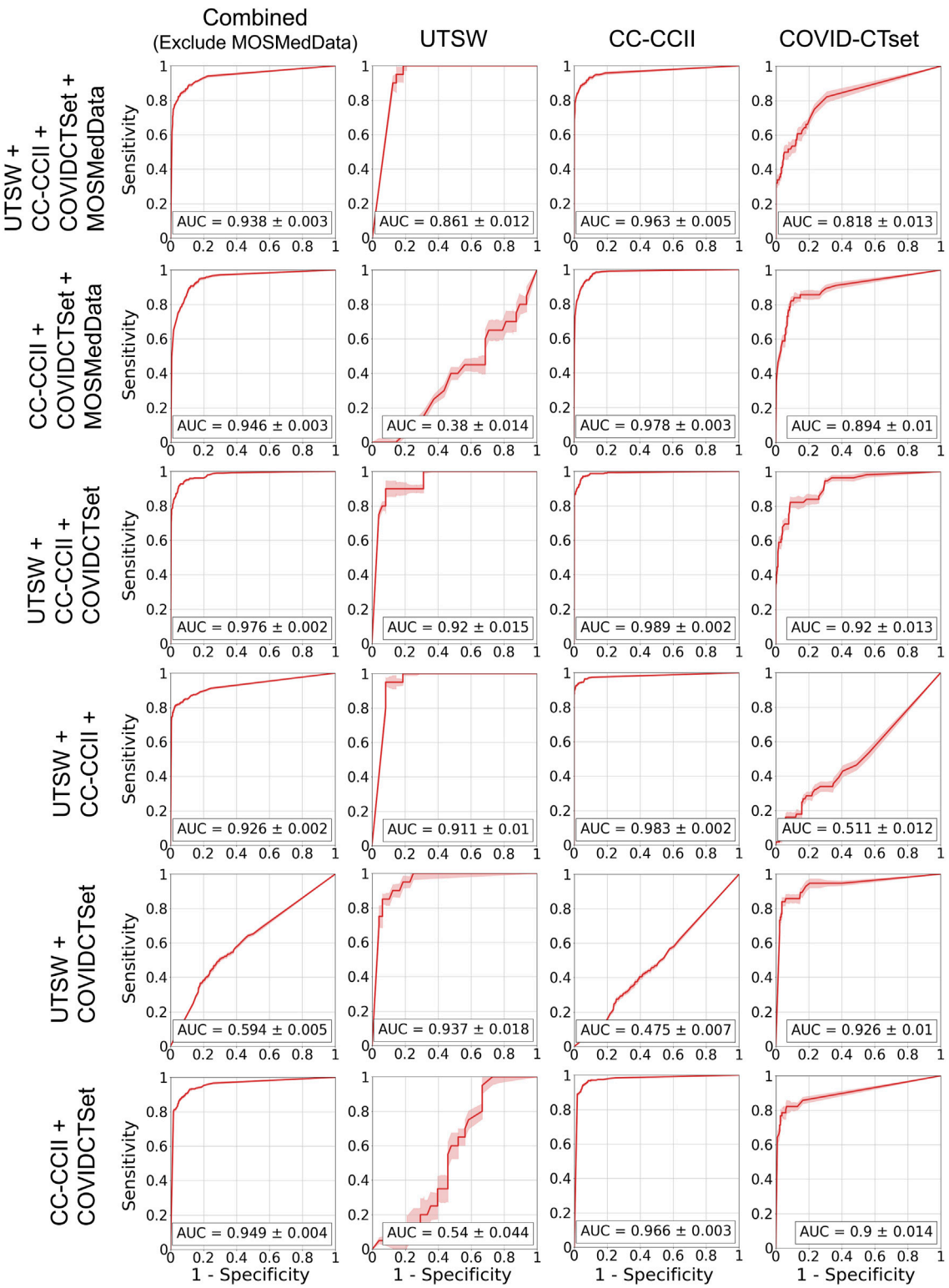


FIGURE 6 | ROC curves on the test data for the models that trained on multiple datasets. Each row represents the datasets that the model was trained on, and each column represents the datasets that the model was evaluated on. The error band and the error value in the reported AUC represent 1 standard deviation.

In contrast, the COVID-CTset dataset consistently yielded models that had the poorest performance. One potential reason is possibly its lack of variability of data to train on. For example, the UTSW dataset had COVID-19–negative scans that also included infected patients and the CC-CCII dataset had COVID-19–negative scans with common pneumonia. This may have helped the model further distinguish the nuances between COVID-19–positive and COVID-19–negative patients but with other presenting diseases. We plan to further identify and investigate these sources of biases in detail as part of a future study.

Although this study did not fully explore the possible techniques to improve robustness and prevent overfitting, it may serve as a baseline for future model generalization studies that use medical data for the clinical implementation of COVID-19–related classification models. We will continue to explore the limits of model generalization with respect to improving the algorithm and to the intra- and inter-source data variability, regarding the identification of COVID-19–positive patients by their medical data. As a whole, the deep learning models achieved a high performance on the unseen test set from the same distribution that they were trained on, which indicates that we did not have a typical overfitting problem with the training data. The low performance on datasets that the models had never seen before may actually be an indicator that the problem is not in the approach to the initial algorithm development—the problem may be the transfer and deployment of the algorithm to a new clinical setting. Creating a globally generalizable algorithm is a tall order, when people around the world have vastly different demographics and data collection protocols. With limited data and learning time, these AI algorithms are bound to fail when they encounter a unique data distribution they have never seen before. These results underscore the limited versatility of AI algorithms which may hamper the widespread adoption of AI algorithms for automated diagnosis of radiology images. This is in contrast to radiologists who in general can easily adapt to new clinical practices quickly. Perhaps we need to recalibrate our mindset with regard to the expectation for these AI algorithms—we should expect that these AI algorithms will always need to be fine-tuned to the local distribution when implemented and deployed in a specific clinical setting, then need to be retuned over time as distributions inevitably shift, either through demographic shifts or through the advancement of new treatment technologies. Transfer learning and continuous learning techniques (Torrey and Shavlik, 2010) are active fields of research and may become critical components to rapidly transfer, deploy, and maintain an AI model into the clinic.

AI tools designed for automatic identification of diseases on CT datasets, such as COVID-19, will only succeed if they can prove their robustness against a wide array of patient populations, scan protocols, and image quality. Notwithstanding, they hold the promise of becoming a powerful resource for identifying diseases, where time to detection is a critical variable. In the case of COVID-19, it is well known that many cases are asymptomatic, of which up to 54% will present abnormalities on chest CT (Inui et al., 2020). Thus, COVID-19 can be incidentally found on routine imaging. Timely identification of incidental cases of COVID-19 on chest CT by AI tools could lead to adequate prioritization of scans for reporting, resulting in prompt initiation of disease tracking and control measures. Moreover, the model architecture developed in this work can also serve as a template for similar tools tailored for detecting other clinical conditions.

The deep learning models were capable of identifying COVID-19–positive patients when the testing data was in the same dataset as the training data, whether the model was trained on a single dataset or on multiple datasets. However, we found a poor performance, close to random guessing, when models were evaluated on datasets that they had never seen. This is likely due to different factors, such as patient demographics, image acquisition methods/protocols, or diagnostic methods, causing a data shift between different countries' data. This lack of generalization for the identification of COVID-19–positive patients may not necessarily mean that the models were trained poorly, but rather the distribution of the training data may be too different from the evaluation data. Transfer learning and continuous learning may become imperative tools for tuning and deploying a model in a new clinical setting.

DATA AVAILABILITY STATEMENT

UTSW dataset is non-public. In accordance with HIPAA policy, access to the dataset will be granted on a case-by-case basis upon submission of a request to the corresponding authors and the institution. The DL models and related code developed in this study are available upon request for non-commercial research purposes.

AUTHOR CONTRIBUTIONS

SJ and DN conceived initial conceptual ideas. All authors provided essential feedback in shaping the research direction. DN curated and characterized the public dataset, designed the neural network model, developed the training methodology, trained the various models, evaluated the models' performances, and took lead in the manuscript writing. FK worked closely with DN for the internal dataset curation and characterization. JT and YY provided additional computational support and infrastructure for model performance characterization. FK, YN, PI, RP, and SJ provided additional feedback and analysis of the results and their implications on clinical practice. SJ worked closely with everyone and supervised the overall project direction. All authors provided substantial feedback on the manuscript and gave final approval for the publication.

ACKNOWLEDGMENTS

We would like to thank Jonathan Feinberg for editing the manuscript.

REFERENCES

Ali, A., Shaharabany, T., and Wolf, L. (2021). Explainability Guided Multi-Site COVID-19 CT Classification. *arXiv preprint arXiv*. 2103, 13677. doi:10.1186/s43055-020-00266-3

Ali, R. M. M., and Ghonimy, M. B. I. (2020). Radiological Findings Spectrum of Asymptomatic Coronavirus (COVID-19) Patients. *Egypt. J. Radiol. Nucl. Med.* 51, 1–6. doi:10.1186/s43055-020-00266-3

Apostolopoulos, I. D., and Mpesiana, T. A. (2020). Covid-19: Automatic Detection from X-ray Images Utilizing Transfer Learning with Convolutional Neural Networks. *Phys. Eng. Sci. Med.* 43, 635–640. doi:10.1007/s13246-020-00865-4

Bachtiger, P., Peters, N. S., and Walsh, S. L. (2020). Machine Learning for COVID-19-Asking the Right Questions. *The Lancet Digital Health*. 2, e391–e392. doi:10.1016/s2589-7500(20)30162-x

Barish, M., Bolourani, S., Lau, L. F., Shah, S., and Zanos, T. P. (2021). External Validation Demonstrates Limited Clinical Utility of the Interpretable Mortality Prediction Model for Patients with COVID-19. *Nat. Machine Intelligence* 3 (1), 25–27. doi:10.1038/s42256-020-00254-2

FLARE (2020). *Favipiravir +/- Lopinavir: A RCT of Early Antivirals (FLARE)*. Bethesda, MD: NIH U.S. National Libratiy of Medicn. Available at: <https://clinicaltrials.gov/ct2/show/NCT04499677>

Gal, Y., and Ghahramani, Z. (2016). "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in *International Conference on Machine Learning*, 1050–1059. New York City, NY: Proceedings of Machine Learning Research.

Ghiasi, G., Lin, T.-Y., and Le, Q. V. (2018). "Dropblock: A Regularization Method for Convolutional Networks," in *Advances in Neural Information Processing Systems*. 10727–10737. Cambridge, MA: MIT Press.

Guan, W.-j., Ni, Z.-Y., Hu, Y., Liang, W.-H., Ou, C.-Q., He, J. -X., et al. (2020). Clinical Characteristics of Coronavirus Disease 2019 in China. *New Engl. J. Med.* 382, 1708–1720. doi:10.1056/NEJMoa2002032

Hemdan, E. E.-D., Shouman, M. A., and Karar, M. E. (2020). Covidx-net: A Framework of Deep Learning Classifiers to Diagnose Covid-19 in X-ray Images. *arXiv preprint arXiv* 2003, 11055.

Inui, S., Fujikawa, A., Jitsu, M., Kunishima, N., Watanabe, S., Suzuki, Y., et al. (2020). Chest CT Findings in Cases from the Cruise Ship diamond Princess with Coronavirus Disease (COVID-19). *Radiol. Cardiothorac. Imaging*. 2, e200110. doi:10.1148/ryct.2020200110

Jin, C., Chen, W., Cao, Y., Xu, Z., Tan, Z., Zhang, X., Deng, L., et al. (2020). Development and Evaluation of an Artificial Intelligence System for COVID-19 Diagnosis. *Nat. Commun.* 11, 1–14. doi:10.1038/s41467-020-18685-1

Kingma, D., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv*. 1412, 6980

Laghi, A. (2020). Cautions about Radiologic Diagnosis of COVID-19 Infection Driven by Artificial Intelligence. *The Lancet Digital Health*. 2, e225. doi:10.1016/s2589-7500(20)30079-0

LeCun, Y., and Bengio, Y. (1995). Convolutional Networks for Images, Speech, and Time Series. *The handbook Brain Theor. Neural networks*. 3361, 1995.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., et al. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* 1, 541–551. doi:10.1162/neco.1989.1.4.541

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based Learning Applied to Document Recognition. *Proc. IEEE.* 86, 2278–2324. doi:10.1109/5.726791

LeCun, Y., Haffner, P., Bottou, L., and Bengio, Y. (1999). *Shape, Contour and Grouping in Computer Vision 319-345*. Berlin, Heidelberg: Springer.

McCullough, P. A., Alexander, P. E., Armstrong, R., Arvinte, C., Bain, A. F., Bartlett, R. P., et al. (2020a). Multifaceted Highly Targeted Sequential Multidrug Treatment of Early Ambulatory High-Risk SARS-CoV-2 Infection (COVID-19). *Rev. Cardiovasc. Med.* 21, 517. doi:10.31083/j.rcm.2020.04.264

McCullough, P. A., Kelly, R. J., Ruocco, G., Lerma, E., Tumlin, J., Wheelan, K. R., et al. (2020b). Pathophysiological Basis and Rationale for Early Outpatient Treatment of SARS-CoV-2 (COVID-19) Infection. *Am. J. Med.* 134, 16-22. doi:10.1016/j.amjmed.2020.07.003

Morozov, S., Andreychenko, A., Pavlov, N., Vladzymyrskyy, A., Ledikhova, N., Gombolevskiy, V., et al. (2020). MosMedData: Chest CT Scans with COVID-19 Related Findings Dataset. *medRxiv*. doi:10.1101/2020.05.20.20100362

Narin, A., Kaya, C., and Pamuk, Z. (2021). Automatic Detection of Coronavirus Disease (Covid-19) Using X-ray Images and Deep Convolutional Neural Networks. *Pattern Anal. Appl.*. doi:10.1007/s10044-021-00984-y

Naudé, W. (2020). Artificial Intelligence vs COVID-19: Limitations, Constraints and Pitfalls. *AI Soc.* 35, 761–765. doi:10.1007/s00146-020-00978-0

Oh, Y., Park, S., and Ye, J. C. (2020). Deep Learning Covid-19 Features on Cxr Using Limited Training Data Sets. *IEEE Trans. Med. Imaging*. 39, 2688–2700. doi:10.1109/tmi.2020.2993291

Ozturk, T., Talo, M., Yildirim, E. A., Baloglu, U. B., Yildirim, O., and Rajendra Acharya, U. (2020). Automated Detection of COVID-19 Cases Using Deep Neural Networks with X-ray Images. *Comput. Biol. Med.* 121, 103792. doi:10.1016/j.compbiomed.2020.103792

Pérez-Peña, R. (2020). *Virus Has Killed 1 Million Worldwide*.

Quanjel, M. J., van Holten, T. C., Gunst-van der Vliet, P. C., Wielaard, J., Karakaya, B., Sohne, M., et al. (2021). Replication of a Mortality Prediction Model in Dutch Patients with COVID-19. *Nat. Mach. Intell.* 3, 23–24. doi:10.1038/s42256-020-00253-3

Rahimzadeh, M., Attar, A., and Sakhaei, S. M. (2021). A Fully Automated Deep Learning-Based Network for Detecting COVID-19 from a New and Large Lung CT Scan Dataset. *Biomed. Signal Process. Control* 68, 102588. doi:10.1016/j.bspc.2021.102588

Sethy, P. K., Behera, S. K., Ratha, P. K., and Biswas, P. (2020). Detection of Coronavirus Disease (Covid-19) Based on Deep Features. *Int. J. Math. Eng. Manag. Sci.*. 5 (4), 643–651. doi:10.33889/ijmems.2020.5.4.052

Shibly, K. H., Dey, S. K., Islam, M. T.-U., and Rahman, M. M. (2020). COVID Faster R–CNN: A Novel Framework to Diagnose Novel Coronavirus Disease (COVID-19) in X-Ray images. *Inform. Med. Unlocked* 20, 100405. doi:10.1016/j.imu.2020.100405

Song, X. (2021). "Augmented Multi-center Graph Convolutional Network for COVID-19 Diagnosis," in *IEEE Transactions on Industrial Informatics* 17, 6499–6509. doi:10.1109/TII.2021.3056686

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a Simple Way to Prevent Neural Networks from Overfitting. *J. machine Learn. Res.* 15, 1929–1958.

Topol, E. J. (2020). Is My Cough COVID-19? *The Lancet*. 396, 1874. doi:10.1016/s0140-6736(20)32589-7

Torrey, L., and Shavlik, J. (2010). *Handbook Of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques 242-264*. Hershey, Pennsylvania: IGI global.

Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., et al. (2020). A Deep Learning Algorithm Using CT Images to Screen for Corona Virus Disease (COVID-19). *Eur. Radiol.*, 1–9. doi:10.1007/s00330-021-07715-1

Wang, L., Lin, Z. Q., and Wong, A. (2020). Covid-net: A Tailored Deep Convolutional Neural Network Design for Detection of Covid-19 Cases from Chest X-ray Images. *Scientific Rep.* 10, 1–12. doi:10.1038/s41598-020-76550-z

Wang, Z., Liu, Q., and Dou, Q. (2020). Contrastive Cross-Site Learning with Redesigned Net for COVID-19 CT Classification. *IEEE J. Biomed. Health Inform.* 24, 2806–2813. doi:10.1109/jbhi.2020.3023246

Wu, Y., and He, K. (2018). "Group Normalization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–19. New York City, NY: Springer.

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., et al. (2020). Prediction Models for Diagnosis and Prognosis of Covid-19 Infection: Systematic Review and Critical Appraisal. *bmj* 369, m1328. doi:10.1136/bmj.m1328

Xie, Y., and Richmond, D. (2018). "Pre-Training on Grayscale Imagenet Improves Medical Image Classification," in *Proceedings Of the European Conference On Computer Vision (ECCV) Workshops*. New York City, NY: Springer.

Xu, X., Jiang, X., Ma, C., Du, P., Li, X., Lv, S., et al. (2020). Deep Learning System to Screen Coronavirus Disease 2019 Pneumonia. *Engineering* 6, 1122–1129. doi:10.1016/j.eng.2020.04.010

Yan, L. (2020). An Interpretable Mortality Prediction Model for COVID-19 Patients. *Nat. Machine Intelligence* 2 (5), 283–288. doi:10.1038/s42256-020-0180-7

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., Oermann, E. K., et al. (2018). Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: a Cross-Sectional Study. *PLoS Med.* 15, e1002683. doi:10.1371/journal.pmed.1002683

Zhang, K., Liu, X., Shen, J., Li, Z., Ye Sang, Y., Wu, X., et al. (2020). Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of Covid-19 Pneumonia Using Computed Tomography. *Cell*. 181 (6), 1423-1433.e11. doi:10.1016/j.cell.2020.04.045

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 03 September 2021
doi: 10.3389/frai.2021.672050

Reliable and Interpretable Mortality Prediction With Strong Foresight in COVID-19 Patients: An International Study From China and Germany

Tao Bai[1†], Xue Zhu[2†], Xiang Zhou[3], Denise Grathwohl[3], Pengshuo Yang[2], Yuguo Zha[2], Yu Jin[1], Hui Chong[2], Qingyang Yu[2], Nora Isberner[3], Dongke Wang[1], Lei Zhang[1], K. Martin Kortüm[3], Jun Song[1], Leo Rasche[3], Hermann Einsele[3], Kang Ning[2] and Xiaohua Hou[1]**

[1]Division of Gastroenterology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [2]Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China, [3]Department of Internal Medicine II, University Hospital of Würzburg, Würzburg, Germany

OPEN ACCESS

Edited by:
Jake Y. Chen,
University of Alabama at Birmingham, United States

Reviewed by:
Dongxiao Zhu,
Wayne State University, United States
Pengwei Hu,
Merck, Germany

***Correspondence:**
Kang Ning
ningkang@hust.edu.cn
Xiaohua Hou
houxh@hust.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: 26 February 2021
Accepted: 26 July 2021
Published: 03 September 2021

Citation:
Bai T, Zhu X, Zhou X, Grathwohl D, Yang P, Zha Y, Jin Y, Chong H, Yu Q, Isberner N, Wang D, Zhang L, Kortüm KM, Song J, Rasche L, Einsele H, Ning K and Hou X (2021) Reliable and Interpretable Mortality Prediction With Strong Foresight in COVID-19 Patients: An International Study From China and Germany. Front. Artif. Intell. 4:672050. doi: 10.3389/frai.2021.672050

Cohort-independent robust mortality prediction model in patients with COVID-19 infection is not yet established. To build up a reliable, interpretable mortality prediction model with strong foresight, we have performed an international, bi-institutional study from China (Wuhan cohort, collected from January to March) and Germany (Würzburg cohort, collected from March to September). A Random Forest-based machine learning approach was applied to 1,352 patients from the Wuhan cohort, generating a mortality prediction model based on their clinical features. The results showed that five clinical features at admission, including lymphocyte (%), neutrophil count, C-reactive protein, lactate dehydrogenase, and α-hydroxybutyrate dehydrogenase, could be used for mortality prediction of COVID-19 patients with more than 91% accuracy and 99% AUC. Additionally, the time-series analysis revealed that the predictive model based on these clinical features is very robust over time when patients are in the hospital, indicating the strong association of these five clinical features with the progression of treatment as well. Moreover, for different preexisting diseases, this model also demonstrated high predictive power. Finally, the mortality prediction model has been applied to the independent Würzburg cohort, resulting in high prediction accuracy (with above 90% accuracy and 85% AUC) as well, indicating the robustness of the model in different cohorts. In summary, this study has established the mortality prediction model that allowed early classification of COVID-19 patients, not only at admission but also along the treatment timeline, not only cohort-independent but also highly interpretable. This model represents a valuable tool for triaging and optimizing the resources in COVID-19 patients.

Keywords: COVID-19, Wuhan cohort, Würzburg cohort, mortality prediction model, reliability, interpretability, foresight

INTRODUCTION

The pandemic of coronavirus disease 2019 (COVID-19) has become a public health emergency of international concern (Salyer et al., 2021; Sirleaf and Clark, 2021; Watson and Lilford, 2021). As of July 12, 2021, 187, 796, 841 confirmed infection cases have been reported by the World Health Organization, with a global mortality rate of 2.16% (https://covid19.who.int/). Even worse, the incidence of COVID-19 is continuously increasing worldwide, and areas already under control are likely to relapse (Setti et al., 2020). The proportion of critically ill COVID-19 patients is 18.5% (Epidemiology Working Group for Ncip Epidemic Response CCfDC, Prevention, 2020), and this high proportion of severe cases has put enormous pressure on medical systems, resulting in a serious shortage of medical resources (Rasmussen et al., 2020; Ammar et al., 2021; Wahlster et al., 2021).

In recent years, machine learning methods used for large clinical data analysis have been sprung up (Liang et al., 2020; Wu et al., 2020; Xiao et al., 2020; Zhu et al., 2020; Gomes and Serra, 2021; Ikemura et al., 2021; Wang et al., 2021). Yan et al. used the XGBoost classifier (Chen and Guestrin, 2016) to predict the outcome of 485 patients using the final samples at discharge, and they found three blood features that could be used as predictors, providing important evidence for clinical decision-making and patient management (Liang et al., 2020). Xiao et al. have used the HNC-LL score that considered hypertension, neutrophil count, C-reactive protein (CRP), lymphocyte count, and lactate dehydrogenase (LDH) to predict the severity of COVID-19 with AUC higher than 0.82 based on 442 patients (Xiao et al., 2020). Liang et al. developed a deep learning survival Cox model for 1,590 patients' triage, which was based on four clinical features and six phenotypic characteristics, to ensure patients at the greatest risk for severe illness receive appropriate care as early as possible (Liang et al., 2020). Wu et al. also used the Cox model to investigate the key risk factors and predicted the mortality rate of 21,392 COVID-19 patients based on demographic, clinical, and laboratory features and found that the mortality rate increased with time, especially for these critically ill patients (Wu et al., 2020).

Unfortunately, although the clinical features of COVID-19 patients have been reported in several recent publications (Gupta et al., 2020a; Xu et al., 2020), such as decreased lymphocytes and elevated CRP (Gupta et al., 2020a; Xu et al., 2020), the predictive powers and interpretations of these clinical features remain unclear. Additionally, since progression and outcome are critical for COVID-19 patients (Liang et al., 2020; Risch, 2020), timely monitoring from admission to outcome also has important clinical significance, making it possible to adjust treatment regimens in time, but this process is not entirely clear. Moreover, the foresight of a predictive model, as to how many days before discharge these features could accurately predict the patients' outcome, remains elusive. However, the association of these clinical features with phenotypic characteristics is also unclear. The robustness of the mortality prediction model along the timeline and the predictive power considering different preexisting diseases also need further exploration. Therefore, we performed this international, bi-institutional study to establish a mortality prediction model with the aim of early triaging and optimizing the resources.

METHODS

Ethical Approval

This study was approved by the Ethics Committee of Union Hospital, Tongji Medical College, Huazhong University of Science and Technology. Due to the retrospective nature of this study, the local institutional review board of the University of Würzburg waived the requirement for additional approval. This study was performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments.

Sample Description

Clinical data were collected from 1,441 COVID-19 patients from January 28, 2020, to March 29, 2020, at Wuhan Union Hospital (also called Wuhan cohort), China, for model development. Moreover, 96 patients with confirmed COVID-19 disease were collected from the University Hospital of Würzburg (also called Würzburg cohort), Germany, from March 6, 2020, to September 14, 2020, for independent test.

For the Wuhan cohort, more than 300 clinical features from hospital laboratory tests were recorded, and most patients have multiple sets of clinical features during their stay in the hospital. In addition, physical examinations, such as height, weight, temperature, sphygmus, systolic/diastolic pressure, respiratory rate, and heart rate, were performed upon admission of these COVID-19 patients. For robust analysis, clinical features that covered less than 30 samples, as well as samples containing fewer than three clinical features, were discarded (**Figure 1A**). After filtering out low-quality records, 1,352 patients and 130 clinical features were selected for systematic analysis. The average age of these patients was 58.22 (standard error: 14.90), and 50.52% of them were male, indicating a balanced gender. The minimal, maximal, and median duration from admission to discharge of the 1,352 patients is 0, 55, and 10 days, respectively. Among all of 1,352 COVID-19 patients, 1,221 patients survived and 131 died (**Supplementary Table S1**).

Clinical features (**Figure 1B**) from hospital laboratory tests were primarily composed of two parts: 101 numerical features, such as LDH and CRP, and 29 binary features, such as ABO blood type, Mp-IgM, and Mp-IgG. These clinical features were considered as candidate biomarkers for COVID19 mortality prognosis.

Phenotypic characteristics at admission (**Figure 1B**) were primarily composed of two parts: numerical and binary phenotypic characteristics. The numerical phenotypic characteristics included age, height, weight, temperature, sphygmus, systolic/diastolic pressure, respiratory rate, heart rate, and clinical classification. Binary phenotypic characteristics included records of gender, smoking status, and blood type.

Recent studies have already reported that the outcome of COVID-19 patients is greatly influenced by whether the

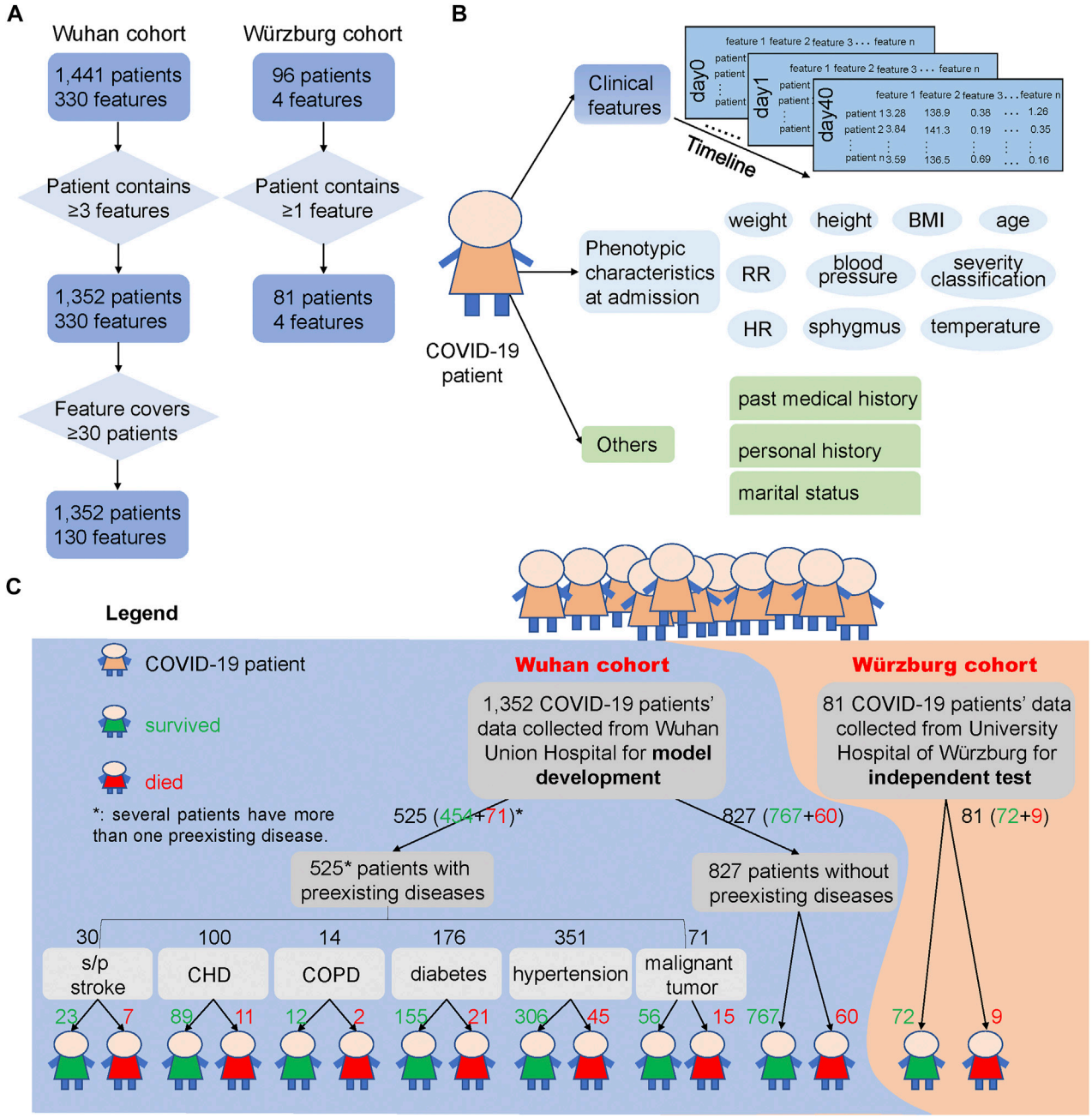


FIGURE 1 | COVID-19 patients and their clinical feature filtering process, phenotypic characteristics and clinical features used in this study, and the outcome of the two cohorts. **(A)** Process of filtering low-quality samples of the two cohorts. Here, 330 features in the Wuhan cohort were the union of 1,441 patients' clinical features and 130 features were the union of the 1,352 patients' clinical features after filtering. "≥3 features" says that the patients from the Wuhan cohort should contain at least three clinical features during the hospital stays, and "≥30 patients" says that the clinical features that collected from clinical laboratory should cover at least 30 patients and thus could be used for subsequent analysis. In the Würzburg cohort, "≥1 feature" says that the patient should contain at least one of the features from these four clinical features: lymphocyte (%), neutrophil count, LDH, and CRP. **(B)** Different types of clinical features and phenotypic characteristics used in the two cohorts. We used clinical features from hospital laboratory tests for developing the prediction model, and these clinical features were also used to test the association with phenotypic characteristics and other records. **(C)** Overview of samples used for model development and independent test. Samples of 1,352 COVID-19 patients from Wuhan Union Hospital (Wuhan cohort, the blue background) were used for building and testing the mortality prediction model, while samples of 81 COVID-19 patients from Germany (Würzburg cohort, the orange background) were used for independent test of the mortality prediction model. The green number represents the number of patients who survived from COVID-19, while the red number means the number of patients who died from COVID-19. Note that several patients have more than one preexisting disease.

patient has a preexisting disease (Azevedo et al., 2020; Zhou et al., 2020a; Williamson et al., 2020), such as CHD (Mai et al., 2020), hypertension (Itoh, 2020), and diabetes (Gupta et al., 2020b). Here, we divided 1,352 COVID-19 patients into seven groups: s/p stroke (23 survived and seven died), CHD (89, 11), chronic obstructive pulmonary disease (COPD) (12, 2), diabetes (155, 21), hypertension (306, 45), malignant tumor (56, 15), and those without preexisting diseases (767, 60) according to their past medical history (**Figure 1C**).

For the 96 patients in the Würzburg cohort, we have filtered out the patient who has not a single clinical feature among the four clinical features (lymphocyte (%), neutrophil count, LDH, and CRP) (**Figure 1A**). After this process, 81 samples were retained and utilized for independent test. For these 81 patients, their phenotypic characteristics including systolic pressure, diastolic pressure, temperature, heart rate, SpO_2, age, and respiratory rate were also used for analysis. The average age of these patients was 67.15 years (standard error: 15.17), which was significantly higher than that of patients in the Wuhan cohort (t-test, $p = 0.0005$). 62.96% of them are male, 53.67% of them have respiratory failure, and 41.46% of them need mechanical ventilation. Among them, 72 survived and nine died from COVID-19 (**Supplementary Table S2**).

Severity Classification

According to the diagnosis and treatment of pneumonia infected by the new novel coronavirus (the trial seventh edition) (National Health Commission of the People's Republic of China, 2020), the patient's severity classification was divided into three classifications, general, severe, and critical, according to their symptoms at admission. In this work, among 1,352 patients from the Wuhan cohort, 896 were in general, 393 were in severe, and 63 were in critical. For the Würzburg cohort, 24 were in general, 35 were in severe, and 22 were in critical. Here, we defined severity classification as follows: general as 1, severe as 2, and critical as 3.

Clinical Feature Profiling

Using patient samples at admission, all numerical clinical features were normalized to a range [0, 1]. These normalized data with an average abundance ≥0.001 were illustrated as boxplots using the R package "ggplot2". To illustrate differences between patients who survived and died, as well as between patients with or without preexisting diseases, principal coordinate analysis (PCoA) was performed using all patients' numerical clinical features at admission based on the Jaccard coefficient for distance measurement using the R package "vegan".

Feature Selection and Development of a Prediction Model Utilizing Clinical Features

To identify the most important clinical features that reflect differences among the samples, feature selection was employed for a deeper understanding of COVID-19 infection. We assessed the contribution of each clinical feature to facilitate the decisions of the algorithm. Considering both MeanDecreaseAccuracy and MeanDecreaseGini, the top five discriminatory clinical features were selected. Different Random Forest (RF) models were tested on the top five important clinical features, as well as their different combinations according to their importance.

To develop a mortality prediction model that is capable of distinguishing the outcome of COVID-19 patients, RF analysis was performed by randomForest() function in R (package "randomForest"). For the sample size larger than 100, we randomly selected 90% of samples as training set and 10% of samples as testing set using sample() function with replacement. In this process, replace parameter was set as true, which specifies using the Bootstrap method for random sampling. For each model, based on each training set, the important parameters ntree (number of decision trees contained in the RF model) and mtry (variable sampling values for each iteration) were trained and estimated with the out-of-bag (OOB) value. The importance was set as true for calculating the importance of each variable in the model, which was mainly used in conjunction with the importance() function. The proximity parameters were set as true for calculating the proximity matrix of the model, which is mainly used in conjunction with the MDSplot() function to realize the visualization of random forest. The na.action parameter specifies the methods for handling the missing values and was set as na.omi (that is, delete the samples with missing values of all features). Other parameters were set as default. A traversal search was performed on all clinical features to obtain the minimum OOB value. The value of mtry was determined by the OOB value (that is, the index of the minimum OOB value). Then, combining the outcome of COVID-19 patients, the mtry value was iterated to obtain an optimal ntree. This process was iterated 15,000 times or more to construct the most accurate model. When the error tree approaches stable, the minimum number of trees was the best value for ntree. This trained model was used for predicting the outcome of the testing set.

Evaluation of Prediction Models

To evaluate the performance of the RF model, we used several standard statistic parameters: accuracy, precision, sensitivity or recall, specificity, and F1 scores. Here, we defined the prediction result: survived-survived as TP and died-died as TN. The formulas of the parameters mentioned above are defined as follows:

$$\text{accuracy} = (TP + TN)/(TP + TN + FP + FN), \quad (1)$$

$$\text{precision} = TP/(TP + FP), \quad (2)$$

$$\text{recall} = TP/(TP + FN), \quad (3)$$

$$\text{specificity} = TN/(TN + FP), \quad (4)$$

$$F1 = 2^{*}\,\text{precision}^{*}\text{recall}/(\text{precision} + \text{recall}), \quad (5)$$

where TP, TN, FP, and FN stand for true-positive, true-negative, false-positive, and false-negative rates, respectively.

Correlation Analysis Between Phenotypic Characteristics and Clinical Features

To better understand the relationship between phenotypic characteristics and the mortality rate of patients, we used the Pearson coefficient to examine the correlation between

phenotypic characteristics and clinical features. Again, we organized these correlation values along the timeline to identify the dynamics of such correlations during treatment progression.

Evaluation of Prediction Models along the Timeline

Most patients have multiple sets of clinical features during their stay in the hospital, allowing for a series of mortality prediction models along the timeline. Here, we ordered these mortality prediction models in two directions: from admission forward to outcome to directly provide day-to-day guidance for clinics and from outcome backward to admission to evaluate the robustness and prediction power of the model against the time of hospital stay.

Development of High-Quality and Interpretable Binary Decision Tree for Clinical Diagnosis

Based on the five selected features, we aimed to develop a high-quality decision tree for clinical diagnosis. To train the RF model, the dataset was randomly separated into two groups: the training set (90% of entries) was applied to construct the mortality prediction model, and the testing set (10% of entries) was applied to validate the mortality prediction model. For datasets with a sample size of less than 100, we used 70% of the dataset for training and 30% for testing to reduce the contingency error. This process was iterated 15,000 times to construct the most accurate model. The most discriminative clinical feature was used as the root node of this binary decision tree, and the child nodes were hierarchically formed according to their distinguishing ability until all samples are completely distinguished. Finally, the decision tree was visualized by rpart() function in R (package "party").

Development of a Prediction Model for Different Preexisting Diseases

Considering the influence of preexisting diseases on the outcome in COVID-19 patients, we also used the first samples of patients with preexisting diseases as a training dataset to build the mortality prediction models: s/p stroke, CHD, COPD, diabetes, hypertension, malignant tumor, and those without preexisting diseases. For a dataset with a sample size larger than 100, we used 90% of the dataset for training and 10% for validation. For a dataset with a sample size smaller than 100, we used 70% of the dataset for training and 30% for testing to validate the model to reduce the contingency error.

Independent Test of the Mortality Prediction Model Using the Würzburg Cohort

To examine the reliability, interpretability, and foresight of our mortality prediction model developed based on the Wuhan cohort, 81 samples at admission from the Würzburg cohort were used for independent test. Pearson coefficient was also used to evaluate the association between the four clinical features (lymphocyte (%), neutrophil count, LDH, and CRP) and phenotypic characteristics (systolic pressure, diastolic pressure, temperature, heart rate, SpO_2, age, and respiratory rate).

RESULTS

In this study, we have recruited two independent cohorts from China (the Wuhan cohort) and Germany (the Würzburg cohort) for building and testing a mortality prediction model, respectively. The Wuhan cohort contained 1,352 COVID-19 patients from Wuhan Union Hospital, and it has been utilized for establishing a multi-feature and time-series aware machine learning models. The Würzburg cohort consists of 81 COVID-19 patients and has been used as an independent validation cohort.

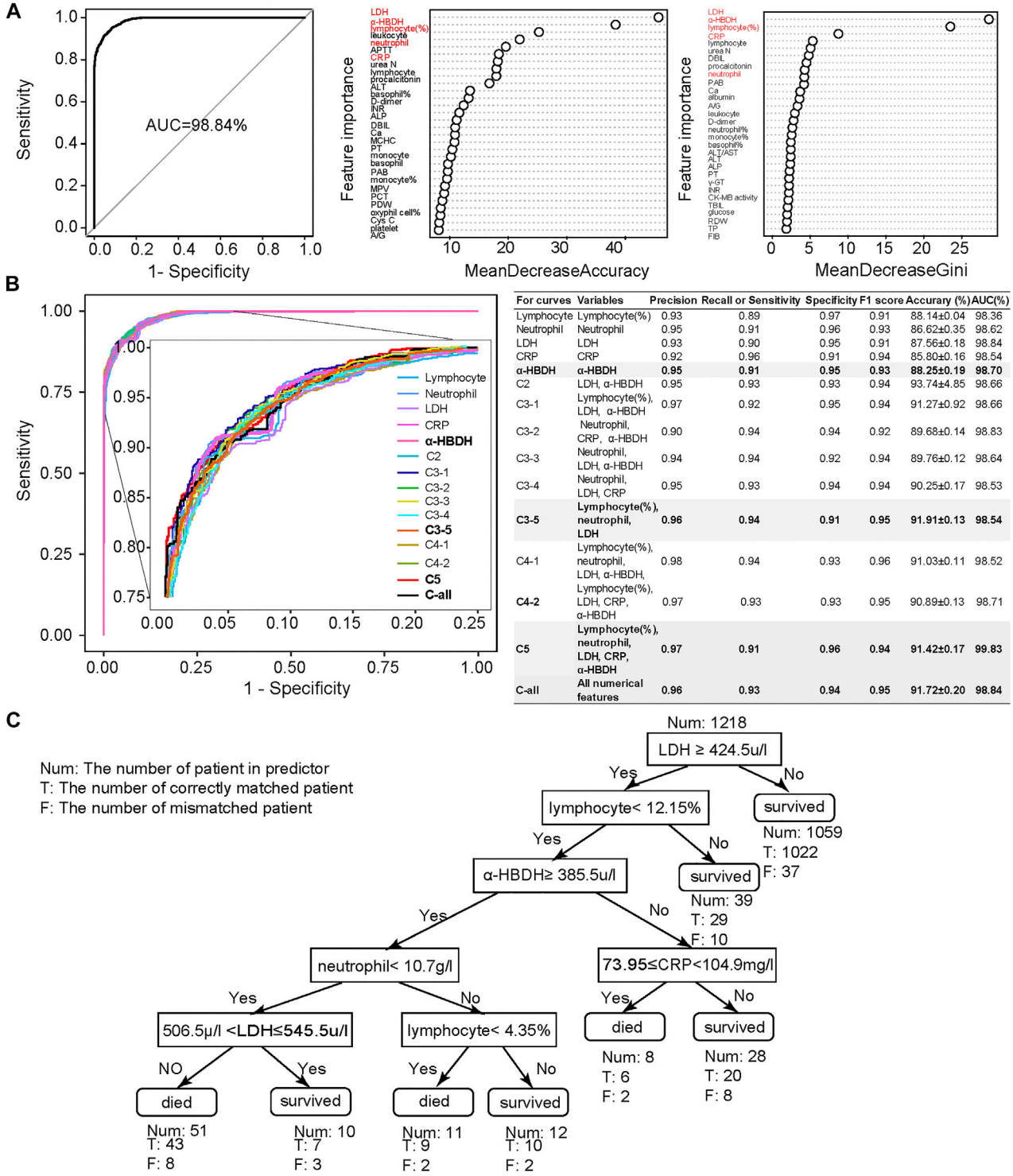
Data Resource and General Profiles of COVID-19 Patients from Wuhan Cohort

1,352 patients were enrolled in the Wuhan cohort, who had more than three clinical features (such as neutrophil count, CRP, lymphocyte count, LDH, albumin, direct bilirubin, and creatine kinase) (Liang et al., 2020; Xiao et al., 2020) and detailed medicinal records from January 28, 2020, to March 29, 2020. The distribution of the number of patients with clinical laboratory tests on a daily basis, as well as the total number of diagnoses for each patient, is shown in **Supplementary Figure S1**. Among them, the mortality rates in patients with preexisting diseases: s/p stroke, coronary heart disease (CHD), chronic obstructive pulmonary disease (COPD), diabetes, hypertension, and malignant tumor were 23.33, 11.00, 14.29, 11.93, 12.82, and 21.13%, respectively (**Supplementary Figures S2A,B**). These mortality rates were significantly higher (*t*-test, $p < 0.001$) than those in patients without preexisting diseases (mortality rate: 7.26%). PCoA showed that if we used all clinical features, these patients cannot be clearly separated (**Supplementary Figure S2C**). In addition, these patients could not be separated by whether they had a preexisting disease or not (**Supplementary Figures S2D–J**). This highlights the importance of clinical feature selection and developing the mortality prediction models to differentiate patients.

Development and Evaluation of Clinical Feature Selection and Mortality Prediction Model for Early Prognosis Based on Wuhan Cohort

We first developed a mortality prediction model based on patients' samples at admission, since such prediction is of paramount importance in clinics (Risch, 2020). This model took the clinical features and outcomes into consideration, aiming to optimize the medical resources, as well as preemptive therapy.

Before developing a mortality prediction model, we divided the 130 clinical features into two parts: 101 numerical features



For curves	Variables	Precision	Recall or Sensitivity	Specificity	F1 score	Accuray (%)	AUC(%)
Lymphocyte	Lymphocyte(%)	0.93	0.89	0.97	0.91	88.14±0.04	98.36
Neutrophil	Neutrophil	0.95	0.91	0.96	0.93	86.62±0.35	98.62
LDH	LDH	0.93	0.90	0.95	0.91	87.56±0.18	98.84
CRP	CRP	0.92	0.96	0.91	0.94	85.80±0.16	98.54
α-HBDH	**α-HBDH**	**0.95**	**0.91**	**0.95**	**0.93**	**88.25±0.19**	**98.70**
C2	LDH, α-HBDH	0.95	0.93	0.93	0.94	93.74±4.85	98.66
C3-1	Lymphocyte(%), LDH, α-HBDH	0.97	0.92	0.95	0.94	91.27±0.92	98.66
C3-2	Neutrophil, CRP, α-HBDH	0.90	0.94	0.94	0.92	89.68±0.14	98.83
C3-3	Neutrophil, LDH, α-HBDH	0.94	0.94	0.92	0.94	89.76±0.12	98.64
C3-4	Neutrophil, LDH, CRP	0.95	0.93	0.94	0.94	90.25±0.17	98.53
C3-5	**Lymphocyte(%), neutrophil, LDH**	**0.96**	**0.94**	**0.91**	**0.95**	**91.91±0.13**	**98.54**
C4-1	Lymphocyte(%), neutrophil, LDH, α-HBDH,	0.98	0.94	0.93	0.96	91.03±0.11	98.52
C4-2	Lymphocyte(%), LDH, CRP, α-HBDH	0.97	0.93	0.93	0.95	90.89±0.13	98.71
C5	**Lymphocyte(%), neutrophil, LDH, CRP, α-HBDH**	**0.97**	**0.91**	**0.96**	**0.94**	**91.42±0.17**	**99.83**
C-all	**All numerical features**	**0.96**	**0.93**	**0.94**	**0.95**	**91.72±0.20**	**98.84**

FIGURE 2 | Clinical feature selection and mortality prediction results utilizing the first samples at admission. **(A)** Clinical feature selection based on 1,352 samples and 101 numerical features. Features were ranked by MeanDecreasedAccuracy and MeanDecreasedGini according to their importance. Considering both of these import RF parameters, we selected five important clinical features: lymphocyte (%), neutrophil count, LDH, CRP, and α-HBDH. **(B)** Comparison of receiver operating characteristic (ROC) curves and diagnostic performance of all numerical clinical features, subfeature combinations, as well as each selected single feature, using the first samples at admission (also referred to as admission-day 0). **(C)** The binary decision tree for predicting the outcome of COVID-19 patients based on the five selected clinical features at admission (admission-day 0). Annotations: Num: the number of patients in the predictor; T: the number of correctly matched patients; F: the number of mismatched patients. Here, the Num above the root node indicates the total samples used for building the binary decision tree.

and 29 binary features. For numerical features, those features which are identified with the average abundance ≥0.001 are shown in **Supplementary Figure S3**. 101 numerical clinical features with at least 30 samples' coverage were considered as the outcome predictors and were used to build the mortality prediction model. We used 90% of the samples for model training and 10% for testing to validate the model.

Combined MeanDecreaseAccuracy and MeanDecreaseGini (**Figure 2A**), lymphocyte (%), neutrophil count, C-reactive protein (CRP), lactic acid dehydrogenase (LDH), and α-hydroxybutyric dehydrogenase (α-HBDH) were selected for developing an optimized model, where lymphocyte (%) is an immune disorder indicator (Trowell, 1947), neutrophil count represents infection (Xie et al., 2020), CRP represents inflammatory response (Vermeire et al., 2004; Sabrina et al., 2012), and both LDH and α-HBDH represent tissue lesions (Sanwald and Kirk, 1966; Kishaba et al., 2014).

We then used these five selected numerical clinical features (lymphocyte (%), neutrophil count, CRP, LDH, and α-HBDH), as well as different combinations of the subset of these five clinical features according to their importance, for prediction (**Figure 2B**). Results showed that the performance of these five clinical features could be comparable to the results predicted by all numerical features. Considering the F1 score, accuracy, and AUC, the combination of lymphocyte (%), neutrophil count, and LDH also showed high performance, especially the performance of α-HBDH used alone (bold in **Figure 2B**). Several specified combinations of three out of these five clinical features, such as the combination of lymphocyte (%), neutrophil count, and LDH, also reached more than 91% accuracy and 99% AUC at admission. However, in clinics, these five features covered more types of clinical symptoms: lymphocyte (%) is an immune disorder indicator (Trowell, 1947), neutrophil count represents infection (Xie et al., 2020), CRP represents inflammatory response (Vermeire et al., 2004; Sabrina et al., 2012), and both LDH and α-HBDH represent tissue lesions (Sanwald and Kirk, 1966; Kishaba et al., 2014). Thus, we confirmed these five clinical features as credible biomarkers.

To benchmark with other classification algorithms, we also used FEAST (an expectation–maximization-based unsupervised learning method) (Shenhav et al., 2019) and JSD (Jensen–Shannon divergence) methods (Lin, 1991) to predict the outcome of COVID-19 patients based on all features, the top five features, and the top three features (**Supplementary Figure S4**). Results demonstrated that the RF model was more or equally credible for constructing the mortality prediction model. The neural network (Kriegeskorte and Golan, 2019) with two hidden layers (the first layer has 128 neurons and the second layer has eight neurons) also illustrated that RF model based on the combination of lymphocyte (%), neutrophil, LDH, CRP, and α-HBDH could best predict the outcome of COVID-19 patients (**Supplementary Figure S5**). Moreover, all three methods (RF, FEAST, JSD, and neural network) showed the best distinguishing power when using the top five clinical features to construct the model.

We also used the binary clinical features (such as urine occult blood, blood type, and COVID-19 nucleic acid) to build the mortality prediction model (**Supplementary Figure S6A**). Based on the contribution of each feature, we selected urine protein (UPRO), urine occult blood (UOB), monospecific antibodies of blood type (Ab-monospecific-B), ABO blood type (ABO), and ketones (KET) for further model improvement. Their different combinations and performance are shown in **Supplementary Figure S6B**. Among them, the combination of UPRO, UOB, and KET (accuracy = 99.61%; AUC = 99.96%) was outstanding from the others, followed by UPRO, all binary features, and the combination of these five features.

Finally, we emphasized that all of the above results were based on the first samples at admission, since it is more important for the clinical prediction to utilize these samples. It was noticed that a recently published study used the final samples of COVID-19 patients for predicting their outcome (Yan et al., 2020), and we also used the final samples in the Wuhan cohort to assess our model based on five selected features (**Supplementary Figure S7**), with results showing high prediction accuracy. Yet, the prediction accuracy and AUC based on first samples at admission (**Figure 2B**) were comparable to those based on these final samples for the Wuhan cohort. These results confirmed again that patients with a high mortality rate could be accurately predicted at admission, which could be used for prioritizing critically ill patients to potentially reduce the mortality rate.

Clinical Features Have Profound Association with Phenotypic Characteristics in the Wuhan Cohort

Notable correlations were observed between phenotypic characteristics and clinical features associated with COVID-19 (**Figure 3** and **Supplementary Figures S8, S9**). Among 101 numerical clinical features, many of them have shown significant correlations with age, respiratory rate, and severity classification of patients. Expect for lymphocyte (%), neutrophil count, LDH, CRP, and α-HBDH were positively correlated with age ($p < 0.05$) along the timeline. Since the above analyses also confirmed that these five clinical features are tightly associated with patient outcomes (**Figure 2**), these associations partially verified the fact that elder patients were more likely to die from COVID-19. LDH, CRP, and α-HBDH were also positively correlated with respiratory rate and severity classification ($p < 0.05$) in patients (896 were in general, 392 were in severe, and 63 were in critical), illustrating the importance of these phenotypic characteristics on outcome in COVID-19 patients. The result also showed dynamic changes in the associations of these clinical features with phenotypic characteristics over time, especially for the five clinical features used for model prediction.

Time-Series Analysis Reveals That the Mortality Prediction Model Is Very Robust along the Timeline

Evaluation of the mortality prediction model along the timeline forward from admission day as the start point: Because these clinical features are dynamic along the timeline, and in clinics, the

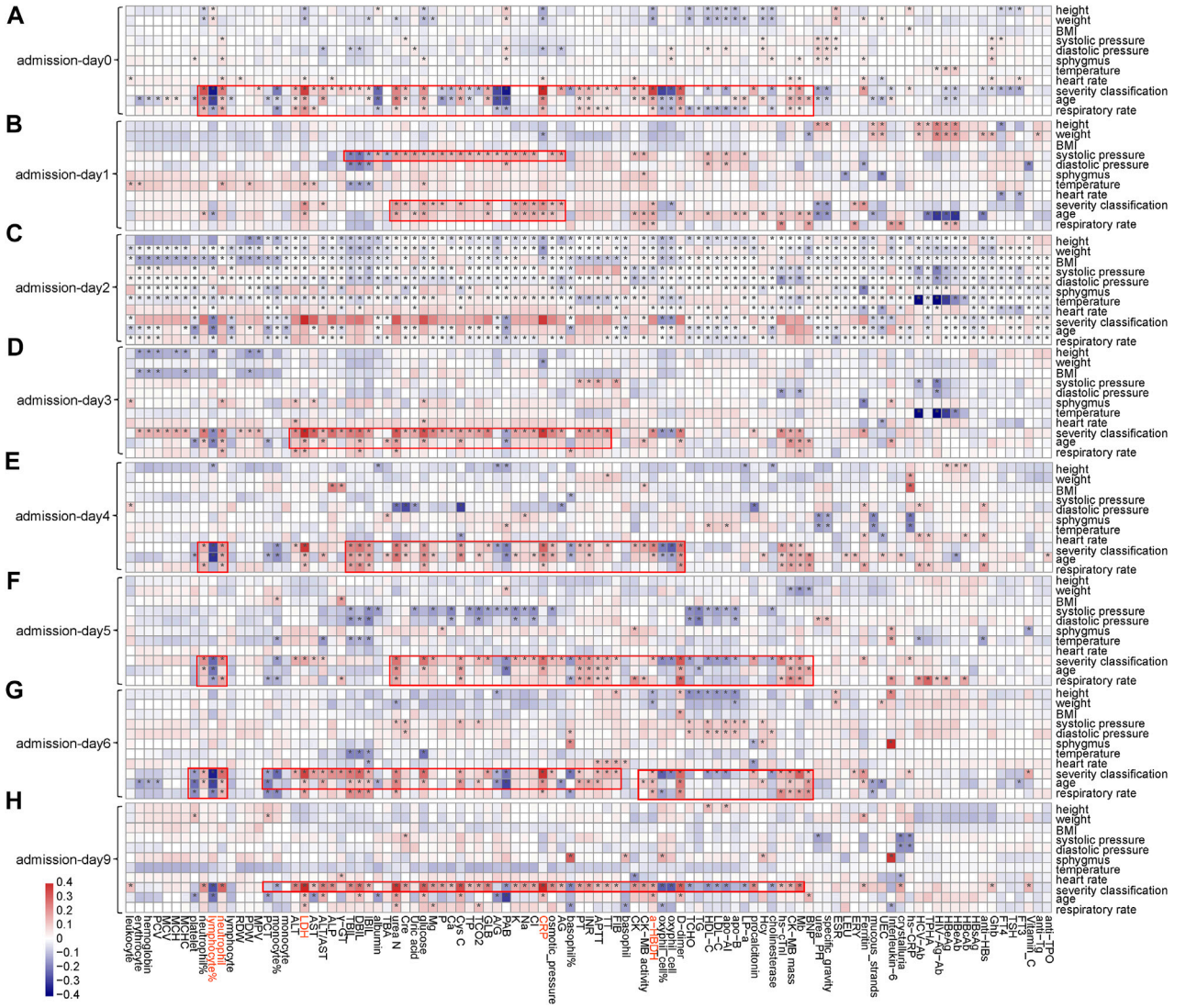


FIGURE 3 | Associations between clinical features and phenotypic characteristics based on several representative time points. **(A)** Based on admission day (admission-day 0). **(B)** Based on the third day after admission (admission-day 1). **(C)** Based on the sixth day after admission (admission-day 2). **(D)** Based on the ninth day after admission (admission-day 3). **(E)** Based on the admission day (admission-day 4). **(F)** Based on the third day after admission (admission-day 5). **(G)** Based on the sixth day after admission (admission-day 6). **(H)** Based on the ninth day after admission (admission-day 9). Note: *represents a significant correlation between a phenotypic characteristic and a clinical feature (Pearson correlation: $p < 0.05$).

progression and outcome of patients are critical (Liang et al., 2020; Risch, 2020). Therefore, we used the admission day of each patient as the start point and built mortality prediction models day by day after admission along the timeline. The number of samples enrolled on a daily basis is shown in **Figure 4A** from admission-day 0. We used 90% of the dataset for training and 10% for testing. For datasets with a sample size of less than 100, we used 70% of the dataset for training and 30% as a test set for validation. Since the sample number was less than 50 for patients who stayed in the hospital longer than 40 days, we only used the dataset from admission-day 0 to admission-day 40 to build the time-series mortality prediction models. Results confirmed that our mortality prediction model was very robust over time, suggesting that according to the prediction outcome of patients, clinics could adjust the treatment plan at any time, which could provide higher quality treatment for patients.

Evaluation of the mortality prediction model along the timeline backward from discharge day as the start point: To prove the robustness of our mortality prediction model and how many days in advance it could predict the outcome of COVID-19 patients, we used the discharge day of each patient as the start point. Prediction accuracies were evaluated backward day by day (**Figure 4B**) from discharge-day 0. The mortality prediction model based on five clinical features also reached more than 91% accuracy and 99% AUC (usually 10 days or more in advance of the outcome) (**Figure 4B**), confirming this mortality prediction

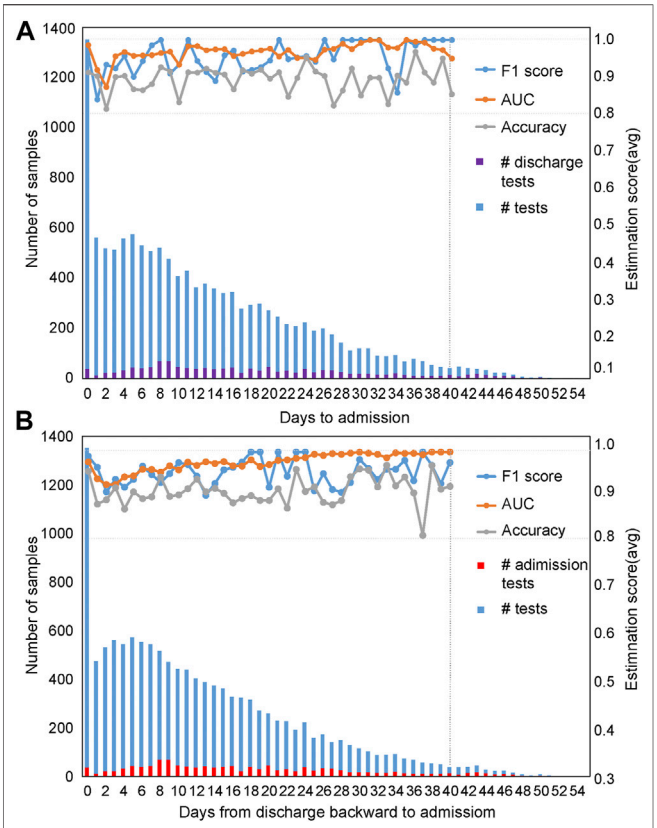


FIGURE 4 | Evaluation of prediction results based on time series with five selected features. The prediction results were evaluated based on time series using the admission-day 0 **(A)** and the discharge-day 0 **(B)** as start points. The first *y*-axis represents the number of samples enrolled, while the second *y*-axis represents the estimation scores. The purple bar represents the number of patients discharged from the hospital on that day, while the red bar indicates the number of patients admitted to the hospital on that day.

model is very robust over time when patients were in the hospital and indicating the strong association of these five clinical features with the progression of treatment.

The Highly Accurate and Interpretable Binary Decision Tree for Clinical Diagnosis

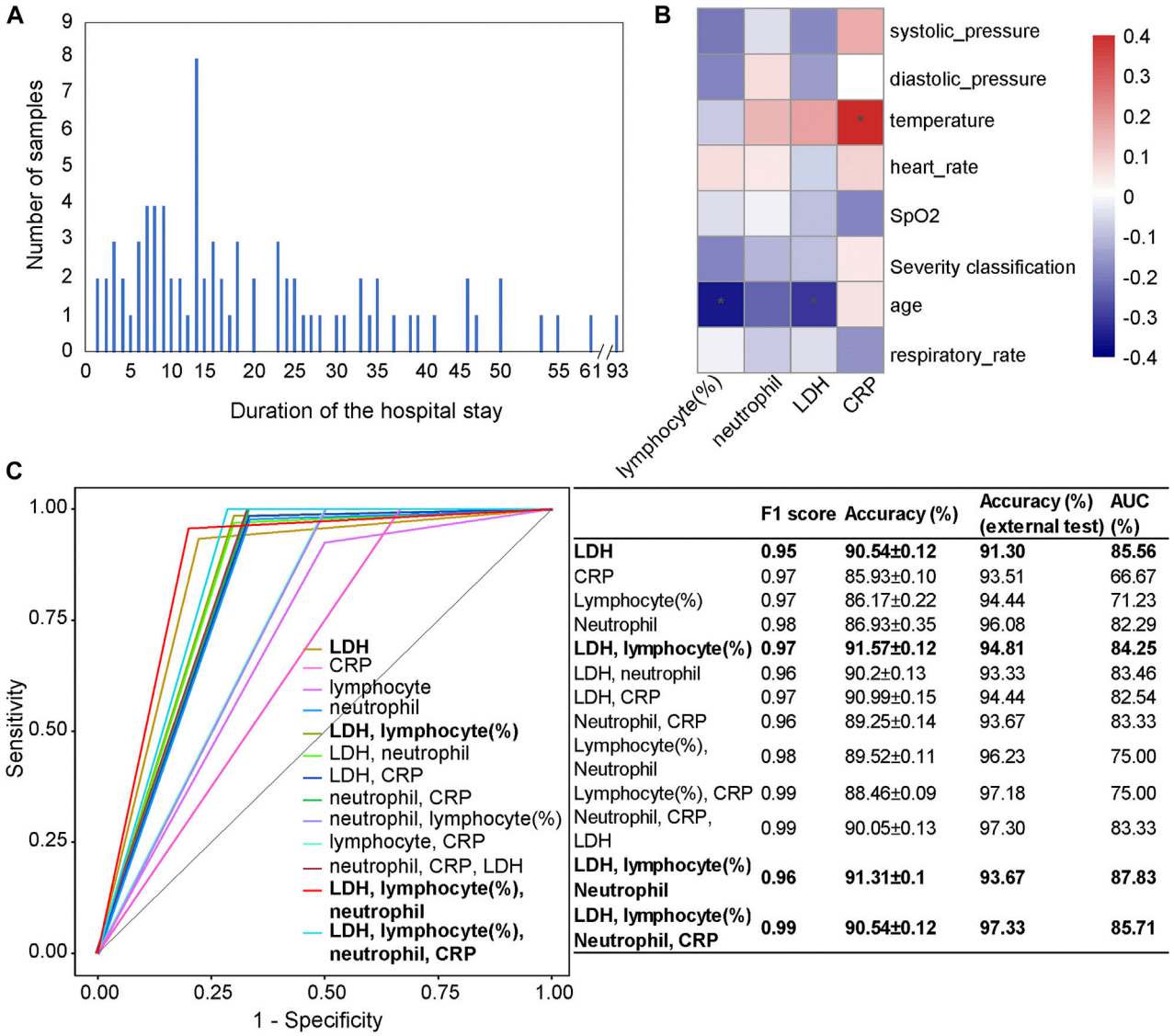
To make the prediction interpretable, we also generated a series of decision trees (along the timeline) for assisting clinical diagnosis based on the Wuhan cohort. The decision tree is hierarchically organized by the distinguishing ability of these five clinical features based on the first samples at admission (**Figure 2C**). LDH could distinguish 87% of samples with more than 96% accuracy and was used as the root node of this decision tree. The remaining 13% of samples were differentiated by a combination of these five clinical features. The binary decision tree of the final samples at discharge was simpler than that of the first samples at admission (**Supplementary Figure S10E**). The decision trees based on other time points are shown in **Supplementary Figures S10A–D**, confirming that using these five clinical features was more comprehensive and precise. These results also suggested that the mortality prediction model based on the admission samples, rather than the discharge samples, could already provide outcome prediction and clinical guidance for personalized treatment with high fidelity.

The binary decision tree, either based on samples at admission or based on discharge, was also highly interpretable for clinical diagnosis. The elevated LDH was associated with patients' death: LDH larger than 445 u/l was a significant risk factor related to death in cases with severe COVID-19 (Zhou et al., 2020b; Li et al., 2020), which was consistent with our results. The increased level of α-HBDH was also found as a critical risk factor associated with the severity of COVID-19 patients (Dong et al., 2020). The decreased amount of lymphocyte (lymphopenia) and neutrophil (neutrophilia), together with the increased number of CRP and LDH, showed the immunological response to the virus, followed by severe virus infection (Frater et al., 2020; Lippi and Plebani, 2020). In summary, current published clinical evidence could well support our decision tree.

Prediction Power Considering Different Preexisting Diseases

For different preexisting diseases, the clinical features that can accurately mark the COVID-19 patients' outcomes are generally different. Previous studies have shown that preexisting disease increases the risk of COVID-19 mortality rate (Williamson et al., 2020). We also used the six preexisting diseases to evaluate the mortality prediction model based on the five selected clinical features (**Supplementary Figure S11**).

Out of the five selected clinical features, feature combinations should be different for each of the different preexisting diseases. Therefore, for each of the preexisting diseases, we performed feature importance evaluation before mortality prediction model evaluation. For patients with s/p stroke (**Supplementary Figure S11A**), considering F1 score, accuracy, and AUC, the combination of LDH, CRP, and α-HBDH showed the highest performance, followed by the combination of LDH and α-HBDH, then all five features. The results for patients with CHD are illustrated in **Supplementary Figure S11B**. Except for using the five features, the combination of LDH, CRP, and α-HBDH showed the highest performance. For patients with COPD, a combination of neutrophil count, lymphocyte (%), and LDH showed the highest performance (**Supplementary Figure S11C**). For patients with diabetes, among all combinations of clinical features, LDH showed the highest performance (**Supplementary Figure S11D**), indicating that LDH could be used to distinguish the outcome of COVID-19 patients. For patients with hypertension, results indicated that a combination of neutrophil count, lymphocyte (%), and LDH could be used as biomarkers for predicting the outcome of COVID-19 patients with hypertension (**Supplementary Figure S11E**). For patients with malignant tumor, the combination of all five features showed the highest performance, followed by the combination of neutrophil count and lymphocyte (%) (**Supplementary Figure S11F**). For patients without preexisting diseases, results showed that we can use lymphocyte (%), LDH, CRP, and α-HBDH to accurately predict the outcome of these patients (**Supplementary Figure S11G**).



	F1 score	Accuracy (%)	Accuracy (%) (external test)	AUC (%)
LDH	**0.95**	**90.54±0.12**	**91.30**	**85.56**
CRP	0.97	85.93±0.10	93.51	66.67
Lymphocyte(%)	0.97	86.17±0.22	94.44	71.23
Neutrophil	0.98	86.93±0.35	96.08	82.29
LDH, lymphocyte(%)	**0.97**	**91.57±0.12**	**94.81**	**84.25**
LDH, neutrophil	0.96	90.2±0.13	93.33	83.46
LDH, CRP	0.97	90.99±0.15	94.44	82.54
Neutrophil, CRP	0.96	89.25±0.14	93.67	83.33
Lymphocyte(%), Neutrophil	0.98	89.52±0.11	96.23	75.00
Lymphocyte(%), CRP	0.99	88.46±0.09	97.18	75.00
Neutrophil, CRP, LDH	0.99	90.05±0.13	97.30	83.33
LDH, lymphocyte(%) Neutrophil	**0.96**	**91.31±0.1**	**93.67**	**87.83**
LDH, lymphocyte(%) Neutrophil, CRP	**0.99**	**90.54±0.12**	**97.33**	**85.71**

FIGURE 5 | Independent test results on the Würzburg cohort utilizing the first samples at admission. **(A)** The number of patients categorized by their duration of the hospital stay. **(B)** The associations between clinical features (LDH, CRP, lymphocyte (%), and neutrophil count) and phenotypic characteristics. **(C)** Comparison of receiver operating characteristic (ROC) curves and diagnostic performance of four clinical features, the combinations of a subset of features, as well as each selected single feature, using the first samples at admission. Note: * represents a significant correlation between a phenotypic characteristic and a clinical feature (Pearson correlation: $p < 0.05$).

Evaluation of the Mortality Prediction Model Using the Independent Würzburg Cohort

The reliability, interpretability, and foresight of our mortality prediction model were further confirmed in another independent cohort collected from Germany, the Würzburg cohort, with samples collected from March to September 2020 (**Figure 5**). For the patients in the Würzburg cohort, their duration of stay in the hospital is usually 5–20 days (**Figure 5A**). All samples used in the Würzburg cohort were the patient samples at admission.

We used four clinical features (lymphocyte (%), neutrophil count, LDH, and CRP), as well as their different combinations to test our mortality prediction model (**Figure 5C**). Considering F1 score, accuracy, and AUC, the combination of LDH, lymphocyte (%), neutrophil count, and CRP (accuracy = 97.33%; AUC = 85.71%) showed the highest performance among different combinations. Other combinations, such as the combination of LDH, lymphocyte (%), and neutrophil count (accuracy = 93.67%, AUC = 87.83%) and the combination of LDH and lymphocyte (%) (accuracy = 94.81%, AUC = 84.25%), also performed well. When only one clinical feature was used, LDH (accuracy = 91.30%, AUC = 85.56%) showed the highest performance, which was consistent with the results on the Wuhan cohort and a previous study (Liang et al., 2020).

From the Pearson correlation analysis (**Figure 5B**) between these four clinical features (LDH, lymphocyte (%), neutrophil count, and CRP) and the phenotypic characteristics (systolic pressure, diastolic pressure, temperature, heart rate, SpO_2, age, and respiratory rate), we could observe that there was a significantly negative correlation between lymphocyte (%) and age ($p < 0.05$), which was consistent with the general pattern of COVID-19 patients. CRP was significantly positively correlated with temperature ($p < 0.05$), which was consistent with the result in the Wuhan cohort.

Furthermore, as the duration of stay in the hospital of patients is usually 5–20 days, the strong foresight of the mortality prediction model has again been validated on the Würzburg cohort. Furthermore, one male patient aged 54 has a hospital stay of 93 days before recovery, and our mortality prediction model has successfully predicted his outcome.

DISCUSSIONS AND CONCLUSION

Our study enrolled two independent cohorts of COVID-19 patients for reliable, interpretable, and universal mortality model evaluation. Through multiple analyses including RF analysis, association analysis, time-series analysis, etc., the mortality prediction model was established, evaluated, and achieved clinically creditable prediction power on the Wuhan cohort and Würzburg cohort.

The mortality prediction model proposed in this study could help identify critically ill patients early and provide preferential treatment for each individual. Firstly, the five important clinical features (lymphocyte (%), neutrophil count, CRP, LDH, and α-HBDH) were identified. These five features could reflect several important aspects of disease development, such as viral infection (Trowell, 1947), coexistence of other infections (Xie et al., 2020), immune reaction during pneumonia (Sabrina et al., 2012), the severity of inflammatory response (Vermeire et al., 2004), tissue/cell damage, and cardiac injury (Sanwald and Kirk, 1966; Kishaba et al., 2014), which could provide more information to monitor the progression of patients. Secondly, these five features could be used for predicting the outcome of COVID-19 patients with high accuracy. Thirdly, the foresight of the mortality prediction model was strong up to as early as 40 days or more before discharge. This indicates that our model could allow resource optimization to be conducted many days ahead, and physicians can make a preliminary judgment on the prognosis of patients according to this model to prompt the choice of clinical intervention in later stages.

Our mortality prediction model shows superior prediction power at different time points during the course of the disease. Robust prediction power at different time points (**Figure 4**) also suggests that the mortality prediction model provides important indicators for disease monitoring, indicating early clinical intervention for clinical treatment. Our mortality prediction model also shows superior prediction power for different preexisting diseases of patients, indicating the robustness of the mortality prediction model. These results could serve well as the basis for personalized treatment of COVID-19 patients.

Our finding in the Wuhan cohort (model development) has also been tested in an independent cohort from Germany (Würzburg cohort). Although the international aspects such as the ethnicities, healthcare systems, hygienic measures, local regulations, and management strategies, as well as their average age (*t*-test, $p = 0.0005$), are different in these two cohorts, our mortality prediction model has also shown the high prediction power in tens of days ahead of patients' discharge, underlining the robustness and the foresight of this model.

The second COVID-19 wave in Europe is ongoing. This mortality prediction model has been validated at a European center and might provide a useful instrument for triaging the patients and optimizing the resources. Because we have a series of mortality prediction models with constant high accuracy along with the whole duration of patients' stay in the hospital, we could adjust treatment for possibly serious patients on a day-to-day basis to reduce the mortality rate of patients with COVID-19 as much as possible. In addition, our study also provides new insight into the mortality prediction model's application value in other infectious disease outbreaks in the future.

In conclusion, this study has established a mortality prediction model that allowed early classification of COVID-19 towards personalized treatment in these patients, not only at admission but also along the treatment timeline. This model may represent a valuable tool for triaging and optimizing the resources in patients with COVID-19 infection worldwide.

DATA AVAILABILITY STATEMENT

The raw data used in this study are not publicly available due to the confidential policy of National Health Commission of China, as well as the General Data Protection Regulation (GDPR) of European Union (EU), but are available from the corresponding author Xiaohua Hou upon reasonable request.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, and University of Würzburg.

AUTHOR CONTRIBUTIONS

XH and KN designed and conceived the study. TB, DW, LZ, JS, XZ, YJ, DG, NI, KK, LR, and HE collected the data. KN, XZ, and PY developed the model. KN, XZ, PY, YZ, and HC analyzed the data. KN, TB, XZ, and PY wrote the manuscript. XH, KN, TB, XZ, PY, XZ, HC, QY, LR, and HE revised the manuscript. All authors have read and approved the final manuscript.

FUNDING

This work was partially supported by the National Science Foundation of China (grant numbers: 81573702, 81774008,

31871334, and 31671374), the National Key Research and Development Program of China (grant number: 2018YFC0910502), and Urgent projects of scientific and technological research on COVID-19 funded by Hubei Province (2020FCA014).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2021.672050/full#supplementary-material

REFERENCES

Ammar, M. A., Sacha, G. L., Welch, S. C., Bass, S. N., Kane-Gill, S. L., Duggal, A., et al. (2021). Sedation, Analgesia, and Paralysis in COVID-19 Patients in the Setting of Drug Shortages. *J. Intensive Care Med.* 36 (2), 157–174. doi:10.1177/0885066620951426

Azevedo, R. B., Botelho, B. G., Hollanda, J. V. G., Ferreira, L. V. L., Junqueira de Andrade, L. Z., Oei, S., et al. (2020). Covid-19 and the Cardiovascular System: a Comprehensive Review. *J. Hum. Hypertens.* doi:10.36660/ijcs.20200150

Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: Association for Computing Machinery, 785–794.

Dong, Y., Zhou, H., Li, M., Zhang, Z., Guo, W., Yu, T., et al. (2020). A Novel Simple Scoring Model for Predicting Severity of Patients With SARS-CoV-2 Infection. *Transboundary emerging Dis.* 67 (6), 2823–2829. doi:10.1111/tbed.13651

Epidemiology Working Group for Ncip Epidemic Response CCfDC, Prevention (2020). The Epidemiological Characteristics of an Outbreak of 2019 Novel Coronavirus Diseases (COVID-19) in China. *Zhonghua Liu Xing Bing Xue Za Zhi.* 41 (2), 145–151. doi:10.3760/cma.j.issn.0254-6450.2020.02.003

Frater, J. L., Zini, G., d'Onofrio, G., and Rogers, H. J. (2020). COVID-19 and the Clinical Hematology Laboratory. *Int. J. Lab. Hematol.* 42 (Suppl. 1), 11–18. doi:10.1111/ijlh.13229

Gomes, D. C. d. S., and Serra, G. L. d. O. (2021). Machine Learning Model for Computational Tracking and Forecasting the COVID-19 Dynamic Propagation. *IEEE J. Biomed. Health Inform.* 25 (3), 615–622. doi:10.1109/jbhi.2021.3052134

Gupta, S., Hayek, S. S., Wang, W., Chan, L., Mathews, K. S., Melamed, M. L., et al. (2020a). Factors Associated With Death in Critically Ill Patients With Coronavirus Disease 2019 in the US. *JAMA Intern. Med.* 180 (11), 1436–1447. doi:10.1001/jamainternmed.2020.3596

Gupta, R., Hussain, A., and Misra, A. (2020b). Diabetes and COVID-19: Evidence, Current Status and Unanswered Research Questions. *Eur. J. Clin. Nutr.* 74 (6), 864–870. doi:10.1038/s41430-020-0652-1

Ikemura, K., Bellin, E., Yagi, Y., Billett, H., Saada, M., Simone, K., et al. (2021). Using Automated Machine Learning to Predict the Mortality of Patients with COVID-19: Prediction Model Development Study. *J. Med. Internet Res.* 23 (2), e23458. doi:10.2196/23458

Itoh, H. (2020). A New normal for Hypertension Medicine with Coronavirus Disease-2019 (COVID-19): Proposal from the President of the Japanese Society of Hypertension. *Hypertens. Res.* 43 (9), 857–858. doi:10.1038/s41440-020-0497-y

Kishaba, T., Tamaki, H., Shimaoka, Y., Fukuyama, H., and Yamashiro, S. (2014). Staging of Acute Exacerbation in Patients With Idiopathic Pulmonary Fibrosis. *Lung.* 192 (1), 141–149. doi:10.1007/s00408-013-9530-0

Kriegeskorte, N., and Golan, T. (2019). Neural Network Models and Deep Learning. *Curr. Biol.* 29 (7), R231–r236. doi:10.1016/j.cub.2019.02.034

Li, X., Xu, S., Yu, M., Wang, K., Tao, Y., Zhou, Y., et al. (2020). Risk Factors for Severity and Mortality in Adult COVID-19 Inpatients in Wuhan. *J. Allergy Clin. Immunol.* 146 (1), 110–118. doi:10.1016/j.jaci.2020.04.006

Liang, W., Yao, J., Chen, A., Lv, Q., Zanin, M., Liu, J., et al. (2020). Early Triage of Critically Ill COVID-19 Patients Using Deep Learning. *Nat. Commun.* 11 (1), 3543. doi:10.1038/s41467-020-17280-8

Lin, J. (1991). Divergence Measures Based on the Shannon Entropy. *IEEE Trans. Inform. Theor.* 37 (1), 145–151. doi:10.1109/18.61115

Lippi, G., and Plebani, M. (2020). The Critical Role of Laboratory Medicine During Coronavirus Disease 2019 (COVID-19) and Other Viral Outbreaks. *Clin. Chem. Lab. Med.* 58 (7), 1063–1069. doi:10.1515/cclm-2020-0240

Mai, F., Del Pinto, R., and Ferri, C. (2020). COVID-19 and Cardiovascular Diseases. *J. Cardiol.* 76 (5), 453–458. doi:10.1016/j.jjcc.2020.07.013

National Health Commission of the People's Republic of China (2020). Diagnosis and Treatment of Pneumonia Infected by the New Novel Coronavirus (The Trial Fifth Edition) Medical Letter From the National Health Office.

Rasmussen, S., Sperling, P., Poulsen, M. S., Emmersen, J., and Andersen, S. (2020). Medical Students for Health-Care Staff Shortages during the COVID-19 Pandemic. *The Lancet.* 395 (10234), e79–e80. doi:10.1016/s0140-6736(20)30923-5

Risch, H. A. (2020). Early Outpatient Treatment of Symptomatic, High-Risk Covid-19 Patients that Should Be Ramped-Up Immediately as Key to the Pandemic Crisis. *Am. J. Epidemiol.* 189 (11), 1218–1226. doi:10.1093/aje/kwaa093

Sabrina, P., Pierre-Frederic, K., and Nicolas, V. (2012). CRP Pro-inflammatory Signalling in Atherosclerosis: Myth or Reality? *Curr. Signal Transduction Ther.* 7 (2), 142–148. doi:10.2174/157436212800376681

Salyer, S. J., Maeda, J., Sembuche, S., Kebede, Y., Tshangela, A., Moussif, M., et al. (2021). The First and Second Waves of the COVID-19 Pandemic in Africa: a Cross-Sectional Study. *The Lancet.* 397 (10281), 1265–1275. doi:10.1016/s0140-6736(21)00632-2

Sanwald, R., and Kirk, J. E. (1966). α-Hydroxybutyric Dehydrogenase Activity of Human Vascular Tissue. *Nature.* 209 (5026), 912–913. doi:10.1038/209912a0

Setti, L., Passarini, F., De Gennaro, G., Barbieri, P., Pallavicini, A., Ruscio, M., et al. (2020). Searching for SARS-COV-2 on Particulate Matter: A Possible Early Indicator of COVID-19 Epidemic Recurrence. *Int. J. Environ. Res. Public Health.* 17 (9), 2986. doi:10.3390/ijerph17092986

Shenhav, L., Thompson, M., Joseph, T. A., Briscoe, L., Furman, O., Bogumil, D., et al. (2019). FEAST: Fast Expectation-Maximization for Microbial Source Tracking. *Nat. Methods.* 16 (7), 627–632. doi:10.1038/s41592-019-0431-x

Sirleaf, E. J., and Clark, H. (2021). Report of the Independent Panel for Pandemic Preparedness and Response: Making COVID-19 the Last Pandemic. *The Lancet.* 398 (10295), 101–103. doi:10.1016/s0140-6736(21)01095-3

Trowell, O. A. (1947). Function of the Lymphocyte. *Nature.* 160 (4076), 845–846. doi:10.1038/160845a0

Vermeire, S., Van Assche, G., and Rutgeerts, P. (2004). C-Reactive Protein as a Marker for Inflammatory Bowel Disease. *Inflamm. Bowel Dis.* 10 (5), 661–665. doi:10.1097/00054725-200409000-00026

Wahlster, S., Sharma, M., Lewis, A. K., Patel, P. V., Hartog, C. S., Jannotta, G., et al. (2021). The Coronavirus Disease 2019 Pandemic's Effect on Critical Care Resources and Health-Care Providers. *Chest.* 159 (2), 619–633. doi:10.1016/j.chest.2020.09.070

Wang, H., Wang, L., Lee, E. H., Zheng, J., Zhang, W., Halabi, S., et al. (2021). Decoding COVID-19 Pneumonia: Comparison of Deep Learning and Radiomics CT Image Signatures. *Eur. J. Nucl. Med. Mol. Imaging.* 48 (5), 1478–1486. doi:10.1007/s00259-020-05075-4

Watson, S. I., and Lilford, R. J. (2021). Global COVID-19 Vaccine Roll-Out: Time to Randomise Vaccine Allocation?. *The Lancet.* 397 (10287), 1804–1805. doi:10.1016/s0140-6736(21)00895-3

Williamson, E. J., Walker, A. J., Bhaskaran, K., Bacon, S., Bates, C., Morton, C. E., et al. (2020). Factors Associated with COVID-19-Related Death Using OpenSAFELY. *Nature.* 584 (7821), 430–436. doi:10.1038/s41586-020-2521-4

Wu, R., Ai, S., Cai, J., Zhang, S., Qian, Z., Zhang, Y., et al. (2020). Predictive Model and Risk Factors for Case Fatality of COVID-19: A Cohort of 21,392 Cases in Hubei, China. *The Innovation.* 1 (2), 100022. doi:10.1016/j.xinn.2020.100022

Xiao, L.-s., Zhang, W.-F., Gong, M.-c., Zhang, Y.-p., Chen, L.-y., Zhu, H.-b., et al. (2020). Development and Validation of the HNC-LL Score for Predicting the Severity of Coronavirus Disease 2019. *EBioMedicine.* 57, 102880. doi:10.1016/j.ebiom.2020.102880

Xie, X., Shi, Q., Wu, P., Zhang, X., Kambara, H., Su, J., et al. (2020). Single-Cell Transcriptome Profiling Reveals Neutrophil Heterogeneity in Homeostasis and Infection. *Nat. Immunol.* 21 (9), 1119–1133. doi:10.1038/s41590-020-0736-z

Xu, X.-W., Wu, X.-X., Jiang, X.-G., Xu, K.-J., Ying, L.-J., Ma, C.-L., et al. (2020). Clinical Findings in a Group of Patients Infected With the 2019 Novel Coronavirus (SARS-Cov-2) Outside of Wuhan, China: Retrospective Case Series. *Bmj.* 368, m606. doi:10.1136/bmj.m606

Yan, L., Zhang, H.-T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., et al. (2020). An Interpretable Mortality Prediction Model for COVID-19 Patients. *Nat. Mach Intell.* 2 (5), 283–288. doi:10.1038/s42256-020-0180-7

Zhou, X., Bai, T., Meckel, K., Song, J., Jin, Y., Kortüm, K. M., et al. (2020a). COVID-19 Infection in Patients with Multiple Myeloma: a German-Chinese Experience from Würzburg and Wuhan. *Ann. Hematol.* 100 (3), 843–846. doi:10.1007/s00277-020-04184-2

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., et al. (2020b). Clinical Course and Risk Factors for Mortality of Adult Inpatients with COVID-19 in Wuhan, China: a Retrospective Cohort Study. *The Lancet.* 395 (10229), 1054–1062. doi:10.1016/s0140-6736(20)30566-3

Zhu, W., Xie, L., Han, J., and Guo, X. (2020). The Application of Deep Learning in Cancer Prognosis Prediction. *Cancers.* 12 (3). doi:10.3390/cancers12030603

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 13 September 2021
doi: 10.3389/frai.2021.684609

Insights Into Co-Morbidity and Other Risk Factors Related to COVID-19 Within Ontario, Canada

Brett Snider, Bhumi Patel and Edward McBean*

University of Guelph, Guelph, ON, Canada

The worldwide rapid spread of the severe acute respiratory syndrome coronavirus 2 has affected millions of individuals and caused unprecedented medical challenges by putting healthcare services under high pressure. Given the global increase in number of cases and mortalities due to the current COVID-19 pandemic, it is critical to identify predictive features that assist identification of individuals most at-risk of COVID-19 mortality and thus, enable planning for effective usage of medical resources. The impact of individual variables in an XGBoost artificial intelligence model, applied to a dataset containing 57,390 individual COVID-19 cases and 2,822 patient deaths in Ontario, is explored with the use of SHapley Additive exPlanations values. The most important variables were found to be: age, date of the positive test, sex, income, dementia plus many more that were considered. The utility of SHapley Additive exPlanations dependency graphs is used to provide greater interpretation of the black-box XGBoost mortality prediction model, allowing focus on the non-linear relationships to improve insights. A "Test-date Dependency" plot indicates mortality risk dropped substantially over time, as likely a result of the improved treatment being developed within the medical system. As well, the findings indicate that people of lower income and people from more ethnically diverse communities, face an increased mortality risk due to COVID-19 within Ontario. These findings will help guide clinical decision-making for patients with COVID-19.

Keywords: artificial intelligence, COVID-19, SHAP (shapley additive explanation), XGBoost (extreme gradient boosting), mortality, co-morbidity

OPEN ACCESS

Edited by:
Da Yan,
University of Alabama at Birmingham, United States

Reviewed by:
Jingyi Zheng,
Auburn University, United States
Ryan Lee Melvin,
University of Alabama at Birmingham, United States
Ramaraju Rudraraju,
University of Alabama at Birmingham, United States

****Correspondence:***
Brett Snider
bsnide01@uoguelph.ca

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *23 March 2021*
Accepted: *26 May 2021*
Published: *13 September 2021*

Citation:
Snider B, Patel B and McBean E (2021) Insights Into Co-Morbidity and Other Risk Factors Related to COVID-19 Within Ontario, Canada. Front. Artif. Intell. 4:684609. doi: 10.3389/frai.2021.684609

INTRODUCTION

With issues of the second wave of the COVID-19 pandemic ongoing in 2021 and the world in a continuing crisis, interest continues to escalate to improve the understanding of features resulting in virus caseload increases. In response, of particular interest are opportunities to improve modeling prediction capabilities which can provide more accurate information as it becomes available from the first and second waves of COVID-19. In this regard, until recently, data security and privacy issues have limited accessibility to alternate and detailed data sources, but opportunities are opening up and showing real potential. As an example, improved access to Ministry of Health for Ontario, enabled Snider et al. (2021) to develop powerful artificial intelligence (AI) models that are now able to predict mortality and recovery of COVID-19 patients with a high degree of accuracy; the models developed were based on data from Ontario Health Data Platform (February 22, 2020–October 20, 2020), utilizing extensive and detailed data for 57,390 individual COVID-19 cases.

AI models in general, and Snider et al. (2021) in particular, provide dimensions including the ability to uncover and understand the value of an array of "base" information, including co-morbidity data, that influence mortality rates including at the case-by-case level. Findings on the risks of mortality for individual patients have the potential to influence many important actions such as helping identify "most at-risk populations" thus providing insights on hospitalizations/medical strategies and opportunities to aid delivery of COVID-19 vaccination priority strategies in the future.

The findings and predictions made available from use of logistic regression and other AI models, have excellent potential, when caseload data are available. Specifically, the models of Snider et al. (2021) demonstrated excellent discrimination with all model's area under the curve (AUC) exceeding 0.948, with the greatest being 0.956 for an XGBoost (Extreme Gradient Boosting) model. Hence, this paper advances the knowledge in mortality risk of COVID-19 patients in Ontario, Canada, by calculating and exploring SHapley Additive exPlanations (SHAP) values of parameters used for the XGBoost AI model developed by Snider et al. (2021). Most importantly, these models provide specifics on the causative/impactful inter-relationships, which allow extraction of additional information from datasets and exceed the information provided by logistics models since logistic models assume a specific type of relationship between input and output, whereas the machine learning models allow capture of a more flexible relationship. In order to see the exact form of the relationship, SHAP dependance plots were made and analyzed for 4 principal features driving the XGBoost mortality prediction model. Hence, these provide indications detailing the importance of the individual variables that can be used to characterize co-morbidities that can be important indicators as to whom may be most susceptible to mortality and more likely to be in need of intensive medical needs, arising from the COVID-19 virus. Also, the relationships identified using this approach between parameters such as co-morbidities and other risk factors associated with COVID-19, and the corresponding impact on the mortality prediction XGBoost model, provide information which can be of great value in designing effective non-pharmaceutical interventions (NPIs) and vaccination schedules.

REVIEW OF TECHNICAL LITERATURE

The impressive predictive capabilities of AI have resulted in AI models being adopted across a wide range of disciplines. Their excellent performance in some areas of investigation arises largely due to the ability of AI models to identify and to model complex patterns between input variables and the predicted output. However, the AI model's complexity often makes it difficult to identify the relationships between the input variables and the output, resulting in most advanced AI models being classified as "black-boxes".

These so-called black-box models can be very accurate in their predictions but leave the users wondering how individual factors contribute to the model's final prediction. A number of dynamic and statistical models of COVID-19 outbreaks including SEIR models (which assign individuals to the susceptible (S), exposed (E), infected (I), and recovered (R) classes) have previously been used to study and analyze transmission (Hellewell et al., 2020; Tuite and Fisman, 2020; Kucharski at al., 2020). However, these epidemiological models require values for unknown parameters and rely on many assumptions (Hu at al., 2020). Interest in understanding how the individual factors contribute has resulted in a variety of interpretable machine learning techniques being developed in recent years to assist in the interpretation of the impact of specific input variables on the final prediction (Molnar, 2019). This information is critical in promoting the gaining of trust in the AI model, as well as providing insights into which variables are important, and identifying key relationships that influence the AI models' final prediction.

AI models have played a major role during the COVID-19 pandemic, through COVID-19 case identification, predicting transmission scenarios, and identifying the mortality risks of specific COVID-19 patients (see e.g., Li et al., 2020a; Boulle et al., 2020; Li et al., 2020b; Dhamodharavadhani et al., 2020). A significant focus has been placed on ensuring these models are interpretable, to allow a better understanding of the factors contributing to the predictions of patients' outcomes, and to help inform responses.

Some researchers have selected AI models that are interpretable by design, such as logistic regression and decision trees. Yan et al. (2020) used decision trees and blood samples to interpret and identify mortality prediction for COVID-19 patients using blood samples. Fisman et al. (2020) used logistic regression models to predict mortality risk of COVID-19 patients; their logistic regression model quantifies the weight of each input variable to the final prediction, making it straightforward to determine how the model is calculating the overall COVID-19 mortality risk. A similar study by Quiroz et al. (2021) developed a logistic regression model using clinical and imaging data from two hospitals in Hubei, China, for automated severity assessment of COVID-19 for individual patients, obtaining an AUC of 0.950 using a combination of clinical and imaging features. They interpreted the importance of features using SHAP values and found patients in severe conditions had co-morbidities which included cardiovascular disease, diabetes, hypertension and cancer which is similar to findings obtained from previous studies (see e.g., Petrilli et al., 2020; Richardson et al., 2020; Shi et al., 2020; Siordia, 2020). Thus, interpretable machine learning techniques help address the most significant limitation of machine learning i.e., the lack of transparency due to its' black box nature, however, there are trade-offs between the accuracy of predictions and interpretability with such models (Du et al., 2020). Overall, interpretable AI algorithms such as logistic regression and decision trees allow for the user to identify the weights associated with the model's input variables, but these approaches are often less accurate compared to black-box models (see e.g., Murdoch et al., 2019; Snider et al., 2021).

TABLE 1 | Characteristics of 57,390 Ontario residents with COVID-19.

Variable	Description	Range of values
Age	Age in years, as of Jan 1, 2020	0–105
Test date	Test date	Feb–Oct 2020
Sex	Indicator variable for sex	26,861 (M = 1, F = 0)
Hypertension	Chronic hypertension, as of Jan 1, 2020	15,778 (0,1)
LTC resident	LTC resident, as of Jan 1, 2020	5,179 (0,1)
Chronic_dementia	Chronic dementia diagnosed, as of Jan 1, 2020	4,746 (0,1)
Chronic_odd	Chronic diabetes diagnosed as of Jan 1, 2020	9,002 (0,1)
Ethnic concentration quint.	Calculated from Ontario marginalization index, based on census designation. Refers to visible minorities and/or recent immigrants	(0–5)[a]
Commuter concentration quint	% Of people that commute within census designated area - converted to quintiles	(0–5)[a]
Median income quint.	Median income within census-designated area - converted to quintiles	(0–5)[a]
Charl	Charlson co-morbidity index. Only 2,059 patients with charl above 0.	(0–10)
Household size quint.	Avg. Household size within census-designated area - converted to quintiles (5 being the highest, 0 = missing DA info).	(0–5)
CKD	Chronic kidney disease.	2,523 (0,1)
Cancer	Cancer index	2,995 (0–1)
Chronic_copd	Chronic obstructive pulmonary disease	4,030 (0–1)
Chronic_asthma	Asthma	9,100 (0–1)
Chronic_chf	Congestive heart failure	2,257 (0–1)
Stroke	If patient suffered a stroke previous to Jan 1, 2020	1,016 (0–1)
Cardiac ISCH	Cardiac ischemia	1,916 (0–1)
Rural	Indicator if a patient lives in a rural residence	1,746 (0–1)
Chronic_ra	Rheumatoid arthritis	567 (0–1)
Tia	Transient Ischemic Attack	722 (0–1)
immuno_comp	Immuno-compromised	237 (0–1)
Thala	History of Thalassemia	36 (0–1)
Cases recovered		54,568
Cases died		2,822

[a](0 referring to missing information).

For a critical discussion in a clinical context, see the work by Christodoulou et al. (2019).

Another technique is to apply model agnostic interpretation methods to black-box models to investigate the relationship between inputs and the model's prediction. A leading agnostic method to interpret black box AI models is through the use of SHAP values (Molnar, 2019). Barda et al. (2020) explored their black-box mortality prediction model for Israel's COVID-19 patients using SHAP values to estimate the contribution of individual features to the overall model predictions. The calculated SHAP values identified the importance of several demographic attributes that the model determined important in predicting COVID-19 mortality (for example, age and cardiovascular disease) but the model used by Barda et al. (2020) has limited individual-level data, making it difficult to explore key relationships between COVID-19 patients and mortality, such as income level and ethnicity.

MATERIALS AND METHODS

The following sections describe the datasets and models developed by Snider et al. (2021) to predict mortality risk of COVID-19 patients in Ontario, Canada. The SHAP value methodology and application used to explore the black-box prediction models are then outlined.

Dataset Description

The Ontario Health Data Platform (OHDP) was used in this research to assemble extensive data regarding COVID-19 patients within Ontario. The OHDP dataset contains epidemiological and demographic information, recovery/mortality outcome information and co-morbidities of individuals residing in Ontario. The attributes which proved most useful by the AI models are listed in **Table 1**. Co-morbidities and age were collected from patient health records as of January 1, 2020; hence, diagnosis of additional medical conditions after this date were excluded. Of the 57,390 cases included in the dataset, 2,822 patients died of COVID-19 and the remaining 54,568 either recovered from COVID-19 or remained hospitalized as of January 1, 2021. Several input variables were derived using 2016 Canadian census data for the designated area of the individual patients. Canadian census location information is based on a size of approximately three blocks and hence is able to capture representation of ethnicity, income level and other social differences, and can therefore be considered robust. The census data includes: ethnic concentration (of residential area), commuter concentration, median income and household size (these values are unlikely to change significantly between date of census and start of pandemic). These values were converted into quintiles (division of the population into five equal-sized groups according to the distribution of input variables) with 1 being the lowest quintile, and 5 being the highest. Individuals with missing data were not included in these analyses. It is noted

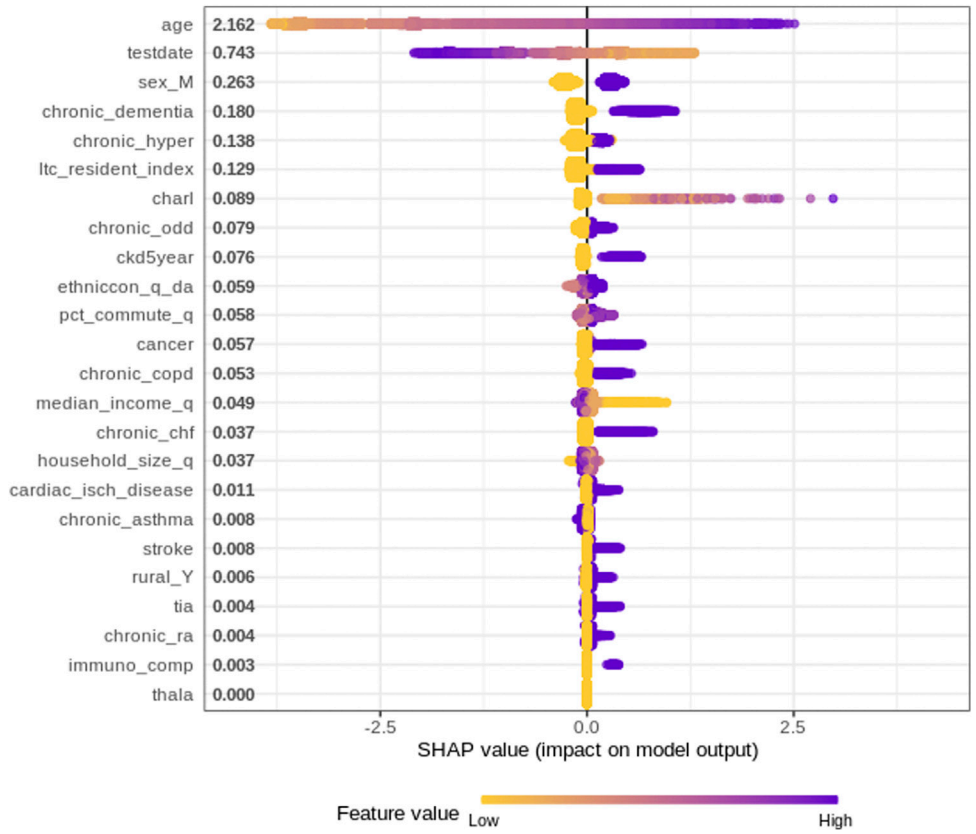


FIGURE 1 | SHAP summary plot for XGBoost model.

that long-term care (LTC) residents in Ontario did not include census-designated area information and therefore, data for the LTC residents were represented with a zero value.

Model Development

Snider et al. (2021) compared three black-box machine learning models which were 1) Artificial Neural Network (Venables and Ripley, 2002), 2) Random Forest (Wiens and Shenoy, 2018), 3) Extreme gradient boosting decision tree—XGBoost (Chen et al., 2021) and one interpretable machine learning model which was logistic regression (Venables and Ripley, 2002). These models were adopted because of their high accuracy in binary classification problems as well as their prevalence/adoption in previous literature. Prior to model calibration, the dataset was randomly split into two segments, namely an 80% training dataset and a 20% testing dataset where each model was calibrated using the training dataset and assessed for accuracy using the testing dataset. A grid search approach was used to adjust the hyper-parameters of the models using a 10-fold cross-validation technique repeated three times per model and optimized to produce the maximum area under the receiver operating characteristic curve (Area Under Curve, or AUC). The XGBoost model was determined to be the most accurate model, having an AUC of 0.956. Therefore, this paper explores the XGBoost model's relationships between the input variables and the predicted mortality risk by calculating SHAP values for each attribute and patient included in the training dataset. Features such as the public health unit of individual cases from the same locality/region were excluded when training the model as such parameters could cause problems if a particular region has a higher number of patients compared to others.

Shapley Additive Explanation Values

To explore the impact of each variable on the XGBoost's mortality model prediction, SHAP values have been used. SHAP values determine the importance of a feature by comparing what a model predicts with and without the feature for each observation within the training data. Specifically, the SHAP values represent the final AI model's prediction using the following equation:

$$y_i = y_{base} + f(x_{i1}) + f(x_{i2}) + \ldots + f(x_{iN})$$

The i^{th} sample (or patient) is defined as x_i, the N represents the final feature (or input parameter) for the i^{th} sample (as defined by x_{iN}). The predicted value of the AI model is y_i and the reference value, or mean value of the target sample variable, is defined as y_{base}. The function $f(x_{ij})$ is the calculated SHAP value of x_{ij}. The SHAP values are calculated using SHAP for XGBoost R package (Liu and Just, 2020) and present the variable contribution on a

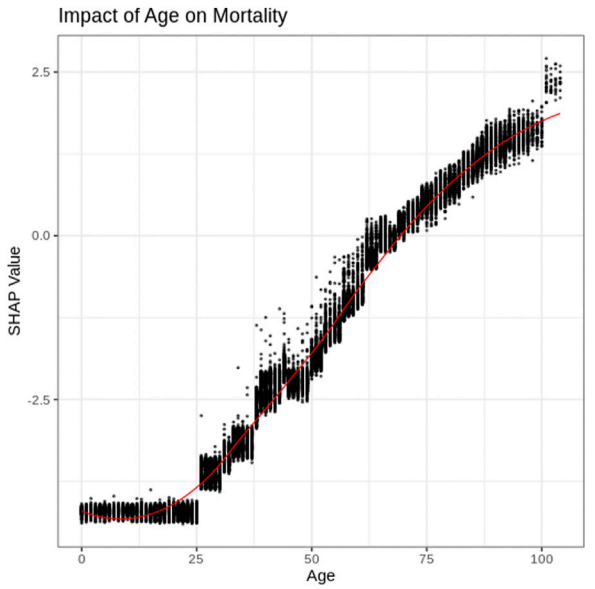


FIGURE 2 | SHAP plot for Age.

log-odds scale (logarithm of the ratio of high mortality risk to low mortality risk).

RESULTS AND DISCUSSION

Figure 1 plots the SHAP value for each individual patient within the training dataset for each input variable. The input variables, as listed on the y-axis, are ranked from most important (top) to least important (bottom) with their mean absolute SHAP value indicated next to the name in **Figure 1**. The X axis represents the SHAP value associated with each variable and patient within the training dataset (i.e., there is a plotted point for each case based on the influence that the variable has on the prediction of that case). The color indicates whether the individual patients' input variable value was high (purple) or low (yellow). For example, in **Figure 1** a "high" age has a high and "positive" impact on predicting mortality. The "high" comes from the purple color and the "positive" impact is shown along the X axis. Note, a range of SHAP values exists for each input variable value based on the SHAP values calculated for each observation, and how they independently contribute to the machine learning model's predictions.

Overall, age is unquestionably the most important variable for the XGBoost model. As a patient's age increases (approaches purple), the SHAP value impact increases, with a very high age being associated with an additional 2.5 increase in log-odds. The test-date when someone tested positive for COVID-19 also has a strong impact on overall mortality risk; as the positive test date increases (i.e., later on during the pandemic), the risk of mortality decreases.

The impact of the SHAP values are easily identified for binary variables, such as sex, hypertension, whether or not a patient was an LTC resident, and dementia. Being a Male (i.e., Sex = 1) has an additional 0.25 increase in log-odds, which indicates males have an increased risk of mortality. Similar increases are also identified with people having hypertension. An "LTC residence" designation is also associated with a significant increase in mortality, which is consistent with reported large numbers of outbreaks and deaths of individuals living in LTC homes. Chronic dementia is the co-morbidity associated with the largest increase in mortality.

Age

The impact of a patient's age on the AI model's mortality risk prediction can be further explored using a SHAP dependency graph. **Figure 2** depicts the SHAP values associated with patient ages within the training dataset. As further explanation of the results, a patient of <20 years of age is associated with a significant decrease in mortality risk; alternatively, as age increases, the risk of mortality increases. The non-linear shape of this figure, as well as the range of values for similar age highlight some of the advantages of the more complex AI models compared to less complex models such as logistic regression. Specifically, the XGBoost model examined here is able to identify complex patterns, as well as interaction effects that are often difficult for regression models (for example, logistic regression) to identify.

Test Date

Figure 3 depicts the SHAP values associated with the "day the patient tested positive" for COVID-19. **Figure 3** indicates that residents of Ontario who tested positive for COVID-19 early on in the pandemic (e.g., April 2020) had an increased risk of mortality. From the data, mortality risk decreased for those individuals with later positive-test dates with a substantial decrease in mortality being associated with more recent

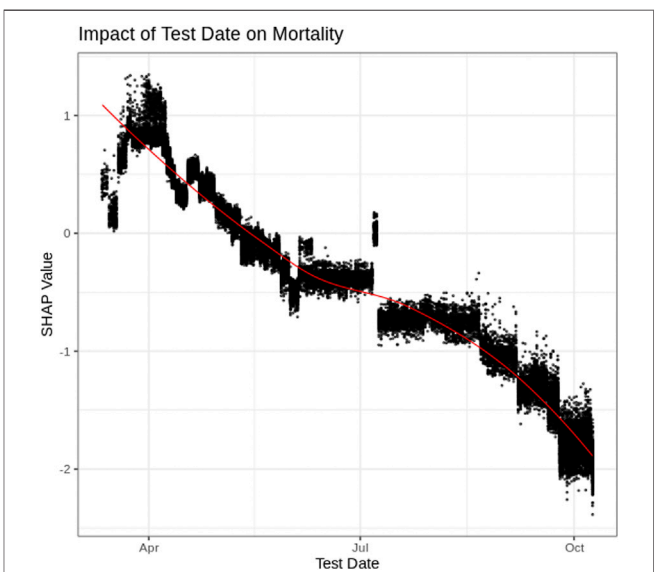


FIGURE 3 | SHAP plot for Test Date.

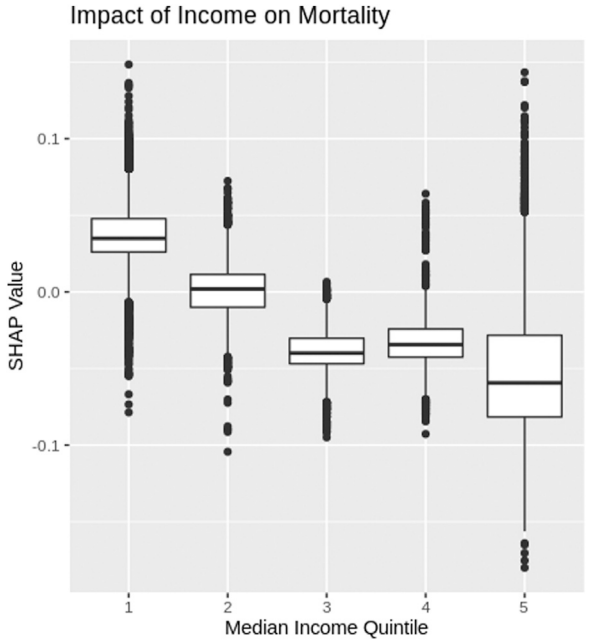


FIGURE 4 | SHAP box-plot for median income.

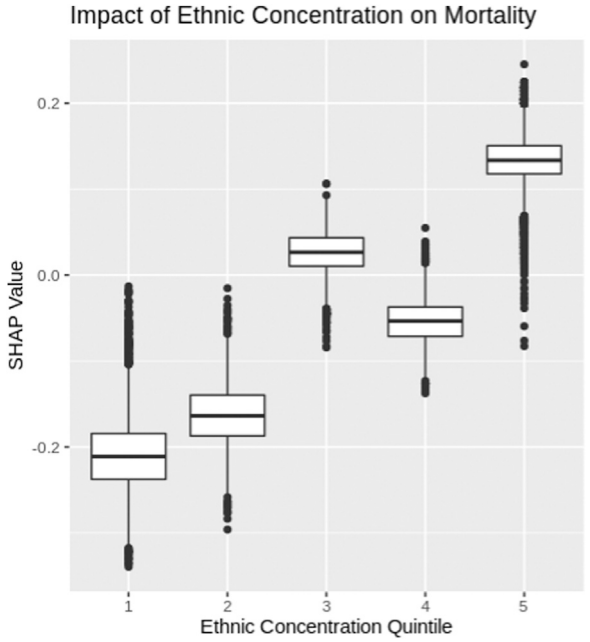


FIGURE 5 | SHAP box-plot for ethnic concentration.

months (e.g., September and October of 2020). Comparing positive test rates (% of tests performed that were positive) over the same time period identifies that "positivity rates increased during the period of substantial decrease in mortality" risk (October–December) (Public Health Ontario, 2021). This indicates that the decrease in mortality is unlikely a result of less severe cases being identified since positivity rates increased in October, while the associated risk decreased. Therefore, the decrease in mortality associated with later test date is considered more likely associated with improved treatment within the medical system (Robinson, 2021).

Income

The SHAP values for each median income quintile, based on census designated area, are depicted as box plots in **Figure 4** (note income data were not available for LTC residents and therefore, LTC resident data were not included in **Figure 4**). COVID-19 patients who come from a census area with the lowest median income quintile have a higher risk of mortality. As the median income increases, **Figure 4** shows the risk of COVID-19 mortality decreases.

Ethnicity

Ethnic concentrations were calculated based on 2016 census data for each designated area using the methodology outlined in the Ontario Marginalization index (Public Health Ontario, 2018). Specifically, ethnic concentration refers to the proportion of the population within a designated area who are recent immigrants or belong to a visible minority. The ethnic concentration was then segmented into quintiles and the SHAP values for each quintile are depicted using box plots in **Figure 5**. COVID-19 patients from census areas with high ethnic concentrations experience higher levels of mortality risk, while patients from neighborhoods with low ethnic concentrations experience lower levels of mortality risk. The ethnic and income factor results further highlight that COVID-19 has a greater impact among marginalized communities within Ontario, Canada.

CONCLUSION

This paper explored an advanced AI mortality prediction model for COVID-19 patients within Ontario, Canada. Specifically, SHAP values were calculated and examined in order to uncover the relationships identified by the XGBoost model used by Snider et al. (2021). Several key findings are identified through this research. First, by examining the average SHAP value for each variable, key attributes related to mortality risk are identified (**Figure 1**). Age and test date are determined to be the leading factors that influence the mortality risk of COVID-19 patients in Ontario but also identified as important were sex, dementia, ethnicity, etc. at lesser degrees of importance.

SHAP dependency graphs are shown to provide very useful interpretation of the black-box XGBoost mortality prediction model. This paper explores four key attributes using SHAP dependency graphs: patients' age, test-date, income and ethnic concentration. The Age SHAP dependency plot highlights the non-linear relationship between the patients age and risk of COVID-19 mortality, highlighting the significant increase in mortality risk associated with older patients with COVID-19 in Ontario. The Test-date dependency plot indicates mortality risk has dropped substantially within Ontario since the start of

the pandemic. The SHAP values for income and ethnic quintiles suggests people of lower income and higher ethnic concentrations face an increased mortality risk due to COVID-19 within Ontario. Further exploration into these trends will be important as vaccinations become more widespread around the world, and with variants of concern becoming more common.

Overall, AI models have and will continue to play a major role in understanding and combating the COVID-19 pandemic. However, to build trust in these models and to gain further insight, a strong emphasis must be placed on ensuring the results from these models are interpretable. SHAP values are shown to be a useful tool to "open up" some of the more complex black box AI models and uncover the key patterns being modeled. The findings gathered from the model exploration performed in this paper further adds to the literature regarding mortality risks associated with COVID-19 patients and will help guide strategic interventions and vaccination schedules.

DATA AVAILABILITY STATEMENT

These datasets were linked using unique encoded identifiers and analyzed at ICES. The use of the data in this project is authorized under section 45 of Ontario's Personal Health Information Protection Act (PHIPA) and does not require review by a Research Ethics Board. Access to datasets: The dataset from this study is held securely in coded form at ICES. While legal data sharing agreements between ICES and data providers (e.g., healthcare organizations and government) prohibit ICES from making the dataset publicly available, access may be granted to those who meet pre-specified criteria for confidential access, available at www.ices.on.ca/DAS (email: das@ices.on.ca). The full dataset creation plan and underlying analytic code are available from the authors upon request, understanding that the computer programs may rely upon coding templates or macros that are unique to ICES and are therefore either inaccessible or may require modification.

FUNDING

This work was supported by Natural Sciences and Engineering Research Council of Canada Alliance Special COVID-19 program (grant number 401636) and University of Guelph Research Leadership Chair Professor funding.

ACKNOWLEDGMENTS

This study was supported by ICES, which is funded by an annual grant from the Ontario Ministry of Health (MOH) and the Ministry of Long-Term Care (MLTC). This study was supported by the Ontario Health Data Platform (OHDP), a Province of Ontario initiative to support Ontario's ongoing response to COVID-19 and its related impacts. We thank IQVIA Solutions Canada Inc. for use of their Drug Information Database. Ontario Community Health Profiles Partnership (OCHPP) created Ontario Marginalization Index (ON-Marg) which is a source for this paper as ON-Marg is used to understand inequalities in health and other social problems related to health among either population groups or geographic areas across Ontario. Postal Code Conversion File and census data were adapted from Statistics Canada. This does not constitute an endorsement by Statistics Canada of this product.

REFERENCES

Barda, N., Riesel, D., Akriv, A., Levy, J., Finkel, U., Yona, G., et al. (2020). Developing a COVID-19 Mortality Risk Prediction Model when Individual-Level Data Are Not Available. *Nat. Commun.* 11, 4439. doi:10.1038/s41467-020-18297-9

Boulle, A., Davies, M.-A., Hussey, H., Ismail, M., Morden, E., Vundle, Z., et al. (2020). Risk Factors for COVID-19 Death in a Population Cohort Study from the Western Cape Province, South Africa. *Clin. Infect. Dis.* doi:10.1093/cid/ciaa1198

Chen, T., He, T., Benesty, M., Khotilovi, V., Tang, Y., Cho, H., et al. (2021). Xgboost: Extreme Gradient Boosting Version 1.0.0.2. The Comprehensive R Archive Network CRAN. Available at: https://CRAN.R-project.org/package=xgboost (Accessed January 25, 2021).

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., and Van Calster, B. (2019). A Systematic Review Shows No Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models. *J. Clin. Epidemiol.* 110, 12–22. doi:10.1016/j.jclinepi.2019.02.004

Dhamodharavadhani, S., Rathipriya, R., and Chatterjee, J. M. (2020). COVID-19 Mortality Rate Prediction for India Using Statistical Neural Network Models. *Front. Public Health* 8, 441. doi:10.3389/fpubh.2020.00441

Du, M., Liu, N., and Hu, X. (2019). Techniques for Interpretable Machine Learning. *Commun. ACM* 63 (1), 68–77. doi:10.1145/3359786

Fisman, D. N., Greer, A. L., Hillmer, M., and Tuite, R. (2020). Derivation and Validation of Clinical Prediction Rules for COVID-19 Mortality in Ontario, Canada. *Open Forum Infect. Dis.* 7, 11. doi:10.1093/ofid/ofaa463

Hellewell, J., Abbott, S., Gimma, A., Bosse, N. I., Jarvis, C. I., Russell, T. W., et al. (2020). Feasibility of Controlling COVID-19 Outbreaks by Isolation of Cases and Contacts. *Lancet Glob. Health* 8, e488–e496. doi:10.1016/S2214-109X(20)30074-7

Hu, Z., Ge, Q., Li, S., Boerwinkle, E., Jin, L., and Xiong, M. (2020). Forecasting and Evaluating Multiple Interventions for COVID-19 Worldwide. *Front. Artif. Intell.* 3, 41. doi:10.3389/frai.2020.00041

Kucharski, A., Russell, T., Diamond, C., Liu, Y., Edmunds, J., et al.CMMID nCo V working group (2020). Analysis and Projections of Transmission Dynamics of nCoV in Wuhan. Available online at: https://cmmid.github.io/ncov/wuhan_early_dynamics/index.html (accessed April 19,2021).

Li, H., Wang, S., Zhong, F., Bao, W., Li, Y., Liu, L., et al. (2020a). Age-Dependent Risks of Incidence and Mortality of COVID-19 in Hubei Province and Other Parts of China. *Front. Med.* 7, 190. doi:10.3389/fmed.2020.00190

Li, X., Xu, S., Yu, M., Wang, K., Tao, Y., Zhou, Y., et al. (2020b). Risk Factors for Severity and Mortality in Adult COVID-19 Inpatients in Wuhan. *J. Allergy Clin. Immunol.* 146 (1), 110–118. doi:10.1016/j.jaci.2020.04.006

Liu, Y., and Just, A. (2020). SHAPforxgboost: SHAP Plots for 'XGBoost'. R Package Version 0.1.0. (Github). Available at: https://github.com/liuyanguu/SHAPforxgboost/(Accessed January 26, 2021).

Molnar, C. (2019). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. (Github). Available at: https://christophm.github.io/interpretable-ml-book/(Accessed August 30, 2020).

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, Methods, and Applications in Interpretable Machine Learning. *Proc. Natl. Acad. Sci. USA* 116 (44), 22071–22080. doi:10.1073/pnas.1900654116

Petrilli, C. M., Jones, S. A., Yang, J., Rajagopalan, H., O'Donnell, L., Chernyak, Y., et al. (2020). Factors Associated with Hospital Admission and Critical Illness Among 5279 People with Coronavirus Disease 2019 in New York City: Prospective Cohort Study. *BMJ* 22, m1966. doi:10.1136/bmj.m1966

Public Health Ontario (2021). Ontario COVID-19 Data Tool. Available at: https://www.publichealthontario.ca/en/data-and-analysis/infectious-disease/covid-19-data-surveillance/covid-19-data-tool?tab=labTests (Accessed March 22 2021).

Public Health Ontario (2018). Ontario Marginalization Index. Available at: https://www.publichealthontario.ca/-/media/documents/O/2017/on-marg-userguide.pdf?la=en.

Quiroz, J. C., Feng, Y.-Z., Cheng, Z.-Y., Rezazadegan, D., Chen, P.-K., Lin, Q.-T., et al. (2021). Development and Validation of a Machine Learning Approach for Automated Severity Assessment of COVID-19 Based on Clinical and Imaging Data: Retrospective Study. *JMIR Med. Inform.* 9, e24572. doi:10.2196/24572

Richardson, S., Hirsch, J. S., Narasimhan, M., Crawford, J. M., McGinn, T., Davidson, K. W., et al. (2020). Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. *JAMA* 323 (20), 2052–2059. doi:10.1001/jama.2020.6775

Robinson, J. (2021). Everything You Need to Know about the COVID-19 Therapy Trials. Available at: https://pharmaceutical-journal.com/article/feature/everything-you-need-to-know-about-the-covid-19-therapy-trials Accessed March 22 2021.

Shi, H., Han, X., Jiang, N., Cao, Y., Alwalid, O., Gu, J., et al. (2020). Radiological Findings from 81 Patients with COVID-19 Pneumonia in Wuhan, China: a Descriptive Study. *Lancet Infect. Dis.* 20 (4), 425–434. doi:10.1016/S1473-3099(20)30086-4

Siordia, J. A. (2020). Epidemiology and Clinical Features of COVID-19: A Review of Current Literature. *J. Clin. Virol.* 127, 104357. doi:10.1016/j.jcv.2020.104357

Snider, B., McBean, E., Yawney, J., Gadsden, S., and Patel, B. (2021). Identification of Variable Importance for Predictions of Mortality from COVID-19 Using AI Models for Ontario, Canada. *Front. Public Health*.

Tuite, A. R., and Fisman, D. N. (2020). Reporting, Epidemic Growth, and Reproduction Numbers for the 2019 Novel Coronavirus (2019-nCoV) Epidemic. *Ann. Intern. Med.* 172, 567–568. doi:10.7326/M20-0358

Venables, W., and Ripley, B. (2002). *Modern Applied Statistics with S*. Fourth Edition. New York: Springer. doi:10.1007/978-0-387-21706-2 ISBN 0-387-95457-0).

Wiens, J., and Shenoy, E. S. (2018). Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. *Clin. Infect. Dis.* 66 (1), 149–153. doi:10.1093/cid/cix731

Yan, L., Zhang, H.-T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., et al. (2020). An Interpretable Mortality Prediction Model for COVID-19 Patients. *Nat. Mach Intell.* 2, 283–288. doi:10.1038/s42256-020-0180-7

Author Disclaimer: The analyses, opinions, results, and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by ICES, the OHDP, its partners, or the Province of Ontario is intended or should be inferred. Parts of this material are based on data and/or information compiled and provided by CIHI. However, the analyses, conclusions, opinions, and statements expressed in the material are those of the author(s), and not necessarily those of CIHI. Parts of this material are based on data and information provided by Cancer Care Ontario (CCO). The opinions, results, view, and conclusions reported in this paper are those of the authors and do not necessarily reflect those of CCO. No endorsement by CCO is intended or should be inferred.

frontiers in Artificial Intelligence

CORRECTION
published: 13 September 2021
doi: 10.3389/frai.2021.759022

Corrigendum: Insights Into Co-Morbidity and Other Risk Factors Related to COVID-19 Within Ontario, Canada

Brett Snider, Bhumi Patel and Edward McBean*

School of Engineering, University of Guelph, Guelph, ON, Canada

Keywords: artificial intelligence, COVID-19, SHAP (shapley additive explanation), XGBoost (extreme gradient boosting), mortality, co-morbidity

A Corrigendum on

Insights Into Co-Morbidity and Other Risk Factors Related to COVID-19 Within Ontario, Canada

by Snider, B., Patel, B., and McBean, E. (2021). Front. Artif. Intell. 4:684609. doi: 10.3389/frai.2021.684609

In the original article, there was an error. The dataset we analyzed requires very specific language regarding acknowledgements. We have therefore adjusted the corresponding text in the acknowledgments, disclaimer and data availability sections

A correction has been made to the **Acknowledgments** section:

"This study was supported by ICES, which is funded by an annual grant from the Ontario Ministry of Health (MOH) and the Ministry of Long-Term Care (MLTC). This study was supported by the Ontario Health Data Platform (OHDP), a Province of Ontario initiative to support Ontario's ongoing response to COVID-19 and its related impacts. We thank IQVIA Solutions Canada Inc. for use of their Drug Information Database. Ontario Community Health Profiles Partnership (OCHPP) created Ontario Marginalization Index (ON-Marg) which is a source for this paper as ON-Marg is used to understand inequalities in health and other social problems related to health among either population groups or geographic areas across Ontario. Postal Code Conversion File and census data were adapted from Statistics Canada. This does not constitute an endorsement by Statistics Canada of this product."

A correction has been made to the **Author's Disclaimer** section:

"The analyses, opinions, results, and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by ICES, the OHDP, its partners, or the Province of Ontario is intended or should be inferred. Parts of this material are based on data and/or information compiled and provided by CIHI. However, the analyses, conclusions, opinions, and statements expressed in the material are those of the author(s), and not necessarily those of CIHI.
Parts of this material are based on data and information provided by Cancer Care Ontario (CCO). The opinions, results, view, and conclusions reported in this paper are those of the

Approved by:
Frontiers Editorial Office, Frontiers Media SA, Switzerland

****Correspondence:***
Brett Snider
bsnide01@uoguelph.ca

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *15 August 2021*
Accepted: *18 August 2021*
Published: *13 September 2021*

Citation:
Snider B, Patel B and McBean E (2021) Corrigendum: Insights Into Co-Morbidity and Other Risk Factors Related to COVID-19 Within Ontario, Canada. Front. Artif. Intell. 4:759022. doi: 10.3389/frai.2021.759022

authors and do not necessarily reflect those of CCO. No endorsement by CCO is intended or should be inferred."

A correction has been made to the **Data Availability Statement** section:

"These datasets were linked using unique encoded identifiers and analyzed at ICES. The use of the data in this project is authorized under Section 45 of Ontario's Personal Health Information Protection Act (PHIPA) and does not require review by a Research Ethics Board.
Access to datasets: The dataset from this study is held securely in coded form at ICES. While legal data sharing agreements between ICES and data providers (e.g., healthcare organizations and government) prohibit ICES from making the dataset publicly available, access may be granted to those who meet pre-specified criteria for confidential access, available at www.ices.on.ca/DAS (email: das@ices.on.ca). The full dataset creation plan and underlying analytic code are available from the authors upon request, understanding that the computer programs may rely upon coding templates or macros that are unique to ICES and are therefore either inaccessible or may require modification."

The authors apologize for this error and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

frontiers in Big Data

ORIGINAL RESEARCH
published: 22 October 2021
doi: 10.3389/fdata.2021.718351

Modelling Representative Population Mobility for COVID-19 Spatial Transmission in South Africa

A. Potgieter[1], I. N. Fabris-Rotelli[1], Z. Kimmie[2], N. Dudeni-Tlhone[3], J. P. Holloway[3], C. Janse van Rensburg[4], R. N. Thiede[1], P. Debba[5,6], R. Manjoo-Docrat[6], N. Abdelatif[4] and S. Khuluse-Makhanya[7,8]*

[1]Department of Statistics, University of Pretoria, Pretoria, South Africa, [2]Foundation of Human Rights, Johannesburg, South Africa, [3]Operational Intelligence, NextGen Enterprises and Institutions, Council for Scientific and Industrial Research, Pretoria, South Africa, [4]Biostatistics Research Unit, South African Medical Research Council, Cape Town, South Africa, [5]Inclusive Smart Settlements and Regions, Smart Places, Council for Scientific and Industrial Research, Pretoria, South Africa, [6]Department of Statistics and Actuarial Science, University of Witwatersrand, Johannesburg, South Africa, [7]IBM Research, Johannesburg, South Africa, [8]College of Graduate Studies, University of South Africa, Johannesburg, South Africa

OPEN ACCESS

Edited by:
Hsiang-Yun Wu, Vienna University of Technology, Austria

Reviewed by:
Jun-Li Lu, University of Tsukuba, Japan
Sunyoung Jang, Princeton Radiation Oncology Center, United States

***Correspondence:**
I. N. Fabris-Rotelli
inger.fabris-rotelli@up.ac.za

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Big Data

Received: *31 May 2021*
Accepted: *06 September 2021*
Published: *22 October 2021*

Citation:
Potgieter A, Fabris-Rotelli IN, Kimmie Z, Dudeni-Tlhone N, Holloway JP, Janse van Rensburg C, Thiede RN, Debba P, Manjoo-Docrat R, Abdelatif N and Khuluse-Makhanya S (2021) Modelling Representative Population Mobility for COVID-19 Spatial Transmission in South Africa. Front. Big Data 4:718351. doi: 10.3389/fdata.2021.718351

The COVID-19 pandemic starting in the first half of 2020 has changed the lives of everyone across the world. Reduced mobility was essential due to it being the largest impact possible against the spread of the little understood SARS-CoV-2 virus. To understand the spread, a comprehension of human mobility patterns is needed. The use of mobility data in modelling is thus essential to capture the intrinsic spread through the population. It is necessary to determine to what extent mobility data sources convey the same message of mobility within a region. This paper compares different mobility data sources by constructing spatial weight matrices at a variety of spatial resolutions and further compares the results through hierarchical clustering. We consider four methods for constructing spatial weight matrices representing mobility between spatial units, taking into account distance between spatial units as well as spatial covariates. This provides insight for the user into which data provides what type of information and in what situations a particular data source is most useful.

Keywords: COVID-19, spatial, mobility, spatial weight matrices, principal component analysis, hierarchical clustering

1 INTRODUCTION

The COVID-19 pandemic starting in the first half of 2020 has changed the lives of everyone across the world. From working from home at all hours, using less public and personal transport, home-schooling under lock down, to economic strife and anxiety; predicting such changes would have been near impossible a priori. By far the largest impact, aside from the economic troubles many find themselves in, is reduced mobility. Daily commuting has been much reduced due to various lockdown measures internationally. In addition, international flights and cross border travel was restricted for significant periods of time, even between regions in some countries.

Reduced mobility was essential, however, due to it being the largest impact possible against the spread of the little understood SARS-CoV-2 virus. Social distancing and stay at home instructions were understood and implemented internationally. These instructions were seen as the best protection for the individual, as well as being the means to prevent overload on the hospital systems, which would otherwise result in inflated death rates. These protection mechanisms are

formed on an understanding of the basic nature of the spatial spread of the virus. A virus spreads via a host, whom it relies on to move amongst other susceptibles. The more movement and interaction performed by the host, the more the virus is able to spread. It is thus imperative to incorporate a spatial element when modelling the spread of the COVID-19 pandemic. Herein, we focus on modelling the mobility spatially.

Quantifying mobility patterns of people facilitates a more accurate understanding of the spread of the disease. An individual's ability to physically "lock down" and stay at home was affected by economic inequality, as shown in a US study (Huang et al., 2021). In South Africa, this economic inequality is extreme, with the World Bank recognising South Africa, in 2019, as having the worst inequality in the world[1].

While the strict lockdown introduced by the South African government from March 27, 2020 delayed the first wave, the mobility was by no means completely reduced due to many living day-to-day for food. Food parcel queues from food donations were a large focus during the first half of the pandemic in South Africa, as the risk of contracting COVID-19 was overridden by the need for food. Such queues, and the use of public transport during these times, heightened the transmission risk of COVID-19 in South Africa, even while lockdown rules were in place. A full lockdown was therefore not possible, and spatial interaction continued between individuals from different regions across South Africa. Modelling regions in isolation will therefore not capture the influence of this mobility on the spread of COVID-19 in South Africa. The use of mobility data in modelling COVID-19 is thus essential to capture the intrinsic spread through the population. A common source is mobile phone location data, which has been utilized previously for epidemiological modelling (Cummings et al., 2004; Wesolowski et al., 2012; Bengtsson et al., 2015; Wesolowski et al., 2015; Finger et al., 2016; Ruktanonchai et al., 2016). However, this data is difficult to obtain due to increasing privacy concerns worldwide. In addition, there are often a number of network providers in a region, each with certain market share. Without data access from all, or at least, the largest providers, representativeness and mobile phone penetration will be limited and should be used with caution. Other sources of mobility data are published by Facebook and Google. The spatial resolution of these is lower, however. In this paper we focus on mobile phone and Facebook mobility data, which has higher spatial resolution than the Google alternative.

It is necessary to determine to what extent different sources of mobility data, at differing spatial resolutions, convey the same message of mobility within a region. In this paper we demonstrate, through the use of principal component analysis as well as hierarchical clustering, how different sources of spatial mobility data at various resolutions can lead to different conclusions with regards to spatial unit connectivity. Spatial connectivity is an essential first step in spatial modeling, providing a quantification of the spatial dependency between spatial units. Herein, we compare the calculation of a number of spatial weight matrices in quantifying how spatial units relate. We also discuss the advantages of different sources and how they can be harnessed when modelling the spread of a virus. We do this by using principal component analysis in order to condense the information that can be gained from a spatial weight matrix and then using hierarchical clustering to identify the strongest spatial associations and to essentially put on display what type of relationships the spatial weight matrix is identifying. This is to the best of our knowledge the first time this exact combination has been used for this purpose.

The mobility data available for South Africa is presented in **Section 2**. The methodology for constructing connectivity matrices is developed in **Section 3**, and the results are presented in **Section 4**. **Section 5** provides a discussion and **Section 6** concludes.

TABLE 1 | South Africa's administrative boundaries.

Administrative level	Spatial unit name	Number of spatial units
0	Country	1
1	Province	9
2	District municipality	52
3	Local municipality	213
4	Ward	4,392

2 DATA

Available mobility data is at different resolutions. For the case of South Africa, the administrative divisions of the country are summarised in **Table 1**. In order of increasing spatial resolution these are country, province, district municipality, local municipality, and ward, labelled as administrative levels 0 through 4 respectively. To facilitate the comparison of different sources of spatial information, it is first necessary to aggregate the data from each source to the same spatial resolution. Increasing the resolution of spatial data can be achieved through methods such as small area estimation or spatial micro-simulation (see e.g. (Ballas et al., 2005; Pfeffermann, 2013)). These methods are somewhat involved and require the use of auxiliary information or assumptions that are unlikely to be true. In this paper we investigate aggregating down to the lowest spatial resolution used by our data sources. While this is relatively straightforward to accomplish, it potentially results in the loss of information.

Mobility data are used to understand various issues ranging from epidemic modelling, transport planning and management, communication network improvement and urban planning (Asgari et al., 2013; Zhou et al., 2018). Asgari et al. (2013) indicates that mobility goes far beyond mere geographical movement of humans, but provides a comprehensive perspective on human interactions that could be considered from spatial, temporal, and contextual aspects. Human mobility is one of the aspects of mobility that gained attention from the global spread of infectious diseases as with the recent COVID-19 pandemic. A variety of technologies including

[1]https://povertydata.worldbank.org/Poverty/Home (Accessed May 2021)

TABLE 2 | South Africa's lockdown levels and dates.

Level	Date	Restrictions
Business as usual	March 1, 2020–March 26, 2020	No restrictions
Level 5	March 27, 2020–April 30, 2020	Essential services only otherwise all confined to place of residence. No inter-provincial movement, except for transportation of goods and exceptional circumstances e.g. funerals. Public and private transport restricted to certain times of the day with limitations on vehicle capacity
Level 4	May 1, 2020–May 31, 2020	More sectors permitted with restrictions, including mining, and partial e-Commerce allowed. Public places (such as religious, cultural, recreational facilities) and the tourism sector remain closed and gatherings prohibited. All confined to place of residence from 8pm to 5am. No local (between metropolitan areas or districts) or inter-provincial movement of people, except for permitted reasons e.g. returning for alert level 4 operations. All borders remain closed except for designated ports of entry for restricted home affairs operations and for the transportation of fuel, cargo and goods. Public and private transport may operate at all times of the day, with limitations on vehicle capacity
Level 3	June 1, 2020–August 17, 2020	More sectors permitted including take away restaurants, e-commerce and delivery services and global business services. Public places and tourism opened and gatherings and sporting activities permitted but all subject to restrictions. All confined to place of residence from 11pm to 4am. No inter-provincial movement of people, except for transportation of goods, exceptional circumstances and other permitted reasons. Public and private transport may operate at all times of the day, with limitations on vehicle capacity

navigation sensors, wireless technologies, and cellular communication technologies are used to position humans in space (Toch et al., 2019). A study by Zhou et al. (2018) provides a comprehensive overview of the different types of human mobility patterns data. These include those data types that capture both the wider (city-wide) and minute (building-wide or large structure) human movements, for example, cellular services records (CSRs), surrounding WiFi access point records (SWAPRs), Global Positioning System locations (GPSLs), geotagged social media (GTSM), public transport smart card records (PTSCRs), bluetooth detection records (BDRs), and WiFi probe request records (WFPRs). The analysis methods range from data visualisation to statistical analysis methods (classification and clustering), heuristic logic, graph theory and optimization techniques.

2.1 South Africa's Lockdown Levels

To quell the spread and impact of the COVID-19 pandemic, the South African government instigated one of the strictest lockdowns in the world. This particular lockdown strategy is structured around different "levels" of lockdown, each of which brings different restrictions (with level 5 being the highest and placing restrictions on nearly all forms of travel to all citizens except for those classified as essential workers). The various levels as well as the dates for which they were active are given in **Table 2**. Note that for this paper we only consider the lockdown until the end of Level 3 due to data availability only over this period.

As non-pharmaceutical interventions (such as the lockdown) are eased the population is allowed to become more mobile. Naturally this will have an impact on the transmission rate of the virus and thus this temporal element must be included in some manner. In this paper we split the data temporally on the date ranges given in **Table 2** up to level 4 and set up a spatial weight matrix for each level of lockdown to study how mobility patterns changed. Two mobility data types were available for this research. The first is freely available data shared by Facebook, and the second is mobility data made available by a South African cellular provider for the context of the COVID-19 response in 2020. In **Figure 7** we provide the Google mobility data at country level. We do not use this data in this research as it is only available at administrative level 1, representing low spatial resolution. It is however useful for context providing mobility levels in each various industry sectors. Mobility for residential travel (i.e., individuals remaining at their place of residence) is the only type of travel that saw an increase after the country transitioned into level 5. Grocery and pharmacy travel saw an initial spike shortly before the country went into level 5 (possibly attributed to panic-buying). After transitioning to level 5 we see a drastic decrease in all types of travel, with residential travel showcases a slightly downward trend while all other forms of travel have an upward trend. Grocery and pharmacy travel is the quickest to recover to pre-COVID levels while travel to parks and travel stations is the slowest to recover (most likely due to this being for leisure). By the end of the year residential travel is still higher than before any lockdown interventions. **Table 3** provides the average changes over each level as well.

2.2 Facebook Data for Good

Multiple geographically indexed datasets have been made freely available for use by Facebook through their "Facebook data for good" initiative. These datasets serve to aid researchers and policymakers in understanding the spread of COVID-19[2].

This paper utilises one of these available datasets, namely the "Movement range maps" dataset. The data indicates the change in mobility, $F_i^{(t)} \in (-1, 1)$ (which can be interpreted as a percentage (−100, 100)), for a spatial unit i on a given day t over the period March 1, 2020–February 28, 2021 relative to a 1-week baseline calculated in February 2020. The daily values for each district municipality were calculated by determining the number of so-called "Bing tiles"[3] that each inhabitant visited on a given day (place of residence being determined by the location where users most often spend their nights). A bing tile is the term used by Microsoft for a spatial polygon. After incorporating some degree

[2]https://dataforgood.fb.com/(Accessed May 2021)

[3]https://docs.microsoft.com/en-us/bingmaps/articles/bing-maps-tile-system (Accessed May 2021)

TABLE 3 | Average changes in population mobility over lockdown levels using the Google mobility data during 2020.

Level	Date	Retail	Grocery and pharmacy	Parks	Transit stations	Workplaces	Residential
BAU	2 Feb - 26 Mar	–3.49	1.68	–9.39	–5	–0.88	1.71
Level 5	27 Mar - 30 Apr	–73.06	–46.09	–46.86	–78.49	–65.89	33
Level 4	1 May - 31 May	–50.39	–23.45	–39.39	–61.71	–40.58	23.35
Level 3	1 Jun - 17 Aug	–29.53	–10.71	–23.17	–49.72	–28.1	17.17
Level 2	18 Aug onwards	–17.76	–3.34	–23.29	–34.65	–19.78	11.35

of noise, the average number of tiles visited by the inhabitants was determined and expressed relative to the baseline. The full description of how these values were calculated is available in the **Appendix**. The spatial resolution for units of this data are district municipalities, namely at administrative level 2.

Figure 1 shows the aggregated data for district municipalities, with the average across the district municipalities shown in red. The figure demonstrates that the average mobility nationally dropped significantly in late March. This corresponds to when South Africa entered its first hard lockdown on the March 27, 2020 (see **Table 2**). The hard lockdown imposed severe restrictions on travel and constituted a strict stay at home directive. Only essential workers were allowed to leave their homes. Furthermore, the average change in mobility is primarily negative over the entire study period, indicating that mobility patterns remain more constrained than before the hard lockdown. The first COVID-19 case was discovered on March 5, 2020 and the lockdown announcement was made a week later on 15 March. This could explain the drop in mobility already seen from early March.

Notable benefits of using this data are that the data is freely available and could potentially act as a very representative proxy for human mobility, as Facebook services are not constrained to specific mobile network providers. In addition, all the cellular network providers in South Africa provide a free version of Facebook called Facebook Zero. Even though it is known that not all South Africans have a Facebook account, the Facebook mobility data may provide an acceptable level of representativeness for mobility within the country since the population of South Africa is considered significantly young[4]. It is also clear that a large amount of the original data was censored in order to preserve user privacy and thus the data is at a sparse level of spatial resolution (administrative level 2). The data is also not specific with regards to the direction of spatial mobility. Daily observations only indicate whether individuals were more or less mobile in a district municipality and do not indicate the spatial units towards which this mobility was directed.

2.3 Mobile Network Data

The growing popularity and widespread use of mobile devices has led to massive amounts of data being produced at any given point in time all around the world. Mobile phone data can be collected either passively by mobile services providers or through the use of mobile applications. The ease with which such large quantities of data can be gathered makes cellular data attractive for researchers. Mobile devices operate by sending and receiving information from cellphone towers. When interacting with a cellphone tower we say that a phone has "pinged" off a cellphone tower. A mobile device may ping off a cellphone tower by sending or receiving any kind of information, be it a phone call, text message or application notification. The mobile network data obtained for this research is obtained using the number of users whose mobile devices pinged off a cellphone tower within one ward (administrative level 4) on a given day and then later that day pinged off a cellphone tower in a different ward.

Mobile phone data has been used numerous times in the field of spatial epidemiology to model the spread of various diseases, including cholera (Bengtsson et al., 2015; Finger et al., 2016), dengue (Cummings et al., 2004; Wesolowski et al., 2015) and malaria (Wesolowski et al., 2012; Ruktanonchai et al., 2016). Following the outbreak of the COVID-19 pandemic, the governments of various countries across the world began collecting cellular device user data in an attempt to aid the conception and implementation of non-pharmaceutical interventions (Ekong et al., 2020; Oliver et al., 2020; Peixoto et al., 2020; Varsavsky et al., 2021). This data has since been used by researchers to clearly establish a correlation between population mobility and COVID-19 case numbers (Gao et al., 2020; Peixoto et al., 2020; Xiong et al., 2020; Zhou et al., 2020).

Limitations of mobile phone data exist. First and foremost of these is the issue of user privacy. Mobile phone data could potentially be misused to identify specific individuals and thus cellular providers are often hesitant to provide researchers with such data (Grantz et al., 2020; Oliver et al., 2020). Such data is often aggregated to a low spatial resolution to prevent this as well as reduce noise, but this comes at the cost of some data specificity. Another potential drawback of mobile phone data is high computational cost. For high mobile phone penetration rates, mobile phone data may consist of a number of entries in the order of billions. The computational cost of processing such datasets is prohibitive, potentially preventing analysis.

For this paper, anonymised mobile phone data was obtained from a local mobile network provider. In South Africa, the mobile phone penetration level is estimated to be as high as 95%[5]. The

[4]Mid-2021 Statistics South Africa Population Report http://www.statssa.gov.za/publications/P0302/P03022021.pdf (Accessed August 2021)

[5]See https://www.geopoll.com/blog/mobile-penetration-south-africa/and https://www.icasa.org.za/uploads/files/State-of-the-ICT-Sector-Report-March-2020.pdf (Accessed May 2021)

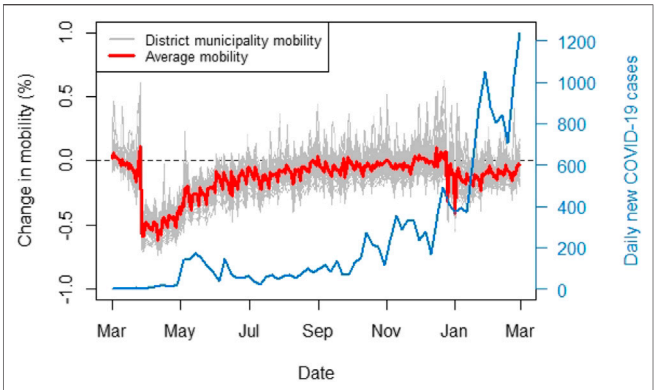


FIGURE 1 | "Facebook for good" movement range maps data (March 1, 2020–February 28, 2021) relative to a baseline calculated in a week of February 2020.

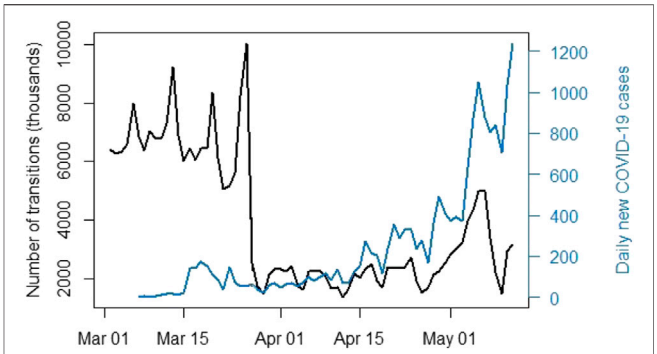


FIGURE 2 | Number of individual transitions between wards using the available mobile phone data (March 2, 2020–May 12, 2020).

provider utilised in this paper is one of the largest providers in the country, with an estimated market share of 42%.

The data provides the number of mobile phone users $m_{ij}^{(t)}$ that travelled to ward j from ward i on day t for the period 2 March - May 12, 2020. The data is at administrative level 4, which is the highest spatial resolution reasonably possible while preserving some level of privacy of exact user location. To compare insights gained from this data and the Facebook dataset in **Section 2.2**, it would first be necessary to aggregate the mobile phone data to the same spatial resolution which is administrative level 2. In South Africa, each ward has a unique 8-digit ID code. The first three digits of this code indicates the district municipality that the ward is a part of. For example, the ward ID 9344007 indicates that the ward is part of the district municipality with code 934. In order to aggregate the data to district municipality level, one could replace the ward IDs of the observations with their district municipality codes (i.e. only the first three digits), whereupon rows with identical origin and destination codes would be discarded. The mobile phone data at administrative level 2 is thus given by

$$M_{I,J}^{(t)} = \sum_{i \in I, j \in J} m_{ij}^{(t)},$$

where I and J are district municipalities and i and j are wards as previously indicated. Transitions contained within a single district municipality are thus discarded. Analysis revealed that this caused an average of 26% of daily observations to be discarded. The retained data is displayed in **Figure 2**. The representation differs to that of **Figure 3** as the data provides transitions between regions in this case. We once again notice a sharp decline in population mobility in late March.

The population of South Africa (mid-2021) is approximately 60.14 million[6], and yet the highest total number of inter-district municipality transitions on any given day was approximately 10 million (seen in **Figure 2**). It should be noted that the same individual can be responsible for multiple transitions and that some individuals could potentially possess multiple cellular devices. Literature does exist on the use of mobile phone data to estimate population numbers, see e.g. (Sakarovitch et al., 2018). Doing so is not within the scope of the research presented here but would be of value in testing mobile phone representability. Despite the quality of available hardware[7], this process proved highly computationally expensive due to the number of comparisons that need to be run on billions of lines of data in order to create a spatial weight matrix for each day in the time period.

3 METHODOLOGY

3.1 Literature Review

When a particular phenomenon exhibits evidence of spatial dependence, this dependency must be taken into account when modelling to minimise the risk of producing biased results (Stakhovych and Bijmolt, 2009; Ejigu and Wencheko, 2020). In the case of an infectious disease that is spread through physical contact and near proximity, it is clear that locations that are situated closer together (or rather the inhabitants of these locations) will play a larger role in determining their respective infection rates than locations that are farther apart. To incorporate this fact, spatial models allow spatial units to be more strongly (or weakly) correlated with one another based on some select criteria that is deemed suitable for the phenomenon being modelled. This is achieved through the use of a spatial weight matrix (sometimes called a "spatial mobility matrix") usually denoted by $\boldsymbol{W}$ (Bavaud, 1998; Getis and Aldstadt, 2004; Aldstadt and Getis, 2006; Stakhovych and Bijmolt, 2009; Anselin, 2013; Ejigu and Wencheko, 2020).

Definition 1 (Spatial weight matrix). Let $S = \{1, 2, \ldots, n\}$ be a set of spatial units. A spatial weight matrix (Bavaud, 1998; Getis and Aldstadt, 2004; Stakhovych and Bijmolt, 2009; Anselin, 2013) is an $n \times n$ matrix $\boldsymbol{W} = [w_{ij}]$ satisfying $w_{ij} \geq 0$ and $\sum_{j=1}^{n} w_{ij} = 1 \quad \forall \quad i \in S$.

[6]Mid-2021 Statistics South Africa Population Report http://www.statssa.gov.za/publications/P0302/P03022021.pdf (Accessed August 2021)

[7]All analysis presented here was performed on a desktop computer running Intel Core i7 with a clock speed of 3.40GHz, a 64-bit operating system and 64 GB of installed memory

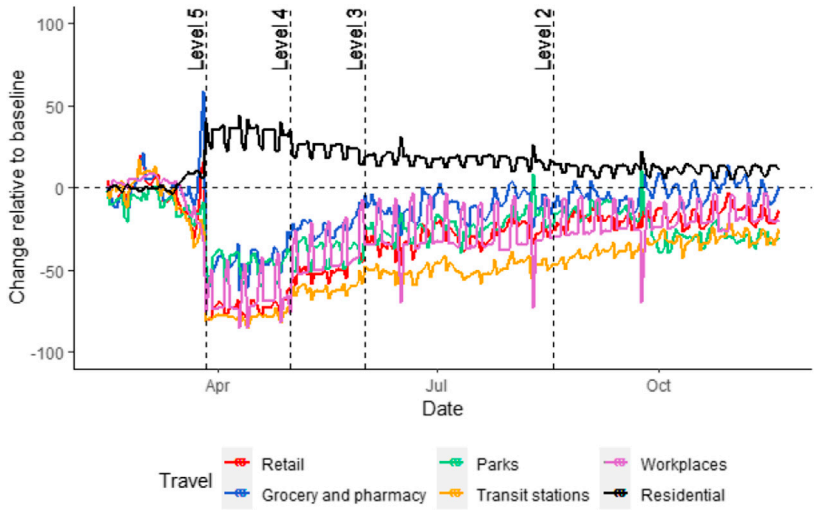


FIGURE 3 | Google mobility report data for February 15, 2020–November 20, 2020 (transitions to different levels of lockdown indicated by vertical reference lines).

This matrix is formally defined as an expression of spatial dependency between spatial units (Bavaud, 1998; Getis and Aldstadt, 2004; Stakhovych and Bijmolt, 2009; Anselin, 2013). Simply put, the spatial weight matrix is constructed in such a way so that entry w_{ij} quantifies the amount of spatial influence that spatial unit i exerts on spatial unit j (Bavaud, 1998; Getis and Aldstadt, 2004; Stakhovych and Bijmolt, 2009; Anselin, 2013).

Such matrices are frequently restricted to being symmetrical to simplify estimation. However, symmetry is not required and can result in a less realistic representation of spatial dependency (Bavaud, 1998; Getis and Aldstadt, 2004; Stakhovych and Bijmolt, 2009; Anselin, 2013). Another common convention is that $w_{ii} = 0$ for all i to exclude the possibility of so-called "self-influence" (Bavaud, 1998; Getis and Aldstadt, 2004; Stakhovych and Bijmolt, 2009). Non-zero diagonal entries can however be included and are interpreted as quantifying the resistance that each spatial unit has against influence from the other spatial units (Bavaud, 1998; Anselin, 2013). Performing row-standardisation on the matrix allows the connectivity of different spatial units to be compared (Bavaud, 1998; Getis and Aldstadt, 2004).

Spatial weight matrices are most commonly used in the fields of econometrics and spatial statistics (Anselin, 2013). Recently however, they have become popular in the field of spatial epidemiology and have been used to model various diseases including dengue, malaria, foot and mouth disease (Brown et al., 2016; Malik et al., 2016; Brown et al., 2018; Suryowati et al., 2018) and most recently COVID-19 (Huang et al., 2020; Tagliazucchi et al., 2020). There are relatively few established guidelines with regards to constructing a spatial weight matrix (Bavaud, 1998; Aldstadt and Getis, 2006; Stakhovych and Bijmolt, 2009; Ejigu and Wencheko, 2020), however, the construction of these matrices has seen some advancement, with greater emphasis being placed on creating matrices that offer an accurate representation of human mobility. Simpler models rely on measures such as distance, contiguity or adjacency (Aldstadt and Getis, 2006; Stakhovych and Bijmolt, 2009; Anselin, 2013; Brown et al., 2016; Malik et al., 2016; Brown et al., 2018; Suryowati et al., 2018; Ejigu and Wencheko, 2020; Huang et al., 2020) while more complex ones are able to use mobile phone data (Huang et al., 2020) and geostatistical information (Getis and Aldstadt, 2004; Aldstadt and Getis, 2006). Accurately specifying these matrices is a non-trivial problem, as discussed in (Ejigu and Wencheko, 2020). Most recently, Ejigu et al. proposed a methodology through which both distance and covariate information can be utilized (Ejigu and Wencheko, 2020).

Given the importance of correctly specifying the spatial weight matrix, and the fact that there are often multiple sources of spatial data available on hand, it becomes necessary to develop some means of comparing spatial weight matrices. Specifically, it is necessary to compare the insights that can be derived from different spatial weight matrix definitions. In recent years this comparison has been achieved either through the use of measures of spatial autocorrelation, such as Moran's I (Suryowati et al., 2018), or through more specialised methods local to the field of spatial statistics (Gao et al., 2018; Jin et al., 2020). In this paper, we adapt an idea initially presented by Garrison and Marble (Garrison and Marble, 1964), whereby principal component analysis is used to reduce the dimensionality of candidate spatial weight matrices. We then introduce the use of hierarchical clustering to derive a clustering solution for the spatial unit principal scores. This allows for a more informative comparison of the information provided by these connectivity matrices, as opposed to simply comparing their structure visually.

3.2 Spatial weight Matrices

Selecting an optimal spatial weight matrix is often reliant on the use of a priori information and experience. In this paper the emphasis is on comparing the implications for different spatial weight matrices and the varying types of spatial associations that they represent. We next discuss the spatial weight matrix construction approaches used in this paper.

3.2.1 Method 1: Distance Method

The exponential distance definition of a spatial mobility matrix is used frequently in studies involving spatial correlation, and is a popular choice in spatial econometrics (Aldstadt and Getis, 2006; Stakhovych and Bijmolt, 2009; Anselin, 2013; Ejigu and Wencheko, 2020). As previously mentioned however, the concepts of distance, contiguity and adjacency do not necessarily offer the most accurate or realistic representation of human mobility. In this paper we include this model in order to draw comparisons between it and more data-driven models. The entries of the spatial weight matrix are given by

$$w_{ij} = \exp(-d_{ij}) \tag{1}$$

where d_{ij} is the Euclidean distance between the centroids of spatial units i and j. Diagonal entries are set to 0 to remove the possibility of so-called "self-influence," and all rows are standardised to sum to 1 to facilitate comparisons between different spatial units. Both of these restrictions were maintained for all matrices in this paper. Under this model, spatial units are most strongly spatially correlated with the spatial units that are closest to them geographically. No temporal component can be incorporated for this method.

3.2.2 Method 2: Mobile Network Method

The mobile network data indicates the number of individuals that travelled from spatial unit i to spatial unit j on a given day t. These entries are used to construct the spatial weight matrix as follows,

$$w_{ij}^{(t)} = M_{ij}^{(t)}. \tag{2}$$

This model expresses spatial weights as a function of the amount of flux (both in and out) occurring at a spatial location, and is sometimes referred to as a spatial interaction matrix (Bavaud, 1998). Spatial units where more (less) individuals travelled to other spatial units will thus have a larger (smaller) effect on other spatial units.

3.2.3 Method 3: Weighted Facebook Data Method

In order to create a spatial mobility matrix using the Facebook data, we use the same approach of Ejigu et al. (Ejigu and Wencheko, 2020). This takes into account proximity as well as covariate information which is spatially dependent. The entries of the spatial weight matrix are given by

$$w_{ij}^{(t)} = \exp\left(-\left(\alpha \cdot |F_i^{(t)} - F_j^{(t)}| + (1-\alpha)\cdot d_{ij}\right)\right) \tag{3}$$

where $F_i^{(t)}$ is the mobility of spatial unit i at time t, scaled by population size (the covariate information), d_{ij} is the Euclidean distance between the centroids of spatial units i and j, and $\alpha \in (0, 1)$ is a control parameter indicating the amount of weight that should be given to the covariate term. The control parameter α was set to 0.6 in this paper to allow for the covariate data to play a slightly more prominent role in the estimation process without disregarding the importance of distance. The parameter captures that we are making an assumption that the Facebook data can be used to capture transitions between regions even though it is isolated location data. The value of 0.6 gives the weighted calculation a slight nudge towards the Facebook data. Note that if $\alpha = 0$ then the model simplifies to the exponential distance model in **Eq. 1**.

The Facebook mobility data for each district municipality was scaled using population size in order to account for the fact that increased mobility in a given district is more (less) influential to neighbouring districts if the population size is large (small). This was also done in order to restore some of the variation in the data that was likely lost when the data was censored to a lower spatial resolution.

3.2.4 Method 4: Scaled Facebook Data Method

An additional final spatial weight matrix was constructed based on further variation of the exponential distance model. For this matrix, the rows of the exponential distance matrix are scaled using the (unscaled) Facebook mobility data. For example, if the mobility within district municipality i was 20% lower than the baseline, then the entire row i is multiplied by 0.8. Each entry in the exponential distance matrix is thus scaled by some number in **(0,2)**. The entries in the matrix are given by

$$w_{ij}^{(t)} = \left(1 + F_i^{(t)}\right) \cdot \exp(-d_{ij}). \tag{4}$$

This construction allows the exponential distance matrix to be scaled such that the spatial influence of more (less) mobile district municipalities is increased (decreased). This also renders the exponential distance matrix non-symmetric, which should offer a more realistic representation of spatial influence. Methods three and four are a novel approach to constructing connectivity matrices from the Facebook mobility data.

3.3 Principal Component Analysis

Principal component analysis (PCA) is a statistical technique that aims to derive a parsimonious representation of a given dataset by deriving an orthogonal linear transformation of the data (Friedman et al., 2001). In standard PCA, the only hyperparameter that needs to be selected is the number of principal components, which is primarily dependent on the cumulative proportion of variance in the data that the user wishes to retain. For this paper, the number of principal components was chosen such that 75% of the variation in the data was maintained. The full discussion of PCA and its various extensions is left to the existing literature (see e.g. (Friedman et al., 2001)).

3.4 Hierarchical Clustering

Hierarchical clustering is an unsupervised machine learning technique that allows the user to group together data points in an attempt to uncover sets of observations that share similar characteristics (Friedman et al., 2001). This is achieved by procedurally grouping together those observations that are most similar to each other based on some selected measure of dissimilarity, referred to as a "linkage" (Friedman et al., 2001). The number of retained clusters can then be selected either using some measure of cluster (dis)similarity or a pre-selected value. We use agglomerative clustering, which additionally requires the selection of a method through which the dissimilarity of separate clusters is calculated. A full discussion on hierarchical clustering may be found in (Friedman et al., 2001).

Herein, we chose the number of clusters to be identical to the number of principal components. Complete linkage was used to calculate the difference between clusters at each iteration. Single

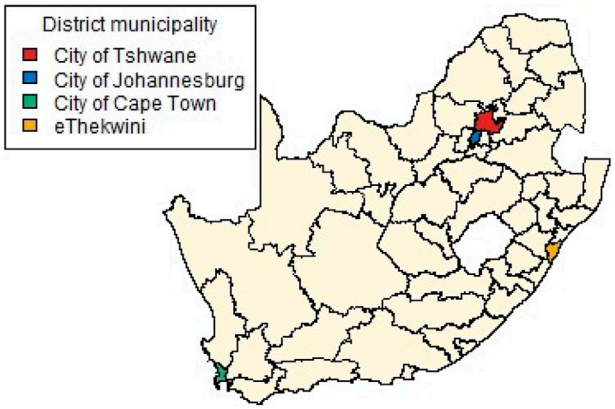


FIGURE 4 | South Africa's district municipality boundaries and locations of four largest cities.

and average linkage displayed a propensity for resulting in clusters that were very large. This was most likely due to the fact that single linkage considers the minimum distance between clusters at each iteration, thus regarding clusters as more similar in general. Complete linkage considers the maximum distance between clusters and thus considers clusters to be more distinct. Average linkage is the average of these two extremes.

4 RESULTS

Figure 4 shows the 52 district municipalities of South Africa. The four largest cities in the country are Tshwane, Johannesburg, Durban and Cape Town, situated in the City of Tshwane, City of Johannesburg, eThekwini and City of Cape Town district municipalities respectively as indicated in colour in **Figure 4**. These four cities are the focal point of economic activity and travel in the country, and it is thus logical that they would play a substantially larger role in the transmission of the virus than other municipalities.

4.1 Method 1: Distance Method

Figure 5A shows the weights (those >5) for the exponential distance weight matrix. Since the entries are calculated based only on the Euclidean distance between the district municipalities (and no additional information), there are no significantly large weights present. As temporal information cannot be included, this method produces only a single spatial weight matrix.

This spatial weight matrix required the largest number of principal components, namely 14, in order to explain 75% of the variation in the data. This is most likely due to the lack of any form of auxiliary data or information that could be used to better describe the relationship of the district municipalities. The result of hierarchical clustering on the principal component observations is given in **Figure 5B**.

4.2 Method 2: Mobile Network Method

Figure 6 shows the spatial weight matrix for every level of lockdown that the mobile phone data spans at administrative level 3. This spatial weight matrix identifies very strong spatial associations over relatively shorter distances (indicated by the yellow lines). These strong correlations appear to cluster around the edges of the country, with locations in the centre of the country displaying less spatial association overall.

We note that there are strong spatial associations that do not appear to be associated with any of the four major cities in the country. In particular, we note strong associations in the North-Western region of the country as well as some spatial associations across Lesotho (a neighbouring country that is landlocked by South Africa, shown in **Figure 6D**). The spatial weight matrices for the mobile network data were also aggregated to administrative level 2, shown at **Figure 7**. While some strong spatial associations can still be identified around the country's borders, many previously identified associations (including several significant associations spanning across the neighboring country of Lesotho) are now negligible. It is clear that while this lower spatial resolution does capture some of the spatial associations present in the data, much information is lost when aggregating between spatial resolutions.

A notable drawback of data being at such a high spatial resolution is that it becomes very difficult to cluster locations in a meaningful way. At administrative level 3 there are 213

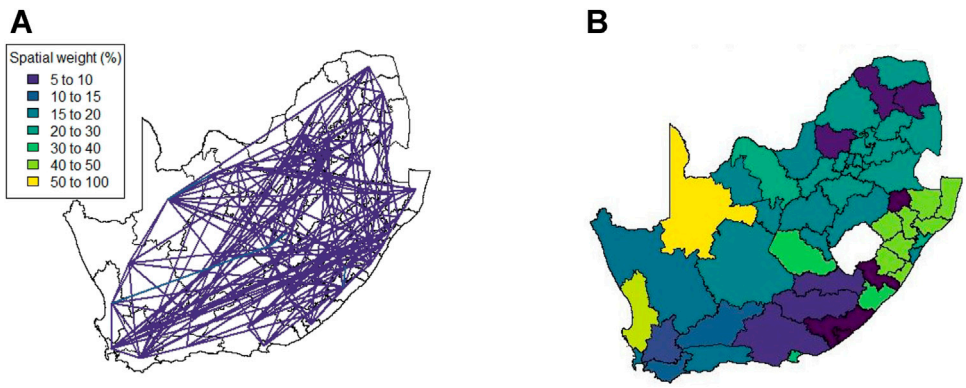


FIGURE 5 | Method 1 **(A)** Spatial weights (weights ≤5 not shown), **(B)** Complete linkage clustering (14 clusters indicated by colours).

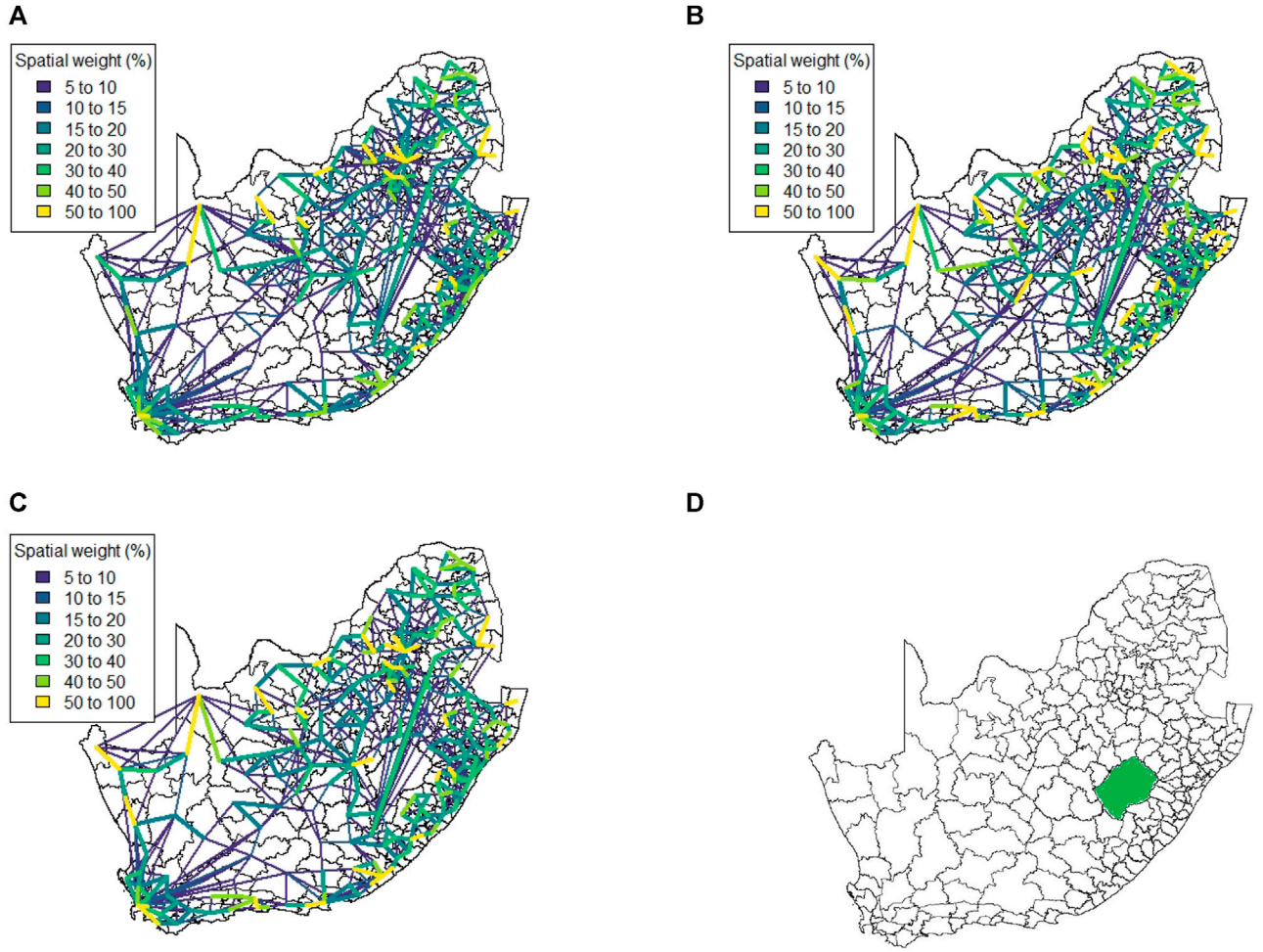


FIGURE 6 | Method 2 spatial weight matrix entries (weights ≤5 not shown) **(A)** Business as usual, **(B)** Level 5, **(C)** Level 4, and **(D)** South Africa at local municipality level (neighboring country Lesotho in green).

spatial units to consider. In order to explain just 75% of the variation in this data one requires approximately 70 principal components. Such a high number of clusters does not lend itself to easy interpretation and thus it is necessary to aggregate to a lower spatial resolution to render analysis feasible. When aggregating to administrative level 2 we find that 20 principal components are required to retain 75% of the variation present in the data. This is most likely due to the fact that the mobile network exhibits far greater daily variation than our data sources. **Figure 8** shows the clustering solution.

4.3 Method 3: Weighted Facebook Data Method

This matrix construction incorporates both the Facebook population mobility data and the population size for each district municipality into the spatial weights for each district municipality pair. **Figure 9** shows the resulting matrix for each level of lockdown. By allowing both mobility and population size to play a role in this matrix, the strong spatial association between the four largest cities in South Africa is identified, despite the large geographical distance between them. If only Euclidean distance had been taken into account, this association would have been missed, as with Method 1. This spatial weight matrix required nine principal components to explain 75% of the variation in the data. **Figure 10** shows the results of applying hierarchical clustering to the principal component observations.

4.4 Method 4: Scaled Facebook Data Method

This spatial weight matrix was constructed as a potentially more realistic alternative to the exponential distance matrix. Despite containing a temporal element (in the form of daily mobility measurements retrieved from the Facebook data), the results for this matrix do not show any significant change across the various levels of lockdown. **Figure 11** visualises the spatial weight matrix. Clustering performed on this matrix was more

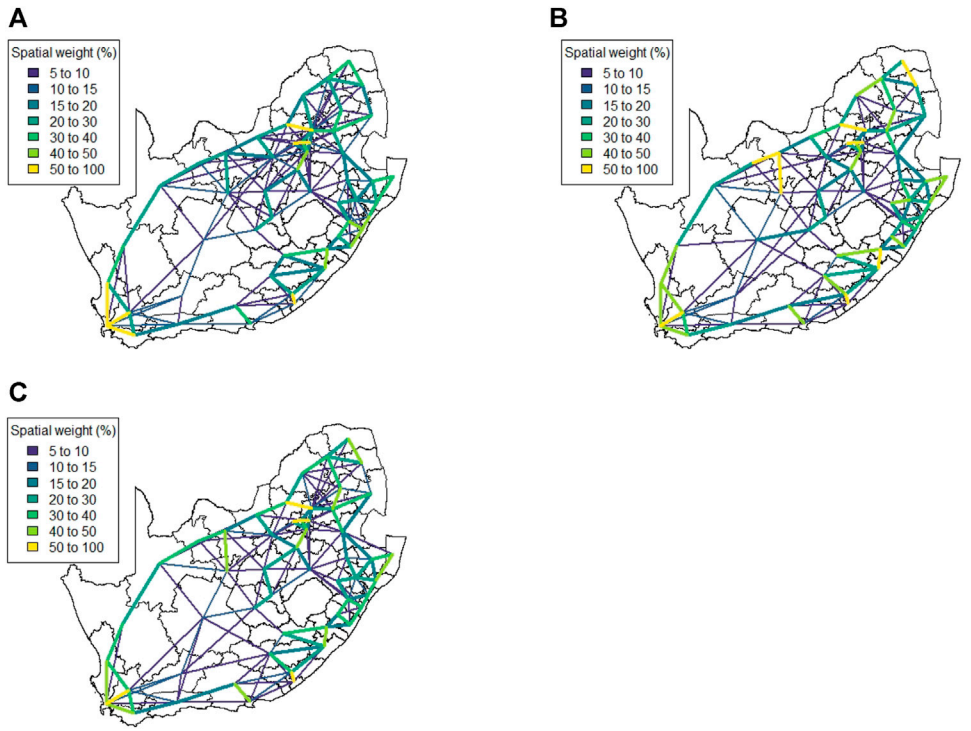


FIGURE 7 | Method 2 spatial weight matrix entries (weights ≤5 not shown) **(A)** Business as usual, **(B)** Level 5, **(C)** Level 4 at district municipality level.

successful and intuitive. Only seven components were required to explain 75% of the variation in the data. **Figure 11** shows the clustering solution.

5 DISCUSSION

The results in **Section 4** illustrate a number of ways to construct spatial weight matrices from mobility data. For the standard exponential distance method (Method 1), it is clear from **Figure 5** that the clustering solution on this spatial weight matrix is not ideal. There are far too many clusters and the clustering solution reveals no clear interpretation. Although the initial matrix construction used only the distances between district municipalities, district municipalities that were located closer together were not generally clustered together. The entries of the spatial weight matrix constructed using the mobile network data (Method 2), shown in **Figures 6**, **7**, reveal strong spatial associations over relatively short distances. The four focal largest cities in the country are clearly identified as hubs for high mobility but there are other regions, particularly those situated on or near the borders of the country, that showcase highly concentrated mobility. A possible explanation for these strong spatial associations being observed far away from cities is the existence of mining activity in these areas. Given that South Africa has a very large and widespread mining sector, it seems only reasonable that any model with a spatial element should strive to incorporate these associations. The clustering solution for this spatial weight matrix, shown in **Figure 8**, is distinct from the other solutions in this paper in that distance is clearly not a key role player in deciding which spatial units are clustered together. Many spatial units that are situated close to one another in geographical space are not clustered together, and some spatial units are even placed into their own clusters despite having many spatial neighbours. It can be argued that this clustering solution is a more realistic reflection of the amount of travel between spatial units. The reason for this is that locations being situated closer together does not always imply that there is a higher degree of travel between these locations. The strong local connectivities picked up by this method are useful for epidemiological modelling, for example, prediction of case number hotspot movement into spatial units of higher likelihood of mobility.

The four largest cities in South Africa are Tshwane, Johannesburg, Cape Town and Durban, situated in the City of Tshwane, City of Johannesburg, eThekwini and City of Cape Town district municipalities respectively, as shown in **Figure 4**. The results in **Figure 9** (method 3) show a large spatial association between these locations prior to the implementation of level 5 lockdown. Under level 5 restrictions, when the spatial influence of most district municipalities decreased, the spatial influence between these four locations became more pronounced by comparison. This most likely indicates that while smaller district municipalities were less active due to restrictions, these four were comparatively more active and still saw a sizable amount of travel between them. This seems feasible, given that these locations are the focal points for economic activity in the country and thus could not reasonably become “immobile”. As restrictions were lifted, these spatial weights were still significantly larger than those for other district municipalities, indicating that,

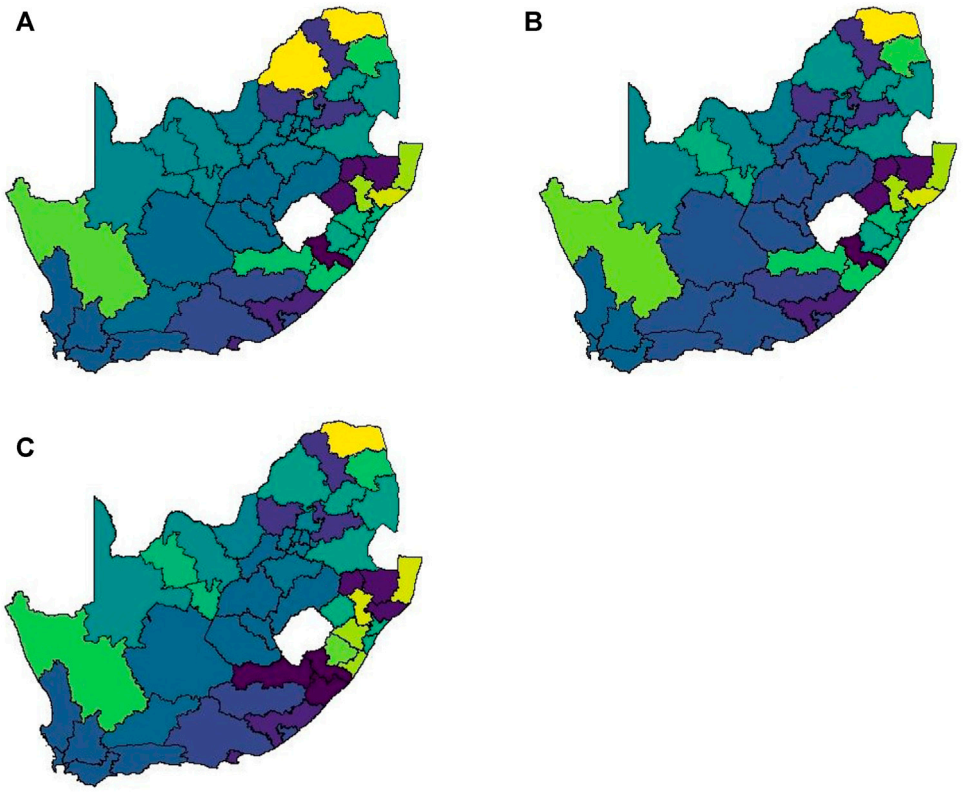


FIGURE 8 | Method 2 complete linkage clustering results (20 clusters) **(A)** Business as usual, **(B)** Level 5 and **(C)** Level 4.

despite restrictions being eased, the spatial influence between these four places is still significantly stronger than before the lockdown. It is also apparent that the spatial influence between less influential district municipalities has not returned to the level that they were during business as usual (pre-lockdown). **Figure 10** shows that the district municipalities housing the four largest cities are all either clustered together or in clusters of their own. Other district municipalities are generally clustered together based on the distance between them. This clustering solution indicates that the four largest cities are significantly different from the locations around them. This spatial weight matrix is thus able to pinpoint the fact that these locations play a potentially larger role in spatially-dependent phenomena such as the spread of a virus. The effect in epidemiological modelling allows for longer range spatial dependency, for example, spread of the virus by daily flights between major city hubs. This is not captured by Method 2.

The clustering results for Method 4, shown in **Figure 11**, do not display any significant changes over the various levels of lockdown. **Figure 11** also shows that the clusters that are formed for this spatial weight matrix are clearly based primarily on distance, but illustrates that the auxiliary Facebook data aids in constructing more finite and sensible clusters. Interestingly, we notice a district municipality that has been classified into a cluster on its own. When inspecting the results for the other spatial weight matrices we note that this district municipality has previously also been identified as its own cluster and was shown to have strong spatial associations for Method 2. Upon further inspection we note this district municipality houses several mines. Similarly to Method 2, this spatial weight matrix is able to identify location associations that go unnoticed when relying on simple concepts such as Euclidean distance. This method may not be useful alone in epidemiological modelling and should most likely be used in conjunction with either Method 2 or 3.

This paper shows that different representations of spatial data can offer a variety of insights and capture different relationships in the data. For example, the spatial weight matrix created using Method three data emphasises the prominent role of focal points in population activity. However, the spatial weight matrix constructed using Method four offers a scaled and smoothed way to use distance to indicate which locations have a higher spatial influence on one another. These two spatial weight matrices use the same spatial data (i.e. the Facebook for good data), but offer vastly different interpretations of spatial influence. Finally, the interpretations that were able to be made from the mobile phone data indicates that there are many potentially strong spatial associations at shorter distances that can only be identified when inspecting data at a high

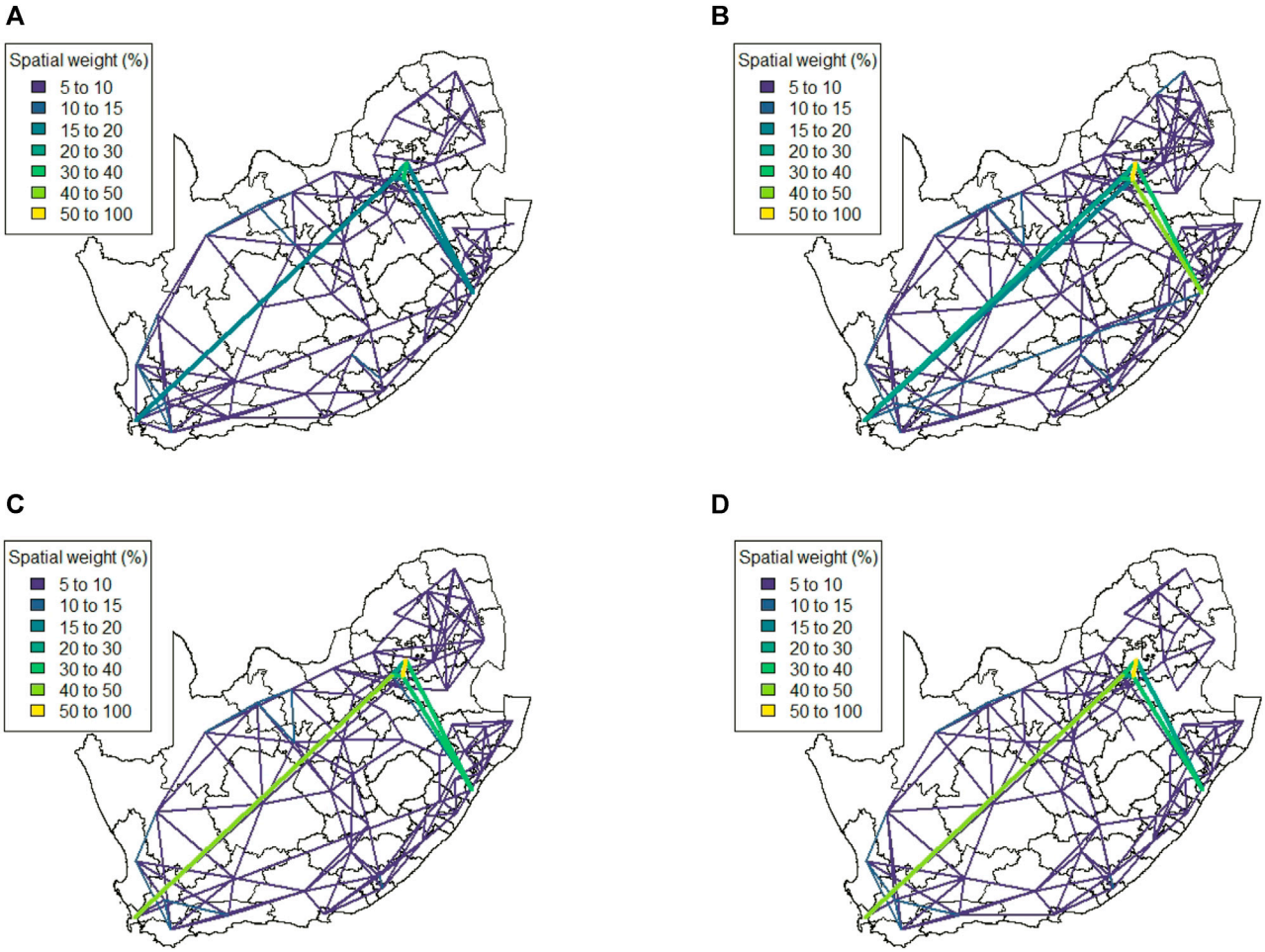


FIGURE 9 | Method 3 spatial weight matrix entries (weights ≤5 not shown) **(A)** Business as usual, **(B)** Level 5, **(C)** Level 4, and **(D)** Level 3.

spatial resolution. **Table 4** provides a summary of the methods used in this paper, their strengths and weaknesses, and their usability based on the results. Each of these representations can be seen as valid and are complementary with regards to the insight they offer. Depending on the specific phenomenon under study, an argument could be made their usability based on observed patterns from the results, as in the case of a pandemic such as COVID-19, which affects not only congregated communities but allows for consequences to be felt across an entire country.

Understanding mobility during the current pandemic is essential. Both the reduction in mobility as well as retained mobility need to be well understood, and depend on reliable data collection. As shown here, data are collected in different ways and are also made available in a variety of formats. Mobility is distributionally different across strata of a region's demographics, with more mobile locations likely to result in higher disease transmission. Higher resolution mobility data is important to capture these differences in more detail. Even so, the spatial resolution at district municipality captures these nuances of the movement under each lockdown level, and shows that significant movement still took place due to the vulnerability of a large portion of South Africa's population.

The possibility of micro-spatial estimation (small area estimation) is something to investigate further. Making use of demographic covariates, transport networks and as well as mobile network coverage maps could provide connectivity matrices at higher spatial resolution, ideally at ward level. Estimation at higher spatial resolution could be done by making use of a number of lower spatial resolution sources. This allows for micro-scale modelling of COVID-19 spread and will allow for privacy while increasing spatial resolution and providing deeper coverage in a region. Google mobility data is also available[8] but only at provincial level (administration level 1) for South Africa. This spatial resolution is too low to consider estimation down to ward level, especially if alternative mobility data is available at administrative level 2. However, one could also combine mobility

[8]https://www.google.com/covid19/mobility/(Accessed May 2021)

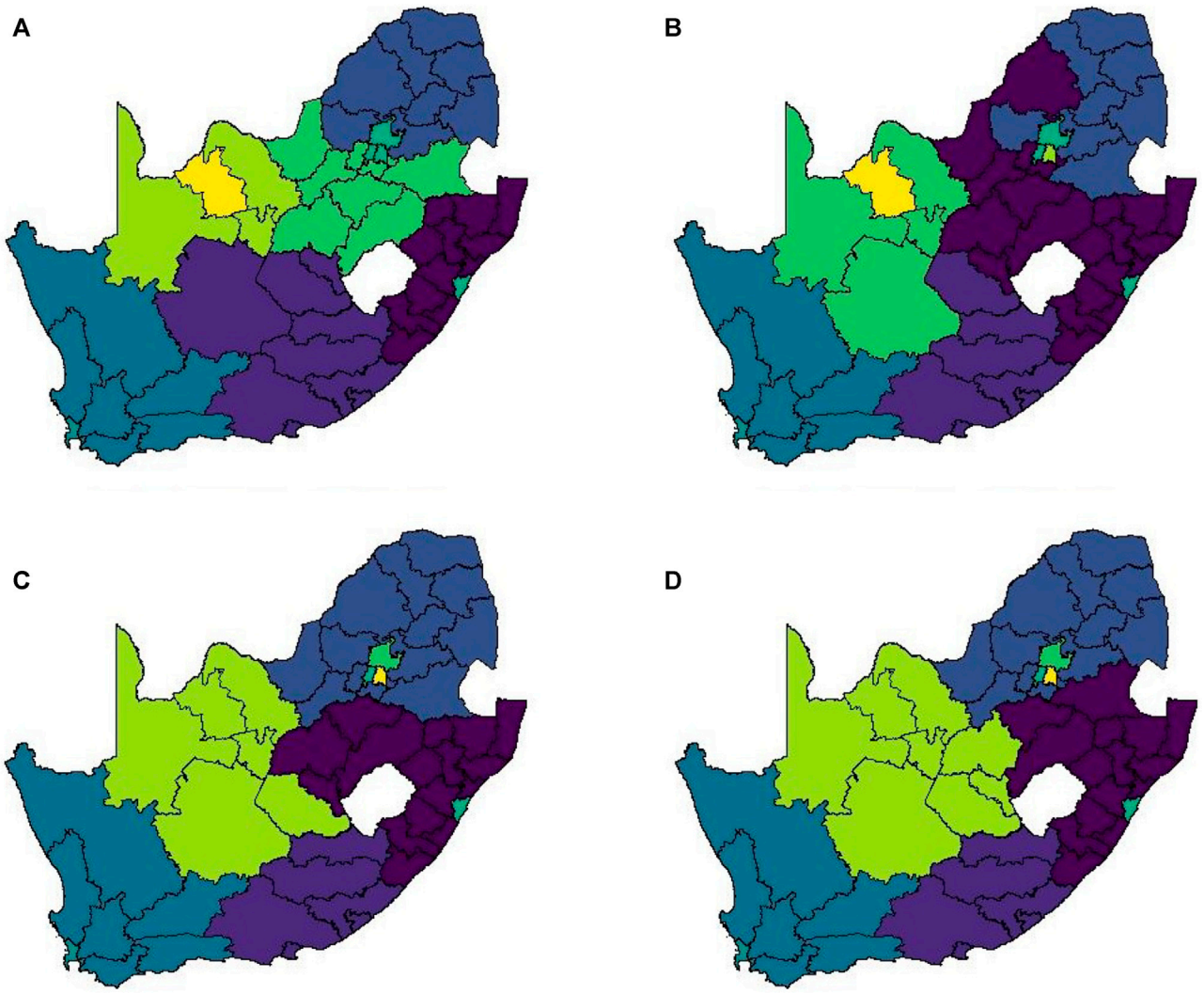


FIGURE 10 | Method 3 complete linkage clustering results (9 clusters) **(A)** Business as usual, **(B)** Level 5, **(C)** Level 4, and **(D)** Level 3.

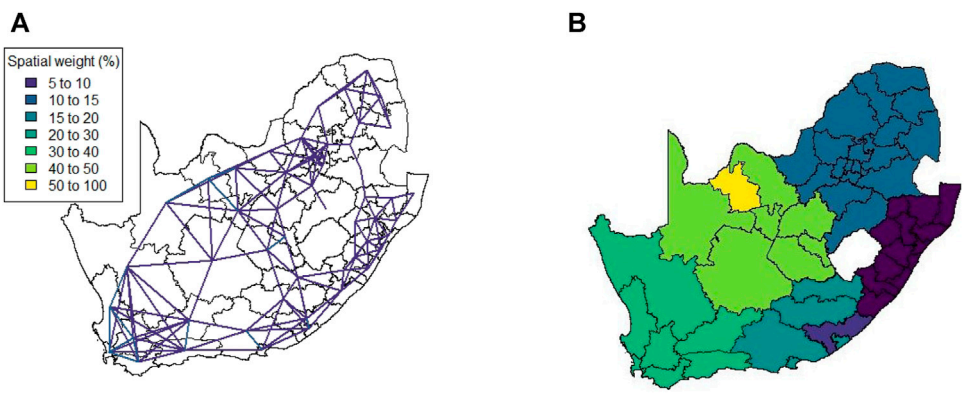


FIGURE 11 | Method 4 **(A)** Spatial weights (weights ≤5 not shown), **(B)** Complete linkage clustering (7 clusters indicated by colours).

TABLE 4 | Spatial weight matrices comparison.

Spatial weight matrix	*Pro*	*Con*	*Interpretation/Contribution*
Method 1 - Distance	Simple to construct and understand Used often in literature	Less realistic Inadequate for clustering Lacks temporal element	Convenient to use and easy to understand and interpret. Not realistic enough for real insight
Method 2 - Mobile network	High spatial resolution Large amounts generated passively by mobile device users	Computationally expensive Difficult to obtain Not representative Privacy concerns	Captures strong spatial associations over relatively short distances. Allows for the identification of patterns potentially missed by other methods
Method 3 - Weighted Facebook data	Freely available data Potentially more representative	Low spatial resolution Lacks specificity	Captures association between focal points of human activity regardless of distance
Method 4 - Scaled Facebook data	Simple to construct and understand. Freely available data Potentially more representative	Lacks temporal elements Low spatial resolution	Adds additional information to previously simplistic model. Additional information improves clustering

data at different spatial resolutions in a way that takes advantage of the strengths of each dataset.

The computational aspects of dealing with mobility data should not be overlooked. Spatial weight matrices can become very large, depending on the number of spatial regions under consideration. Herein the matrices were not sparse, meaning that sparse representations could not be used. Sparse representations could be investigated for high spatial resolution modelling.

To quantify the similarity between the different spatial weight matrices, one might consider the use of simple parametric measures of correlation such as Pearson's correlation coefficient. However, given that there are a total of 52 spatial units (at a district municipality level) and the weights between many spatial unit pairs are negligible, the spatial weight matrices can be regarded as zero-inflated. In addition to making no allowance for the spatial nature of the data, namely the spatial dependency, standard measures of correlation would also deliver biased results. Future research could investigate methods for comparison of spatial weight matrices via appropriate correlation calculations or other techniques.

6 CONCLUSION

COVID-19 spreads spatially and thus the importance of mobility data for COVID-19 modeling should not be disregarded. Ideally, the raw data from the mobile network providers and Facebook, if available, could provide individual movements, allowing for accurate construction of spatial weight matrices. This data could be anonymised and shared. However, instead the methods proposed here can be made use of. The use of movement data in epidemiology is becoming an important covariate to include, without which the spread can only be modelled in isolated regions. Social interactions between human beings are unavoidable. Simple spatial weight matrix construction techniques, such as only taking into account distances, are not always ideal when the spatial associations being captured are dependent on covariates which are not only proximity based. This is made clear by the observed poor performance of Method 1 when it was used as the basis of clustering. The methods presented herein and the results shown also enable epidemiological modellers in considering how to incorporate spatial relationships in models. This is seldom done due to limited mobility information as well as modelling complexities it introduces. However, the improved accuracy in model outcomes will ultimately balance out computational complexities. The paper provides insights into mobility data availability, representability as well as construction for use in spatial modelling. Future research should investigate estimation to a higher spatial resolution using multiple data sources as well as the effect of spatial resolution in spatial epidemiological modelling.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The mobile network data used in this study is not directly available without approval, so cannot be shared directly with the paper. Requests to access these datasets should be directed to https://research.fb.com/blog/2020/06/protecting-privacy-in-facebook-mobility-data-during-the-covid-19-response.

AUTHOR CONTRIBUTIONS

Conceptualisation: All; Data Curation: AP, IF-R, ZK, PD; Formal Analysis: AP, IF-R; Funding Acquisition: PD; Investigation: AP, IF-R; Methodology: AP, IF-R; Project Administration: ZK, PD; Writing -original draft: AP, IF-R; Writing -review editing: All.

FUNDING

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and

are not necessarily to be attributed to the NRF. This research is also funded by Canada's International Development Research Centre (IDRC) (Grant No. 109559-001). The use of the Centre for High Performance Computing (www.chpc.ac.za) also made this work possible.

ACKNOWLEDGMENTS

We thank Gerbrand Mans from the CSIR for data aggregation assistance as well a Bruce Medallo from WITS for mobility data advice.

REFERENCES

Aldstadt, J., and Getis, A. (2006). Using AMOEBA to Create a Spatial Weights Matrix and Identify Spatial Clusters. *Geographical Anal.* 38, 327–343. doi:10.1111/j.1538-4632.2006.00689.x

Anselin, L. (2013). *Spatial Econometrics: Methods and Models, Vol. 4*. Springer Science & Business Media.

Asgari, F., Gauthier, V., and Becker, M. (2013). *A Survey on Human Mobility and its Applications*. arXiv preprint arXiv:1307.0814.

Ballas, D., Clarke, G., Dorling, D., Eyre, H., Thomas, B., and Rossiter, D. (2005). Simbritain: a Spatial Microsimulation Approach to Population Dynamics. *Popul. Space Place* 11, 13–34. doi:10.1002/psp.351

Bavaud, F. (1998). Models for Spatial Weights: a Systematic Look. *Geographical Anal.* 30, 153–171. doi:10.1111/j.1538-4632.1998.tb00394.x

Bengtsson, L., Gaudart, J., Lu, X., Moore, S., Wetter, E., Sallah, K., et al. (2015). Using mobile Phone Data to Predict the Spatial Spread of Cholera. *Sci. Rep.* 5, 8923–8925. doi:10.1038/srep08923

Brown, G. D., Oleson, J. J., and Porter, A. T. (2016). An Empirically Adjusted Approach to Reproductive Number Estimation for Stochastic Compartmental Models: A Case Study of Two Ebola Outbreaks. *Biom* 72, 335–343. doi:10.1111/biom.12432

Brown, G. D., Porter, A. T., Oleson, J. J., and Hinman, J. A. (2018). Approximate Bayesian Computation for Spatial SEIR(S) Epidemic Models. *Spat. Spatio-Temporal Epidemiol.* 24, 27–37. doi:10.1016/j.sste.2017.11.001

Cummings, D. A. T., Irizarry, R. A., Huang, N. E., Endy, T. P., Nisalak, A., Ungchusak, K., et al. (2004). Travelling Waves in the Occurrence of Dengue Haemorrhagic Fever in Thailand. *Nature* 427, 344–347. doi:10.1038/nature02225

Ejigu, B. A., and Wencheko, E. (2020). Introducing Covariate Dependent Weighting Matrices in Fitting Autoregressive Models and Measuring Spatio-Environmental Autocorrelation. *Spat. Stat.* 38, 100454. doi:10.1016/j.spasta.2020.100454

Ekong, I., Chukwu, E., and Chukwu, M. (2020). Covid-19 mobile Positioning Data Contact Tracing and Patient Privacy Regulations: Exploratory Search of Global Response Strategies and the Use of Digital Tools in Nigeria. *JMIR Mhealth Uhealth* 8, e19139. doi:10.2196/19139

Finger, F., Genolet, T., Mari, L., de Magny, G. C., Manga, N. M., Rinaldo, A., et al. (2016). Mobile Phone Data Highlights the Role of Mass Gatherings in the Spreading of Cholera Outbreaks. *Proc. Natl. Acad. Sci. USA* 113, 6421–6426. doi:10.1073/pnas.1522305113

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning, Vol. 1*. New York: Springer series in statistics.

Gao, S., Rao, J., Kang, Y., Liang, Y., Kruse, J., Dopfer, D., et al. (2020). Association of mobile Phone Location Data Indications of Travel and Stay-At-home Mandates with COVID-19 Infection Rates in the US. *JAMA Netw. Open* 3, e2020485. doi:10.1001/jamanetworkopen.2020.20485

Gao, Y., Li, T., Wang, S., Jeong, M.-H., and Soltani, K. (2018). A Multidimensional Spatial Scan Statistics Approach to Movement Pattern Comparison. *Int. J. Geographical Inf. Sci.* 32, 1304–1325. doi:10.1080/13658816.2018.1426859

Garrison, W. L., and Marble, D. F. (1964). Factor-analytic Study of the Connectivity of a Transportation Network. *Pap. Reg. Sci. Assoc.* 12, 231–238. doi:10.1007/bf01941256

Getis, A., and Aldstadt, J. (2004). Constructing the Spatial Weights Matrix Using a Local Statistic. *Geographical Anal.* 36, 90–104. doi:10.1111/j.1538-4632.2004.tb01127.x

Grantz, K. H., Meredith, H. R., Cummings, D. A. T., Metcalf, C. J. E., Grenfell, B. T., Giles, J. R., et al. (2020). The Use of mobile Phone Data to Inform Analysis of COVID-19 Pandemic Epidemiology. *Nat. Commun.* 11, 4961–4968. doi:10.1038/s41467-020-18190-5

Huang, R., Liu, M., and Ding, Y. (2020). Spatial-temporal Distribution of Covid-19 in China and its Prediction: A Data-Driven Modeling Analysis. *J. Infect. Dev. Ctries* 14, 246–253. doi:10.3855/jidc.12585

Huang, X., Li, Z., Jiang, Y., Ye, X., Deng, C., Zhang, J., et al. (2021). The Characteristics of Multi-Source Mobility Datasets and How They Reveal the Luxury Nature of Social Distancing in the U.S. During the COVID-19 Pandemic. *Int. J. Digital Earth* 14, 424–442. doi:10.1080/17538947.2021.1886358

Jin, C., Nara, A., Yang, J. A., and Tsou, M. H. (2020). Similarity Measurement on Human Mobility Data with Spatially Weighted Structural Similarity index (SpSSIM). *Trans. GIS* 24, 104–122. doi:10.1111/tgis.12590

Malik, R., Deardon, R., and Kwong, G. P. S. (2016). Parameterizing Spatial Models of Infectious Disease Transmission that Incorporate Infection Time Uncertainty Using Sampling-Based Likelihood Approximations. *PLoS One* 11, e0146253. doi:10.1371/journal.pone.0146253

Oliver, N., Lepri, B., Sterly, H., Lambiotte, R., Deletaille, S., De Nadai, M., et al. (2020). Mobile Phone Data for Informing Public Health Actions across the COVID-19 Pandemic Life Cycle. *Sci. Adv.* 6 (23), eabc0764. doi:10.1126/sciadv.abc0764

Peixoto, P. S., Marcondes, D., Peixoto, C., and Oliva, S. M. (2020). Modeling Future Spread of Infections via mobile Geolocation Data and Population Dynamics. An Application to COVID-19 in Brazil. *PLoS One* 15, e0235732. doi:10.1371/journal.pone.0235732

Pfeffermann, D. (2013). New Important Developments in Small Area Estimation. *Stat. Sci.* 28, 40–68. doi:10.1214/12-sts395

Ruktanonchai, N. W., DeLeenheer, P., Tatem, A. J., Alegana, V. A., Caughlin, T. T., zu Erbach-Schoenberg, E., et al. (2016). Identifying Malaria Transmission Foci for Elimination Using Human Mobility Data. *Plos Comput. Biol.* 12, e1004846. doi:10.1371/journal.pcbi.1004846

Sakarovitch, B., Bellefon, M. Pd., Givord, P., and Vanhoof, M. (2018). Estimating the Residential Population from mobile Phone Data, an Initial Exploration. *Economie et Statistique* 505, 109–132. doi:10.24187/ecostat.2018.505d.1968

Stakhovych, S., and Bijmolt, T. H. A. (2009). Specification of Spatial Models: A Simulation Study on Weights Matrices. *Pap. Reg. Sci.* 88, 389–408. doi:10.1111/j.1435-5957.2008.00213.x

Suryowati, K., Bekti, R. D., and Faradila, A. (2018). A Comparison of Weights Matrices on Computation of Dengue Spatial Autocorrelation. *IOP Conf. Ser. Mater. Sci. Eng.* 335, 012052. doi:10.1088/1757-899x/335/1/012052

Tagliazucchi, E., Balenzuela, P., Travizano, M., Mindlin, G. B., and Mininni, P. D. (2020). Lessons from Being Challenged by COVID-19. *Chaos, Solitons & Fractals* 137, 109923. doi:10.1016/j.chaos.2020.109923

Toch, E., Lerner, B., Ben-Zion, E., and Ben-Gal, I. (2019). Analyzing Large-Scale Human Mobility Data: a Survey of Machine Learning Methods and Applications. *Knowl Inf. Syst.* 58, 501–523. doi:10.1007/s10115-018-1186-x

Varsavsky, T., Graham, M. S., Canas, L. S., Ganesh, S., Pujol, J. C., Sudre, C. H., et al. (2021). Detecting COVID-19 Infection Hotspots in England Using Large-Scale Self-Reported Data from a mobile Application: a Prospective, Observational Study. *The Lancet Public Health* 6, e21–e29. doi:10.1016/s2468-2667(20)30269-3

Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., et al. (2012). Quantifying the Impact of Human Mobility on Malaria. *Science* 338, 267–270. doi:10.1126/science.1223467

Wesolowski, A., Qureshi, T., Boni, M. F., Sundsøy, P. R., Johansson, M. A., Rasheed, S. B., et al. (2015). Impact of Human Mobility on the Emergence of Dengue Epidemics in Pakistan. *Proc. Natl. Acad. Sci. USA* 112, 11887–11892. doi:10.1073/pnas.1504964112

Xiong, C., Hu, S., Yang, M., Luo, W., and Zhang, L. (2020). Mobile Device Data Reveal the Dynamics in a Positive Relationship between Human Mobility and COVID-19 Infections. *Proc. Natl. Acad. Sci. USA* 117, 27087–27089. doi:10.1073/pnas.2010836117

Zhou, Y., Lau, B. P. L., Yuen, C., Tunçer, B., and Wilhelm, E. (2018). Understanding Urban Human Mobility through Crowdsensed Data. *IEEE Commun. Mag.* 56, 52–59. doi:10.1109/mcom.2018.1700569

Zhou, Y., Xu, R., Hu, D., Yue, Y., Li, Q., and Xia, J. (2020). Effects of Human Mobility Restrictions on the Spread of COVID-19 in Shenzhen, China: a Modelling Study Using mobile Phone Data. *The Lancet Digital Health* 2, e417. doi:10.1016/s2589-7500(20)30165-5

Conflict of Interest: Author SK-M was employed by IBM, South Africa.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

APPENDIX

Facebook for good data calculation.

Let u represent a single individual and $U_{t,i}$ represent district municipality i at time t. The total number of Bing tiles visited by inhabitants of district municipality i is then

$$\text{total\_tiles}\,(U_{t,i}) = \sum_{u \in U_{t,i}} \min\left(\text{tiles}\,(u), 200\right).$$

Note that the maximum number of Bing tiles visited that a single individual can contribute is restricted to 200. In order to preserve user privacy, an error term was included by drawing from a Laplace distribution with parameters 0 and $\frac{F}{\epsilon}$ where F = sensitivity parameter and ϵ = noise parameter as follows

$$\text{total\_tiles}'\,(U_{t,i}) = \text{total\_tiles}\,(U_{t,i}) + \text{Laplace}\left(0, \frac{F}{\epsilon}\right).$$

The average number of tiles per district municipality was then calculated as

$$\text{avg\_tiles}'\,(U_{t,i}) = \frac{\text{total\_tiles}'\,(U_{t,i})}{|U_{t,i}|}.$$

The mobility value for each district municipality and for each day was then finally expressed with respect to the baseline as

$$F_i^{(t)} = \frac{\text{avg\_tiles}\,(U_{t,i}) - \text{baseline\_avg\_tiles}'\,(i, \text{day\_of\_the\_week}\,(t))}{\text{baseline\_avg\_tiles}'\,(i, \text{day\_of\_the\_week}\,(t))}.$$

For further details regarding this data see https://research.fb.com/blog/2020/06/protecting-privacy-in-facebook-mobility-data-during-the-covid-19-response/.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership