



OPEN Disease progression associated cytokines in COVID-19 patients with deteriorating and recovering health conditions

Eonyong Han¹, Sohyun Youn¹, Ki Tae Kwon², Sang Cheol Kim³, Hye-Yeong Jo³✉ & Inuk Jung¹✉

Understanding the immune response to COVID-19 is challenging due to its high variability among individuals. To identify differentially expressed cytokines between the deteriorating and recovering phases, we analyzed the Electronic Health Records (EHR) and cytokine profile data in a COVID-19 cohort of 444 infected patients and 145 non-infected healthy individuals. We categorized each patient's progression into Deterioration Phase (DP) and Recovery Phase (RP) using longitudinal neutrophil, lymphocyte and lactate dehydrogenase levels. A random forest model was built using healthy and severe patients to compute the contribution of each cytokine toward disease progression using Shapley Additive Explanations (SHAP). SHAP values were used for supervised clustering to identify DP and RP-related samples and their associated cytokines. The identified clusters effectively discriminated DP and RP samples, suggesting that the cytokine profiles differed between deteriorating and recovering health conditions. Especially, CXCL10, GDF15, PTX3, and TNFSF10 were differentially expressed between the DP and RP samples, which are involved in the JAK-STAT, NF- κ B, and MAPK signaling pathways contributing to the inflammatory response. Collectively, we characterized the immune response in terms of disease progression of COVID-19 with deteriorating and recovering health conditions.

Keywords Cytokine, Disease progression, Longitudinal, Severity, COVID-19

The emergence of COVID-19, caused by the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has led to an unprecedented global health crisis, infecting over 500 million people globally. The pandemic's toll on healthcare systems, economies, and societies has been profound, necessitating an urgent and coordinated global response. COVID-19's clinical presentation ranges from mild symptoms like fever and cough to severe complications such as pneumonia and acute respiratory distress syndrome (ARDS). More importantly, its clinical outcome is highly variable among individuals, which has been a key challenge in managing the pandemic, with severe cases often requiring intensive care and mechanical ventilation. Several risk factors, including age¹, gender², and comorbidities³ such as diabetes and obesity, have been associated with an increased risk of severe outcomes. Therefore, many studies have made effort to provide a measurement scale reflecting a patients severity level using their Electronic Health Records (EHR) in hopes to efficiently manage clinical resources^{4–6}. Hence, it is an important task to characterize the pathological progression and its associated immune response of COVID-19 in terms of severity.

The pathologic progression to COVID-19 is deeply intertwined with the immune response. Initially, the body's immune system combats the virus by triggering both the innate and adaptive immune mechanisms. The identification of intermediate stages in COVID-19 progression is of significant clinical and research importance. This approach facilitates targeted interventions, potentially improving patient outcomes through early anticipation and prevention of disease exacerbation via identifying immune response types⁷. Furthermore, it contributes to efficient healthcare resource allocation by aiding the prediction of intensive care requirements, a critical factor during pandemic surges. Moreover, in severe cases, the usual immune response can become

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

²Kyungpook National University Chilgok Hospital, 807 Hoguk-ro, Buk-gu, Daegu 41404, Republic of Korea.

³Division of Healthcare and Artificial Intelligence, Department of Precision Medicine, Korea National Institute of Health, Korea Disease Control and Prevention Agency, Osong, Cheongju-si 28159, Republic of Korea. ✉email: jhy1227@korea.kr; inukjung@knu.ac.kr

dysregulated, leading to a hyperinflammatory state known as the cytokine storm⁸. This surge in cytokine production, involving key molecules like IL-6, TNF- α , and IL-1 β , can exacerbate tissue damage and contribute to the progression from mild to severe disease. The COVID-19 pandemic has spurred extensive research, involving diagnosis, mortality and recovery prediction. While the methods differ, the studies commonly strive to provide means for early intervention strategies, potentially reducing the severity and duration of the illness^{9–12}. In the case of the COVID-19 diagnosis, over 160 machine learning (ML) based methods were proposed using various types of data and sources from both public and private sectors^{13,14}. The majority of severity prediction studies are in the domains of mortality prediction¹⁵ and identifying recovery related patterns¹⁶. Such studies often overlook a comprehensive view of the disease's progression, leading to a segmented understanding. In mortality prediction, numerous studies have utilized clinical and laboratory data to forecast patient outcomes and identify patients at high risk. These studies are key to managing healthcare resources efficiently. However, they tend to concentrate on predicting death as an outcome without encompassing the full trajectory of the disease from initial infection to recovery or death. Parallel to this, research on recovery patterns in COVID-19 patients has been integral in understanding the factors affecting patient recuperation. A study at the Assosa COVID-19 treatment center in Ethiopia analyzed recovery times and their influencing factors¹⁷, providing insights into the variability of recovery duration and the impact of clinical factors. However, such studies mainly focus on the endpoint of recovery, without extensively exploring the intermediate stages of the disease.

Our study addresses the challenge of categorizing COVID-19 patient statuses by proposing an approach that views the disease's progression as a continuous spectrum, spanning from critically ill to fully recovered states. While our goal is to map each patient's status accurately, we acknowledge the complexity and variability inherent in this task. Our approach aims to characterize the spectrum of disease progression by establishing reference points that represent the extremes of disease severity, rather than attempting to precisely categorize each patient's severity status. To categorize the disease progression of COVID-19 patients, we introduce the concept of Pathological Progression Groups (PPGs), which encapsulate the dynamic shifts in a patient's status following SARS-CoV-2 infection. Considering the natural process of the host immune response over the viral infection status, PPGs are composed of two phases: the Deterioration Phase (DP) and the Recovery Phase (RP). This concept is expected to encompass the full course of the disease progression. Using cytokine expression data from a diverse patient group, we constructed a random forest (RF) model for classifying non-COVID-19 healthy and severe COVID-19 patients. Based on the model, SHAP values were computed for each cytokine on the exploration dataset. Since the RF was trained with healthy and severe patients, any samples with a severity in between are expected to be placed between zero and one. For a severe sample, it will output a probability of close to 1 and a probability close to zero for healthy or mild severe patients. Thus, this conceptual framework enables us to map each patient's sample on a spectrum that ranges from critically ill to fully recovered states. Collectively, we investigated and characterized the dynamics of the immune response to COVID-19 in terms of deterioration and recovery using clinical and cytokine expression data.

Materials and methods

Data summary and preprocessing

Cytokine expression profiles and EHR samples of 145 healthy (i.e., non-infectious) and 444 COVID-19 patients were collected. For the COVID-19 infectious group, the omics and clinical data were sampled in a longitudinal manner, with each contributing data from 1 to 7 different time points. The COVID-19 dataset is comprised of patients infected with various SARS-CoV-2 strains, predominantly Delta (AY.69, 47.9%) and Beta (B.1.497, 45.5%) variants. Strain information was available for 396 out of 444 patients. They were collected from three institutions located in South Korea: Chungnam National University Hospital, Seoul Medical Center, and Samsung Medical Center. The dataset includes cytokine expression samples from plasma and laboratory data (e.g., neutrophil and lymphocyte counts) from blood samples, alongside the patient matched clinical data. Cytokine profiles were measured by the Korea National Institute of Health (KNIH) using the Luminex MAGPIX system with a customized panel, following a standardized protocol. The EHR data, detailing clinical therapy and patient information, was recorded at various stages of their hospital stay, with durations ranging from 5 to 25 days. Blood samples, providing cytokine data, were collected at different intervals during and post-hospitalization. Cytokine measurements were taken at regular intervals during each patient's hospital stay and after discharge. On average, these measurements were taken every 5.7 days, with intervals ranging from 1 to 35 days. Such sampling frequency was chosen based on previous studies of cytokine dynamics in acute viral infections, which suggest that significant changes in cytokine levels typically occur over a period of days rather than weeks¹⁸. As a result, the dataset is comprised of a total of 1844 time point specific EHR samples, 1159 time points specific cytokine profile. The cytokine profile consists of 191 cytokines, including IL17, IL25, and CXCL11, associated with SARS-CoV-2 infection. Table 1 presents a summary of the demographics and clinical characteristics of the COVID-19 cohort collected between February 2020 and July 2022 in the Republic of Korea. The initial dataset included 191 cytokines, selected based on their known association with SARS-CoV-2 infection and relevance to immune response pathways. Given the presence of 13,203 missing data points, a stringent criterion was applied: only cytokines with less than 15% missing values were included for further analysis. Consequently, 166 cytokines met this criterion. Missing values were imputed using a Random Forest regressor via the MissForest package (version 1.5) in R (version 4.2.0)¹⁹, trained over ten iterations to ensure the dataset's completeness and reliability.

Constructing pathological progression groups

To reflect the pathological progression of a patient, we defined the concept PPG, which encapsulates the dynamic shifts in a patient's pathological progression status post SARS-CoV-2 infection. Under the natural immune response to the virus, it is typical for patients to experience an exacerbation during the disease's initial phase, known as the pathogenic burden²⁰. In this early stage, the innate immune system is activated, discerning

Category	Description
Sample size	444 patients/1159 samples
Age distribution	52 ± 16.656
Gender distribution	245 males/199 females
Disease severity	Moderate/Severe
Sampling dates	2020-02-25 ~ 2022-07-28
Geographic location	The Republic of Korea
Data source	National Institute of Health

Table 1. Summarized statics of the COVID-19 cohort samples. Demographics and clinical data of 444 patients (1159 samples), including age, gender, disease severity, study period (2020-02-25 to 2022-07-28), and location (Republic of Korea), sourced from the National Institute of Health.

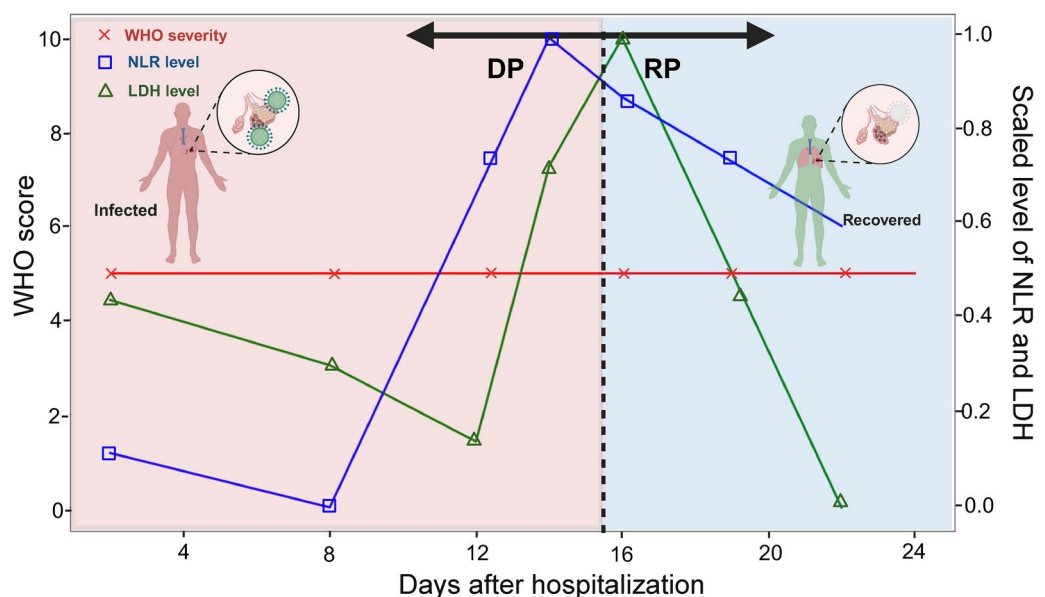


Fig. 1. The conceptual illustration of the DP and RP PPGs. The transition of COVID-19 patients through different stages of the disease is shown. The region colored in red refers to the duration of a patient with deteriorating health conditions, whereas the blue region represents the recovery of the patient.

and counteracting the pathogen. Subsequently, as strategies to combat the virus are solidified, the adaptive immune system takes precedence, heralding a transition towards restoring the immune system's homeostasis and alleviating the pathogenic burden. PPGs are used to delineate a patients pathological progression so that biomarkers specific to recovery or deterioration can be detected. For simplicity, the PPG was categorized into two groups: the DP and the RP. The conceptual depiction of a standard pathological progression is demonstrated in Fig. 1, which assumes that a patient gets ill but recovers afterwards.

The challenge lies in gauging the disease's dynamism, as its severity can fluctuate. A prevalent metric employed for this purpose is the World Health Organization (WHO) ordinal score system²¹, which is based on clinical features with a specific emphasis on oxygen therapy-related information. However, a significant fraction of patients display static clinical patterns throughout their hospitalization, making the demarcation of PPG challenging in the absence of discernible differences (Fig. 1). Hence, we incorporated the Neutrophil-to-Lymphocyte Ratio (NLR) and lactate dehydrogenase (LDH) levels as markers, which are used for labeling PPGs^{22,23}. These biological markers not only furnish insights into immune responses but also facilitate the classification of the disease's progression, especially in cases presenting limited change clinical status. Laboratory markers were measured multiple times during the patients' hospitalization, including at admission, during hospitalization, and at discharge. Given that these indicators are integral to routine laboratory data and are consistently procured during hospitalization, they serve as reliable tools in tracing the disease's phases. For each patient, DP and RP intervals were manually labeled by observing the longitudinal NLR and LDH levels. An interval (i.e. in units of days) of a hospitalized patient was labeled as DP if the NLR and LDH levels elevated. Similarly, an interval was labeled as RP if the NLR and LDH levels decreased. This sample-wise classification allowed us to capture the dynamic changes in patient status more sensitively. Consequently, the same patient could be classified into both DP and RP if their condition exhibited both upward and downward trends during their hospitalization. The reason for such manual labeling was because the NLR and LDH levels did not always

show consensus with a patient’s severity level as shown in Supplementary Fig. S1. Generally, an interval is labeled as RP if either NLR or LDH levels started to decrease. However, if one marker showed a significant increase while the other decreased, we maintained the DP classification to avoid premature categorization of recovery. However, we showed that the quality of our manually labeled PPGs were sufficient to capture DP and RP specific cytokines. The PPG was further split into severity level specific groups: moderate Deterioration Phase/Recovery Phase (mDP/mRP) and severe Deterioration Phase/Recovery Phase (sDP/sRP). The severe DP and RP groups correspond to samples with pronounced disease severity at any instance during hospitalization, typified by a WHO scale of 6 or above, which is a conventional demarcation of severe illness²¹. Similarly, the moderate DP and RP groups correspond to samples with a WHO scale less than 6. The WHO scale can have a range of one to ten depending on the type of pathological state or burden of a patient as described in Supplementary Table S1. As described, the PPGs were further split to by the WHO scale to distinguish moderate and severe status specific cytokines in terms of progression. As a result, cytokine samples that were collected within a DP or RP interval were labeled as the interval’s PPG. The statistics of PPGs are summarized in Table 2, including the number of samples, patients, gender distribution, age, WHO scale and related clinical features. After the PPG labeling process, we focused on 112 patients (81 moderate and 31 severe) for downstream analysis, as these patients exhibited clear DP or RP patterns. Since time-course samples were available per patient, the total number of moderate and severe samples were 276 and 128, respectively. This approach, focusing on patients with clearly identifiable DP or RP phases, allows us to capture the dynamic nature of disease progression more effectively and ensures that our analysis is based on samples with well-defined progression characteristics.

Supervised clustering and PPG analysis

The objective of our study is to identify differentially expressed cytokines between the different PPGs. Instead of directly comparing the raw cytokine levels of the PPGs, it is advantageous to measure the feature importance of cytokines and compare them between the PPGs for enhanced interpretation of the result. SHAP values offer a consistent and fair metric for assessing the importance of features across various machine learning models²⁴ and transcriptomic studies²⁵. These values distribute the contribution of each cytokine to individual predictions, enhancing the interpretability of the model. Often, the computed SHAP values are used to cluster samples that exhibit similar feature importance profiles^{26,27}. Hence, for robust results, they are computed based on a classification model that was built using a well curated dataset with clear difference between the target prediction groups.

Accordingly, a RF model was constructed using the expression of 166 cytokines, targeting the WHO ordinal scale to distinguish between two distinct groups: Healthy (n = 25) and Severe (n = 35). The healthy category represents non-infected healthy samples, indicative of the zero pathological burden. The RF model for discriminating the healthy from severe COVID-19 patients constructed using 100 trees with Gini impurity as the splitting criterion. The minimum number of samples required to split an internal node was set to 2 and the minimum number of samples required to be at a leaf node to 1. These parameters were chosen to balance model complexity with performance and to mitigate overfitting. We focused on these two extreme categories to establish clear boundaries of the severity spectrum, allowing us to model severity as a continuum rather than discrete classes. The selected 25 healthy donors were to match the time period of the 35 severe patients, ensuring consistency and avoiding periodic differences among the training data. In contrast, the severe category encapsulates the peak severe status of COVID-19 infection. Severe cases were defined as those with a WHO ordinal scale score 6 or above, in line with established clinical criteria. This selection was made to ensure that the RF model could learn from well-defined extremes of disease severity, thereby improving the model’s accuracy and reliability. By using these clear-cut cases, the model could effectively compute SHAP values, which were then applied to the broader and more variable exploration cohort for computing the SHAP values, or feature importance, of cytokines, which are used for the supervised clustering of the exploration dataset that is composed of cytokines samples from 81 moderate and 31 severe COVID-19 infected patients. The selection of the two

	Moderate		Severe	
	DP	RP	DP	RP
No. of samples	94	182	53	75
No. of patients	52	72	28	27
Male/female	26/26	39/33	17/11	16/11
Age	55.02 ± 15.69	54.75 ± 15.04	63.57 ± 11.29	60.7 ± 15.52
WHO scale	4.24 ± 0.43	4.29 ± 0.45	6.39 ± 1.34	6.29 ± 1.34
Neutrophil	67.49 ± 13.57	67.68 ± 12.72	84.99 ± 8.98	83.91 ± 9.05
Lymphocyte	22.79 ± 11.3	22.93 ± 10.58	8.21 ± 6.6	9.08 ± 6.66
LDH, U/L	447.26 ± 221.44	710.85 ± 974.88	1519.1 ± 1384.94	1761.23 ± 1311.73
D-dimers	650.92 ± 899.32	467.21 ± 241.41	961.66 ± 566.06	888.49 ± 509.09
Days of hospitalization	15.9 ± 5.38	15.8 ± 6.08	20.4 ± 8.67	20.5 ± 9.12

Table 2. Patient statistics by PPG categories. The distribution and clinical characteristics of patients in the DP and RP are shown per severeness. The WHO scale, neutrophil count, and LDH levels were similar within the same severity category, even when comparing different PPGs. However, they exhibited different levels across the severity groups, with higher values observed in the severe patients.

extremes are expected to capture the associated cytokines characteristics so that the cytokines SHAP values from the exploration dataset will be placed between them according to their severity level, and thus serving as a continuum of the disease progression. The schematic workflow of analysis is depicted in Fig. 2.

For a given cytokine sample, the prediction output of the RF model is decomposable into a sum of SHAP values for all cytokines and a base value $f(x_0)$. Thus, $f(x_0)$ represents the model's output in the absence of any cytokines, essentially the expected model output across the dataset and computed as follows:

$$f(x) = f(x_0) + \sum_{i=1}^M \gamma_i. \quad (1)$$

Here, $f(x)$ and $f(x_0)$ refer to the model's severity and to the base value, or the model's average prediction over the dataset, respectively. γ_i denotes the SHAP value for cytokine i , reflecting the cytokine's contribution to the prediction deviation from the base value. At last, M is the number of total cytokines in the data, which is 166 in our case. In essence, $f(x)$ is the severity prediction of the model for a cytokine sample x , and each γ_i signifies the impact of the respective cytokine. It implies that a model's prediction is the cumulative effect of each cytokine expression level, starting from the average model prediction as the reference point $f(x_0)$. This additive explanation model is essential for interpreting SHAP values and is fundamental to our analysis of feature importance. The SHAP value γ_i of a cytokine i is computed as follows:

$$\gamma_i(v) = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [v(S \cup \{i\}) - v(S)], \quad (2)$$

where v refers to the coalition value function that assigns a prediction value to each coalition of cytokine, S refers to a subset of cytokines, excluding cytokine i and M refers to the total number of cytokines, which is 166 in our study. Collectively, $\gamma_i(v)$ represents the average marginal contribution of cytokine i in the prediction model across all possible cytokine subsets S . The coalition value function $v(S)$ for a subset S of cytokines is defined as the expected output of the model when only the cytokines in S are known:

$$v(S) = \mathbb{E}[f(x)|x_S], \quad (3)$$

where $\mathbb{E}[f(x)|x_S]$ represents the expected value of the model's prediction given that the cytokines in S are known. This involves averaging the model's predictions over all possible values of the cytokines not in S , according to their distribution.

To identify samples that showed significantly different SHAP values, the density based clustering algorithm DBSCAN was applied on the UMAP of the SHAP value samples. DBSCAN was selected for its capacity to identify high-density areas separated by regions of low density, thus distinguishing clusters with dissimilar SHAP profiles. From the DBSCAN result, clusters were evaluated for homogeneity in terms of PPG. Clusters with high homogeneity were selected for downstream analysis, which are expected to well capture the differences between the DP and RP groups. The optimization of the DBSCAN algorithm was contingent on the calibration of two parameters: the minimum number of samples (MinPts) required to establish a dense region, and the neighborhood size (ϵ), which dictates the proximity required for points to be considered part of a cluster. Iterative refinement led to the determination that an ϵ value of 0.8 and a MinPts threshold of 20 were optimal for our dataset, resulting in the formation of distinct and meaningful clusters. This process was facilitated by the DBSCAN implementation available in the sklearn Python package²⁸. To ascertain clusters of analytical relevance, particularly those reflecting the nuances of the RP and DP, including their respective severity, we performed a purity assessment. To ensure the quality and robustness of the selected clusters, we performed a manual review process. Each cluster was carefully examined to assess the homogeneity and purity of the samples

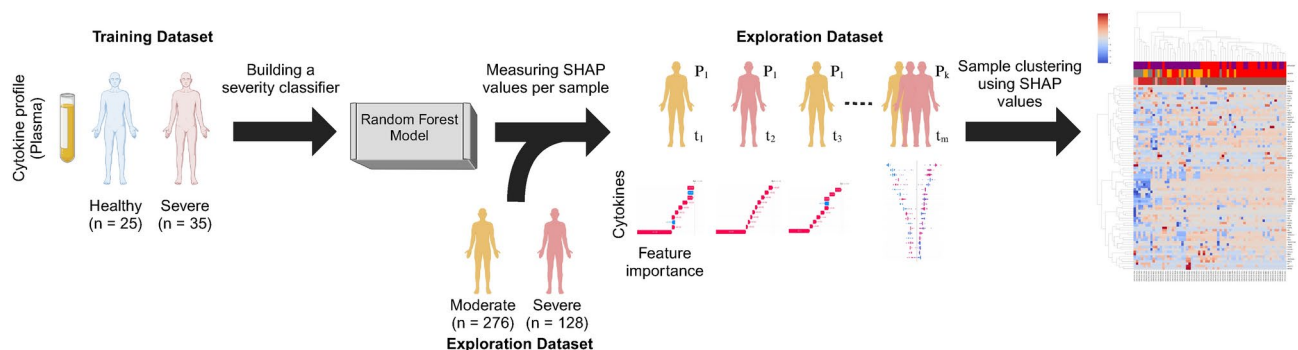


Fig. 2. The analysis workflow. First, a severity classifier using healthy and severe COVID-19 patients was built. Then, the model was used to compute the SHAP values of additional input data (i.e., non-training data). The computed SHAP values were then subject to clustering to identify PPGs associated with disease progression.

in terms of PPG representation. This manual review allowed us to identify and address any potential outliers or anomalies that could affect the subsequent analysis. By applying this quality control step, we aimed to select clusters that were representative of the specific PPG subgroups and provided a reliable basis for understanding the immunological responses in COVID-19 progression. To provide a comprehensive comparison between cytokine expression and cytokine SHAP value clustering, we also applied DBSCAN to the preprocessed cytokine expression value. For comparison, we used the same 77 cytokines that showed significant SHAP values from the measurement. To address the high variance in raw cytokine expression value, we applied two preprocessing steps: a) log transformation and b) quantile normalization to standardize the scale across different samples.

Finally, we focused on the statistical analysis of the Shapley values derived from the DBSCAN clusters. This analysis identified significant cytokines influencing the PPGs subgroups. To achieve this, we employed the *t*-test, a robust statistical tool, to evaluate the differences in feature importance across the PPG subgroups. The *t*-test was performed using the `pingouin.ttest` function, which automatically applies Welch's correction for unequal variances when sample sizes are unequal, as recommended by Zimmerman²⁹. To account for multiple testing, we applied the Benjamini–Hochberg procedure to control the false discovery rate at 0.05 across all *t*-tests comparing SHAP values between DP and RP groups. The results from these *t*-tests provided insights into the cytokine set that significantly contributes to the differentiation of disease severity in COVID-19 patients. Furthermore, to explore the interactions and pathways involving the key cytokines identified in the analysis, we performed a protein-protein interaction (PPI) network analysis using the STRING database (version 12.0)³⁰. The analysis was based on the statistical results obtained from the *t*-test result.

Results

PPG labeling results

The analysis of COVID-19 progression focused on various factors, including patient demographics, WHO scale, and laboratory parameters. The data presented in Table 2 expressed the diversity of pathological responses within our cohort and support the potential utility of the PPG approach for categorizing disease severity. Consistent disease severity within specific groups was observed, as indicated by stable WHO scale across different disease phases. Additionally, our findings suggest a correlation between age and disease severity, with older patients more likely to experience severe outcomes. However, the variance in other biomarkers, such as neutrophil and lymphocyte counts, was not pronounced, despite a slight increase in LDH levels between disease phases. This observation suggests that further research into additional biomarkers may be beneficial for a more comprehensive understanding of COVID-19 progression. It is observed that the hospitalization days across severity groups were similar, with severe cases averaging 20 days and moderate cases averaging 15 days of hospitalization.

To validate the PPG labeling approach, correlation analysis was done between the clinical indicators (i.e., NLR, LDH, WHO score) and the RF model's severe probability. Significant positive correlations between these measures (Supplementary Fig. S2) were observed, supporting the consistency of the PPG labeling with established clinical indicators. Notably, WHO score showed strong correlations with NLR ($r = 0.653$, $p < 0.001$) and LDH ($r = 0.634$, $p < 0.001$). The model's severe probability also correlated significantly with WHO score ($r = 0.569$, $p < 0.001$), NLR ($r = 0.463$, $p < 0.001$), and LDH ($r = 0.508$, $p < 0.001$). These results provide additional support for the validity of the PPG labeling approach and its alignment with established measures of disease severity.

Evaluation of the severity prediction RF model

The RF model was used to classify severe COVID-19 and complete healthy patients, which was used to compute the SHAP values. The WHO ordinal scale was used to differentiate the two groups. As expected, the model was able to distinguish the severe from the non-infected samples by 100% accuracy (Fig. 3b). 5-fold cross-validation on the training dataset maintained 100% accuracy across all folds, reflecting the clear separation between the extreme cases. When applied to the exploration dataset, including moderate cases not in the training data, the model predicted severity probabilities that correlated well with the WHO scale (Fig. 3a). This demonstrates the model's ability to generalize to intermediate severity levels and capture the continuous nature of disease progression. Analysis of the SARS-CoV-2 lineages showed that the training dataset predominantly consisted of Delta variants, while the exploration dataset had a significant presence of Beta variants. Despite such lineage heterogeneity, the model's prediction for mild and moderate cases fell within continuum between the two extremes. For instance, cases with WHO scores of 3–5, typically considered mild to moderate, showed a range of severity probabilities, reflecting the differences in their clinical presentations. Collectively, the RF model was able to predict the severity level of samples correctly between the spectrum of being healthy and critically ill. Then, the SHAP values for all samples were computed in reference to the trained RF model.

To gain insights of the factors influencing the RF model's predictions, we analyzed the predictor importance using both RF feature importance scores and SHAP values. Figure 4 presents the dual view of predictor importance. The SHAP summary plot in Fig. 4a displays the distribution of SHAP values for the top 20 predictors, with each point representing a single sample. The color of each point indicates whether that feature value was high (red) or low (blue) for that sample. The random forest feature importance scores (Fig. 4b) highlight the overall impact of each predictor on model decisions across all samples. Notably, CD274 (PD-L1) and KIT emerged as the top two predictors, suggesting their crucial role in determining COVID-19 severity.

Cytokines expressed differently in DP and RP groups

Before the SHAP value analysis, we first examined the cytokine expression levels values for different PPGs and severity groups. As shown in Fig. 5, the clustermap of cytokine profiles did not exhibit clear separation among PPGs. While some similarity was observed within severity groups, the PPGs remained largely indistinguishable based on the raw cytokine expression data alone. This suggests that raw cytokine profiles alone may not be

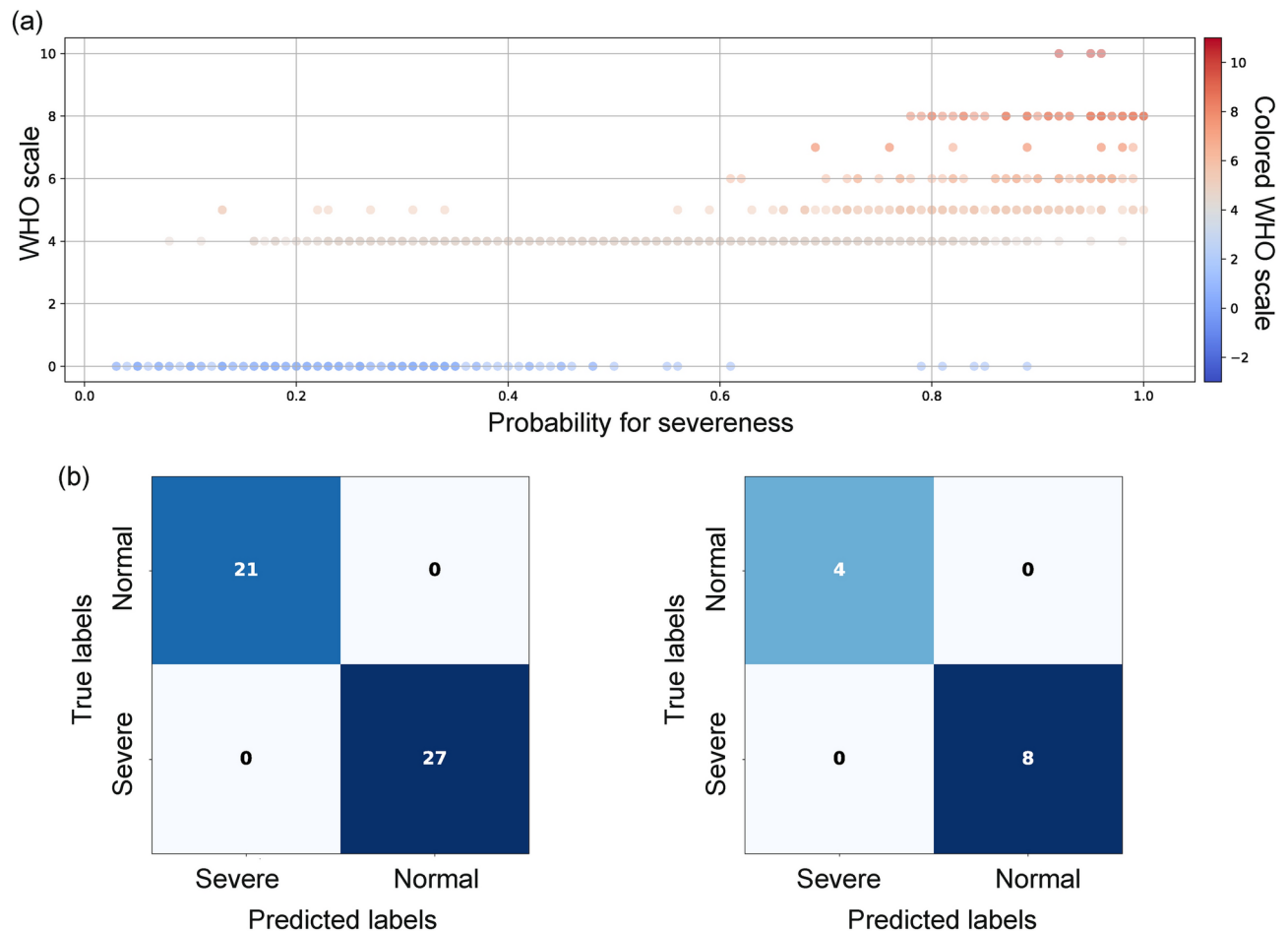


Fig. 3. Evaluation of RF model. **(a)** The scatter plot of the severe probability as predicted by the model, shows that the samples in the exploration dataset were well predicted and distributed along the severity probability scale in correlation to their WHO scales. It shows that the model would correctly predict new input samples correctly within the range of 0 to 1. For example, samples with moderate severeness will likely be predicted with a probability for severeness between 0.4 and 0.8 **(b)** The confusion matrix for the training set (left) and test set (right) shows that the model yielded high in both sensitivity and specificity.

sufficient to fully capture the complexity of disease progression and accurately categorize patients by their disease progression stage. To address such limitation, the clustering results from both cytokine expression data and SHAP values were compared. The SHAP value-based approach yielded a clearer separation between the DP and RP samples. Despite this initial observation, the presence of mixed cytokines among normal samples and patient PPGs, including mDP, sDP, and RP groups, pointed to the complexity of disease progression (Supplementary Fig. S3). Given these findings, we decided to narrow our focus. Rather than broad categorizations, we concentrate on specific data segments that are particularly representative of our target subgroups. This strategy aims to search for cytokines essential to understand the progression of the pathology.

The density-based clustering algorithm, DBSCAN, was applied to the dimensions 1 and 2 of the UMAP, which resulted in 11 distinct clusters (Fig. 6a). These clusters exhibited clear separations, with some demonstrating representative membership characteristics. From these, we selected clusters 4, 5, 6, 8, and 9 for further analysis based on their composition of PPGs and severity categories. Figure 6b illustrates the composition of these selected clusters (i.e., 4, 5, 6, 8, and 9) in terms of PPGs and severity categories, providing transparency in our cluster selection process. Clusters 4, 5, and 6 were chosen for moderate PPG analysis (mDP vs. mRP), while clusters 4, 8, and 9 were selected for severe PPG analysis (sDP vs. sRP). In Fig. 6, yellow stars indicates clusters selected for mDP vs. mRP comparison (4, 5, 6), red stars for sDP vs. sRP comparison (8, 9), and the star-shaped marker with both colors (cluster 4) indicates the cluster used in both comparisons. By thoroughly reviewing each selected cluster and assessing the homogeneity and purity of the samples, we ensured that they well represented mDP, mRP, sDP, and sRP subgroups. This approach allowed us to focus on the most informative and reliable clusters, minimizing the impact of potential outliers or anomalies. For mDP and mRP analysis, a total of 195 samples were selected from clusters 4, 5, and 6. In the severe comparison between sDP and sRP, 64 samples were selected from clusters 4, 8, and 9. These selected samples formed the basis for the subsequent analysis of key cytokines and their potential roles in disease progression within each severity group. The selected clusters were used to plot cluster map for the moderate and severe groups, where the RP and DP differences were now clearly

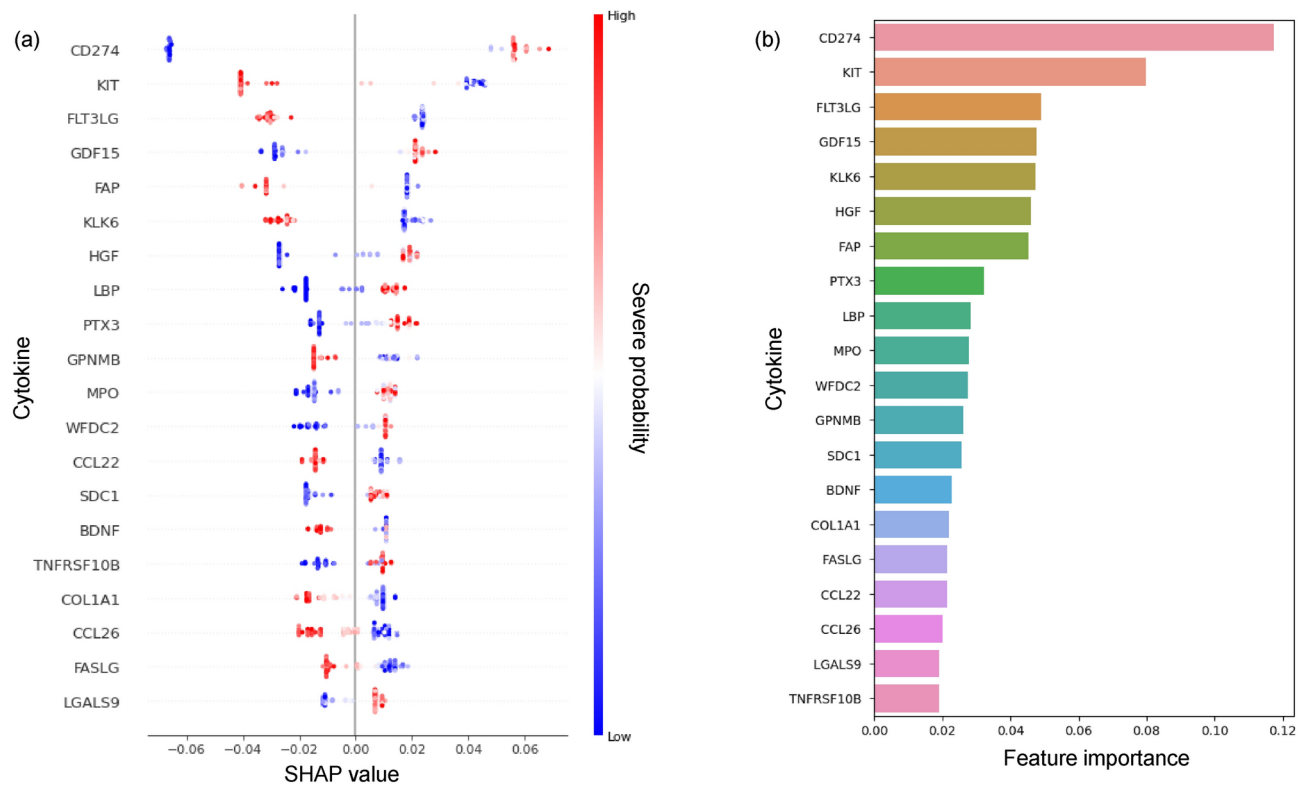


Fig. 4. Feature importance of the RF model for predicting COVID-19 severity. **(a)** SHAP summary plot showing the impact of the top 20 cytokines on model predictions. Each point represents a single sample, with color indicating high (red) or low (blue) feature values. The x-axis shows the SHAP value, representing the impact on model output. **(b)** Bar plot of feature importance scores from the RF model, highlighting the overall contribution of each cytokine to severity predictions across all samples. CD274 (PD-L1) and KIT emerge as the top two predictors of COVID-19 severity.

observed, as shown in Fig. 6c,d. These clustermaps demonstrate the distinction between mDP and mRP samples, and between sDP and sRP samples, based on the cytokine SHAP values.

PPGs specific to severe COVID-19 patients

In the mDP versus mRP comparison, 77 cytokines were evaluated. A *t*-test was performed to search for statistical differences in SHAP values of the cytokines between mDP and mRP. The analysis revealed 38 cytokines with significantly differing SHAP values. Similarly, in the sDP versus sRP comparison, 22 cytokines showed significant differences in SHAP values (Table 3). This disparity in significant cytokine counts showed the distinct immunological responses in different disease phases.

The analysis focused on cytokines contributing to sDP and sRP, comparing them with moderate PPGs. Adjusted *p*-value rank changes in cytokines between sDP and sRP were measured and compared to the moderate cases, as summarized in Table 3. The Δ Rank indicates the rank difference of a cytokine in the severe PPGs relative to the same cytokine's rank in the moderate PPGs in terms of SHAP value. For instance, GDF15 was placed 20 ranks higher in the severe PPGs compared to the moderate PPGs. CXCL10, GDF15, TNFSF10, and PTX3 cytokines were higher ranked in the severe PPGs. The SHAP value analysis revealed patterns for these cytokines across different PPGs. CXCL10 and GDF15 contributed positively to severity prediction in DP samples, with a more noticeable effect in severe patients. These findings suggest their potential as indicators for disease progression, particularly in the deterioration phase (Fig. 7a,b). PTX3 showed varying patterns, with higher contributions to severity prediction observed in the sDP group (Fig. 7b). The expression profile of PTX3 differed across PPGs, being most expressed in the sDP group (Fig. 8a). This variability in PTX3 expression and its SHAP values suggests its potential utility in distinguishing between different stages of disease progression. TNFSF10 showed lower expression levels in the sDP group compared to other PPGs. The SHAP value analysis indicated an association between TNFSF10 expression and disease severity, with lower values in the severe group. This finding suggests that TNFSF10 may play a role in the innate immune response during the early stages of COVID-19, and its reduced expression could be associated with severe disease progression. PPI network analysis showed that these four cytokines were connected to each other (Fig. 8b). They were also enriched in the "Viral protein interaction with cytokines and cytokine receptor" pathway in the KEGG database³¹.

In addition, the analysis identified ERBB2 and S100A12 as potential markers for COVID-19 severity. MPO and HGF also showed large Δ rank values of + 60 and + 57 respectively, suggesting their potential relevance with severe cases. These cytokines showed differences in SHAP values between DP and RP groups in severe

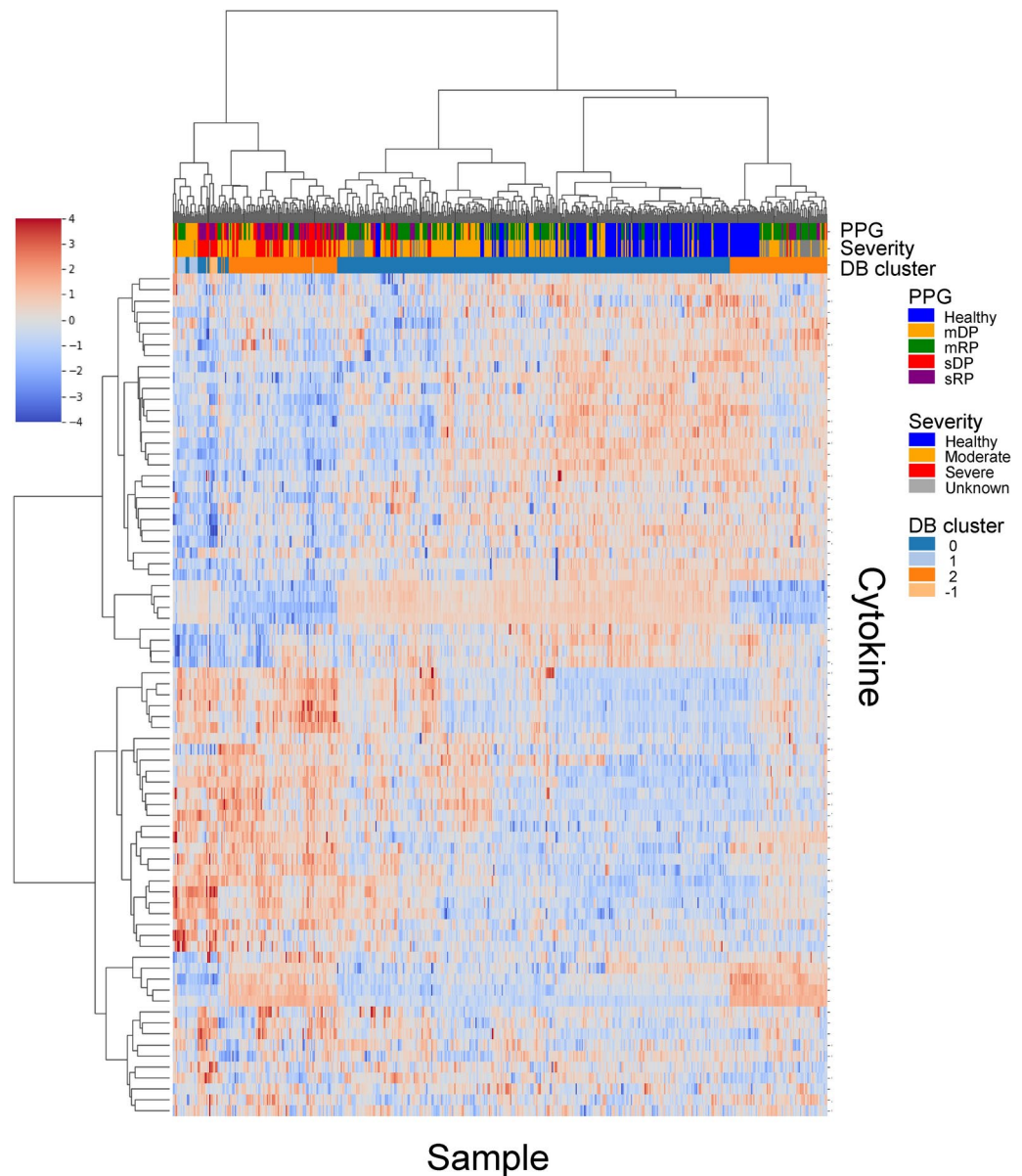


Fig. 5. The clustermap of cytokine expression data across different patient groups. The clusters do not show clear separation between severity levels or PPGs, demonstrating that cytokine expression value alone is insufficient to categorize patients by their disease progression stages. While some similarity is observed within severity groups, the PPGs remain largely indistinguishable. The color bars above the heatmap indicate PPG, Severity, and DBSCAN clusters, with 'Healthy' representing non-infected donors and 'Unknown' referring to discharged patients with incomplete oxygen therapy information but available lab data.

patients, suggesting their potential involvement in pathological pathways of the virus. These results provide a view of cytokine dynamics during COVID-19 progression, indicating potential biomarkers for disease severity and progression. The identification of cytokines differentially expressed between DP and RP, particularly in severe cases, offers insights into the immunological changes occurring during different phases of COVID-19 infection. The list of the cytokines analyzed in this study is provided in the Supplementary Tables S2 and S3.

Discussion

The findings align with previous studies identifying CXCL10, GDF15, PTX3, and TNFSF10 as factors in COVID-19 severity. These cytokines are involved in inflammatory and immune response pathways. CXCL10 has been associated with cytokine storm and severity prediction in COVID-19^{32,33}. Its positive contribution to severity prediction in DP samples supports its potential as a biomarker for disease progression. GDF15, which also contributed positively to severity prediction, is elevated in severe cases, possibly reflecting the extent of inflammatory tissue damage^{34,35}.

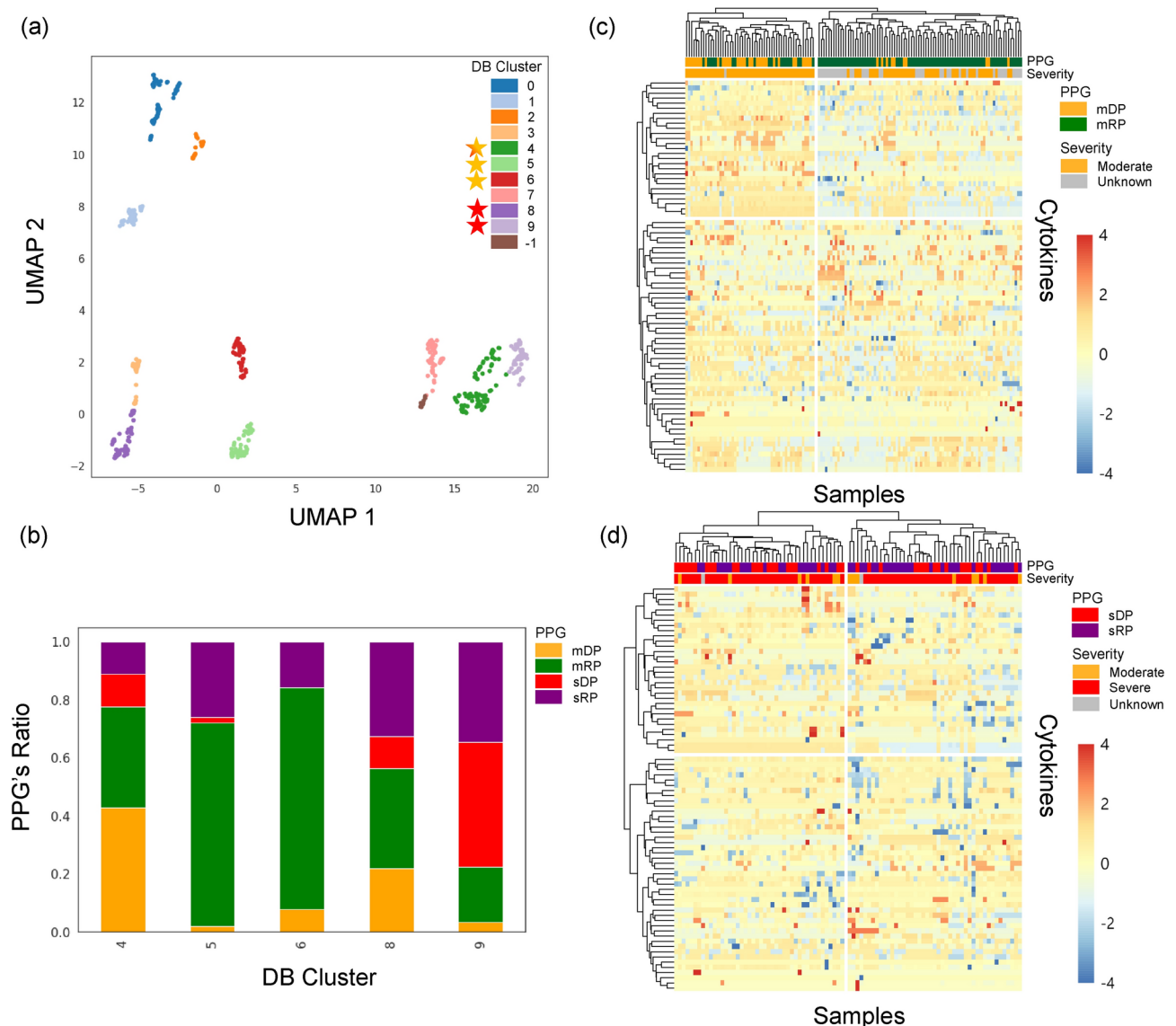


Fig. 6. UMAP embeddings and clustermaps of the cytokines SHAP values. **(a)** The UMAP of all samples colored by DB clusters. Yellow stars mark clusters selected for mDP vs. mRP comparison, red stars for sDP vs. sRP comparison, and star-shaped markers with both colors indicate clusters used in both comparisons. **(b)** The stack bar plot showing the composition of selected DB cluster in terms of PPGs and severity categories. **(c)** The clustermap of the moderate PPGs (mDP and mRP) showing a clear distinction between the two groups based on cytokine SHAP values. **(d)** The clustermap of the severe PPGs (sDP and sRP) depicting a clear pattern between the two groups based on cytokine SHAP values.

PTX3 showed varying patterns, with higher contributions in the sDP group. This aligns with its potential role as an indicator of immune response, particularly in severe cases of COVID-19^{36,37}. The expression patterns of PTX3 across different PPGs suggest its potential in monitoring disease progression. TNFSF10, displaying low expression levels in the sDP group compared to other PPGs, indicates its potential role in preventing severe disease progression^{38–40}. This finding suggests that an innate immune response mediated by TNFSF10 may be involved in mitigating severe outcomes in COVID-19.

By analyzing a comprehensive dataset combining cytokine profiles with electronic health records, key biomarkers that contribute to disease severity and progression have been identified. The cytokines including PTX3, GDF15, CXCL10, and TNFSF10 are differentially expressed between the DP and RP of COVID-19, particularly in patients in sDP. The elevated expression levels of PTX3 and GDF15 observed in patients with sDP may be indicative of abnormal inflammation, which is a characteristic of severe COVID-19^{41,42}. PTX3 has been shown to be a regulator of inflammation and is associated with disease severity and mortality in COVID-19³⁶. Similarly, GDF15 is induced by inflammation and has been proposed as a biomarker for severe COVID-19³⁴. The expression levels of CXCL10 and TNFSF10 exhibited a negative relationship in our study. CXCL10 was highly expressed in patients with sDP, while TNFSF10, a cytokine involved in innate immune responses, showed

Cytokines	P-adj	Δ Rank
CD274	0.000007	–
FLT3LG	0.000008	–
CXCL10	0.000657	+ 10
TNFSF10	0.002169	+ 54
GDF15	0.002182	+ 20
ERBB2	0.002901	+ 31
PTX3	0.003023	+ 8
CSF1	0.003947	–
SERPINA4	0.004319	– 2
PROS1	0.005541	– 1
S100A12	0.005589	– 8
MPO	0.006934	+ 60
HGF	0.007946	+ 57
ST2	0.010413	+ 30
LGALS9	0.012140	– 11
VCAM1	0.014725	– 11
IL3	0.019497	+ 52
LBP	0.029396	– 8
ADAMTS13	0.038749	– 7
FCER2	0.039913	+ 42
BDNF	0.042778	+ 47
TIMP1	0.046956	+ 34

Table 3. Significant cytokines and their rank variations in severe versus moderate PPGs. The results of *t*-tests comparing cytokine SHAP values between the DP and RP groups in severe patients are shown. Here, Δ Rank represents the rank difference of each cytokine between the severe and moderate PPGs in terms of SHAP value. A positive Δ Rank indicates that the cytokine has a higher rank in the severe PPGs compared to the moderate PPGs and vice versa, while ‘–’ indicates no change in rank. For example, CXCL10 was ranked third in severe PPGs and the thirteenth in moderate PPGs, resulting in Δ Rank = + 10.

lower expression levels. This suggests that a lack of early and adequate innate immune response, mediated by TNFSF10, may contribute to the development of severe COVID-19 characterized by heightened inflammation and high levels of CXCL10^{43,44}. Moreover, these cytokines are involved in specific signaling pathways that contribute to the pathophysiology of COVID-19. For instance, CXCL10 is a ligand for the CXCR3 receptor and activates the JAK-STAT signaling pathway leading to the production of pro-inflammatory cytokines and chemokines⁴⁵. TNFSF10 binds to its receptors and activates the NF- κ B and MAPK signaling pathways, which are involved in regulating immune responses and cell survival⁴⁶. These signaling pathways are involved in the excessive inflammatory response and tissue damage observed in severe COVID-19^{47,48}. PTX3 and GDF15 are also involved in the regulation of the NF- κ B and MAPK signaling pathways, respectively^{49,50}.

In addition to the well-established cytokines, ERBB2 and S100A12 were also identified as PPG differing candidates. ERBB2, traditionally associated with oncogenic processes, showed a potential link to COVID-19 severity in the analysis, suggesting its involvement in key pathological pathways of the virus^{51–53}. Similarly, the significant upregulation of S100A12 in severe COVID-19 cases aligns with recent studies identifying it as a marker for severe disease manifestations^{54–56}.

The ability to detect meaningful patterns and relationships in complex biological data shows possibility for advancing the understanding of COVID-19 and other diseases. Further research is needed to validate these findings in larger cohorts and to investigate the mechanistic roles of the identified cytokines in disease progression. It is important to acknowledge the limitations of the current study. The analysis is based on data from a single geographic region, and validation in diverse populations is necessary to generalize the findings. Moreover, integrating the results with transcriptomic data could provide a more comprehensive view of the molecular mechanisms underlying COVID-19 progression.

While our study has identified several cytokines (CXCL10, GDF15, TNFSF10, PTX3) that are relevant for distinguishing DP and RP, we acknowledge that we have not fully developed and validated a comprehensive model to differentiate all four PPGs based on their time-scale dynamics. Thus, limitations remains to be solved.

Conclusion

The integration of SHAP value analysis with machine learning has provided insights into the complexities of COVID-19 progression. Our study identified key cytokines, including PTX3, GDF15, CXCL10, and TNFSF10, that are potentially involved in the pathophysiology of severe COVID-19. These cytokines are associated with heightened inflammation and dysregulated innate immune responses, and they are involved in specific signaling pathways that contribute to the inflammatory response and disease progression. To further advance our

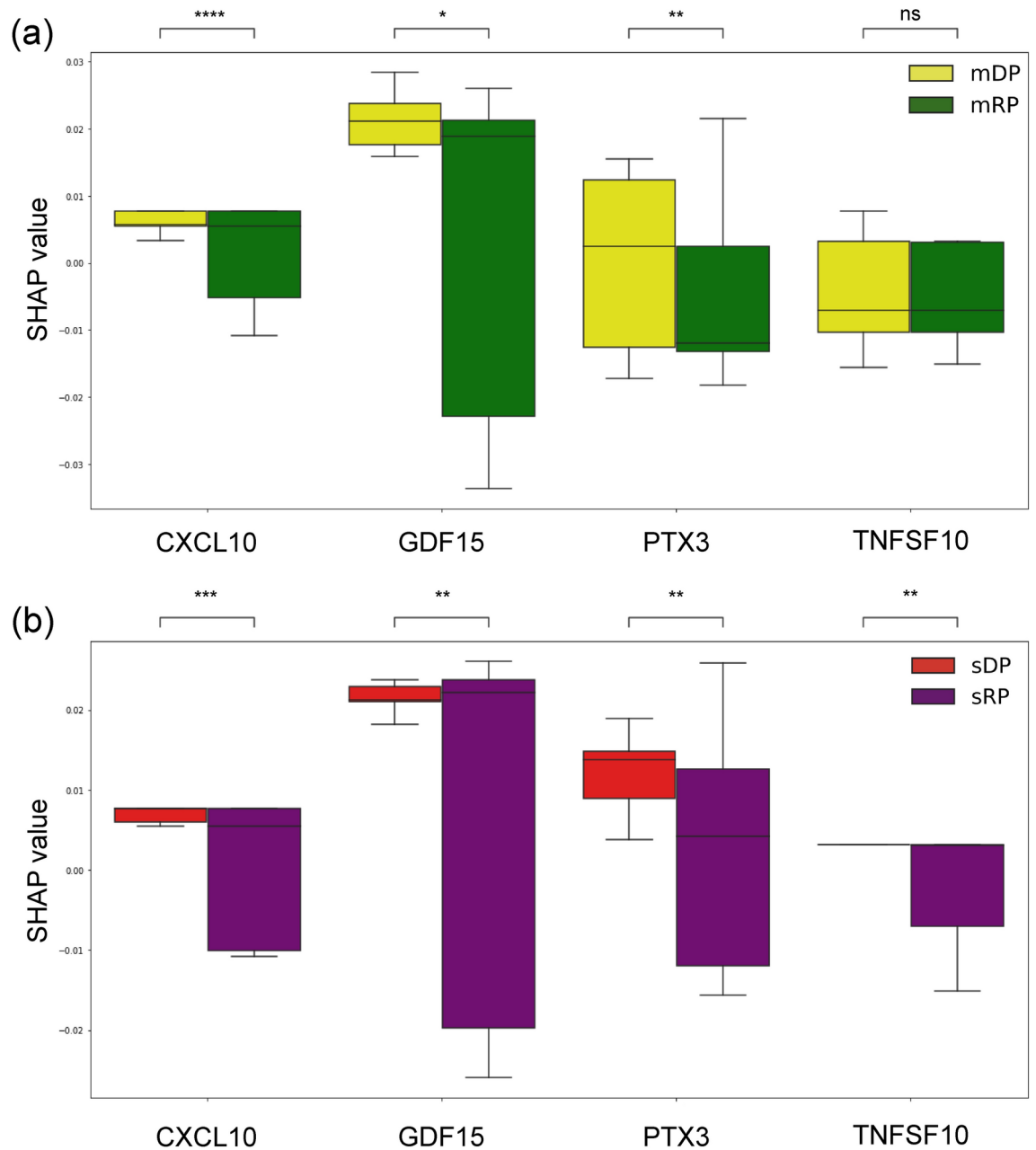


Fig. 7. SHAP values of highly ranked cytokines in severe samples. **(a)** The SHAP values for moderate PPGs, including the mDP and mRP. CXCL10 and GDF15 consistently show positive contributions to severity in mDP. In contrast, PTX3 and TNFSF10 exhibit different patterns compared to the severe PPGs. **(b)** The SHAP values for severe PPGs, including the sDP and sRP. In the severe PPGs comparison, all significant cytokines demonstrate positive contributions to severity, with consistent patterns across sDP and sRP. Notably, when comparing severe PPGs to moderate PPGs, PTX3 and TNFSF10 show robust variation, particularly in sDP, while moderate PPGs display a broader range of SHAP values for these cytokines.

understanding of COVID-19 pathophysiology and develop effective management strategies, future research may focus on validating these findings in a larger and more diverse cohort. A possible approach for future research is the development of a scoring system to quantify the modulation of inflammation based on the expression levels of key cytokines. Addressing the limitations of the current approach will be helpful in offering a more comprehensive understanding of COVID-19 pathophysiology.

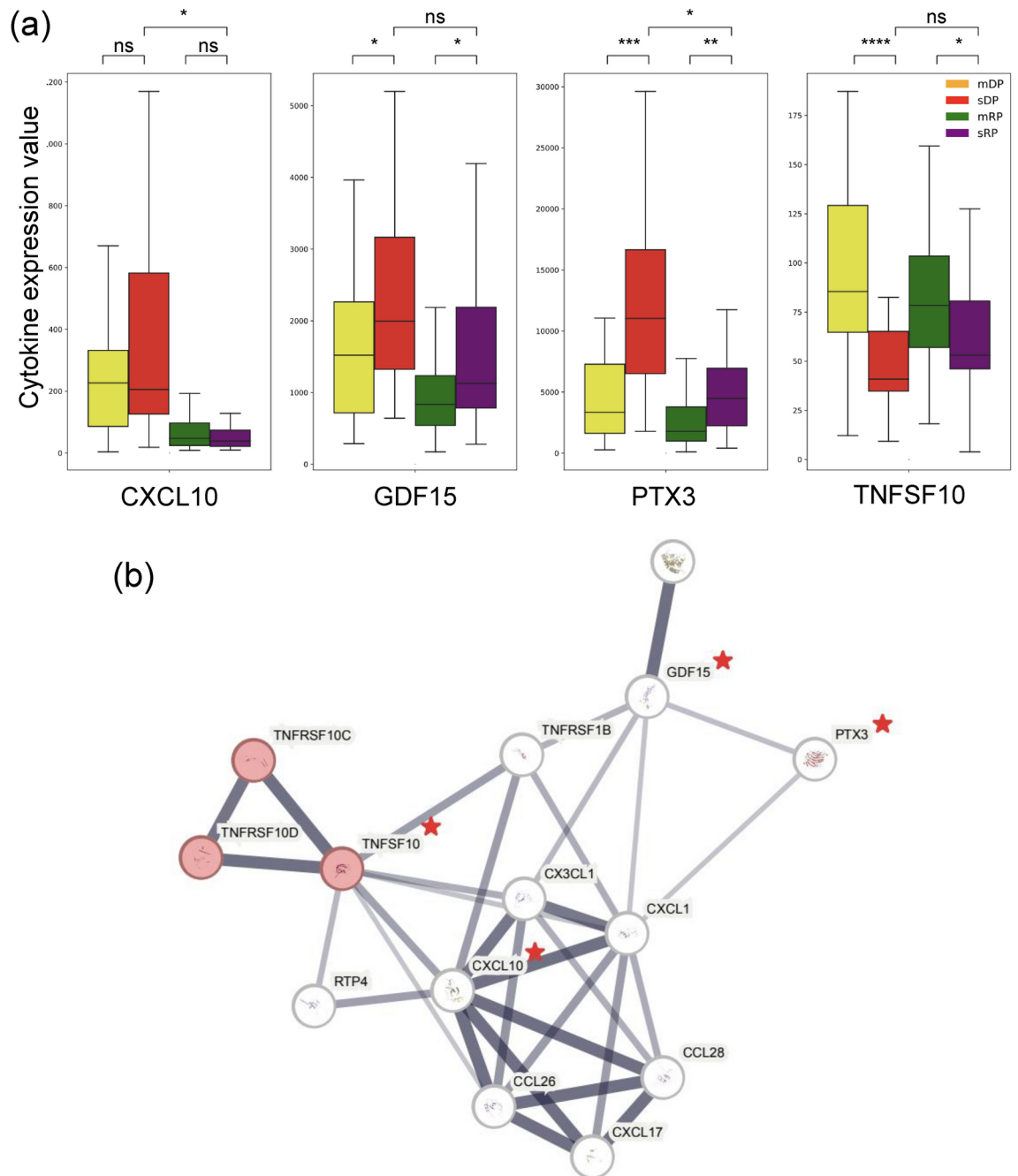


Fig. 8. Expression levels of highly ranked cytokines in severe samples, and their protein-protein interaction network. **(a)** Expression level of CXCL10, GDF15, PTX3, and TNFSF10 showed distinct patterns in the sDP group compared to the other PPGs. Among them, PTX3 showed the most distinctive difference between the PPGs, being most highly expressed in the sDP group. TNFSF10 displays particularly low values in the sDP group compared to the other PPGs. **(b)** PPI network of the four cytokines showed their relationship as associated manner. The four cytokines are marked with a red star. As show, the four cytokines were closely connected to each other.

Data availability

The dataset of the COVID-19 cohort used in this study are available in the Clinical and Omics Data Archive (CODA) database by the accession number CODA_D23017, <https://codan.nih.go.kr>. All code used in this study is publicly available on GitHub at <https://github.com/cobi-git/Cytokine-SHAP-COVID-19>.

Received: 11 April 2024; Accepted: 9 October 2024

Published online: 21 October 2024

References

- Davies, N. G. et al. Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nat. Med.* **26**, 1205–1211 (2020).
- Gebhard, C., Regitz-Zagrosek, V., Neuhauser, H. K., Morgan, R. & Klein, S. L. Impact of sex and gender on COVID-19 outcomes in Europe. *Biol. Sex Differ.* **11**, 1–13 (2020).
- Russell, C. D., Lone, N. I. & Baillie, J. K. Comorbidities, multimorbidity and COVID-19. *Nat. Med.* **29**, 334–343 (2023).
- Rubio-Rivas, M. et al. Who ordinal scale and inflammation risk categories in COVID-19. Comparative study of the severity scales. *J. General Intern. Med.* **37**, 1980–1987 (2022).
- Schwab, P. et al. Real-time prediction of COVID-19 related mortality using electronic health records. *Nat. Commun.* **12**, 1058 (2021).
- Barnett, W. R. et al. Initial mews score to predict ICU admission or transfer of hospitalized patients with COVID-19: A retrospective study. *J. Infect.* **82**, 282–327 (2021).
- Mueller, Y. M. et al. Stratification of hospitalized COVID-19 patients into clinical severity progression groups by immunophenotyping and machine learning. *Nat. Commun.* **13**, 915 (2022).
- Yang, L. et al. The signal pathways and treatment of cytokine storm in COVID-19. *Signal Transduct. Target. Ther.* **6**, 255 (2021).
- Liu, A., Hammond, R., Donnelly, P. D., Kaski, J. C. & Coates, A. R. Effective prognostic and clinical risk stratification in COVID-19 using multimodality biomarkers. *J. Intern. Med.* **294**, 21–46 (2023).
- Tiwari, S. et al. Applications of machine learning approaches to combat covid-19: A survey. *Lessons COVID-19* 263–287 (2022).
- Caron, R. M. & Adegboye, A. R. A. Covid-19: A syndemic requiring an integrated approach for marginalized populations. *Front. Public Health* **9**, 675280 (2021).
- Gomes, R. et al. A comprehensive review of machine learning used to combat COVID-19. *Diagnostics* **12**, 1853 (2022).
- Painuli, D., Mishra, D., Bhardwaj, S. & Aggarwal, M. Forecast and prediction of COVID-19 using machine learning. In *Data Science for COVID-19*, 381–397 (Elsevier, 2021).
- Meraïhi, Y., Gabis, A. B., Mirjalili, S., Ramdane-Cherif, A. & Alsaadi, F. E. Machine learning-based research for COVID-19 detection, diagnosis, and prediction: A survey. *SN Comput. Sci.* **3**, 286 (2022).
- Yan, L. et al. An interpretable mortality prediction model for COVID-19 patients. *Nat. Mach. Intell.* **2**, 283–288 (2020).
- Balloz, T. et al. Recovery and symptom trajectories up to two years after SARS-CoV-2 infection: Population based, longitudinal cohort study. *BMJ* **381**, e074425 (2023).
- Kassie, M. Z., Gobena, M. G., Alemu, Y. M. & Tegegne, A. S. Time to recovery and its determinant factors among patients with COVID-19 in Assosa COVID-19 treatment center, Western Ethiopia. *Pneumonia* **15**, 17 (2023).
- Liu, J. et al. Longitudinal characteristics of lymphocyte responses and cytokine profiles in the peripheral blood of SARS-CoV-2 infected patients. *EBioMedicine* **55**, 102763 (2020).
- Stekhoven, D. J. & Bühlmann, P. Missforest-non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2012).
- Reyes-Silveyra, J. & Mikler, A. R. Modeling immune response and its effect on infectious disease outbreak dynamics. *Theor. Biol. Med. Model.* **13**, 1–21 (2016).
- WHO, N. C. Covid-19 therapeutic trial synopsis. *R & D Blue Print*. <https://www.who.int/publications/i/item/covid-19-therapeutic-trial-synopsis> (2020).
- Prozan, L. et al. Prognostic value of neutrophil-to-lymphocyte ratio in COVID-19 compared with influenza and respiratory syncytial virus infection. *Sci. Rep.* **11**, 21519 (2021).
- 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).
- Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4765–4774 (2017).
- Yap, M. et al. Verifying explainability of a deep learning tissue classifier trained on RNA-seq data. *Sci. Rep.* **11**, 2641 (2021).
- Cooper, A., Doyle, O. & Bourke, A. Supervised clustering for subgroup discovery: An application to COVID-19 symptomatology. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 408–422 (Springer, 2021).
- Clement, T., Nguyen, H. T. T., Kemmerzell, N., Abdelaal, M. & Stjelja, D. Beyond explaining: Xai-based adaptive learning with shap clustering for energy consumption prediction. arXiv preprint [arXiv:2402.04982](https://arxiv.org/abs/2402.04982) (2024).
- Pedregosa, F. et al. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
- Zimmerman, D. W. A note on preliminary tests of equality of variances. *Br. J. Math. Stat. Psychol.* **57**, 173–181 (2004).
- Szklarczyk, D. et al. The string database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2021).
- Kanehisa, M. & Goto, S. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
- Gudowska-Sawczuk, M. & Mroczko, B. What is currently known about the role of CXCL10 in SARS-CoV-2 infection?. *Int. J. Mol. Sci.* **23**, 3673 (2022).
- Callahan, V. et al. The pro-inflammatory chemokines CXCL9, CXCL10 and CXCL11 are upregulated following SARS-CoV-2 infection in an akt-dependent manner. *Viruses* **13**, 1062 (2021).
- Rochette, L., Zeller, M., Cottin, Y. & Vergely, C. GDF15: An emerging modulator of immunity and a strategy in COVID-19 in association with iron metabolism. *Trends Endocrinol. Metab.* **32**, 875–889 (2021).
- Alserawan, L. et al. Growth differentiation factor 15 (GDF-15): A novel biomarker associated with poorer respiratory function in COVID-19. *Diagnostics* **11**, 1998 (2021).
- Brunetta, E. et al. Macrophage expression and prognostic significance of the long pentraxin PTX3 in COVID-19. *Nat. Immunol.* **22**, 19–24 (2021).
- Capra, A. P. et al. The prognostic value of pentraxin-3 in COVID-19 patients: A systematic review and meta-analysis of mortality incidence. *Int. J. Mol. Sci.* **24**, 3537 (2023).
- Shin, G.-C., Kang, H. S., Lee, A. R. & Kim, K.-H. Hepatitis B virus-triggered autophagy targets TNFRSF10B/death receptor 5 for degradation to limit TNFSF10/TRAIL response. *Autophagy* **12**, 2451–2466 (2016).
- Chen, Y., Qin, Y., Fu, Y., Gao, Z. & Deng, Y. Integrated analysis of bulk RNA-seq and single-cell RNA-seq unravels the influences of SARS-CoV-2 infections to cancer patients. *Int. J. Mol. Sci.* **23**, 15698 (2022).
- Li, Q. et al. Immune response in COVID-19: What is next?. *Cell Death Differ.* **29**, 1107–1122 (2022).
- Hu, B., Guo, H., Zhou, P. & Shi, Z.-L. Characteristics of SARS-CoV-2 and COVID-19. *Nat. Rev. Microbiol.* **19**, 141–154 (2021).
- 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).
- Coperchini, F. et al. The cytokine storm in COVID-19: Further advances in our understanding the role of specific chemokines involved. *Cytokine Growth Factor Rev.* **58**, 82–91 (2021).
- Li, S. et al. SARS-CoV-2 triggers inflammatory responses and cell death through caspase-8 activation. *Signal Transduct. Target. Ther.* **5**, 235 (2020).
- Liu, M. et al. CXCL10/IP-10 in infectious diseases pathogenesis and potential therapeutic implications. *Cytokine Growth Factor Rev.* **22**, 121–130 (2011).
- Condamine, T., Ramachandran, I., Youn, J.-I. & Gabrilovich, D. I. Regulation of tumor metastasis by myeloid-derived suppressor cells. *Annu. Rev. Med.* **66**, 97–110 (2015).

47. Catanzaro, M. et al. Immune response in COVID-19: Addressing a pharmacological challenge by targeting pathways triggered by SARS-CoV-2. *Signal Transduct. Target. Ther.* **5**, 84 (2020).
48. Costela-Ruiz, V. J., Illescas-Montes, R., Puerta-Puerta, J. M., Ruiz, C. & Melguizo-Rodríguez, L. SARS-CoV-2 infection: The role of cytokines in COVID-19 disease. *Cytokine Growth Factor Rev.* **54**, 62–75 (2020).
49. Jaillon, S. et al. The long pentraxin PTX3 as a key component of humoral innate immunity and a candidate diagnostic for inflammatory diseases. *Int. Arch. Allergy Immunol.* **165**, 165–178 (2015).
50. Wollert, K. C., Kempf, T. & Wallentin, L. Growth differentiation factor 15 as a biomarker in cardiovascular disease. *Clin. Chem.* **63**, 140–151 (2017).
51. Zhu, Z., Chen, X., Wang, C. & Cheng, L. Novel genes/loci validate the small effect size of ERBB2 in patients with myasthenia gravis. *Proc. Natl. Acad. Sci.* **119**, e2207273119 (2022).
52. Wang, H. et al. Molecular landscape of ERBB2 alterations in 14,956 solid tumors. *Pathol. Oncol. Res.* **28**, 1610360 (2022).
53. Liu, F. et al. Shared mechanisms and crosstalk of COVID-19 and osteoporosis via vitamin D. *Sci. Rep.* **12**, 18147 (2022).
54. Lei, H. A single transcript for the prognosis of disease severity in COVID-19 patients. *Sci. Rep.* **11**, 12174 (2021).
55. Lei, H. A two-gene marker for the two-tiered innate immune response in COVID-19 patients. *PLoS ONE* **18**, e0280392 (2023).
56. Haljasmägi, L. et al. Longitudinal proteomic profiling reveals increased early inflammation and sustained apoptosis proteins in severe COVID-19. *Sci. Rep.* **10**, 20533 (2020).

Acknowledgements

This research was supported by the National Institute of Health research projects (project No. 2024-ER-0801-00). This research was also supported by the project for Infectious Disease Medical Safety, funded by the Ministry of Health and Welfare, South Korea (Grant Number: HG22C0014). The Figs. 1 and 2 were created with BioRender.com.

Author contributions

I.J. conceived the experiments, E.H. and S.Y. conducted the experiments and analysis. H.J., S.C.K. and K.K., revised the manuscript. I.J. and H.J. supervised the research. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-75924-x>.

Correspondence and requests for materials should be addressed to H.-Y.J. or I.J.

Reprints and permissions information is available at 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) 2024