Infect Dis Ther (2022) 11:787–805
https://doi.org/10.1007/s40121-022-00600-4

ORIGINAL RESEARCH

Machine Learning-Based COVID-19 Patients Triage Algorithm Using Patient-Generated Health Data from Nationwide Multicenter Database

Min Sue Park · Hyeontae Jo · Haeun Lee · Se Young Jung · Hyung Ju Hwang

Received: January 5, 2022 / Accepted: January 28, 2022 / Published online: February 16, 2022
© The Author(s) 2022

ABSTRACT

Introduction: A prompt severity assessment model of patients with confirmed infectious diseases could enable efficient diagnosis while alleviating burden on the medical system. This study aims to develop a SARS-CoV-2 severity assessment model and establish a medical system that allows patients to check the severity of their cases and informs them to visit the appropriate clinic center on the basis of past treatment data of other patients with similar severity levels.

Methods: This paper provides the development processes of a severity assessment model using machine learning techniques and its application on SARS-CoV-2-infected patients. The proposed model is trained on a nationwide data set provided by a Korean government agency and only requires patients' basic personal data, allowing them to judge the severity of their own cases. After modeling, the boosting-based decision tree model was selected as the classifier while mortality rate was interpreted as the probability score. The data set was collected from all Korean citizens with confirmed COVID-19 between February 2020 and July 2021 (N = 149,471).

Min Sue Park and Hyeontae Jo contributed equally and are co-first authors.

M. S. Park · H. J. Hwang
Department of Mathematics, Pohang University of Science and Technology, Pohang, Republic of Korea

H. Jo
Basic Science Research Institute, Pohang University of Science and Technology, Pohang, Republic of Korea

H. Lee · S. Y. Jung
Office of eHealth Research and Business, Seoul National University Bundang Hospital, Seongnam-si, Republic of Korea

S. Y. Jung
Department of Family Medicine, Seoul National University Bundang Hospital, Seongnam-si, Republic of Korea

H. J. Hwang
Graduate School of Artificial Intelligence, Pohang University of Science and Technology, Pohang, Republic of Korea

H. J. Hwang (✉)
Math Building, Room 317, 77, Cheongam-ro, Nam-gu, Pohang-si, Gyeongsangbuk-do, Republic of Korea
e-mail: hjhwang@postech.ac.kr

S. Y. Jung (✉)
Department of Digital Healthcare, Seoul National University Bundang Hospital, 172 Dolma-ro, Bundang-gu, Seongnam-si 13620, Republic of Korea
e-mail: imsyjung@gmail.com

Results: The experiments achieved high model performance with an approximate precision of 0.923 and area under the curve of receiver operating characteristic (AUROC) score of 0.950 [95% tolerance interval (TI) 0.940–0.958, 95% confidence interval (CI) 0.949–0.950]. Moreover, our experiments identified the most important variables affecting the severity in the model via sensitivity analysis.
Conclusion: A prompt severity assessment model for managing infectious people has been attained through using a nationwide data set. It has demonstrated its superior performance by surpassing that of conventional risk assessments. With the model's high performance and easily accessible features, the triage algorithm is expected to be particularly useful when patients monitor their health status by themselves through smartphone applications.

Keywords: Machine learning; Deep learning; COVID-19; Triage protocol; Mortality; SARS-CoV-2

Key Summary Points

Why carry out this study?

Traditional risk prediction models are limited to identifying the condition of an asymptomatic patient who deteriorates from mild to moderate or extremely severe risk of COVID-19 at triage

Existing disease risk assessment models were developed with limited size data sets, input variables, and unstandardized independent features without specific machine learning algorithms

What was learned from the study?

This prediction model, trained with patient-generated health data (PGHD) from nationwide COVID-19 screening centers, can be globally utilized to monitor hospitalized or quarantined patients with confirmed SARS-CoV-2 infection daily

This risk assessment model, developed with multivariable factors like demographic, geographic, and clinical characteristics of a superior performance, can be successfully deployed to triage patients with COVID-19

INTRODUCTION

Countries such as the UK, Singapore, Germany, Portugal, and Israel—with high vaccination rates—have created strategies for the new normal after COVID-19 [1–3] as many are resuming their pre-COVID-19 lives. However, as the coronavirus mutations cause breakthrough infections, the current vaccine has little effect on reducing the transmission of the virus. The number of confirmed cases in the UK and Singapore has been increasing since October 2021 [4]. The variants put a great burden on the healthcare system of those countries [5]. Thus, it is evermore imperative to ensure medical readiness at a national level by preparing accurate and reasonable patient severity classification criteria and procedures [6].

Over the past year and 10 months, South Korea has experienced four COVID-19 outbreaks, and the occurrence of confirmed cases has been suppressed through the 3T strategy (test, confirmation, investigation, tracking, treatment) and adjustment of social distancing without border blocking and regional blockade [7]. According to the Organization for Economic Cooperation and Development, South Korea has achieved quarantine results without any containment measures, minimizing economic damage, and most effectively blocking the spread of the virus [8]. Although South Korea has been performing relatively well in controlling COVID-19, it had difficulty in managing patients whose clinical condition deteriorated from mild to modulate risk level. In fact, there have been cases where patients died at home or a community treatment center, a facility for isolating asymptomatic and mildly

symptomatic patients with COVID-19, as a result of delayed response [9, 10].

Thus, a risk prediction model that accurately identifies the condition of a patient who deteriorates from mild to moderate or severe risk is required. Furthermore, it is crucial to triage patients with COVID-19 on the basis of the severity of their infection to secure the entire medical system of a nation. For the self-quarantining population of COVID-19, accurate severity assessment tools are necessary to appraise health status every day [11, 12]. Several models have been developed to predict the prognosis of patients with confirmed COVID-19 or the possibility of COVID-19 diagnosis of patients before confirmation. However, there were several problems: (1) the size of the research data sets was too small, (2) the number of input variables was limited, (3) the non-standard variables were difficult to use by other institutions, or (4) the specific method of using the model was not presented. Moreover, to the best of our knowledge, there was no study on the mortality rate of SARS-CoV-2 according to symptoms at national level while there have been several studies conducted on the establishment of a model for predicting COVID-19 confirmation based on nationwide data set with features related to COVID-19. Preventing the spread of COVID-19 has difficult aspects such as requiring not only medical staff but also national action. In contrast, lowering the mortality rate can be effectively managed by medical staff by developing an appropriate triage protocol.

Thus, this study aims to review previous research of prediction models for COVID-19 and develop a model predicting mortality rate of SARS-CoV-2 using nationwide multicenter data, thereby allowing patients to easily predict the severity of COVID-19 by entering their patient-generated health data (PGHD) during quarantine out of hospital.

METHODS

Review of Previous Research

The review of previous research was based on a search of three databases: Google Scholar, PubMed, and medRxiv. The following keywords were searched in combination: severity, machine learning, deep learning, COVID-19, triage protocol, mortality, and SARS-CoV-2.

In this paper, we propose a machine learning model that predicts the mortality of SARS-CoV-2 based on questionnaires completed by patients. This research was approved by the Institutional Review Board of Seoul National University Bundang Hospital (X-2110-717-902). An Informed consent form was not obtained owing to the nature of retrospective studies. The study was performed in accordance with the Helsinki Declaration of 1964 and its later amendments.

Data Source and Study Cohort

The data set was collected from February 2020 to July 2021 by the Korea Disease Control and Prevention Agency (KDCA), a government-affiliated organization, for all Koreans who tested positive for SARS-CoV-2 in polymerase chain reaction (PCR). Our study was approved by the Institutional Review Board of Seoul National University Bundang Hospital (X-2110-717-902). The data set consists of 149,471 patients who tested positive, of whom 2000 died. The data set is labeled according to whether the patient is dead or alive, and it is highly imbalanced (98.7% imbalance ratio). The data set are mainly composed of three types of patient data: (i) basic personal information, (ii) types of first symptoms, and (iii) underlying diseases. A detailed description of these features is given in Tables 1 and 2. As mentioned in the "Introduction", the area of residence is included in the data feature because it affects the degree of virus activation and medicalization scale.

The data was collected from 1382 designated COVID-19 screening centers in South Korea. These centers consist of national safe hospitals (263), dedicated respiratory clinics (518),

Table 1 Baseline characteristics of input features

Type	Variables	*N* (total = 149,471)	%
Basic information	Sex		
	Male	75,073	50.23
	Female	74,398	49.77
	Age	Mean = 44.36	
		(std = 20.27)	
	Area of residence		
	Latitude	Mean = 36.93	
		(std = 0.93)	
	Longitude	Mean = 127.39	
		(std = 0.76)	
	Body temperature (T, °C)		
	$T \leq 36.5$	121,557	81.32
	$36.5 < T < 37.5$	6310	4.22
	$37.5 \leq T < 38.3$	17,227	11.53
	$T \geq 38.3$	4377	2.93
Respiratory symptom	Cough		
	True	34,201	22.88
	False	99,997	66.90
	Sputum		
	True	17,108	11.45
	False	117,090	78.34
	Sore throat		
	True	25,078	16.78
	False	109,120	73.00
	Dyspnea		
	True	1962	1.31
	False	132,236	88.47

Table 1 continued

Type	Variables	*N* (total = 149,471)	%
Non-respiratory symptom	Musculoskeletal pain		
	True	24,017	16.07
	False	110,181	73.71
	Headache		
	True	16,337	10.93
	False	117,861	78.85
	Chill		
	True	17,227	11.53
	False	116,971	78.26
	Ageusia		
	True	4846	3.24
	False	129,352	86.54
	Anosmia		
	True	5498	3.68
	False	128,700	86.10

screening clinics in public health centers (627), temporary screening offices (200), and car mobile screening clinics (15). The process of initial screening, transfer, admission to a hospital or community treatment center (CTC) is presented in Fig. 1.

The triage process of patients with confirmed COVID-19 was initiated on the basis of the severity of their symptoms: asymptomatic to mild, moderate, severe, and critical. Symptoms were assessed by telephone interviews or face-to-face in the first-visit facility, and patients were quarantined at designated facilities according to their severity. Asymptomatic and mildly symptomatic patients were admitted to CTCs. Meanwhile, patients with an aggravated severity were hospitalized at tertiary hospitals. The referral system at each level of medical care aims to allow for patients to be efficiently transferred to a higher level of care before worsening clinical status [13].

The overall process of hospitalization and transfer is presented in Fig. 1.

Data Collection and Measurement

Previous studies revealed that the outbreaks of COVID-19 were associated with latitude, temperature, and humidity measurements, which reflects seasonal variation in the incidence of respiratory viruses [14, 15]. Thus, geographic information of latitude and longitude have been integrated into our model.

Easy-to-measure features are defined as variables such as body temperature, pulse rate, respiratory rate, blood pressure, any symptoms, and past medical history that can be directly collected from patients without much delay.

Table 2 Underlying diseases of study participants

Disease	Count	Total (N = 149,471)	
		N	%
Liver disease[a]	0	148,632	99.44
	1	354	0.24
	2	475	0.32
	3	10	0.01
Cancer[b]	0	147,260	98.52
	1	594	0.4
	2	1423	0.95
	3	187	0.13
	4	5	0.00
	5	2	0.00
Diabetes mellitus	0	139,063	93.04
	1	10,408	6.96
Cardio-cerebrovascular disease[c]	0	127,608	85.37
	1	2165	1.45
	2	18,719	12.52
	3	825	0.55
	4	139	0.09
	5	15	0.01
Renal disease[d]	0	148,698	99.48
	1	758	0.51
	2	15	0.01
Degenerative disease[e]	0	146,945	98.31
	1	2331	1.56
	2	193	0.13
	3	2	0.00

Table 2 continued

Disease	Count	Total (N = 149,471)	
		N	%
Lung disease[f]	0	147,253	98.52
	1	2086	1.40
	2	122	0.08
	3	10	0.01

[a]Liver disease includes hepatitis B, cirrhosis, and any other hepatitis
[b]Cancer includes liver cancer, thyroid cancer, oral cancer, acute myelogenous white blood, ovarian cancer, brain cancer, colon cancer, lymphoma, chronic myelogenous white blood, bladder cancer, esophageal cancer, cancer, stomach cancer, cervical cancer, uterine cancer, prostate cancer, rectal cancer, skin cancer, hematoma, laryngeal cancer, prostate cancer, hematologic cancer, hematoma, and blood cancer
[c]Cardio-cerebrovascular disease includes hypertension, stroke, cerebral infarction, myocardial infarction, myocardial hemorrhage, arteriosclerosis, and angina
[d]Renal disease includes renal failure, renal failure, and glomerular disease
[e]Degenerative diseases include Alzheimer disease, other dementia, and Parkinson disease
[f]Lung disease includes emphysema and any other lung disease

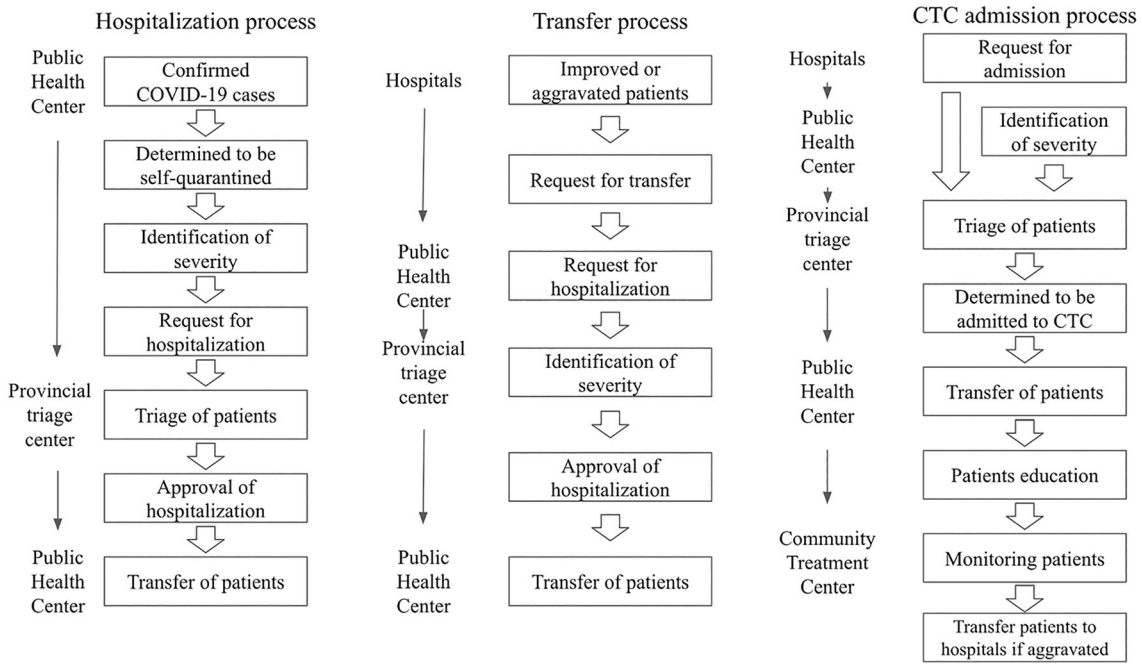

Fig. 1 Management strategy of COVID-19 confirmed cases in South Korea

Outcome Definition

The outcome was defined as deceased cases due to COVID-19 in hospitals, CTCs, and at homes. The mortality cases were collected by the KDCA from national statistics.

Feature Generation

We observed that the structural stability of individual SARS-CoV-2 virus-like particles could be affected by the temperature and humidity of the atmosphere [16]. In addition, hospitalization rates may vary depending on access to medical resources and the severity of previous diseases [17]. For these reasons, we utilized additional features such as the date of the onset of symptoms (in months), the area of residence (in longitude and latitude coordinates), and underlying patient symptoms.

The features of the data set provided by KDCA as follows: sex, age, body temperature, clinical symptoms (cough, sputum, sore throat, dyspnea, musculoskeletal pain, headache, chill, ageusia, anosmia), self-reported underlying diseases. For body temperature (T), we divided patients and categorized them into four subgroups: (1) no fever with $T \leq 36.5$ °C, (2) mild elevation of body temperature with $36.5\ °C < T < 37.5\ °C$, (3) mild fever with $37.5\ °C \leq T < 38.3\ °C$, and (4) overt fever with $T \geq 38.3$ °C. All clinical symptoms have binary values: true or false. Since underlying diseases are self-reported in a free format, we manually classified the reported diseases into seven subgroups: liver disease, cancer, diabetes mellitus, cardio-cerebrovascular disease, renal disease, degenerative disease, and lung disease. Thus, if a patient had lung cancer and liver cancer, they were assigned a value of two to the feature named "cancer" for this patient. This was done to reduce the sparsity of our data set. Since there are so many different diseases, our data set would become very sparse if we treated each disease as a different feature. If a model is naively trained on a given sparse data set, the performance of the model would degenerate; worse still, it could increase the chances of the model wrongly predicting the mortality probability for a patient with a rare disease. Moreover, requiring many features would lower user convenience.

Training and Evaluation

We split the data set into training sets and test sets with an 80:20 ratio, and the model was evaluated on the test set. We used a tree-based gradient boosting machine learning model with binary logistic objectives, XGBoost (XGB) [18]. This model is a decision-tree-based ensemble machine learning model known for its powerful performance in classification problems in various fields [19, 20]. Since this is a tree-based model, it has the advantage of being able to process data with missing values [21]. Another benefit of using gradient boosting algorithms is that they enable straightforward measurement of feature importance scores in prediction by calculating how useful each feature is in the construction of the weak learners within the model. Therefore, this method does not tell us how positively or negatively the features affected the prediction and does not consider the association relations among features in making predictions.

Meanwhile, originating from game theory, the SHapley Additive exPlanations (SHAP) algorithm [22] is used to compute Shapley values [23] for each feature, where each Shapley value represents the impact of the feature to which it is associated and predicted. When used for tree-based models, SHAP has the great advantage of being able to calculate Shapley values relatively quickly. Therefore, we have utilized it to identify the principal features in model prediction.

The model was evaluated on the test set using various metrics, including area under the curve of receiver operating characteristic (AUROC), area under the precision–recall curve (AUPRC), F1 score, precision, sensitivity, and specificity. Moreover, we performed a decision curve analysis on the model. ROC analysis provides information about diagnostic test performance; a ROC curve consists of the true positive (TP) and false positive (FP) rates and demonstrates the discriminatory ability of a

binary classifier system by varying the discriminant thresholds. In other words, the discriminatory ability of the test could be powerful when the vertex of the curve is closer to the upper left (high TP rate and low FP rate). In addition, the baseline for AUROC is always 0.5.

On the other hand, PR curves plot the precision against the recall, and AUPRC is especially useful for imbalanced data in a setting where we focus more on detecting the positive examples. Unlike AUROC, the baseline for AUPRC is equal to the fraction of positives. This means that obtaining an AUPRC of 0.4 on a class with 10% positives is good but obtaining an AUPRC of 0.6 on a class with 80% positives is undesirable [24].

RESULTS

Literature Review

Previous research was classified according to the five classification criteria: (1) type of learning data, (2) type of prediction models, (3) outcome variables, (4) data type, and (5) whether or not easy-to-measure input features were utilized. In terms of modeling and utilizing the prediction models, they have four major components: gathering patients' information such as symptoms, signs, previous medical history; results of imaging studies; and laboratory tests; confirmation of COVID-19 through reverse transcriptase polymerase chain reaction (RT-PCR) test; and triage of confirmed cases. The schematic flow of management for patients with COVID-19 is presented in Fig. 2.

In terms of outcome variables, previous studies were classified into four major classes.

Outcome class 1: diagnosis.

A + B ⇒ 1 (Zoabi, Menni) [25, 26].
B ⇒ 1 (Yanamala) [27].
D ⇒ 1 (Gozes, Song, Jin, Punn) [28–31].
A + B + C + E + 2a ⇒ 1 (Feng) [32].

Outcome class 2: mortality.

F + 1 + 2a + 2b + 2c ⇒ 3a vs 3b (Cifuentes) [33].
A + B + C + E + 1 + 2a ⇒ 3a vs 3b (Her) [34].
C + 1 + 2a + 2b + 2c ⇒ 3a vs 3b (Cho) [35].
C + E + 1 + 2a ⇒ 3a vs 3b (Ikemura) [36].

Outcome class 3: mortality and complication.

B + D + E + 1 + 2a ⇒ (3a + 3c) vs 3b (Shamout) [37].

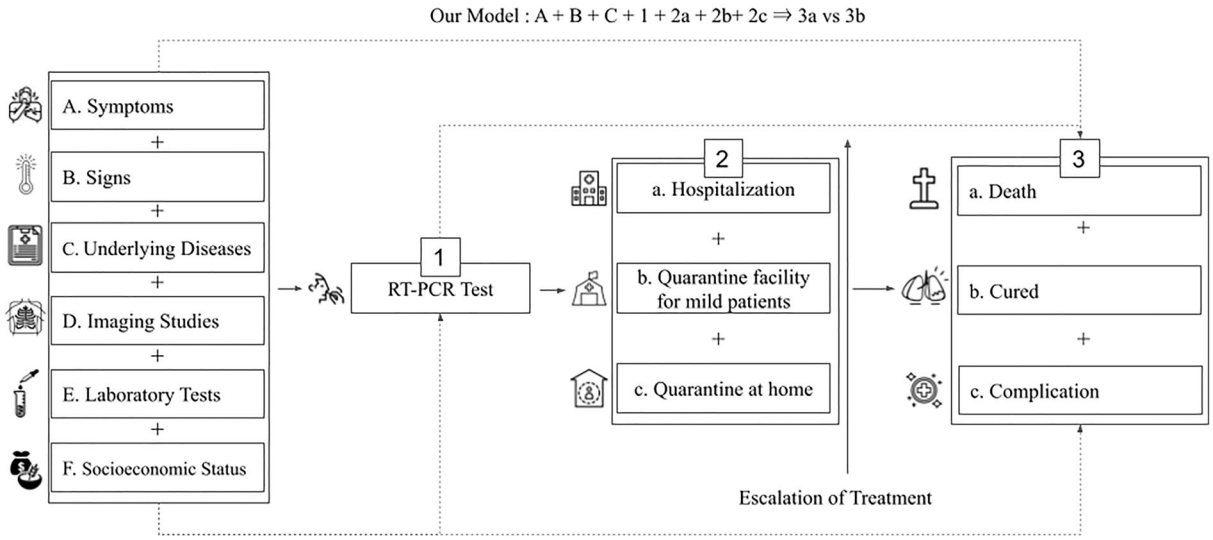

Fig. 2 Classification of the previous prediction models according to the type of learning data and type of prediction models

Table 3 Previous research regarding COVID-19 prediction models

Class	Studies	Prediction type	Outcome variable	Data type	Sample size	Easy-to-measure input features
	Our model	Prognosis	Mortality	Nationwide	149,471	Yes
1	Zoabi et al. [25]	Diagnosis	RT-PCR	Nationwide	99,232	Yes
	Yanamala et al. [27]	Diagnosis	RT-PCR	Local	3883	No
	Gozes et al. [28]	Diagnosis	RT-PCR	Local	157	No
	Song et al. [29]	Diagnosis	RT-PCR	Local	275	No
	Feng et al. [32]	Diagnosis	RT-PCR	Local	164	No
	Jin et al. [30]	Diagnosis	RT-PCR	Local	11,356	No
	Punn et al. [31]	Diagnosis	RT-PCR	Local	1214	No
	Menni et al. [26]	Diagnosis	RT-PCR	Nationwide	2,618,862	Yes
2	Cifuentes et al. [33]	Prognosis	Mortality	Nationwide	1,033,218	Yes
	Cho et al. [35]	Prognosis	Mortality	Nationwide	7590	No
	Ikemura et al. [36]	Prognosis	Mortality	Local	4313	No
	Her et al. [34]	Prognosis	Mortality	Nationwide	5628	No
3	Subudhi et al. [38]	Prognosis	Complication or mortality	Local	10,826	No
	Shamout et al. [37]	Prognosis	Complication or mortality	Local	3661	No
	Marcos et al. [39]	Prognosis	Complication or mortality	Local	1270	No
	Kim et al. [40]	Prognosis	Complication or mortality	Nationwide	4787	Yes
	Su et al. [41]	Prognosis	Complication or mortality	Local	14,418	No
4	Rinderknecht et al. [42]	Prognosis	Complication	Nationwide	15,753	Yes
	Wang et al. [43]	Prognosis	Complication	Local	3008	No

RT-PCR reverse transcription polymerase chain reaction

C + E + 1 + 2a ⇒ (3a + 3c) vs 3b (Subudhi) [38].
A + B + C + E + 1 + 2a ⇒ (3a + 3c) vs 3b (Marcos) [39].
A + B + C + 1 + 2a ⇒ (3a + 3c) vs 3b (Kim) [40].
C + E + 1 + 2a ⇒ (3a + 3c) vs 3b (Su) [41].

Outcome class 4: complication.

A + B + C + 1 + 2a + 2b + 2c ⇒ 3b vs 3c (Rinderknecht) [42].
A + B + C + D + E + 1 + 2a ⇒ 3b vs 3c (Wang) [43].

We reviewed 19 existing studies and classified them by the four classification criteria into the four major outcome classes. The result is presented in Table 3.

The baseline characteristics of the input features used in the research are presented in Tables 1 and 2. The area of residence for each confirmed patient was converted to floating-point variables using the Python Google Maps API client owing to its large scale.

The distribution of longitude and latitude of the study participants is presented in Fig. 3. The *x*-axis represents the latitude–longitude coordinate, while the *y*-axis shows its number of

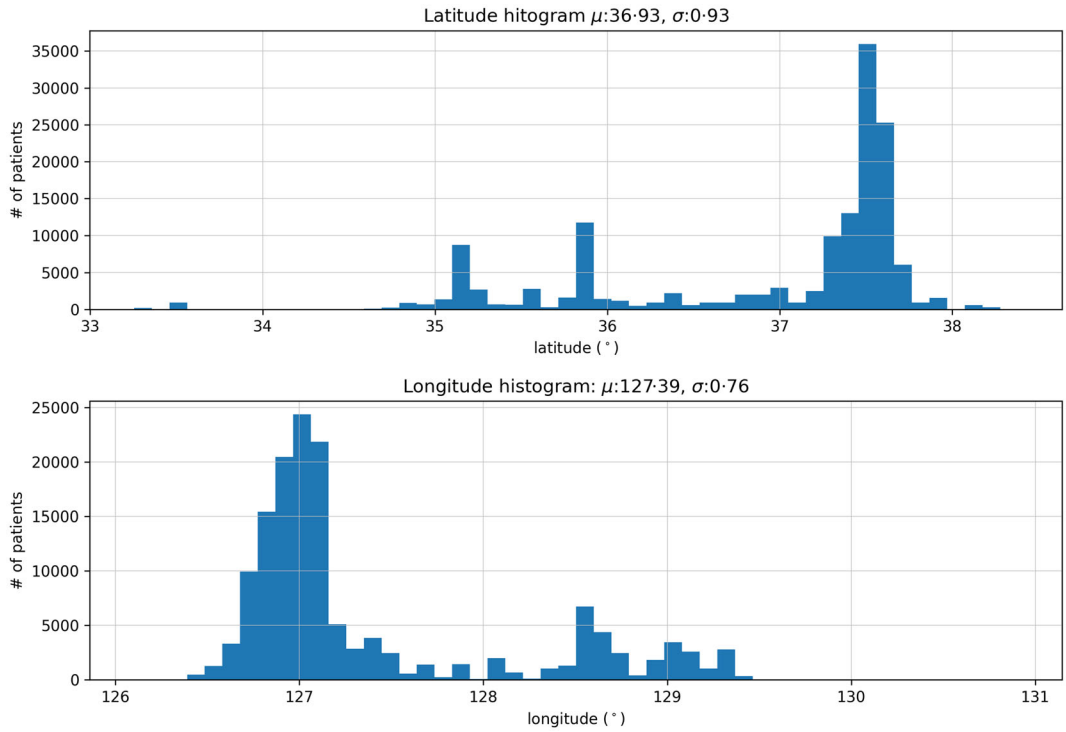

Fig. 3 Histogram of patients' distribution by latitude (top) and longitude (bottom)

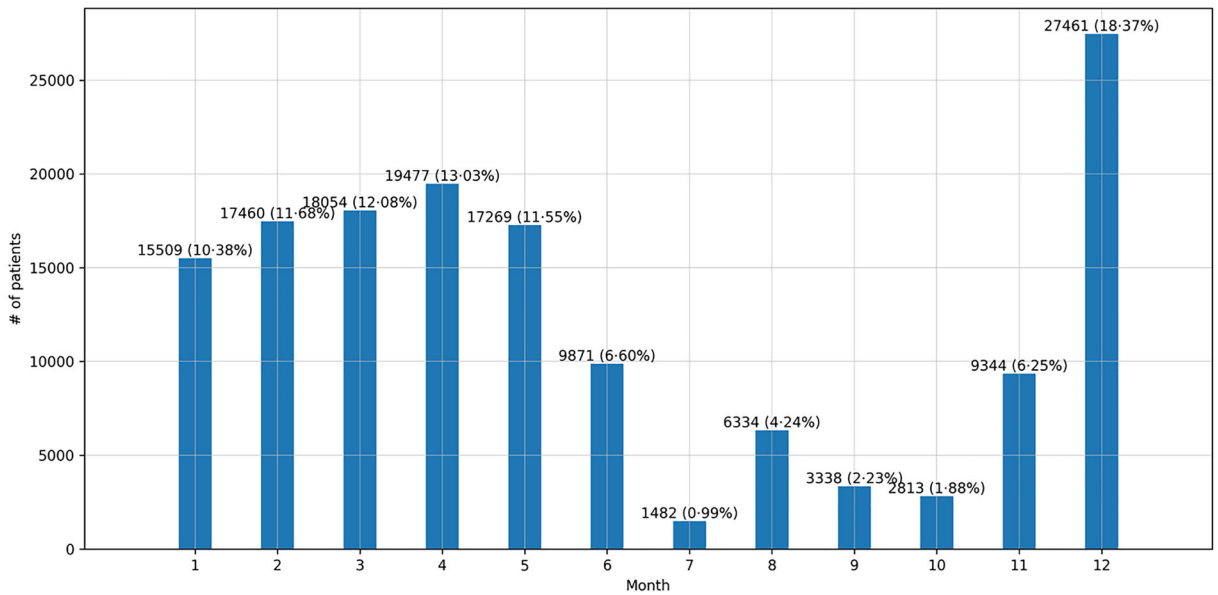

Fig. 4 Cumulative number of confirmed cases per month

patients. The μ and σ in the title denote the mean and the standard deviation, respectively. Even though discrepancies between the actual area of residence and latitude–longitude pair exist, they were ignored because such cases were rare.

The seasonality of the cumulative number of confirmed cases per month is presented in Fig. 4. The height of each bar represents the number of patients in that month. We marked the number of patients and their percentage (%) at the top of the bar.

Model Performance

The proposed model achieved an AUROC score of 0.950 at a 95% tolerance interval (TI) 0.940–0.958 and 95% confidence interval (CI) 0.949–0.950, Youden's index of 0.739, F1 score of 0.861, recall 0.807, precision 0.923, and specificity 0.933. Since the size of the test set was 29,895, and there were 398 positives in the test set, the fraction of positives is 0.013, which is the baseline for the AUPRC score. The model achieved an AUPRC score of 0.268 (with 95% TI 0.225–0.310 and 95% CI 0.266–0.269), greatly outperforming the baseline score of 0.013. The

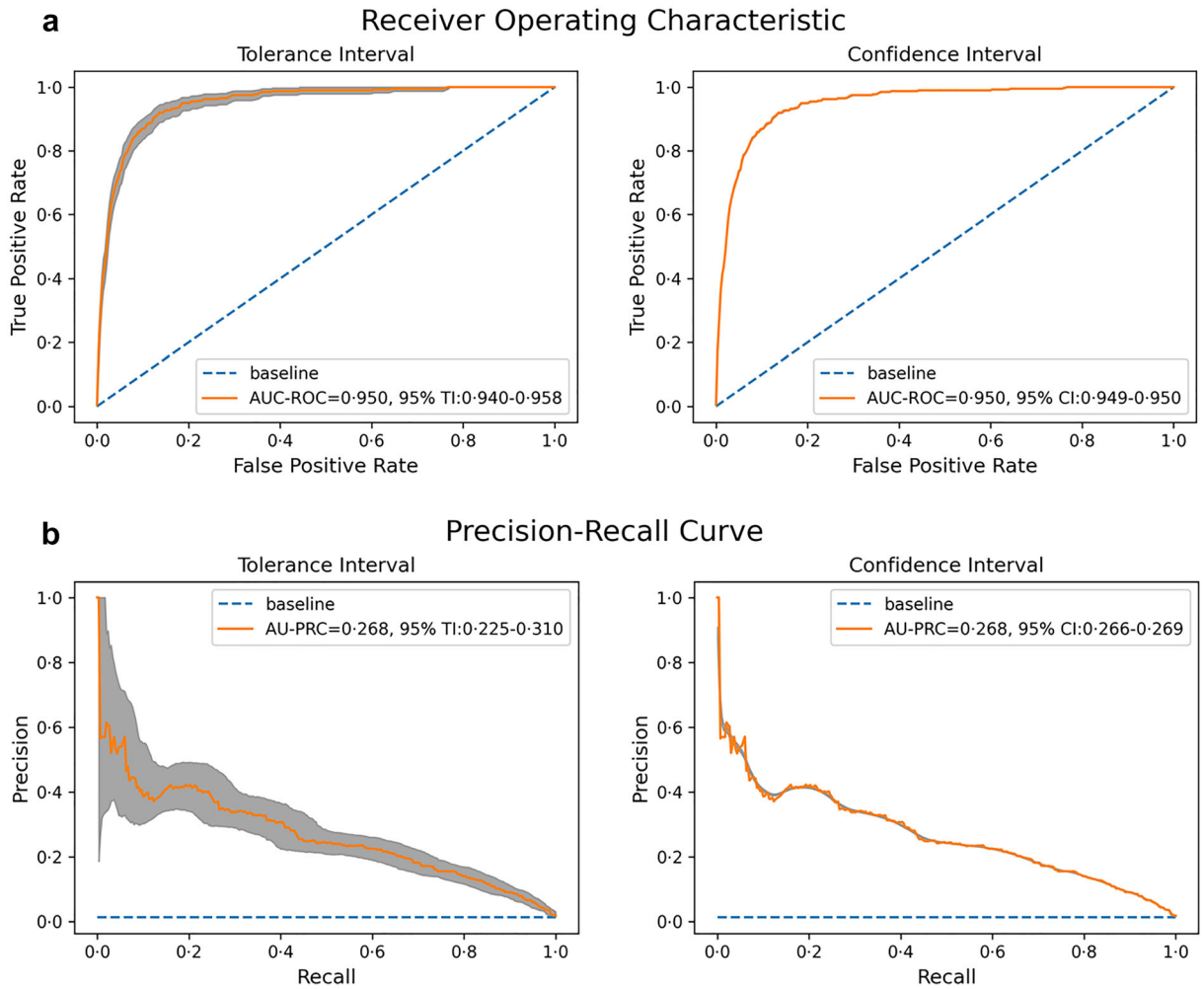

Fig. 5 **a** ROC curve and **b** precision–recall curve. The gray bands around the curves are pointwise 95% TI and 95% CI, which are derived by bootstrapping with 1000 repetitions

Table 4 Performance of four different models

	XGBoost	Light GBM	Random forest	CatBoost
AUPRC	0.268	0.260	0.240	0.261
AUROC	0.950	0.943	0.944	0.947
Precision	0.923	0.925	0.978	0.881
Recall	0.807	0.769	0.025	0.897
F1	0.861	0.840	0.049	0.889
Youden's index	0.739	0.707	0.025	0.776
Specificity	0.933	0.938	0.999	0.879

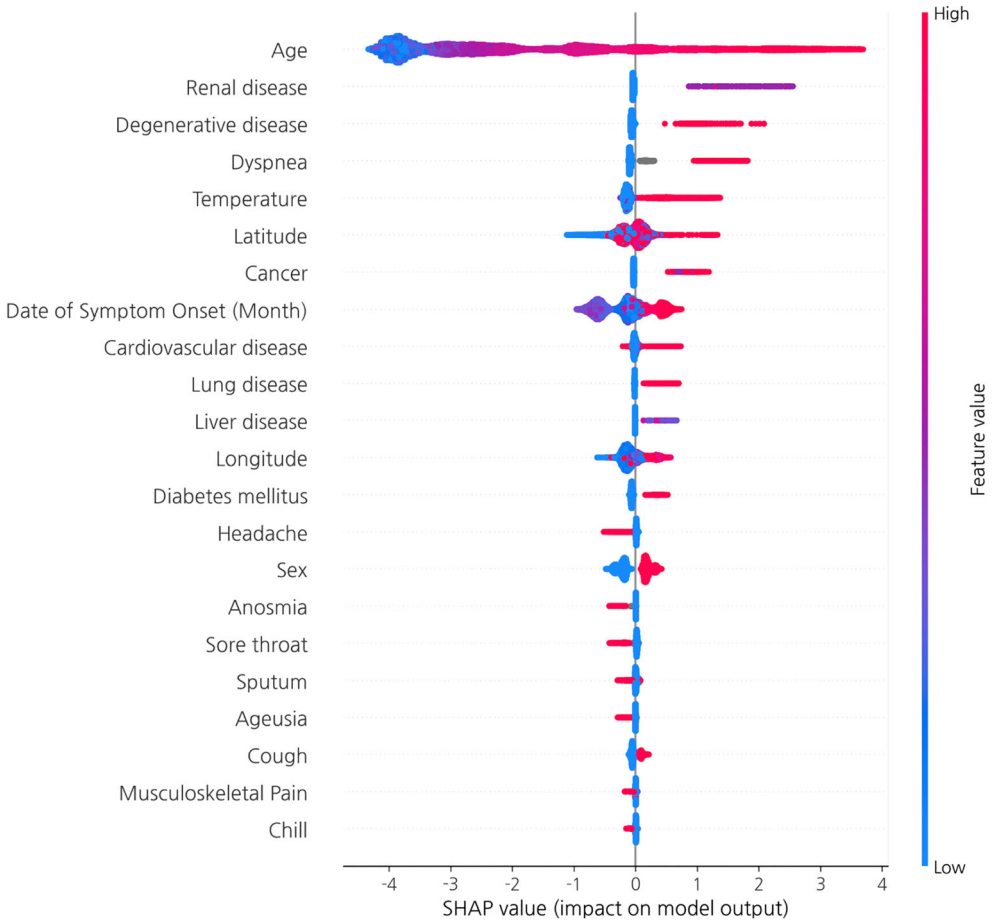

Fig. 6 Feature importance plot

general ROC curve and PR curve are presented in Fig. 5.

We compared the performance of four different models (Table 4). The XGB model achieved the highest scores with an AUROC of 0.950 and AUPRC of 0.268.

Explainability

Feature importance was measured by SHAP, as presented in Fig. 6. Features in the plot are sorted in descending order by their maximum absolute values. A single dot on each row represents the explanation for each patient, and the original feature values are represented by their colors. The SHAP analysis proved age to be the most important relevant risk factor for mortality. Body temperature was also an important risk factor, as were previous diseases before COVID-19 infection, such as renal disease, degenerative disease, cancer, liver, cardiovascular, and lung disease. Among initial symptoms of patients, dyspnea was shown to be an important risk factor. Geographic information is also closely related to the mortality of patients with COVID-19. Higher longitude and latitude are related to high mortality. The northeast region is covered with more mountains than the west or southern region in South Korea while almost all large cities are located in the southern and western parts of the country. In terms of accessibility to acute care facilities, geographic location significantly affects the mortality of patients with acute respiratory diseases [44, 45]. Different weather according to location may also affect the severity of disease or mortality of the patients [46].

Cost–Benefit Analysis

Decision curve analysis (DCA), as depicted in Fig. 7, provides the range of threshold probabilities in which a prediction model shows the value and magnitude of benefit [47]. In the context of this research, the threshold can be used to decide whether a self-quarantined patient should be hospitalized or not. The threshold should be set depending on the medical and economic environment of the country in which the model is implemented. The DCA identified the optimal threshold range in which net benefit does not fall below zero. In our model, the optimal threshold for the DCA ranged from 0 to 0.05.

We also investigated the types of medical institutions visited by patients according to their predicted mortality probabilities, as shown in Fig. 8. First, we divided the test set into three groups: patients with predicted mortality probabilities less than 0.05, those between 0.05 and 0.5, and those greater than 0.5. Then, we analyzed the types of medical institutions that the patients visited first for each group. Since public health centers are the first places where patients receive the PCR test in general, the proportion of public health centers among the medical

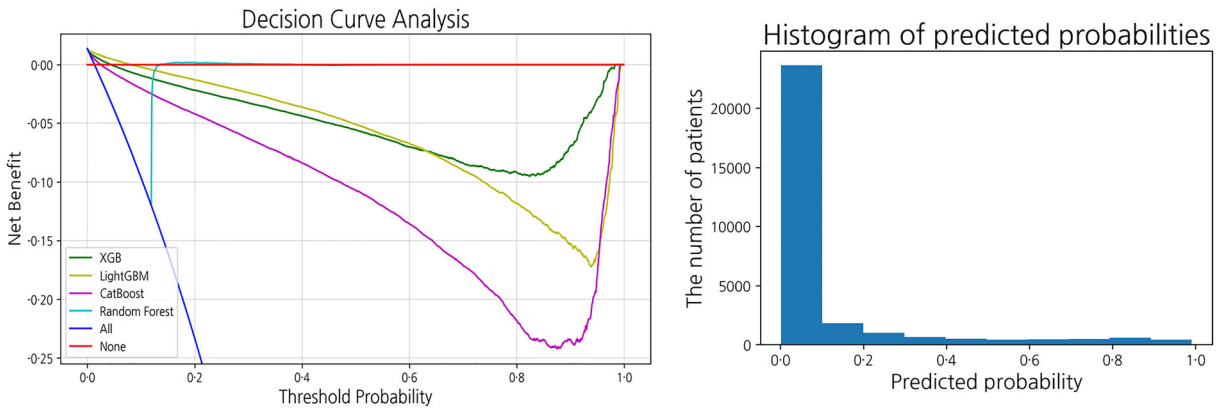

Fig. 7 Decision curve analysis and the histogram of predicted probabilities of the XGB model

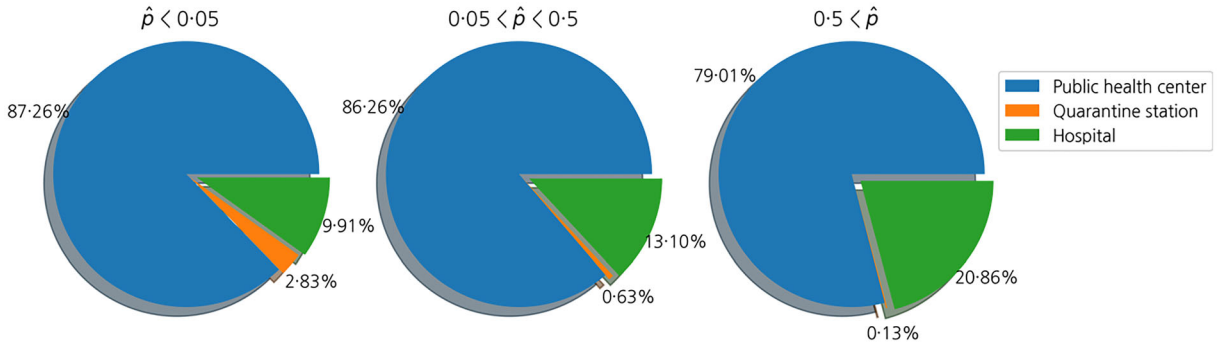

Fig. 8 First-visit facility of patients with COVID-19 according to the patients' mortality probabilities

institutions where patients get treated is great. However, the proportion of hospitals in the pie chart increases if the mortality rate of patients increases, which means more severely infected patients visited hospitals at first than those with less severe cases.

DISCUSSION

In this research, we propose a machine learning model that predicts the prognosis of SARS-CoV-2-infected patients by obtaining 20 basic pieces of PGHD. The model was developed using the data of 149,471 patients from 1382 designated COVID-19 screening centers. Thus, our model can be utilized globally for triaging patients with confirmed SARS-CoV-2 infection at the initial stage and monitoring hospitalized or quarantined patients daily.

The characteristics of SARS-CoV-2 and the related spectrum of signs and symptoms are the subjects of much ongoing research. Initial triage of the patients is crucial to prevent the shutdown of the entire medical system of a country. Thus, there have been many studies on developing patient triage algorithms using easily obtainable signs and symptoms. The model in this study provides a novel method integrating easily obtainable signs and symptoms, along with geographic and seasonal data that reflect characteristics of respiratory viruses, all from the nationwide multicenter database, including hospitalization and mortality data.

Accurate patient triage may lower the burden currently faced by health systems through facilitating optimized management of healthcare resources during future waves of the SARS-CoV-2 pandemic [48]. This is especially important in developing countries with limited resources to maintain essential health services [49].

While reviewing the existing research, we found that most of the previous studies utilized limited data. Furthermore, almost all of them utilized various input features that are not easy to measure. Compared to the previous studies, we adopted two types of demographic information, one geographic location, one sign, nine symptoms, and seven underlying diseases, which are easy to measure. Only body temperature and the nine symptoms are changeable during quarantine and hospitalization. Thus, patients can check the severity of the disease every day with the variable input features. The data for the research was collected from 1382 designated COVID-19 screening centers in South Korea, which means the developed model covered patients with variable clinical characteristics from all over the country. In addition, we adopted longitude and latitude in our model to reflect clinical characteristics of the acute respiratory virus by weather and accessibility of acute care facilities in each region.

Through the result of DCA, users can set a threshold for intervention such as transfer to a higher level of care or medical facility or a thorough examination by doctors. For example, if they are allowed to have a higher false-

positive rate and want to screen necessary patients for intervention as much as possible, they can set the threshold near 0. If they have to save hospital beds for severe patients when medical resources are depleting, they can set the value closer to 0.05.

The SHAP analysis found patients with previous renal, degenerative, or cardiovascular diseases or cancer should be monitored thoroughly. In addition, body temperature and dyspnea should be considered the most important factors to assess aggravation of their health daily.

One of the main limitations of the study is that our model has not yet been extensively applied to the field. Therefore, we could not quantify how efficiently our model could lower the burden on the healthcare system. However, since our model has high performance and is easily accessible, we expect to have positive results and leave this analysis for future work.

CONCLUSION

We developed a model for predicting COVID-19 diagnosis by obtaining 20 basic pieces of PGHD based on nationwide multicenter data reported by KDCA. With the help of COVID-19 vaccination and medicine to be released soon, it will be more important to manage patients under quarantine at home or a facility. Our framework can be implemented and utilized conveniently to triage patients with positive RT-PCR test results as well as enabling them to monitor themselves at home or a quarantine facility.

ACKNOWLEDGEMENTS

We thank the participants of the study.

Funding. This research was supported by the Seoul National University Bundang Hospital Research Fund (Grant #14-2021-0041), the National Research Foundation of Korea Grant funded by the Korean government (NRF-2017R1E1A1A03070105 and NRF-2019R1A5A1028324), the ITRC (Information Technology Research Center) support program (IITP-2018-0-01441), and the Institute for Information & Communications Technology Promotion grant funded by the Korean government (Artificial Intelligence Graduate School Program [POSTECH]; #2019-0-01906). The journal's Rapid Service Fee was funded by KAIST Stochastic Analysis and Application Research Center.

Authorship. All named authors meet the International Committee of Medical Journal Editors (ICMJE) criteria for authorship for this article, take responsibility for the integrity of the work as a whole, and have given their approval for this version to be published.

Author Contributions. MSP and HJ analyzed data and created the models and drafted the entire manuscript as the first authors. HL contributed to the discussion of the results. SYJ and HJH revised the manuscript and supervised the entire process as the corresponding authors.

Disclosures. Min Sue Park, Hyeontae Jo, Haeun Lee, Se Young Jung and Hyung Ju Hwang have nothing to disclose.

Compliance with Ethics Guidelines. This research was approved by the Institutional Review Board of Seoul National University Bundang Hospital (X-2110-717-902). An Informed consent form was not obtained owing to the nature of retrospective studies. The study was performed in accordance with the Helsinki Declaration of 1964 and its later amendments.

Data Availability. Data is not available to the public owing to the regulation of KDCA.

Open Access. This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are

included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc/4.0/.

REFERENCES

1. Korea's COVID-19 vaccination rate ranks 10th among OECD countries. 2021. https://www.koreatimes.co.kr/www/nation/2021/10/119_317546.html. Accessed 8 Nov 2021.

2. Tham D. "New normal": 7 ways Singapore is moving towards living with COVID-19. https://www.channelnewsasia.com/singapore/covid19-dine-vaccination-home-recovery-booster-new-normal-2232901. Accessed 8 Nov 2021.

3. More In Common. The new normal? https://www.moreincommon.com/newnormal/. Accessed 8 Nov 2021.

4. Ray S. "Living with covid" strategy under threat as Singapore, U.K. face sharp rise in cases. Forbes Magazine. 2021 Oct 21. https://www.forbes.com/sites/siladityaray/2021/10/21/living-with-covid-strategy-under-threat-as-singapore-uk-face-sharp-rise-in-cases/?sh=533341063a77. Accessed 8 Nov 2021.

5. Sim D. Can Singapore's health care system cope with its biggest surge in Covid-19 infections? South China Morning Post. 2021. https://www.scmp.com/week-asia/health-environment/article/3154538/can-singapores-health-care-system-cope-its-biggest. Accessed 8 Nov 2021.

6. World Health Organization. Critical preparedness, readiness and response actions for COVID-19: interim guidance, 4 November 2020. World Health Organization; 2020. https://apps.who.int/iris/bitstream/handle/10665/336373/WHO-COVID-19-Community_Actions-2020.5-eng.pdf.

7. Shin H-S, Park H, Kwon JS, et al. National Academy of Medicine of Korea (NAMOK) key statements on COVID-19. J Korean Med Sci. 2021;36(41):e287. https://doi.org/10.3346/jkms.2021.36.e287.

8. Organisation for Economic Co-Operation and Development (OECD). The territorial impact of COVID-19: Managing the crisis across levels of government. 2020 Nov. (OECD Policy Responses to Coronavirus (COVID-19)). https://www.oecd-ilibrary.org/urban-rural-and-regional-development/the-territorial-impact-of-covid-19-managing-the-crisis-across-levels-of-government_d3e314e1-en. Accessed 8 Nov 2021.

9. Choi WS, Kim HS, Kim B, Nam S, Sohn JW. Community treatment centers for isolation of asymptomatic and mildly symptomatic patients with coronavirus disease, South Korea. Emerg Infect Dis. 2020;26(10):2338–45. https://doi.org/10.3201/eid2610.201539.

10. Arin K. Delayed response blamed in COVID-19 patient's death at home. http://www.koreaherald.com/view.php?ud=20211024000249. Accessed 8 Nov 2021.

11. Wang Q, Wang X, Lin H. The role of triage in the prevention and control of COVID-19. Infect Control Hosp Epidemiol. 2020;41(7):772–6.

12. Depuydt P, Guidet B. Triage policy of severe Covid-19 patients: what to do now? Ann Intensive Care. 2021;11(1):18. https://doi.org/10.1186/s13613-020-00770-9.

13. Kim J-H, An JA-R, Min P-K, Bitton A, Gawande AA. How South Korea responded to the Covid-19 outbreak in Daegu. NEJM Catalyst. 2020;1(4). https://doi.org/10.1056/CAT.20.0159.

14. Sajadi MM, Habibzadeh P, Vintzileos A, Shokouhi S, Miralles-Wilhelm F, Amoroso A. Temperature, humidity, and latitude analysis to estimate potential spread and seasonality of coronavirus disease 2019 (COVID-19). JAMA Netw Open. 2020;3(6):e2011834. https://doi.org/10.1001/jamanetworkopen.2020.11834.

15. Tian Y, Wu Q, Li H, Wu Q, Xie Y, Li L, Chen H. Distinct symptoms and underlying comorbidities with latitude and longitude in COVID-19: a systematic review and meta-analysis. Research Square; 2020. https://www.researchsquare.com/article/rs-117666/latest.pdf.

16. Sharma A, Preece B, Swann H, et al. Structural stability of SARS-CoV-2 virus like particles degrades with temperature. Biochem Biophys Res Commun. 2021;534:343–6. https://doi.org/10.1016/j.bbrc.2020.11.080.

17. Riley WJ. Health disparities: gaps in access, quality and affordability of medical care. Trans Am Clin Climatol Assoc. 2012;123:167–72; discussion 172–4. https://www.ncbi.nlm.nih.gov/pubmed/23303983.

18. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery; 2016. p. 785–94. (KDD '16). https://doi.org/10.1145/2939672.2939785. Accessed 12 Aug 2021.

19. Park MS, Son H, Hyun C, Hwang HJ. Explainability of machine learning models for bankruptcy prediction. IEEE Access. 2021;9:124887–99. https://doi.org/10.1109/ACCESS.2021.3110270.

20. Hwang R, Jo H, Kim KS, Hwang HJ. Hybrid model of mathematical and neural network formulations for rolling force and temperature prediction in hot rolling processes. IEEE Access. 2020;8:153123–33. https://doi.org/10.1109/ACCESS.2020.3016725.

21. Josse J, Prost N, Scornet E, Varoquaux G. On the consistency of supervised learning with missing values. arXiv [stat.ML]. 2019. http://arxiv.org/abs/1902.06931.

22. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems. 2017. p. 4768–77. http://papers.nips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

23. Lipovetsky S, Conklin M. Analysis of regression in game theory approach. Appl Stoch Mod Data Anal. 2001;17(4):319–30. https://doi.org/10.1002/asmb.446.

24. Beger A. Precision-Recall Curves. 2016. https://papers.ssrn.com/abstract=2765419. Accessed 18 Aug 2021.

25. Zoabi Y, Deri-Rozov S, Shomron N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. NPJ Digit Med. 2021;4(1):3. https://doi.org/10.1038/s41746-020-00372-6.

26. Menni C, Valdes AM, Freidin MB, et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. Nat Med. 2020;26(7):1037–40. https://doi.org/10.1038/s41591-020-0916-2.

27. Yanamala N, Krishna NH, Hathaway QA, et al. A vital sign-based prediction algorithm for differentiating COVID-19 versus seasonal influenza in hospitalized patients. NPJ Digit Med. 2021;4(1):95. https://doi.org/10.1038/s41746-021-00467-8.

28. Gozes O, Frid-Adar M, Greenspan H, et al. Rapid AI development cycle for the coronavirus (COVID-19) pandemic: initial results for automated detection & patient monitoring using deep learning CT image analysis. arXiv [eess.IV]. 2020. http://arxiv.org/abs/2003.05037.

29. Song Y, Zheng S, Li L, et al. Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images. IEEE/ACM Trans Comput Biol Bioinform. 2021;18(6):2775–80. https://doi.org/10.1109/TCBB.2021.3065361.

30. Jin C, Chen W, Cao Y, et al. Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. Nat Commun. 2020;11(1):1–4. https://doi.org/10.1038/s41467-020-18685-1.

31. Punn NS, Agarwal S. Automated diagnosis of COVID-19 with limited posteroanterior chest X-ray images using fine-tuned deep neural networks. Appl Intell. 2021;51(5):2689–702. https://doi.org/10.1007/s10489-020-01900-3.

32. Feng C, Huang Z, Wang L, et al. A novel triage tool of artificial intelligence assisted diagnosis aid system for suspected COVID-19 pneumonia in fever clinics. SSRN Electron J. https://doi.org/10.2139/ssrn.3551355.

33. Cifuentes MP, Rodriguez-Villamizar LA, Rojas-Botero ML, Alvarez-Moreno CA, Fernández-Niño JA. Socioeconomic inequalities associated with mortality for COVID-19 in Colombia: a cohort nationwide study. J Epidemiol Commun Health. 2021. https://doi.org/10.1136/jech-2020-216275.

34. Her AY, Bhak Y, Jun EJ, et al. A clinical risk score to predict in-hospital mortality from COVID-19 in South Korea. J Korean Med Sci. 2021;36(15): e108. https://doi.org/10.3346/jkms.2021.36.e108.

35. Cho SI, Yoon S, Lee H-J. Impact of comorbidity burden on mortality in patients with COVID-19 using the Korean health insurance database. Sci Rep. 2021;11(1):6375. https://doi.org/10.1038/s41598-021-85813-2.

36. Ikemura K, Bellin E, Yagi Y, et al. Using automated machine learning to predict the mortality of patients with COVID-19: prediction model development study. J Med Internet Res. 2021;23(2): e23458. https://doi.org/10.2196/23458.

37. Shamout FE, Shen Y, Wu N, et al. An artificial intelligence system for predicting the deterioration of COVID-19 patients in the emergency department. NPJ Digit Med. 2021;4(1):80. https://doi.org/10.1038/s41746-021-00453-0.

38. Subudhi S, Verma A, Patel AB, et al. Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. NPJ Digit Med. 2021;4(1):87. https://doi.org/10.1038/s41746-021-00456-x.

39. Marcos M, Belhassen-García M, Sánchez-Puente A, et al. Development of a severity of disease score and classification model by machine learning for

hospitalized COVID-19 patients. PLoS One. 2021;16(4):e0240200. https://doi.org/10.1371/journal.pone.0240200.

40. Kim H-J, Han D, Kim J-H, et al. An easy-to-use machine learning model to predict the prognosis of patients with COVID-19: retrospective cohort study. J Med Internet Res. 2020;22(11):e24225. https://doi.org/10.2196/24225.

41. Su C, Zhang Y, Flory JH, et al. Clinical subphenotypes in COVID-19: derivation, validation, prediction, temporal patterns, and interaction with social determinants of health. NPJ Digit Med. 2021;4(1):110. https://doi.org/10.1038/s41746-021-00481-w.

42. Rinderknecht MD, Klopfenstein Y. Predicting critical state after COVID-19 diagnosis: model development using a large US electronic health record dataset. NPJ Digit Med. 2021;4(1):113. https://doi.org/10.1038/s41746-021-00482-9.

43. Wang J, Liu C, Li J, et al. iCOVID: interpretable deep learning framework for early recovery-time prediction of COVID-19 patients. NPJ Digit Med. 2021;4(1):124. https://doi.org/10.1038/s41746-021-00496-3.

44. Parcha V, Kalra R, Bhatt SP, Berra L, Arora G, Arora P. Trends and geographic variation in acute respiratory failure and ARDS mortality in the United States. Chest. 2021;159(4):1460–72. https://doi.org/10.1016/j.chest.2020.10.042.

45. Kishamawe C, Rumisha SF, Mremi IR, et al. Trends, patterns and causes of respiratory disease mortality among inpatients in Tanzania, 2006–2015. Trop Med Int Health. 2019;24(1):91–100. https://doi.org/10.1111/tmi.13165.

46. Sil A, Kumar VN. Does weather affect the growth rate of COVID-19, a study to comprehend transmission dynamics on human health. J Saf Sci Resilience. 2020;1(1):3–11.

47. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565–74. https://doi.org/10.1177/0272989X06295361.

48. Hamid H, Abid Z, Amir A, Rehman TU, Akram W, Mehboob T. Current burden on healthcare systems in low- and middle-income countries: recommendations for emergency care of COVID-19. Drugs Ther Perspect. 2020;1–3. https://doi.org/10.1007/s40267-020-00766-2.

49. World Health Organization. Maintaining essential health services: operational guidance for the COVID-19 context: interim guidance, 1 June 2020. World Health Organization; 2020. https://apps.who.int/iris/bitstream/handle/10665/332240/WHO-2019-nCoV-essential_health_services-2020.2-rus.pdf.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.