



## OPEN Profiling short-term longitudinal severity progression and associated genes in COVID-19 patients using EHR and single-cell analysis

Seung Hwan Bae<sup>1,11</sup>, Donghee Kim<sup>2,11</sup>, Jaeyeon Jang<sup>1</sup>, A-Sol Kim<sup>3</sup>, Soyeon Hwang<sup>4</sup>, Eunkyung Nam<sup>4</sup>, Sohyun Bae<sup>5</sup>, JiYeon Lee<sup>6</sup>, Ji Sun Kim<sup>7</sup>, Sang Cheol Kim<sup>8</sup>, Hye-Yeong Jo<sup>8</sup>, Kwangsoo Kim<sup>9,10</sup>, Inuk Jung<sup>1</sup>✉ & Ki Tae Kwon<sup>4</sup>✉

Here we propose CovSF, a deep learning model designed to track and forecast short-term severity progression of COVID-19 patients using longitudinal clinical records. The motivation stems from the need for timely medical resource allocation, improved treatment decisions during pandemics, and the understanding of severity progression related immunology. The COVID-19 Severity Forecasting model, CovSF, utilizes 15 clinical features to profile the severity levels of hospital admitted patients and also forecast their severity levels of up to three days ahead. CovSF was trained on a large COVID-19 cohort (n=4,509), achieving an AUROC of 0.92 with 0.85 and 0.89 sensitivity and specificity on an external validation dataset (n=443). The type of oxygen therapy administered was utilized as the target predictive label, which is often used as the severity index. This approach enables the inclusion of a more comprehensive dataset encompassing patients across the full spectrum of severity, rather than restricting the analysis to more narrowly defined outcomes such as ICU admission or mortality. We focused on profiling deteriorating and recovering health conditions, which were validated using patient matched single-cell transcriptomes. Especially, we showed that the immunology significantly differed between the samples during deterioration and recovery, whose severity levels were the same, and thus presenting the importance of longitudinal analysis. We believe that the framework of CovSF can be extended to other respiratory infectious diseases to alleviate the strain of allocating hospital resources, especially in pandemics.

**Keywords** Severity, Progression, Time course, Deep learning, Single-cell

The severity of a hospitalized patient can be measured by observing a set of clinical features collected from the electronic health record (EHR). For example, the National Early Warning Score (NEWS)<sup>1</sup> is a single integer value that indicates the level of illness of a patient at some specific time. NEWS aggregates the measurement of seven clinical features (i.e., respiratory rate, oxygen saturation, oxygen supplement, body temperature, systolic blood pressure, heart rate and consciousness) into a single value. Similarly, the WHO Ordinal Scale (WHO)<sup>2</sup> and the

<sup>1</sup>School of Computer Science and Engineering, Kyungpook National University, Daegu 41566, Republic of Korea.

<sup>2</sup>Seoul National University Hospital Biomedical Research Institute, Seoul 03082, Republic of Korea. <sup>3</sup>Department of Family Medicine, Kyungpook National University Chilgok Hospital, School of Medicine, Kyungpook National University, Daegu 41404, Republic of Korea. <sup>4</sup>Division of Infectious Diseases, Department of Internal Medicine, Kyungpook National University Chilgok Hospital, School of Medicine, Kyungpook National University, Daegu 41404, Republic of Korea. <sup>5</sup>Division of Infectious Diseases, Department of Internal Medicine, School of Medicine, Kyungpook National University Hospital, Kyungpook National University, Daegu 41404, South Korea. <sup>6</sup>Division of Infectious Diseases, Department of Internal Medicine, Keimyung University Dongsan Hospital, Keimyung University School of Medicine, Daegu 41931, Republic of Korea. <sup>7</sup>Department of Medical Information, Kyungpook National University Hospital, Daegu 41944, Korea. <sup>8</sup>Division of Healthcare and Artificial Intelligence, Department of Precision Medicine, Korea National Institute of Health, Korea Disease Control and Prevention Agency, Cheongju 28159, South Korea. <sup>9</sup>Department of Transdisciplinary Medicine, Institute of Convergence Medicine with Innovative Technology, Seoul National University Hospital, Seoul 03080, Republic of Korea. <sup>10</sup>Department of Medicine, College of Medicine, Seoul National University, Seoul 03080, Republic of Korea. <sup>11</sup>These authors contributed equally this work: Seung Hwan Bae and Donghee Kim. ✉email: inukjung@knu.ac.kr; ktkwon@knu.ac.kr

Sequential Organ Failure Assessment (SOFA)<sup>3</sup> are also widely used indexes for describing one's severity but with different purposes. While NEWS is a scoring tool for general purposes that may be applicable to a wide spectrum of diseases, the WHO scale is more tailored to measure the illness of patients with respiratory diseases. SOFA is used to assess the severity of organ dysfunction and predict outcomes in critically ill patients, particularly treated in the intensive care unit (ICU).

When such scores are measured in temporal manner (e.g., hourly or daily), the severity progression of a patient can be observed from which deteriorating or recovering conditions can be determined. Knowing such progression can be helpful in planning the management of hospital resources, providing patients with timely treatment and securing sickbeds or ICU, which becomes difficult during a pandemic outbreak. More importantly, when patient matched omics data is available, gene markers related to deterioration or recovery can be searched to understand their underlying immunological mechanism. EHR based clinical outcome prediction models can alleviate such strain by aiding in preemptive hospital resource allocation or de-allocation via monitoring the patient health status in real time. These models have the potential to forecast mortality risk, thereby enabling prioritized treatment for high-risk patients, and can evolve to identify new risk factors and understand interactions among them for more effective interventions. Addressing this issue requires an understanding of the stages of a disease, such as COVID-19, predicting patient outcomes, and accordingly allocating medical resources.

Longitudinal clinical data is valuable and its analysis is essential to identify and understand disease progression, devise prognosis, and develop early diagnostic methods. Given the complexity of patient conditions, intervention frequency, and the need for timely information, machine learning techniques utilizing clinical data are vital. Therefore, for certain diseases with low mortality, defining the oxygen treatment type or admission to ICU as severity is more appropriate to incorporate time-course characteristics than mortality events<sup>4,5</sup>. Recent analyses<sup>6–8</sup> have employed statistical methods, machine learning and deep learning approaches to predict patient's severity outcomes using non-time course snapshot data. Their target severity outcomes mainly focused on mortality events and less critical events, such as ICU admission and the need for mechanical ventilation. Nevertheless, due to the static data they are not apt for tracking one's severity progression and thus are less accurate at determining an appropriate treatment type at a certain time point. Also, in case of COVID-19, the mortality rate is much lower than the severity rate, which is expected to further decline with accumulated vaccination, affecting the quality of learning mortality events due to the diminishing number of cases<sup>9</sup>.

Other studies have made effort to predict ICU admission or the required type of oxygen treatment in real time by utilizing time-course data. However, they either use too few or too many clinical features making the adoption of their methods in hospitals difficult. For example, only the saturation of partial pressure oxygen (SPO<sub>2</sub>), heart rate and body temperature levels were used to forecast the deterioration of a patient's health condition<sup>10</sup>. Similarly, the C-reactive protein (CRP), neutrophil and lymphocyte levels were used to predict the required oxygen treatment in real-time<sup>11</sup>. In contrast, in two other studies<sup>12,13</sup>, 66 and 54 features were used to forecast the mortality and severity of COVID-19 patients. While the previous two studies utilized longitudinal features the latter one used a mixture of static and dynamic features. Here, static features are referred to as clinical features that unlikely change during the admission, such as age, height or ethnicity (i.e., mostly demographics). Dynamic features are referred as the ones that are highly responsive to an infection and changes over time during the admission, such as breath, body temperature, white blood cell (WBC) counts, neutrophil, or CRP. In addition, the number of input features should be minimal but informative and frequently measured to make a prediction model as applicable and general as possible to be used across different hospitals. The majority of previous studies lack an external validation dataset or sufficient number of patients, thus the generalizability of their models remains inadequately verified<sup>10–12,14</sup>.

Here, we developed a severity scoring method, COVID-19 Severity Forecast (CovSF), for COVID-19 patients that is able to 1) measure the illness of a patient in real-time, 2) forecast one's severity level up to three days after and 3) profile the severity progression (i.e., deterioration or recovery). CovSF takes 15 clinical features measured at a particular day as input of up to five days and outputs the severity level of each of the input days followed by the predicted severity levels of the next three days. CovSF was trained on longitudinal EHR data collected from a large COVID-19 cohort (n=10,627) in South Korea to learn the required oxygen treatment type that indirectly reflect one's severity. The performance of CovSF was validated using an external cohort of 459 COVID-19 patients and 189 community-acquired pneumonia patients, which was used to exploit the possibility to scale our method to other respiratory diseases. Furthermore, the severity score and progression output by CovSF was validated using patient-matched single-cell transcriptomes to present that the expression level of known severity progression-related genes significantly correlated with the CovSF output scores. While this study focused on COVID-19, it showed that the relatively less costly EHR data can be used to identify the temporal progression of a disease and therefore stratify patients to search for progression related biomarkers in the complex and costly omics samples. The biomarkers may serve as a sources for understanding the underlying biology, development of drug targets or diagnostic markers.

## Results

### Data collection and clinical feature selection

Longitudinal EHR data were collected from three different cohorts across South Korea: 1) Kyungpook National University (KNU, n=10,672), 2) the Korea National Institute of Health (KNIH, n=459), and 3) the Non-COVID-19 Community Acquired Pneumonia (NCCAP, n=189). The KNU and KNIH cohorts is comprised of COVID-19 patients, whereas the NCCAP is composed of non-COVID-19 community-acquired pneumonia patients. The KNU data were collected from Kyungpook National University Hospital and Kyungpook National University Chilgok Hospital. The KNIH data were collected from Chungnam National University Hospital, Seoul Medical Center and Samsung Medical Center. In addition to the clinical data, patient matched single-cell

RNA-sequencing (scRNA-seq) samples were collected from the KNIH cohort<sup>15</sup>. The NCCAP data was collected from Kyungpook National University Chilgok Hospital to assess if the proposed approach may be extended to other infectious diseases.

Among the many clinical features in the three cohorts, 15 were carefully chosen to train the CovSF model, including six vital signs and nine laboratory test characteristics. The vital signs include body temperature (BDTEMP), systolic blood pressure (SBP), diastolic blood pressure (DBP), heart rate (i.e.pulse), respiratory rate (i.e.breath), and SPO<sub>2</sub>, and the laboratory features consist of blood urea nitrogen (BUN), creatinine, hemoglobin, lactate dehydrogenase (LDH), lymphocyte, neutrophil, platelet count, WBC counts, and CRP, where the neutrophil and lymphocyte features were merged into a single neutrophil-to-lymphocyte ratio (NLR) feature. The oxygen treatment type was used as the severity index because of its ability to inform clinical decision and its alignment with recommended drug prescription criteria<sup>16</sup>. Furthermore, due to the Korean government's strict regulation of hospital bed allocation and medication distribution based on patient severity, including the level of oxygen therapy, all healthcare institutions and physicians were required to comply with governmental directives. These policies also served to minimize variations in practice among institutions and physicians. The oxygen treatment is constituted of five types ordered by severity: room air (no oxygen treatment), nasal, mask, High-Flow Nasal Cannula (HFNC) and mechanical ventilation. The oxygen treatment types were further reduced into two groups to be used as the predictor labels in a binary classification problem: 1) mild treatment (i.e., room air, nasal treatment) and 2) severe treatment (i.e., mask, HFNC and mechanical ventilation). The differences in clinical features among oxygen treatment types are clearly demonstrated in our datasets (Figure S1). Preprocessing of the three cohorts for model development and assessment is summarized in Figure 1A, and their statistics, including the final number of patients are described in Table 1.

### Framework of CovSF

As illustrated in Figure 1B, CovSF takes an input sequence (blue boxes) of 15 clinical features that are measured longitudinally from up to 4 days prior to the present, to score the severity of the present and also predict the patient's severity for the upcoming 3 days (green boxes). Due to model architecture's flexibility, CovSF is able to forecast one's severity even using a single day as input, which is advantageous when an input sequence is inevitably short but severity predictions are still required for timely treatment (e.g., the first day at admission). As the input sequence is input to CovSF on a daily basis during the patient's hospitalization, multiple predictions can be made for a particular day. The final CovSF score is calculated by averaging these outputs for each day. The CovSF score ranges from 0 to 1, where a higher score indicates higher severity. Thus, the CovSF score provides an overall severity assessment of a COVID-19 patient, reflecting the longitudinal trend and relations between the clinical features embedded in high-dimension.

Based on the score, each time point of a patient was stratified into one of the three severity groups according to the grid search results for optimal performance (Figure S2): 1) mild (Mi,  $\text{CovSF} < 0.2$ ), 2) moderate (M,  $0.2 \leq \text{CovSF} < 0.5$ ) and 3) severe (S,  $\text{CovSF} \geq 0.5$ ). The time point signifies one unit per day, representing the hospitalized day of the patient and serving as a unit for model input. The Severity of patients might change from admission to discharge, with three major type of progressions: 1) Constant (CONS) 2) deteriorating phase (DP) and 3) recovering phase (RP). Here, CONS refers to patients with a static severity level that do not change during the admission. In the other hand, a patient with a full DP will experience the following sequence of severity levels, Mi-M-S, whereas an RP patient will experience a severity sequence of S-M-Mi. Detailed information on the above severity and progression groups is presented in Table S1.

Briefly, We present the CovSF outputs and scores of two patients selected from the KNIH cohort, one with DP and another with RP health conditions. Figure 1C depicts the CovSF severity scores across the hospitalization, along with feature values and their importance, in longitudinal manner. CovSF demonstrated more reasonable inference of severity compared to the actually received oxygen treatment type, especially by capturing elevating severity five days earlier in the DP patient (i.e., left) before the actual severe treatment began. Feature values that highly correlated with the CovSF score was observed in both patients, supported by the disparity from the feature's baseline, which was defined as the average of time points for patients who consistently remained in the Mi group from admission to discharge. Likewise, the CovSF was verified in terms of clinical and transcriptomic profiles in the following steps, and its performance for severity prediction was also evaluated using several metrics.

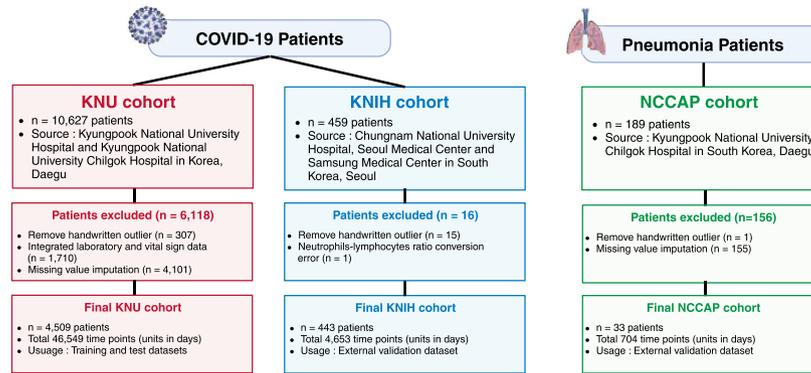
### Performance evaluation of CovSF

As CovSF is able to 1) forecast the severity of up to three days and 2) determine the severity progression, the performance of CovSF was evaluated accordingly.

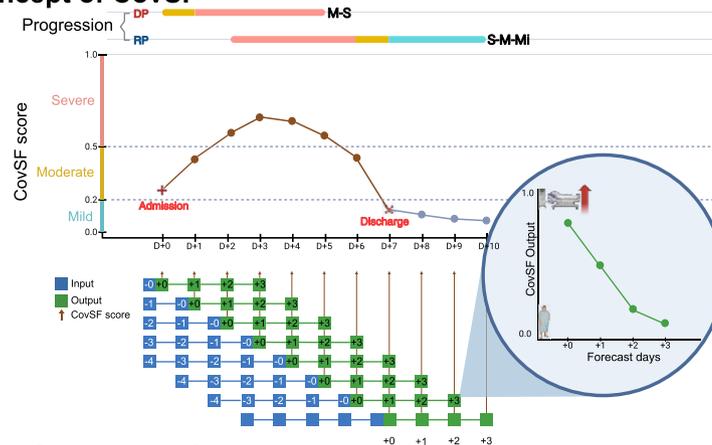
First, the performance metrics of predicting the required mild or severe oxygen treatment was measured. Two deep learning models (i.e., CovSF and Vanilla) and two linear models were implemented for performance comparison. The Area Under Receiver Operating Characteristic (AUROC), sensitivity, and specificity were measured for each model at each severity forecast day using the external validation cohorts KNIH and NCCAP, presented in Table S2. In general, the deep learning models demonstrated overall high performance. In the KNUH training cohort, they achieved an AUROC of 99%, with 94% sensitivity and 97% specificity, while in the KNIH external validation cohort, the AUROC was 91%, sensitivity 84%, and specificity 86%, on average. The CovSF yielded the highest overall performance across the three cohorts, including the NCCAP external validation cohort, demonstrating the potential to extend our approach to other target infectious diseases.

Next, CovSF was further evaluated to determine whether it could forecast changes in the required oxygen treatment type, thus demonstrating the ability to allocate timely and appropriate treatment. For such purpose, we define a transition point (TP) as a time point where a change of treatment type occurred, such as from room air to mechanical ventilation treatment or vice versa. Figure S3A illustrates how the CovSF model behaved

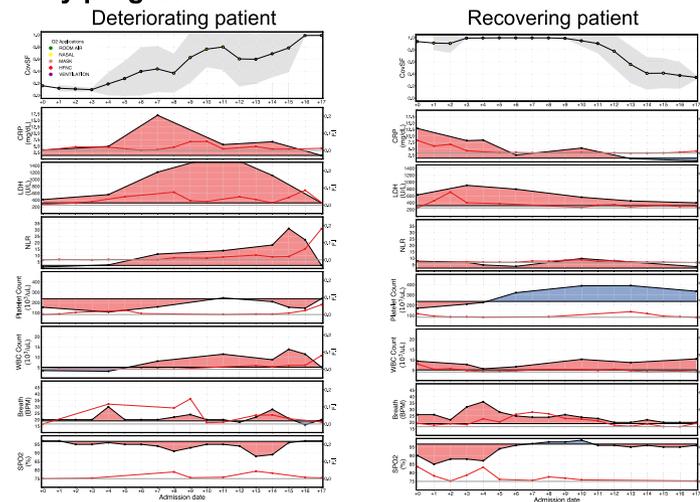
### A Cohort datasets



### B Concept of CovSF



### C Severity progression



**Fig. 1.** Demonstration of the overall study design. (A) The number of patients and time points collected and excluded during data preprocessing across the three cohorts. (B) The conceptual depiction of CovSF predicting one’s severity level (green box) based on the given clinical input (blue box). (C) CovSF scores, feature values, and feature importance for deteriorating and recovering patients in the KNIH cohort. The grey area beside the CovSF score represents the output variability (i.e. standard deviation) arising from multiple predictions on the same day. Feature values closer to severe than the baseline are shaded in red, and those closer to the mild are shaded in blue. Feature importance (Fi.) is shown by the red line.

near the TP within the KNUH and KNIH cohorts, with its outputs limited to the days before the TP was input into CovSF. In case when oxygen treatment changed from mild to severe (left), it can be seen that the output severity score was able to correctly predict such change before the actual treatment change took place. The median outputs before the TP was under 0.5, while the CovSF was predicted to gradually exceed 0.5 at TP and afterwards. This was similarly observed for recovering patients (right) who received severe oxygen treatment but

	KNUH	KNIH	NCCAP
	Train	External validation	
	Internal validation		
	COVID-19	Pneumonia	
	Feb.2020 - Oct.2022	Jan.2021 - Sep.2022	Mar.2020 - Feb.2023
<b>Demographics</b>	(Baseline: # of patients)		
Number of patients (n)	4,509	443	33
Age	62.45 ± 18.81	52.64±16.55	71.31±11.96
Gender (male, n, %)	2,232 (49.50)	244 (55.08)	23 (71.88)
Weight (kg)	63.45 ± 21.57	67.60±14.46	-
Height (cm)	163.53 ± 26.35	165.09±9.53	-
Mortality (n,%)	375 (8.32)	5 (1.13)	8 (25)
Clinical features	(Baseline: # of admission timepoints)		
<b>Vital signal</b>			
Temperature (°C)	36.54 ± 0.33	37.23 ± 0.73	36.57 ± 0.32
Systolic Blood Pressure (mmHg)	124.21 ± 15.22	114.46 ± 15.42	122.60 ± 17.15
Diastolic Blood Pressure (mmHg)	73.38 ± 9.75	72.62 ± 11.50	72.29 ± 10.99
Pulse (BPM)	81.16 ± 14.46	83.06 ± 14.72	91.59 ± 13.39
Breath (BPM)	20.27 ± 2.69	20.11 ± 3.67	19.36 ± 3.74
Saturation of Partial pressure Oxygen (%)	97.77 ± 1.76	96.55 ± 12.33	98.19 ± 1.50
Laboratory test			
Blood Urea Nitrogen (mg/dL)	28.59 ± 23.48	17.03 ± 12.12	33.15 ± 26.35
Creatinine (mg/dL)	1.46 ± 1.98	0.97 ± 1.27	1.72 ± 1.87
Hemoglobin (g/dL)	11.49 ± 2.26	12.83 ± 1.96	9.69 ± 1.84
Lactate Dehydrogenase (U/L)	341.97 ± 307.75	474.14 ± 367.25	455.36 ± 551.78
Lymphocytes (%)	18.17 ± 13.20	23.27 ± 12.98	12.42 ± 9.27
Neutrophils (%)	71.76 ± 16.64	66.79 ± 15.60	78.34 ± 13.40
Platelet count (10 <sup>3</sup> /uL)	214.93 ± 116.01	227.53 ± 96.53	243.98 ± 158.32
White Blood Cell count (10 <sup>3</sup> /uL)	8.83 ± 6.72	6.98 ± 4.66	10.92 ± 13.78
C-Reactive Protein (mg/dL)	4.25 ± 5.82	2.57 ± 3.96	7.40 ± 6.87
<b>O2 treatments (n, %)</b>	(Baseline: # of admission timepoints)		
Room air	29,252 (62.85%)	3,576 (76.85%)	165 (23.44%)
Nasal	8,074 (17.35%)	528 (11.35%)	286 (40.62%)
Mask	769 (1.65%)	45 (0.97%)	23 (3.27%)
High Flow Nasal Cannula	4,918 (10.57%)	249 (5.35%)	117 (16.62%)
Ventilation tube	3,529 (7.58%)	255 (5.48%)	113 (16.05%)

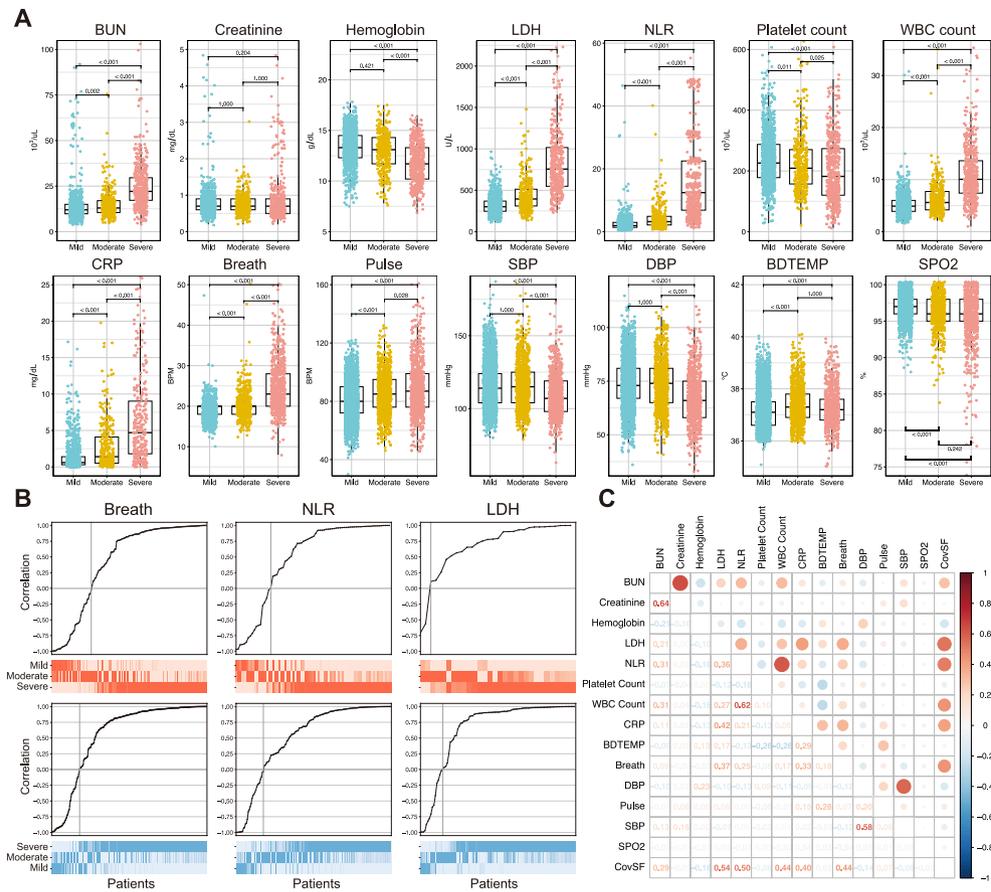
**Table 1.** Baseline characteristics of the three cohorts, focusing on patients and in-hospitalized time points after preprocessing. For categorical features, their occurrences are described, while for numerical features, their mean and standard deviation are specified.

then mild oxygen treatment without observing the actual treatment type at TP. Collectively, CovSF proved to be effective in predicting the severity 3 days prior the severity progression, or treatment type, changed and well reflected the treatment types decided by the professional medical practitioners at the hospital.

To understand the behavior of the CovSF model, we measured how each input feature contributed to a prediction as shown in Figure S3B. Among the laboratory features, LDH and NLR showed to be the highest predictors in the COVID-19 cohorts, while breath was the highest predictor among the vital sign features followed by SPO<sub>2</sub>, which is agreeable since COVID-19 is a respiratory disease. LDH and NLR were the next important features in predicting the progression to severe conditions (left), which were frequently reported as important severity and mortality risk predictors in previous studies<sup>17,18</sup>. Interestingly, for pneumonia, creatinine emerged as the most significant severity predictor, surpassing LDH and NLR, which also aligns with previous studies<sup>19</sup>. While COVID-19 and pneumonia are both respiratory diseases, the feature importance results show that significant difference exist between the two. Collectively, we observed that CovSF predicted severity with reasoning similar to that of clinicians, while also capturing valuable representations that effectively reflected the complex interactions between clinical features acquired over a course of multiple days.

### Application of CovSF on discovering severity progression related clinical features

To first compare the differences between the three severity groups, patients who classified as CONS groups were subject for analysis. The clinical feature values in the CONS groups are shown in Figure 2A. Every feature, except creatinine, showed significantly different levels between a pair of severity groups. There are a plethora



**Fig. 2.** Clinical characteristics across the severity and progression groups. Values that were interpolated during preprocessing were excluded from the analysis. **(A)** Box plots and results from statistical tests between the CONS groups for each clinical feature in the KNIH cohort. **(B)** Correlations derived from feature values and CovSF, averaged by each severity group during progression, in KNIH and KNUH. We calculated the correlation between the CovSF score and feature values across three severity levels (Mi,M,S) and (S,M,Mi), using their average values from each time point within each level to form a three-dimensional vector. The upper figure represents Mi-M-S subgroup in DP groups, while the lower figure represents S-M-Mi subgroup (RP). **(C)** The correlation plot between feature values and CovSF scores across all time points in DP and RP groups of KNIH.

of studies regarding each feature’s association to COVID-19 clinical outcomes, which well aligned with our results in general. For example, the S group showed significantly low SPO2 levels compared to the Mi group, which is commonly recognized as predictors for the need for mechanical ventilation<sup>20,21</sup>. It was shown that an LDH level higher than approximately 360 U/L provides the highest sensitivity for predicting COVID-19-related death<sup>17</sup>, which also aligned with our findings. Here, we observed that the LDH levels were distributed between 350 U/L and 500 U/L for moderate cases, and exceeded 500 U/L for severe cases<sup>22</sup>, which was clearly observed in our results. At normal conditions, NLR ranges from 1 to 3, whereas an NLR of 6 to 9 indicates mild stress, and an NLR over 9 suggests critically ill patients<sup>23,24</sup>, which all agreed with our CovSF scoring scheme. CRP is well known as a biomarker representing the degree of inflammation, and it acts independently, unaffected by other factors such as age and sex<sup>25,26</sup>. Also, platelet count is related to inflammation, and an decreased platelet count can be a sign of inflammation<sup>27</sup>. CRP levels were distributed from high to low across the S to Mi groups, whereas platelet counts showed an inverse pattern. As our study was conducted on a large cohort, it complements previous research on CRP and platelet count by reproducing their results, addressing the limitations of smaller sample sizes. Collectively, the CovSF scoring scheme for stratifying patients to the three severity groups showed to well reflect the findings of previous studies and can be easily applied to real world situations. It is also evident that incorporating laboratory features is important to elucidate patients’ immune responses and inflammation levels, along with vital sign features that effectively reflect dynamic symptoms, showing advantage over the other scoring methods, such as NEWS, WHO and SOFA.

While understanding the static characteristics or a snapshot of a patient’s severity is important, identifying trends in disease progression as severity evolves in real time is equally crucial. After the pandemic, it became evident that COVID-19 involves heterogeneous multi-organ responses, with severe outcomes triggered by various factors such as comorbidities, age, and sex<sup>28,29</sup>. Given the heterogeneity of clinical features and the multivariate inputs used in our model, it is unlikely that clinical features would change with the same tendency

simultaneously. For example, some DP patients may exhibit decreasing WBC counts while having elevated LDH levels. However, some clinical features displayed a high correlation with the CovSF scores among patients who experienced Mi-M-S (DP) or S-M-Mi (RP). As shown in Figure 2B, patients showed an increase in LDH, NLR, and breath levels during DP and decreases during RP, respectively. Based on the Figure S4, interestingly, more than half of the patients exhibited decreased hemoglobin levels<sup>30</sup> and increased WBC<sup>31</sup> during DP. However, during RP, the body temperature and BUN levels<sup>32</sup> decreased. The majority of features lack studies conducted in a longitudinal manner, as research usually focuses on data collected at admission. Further research using defined progression, such as CovSF, may provide a better understanding of diseases and enable the appropriate use of medicines to control progression-related features. Besides the Mi-M-S and S-M-Mi groups, correlations obtained between the CovSF scores and feature values, targeting all time points included in DP and RP groups in KNIH, are shown in Figure 2C.

### Application of CovSF on discovering severity progression related genes

The phenotypical characteristics observed in the clinical data are the results of their underlying biological immune response, which can be observed through the transcriptome data by quantifying the RNA expression level of immune related genes. Here, we identified Progression-related DEGs (PDGs), referring to genes whose expression levels exhibited increasing or decreasing patterns along DP and RP. PDGs were defined as differentially expressed genes (DEGs) between the Mi-S and S-Mi samples of DP and RP groups, respectively, which also showed significant correlation with the CovSF score. The detailed description of PDGs is provided in the Methods section. The PDGs were examined at the single-cell level, particularly in monocytes, as immune response to COVID-19 was strongly observed in them<sup>33</sup>. Additionally, patients with at least one sample at each severity stage during progression were included in the correlation analyses, thereby enhancing generalizability (Figure S5 and Figure S6.)

The single-cell transcriptome of the four patients with DP is depicted as a Uniform Manifold Approximation and Projection (UMAP) in Figure 3A. The Amphiregulin (*AREG*) gene is one of the 78 PDGs whose expression level increased as the patients' condition worsened. Its expression elevated in monocytes over time, progressing from Mi to M and to S stages as shown in Figure 3B. *AREG* is known to be involved in wound repair and inflammation resolution, induced by IFN-I signaling, and associated with immunosuppressive environments in chronic infections and cancer, leading to increased expression in severe COVID-19 patients<sup>34</sup>. While a significant correlation between *AREG* and severity score was not observed in the RP group, *AREG* exhibited a decreasing trend in expression as severity scores declined as shown in Figure 3E. In other words, the *AREG* gene demonstrates an expression level progression similar to the patients' clinical severity, and we identified *RNASE1* and *RNASE2* as additional genes with comparable patterns. Both genes exhibited decreased expression levels as patients' conditions improved; however, their overall trend in the DP and RP groups reflected the progression of severity scores (Figure S7). In addition, previous studies have shown that both genes show differential expression levels in monocytes, with higher expression observed in progressive (severe) groups compared to stable (mild) groups<sup>34</sup>. In contrast, the expression level of Interferon Alpha Inducible Protein 27 (*IFI27*), another PDG, decreased as the condition of the patients worsened.

The single-cell transcriptome of the six RP patients is depicted in Figure 3C. *IFI27* showed reducing expression levels during RP as shown in Figure 3D. Interestingly, instead of showing an opposite trend in DP, the trend was similar, and thus exhibited a continuously decreasing pattern from DP to RP. Since *IFI27* belongs to the interferon stimulated gene (ISG) set and is known to be an early responder to an infection, other ISGs were searched and found that many of them followed such characteristics<sup>35,36</sup>. While it is associated with the initial stage of COVID-19, studies indicate that its expression decreases over time, showing no direct correlation with longitudinal changes in severity scores<sup>37</sup>. This aligns with our findings, where *IFI27* expression declined during the DP despite increasing severity scores and continued to decrease in the RP as severity scores decreased. Similarly, other ISGs identified in our analysis, including *ISG15*, *ISG20*, *IFI35*, *IFI44*, *IFI44L*, *IFI6*, *IFIH1*, *IFIT1-5*, *OAS1-3*, and *OASL*, also showed declining patterns during DP but showed constant progression in the RP (Figure S8). This suggests that ISGs primarily reflect early or acute immune activation rather than sustained severity dynamics.

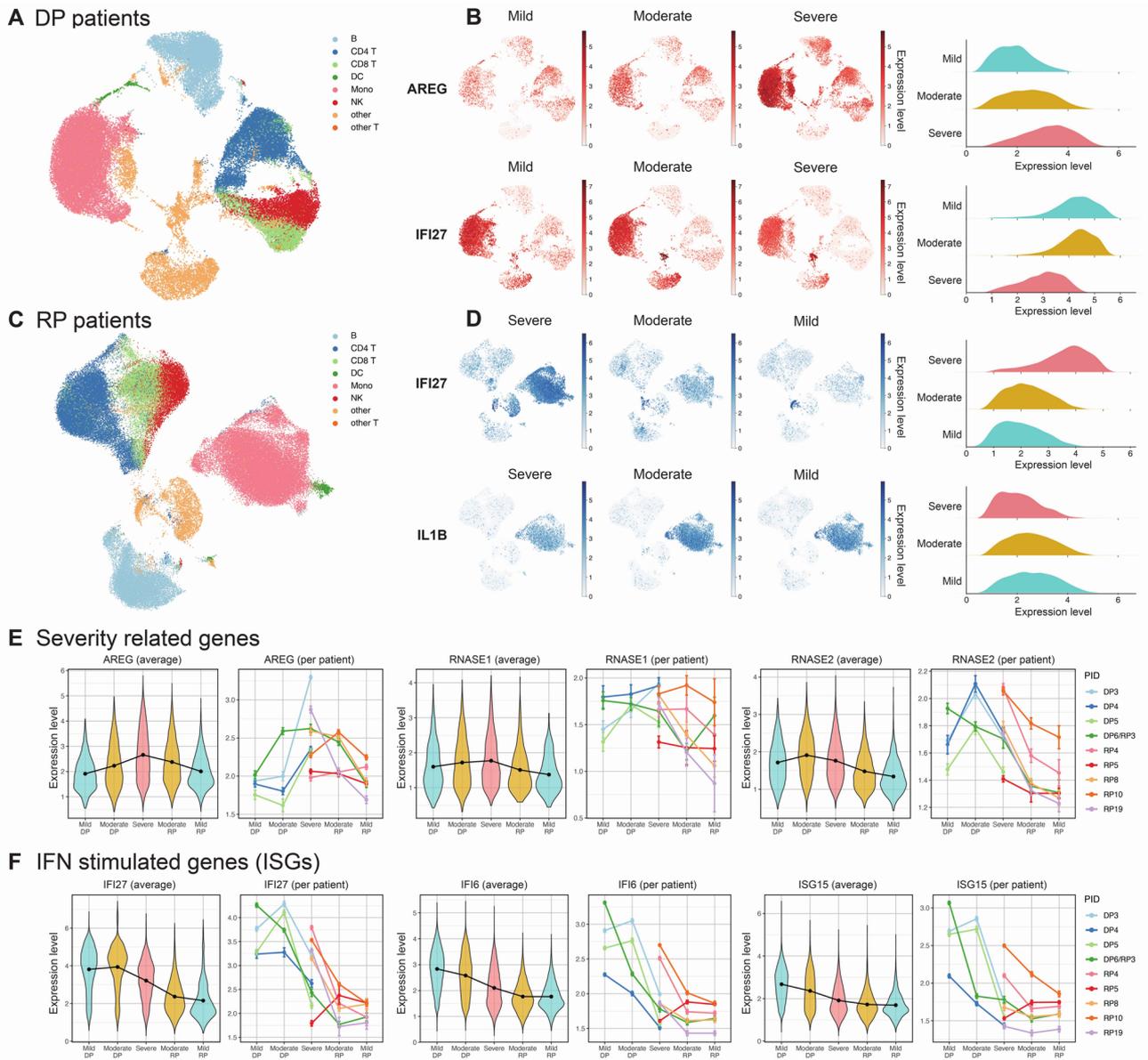
In contrast, Interleukin-1 beta (*IL-1B*) showed elevating expression levels. *IL-1B* is an inflammatory cytokine that plays a dual role in driving inflammation and tissue recovery, including lung alveolar stem cell-mediated regeneration during recovery from influenza virus-induced lung damage<sup>38</sup>. Previous studies have demonstrated that *IL-1B* expression patterns increase approximately two weeks post-symptom onset in COVID-19 patients<sup>39,40</sup>. This observation aligns with our findings, considering the timing of RP sample collection, further supporting the association of *IL-1B* with recovery processes in COVID-19. Similarly, *TNF*, another cytokine with a role similar to that of *IL-1B*, demonstrated expression patterns consistent with increased levels during the RP in our study (Figure S7).

## Discussion

### Transcriptome response differs during deterioration and recovery

Although we have shown use cases of CovSF in context of severity analysis, we like to further highlight the strength of longitudinal severity progression analysis. While genes are expected to have similar expression profiles within the same severity group we show that this is only partially true. Here we show that the transcriptomic profiles differ between the DP and RP samples even when their severity levels were the same.

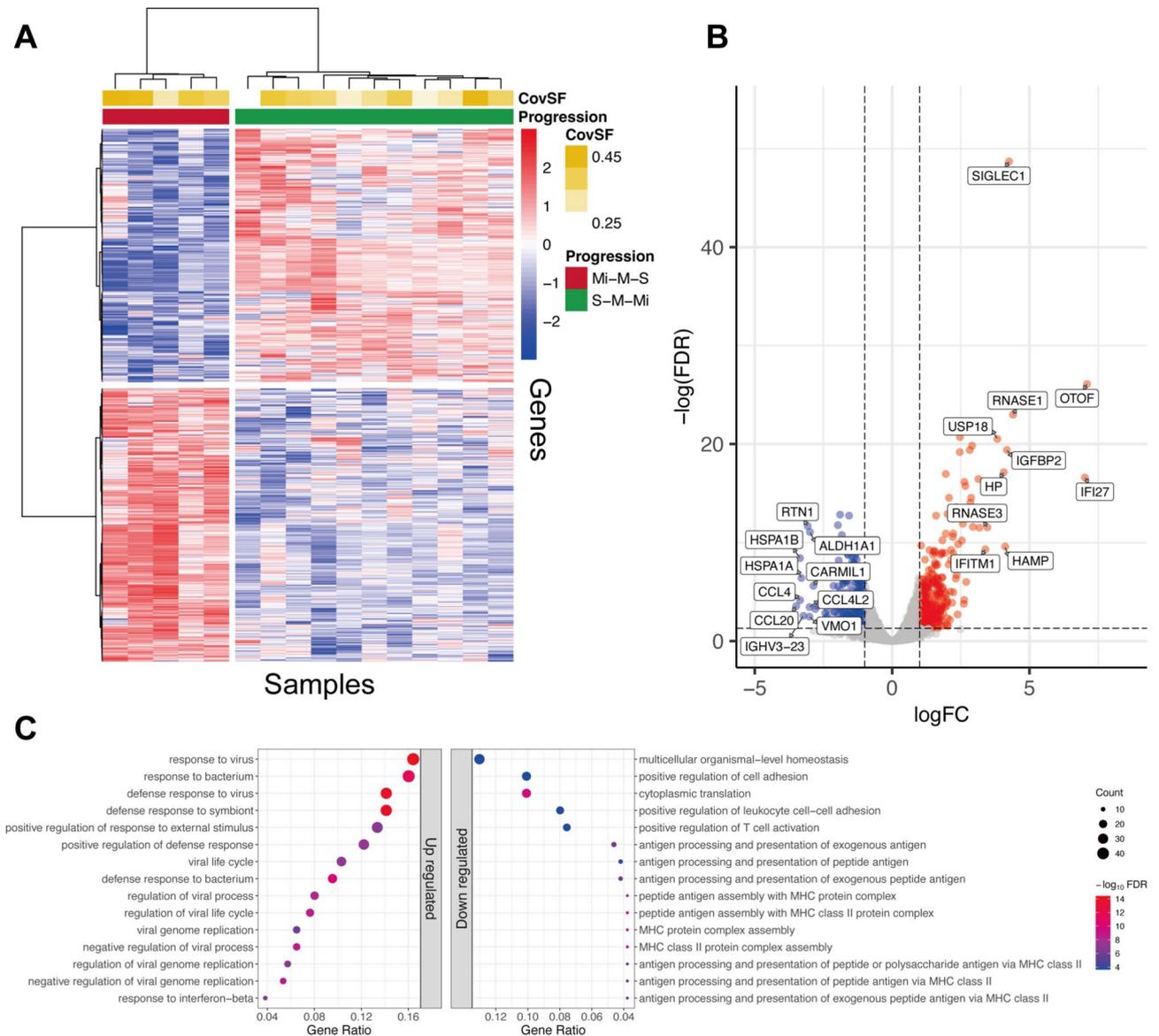
For the DP (Mi-M-S) and RP (S-M-Mi) patients, DEG analysis was performed between the M samples in the two progression groups. 316 and 293 genes were found to be significantly up and down regulated in the moderately severe DP monocytes compared to the moderately severe RP monocytes, which number is even greater than the DEGs found between the different severity levels (i.e., Mi vs S or S vs Mi). The overall analysis



**Fig. 3.** Single-cell analysis of PDGs for each DP (Mi-M-S) and RP (S-M-Mi) group. **(A)** UMAP of samples from DP patients, colored by cell types. **(B)** UMAP and ridge plots of *AREG* and *IFI27*, whose expression levels increase and decrease, respectively, during DP. **(C)** UMAP of samples from RP patients, colored by cell types. **(D)** UMAP and ridge plots of *IFI27* and *IL-1B*, whose expression levels decrease and increase, respectively, during RP. **(E)** Boxplots for the expression levels of genes with tendencies related to severity progression. **(F)** Boxplots for the expression levels of ISGs. To enhance visualization, cells with zero expression for each gene are excluded from the ridge and boxplots.

of up-regulated and down-regulated genes is presented in Figure 4, along with the associated genes and enriched pathways. The list of PDGs and the DEG results, comparing samples with M severity between DP and RP across different cell types are provided in the Table S3 and S4.

Such results were expectable, since the immunological response would definitely differ during deterioration and recovery. During deterioration, or early stages of a viral infection, the immune response will focus to detect, limit and control the spread of the virus, which responses are rapid and highly coordinated. During early infection, the immune system induces the expression of interferons (IFNs), which are translated into action by ISGs. Recent studies have shown that patients or mice with dysregulated interferon responses experienced severe clinical outcomes, whereas mild cases exhibited relatively earlier activation of interferon responses, leading to viral clearance<sup>41</sup>. This aligns well with our results, as pattern recognition receptors, which detect viruses and initiate immune responses, such as *DDX58*, *DHX58*, *ZBP1*, and *AIM2*, were up-regulated in the M of DP patients. Dozens of ISGs also showed high expression levels, including *IFI27*, *IFI6*, *IFI44L*, *IFITM3*,



**Fig. 4.** DEG results comparing the M samples between DP (Mi-M-S) and RP (S-M-Mi) groups. **(A)** Heatmap showing the expression levels of DEGs, with each sample labeled by progression. **(B)** Volcano plot of DEGs, with the top 10 up-regulated and top 10 down-regulated genes having the highest absolute logFC are labeled. **(C)** Gene Ontology Biological Process enrichment analysis of up-regulated and down-regulated genes.

*OAS1*, *OAS2*, *OAS3*, *OASL*, *MX1*, and *MX2*. As a result, IFNs were stimulated even in M beyond Mi, which can be considered dysregulated interferon responses. It gave rise to excessively high levels of cytokines, acting in a pro-inflammatory manner, commonly known as a cytokine storm. Pro-inflammatory related genes showed up-regulated expression levels, with *IL6* and *IL10* being supported by many studies consistent with ours<sup>42,43</sup>. Interestingly, *S100A8* and *S100A9* also cause cytokine storms, and a recent study demonstrated that their elevation leads to critical outcomes in patients, suggesting their potential as biomarkers<sup>44,45</sup>. As a result, it might be assumed that these activated mechanisms in M resulted in lung injury, such as acute respiratory distress syndrome and multi-organ damage, coinciding with our clinical analysis.

In contrast, during recovery the immune system will promote tissue repair and resolution to restore damaged tissue, clearing remaining pathogens and finally re-establish homeostasis, which biological mechanism clearly differs from the above. Of course, the adaptive immune response would also take place during or after recovery. The *PROS1* gene encodes Protein S, which is located in basal cells and plays a role in blood coagulation. Although the relationship between *PROS1* and COVID-19 has not been extensively investigated, Simakou et al.<sup>46</sup> suggested that *PROS1* down-regulates *CXCL10/11* and *S100A8/9*, contributing to cell generation and protecting against inflammation. This aligns with our results, indicating that the implications of the cytokine storm were alleviated as *PROS1* was expressed during RP.

CovSF is a deep learning model developed with the support of the Ministry of Health and Welfare of South Korea to overcome difficulties caused by a shortage of ICUs during a pandemic. CovSF was trained using time-series EHR data, utilizing 15 routinely measured clinical features to enhance adaptability across institutions and predict the required oxygen treatment three days in advance. CovSF achieved an AUROC of up to 0.92 in the KNIH external validation cohorts, raising expectations for enhanced sensitivity and specificity through an appropriate binary threshold to distinguish between mild and severe oxygen treatment. Furthermore, the model's output reflected the complex interactions between features in a longitudinal manner, enabling us to define a severity score that facilitates stratifying patient severity and its progression for further downstream analysis. From a clinical perspective, we validated the scores by identifying the presence of a significant tendency or difference in actively researched features between severity or progression groups, further supporting other studies that lack large participant cohorts. Additionally, despite the model being driven by clinical features, we suggested PDGs whose expression levels increased or decreased during severity progression and DEGs, observed between the same severity but different progression, based on patient-matched scRNA-seq samples. We confirmed that some PDGs and DEGs conform to existing studies and highlighted the importance of conducting further research on them. The COVID-19 pandemic, which had a critical impact on the global healthcare system, has ended. However, further investigation into biological mechanism or biomarkers of COVID-19 remains vital due to its multi-organ response and heterogeneity, particularly in severe cases<sup>47</sup>. Collectively, a severity assessment system like CovSF, particularly covered with the time axis, may provide an innovative and comprehensive analysis perspective, like severity progression. We are looking forward to our strategy, which started with a severity prediction model based on clinical data and expanded to transcriptomics, being leveraged for other infectious or respiratory diseases.

It is essential to recognize the limitations of our study, especially when predicting and gauging patient severity. We only utilized timely measured clinical features, but static information might have played a role in severity and its prediction. For example, Age is known to be a major factor contributing to severe outcomes in COVID-19, as the elderly have shown greater vulnerability in terms of inflammation and immune function<sup>48</sup>. We also reviewed prediction performance in KNIH cohort across different age groups (Figure S9C). A slight disparity was observed in sensitivity and specificity for the relatively oldest and youngest age groups, warranting some caution. However, the overall performance, represented by an AUROC of approximately 0.93 across all age groups, suggests that it can be addressed by applying different thresholds for each age group. Additionally, patients with comorbidities are relatively more likely to experience severe outcomes and progression than others<sup>49</sup>. Hypertension, diabetes, cardiovascular diseases, and cerebrovascular diseases have been reported as significant conditions associated with severe manifestations of COVID-19<sup>50</sup>. A total of 82 patients had 51 comorbidities in the KNIH cohort, where Table S5 describes how CovSF discerned severity for each disease type compared with the oxygen treatment they actually received. On the whole, our score accurately assessed the severity of patients with comorbidities, closely matching oxygen treatment, besides for metabolic/endocrine disorders and respiratory disorders. However, it was somewhat mismatched with cardiovascular and circulatory disorders, requiring additional analysis and caution. At last, since our model was trained and validated on the datasets including only individuals of Korean ethnicity, future studies should include more diverse ethnic groups to enhance its generalizability. Additionally, our study was conducted in a retrospective manner, then we aim to test the model by linking it with a bed allocation system in the hospital where the data were collected. We also provide the code to operate CovSF in the terminal, addressing these issues.

## Methods

### Data preprocessing

The clinical data is comprised of a wide range of features that typically vary between different hospitals. To make the model as applicable as possible in real world, we selected clinical features that are related to symptoms of infection and are highly likely to be routinely measured in general hospitals. Since the objective is to learn the dynamics of the selected features, which is vital for forecasting a patient's severity progression, only features measured in time-series were considered. Six features related to vital signals (i.e., body temperature, SBP, DBP, SPO<sub>2</sub>, pulse and breath) and 9 laboratory test features (i.e., BUN, creatinine, hemoglobin, LDH, neutrophil, lymphocyte, platelet count, WBC count and CRP) were selected to train CovSF. A total of five oxygen treatment types were used as the model's prediction labels, categorized into mild ("room air" and "nasal") and severe ("mask," "HFNC," and "ventilation") treatment groups for binary classification.

The raw clinical datasets collected from the three cohorts were preprocessed through three phases, which were used to train, validate and test CovSF. In the first phase, the feature names and their units were standardized across the three cohorts. Here, outpatients were excluded, so that only inpatients with admission and discharge records over multiple days were included. During this process, patients transferred between Kyungpook National University Chilgok Hospital (KNUCH) and Kyungpook National University Hospital (KNUH) were observed, whose clinical data were integrated as one. The KNU cohort consist of patients from the KNUCH and KNUH institutions. Furthermore, outliers (e.g. hand-written error) from each feature were filtered out. The output of the first phase contains the values of the 15 clinical features measured at multiple time points, including the days of admission and discharge for each patient. In the second phase, the output of the first phase was converted into longitudinal data in units of days to compose a data structure of patient × date × feature. For cases where vital signals were measured multiple times during a single day, the feature values were averaged so that only a single data entry exists for a single day. For these cases, the worst severe oxygen treatment type in that day was selected as the target prediction label. In the third phase, missing values were imputed, which mainly consisted of laboratory test features. Compared to the daily measured vital signs, laboratory tests are performed with multiple days interval, thus yielding missing values when matched to the vital sign data. Thus, the laboratory test features were adjusted to daily units by inserting empty vectors between data points to match the timeline with the

vital sign features, which were subject for imputation. To avoid excessive imputation, patients with at least two laboratory test measurements in the final phase were retained. For the retained dataset, linear interpolation was conducted by filling any missing information between two adjacent data points for each variable individually. Finally, neutrophils and lymphocytes were converted into the NLR ratio by dividing neutrophils by lymphocytes. Collectively, longitudinal clinical datasets were collected for 4,509, 443, and 33 patients from the KNU, KNIH and NCCAP cohorts, respectively. KNU cohort was used as the training and internal validation set for CovSF. 10-fold cross-validation was employed to ensure that each set maintained a balanced ratio of labels, stratified by the oxygen treatment level for each patient<sup>51</sup>. During this process, all feature values were normalized using a robust scaler, which transforms features by subtracting the median and dividing by the interquartile range. This fitted scaler was subsequently applied to the external validation set during the testing phase.

#### Model development

CovSF takes 15 clinical features of a COVID-19 infected patient of up to five days to forecast the patient's severity to up to three days after. We first define  $X$  as a set of sequences of the clinical features that are measured in units of days of a single patient. A  $X_{ij}$  refers to a sequential measurement of the 15 clinical features starting from day  $i$  until day  $j$  with a fixed interval of one day. Then, the model was trained using clinical data  $X_{ij}$ , representing daily observations with lengths ranging from  $i$  to  $j$ , to forecast the severity levels at days  $j$  to  $j + 3$ , which we denote as  $Y_j^0, Y_j^1, Y_j^2$  and  $Y_j^3$ . Thus,  $Y_j^d$  refers to the probability of requiring severe oxygen treatment at day  $j + d$ , where  $d = 0, \dots, 3$ . For each patient, every possible  $X_{ij}$  of length from 2 to 5 in the training dataset (i.e., KNU) were collected by sliding-window with an interval of 1, which were used to train CovSF. At last, prediction labels, comprising mild and severe oxygen treatment groups, were assigned as 0 and 1, respectively.

CovSF is a deep-learning based model, whose structure is made of an encoder and a decoder<sup>52</sup>. Both the encoder and the decoder have their own recurrent neural component. Most of the recurrent neural components, such as Recurrent Neural Network (RNN)<sup>53</sup>, Long short-term memory (LSTM)<sup>54</sup> and Gated Recurrent Unit (GRU)<sup>55</sup>, recursively update their hidden states at each input time point, using their previous hidden states and current input vectors. We utilized GRU as the main recurrent neural component for the encoder and decoder to mitigate the long-term dependency issue commonly encountered in RNNs, while requiring significantly fewer parameters than LSTM. GRU updates its hidden states as below.

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \quad (1)$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \quad (2)$$

$$\underline{h}_t = \tanh(W x_t + r_t \circ U h_{t-1}) \quad (3)$$

$$h_t = (1 - z_t) \circ \underline{h}_t + z_t \circ h_{t-1} \quad (4)$$

Here,  $x_t$  is the input vector at the current time point  $t$ , which is a single data entry of  $X_{ij}$ , where  $t = i, \dots, j$ .  $h_t$  and  $h_{t-1}$  represents the hidden states of the current and previous time point  $t$  and  $t - 1$ , respectively.  $W, W_z, W_r, U_z, U_r$  and  $U$  are the trainable parameters. The GRU has two gates, a reset gate that decides how much information to forget from the previous time point (Equation 1). The closer  $r_t$  is to 0, the more it forgets from the previous information. The other is the update gate, which determines how much previous information to retain (Equation 2). The closer  $z_t$  is to 1, the more it passes information from the previous steps and vice versa. Then GRU calculates the candidate hidden states  $\underline{h}_t$  by performing the point-wise product with  $r_t$  (Equation 3). Finally, GRU updates its hidden states  $h_t$ , where the term  $(1 - z_t) \circ \underline{h}_t$  adjusts how much new information to incorporate, and  $z_t \circ h_{t-1}$  decides how much of the previous hidden states to retain (Equation 4). Consequently, the model's encoder and decoder operate as follows:

$$h_j = Encoder(X_{ij}) \quad (5)$$

$$Y_j^d, h_j^d = \begin{cases} Decoder(v_{1 \times 2}, h_j) & \text{if } d = 0 \\ Decoder(Y_j^{(d-1)}, h_j^{(d-1)}) & \text{if } 0 < d \leq 3 \end{cases} \quad (6)$$

When encoder receives an input sequence  $X_{ij}$ , the GRU within the encoder maps it to the latent-space, or hidden state,  $h_j$  by recursively updating its hidden states (Equation 5).  $h_j$  is used as the initial hidden states of the GRU within the decoder when  $d = 0$  in addition with the zero vector  $v_{1 \times 2}$ , since there are no previous predictions.  $h_j^{(d-1)}$  is used as the hidden states of the GRU for any  $0 < d \leq 3$ , which implies that the previous information is considered for predicting  $Y_j^d$  as defined in (Equation 6), where  $d$  refers to the number of days after day  $j$ . So, when the decoder predicts  $Y_j^d$ , the predicted value  $Y_j^{(d-1)}$  of the previous time step  $d - 1$  is input to GRU. For each prediction to generate the probability of requiring severe oxygen treatments,  $h_j^d$  passes through a classification unit consisting of a Gaussian Error Linear Unit<sup>56</sup> activation function between two fully connected layers, followed by a softmax function after the final layer to map the last two-dimensional vector values into probabilities for mild and severe oxygen treatments. In the CovSF, this classification unit is shared across all time points for prediction.

Various statistical and deep learning methods were proposed to similarly predict one's severity level in terms of oxygen treatment type. However, they either require a very small set ( $<4$ )<sup>10,11</sup> or a very large set ( $>50$ )<sup>12,13</sup> of clinical features. Practically, we find that it is advantageous to leverage as many clinical features as possible that are routinely measured upon admission even in pandemic situations, which falls below twenty features. Hence, the aforementioned methods do not fit such criteria and were not considered for comparison. Therefore,

to assess the model's generalization, we generated another model (i.e., Vanilla), an ensemble of independent GRU models, where each GRU receives the same input sequence but predicts different time steps, separately. Thus, a total of four GRUs, one for each  $d$ . For each ensemble, a classification unit with the same structure as in the CovSF exists independently, differing from the CovSF. Collectively, two GRU models were developed and trained, with validation performed across different recurrent neural components, including GRU, LSTM, and RNN, for each model (Figure S9E). Additionally, to compare the performance of linear models, we included two regressors, logistic regression and LASSO logistic regression. They require a fixed input size and can predict severity for only one day, so we used five-day data for training and validation and structured a separate sub-model for each prediction day to compose the complete model.

For training the deep learning models, we used RAdam<sup>57</sup> as the optimizer, starting with a learning rate of 0.001, and applied a ReduceLRonPlateau scheduler with a patience of 7. These were performed over 70 epochs with a batch size of 32, but early stopping was employed if there was no reduction in internal validation loss for 10 consecutive epochs. We conducted training with Asymmetric Loss (ASL)<sup>58</sup>, which is similar to focal loss but assigns different weights to positive and negative samples to effectively address class imbalance issues.

$$L_+ = (1 - p)^{\gamma^+} \log(p) \quad (7)$$

$$L_- = p_m^{\gamma^-} \log(1 - p_m) \quad (8)$$

$$p_m = \max(p - m, 0) \quad (9)$$

Among the positive labels, which represent the severe oxygen treatment groups in our study, the loss is more heavily penalized with weights  $(1 - p)^{\gamma^+}$  for rare positive samples that are difficult to predict (Equation 7). Likewise, loss for negative labels (i.e mild oxygen treatment groups) were weighted by  $p_m^{\gamma^-}$  (Equation 8). To prevent over-penalization of well-classified negative samples, a clipping mechanism is applied, using  $p_m$  instead of  $p$  at negative cases (Equation 9).

In the CovSF based on GRU with baseline parameters,  $\gamma^-$  and the clipping threshold ( $m$ ) were selected by grid search, optimizing the average of AUROC and F1 scores across 10 folds. Finally, under the same process but using the selected loss parameters, we also determined the parameters for the GRU within the CovSF. Specific metrics for the ASL loss and the model are presented in Figure S9 A and B, respectively, showing that the final model consisted of two layers with 32 hidden units, trained with 0.075 clip and 1.5  $\gamma^-$  in the ASL Loss. Other models were also trained using the same parameters and environments. Finally, We conducted an ablation study to validate our feature selection. The CovSF with GRU was trained on datasets consisting of only laboratory or vital sign features, and its performance is shown in Figure S9F.

#### CovSF score definition and subtyping

A maximum of four predictions can be made for the  $j$ th day after admission, using  $Y_{j-3}^3, Y_{j-2}^2, Y_{j-1}^1, Y_j^0$  with sequences  $X_{i(j-3)}, X_{i(j-2)}, X_{i(j-1)}$  and  $X_{ij}$ . The average of  $n$  predictions was taken as the final prediction for day  $j$  for some input  $X_{ij}$ , when multiple predictions were available as shown in Equation 10. We defined this average score as the CovSF Score, which effectively reflects multiple predictions of the probability for requiring severe oxygen treatment using a robust deep learning model. Its range is also from 0 to 1 and can be assigned to every hospitalized day of the in-patients we used.

$$\text{score}(j) = \frac{\sum_{d=0}^{n-1} Y_{j-d}^d}{n} \quad (10)$$

$$n = \min(j + 1, 4) \quad (11)$$

Here, each  $Y_{j-d}^d$  is computed by  $X_{i(j-d)}$  with any  $j - 4 < i \leq j$ . Since the maximum length of  $X_{ij}$  is restricted to 5,  $n$  takes a range from 1 to 4 as shown in Equation 11.

Then, in the KNUH cohort, we determined the threshold to discriminate between the mild and moderate groups using the F1-Score, calculated based on labels where 0 included ROOM AIR and 1 included others, as oxygen treatment devices are essential starting from nasal treatment. Similarly, the threshold separating the moderate group from the severe group is selected in the same manner, with ROOM AIR and Nasal set as 0 and others as 1, consistent with the model training. Specific evaluation results for selecting the severity threshold are presented in Figure S2. As a result, time points where CovSF is below 0.2 are classified as Mi, those below 0.5 as M, and the rest as S.

In a longitudinal manner, it is possible to list the severity levels that a patient experienced from admission to discharge. Then, we removed successive duplicate severity levels so that, for example, if a patient experienced severity levels as [M, M, Mi, Mi, Mi, M, S, S], it was converted to [M, Mi, M, S] first. After that, by observing 2 or 3 adjacent severity levels in sequences, we extracted one recovering progression (M-Mi), where the patient's severity decreased, and one deteriorating progression (M-Mi-S), where the severity conversely increased. In these manners, we broadly identified three groups as CONS, DP, and RP in the KNUH and KNIH cohorts. The CONS group consists of patients who belong to only one severity level during hospitalization, specifically classified into Mi, M, and S subgroups. The DP group comprises the Mi-M, M-S, and Mi-M-S subgroups, while the RP group includes subgroups named M-Mi, S-M, and S-M-Mi. The number of patients belonging to each subgroup, and the statistics for the length of admission days for each severity within each subgroup are presented in Table S1. Especially for the KNIH cohort, the count of scRNA-seq samples retrieved from time points belonging to certain subgroups are included.

### Feature importance

To understand the prediction in terms of clinical features, the contribution of each feature towards making a prediction was measured, which is often referred to as the feature importance. The feature importance provides insight on what features are related to the outcomes of a prediction, which is the severity in our case. The Integrated Gradients (IG)<sup>59,60</sup> was used to measure the importance of each feature at each day from  $i$  to  $j$  for an input  $X_{ij}$ . The IG of the feature  $z$  is calculated as in Equation 12.

$$IG_z(X_{ij}) \approx (x_z - x'_z) \times \int_{\alpha=0}^1 \frac{\partial F(X'_{ij} + \alpha \cdot (X_{ij} - X'_{ij}))}{\partial x_z} d\alpha \quad (12)$$

Here,  $F$  is a function that maps the input  $X_{ij}$  to an interval of  $[0, 1]$ , signifying a deep network.  $IG_z$  computes the integral of the gradients along the straight line path from the baseline  $X'_{ij}$  to the input for the input value of feature  $z$ . The IG for each clinical feature  $z$  across all possible input sequences within three cohorts are computed. Here, the baseline was set as the mean value of each clinical feature at time points where severity was mild in our CovSF measures, and it was determined individually for each cohort. After calculation using IG, we observed each feature's contribution to the increased requirement for severe oxygen treatments, and when multiple IGs were evaluated to the same days, we simply used the averaged values.

### Single-cell RNA-seq analysis

For downstream analysis across different cell types, cell-type annotation for each sample was required. We conducted the pre-processing using the R package Seurat<sup>61</sup> in a sample-wise manner. Initially, cells that expressed fewer than 200 genes or exhibited mitochondrial gene expression exceeding 15% were excluded to remove low-quality cells. Subsequently, to rectify potential quality issues related to the technical aspects of sequencing, we employed the SCTransform function to perform quality correction via linear regression on each of the filtered sample matrices. Then, 3,000 highly variable genes (HVG) shared across the samples were selected by using the SelectIntegrationFeatures function. Finally, Azimuth's automated cell-type annotation<sup>62</sup> was utilized for cell-type annotation, and cells with a mapping or predicted score below 0.6 were excluded. Based on these cell types, pseudo-bulk were generated for differentially DEG analysis, and these label were also utilized to distinguish cell types for cell-level analysis.

DEG analysis was incorporated as one of the steps to identify PDGs and was also used to compare samples with M severity between Mi-M-S (DP) and S-M-MI (RP) groups. We followed the standard edgeR<sup>63</sup> workflow. First, we removed genes with low expression levels across samples using the filterByExpr function and normalized the expression values using the Trimmed Mean of M-values normalization method. We estimated each gene's dispersion by computing a robust estimate of the negative binomial dispersion using the estimateGLMRobustDisp function. Then, these dispersion parameters were utilized to fit the generalized linear model, with severity groups and, additionally, the sex information of patients who generated each sample considered as independent variables in these steps. Finally, we performed a likelihood ratio test for each gene to identify statistical differences in expression levels between the reference severity groups and the target severity groups, with p-values adjusted for multiple testing to control the false discovery rate (FDR). DEGs used in this study were selected based on the criteria of  $|\log_2 \text{FoldChange}| \geq 1$  and  $FDR \leq 0.05$ . Based on the DEGs, we performed pathway enrichment analysis using the ClusterProfiler<sup>64</sup> package on datasets including Biological Process, Kyoto Encyclopedia of Genes and Genomes, and Reactome Pathway<sup>65</sup>.

### Progression related DEGs (PDGs)

We defined DEGs as PDGs whose expression levels significantly increase or decrease during progression and highly correlated with the CovSF score. First, DEG analysis was performed on each of the Mi-M-S (DP) and S-M-Mi (RP) groups in the pseudobulk monocytes. Patients who had at least one matched scRNA-seq sample during the progression were selected. For DP patients, DEGs were searched between the Mi and S samples, and the S and Mi samples for RP patients, respectively. In the DP patients, a total of 152 genes were up and 314 down regulated in S compared to Mi (Figure S6B). In RP patients, a total of 30 genes were up and 31 down regulated in Mi compared to S (Figure S6E). The DEGs showed to well discriminate the severity groups in both the DP and RP patients with clear difference in CovSF scores.

Next, genes whose expression level highly correlated with CovSF were searched within DP and RP patients with scRNA-seq samples covering the entire progression, resulting in four DP and six RP patients. As a result, 762 and 783 genes showing significant positive and negative correlation with CovSF in the DP patients were searched (Figure S6A). A total of 362 and 214 genes with significant positive and negative correlation with CovSF were searched in the RP patients (Figure S6D). The intersection between the DEGs and correlating genes are shown in Figure S6C and F, which were further investigated for the analysis of severity progression. As a result, 311 and 22 genes were found in the DP and RP groups, respectively, which we refer to as PDGs. We find that the difference in gene numbers between the DP and RP groups stem from the heterogeneous nature of the recovery that highly vary from person to person, which may be the reason for the smaller number of genes searched in the RP group.

## Quantification and statistical analysis

### Performance evaluation

All models were trained and tested by 10-fold cross-validation, with the training set and internal validation set split in a 9:1 ratio in the KNUH cohort. We evaluated the two models across the KNU, KNIH and NCCAP cohorts. Using the sliding-window technique, we generated all possible input sequences  $X_{ij}$ , where their length

can range from 1 to 5 and obtained the output predictions of each input sequence. So, even in cases where the length of the input sequence is shorter than the maximum input sequence length  $k=5$  (i.e., days after admission is shorter than 5), all were incorporated into the performance evaluation. For each time step  $d$  in the predict sequences, we calculated performance metrics for binary classification using a threshold of 0.5. We measured the AUROC, sensitivity, specificity. The sensitivity and specificity were derived from Equation 13. The terms TP, TN, FP and FN represent True Positive, True Negative, False Positive, False Negative, respectively. The average and 95% confidence interval of metrics for each trained model are presented in Table S2. In detail, the performance metrics of the main model for different input sequence lengths is shown in Figure S9D.

$$\text{sensitivity} = \frac{TP}{FN + TP}, \text{ specificity} = \frac{TN}{TN + FP} \quad (13)$$

#### Statistical testing

In clinical analysis, we utilized the Mann-Whitney U test to identify clinical features with significant differences between severity groups stratified by the CovSF score. This was chosen because, according to the Anderson-Darling test, all features across severity groups, except for SBP in the moderate group, did not follow a normal distribution. Then, Bonferroni correction was applied to adjust these p-values, providing more precise statistical results. To examine the relationships between CovSF and feature values or gene expression levels, correlation value was employed. When obtaining the correlation between CovSF and gene expression levels for each gene on a patient-wise basis, we used a log<sub>1</sub>p transformation after counts-per-million (CPM) normalization.

#### Data availability

The dataset of the KNIH cohort used in this study are available in the Clinical and Omics Data Archive (CODA, <https://coda.nih.go.kr>) database by the accession number CODA\_D23017. The CovSF model is made publicly available in the following github repository including a tutorial with example data, <https://github.com/cobi-git/CovSF>.

Received: 19 March 2025; Accepted: 17 June 2025

Published online: 01 July 2025

#### References

- Smith, G. B. et al. The national early warning score 2 (news2). *Clin. Med.* **19**, 260 (2019).
- Marshall, J. C. et al. A minimal common outcome measure set for COVID-19 clinical research. *Lancet Infect. Dis.* **20**, e192–e197 (2020).
- Jones, A. E., Trzeciak, S. & Kline, J. A. The sequential organ failure assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation. *Crit. Care Med.* **37**, 1649–1654 (2009).
- Long, L. et al. Effect of early oxygen therapy and antiviral treatment on disease progression in patients with covid-19: A retrospective study of medical charts in China. *PLoS Negl. Trop. Dis.* **15**, e0009051 (2021).
- Chang, R., Elhousseiny, K. M., Yeh, Y.-C. & Sun, W.-Z. Covid-19 icu and mechanical ventilation patient characteristics and outcomes—a systematic review and meta-analysis. *PLoS One* **16**, e0246318 (2021).
- Baik, S.-M., Lee, M., Hong, K.-S. & Park, D.-J. Development of machine-learning model to predict covid-19 mortality: Application of ensemble model and regarding feature impacts. *Diagnostics* **12**, 1464 (2022).
- Booth, A. L., Abels, E. & McCaffrey, P. Development of a prognostic model for mortality in COVID-19 infection using machine learning. *Mod. Pathol.* **34**, 522–531 (2021).
- Barough, S. S. et al. Generalizable machine learning approach for COVID-19 mortality risk prediction using on-admission clinical and laboratory features. *Sci. Rep.* **13**, 2399 (2023).
- Horita, N. & Fukumoto, T. Global case fatality rate from COVID-19 has decreased by 96.8% during 2.5 years of the pandemic. *J. Med. Virol.* <https://doi.org/10.1002/jmv.28231> (2023).
- Mehrdad, S., Shamout, F. E., Wang, Y. & Atashzar, S. F. Deep learning for deterioration prediction of COVID-19 patients based on time-series of three vital signs. *Sci. Rep.* **13**, 9968 (2023).
- Lee, E. E. et al. Prediction of oxygen requirement in covid-19 patients using dynamic change of inflammatory markers: Crp, hypertension, age, neutrophil and lymphocyte (chanel). *Sci. Reports* **11**, 13026 (2021).
- Park, H. et al. In-hospital real-time prediction of COVID-19 severity regardless of disease phase using electronic health records. *PLoS One* **19**, e0294362 (2024).
- Schwab, P. et al. Real-time prediction of COVID-19 related mortality using electronic health records. *Nat. Commun.* **12**, 1058 (2021).
- Zhou, K. et al. Eleven routine clinical features predict COVID-19 severity uncovered by machine learning of longitudinal measurements. *Comput. Struct. Biotechnol. J.* **19**, 3640–3649 (2021).
- Jo, H.-Y. et al. Establishment of the large-scale longitudinal multi-omics dataset in COVID-19 patients: Data profile and biospecimen. *BMB Rep.* **55**, 465 (2022).
- COVID-19 Treatment Guidelines Panel. Coronavirus Disease 2019 (COVID-19) Treatment Guidelines. Tech. Rep., National Institutes of Health (2025). Accessed: May 25, 2025.
- Li, C. et al. Elevated lactate dehydrogenase (ldh) level as an independent risk factor for the severity and mortality of covid-19. *Aging (Albany NY)* **12**, 15670 (2020).
- Kosidlo, J. W., Wolszczak-Biedrzycka, B., Matowicka-Karna, J., Dymicka-Piekarska, V. & Dorf, J. Clinical significance and diagnostic utility of nlr, lmr, plr and sii in the course of covid-19: a literature review. *J. inflammation research* 539–562 (2023).
- Huang, Y. et al. Diagnostic value of blood parameters for community-acquired pneumonia. *Int. Immunopharmacol.* **64**, 10–15 (2018).
- Mukhtar, A. et al. Admission spo<sub>2</sub> and rox index predict outcome in patients with covid-19. *Am. J. Emerg. Med.* **50**, 106–110 (2021).
- Satici, M. O. et al. The role of a noninvasive index 'spo<sub>2</sub>/fio<sub>2</sub>' in predicting mortality among patients with covid-19 pneumonia. *Am. J. Emerg. Med.* **57**, 54–59 (2022).
- Henry, B. M. et al. Lactate dehydrogenase levels predict coronavirus disease 2019 (COVID-19) severity and mortality: A pooled analysis. *Am. J. Emerg. Med.* **38**, 1722–1726 (2020).
- Zahorec, R. Neutrophil-to-lymphocyte ratio, past, present and future perspectives. *Bratisl. Lek. Listy* **122**, 474–488 (2021).

24. Yang, A.-P., Liu, J.-P., Tao, W.-Q. & Li, H.-M. The diagnostic and predictive role of NLR, D-NLR and PLR in COVID-19 patients. *Int. Immunopharmacol.* **84**, 106504 (2020).
25. Smilowitz, N. R. et al. C-reactive protein and clinical outcomes in patients with covid-19. *Eur. Heart J.* **42**, 2270–2279 (2021).
26. Wang, L. C-reactive protein levels in the early stage of covid-19. *Med. Mal. Infect.* **50**, 332–334 (2020).
27. Barrett, T. J. et al. Platelets contribute to disease severity in COVID-19. *J. Thromb. Haemost.* **19**, 3139–3153 (2021).
28. Zaim, S., Chong, J. H., Sankaranarayanan, V. & Harky, A. Covid-19 and multiorgan response. *Curr. Probl. Cardiol.* **45**, 100618 (2020).
29. Thakur, V. et al. Multi-organ involvement in COVID-19: Beyond pulmonary manifestations. *J. Clin. Med.* **10**, 446 (2021).
30. Lippi, G. & Mattiuzzi, C. Hemoglobin value may be decreased in patients with severe coronavirus disease 2019. *Hematol. Transfus. Cell Ther.* **42**, 116–117 (2020).
31. Zhu, B. et al. Correlation between white blood cell count at admission and mortality in covid-19 patients: a retrospective study. *BMC infectious diseases* **21**, 1–5 (2021).
32. Küçükceran, K., Ayrancı, M. K., Girişgin, A. S., Koçak, S. & Dündar, Z. D. The role of the bun/albumin ratio in predicting mortality in COVID-19 patients in the emergency department. *Am. J. Emerg. Med.* **48**, 33–37 (2021).
33. Merad, M. & Martin, J. C. Pathological inflammation in patients with covid-19: A key role for monocytes and macrophages. *Nat. Rev. Immunol.* **20**, 355–362 (2020).
34. Unterman, A. et al. Single-cell multi-omics reveals dyssynchrony of the innate and adaptive immune system in progressive covid-19. *Nat. Commun.* **13**, 440 (2022).
35. Shojaei, M. et al. Ifi27 transcription is an early predictor for COVID-19 outcomes, a multi-cohort observational study. *Front. Immunol.* <https://doi.org/10.3389/fimmu.2022.1060438> (2023).
36. Villamayor, L. et al. The IFN-stimulated gene *ifi27* counteracts innate immune responses after viral infections by interfering with RIG-I signaling. *Front. Microbiol.* <https://doi.org/10.3389/fmicb.2023.1176177> (2023).
37. Lei, H. A two-gene marker for the two-tiered innate immune response in covid-19 patients. *PLOS ONE* **18**, 1–21 (2023).
38. Katsura, H., Kobayashi, Y., Tata, P. R. & Hogan, B. L. Il-1 and tnfa contribute to the inflammatory niche to enhance alveolar regeneration. *Stem Cell Reports* **12**, 657–666 (2019).
39. Ong, E. Z. et al. A dynamic immune response shapes covid-19 progression. *Cell Host & Microbe* **27**, 879–882.e2 (2020).
40. Bell, L. C. et al. Transcriptional response modules characterize il-1b and il-6 activity in covid-19. *iScience* **24**, 101896 (2021).
41. Eskandarian Boroujeni, M. et al. Dysregulated interferon response and immune hyperactivation in severe covid-19: Targeting stats as a novel therapeutic strategy. *Front. Immunol.* **13**, 888897 (2022).
42. Coomes, E. A. & Haghbayan, H. Interleukin-6 in COVID-19: A systematic review and meta-analysis. *Rev. Med. Virol.* **30**, 1–9 (2020).
43. Lu, L., Zhang, H., Dauphars, D. J. & He, Y.-W. A potential role of interleukin 10 in COVID-19 pathogenesis. *Trends Immunol.* **42**, 3–5 (2021).
44. Mellett, L. & Khader, S. A. S100a8/a9 in COVID-19 pathogenesis: Impact on clinical outcomes. *Cytokine Growth Factor Rev.* **63**, 90–97 (2022).
45. Chen, L. et al. Elevated serum levels of s100a8/a9 and hmgb1 at hospital admission are correlated with inferior clinical outcomes in covid-19 patients. *Cell. Mol. Immunol.* **17**, 992–994 (2020).
46. Simakou, T. et al. Prosl released by human lung basal cells upon sars-cov-2 infection facilitates epithelial cell repair and limits inflammation. *bioRxiv* 2024–09 (2024).
47. Nalbadian, A. et al. Post-acute covid-19 syndrome. *Nat. Med.* **27**, 601–615 (2021).
48. Liu, K., Chen, Y., Lin, R. & Han, K. Clinical features of COVID-19 in elderly patients: A comparison with young and middle-aged patients. *J. Infect.* **80**, e14–e18 (2020).
49. Sanyaolu, A. et al. Comorbidity and its impact on patients with COVID-19. *SN Compr. Clin. Med.* **2**, 1069–1076 (2020).
50. Wang, B., Li, R., Lu, Z. & Huang, Y. Does comorbidity increase the risk of patients with COVID-19: Evidence from meta-analysis. *Aging* **12**, 6049 (2020).
51. Sechidis, K., Tsoumakas, G., spsampsps Vlahavas, I. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5–9, 2011, Proceedings, Part III* **22**, 145–158 (Springer, 2011).
52. Sutskever, I., Vinyals, O. & Le, Q. V. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 3104–3112 (2014).
53. Williams, R. J. & Zipser, D. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.* **1**, 270–280 (1989).
54. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
55. Cho, K. et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734 (2014).
56. Hendrycks, D. & Gimpel, K. Gaussian error linear units (gelus). arXiv preprint [arXiv:1606.08415](https://arxiv.org/abs/1606.08415) (2016).
57. Liu, L. et al. On the variance of the adaptive learning rate and beyond. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)* (2020).
58. Ridnik, T. et al. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 82–91 (2021).
59. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 3319–3328 (2017).
60. Kokhlikyan, N. et al. Captum: A unified and generic model interpretability library for pytorch. *Preprint at arXiv* (2020).
61. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
62. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
63. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. Edger: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
64. Xu, S. et al. Using clusterprofiler to characterize multiomics data. *Nat. Protoc.* <https://doi.org/10.1038/s41596-024-01020-z> (2024).
65. Yu, G. & He, Q.-Y. Reactomepa: An r/bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.* **12**, 477–479 (2016).

## Acknowledgements

This work was supported by the project for Infectious Disease Medical Safety, funded by the Ministry of Health and Welfare, South Korea (grant number: RS-2022-KH124555 (HG22C0014)), the Korea National Institute of Health (KNIH) research project (project no. 2024-ER-0801-01) and the Development of heterogeneous health-care data and artificial intelligence project (project No. 2024-NI-009-00). This research was also supported by a grant of the project for ‘Research and Development for Enhancing Infectious Disease Response Capacity in Medical&Healthcare settings’, funded by the Korea Disease Control and Prevention Agency, the Ministry of Health & Welfare, Republic of Korea (grant number : RS-2025-02307351). Every dataset used in this research

study has been reviewed and deemed ethically acceptable by their corresponding Institutional Review Boards (IRB).

### Author contributions

I.J. and K.K. conceptualized, supervised and wrote the manuscript. S.B., D.K. and I.J. developed the methodology. J.K., K.K., A.K., S.H., E.N., S.B., J.L., S.K., H.J. collected the clinical data. S.K., H.J. collected the multi-omics data. D.K., J.J., S.H., K.K. performed the analysis. All authors took part in drafting and reviewing the manuscript.

### Declarations

### Competing interests

All authors declare no competing interests.

### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-07793-x>.

**Correspondence** and requests for materials should be addressed to I.J. or K.T.K.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nd/4.0/>.

© The Author(s) 2025