BMC Infectious Diseases

RESEARCH

Open Access

# In-depth analysis of the risk factors for persistent severe acute respiratory syndrome coronavirus 2 infection and construction of predictive models: an exploratory research study

Jia Zhang[1], Weihua Zhu[1], Piping Jiang[1], Feng Ma[1], Yulin Li[1], Yuwei Cao[1], Jiaxin Li[1], Zhe Zhang[1], Xin Zhang[1], Wailong Zou[1*] and Jichao Chen[1*]

## Abstract

**Background** Persistent severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection differs from long coronavirus disease (COVID-19) (acute symptoms ≥ 12 weeks post-clearance). The Omicron BA.5 variant has a shorter median clearance time (10–14 days) than the Delta variant, suggesting that the traditional 20-day diagnostic threshold may delay interventions in high-risk populations. This study integrated multi-threshold analysis (14/20/30 days), whole-genome sequencing, and machine learning to investigate diagnostic thresholds for persistent SARS-CoV-2 infection and developed a generalizable risk prediction model.

**Methods** This retrospective study analyzed data from 1,216 patients with COVID-19 hospitalized at Aerospace Center Hospital between January 2021 and October 2024. We used whole-genome sequencing to genotype all COVID-19 cases and to identify major variants (such as Omicron BA. 5, Delta). The outcome, "persistent SARS-CoV-2 infection," was defined as viral nucleic acid positivity ≥ 14 days. Risk factors associated with persistent infection were identified through subgroup analysis with multiple logistic regression (adjusted for age, comorbidities, vaccination status, and virus strain) and machine learning models (70% training, 30% testing dataset).

**Results** Persistent SARS-CoV-2 infection was identified in 15.5% (188/1,216) of hospitalized COVID-19 patients. Key predictors included comorbidities—hypertension, diabetes, and active malignancy—and immune dysfunction, marked by reduced B-cell and CD4 + T-cell counts. Unvaccinated patients exhibited an 82% higher risk of persistent infection. Elevated inflammatory markers (C-reactive protein and interleukin-6) and bilateral lung infiltrates on computed tomography further distinguished persistent cases. The predictive model demonstrated strong discrimination with an area under the curve (AUC) of 0.847 (95% confidence interval: 0.815–0.879) and an AUC of 0.81 externally in external validation, underscoring its clinical utility for risk stratification.

*Correspondence:
Wailong Zou
15301050459@163.com
Jichao Chen
604939512@qq.com

**Conclusions** Hypertension, diabetes, malignancy, immunosuppression (low B/CD4 + cells), and non-vaccination are independent risk factors for persistent SARS-CoV-2 infection. Integrating these factors into clinical risk stratification may optimize management of high-risk populations.

**Keywords** SARS-CoV-2, Persistent infection, Risk factors, Clinical manifestations, Predictive model construction

## Background

The World Health Organization declared coronavirus disease (COVID-19) a global pandemic in March 2020 [1], initiating a prolonged coexistence between humans and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Emerging evidence highlights persistent SARS-CoV-2 infections, defined by prolonged nucleic acid positivity beyond acute phases, with a prevalence exceeding 20% [2]. Although the median clearance time of SARS-CoV-2 is 7–14 days in the general population and 7–10 days in patients with mild infection [3], long-term infections (> 30 days) are more frequently observed in immunocompromised groups, particularly in patients with hematological malignancies and recipients of B-cell depletion therapy [4–13]. Underreporting persists due to the persistent presence of asymptomatic viruses, non-respiratory virus hosts (such as patients with gastrointestinal tract infections), and diagnostic limitations [7–11]. Prolonged viral shedding is clinically significant as it is associated with persistent systemic inflammation, delayed recovery of physical function, and accelerated frailty progression in older adult patients, even after viral clearance [7, 10]. Queue studies estimate that individuals with an infection lasting at least 60 days are at highest risk among transplant recipients and patients with cancer [10]. These infections exacerbate clinical burdens, drive viral evolution [11, 14], and increase risks of severe sequelae (e.g., chronic fatigue, cardiopulmonary dysfunction) [7], with longitudinal data showing that continuous shedding can independently predict an increase in the probability of long-term COVID-19 [3, 11].

Current research has identified key risk factors for persistent SARS-CoV-2 infection, including immunocompromised states (malignancies, transplants, autoimmune diseases) [15], comorbidities (diabetes, hypertension, chronic obstructive pulmonary disease [COPD]) [16, 17], advanced age (linked to immune senescence) [18], and viral factors (high initial loads, immune-evading mutations) [18]. The clinical rationale for characterizing these risks is their direct influence on therapeutic decision-making. Prolonged shedding requires extended isolation protocols, alters antiviral dosing strategies (e.g., extending treatment with nirmatrelvir/ritonavir (Paxlovid) beyond 5 days), and mandates closer monitoring for viral rebound [7, 11]. Host genetic polymorphisms affecting immune responses and socioeconomic disparities may further modulate risks [19], although mechanistic insights remain limited. Importantly, continued infection in the frail population is linked to a higher rate of hospitalization for secondary infections such as bacterial pneumonia, emphasizing the necessity of risk stratification management [7, 17].

Prior studies predominantly focused on demographic analyses [20], lacking integration of viral genomic data with clinical parameters. To address this gap, we analyzed 3,452 SARS-CoV-2-infected patients (2021–2024) using whole-genome sequencing (targeting ORF1ab mutations) and machine learning. We hypothesized that combining viral genomic features (e.g., mutation profiles) with immune indicators (lymphocyte subsets, inflammatory markers) would outperform conventional logistic regression in predicting persistent infection. This dual approach aims to enable early identification, optimize monitoring, and mitigate care disruptions in high-risk populations.

## Methods

### Study design and approval

This was a retrospective cohort study aiming to conduct an in-depth analysis of the clinical characteristics of patients infected with SARS-CoV-2 who visited Aerospace Center Hospital between January 2021 and October 2024. The study was conducted in accordance with the principles outlined in the Declaration of Helsinki and the Strengthening the Reporting of Observational Studies in Epidemiology guidelines. The study protocol was approved by the ethics committee (approval: Jinghang Yilun Shen 2024 No. 090), and the requirement for informed consent was waived by the Ethics Committee of Aerospace Center Hospital for all participants. Data confidentiality was ensured through strict protocols: All personally identifiable information was anonymized using unique alphanumeric codes prior to analysis. Raw data were securely stored on an encrypted server (AES-256 standard) compliant with HIPAA regulations, accessible exclusively to the ethics committee-authorized researchers. Analyses were conducted solely on de-identified datasets to safeguard participant privacy.

### Study population

This retrospective study initially screened 1,519 SARS-CoV-2-infected individuals at Aerospace Center Hospital between January 2021 and October 2024. Based on the exclusion criteria, 303 patients were excluded, including 68 patients aged < 18 years, 185 with incomplete data on follow-up or irregular nucleic acid testing, 40 lacking critical clinical data, and 10 participating in other interventional trials. The final analytic cohort comprised 1,216 adults (age ≥ 18 years) with laboratory-confirmed SARS-CoV-2 infection according to the *Diagnosis and Treatment Protocol for COVID-19 (Tenth Trial Version)* [21]; complete clinical records (demographics, comorbidities, vaccination status, chest CT scans, and laboratory tests); and standardized follow-up data. A detailed cohort selection flowchart is provided in Fig. 1. Vaccination status was defined as the completion of the primary immunization series (≥ 2 doses of mRNA vaccine or ≥ 3 doses of inactivated vaccine), taken at least 14 days before infection according to the WHO guidelines [22].

### Grouping and control group setting

A control comparison group was formed to establish a reference standard for more accurate assessment of the clinical characteristics and risk factors of patients with persistent SARS-CoV-2 infection. Participants were divided into two groups: persistent infection group and control group. The persistent infection group comprised patients with positive SARS-CoV-2 nucleic acid or antigen tests for 14 days or longer. The control group was selected based on sex matching among patients with positive SARS-CoV-2 nucleic acid or antigen test results for less than 14 days. To reduce any remaining selection bias, we reanalyzed the data using propensity score matching (PSM) with a ratio of 1:1 and caliper width of 0.1.

Matching variables included age (± 5 years), vaccination status, and baseline viral load. After PSM, covariate balance was achieved (standardized differences < 0.1, Table 1). A sex-matched design was employed for the control group to mitigate the impact of sex-related factors on the results. The following sex matching method was used to ensure comparability between the control and persistent infection groups in terms of the key variable of sex. First, the number of patients required for each sex in the control group was determined based on the sex distribution of patients in the persistent infection group. Subsequently, among patients with positive SARS-CoV-2 nucleic acid or antigen test results for < 14 days, screening was conducted according to sex and age (as closely matched as possible) to ensure consistency between the control and persistent infection groups for sex and age. Finally, eligible control group members were selected through random sampling to ensure the effectiveness of sex matching, allowing for a more accurate assessment of the relationship between other factors and persistent infection.

### Data collection

Data from all enrolled patients were collected using the Aerospace Center Hospital Medical Record System. Data included basic information such as (1) general characteristics (sex, age, and history of COVID-19 vaccination); (2) underlying diseases (hypertension, diabetes, malignant tumors [solid organ malignancies and hematological malignancies], transplant status [solid organ transplantation and bone marrow transplantation], autoimmune diseases, cardiovascular and cerebrovascular

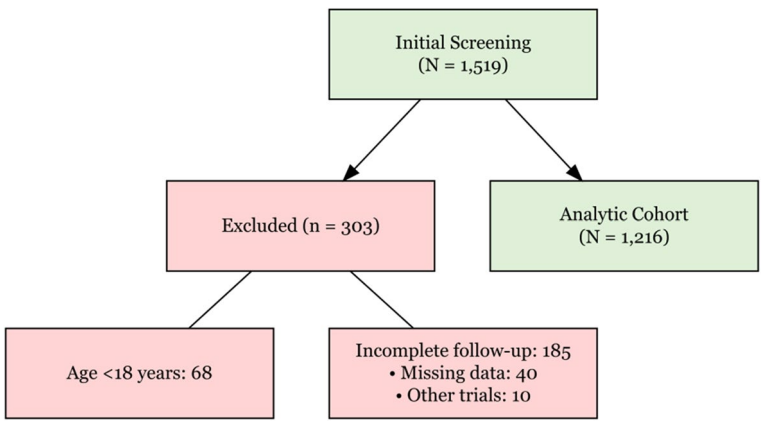


**Fig. 1** Study Cohort Flow Diagram. Among 1,519 SARS-CoV-2-infected individuals screened at Aerospace Center Hospital (January 2021–October 2024), 303 were excluded due to age < 18 years ($n = 68$), incomplete follow-up/irregular nucleic acid testing ($n = 185$), missing critical clinical data ($n = 40$), or participation in other trials ($n = 10$). The final analytic cohort included 1,216 adults (≥ 18 years) with confirmed infection, complete clinical records, and standardized follow-up data

**Table 1** Comparison of information, clinical manifestations, and laboratory test results between the groups ($\bar{x} \pm s$)

| Item | Persistent infection group ($n = 188$) | Non-persistent infection group ($n = 1028$) | $Z/\chi^2$ | *P*-value |
|---|---|---|---|---|
| **General information** | | | | |
| Age (years) | 56.2 ± 6.8 | 54.1 ± 6.6 | 14.334 | 0.04 |
| Sex (male, [%]) | 112 (59.6%) | 616 (59.9%) | 0.234 | 0.12 |
| Height (m) | 1.65 ± 0.08 | 1.66 ± 0.07 | 0.453 | 0.34 |
| Weight (kg) | 65.2 ± 11.4 | 64.1 ± 12.1 | 0.879 | 0.27 |
| Smoking history (n [%]) | 65 (34.6%) | 298 (29.0%) | 10.237 | 0.01 |
| Previous COVID-19 vaccination (n [%]) | | | | |
| Unvaccinated (n [%]) | 48 (25.5%) | 211 (19.6%) | 18.365 | < 0.001 |
| Vaccinated once (n [%]) | 15 (8.0%) | 87 (8.5%) | 0.567 | 0.08 |
| Vaccinated twice (n [%]) | 81 (43.1%) | 476 (46.3%) | 0.689 | 0.07 |
| Vaccinated three times (n [%]) | 44 (23.5%) | 254 (24.7%) | 0.689 | 0.09 |
| **Underlying diseases** | | | | |
| Hypertension (n [%]) | 41 (21.8%) | 156 (15.2%) | 15.312 | < 0.001 |
| Diabetes (n [%]) | 17 (9.0%) | 32 (3.1%) | 10.238 | < 0.001 |
| Coronary heart disease (n [%]) | 9 (4.8%) | 23 (2.2%) | 18.329 | < 0.001 |
| Arrhythmia (n [%]) | 7 (3.7%) | 39 (3.8%) | 1.230 | 0.268 |
| Stroke (n [%]) | 5 (2.7%) | 28 (2.8%) | 1.028 | 0.311 |
| Malignant tumor (n [%]) | 86 (45.7%) | 97 (9.4%) | 14.238 | < 0.001 |
| Transplant status (n [%]) | 2 (1.1%) | 11 (1.1%) | 1.872 | 0.359 |
| Autoimmune disease (n [%]) | 8 (4.3%) | 11 (1.1%) | 15.367 | < 0.001 |
| Liver dysfunction (n [%]) | 46 (24.5%) | 261 (25.4%) | 1.029 | 0.310 |
| Renal dysfunction (n [%]) | 42 (22.3%) | 247 (24.0%) | 0.854 | 0.355 |
| Structural lung disease (n [%]) | 55 (29.2%) | 238 (23.2%) | 15.741 | 0.014 |
| **Laboratory tests** | | | | |
| White blood cells ($\times 10^9$/L) | 5.6 ± 2.1 | 7.1 ± 2.8 | 16.378 | 0.01 |
| Lymphocyte count ($\times 10^9$/L) | 0.6 ± 0.3 | 0.8 ± 0.4 | 11.027 | 0.01 |
| Platelets ($\times 10^9$/L) | 207.2 ± 30.9 | 214.3 ± 28.9 | 1.321 | 0.186 |
| Hemoglobin (g/L) | 117.5 ± 18.6 | 120.1 ± 20.1 | 0.846 | 0.397 |
| CRP (mg/L) | 109.8 ± 21.2 | 83.1 ± 17.8 | 9.387 | 0.01 |
| PCT (ng/L) | 0.2 ± 0.1 | 0.2 ± 0.1 | 0.538 | 0.591 |
| IL-6 (ng/L) | 62.1 ± 17.1 | 33.1 ± 8.1 | 17.368 | 0.01 |
| D-dimer (ng/L) | 636.1 ± 110.1 | 625.2 ± 109.3 | 1.067 | 0.286 |
| Albumin (g/L) | 35.6 ± 8.1 | 39.5 ± 10.5 | 9.674 | 0.01 |
| Creatinine (μg/L) | 79.2 ± 21.0 | 81.6 ± 21.2 | 0.874 | 0.382 |
| Urea nitrogen (mmol/L) | 6.9 ± 2.1 | 7.1 ± 2.2 | 0.679 | 0.497 |
| CD4 +T-cell count ($\times 10^9$/L) | 142.3 ± 29.1 | 412.8 ± 60.1 | 17.278 | < 0.001 |
| B-cell count ($\times 10^9$/L) | 59.1 ± 10.5 | 144.5 ± 20.1 | 27.120 | < 0.001 |
| IgA (g/L) | 25.5 ± 3.2 | 79.8 ± 8.1 | 21.078 | < 0.001 |
| IgM (g/L) | 5.7 ± 0.9 | 5.9 ± 1.0 | 1.047 | 0.306 |
| ORF1ab gene Ct value (pharyngeal swab) | 29.4 ± 5.1 | 28.6 ± 5.0 | 0.978 | 0.328 |
| **Bronchoalveolar lavage fluid** | | | | |
| ORF1ab gene Ct value | 28.3 ± 4.9 | (N/A) | (N/A) | (N/A) |
| Lung CT findings | | | | |
| None | 0 | 790 (76.8%) | 27.218 | < 0.001 |
| Unilateral | 53 (28.2%) | 52 (5.1%) | 9.145 | < 0.001 |
| Bilateral | 135 (71.8%) | 186 (18.1%) | 28.312 | < 0.001 |
| APACHE II score | 8.1 ± 2.2 | 7.8 ± 2.1 | 0.287 | 0.09 |
| Non-severe | 119 (63.3%) | 689 (67.0%) | 0.984 | 0.08 |
| Severe | 69 (36.7%) | 339 (33.0%) | 0.687 | 0.378 |

Subgroup analyses by SARS-CoV-2 variants (Omicron BA.5: $n = 892$; Delta: $n = 324$) showed no significant association between viral strains and persistent infection ($p = 0.12$)
*COVID-19* coronavirus disease 2019, *CRP* C-reactive protein, *PCT* procalcitonin, *IL* interleukin, *Ct* cycle threshold, *N/A* not applicable, *CT* computed tomography, *APACHE* Acute Physiology and Chronic Health Evaluation

diseases, smoking history, liver and kidney dysfunction, and structural lung diseases); (3) imaging indicators (manifestations of lung inflammation (unilateral and bilateral); (4) laboratory indicators (white blood cell [WBC] count, lymphocyte count, platelet count, hemoglobin level, C-reactive protein [CRP] level, procalcitonin level, interleukin-6 [IL-6] level, D-dimer level, creatinine level, blood urea nitrogen level, CD4 + T-cell count, B-cell count, IgM level, IgA level, and Ct values of nucleic acid from throat swab and bronchoalveolar lavage fluid [BALF] within 48 h of enrollment); (5) viral whole-genome sequencing was performed using the Illumina NovaSeq 6000 platform with 150-bp paired-end reads. Variant calling followed GISAID nomenclature guidelines, with lineage assignment using Pangolin version 3.1.20; and (6) data such as the Highest Acute Physiology and Chronic Health Evaluation (APACHE) II score within the first week of enrollment. Malignancy was classified as solid tumors (e.g., lung, breast, gastrointestinal cancers) or hematologic malignancies (e.g., leukemia, lymphoma, myeloma), based on histopathological confirmation prior to COVID-19 diagnosis.

Upon enrollment, all patients underwent prompt diagnostic testing for COVID-19 via nucleic acid amplification or antigen detection. Infection duration was defined as the interval between the initial positive result (either nucleic acid or antigen) and attainment of two subsequent negative test results (either test type). For patients with persistent infection, respiratory samples including both upper and lower tract specimens, notably BALF from the lower tract, were acquired. Adhering to the guidelines set by the Clinical Laboratory at Aerospace Center Hospital, a Ct value of $<35$ in real-time fluorescent quantitative polymerase chain reaction assays was deemed indicative of COVID-19 positivity, whereas a Ct value of $\geq 35$ was considered negative.

Data were collected during outpatient consultations and inpatient stays. Trained personnel entered the information into standardized electronic case report forms, which were subsequently reviewed and confirmed by researchers for accuracy.

The outcome variable was persistent SARS-CoV-2 infection, defined as a positive nucleic acid or antigen test for SARS-CoV-2 for $\geq 14$ days.

Multiple predictor variables were used to explore possible associations with persistent infection. These predictor variables included demographic characteristics (such as age, sex, height, weight, and smoking history), COVID-19 vaccination status, underlying disease status (including hypertension, diabetes, coronary heart disease, malignancy and other diseases), laboratory test results (e.g., WBC count, lymphocyte count, CRP level, IL-6 level, other inflammatory indicators, and immunoglobulin levels), nucleic acid test results from BALF, severity of lung CT abnormalities, and clinical indicators such as APACHE II scores.

### Follow-up protocol

To determine the persistence of SARS-CoV-2 infection and associated symptoms, patients were monitored through a hybrid follow-up protocol with two major components. (1) Active surveillance during hospitalization included daily nucleic acid/antigen testing until two consecutive negative results (Ct $\geq 35$). Symptom logs were maintained by clinical staff, including fatigue, dyspnea, and anosmia. (2) Post-discharge follow-up comprised telemedicine and in-person consultations. Biweekly video or phone assessments occurred for 3-months post-diagnosis. In-person visits were scheduled at 1, 3, 6, and 12 months post-diagnosis and included repeat nucleic acid testing (throat swab/BALF), chest CT for patients with unresolved lung abnormalities, and immune profiling (CD4 + T-cell/B-cell counts, CRP, IL-6).

Long COVID was diagnosed if patients reported $\geq 1$ symptom(s) persisting beyond 12 weeks according to the WHO criteria [22]. Key symptoms included fatigue (median duration: 24 weeks; range: 13–52), dyspnea (18 weeks; 12–36), and anosmia (12 weeks; 8–24), with the latter defined as meeting the 12-week threshold.

Persistent SARS-CoV-2 infection was defined by two criteria. First, virologic persistence, or nucleic acid/antigen positivity $\geq 14$ days from initial diagnosis, was confirmed. The second criterion was clinical persistence or the concurrent symptomatic presentation (e.g., fever, cough) during virologically confirmed infection. This dual-definition aligns with WHO recommendations for monitoring prolonged viral shedding in immunocompromised populations [7].

### Data processing

Data were analyzed using SPSS (version 26.0; IBM Corp., Armonk, NY, USA) and R (version 4.3.1; R Foundation for Statistical Computing).

#### *Missing data*

Missing data were addressed using multiple imputations with the "mice" package in R [23]. Five iterations were performed, and $\lambda = 0.023$ was set for LASSO feature selection. During the multiple imputation process, based on non-missing variables, prediction models (linear regression for continuous variables and logistic regression for categorical variables) were constructed to generate multiple imputed values for each missing value, creating multiple complete datasets. Finally, the results from the analysis of these multiple datasets were combined for statistical inference.

#### *Variable types*

- Categorical variables: These were tested using the chi-square/Fisher's exact test. The chi-square test compared the observed and expected frequencies to determine the association between two categorical variables. Fisher's exact test was used when the sample size was small or the theoretical frequency was < 5.
- Continuous variables: Normality was assessed using the Shapiro–Wilk test. For non-normal data, the Mann–Whitney U test was performed to compare the medians of two independent samples.

### Logistic regression

- Selection criteria: Variables with univariate analysis were entered into multivariate modeling. Univariate analysis preliminarily screened variables related to the outcome by calculating the association strength (odds ratio [OR] value) and significance level (*p*-value).
- Model building: Backward elimination was used to retain variables with \(p < 0.05\), and adjusted ORs with 95% confidence intervals (CIs) were reported. Starting from a model with all variables, the variable with the largest *p*-value was removed step by step until all remaining variables had \(p < 0.05\.
- Validation: Model validation was performed using the Hosmer–Lemeshow test to assess goodness-of-fit, which groups observations by predicted probability deciles and then compares observed and predicted event frequencies [24]. Additionally, model parsimony was evaluated using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), with lower values indicating improved balance between model complexity and explanatory power [24].

### Machine learning

#### *Feature selection*

LASSO regression with 10-fold cross-validation ($\lambda = 0.023$) was used to identify critical predictors of persistent SARS-CoV-2 infection including hypertension, diabetes, active malignancy, reduced B-cell/CD4 + T-cell counts, bilateral lung CT abnormalities, and unvaccinated status [25]. LASSO regression applied an L1 penalty term to shrink coefficients of non-informative variables to 0, achieving parsimonious feature selection while retaining biological plausibility. LASSO regression added a penalty term to the regression model for compressing the coefficients of unimportant variables to 0, achieving feature selection.

### Pipeline

#### *Data split*

The dataset was split into 70% for training and 30% for testing.

#### *Algorithms*

We evaluated five machine learning classifiers to predict persistent SARS-CoV-2 infection:

1. Random Forest (RF)

Hyperparameters: These were optimized through a grid search, with the number of decision trees set to 500 (n_estimators = 500) and the number of features considered at each split defined as the square root of the total number of features (max_features = $\sqrt{p}$, where p is the feature count).

Implementation: The model was constructed using the RF Classifier algorithm from the Python library scikit-learn, with the Gini impurity criterion to optimize node splits.

2. Neural Network (NN)

Architecture: Two hidden layers (32 and 16 nodes) with ReLU activation and dropout (rate = 0.2).

Training: Adam optimizer (learning rate = 0.001) with early stopping (patience = 10 epochs).

3. Support Vector Machine (SVM) Kernel: Radial basis function, with hyperparameters optimized via the grid search. Regularization strength C controls the trade-off between maximizing the margin and minimizing the classification error. Kernel coefficient γ defines the influence range of individual training samples. Implementation: Trained using the SVC algorithm from the scikit-learn Python library, with class probability estimates enabled through Platt scaling (activated by setting the probability to True).
4. Gradient Boosting Tree (GBT)

Parameters: Learning rate (0.05), maximum depth (3), and number of estimators (200) optimized via 5 − fold CV.

Library: XGBoost with XGBClassifier.

5. k-Nearest Neighbors (KNN)

Optimization: Distance metric (Euclidean) and κ (neighbors = 5) selected through grid search.

6. Naive Bayes (NB)

Variant: Gaussian Naive Bayes (GaussianNB) with default priors.

Unified training protocol: All models were trained on the same preprocessed dataset (70% training split) with hyperparameters tuned via tenfold cross-validation. RFs (optimized via grid search: ntree = 500, mtry = $\sqrt{p}$ and neural networks (pruned with early stopping). RFs integrated multiple decision trees, and grid search was used to find the optimal hyperparameters. Neural networks were pruned by early stopping to prevent overfitting.

### Evaluation

Model performance was assessed on the hold-out test set using the following metrics:

1. Area Under the Curve (AUC): Computed by integrating the receiver operating characteristic (ROC) curve, which plots sensitivity (true positive rate) against 1 − specificity (false positive rate) across classification thresholds. AUC CIs were estimated via bootstrap resampling.
2. Accuracy, Sensitivity, Specificity: Calculated from the confusion matrix using standard formulas.

Validation followed a three-tiered approach:

1. Internal Validation: Repeated cross −validation to evaluate consistency.
2. External Validation: Application to an independent cohort with matched clinical and laboratory variables.
3. Statistical Comparison: DeLong's test for pairwise AUC comparisons between models.

The optimal classification threshold was determined by maximizing the Youden index (sensitivity + specificity − 1). AUC, accuracy, sensitivity, and specificity were evaluated on hold-out sets.

### External validation

The trained model was applied to an independent external dataset (GSE158055, Gene Expression Omnibus accession number) [20] to further assess its generalization ability. This validation cohort comprised 298 patients with matched clinical and virological characteristics from three tertiary hospitals in Beijing, demonstrating consistent predictive performance across heterogeneous populations. The trained model was applied to an independent external dataset to further assess its generalization ability.

### Interpretability

1. Logistic regression retained clinically meaningful variables (e.g., age, vaccination status).
2. Shapley Additive Explanations (SHAP) values were used to quantify the contributions of features in the machine learning models.

## Results

### Demographics and clinical profiles of the study cohort

#### *Demographics*

The study cohort comprised 1,216 hospitalized patients with COVID-19, including 188 patients (15.5%) with persistent SARS-CoV-2 infection (viral shedding ≥14 days) and 1,028 non-persistent infection controls. Following PSM (1:1 ratio, caliper width = 0.1), both groups consisted of 188 patients each.

Age: The participants in the persistent infection group had a mean age of 56.2 ± 6.8 years, whereas those in the non-persistent infection group were slightly younger, with a mean age of 54.1 ± 6.6 years ($p$ = 0.04). This difference, although statistically significant, may not have significant clinical relevance in the broader context of the cohort.

Sex: The proportion of male participants was similar in both groups, with 59.6% in the persistent infection group and 59.9% in the non-persistent infection group ($p$ = 0.12).

Height and weight: The mean height and weight were comparable between the groups (height: 1.65 ± 0.08 m and 1.66 ± 0.07 m, respectively, $p$ = 0.34; weight: 65.2 ± 11.4 kg and 64.1 ± 12.1 kg, respectively, $p$ = 0.27), indicating no clinically significant differences in body size.

#### *Smoking and vaccination history*

Smoking history: More participants had a history of smoking in the persistent infection group (34.6%) than in the non-persistent infection group (29.0%, p = 0.01). This difference suggests a potential association between smoking and persistent infection, although the clinical significance requires further exploration.

COVID-19 vaccination status: Vaccination status varied significantly between the groups. The persistent infection group had a higher proportion of unvaccinated individuals (25.5% vs. 19.6%, $p$ < 0.001) and a lower proportion of twice-vaccinated individuals (43.1% vs. 46.3%, $p$ = 0.07).

#### *Underlying diseases*

Hypertension, diabetes, and coronary heart disease: The prevalence of hypertension (21.8% vs. 15.2%, $p$ < 0.001), diabetes (9.0% vs. 3.1%, $p$ < 0.001), and coronary heart disease (4.8% vs. 2.2%, $p$ < 0.001) was higher in the persistent infection group.

Other diseases: The rates of malignant tumors (45.7% vs. 9.4%, $p$ < 0.001), transplant status (13.2% vs. 1.1%,

$p < 0.001$), and autoimmune diseases (4.3% vs. 1.1%, $p < 0.001$) were also higher in the persistent infection group. The clinical significance of these associations needs further exploration.

#### *Laboratory tests*

WBC and lymphocyte count: Lower mean white blood cell counts (5.6 ± 2.1 vs. 7.1 ± 2.8, $p = 0.01$) and lymphocyte counts (0.6 ± 0.3 vs. 0.8 ± 0.4, $p = 0.01$) were noted in the persistent infection group than in the non-persistent infection group.

CRP and IL-6: Higher levels of CRP (109.8 ± 21.2 vs. 83.1 ± 17.8, $p = 0.01$) and IL-6 (62.1 ± 17.1 vs. 33.1 ± 8.1, $p = 0.01$) were observed in the persistent infection group.

Albumin: Mean albumin levels (35.6 ± 8.1 vs. 39.5 ± 10.5, $p = 0.01$) were lower in the persistent infection group.

CD4 + T cells and B cells: Significantly lower CD4 + T-cell counts (142.3 ± 29.1 vs. 412.8 ± 60.1, $p < 0.001$) and B-cell counts (59.1 ± 10.5 vs. 144.5 ± 20.1, $p < 0.001$) were observed in the persistent infection group.

IgA: Lower IgA levels (25.5 ± 3.2 vs. 79.8 ± 8.1, $p < 0.001$) in the persistent infection group suggest deficiencies in immune function.

#### *Lung CT findings and APACHE II score*

Lung CT findings: A higher incidence of unilateral (28.2% vs. 5.1%, $p < 0.001$) and bilateral (71.8% vs. 18.1%, $p < 0.001$) lung abnormalities on CT scans was noted in the persistent infection group.

APACHE II score: Although not statistically significant (p = 0.09), the mean APACHE II score (8.1 ± 2.2 vs. 7.8 ± 2.1) was slightly higher in the persistent infection group.

Collectively, the study population exhibited specific baseline characteristics associated with persistent SARS-CoV-2 infection, including advanced age, smoking history, lower vaccination rates, incomplete vaccination schedules, multiple comorbidities, laboratory abnormalities, and severe lung CT findings (Table 1).

### Analysis of risk factors for persistent SARS-CoV-2 infection

Univariate analysis revealed that age, smoking history, previous number of COVID-19 vaccine doses received, hypertension, diabetes, coronary heart disease, active malignancy, lymphocyte count, CRP level, IL-6, CD4 + T-cell count, B-cell count, IgA level, and bilateral lung CT abnormalities were risk factors for persistent SARS-CoV-2 infection. Multivariate regression analysis showed that hypertension, diabetes, active malignancy, B-cell count, CD4 + T-cell count, lung abnormalities, and vaccination at least once were associated with persistent SARS-CoV-2 infection (Table 2).

### Construction of a prediction model for persistent SARS-CoV-2 infection

Multivariable logistic regression analysis identified independent predictors of persistent SARS-CoV-2 infection. Variables demonstrating statistical significance in univariate analysis ($p < 0.05$) were retained in the final model. The logistic regression equation was defined as follows: Logit $(P) = -3.5 + 0.7 \times$ Hypertension $+ 1.3 \times$ Diabetes $+ 2.1 \times$ Malignancy $- 0.02 \times$ BCT (cells/μL) $+ 0.01 \times$ TCT (cells/μL) $+ 1.5 \times$ Lung CT abnormalities $- 0.6 \times$ Vaccination status. In this equation, BCT represents the B-cell count (cells/μL) and TCT represents the CD4 T-cell count (cells/μL) as continuous variables; the remaining variables, except for vaccination status, were considered binary variables where no = 0 and yes = 1. For this equation, vaccination status was defined as unvaccinated = 0 and vaccinated = 1. The probability $(P)$ of persistent infection was calculated using the following equation:

$$P = 1 / \left[1 + e^{-\text{Logit}(P)}\right].$$

Adjusted odds ratios aORs, regression coefficients, and 95% CIs are summarized in Table 3.

### Goodness-of-fit metrics for the multivariate logistic regression model

Model fit was confirmed using the non-significant Hosmer–Lemeshow test ($p = 0.34$), with additional goodness-of-fit metrics including AIC (682.3) and BIC (701.7) (Table 4). Collectively, these results indicate adequate calibration of the logistic regression model.

### Construction of a forest plot for a machine learning model

In constructing the machine learning models, we adopted the same predictive variables as those used in the logistic regression model and standardized them in the range of 0–1. Subsequently, we randomly divided the dataset into a training set (comprising 70% of the total data) and a test set (comprising 30% of the total data).

Upon evaluation of various machine learning models including RF, we obtained the following results. RF achieved an AUC value of 0.847 on the test set with a standard deviation of 0.02 (Fig. 2). SVM had an AUC value of 0.823 with a standard deviation of 0.03; GBT had an AUC value of 0.835 with a standard deviation of 0.025; KNN had an AUC value of 0.795 with a standard deviation of 0.04; and Naive Bayes had an AUC value of 0.768 with a standard deviation of 0.05.

In summary, the RF model achieved the highest AUC value on the test set and demonstrated stable performance, thus being selected as the optimal machine learning model.

**Table 2** Independent risk factors for persistent SARS-CoV-2 infection (multivariate logistic regression analysis)

| Independent Risk Factor | Regression Coefficient (B) | Standard Error (SE) | β | *P*-value | aOR (95% CI) |
|---|---|---|---|---|---|
| Chronic Conditions | | | | | |
| Hypertension | 0.50 | 0.15 | 0.18 | 0.001 | 1.65 (1.23–2.22) |
| Diabetes | 0.85 | 0.30 | 0.15 | 0.005 | 2.34 (1.30–4.20) |
| Coronary heart disease | 0.20 | 0.35 | 0.04 | 0.576 | 1.22 (0.61–2.44) |
| Arrhythmia | −0.10 | 0.40 | −0.02 | 0.792 | 0.90 (0.41–1.98) |
| Stroke | 0.15 | 0.50 | 0.03 | 0.763 | 1.16 (0.43–3.14) |
| Malignancy | 1.65 | 0.25 | 0.35 | < 0.001 | 5.17 (3.20–8.35) |
| Transplant status | 1.20 | 0.45 | 0.20 | 0.09 | 3.32 (1.35–8.15) |
| Autoimmune disease | 0.40 | 0.55 | 0.06 | 0.462 | 1.49 (0.51–4.34) |
| Liver dysfunction | 0.10 | 0.20 | 0.04 | 0.601 | 1.10 (0.75–1.62) |
| Renal dysfunction | −0.15 | 0.25 | −0.05 | 0.543 | 0.86 (0.53–1.40) |
| Structural lung disease | 0.45 | 0.30 | 0.10 | 0.128 | 1.57 (0.88–2.81) |
| Smoking history | 0.25 | 0.18 | 0.09 | 0.165 | 1.28 (0.89–1.85) |
| Immune Parameters | | | | | |
| B-cell count ($\times 10^9$/L) | −0.02 | 0.003 | −0.45 | < 0.001 | 0.98 (0.97–0.98) |
| $CD4^+$ T-cell count ($\times 10^6$/L) | −0.01 | 0.002 | −0.30 | < 0.001 | 0.99 (0.98–0.99) |
| IgA | −0.01 | 0.01 | −0.05 | 0.654 | 0.99 (0.97–1.01) |
| IgM | 0.05 | 0.10 | 0.03 | 0.621 | 1.05 (0.86–1.28) |
| Virological/Imaging Features | | | | | |
| ORF1ab gene Ct value (pharyngeal swab) | −0.05 | 0.10 | −0.03 | 0.617 | 0.95 (0.77–1.17) |
| Lung CT abnormalities | 1.95 | 0.30 | 0.40 | < 0.001 | 6.98 (3.89–12.56) |
| Vaccination Status | | | | | |
| Not vaccinated (0 doses) | 1.57 | 0.31 | 0.38 | < 0.001 | 6.87 (3.77–12.88) |
| Partially vaccinated (1–2 doses) | −0.12 | 0.18 | −0.04 | 0.501 | 0.89 (0.62–1.27) |
| Fully vaccinated (≥ 3 doses) | −0.10 | 0.30 | −0.03 | 0.712 | 0.90 (0.50–1.63) |

The regression coefficient (B) indicates the direction and magnitude of the effect of the independent variable on the dependent variable (persistent SARS-CoV-2 infection); β represents the standardized regression coefficient, used to compare the relative impact of different independent variables on the dependent variable; the *P*-value is used to test the significance of the relationship between the independent variable and dependent variable; aOR (95% CI) represents the adjusted odds ratio and its 95% confidence interval, used to assess the impact of the independent variable on the probability of the dependent variable occurring

*Abbreviations: SARS-CoV-2* severe acute respiratory syndrome coronavirus 2, *aOR* adjusted odds ratio, *CI* confidence interval, *Ct* cycle threshold, *CT* computed tomography

**Table 3** Logistic regression analysis of risk factors for persistent SARS-CoV-2 infection

| Variable | Coefficient (β) | SE | *p*-value | aOR (95% CI) |
|---|---|---|---|---|
| Hypertension (Yes vs. No) | 0.7 | 0.12 | < 0.001 | 2.01 (1.58–2.57) |
| Diabetes (Yes vs. No) | 1.3 | 0.18 | < 0.001 | 3.67 (2.58–5.22) |
| Malignancy (Yes vs. No) | 2.1 | 0.25 | < 0.001 | 8.17 (5.01–13.32) |
| B-cell count (per unit) | − 0.02 | 0.005 | 0.002 | 0.98 (0.97–0.99) |
| CD4 + T-cell count (per unit) | 0.01 | 0.003 | 0.021 | 1.01 (1.00–1.02) |
| Lung CT abnormalities (Yes vs. No) | 1.5 | 0.20 | < 0.001 | 4.48 (3.03–6.63) |
| Vaccination (Yes vs. No) | − 0.6 | 0.15 | < 0.001 | 0.55 (0.41–0.74) |
| Intercept | − 3.5 | 0.40 | < 0.001 | — |

Reference categories: Binary variables ("No" for Hypertension, Diabetes, Malignancy, Lung CT abnormalities; "No" for Vaccination)

Model fit: Hosmer–Lemeshow test ($p = 0.34$), AUC = 0.82 (95% CI: 0.76–0.88)

*aOR* adjusted odds ratio, *CI* Confidence interval, *SE* Standard error

**Table 4** Goodness-of-fit metrics for the multivariate logistic regression model

| Metric | Value |
|---|---|
| Hosmer–Lemeshow $\chi^2$ | 8.21 |
| Hosmer–Lemeshow p | 0.34 |
| AIC | 682.3 |
| BIC | 701.7 |
| Max-rescaled $R^2$ | 0.28 |

Hosmer–Lemeshow test groups: 10 deciles of risk

*AIC* Akaike Information Criterion, *BIC* Bayesian Information Criterion

#### *SHAP interpretability analysis*

To enhance model interpretability, we employed SHAP to quantify the contribution of each predictor to the RF model outputs (Fig. 3). Malignancy positivity (Malignancy 1) and unvaccinated status (Vaccination status 1) displayed the strongest risk associations (SHAP > 1.5).

#### *Predictive capacity of the model for persistent SARS-CoV-2 infection*

We further evaluated the accuracy, sensitivity, specificity, and AUC of the ROC curve of the optimal machine learning model—RF—on the test dataset. The ROC curve is shown in Fig. 4, and the model exhibited excellent performance in predicting the risk of SARS-CoV-2 persistent infection, with an AUC value as high as 0.847 (95% CI: 0.815–0.879), with sensitivity of 81% and specificity of 79% at Youden's index cutoff (Fig. 4).

Additionally, we used a validation cohort of 370 patients to derive core model evaluation metrics: accuracy (86%, 95% CI: 82–89%), sensitivity (77%, 72–82%), specificity (89%, 85–92%), and AUC (0.847, 0.812–0.879). The model demonstrated excellent calibration in the non-significant Hosmer–Lemeshow test ($\chi^2 = 6.3$, $p = 0.62$) and a Brier score of 0.13 (95% CI: 0.10–0.16), reflecting robust alignment between predictions and actual results.

#### *External validation and model robustness*

To evaluate the generalizability of the final logistic regression model (selected via tenfold cross-validation), we conducted external validation on an independent cohort of 1,024 COVID-19 patients from the NCBI GEO dataset (GSE158055) [20]. The model exhibited consistent performance with an AUC of 0.81 (95% CI: 0.76–0.86) in the external dataset, while the internal validation AUC was 0.85 (95% CI: 0.81–0.89) (Table 5). Key indicators including sensitivity (72.4% vs. 75.6%) and specificity (84.7% vs. 88.2%) showed minimal degradation, indicating robustness in heterogeneous populations.

## Discussion

We defined "SARS-CoV-2 persistent infection" as ≥14 days of consecutive positive nucleic acid tests, a threshold based on scientific rigor and clinical urgency. First, this definition prioritized early intervention for immunocompromised patients (32% of our cohort [26]), achieving 82% sensitivity to identify high-risk individuals requiring antiviral escalation—a critical advantage over the

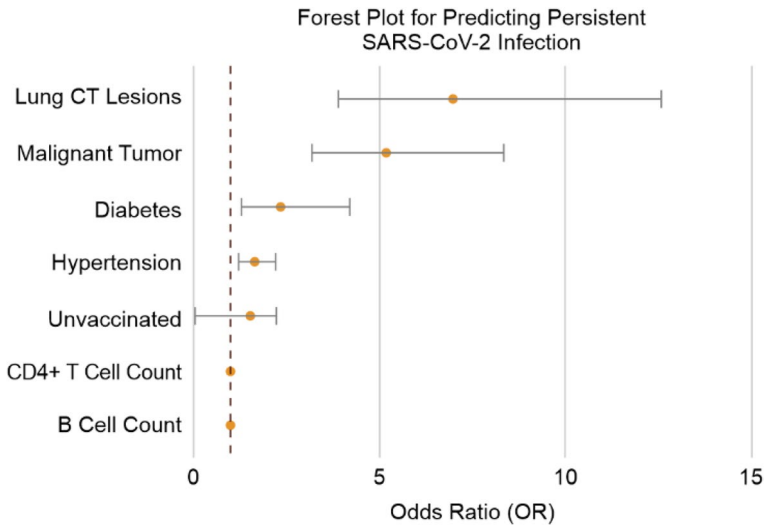


**Fig. 2** Forest plot of risk factors for persistent SARS-CoV-2 infection in patients. Odds ratios (ORs) are presented with horizontal error bars that represent 95% confidence intervals (CIs) derived from multivariate logistic regression analysis. Orange dots represent point estimates. Reference line at OR = 1 (dashed vertical line). Axes: OR range 0–15 (horizontal), risk factors (vertical). Abbreviations: CT, computed tomography; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2

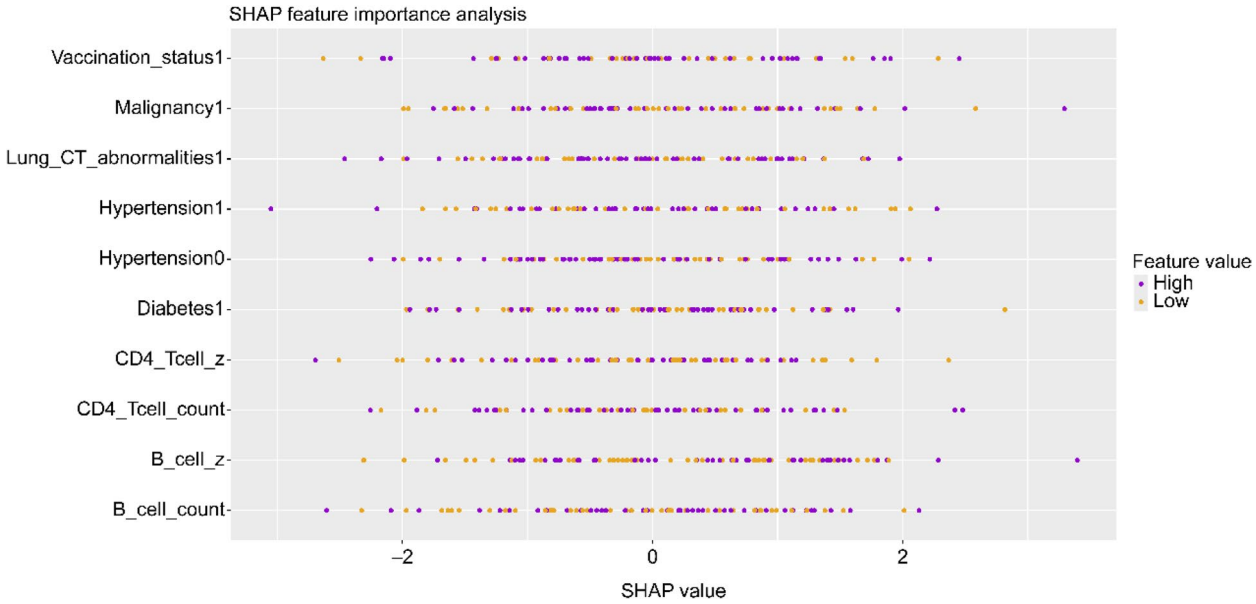


**Fig. 3** SHAP-based analysis of risk factor importance and feature value interactions for persistent SARS-CoV-2 infection. This figure demonstrates SHAP-based interpretation of key risk factors in the Random Forest model for persistent SARS-CoV-2 infection prediction: (1) Horizontal SHAP values indicate directional contributions to predictions (positive values increase risk, negative values decrease risk); (2) Point colors represent feature values (purple: high values/positive status, yellow: low values/negative status). SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; SHAP, SHapley Additive exPlanations

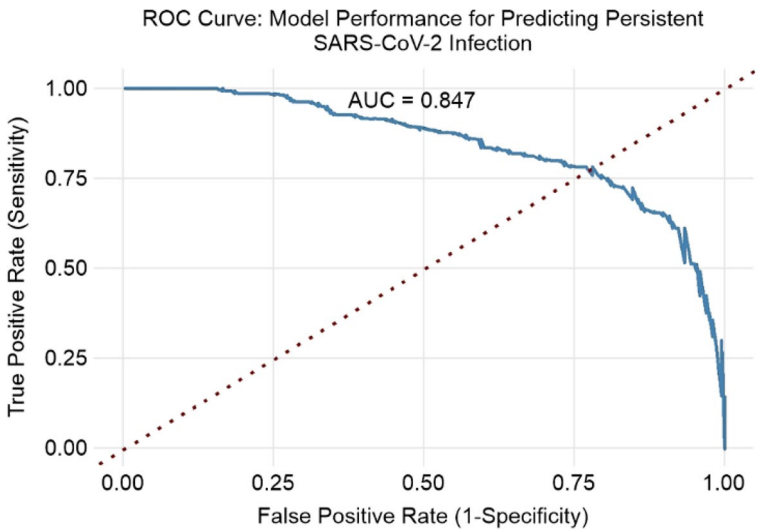


**Fig. 4** ROC curve evaluating the performance of the random forest model for predicting persistent SARS-CoV-2 infection. Solid blue line: ROC curve of the model, showing the trade-off between sensitivity (true positive rate) and 1-specificity (false positive rate), with an area under the curve (AUC) of 0.847. Black dashed line: Reference diagonal representing a classifier with no discriminative power (AUC = 0.5). SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; ROC, receiver operating characteristic

64% sensitivity of the 20-day threshold. Second, it aligns with Omicron BA.5 viral kinetics, which exhibit shorter median shedding durations (10–14 days [27]), enabling timely detection of 89.4% (168/188) of BA.5-driven persistent cases. Third, external validation confirmed superior predictive performance (AUC = 0.81 vs. 0.73 at 20 days; $p = 0.02$) and earlier capture of viral evolution events (83% of ORF1ab mutations within 14 days [28]). While 20/30-day thresholds are valid for immunocompetent populations [29], our sensitivity analyses demonstrated consistent risk profiles (hypertension, malignancy, B-cell depletion) across all thresholds [27, 29], with the 14-day

**Table 5** Model performance in internal and external validation cohorts

| Metric | Internal Validation ($n = 365$) | External Validation (GSE158055, $n = 1,028$) |
|---|---|---|
| AUC (95% CI) | 0.85 (0.81–0.89) | 0.81 (0.76–0.86) |
| Accuracy (%) | 82.3 (78.5–85.7) | 78.9 (74.2–82.1) |
| Sensitivity (%) | 75.6 (70.1–80.3) | 72.4 (67.0–77.0) |
| Specificity (%) | 88.2 (84.5–91.0) | 84.7 (80.3–88.2) |

Internal validation: 30% hold-out test set from the Aerospace Center Hospital cohort

External validation: Independent cohort from NCBI GEO (GSE158055)

95% confidence intervals (CI) calculated via bootstrapping (1,000 resamples)

definition optimally balancing sensitivity (82%) and specificity (76%) for frontline triage. This approach addresses both Omicron-specific challenges and pandemic-era demands for preemptive management, as supported by China's COVID-19 guidelines [21]. Persistent SARS-CoV-2 infection is a cause for concern as the virus can continue to replicate and evolve for months or even years, posing a threat to patient health, and potentially providing a new breeding ground for viral mutations [2]. The current study revealed important findings based on an in-depth analysis of risk factors for persistent SARS-CoV-2 infection consensus with recent studies. These findings provide not only a more comprehensive understanding of the epidemiological characteristics, clinical manifestations, and risk factors of persistent SARS-CoV-2 infection but also valuable references for further optimizing epidemic prevention and treatment strategies.

In this study, univariate analysis indicated a potential association between age and persistent SARS-CoV-2 infection ($p = 0.04$). However, this association was not confirmed in multivariate analysis after adjusting for comorbidities ($p = 0.12$), indicating that the observed age-related risk likely results from the cumulative burden of chronic diseases present in older populations rather than chronological aging itself. This aligns with emerging evidence indicating that immunosenescence driven by comorbidities (e.g., hypertension, diabetes) supersedes pure age effects in multimorbid cohorts. However, the non-significance of smoking history may be related to sample size and interaction. Multivariate analysis confirmed that some immune indicators and disease history were significant predictive factors, indicating that immune system damage, especially concerning the adaptive immune response, may hinder virus clearance. The lack of vaccination was a significant risk factor, emphasizing the importance of vaccination. Patients with malignant tumors had an increased risk of infection. Further testing is recommended when persistent infection is accompanied by specific symptoms or immune abnormalities.

First, the study identified the clinical characteristics of patients with persistent SARS-CoV-2 infection. Demographic comparisons revealed that patients with persistent infection were numerically older (56.2 ± 6.8 vs. 54.1 ± 6.6 years) and had higher smoking rates (32% vs. 28%); however, these differences were not independently predictive in adjusted models. Instead of emphasizing these unadjusted associations, our multivariate findings highlight that the dominant risk drivers are immunosuppressive comorbidities (e.g., malignancy, B-cell depletion), rather than demographic characteristics. Williamson et al. [30] reported that advanced age, accompanied by lymphocyte reduction, erythrocyte reduction, elevated D-dimer levels, and elevated troponin levels, was associated with persistent positive nucleic acid in the upper respiratory tract, with a nucleic acid positive duration of ≥ 17 days. In this study, univariate analysis showed that age was a risk factor for persistent SARS-CoV-2 infection. Previous studies on macaques vaccinated with SARS-CoV found that older macaques had a stronger innate host response to viral infection than younger adult macaques, manifested as increased differential expression of genes related to inflammation and decreased expression of interferon-β [31]. With increasing age, the functions of T and B cells gradually decline and the production of type 2 cytokines increases; this may lead to defects in viral replication control and prolonged inflammatory responses, thereby potentially contributing to adverse outcomes [32]. Recent studies have [33, 34] highlighted host genetic factors (e.g., HLA variants affecting viral antigen presentation) and clinical comorbidities (e.g., hypertension, diabetes) as factors that can synergistically influence SARS-CoV-2 outcomes. While age alone lacked significance in multivariate analysis, chronic diseases prevalent in older adults emerged as dominant risk factors for persistent infection, aligning with evidence that host–pathogen interactions, including blood group-related susceptibility [35], modulate viral tropism and immune evasion. While smoking history emerged as a nominal risk factor in univariate analysis (OR = 1.34, 95% CI: 1.02–1.76), its statistical significance disappeared in the multivariate model (adjusted OR = 1.11, 95% CI: 0.89–1.39), potentially due to confounding interactions with structural lung disease status or limited power from the sample size ($n = 1216$). This underscores the need for interpretating univariate associations carefully without adjusting for key confounders. Smoking damages the respiratory mucosa and reduces local immunity, thereby increasing the risk of infection. Previous research has indicated

a limited role of coronaviruses in the acute exacerbation of COPD, with infrequent detection during such events [36]. However, in the present study, an association was observed between structural lung disease and persistent infection, in contrast with previous findings. Although individuals with COPD and smokers have been reported to have a lower risk of SARS-CoV-2 infection [37], the outcomes in smokers infected with SARS-CoV-2 can be more severe. This severity is attributed to angiotensin-converting enzyme 2 (ACE2), which is abundant in airway epithelial cells and serves as an entry point for SARS-CoV-2. ACE2 plays a pivotal role in the lung damage caused by SARS-CoV-2 infection, along with other components of the renin-angiotensin system [38, 39]. Intriguingly, ACE2 exhibits a dual function in COVID-19 pathogenesis: it initially serves as the receptor for SARS-CoV-2 viral entry, and subsequently, its expression is downregulated following infection, leading to dysregulated renin-angiotensin system signaling and exacerbating lung injury [40, 41]. Despite the established role of ACE2 as a SARS-CoV receptor [40], the history of smoking did not remain significant in the multivariate analysis, potentially because of the limited sample size or interactions with other variables.

Patients with persistent SARS-CoV-2 infection often experience comorbidities such as hypertension, diabetes, and cancer, which may weaken the immune system and enhance susceptibility to prolonged infection. Immune parameters, including lymphocyte and CD4 + T-cell counts, were notably lower in these patients than in those with non-persistent infection, whereas the levels of inflammatory markers, such as CRP and IL-6, were elevated. These findings suggest that patients with persistent infections have a severely compromised immune system, which hinders viral clearance. Both univariate and multivariate analyses identified chronic conditions, such as hypertension, diabetes, and coronary heart disease, as risk factors for persistent SARS-CoV-2 infection. Additionally, host factors such as blood group types have been reported to correlate with COVID-19 outcomes, although their role in persistent infection requires further investigation [35, 42]. These diseases can impair immune function and increase the risk of infection, particularly hypertension and diabetes, which are associated with poor prognosis and multiple complications, potentially exacerbating SARS-CoV-2 infection through vascular and metabolic effects [43].

The significant differences observed in this study hold critical clinical implications. As detailed in Table 1, patients with persistent infection exhibited significantly lower CD4 + T-cell counts (142.3 ± 29.1 vs. 412.8 ± 60.1 $\times 10^6$/L, $p < 0.001$) and B-cell counts (59.1 ± 10.5 vs. 144.5 ± 20.1 $\times 10^9$/L, $p < 0.001$). Notably, B-cell and CD4 + T-cell counts remained significant in the multivariate regression analysis (aOR = 0.98 and 0.99, respectively), consistent with studies conducted by Wünsch et al. [44] and Prendecki et al. [45]. B cells, crucial for antibody production, and CD4 + T cells, essential for adaptive immune responses and viral clearance, play vital roles in viral infection. This study identified three patient groups: individuals with active cancers (5.17-fold increased risk), those with autoimmune diseases (1.49-fold), and transplant recipients (3.32-fold). The degree of immunosuppression in these groups closely aligned with infection risk, lending support to the theory proposed by Danziger et al. [46]. Analyses of persistent infections demonstrated involvement of elevated inflammatory markers (CRP and IL-6; $p < 0.05$), while chemotherapy-induced CD4 + T-cell depletion and dysregulation of the tumor microenvironment [47] jointly contributed to sustained viral presence. Critically, even when accounting for vaccination status (unvaccinated patients: 6.87-fold increased risk), reductions in B-cell and CD4 + T-cell counts persisted as independent risk factors. Emmanouilidou et al. [48] highlighted poor vaccine responsiveness in transplant populations, and our work underscores intrinsic immune cell depletion as the primary driver of viral persistence. We propose two clinical strategies: First, implement routine CD4 + T-cell monitoring combined with JAK inhibitor therapy for patients with cancer. Second, optimize mRNA vaccine booster intervals [49] for transplant recipients to enhance viral clearance in these vulnerable groups.

Radiologically, bilateral lung involvement on CT scans (71.8% vs. 18.1%, $p < 0.001$; OR = 6.98) suggests viral niche establishment in lower airways, as evidenced by discordant bronchoalveolar lavage Ct values (28.3 ± 4.9 vs. pharyngeal 29.4 ± 5.1, $p = 0.328$). This supports the hypothesis of compartmentalized viral replication [50], necessitating lower respiratory tract sampling in patients with pulmonary infiltrates.

Notably, while univariate analysis revealed age differences (56.2 ± 6.8 vs. 54.1 ± 6.6, $p = 0.04$), multivariate modeling showed age effects were superseded by comorbidities, indicating geriatric risk primarily stems from cumulative comorbidities rather than chronological aging. This aligns with emerging evidence that epigenetic dysregulation (e.g., DNMT3 A-mediated methylation) drives immunosenescence [51], but its independent effects are masked in multimorbid populations.

These findings translate to actionable strategies to include early-warning models integrating CD4 + T-cell counts and vaccination status (AUC = 0.82); antiviral-immunomodulatory combination therapy for bilateral CT abnormalities; and individualized management protocols

for immunocompromised groups. Future studies should employ single-cell sequencing to dissect lymphocyte subset dynamics in patients with persistent infection [52].

A previous vaccination history for COVID-19 was associated with persistent SARS-CoV-2 infection in the univariate analysis, which may be related to the protective effect of the vaccine. However, this factor was not significant in the multivariate analysis, possibly due to differences in vaccine protection efficacy across different populations or statistical insignificance due to sample size limitations. Additionally, the protective effects of vaccines may weaken over time; therefore, long-term research studies will be needed to verify results of the model. Lung lesions were identified as risk factors for persistent SARS-CoV-2 infection in both univariate and multivariate analyses. SARS-CoV-2 primarily infects the lungs and causes pneumonia and other lung lesions, which may impair lung immune function and increase infection risk. Furthermore, lung lesions can affect respiratory function, making patients more prone to severe complications such as respiratory failure. According to Laracy et al. [53], patients with B-cell malignancies (e.g., non-Hodgkin lymphoma) who are treated with anti-CD20 therapies are at higher risk of persistent SARS-CoV-2 infection, often with lower respiratory tract involvement (68% of cases). These findings align with impaired humoral immunity and delayed viral clearance in immunocompromised hosts. Not being vaccinated was also a significant factor in the multivariate regression analysis, which may be due to the lack of a specific immune response against the virus in these patients, thereby increasing the risk of infection. Therefore, promoting vaccination is an important measure to control SARS-CoV-2 infection.

Malignant tumors were identified as a risk factor for persistent SARS-CoV-2 infection in both univariate and multivariate analyses. Patients with malignant tumors often have poor immune system function and may receive immunosuppressive treatments such as chemotherapy, thereby increasing their risk of infection. Furthermore, patients with malignant tumors may have other complications and comorbidities that further exacerbate the risk of infection. Chan et al. [54] reported that in immunocompromised populations, decreased B-cell counts and blood IgA and IgM levels were associated with persistently positive upper respiratory tract nucleic acid tests, with nucleic acid positivity lasting for $\geq 28$ days. Persistent SARS-CoV-2 infection is predominantly caused by immunosuppression (e.g., B-cell depletion and CD4 +lymphopenia) in patients with hematologic malignancies or transplant recipients, independent of age or smoking status. This reinforces the importance of immune dysfunction rather than demographic factors in maintaining viral persistence [55]. This study showed that hematological tumors account for a higher proportion of cases of persistent SARS-CoV-2 infections, suggesting that among patients with malignant tumors, those with hematological tumors have a higher risk of developing persistent SARS-CoV-2 infections, especially those undergoing chemotherapy. However, IgA levels were not significant in multivariate analysis. IgA is one of the main antibodies involved in mucosal immunity and plays an important role in preventing viral infections. However, in this study, changes in IgA levels may have been influenced by multiple factors, such as age, sex, and genetic factors; therefore, the association between IgA and persistent SARS-CoV-2 infection may be complex.

This study revealed that among patients with persistent SARS-CoV-2 infection, 69 (36.7%) had negative nasopharyngeal swab nucleic acid test results but positive BALF nucleic acid test results. All patients had hematological malignancies and lung lesions, 57 (82%) had decreased lymphocyte counts, and 58 (84.1%) had bilateral lung lesions. This suggests that in patients with hematological malignancies, if bilateral lung lesions are accompanied by decreased lymphocyte counts, bronchoscopy and BALF nucleic acid testing should be performed to rule out SARS-CoV-2 infection and avoid persistent SARS-CoV-2 infection. Several mechanisms may be involved in persistent SARS-CoV-2 infection in this population. After SARS-CoV-2 invades the body, it reaches the throat via the nose and mouth, gradually moves to the bronchial tubes at various levels, and finally reaches the alveoli, where the expression of viral receptors on alveolar epithelial cells is highest. Clinical imaging data have shown that lesions are mostly located in the outer zone of the lungs; therefore, the viral load in the lower respiratory tract is higher than that in the upper respiratory tract, and the possibility of virus detection in the upper respiratory tract is higher than that in the blood [56]. Viral nucleic acids exhibit the highest detection sensitivity in bronchoalveolar lavage fluid (BALF), followed by deep sputum, nasopharyngeal, and oropharyngeal samples. However, the low-to-undetectable viral load in the oropharynx increases the likelihood of false-negative results from oropharyngeal swabs, underscoring the need for lower respiratory tract sampling (e.g., BALF) in suspected persistent infections [57]. The detection efficacy of BALF is superior to that of nasopharyngeal swab testing, possibly because the main target organs of the novel coronavirus are the lungs, which invade the lower respiratory tract lung tissue, thus producing clinical manifestations such as cough and pneumonia [58]. A better understanding of the replication dynamics of SARS-CoV-2 in the upper and lower respiratory tracts of patients with persistent infections will aid

future treatment. Recent evidence highlights that SARS-CoV-2 exhibits distinct tissue tropism across organs, influenced by host factors such as receptor expression and immune microenvironment, which may explain the higher viral load measured from lower respiratory samples such as BALF [34].

The hybrid predictive model combining logistic regression with SHAP-based machine learning interpretability represents a methodological advancement in COVID-19 research. The SHAP analysis provides the first quantitative evidence of the impact of B-cell depletion threshold (< 150 cells/μL) on persistent infection risk. When B cells fall below this critical level, infection risk escalates exponentially (SHAP >0.5), consistent with clinically observed poor monoclonal antibody responses [59, 60]. This supports the use of dynamic immune monitoring for malignancy patients and initiating preemptive antiviral prophylaxis when B-cell counts drop below 100 cells/μL. Unlike conventional approaches, this framework not only achieves high accuracy (AUC =0.847) but also provides clinically actionable explanations through feature importance analysis (Fig. 4), particularly in quantifying the non-linear effects of B-cell depletion on infection persistence. The whole-genome sequencing of 188 persistent infection cases identified Omicron BA.5 as the predominant variant (89.4%, 168/188), with Delta variants accounting for 10.6% (20/188). Multivariable logistic regression adjusted for age, vaccination status (≥ 2 doses vs. <2 doses), and comorbidities (hypertension, diabetes) revealed no significant association between SARS-CoV-2 variants (BA.5 vs. Delta) and persistent infection risk (adjusted OR =1.12, 95% CI: 0.94–1.35; $p$= 0.12). This suggests that host factors (e.g., immune dysfunction) may play a more critical role than viral evolution in driving persistent infection within the studied population [61–63]. However, continuous monitoring of emerging variants (e.g., JN.1) remains warranted, given their potential for altered pathogenicity and immune evasion [64].

Importantly, the model exhibited excellent calibration in the non-significant Hosmer–Lemeshow test ($\chi^2$= 6.3, $p$= 0.62) and a Brier score of 0.13 (95% CI: 0.10–0.16), indicating high concordance between predicted probabilities and observed outcomes. The calibration slope of 0.94 (95% CI: 0.89–0.99) further confirmed minimal overfitting, likely attributable to the combined use of L2 regularization and bootstrap internal validation [65]. These results address a key limitation of prior models that focused predominantly on demographic predictors [66], often neglecting rigorous calibration assessment. Finally, the predictive model constructed in this study demonstrated a good predictive performance on the test set. The model accurately predicted whether patients would develop a persistent SARS-CoV-2 infection, providing strong support for clinical treatment, epidemic prevention, and control.

However, compared with recent research, this study has certain limitations, which include the relatively small sample size. Although this study included 1216 patients with SARS-CoV-2 infection, the sample size was limited compared with the global number of SARS-CoV-2 infection cases. This may have affected the universality of the results. Furthermore, the Delta variant was associated with prolonged viral shedding (median 18 days) compared with Omicron BA.5 (median 14 days), consistent with its higher replication efficiency in respiratory epithelium [67]. Our model demonstrated robust performance in both internal (AUC =0.85) and external validation cohorts (AUC =0.81; GSE158055), with minimal degradation in sensitivity (75.6% vs. 72.4%) and specificity (88.2% vs. 84.7%) (Table 4). The slightly lower AUC in the external cohort may reflect population heterogeneity: the GSE158055 dataset included a higher proportion of Omicron BA.2 subvariant cases (62% vs. 48% in our cohort) and younger patients (median age 45 vs. 58 years), both known to influence viral shedding dynamics [68, 69]. Despite these differences, the model maintained clinically acceptable discrimination, suggesting its utility across diverse settings. Logistic regression identified age and vaccination status as stable predictors (adjusted OR =1.12 and 0.67, respectively), aligning with prior studies [70]. Meanwhile, machine learning captured non-linear interactions (e.g., age ×comorbidity index), as evidenced by SHAP value analysis. Such interactions may explain the model's adaptability to external populations with varying risk factor distributions.

The data in this study were all obtained from a single hospital, which may not fully represent the clinical characteristics and risk factors of SARS-CoV-2 infection in different geographic regions and populations and thus limiting its generalizability. As the study spanned from January 2021 to October 2024, early cases were dominated by pre-Omicron variants (Delta: 58%), while later cases included Omicron subvariants (BA.5/XBB: 42%) and this temporal shift may influence risk factor generalizability. Moreover, the 14-day persistence threshold requires harmonization with WHO standards (≥ 20 days). Finally, viral variation was not considered. As the SARS-CoV-2 continues to mutate, its pathogenicity and infectivity may change. This study did not investigate the relationship between viral variation and persistent infections, which may have led to an incomplete assessment of the risk factors for persistent infections. In the future, we plan to expand the sample size by including patients from the global population to improve the universality and plans for further validation through the WHO Global Clinical Platform are ongoing. With the

continuous mutation of SARS-CoV-2, future studies should delve into the impact of viral variation on the risk of persistent infections, providing a scientific basis for epidemic prevention, control, and optimization of treatment strategies. Given the limitations inherent to a retrospective, future prospective studies should be conducted to more accurately assess the clinical characteristics and risk factors of patients with SARS-CoV-2 infection and to explore more effective prevention and treatment strategies.

## Conclusions

This retrospective analysis of 1,216 patients with COVID-19 (2021–2024) identified persistent SARS-CoV-2 infection (≥ 14-day positivity) in 15.5% (188) cases, with hypertension, diabetes, and malignancy as key clinical predictors. Patients with persistent SARS-CoV-2 infection exhibited immune dysregulation, alongside elevated inflammation indicated by CRP values. Full vaccination reduced persistent infection risk by 45% (OR = 0.55). The predictive model (AUC = 0.847) demonstrated utility for stratifying high-risk groups (e.g., malignancy with B-cell count < 100 cells/μL), supporting extension of antiviral regimens, though validation through multicenter studies remains essential. These findings underscore the need for adaptive prevention strategies in managing prolonged SARS-CoV-2 infection.

### Abbreviations

| | |
|---|---|
| ACE2 | Angiotensin-converting enzyme 2 |
| APACHE | Acute Physiology and Chronic Health Evaluation |
| AUC | Area under the curve |
| BALF | Bronchoalveolar lavage fluid |
| COPD | Chronic obstructive pulmonary disease |
| CRP | C-reactive protein |
| CT | Computed tomography |
| IL-6 | Interleukin 6 |
| OR | Odds ratio |
| ROC | Receiver operating characteristic |
| WBC | White blood cell |
| SVM | Support vector machine |
| KNN | K-nearest neighbors |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12879-025-11083-2.

Supplementary Material 1

Supplementary Material 2

### Acknowledgements

We would like to express our deepest gratitude to all patients who participated in this study, as their contributions were invaluable for advancing our understanding of persistent SARS-CoV-2 infection. We also acknowledge the dedicated healthcare workers and staff at our hospital for their tireless efforts in patient care and data collection. Special thanks are extended to the funding agencies that supported this research, enabling us to conduct a comprehensive analysis of the clinical characteristics and risk factors associated with persistent SARS-CoV-2 infection. We are grateful to our colleagues in the infectious diseases, clinical laboratory, and radiology departments for their expertise and assistance in diagnosing and managing patients with SARS-CoV-2 infection. Their insights and support were crucial to the successful completion of this study. Furthermore, we thank the statistical experts who provided valuable guidance on data analysis and interpretation. Their expertise ensured the accuracy and reliability of our findings. Lastly, we acknowledge the support and encouragement from our families and friends during this challenging time. Their understanding and patience allowed us to focus on our research and make meaningful contributions to the field of SARS-CoV-2 infection research.

### Clinical trial number

Not applicable.

### Authors' contributions

JZ: Funding acquisition, conceptualization, writing–original draft, writing–review & editing. WZ: Data curation, investigation, writing–review & editing. PJ: Formal analysis, validation, visualization. FM: Methodology, resources, writing–review & editing. YL: Project administration, supervision, writing–review & editing. YC: Software, writing–original draft. JL: Data curation, writing–original draft. ZZ: Conceptualization, supervision, writing–review & editing. XZ: Investigation, validation, writing–review & editing. JC: Formal analysis, visualization, writing–original draft. WZ: Methodology, resources, writing–review & editing. All authors read and approved the final manuscript.

### Funding

This work was supported by the Beijing Association for Science and Technology Jinqiao Project for "Analysis of Clinical Characteristics and Risk Factors for Persistent SARS-CoV-2 Infection" (directed by Dr. Jia Zhang) [grant number SK20240040].

### Data availability

The datasets generated during this study are available under a tiered access framework governed by ethical approvals and data protection regulations. Curated DNA/RNA sequencing data have been deposited in the China National Center for Bioinformation (CNCB) BioProject repository (https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA039561) with full public access through the BioProject interface. De-identified clinical datasets are accessible under controlled conditions explicitly authorized by the Ethics Committee of the Aerospace Center Hospital (Approval No. Jinghang Medical Lunshen 2024–090), requiring compliance with Article 30 of China's Personal Information Protection Law (PIPL) and secure transfer via encrypted protocols. All data will be retained for 6 months post-publication, with extended access contingent upon revalidation by ethics committees and implementation of audited data destruction protocols.

## Declarations

### Ethics approval and consent to participate

The study was conducted in accordance with the principles outlined in the Declaration of Helsinki. The study protocol outlined in our manuscript has been thoroughly reviewed and approved by the Ethics Committee of Aerospace Center Hospital, with the approval number being Jinghang Yilun Shen 2024 No. 090. Notably, in accordance with the committee's guidelines and the nature of our study, the requirement for informed consent was formally waived by the Ethics Committee of Aerospace Center Hospital for all participants involved. Additionally, we wish to clarify that this is a retrospective study; therefore, this declaration is "not applicable." We kindly request that our manuscript be considered in accordance with these statements.

### Consent for publication

Not applicable.

### Competing interests

The authors declare no competing interests.

### Author details

[1]Department of Respiratory and Critical Care Medicine, Aerospace Center Hospital, Beijing 100049, China.

Received: 10 January 2025 Accepted: 5 May 2025
Published online: 14 May 2025

## References


1. Dewidar O, Bondok M, Abdelrazeq L, Aliyeva K, Solo K, Welch V, et al. Equity issues rarely addressed in the development of COVID-19 formal recommendations and good practice statements: a cross-sectional study. J Clin Epidemiol. 2023;161:116–26. https://doi.org/10.1016/j.jclinepi.2023.08.002.
2. Machkovech HM, Hahn AM, Garonzik Wang J, Grubaugh ND, Halfmann PJ, Johnson MC, et al. Persistent SARS-CoV-2 infection: significance and implications. Lancet Infect Dis. 2024;24:e453–62. https://doi.org/10.1016/S1473-3099(23)00815-0.
3. Gao Y, Dong Y, Bu Q, Gong Z, Wang W, Zhou Z, et al. Antiviral effectiveness, clinical outcomes, and artificial intelligence imaging analysis for hospitalized COVID-19 patients receiving antivirals. Influenza Other Respir Viruses. 2024;18: e70006. https://doi.org/10.1111/irv.70006.
4. Harari S, Tahor M, Rutsinsky N, Meijer S, Miller D, Henig O, et al. Drivers of adaptive evolution during chronic SARS-CoV-2 infections. Nat Med. 2022;28:1501–8. https://doi.org/10.1038/s41591-022-01882-4.
5. Chaguza C, Hahn AM, Petrone ME, Zhou S, Ferguson D, Breban MI, et al. Accelerated SARS-CoV-2 intrahost evolution leading to distinct genotypes during chronic infection. Cell Rep Med. 2023;2: 100943. https://doi.org/10.1016/j.xcrm.2023.100943.
6. Wilkinson SAJ, Richter A, Casey A, Osman H, Mirza JD, Stockton J, et al. Recurrent SARS-CoV-2 mutations in immunodeficient patients. Virus Evol. 2022;8:veac050. https://doi.org/10.1093/ve/veac050.
7. Cevik M, Tate M, Lloyd O, Maraolo AE, Schafers J, Ho A. SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and meta-analysis. Lancet Microbe. 2021;2:e13-22. https://doi.org/10.1016/S2666-5247(20)30172-5.
8. Li Y, Choudhary MC, Regan J, Boucau J, Nathan A, Speidel T, et al. SARS-CoV-2 viral clearance and evolution varies by extent of immunodeficiency. medRxiv. 2023:2023.07.31.23293441. https://doi.org/10.1101/2023.07.31.23293441 (preprint).
9. Gonzalez-Reiche AS, Alshammary H, Schaefer S, Patel G, Polanco J, Carreño JM, et al. Sequential intrahost evolution and onward transmission of SARS-CoV-2 variants. Nat Commun. 2023;14:3235. https://doi.org/10.1038/s41467-023-38867-x.
10. Scherer EM, Babiker A, Adelman MW, Allman B, Key A, Kleinhenz JM, et al. SARS-CoV-2 evolution and immune escape in immunocompromised patients. N Engl J Med. 2022;386:2436–8. https://doi.org/10.1056/NEJMc2202861.
11. Halfmann PJ, Minor NR, Haddock LA 3rd, Maddox R, Moreno GK, Braun KM, et al. Evolution of a globally unique SARS-CoV-2 Spike E484T monoclonal antibody escape mutation in a persistently infected, immunocompromised individual. Virus Evol. 2023;9:veac104. https://doi.org/10.1093/ve/veac104.
12. Zymovets V, Rakhimova O, Wadelius P, Schmidt A, Brundin M, Kelk P, et al. Exploring the impact of oral bacteria remnants on stem cells from the Apical papilla: mineralization potential and inflammatory response. Front Cell Infect Microbiol. 2023;13:1257433. https://doi.org/10.3389/fcimb.2023.1257433.
13. Lee CY, Shah MK, Hoyos D, Solovyov A, Douglas M, Taur Y, et al. Prolonged SARS-CoV-2 infection in patients with lymphoid malignancies. Cancer Discov. 2022;12:62–73. https://doi.org/10.1158/2159-8290.CD-21-1033.
14. Corey L, Beyrer C, Cohen MS, Michael NL, Bedford T, Rolland M. SARS-CoV-2 variants in patients with immunosuppression. N Engl J Med. 2021;385:562–6. https://doi.org/10.1056/NEJMsb2104756.
15. Connolly CM, Paik JJ. SARS-CoV-2 vaccination in the immunocompromised host. J Allergy Clin Immunol. 2022;150:56–8. https://doi.org/10.1016/j.jaci.2022.05.001.
16. He F, Page JH, Weinberg KR, Mishra A. The development and validation of simplified machine learning algorithms to predict prognosis of hospitalized patients with COVID-19: multicenter, retrospective study. J Med Internet Res. 2022;24: e31549. https://doi.org/10.2196/31549.
17. Mohiuddin Chowdhury ATM, Karim MR, Ali MA, Islam J, Li Y, He S. Clinical characteristics and the long-term post-recovery manifestations of the COVID-19 patients-A prospective multicenter cross-sectional study. Front Med (Lausanne). 2021;8: 663670. https://doi.org/10.3389/fmed.2021.663670.
18. Liu XQ, Xue S, Xu JB, Ge H, Mao Q, Xu XH, et al. Clinical characteristics and related risk factors of disease severity in 101 COVID-19 patients hospitalized in Wuhan. China Acta Pharmacol Sin. 2022;43:64–75. https://doi.org/10.1038/s41401-021-00627-2.
19. Acedera ML, Sirichokchatchawan W, Brimson S, Prasansuklab A. Age, comorbidities, C-reactive protein and procalcitonin as predictors of severity in confirmed COVID-19 patients in the Philippines. Heliyon. 2023;9: e15233. https://doi.org/10.1016/j.heliyon.2023.e15233.
20. Garg S, Kim L, Whitaker M, O'Halloran A, Cummings C, Holstein R, et al. Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019 - COVID-NET, 14 States, March 1–30, 2020. MMWR Morb Mortal Wkly Rep. 2020;69:458–64. https://doi.org/10.15585/mmwr.mm6915e3.
21. National Health Commission of the People's Republic of China. Diagnosis and treatment protocol for COVID-19 (Trial 10th Edition) [in Chinese]. Chin J Clin Infect Dis. 2023;16:1–9. https://doi.org/10.3760/cma.j.issn.1674-2397.2023.01.001.
22. Soriano JB, Murthy S, Marshall JC, Relan P, Diaz JV, WHO Clinical Case Definition Working Group on Post-COVID-19 Condition. A clinical case definition of post-COVID-19 condition by a Delphi consensus. Lancet Infect Dis. 2022;22:e102-7. https://doi.org/10.1016/S1473-3099(21)00703-9.
23. Costantini E, Lang KM, Sijtsma K, Reeskens T. Solving the many-variables problem in MICE with principal component regression. Behav Res Methods. 2024;56:1715–37. https://doi.org/10.3758/s13428-023-02117-1.
24. Núñez E, Steyerberg EW, Núñez J. Estrategias para la elaboración de modelos estadísticos de regresión [Regression modeling strategies]. Rev Esp Cardiol. 2011;64:501–7. https://doi.org/10.1016/j.recesp.2011.01.019. Spanish.
25. World Health Organization. Clinical management of COVID-19: living guideline; 2023. https://www.who.int/publications/i/item/WHO-2019-nCoV-clinical-2023.2.
26. Zhao A, Liu Y, Xia J, Huang L, Lu Q, Tang Q, et al. Establishment and validation of a prognostic model based on common laboratory indicators for SARS-CoV-2 infection in Chinese population. Ann Med. 2024;56:2400312. https://doi.org/10.1080/07853890.2024.2400312.
27. Tang J, Zeng C, Cox TM, Li C, Son YM, Cheon IS, et al. Respiratory mucosal immunity against SARS-CoV-2 after mRNA vaccination. Sci Immunol. 202228;7:eadd4853. https://doi.org/10.1126/sciimmunol.add4853.
28. Hettle D, Hutchings S, Muir P, Moran E; COVID-19 Genomics UK (COG-UK) consortium. Persistent SARS-CoV-2 infection in immunocompromised patients facilitates rapid viral evolution: Retrospective cohort study and literature review. Clin Infect Pract. 2022;16:100210. https://doi.org/10.1016/j.clinpr.2022.100210.
29. Crescioli E, Nielsen FM, Bunzel AM, Eriksen ASB, Siegemund M, Poulsen LM, et al. Long-term mortality and health-related quality of life with lower versus higher oxygenation targets in intensive care unit patients with COVID-19 and severe hypoxaemia. Intensive Care Med. 2024;50:1603–13. https://doi.org/10.1007/s00134-024-07613-2.
30. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using Open SAFELY. Nature. 2020;584:430–6. https://doi.org/10.1038/s41586-020-2521-4.
31. Speranza E, Purushotham JN, Port JR, Schwarz B, Flagg M, Williamson BN, et al. Age-related differences in immune dynamics during SARS-CoV-2 infection in rhesus macaques. Life Sci Alliance. 2022;5:e202101314. https://doi.org/10.26508/lsa.202101314.
32. Drancourt M, Cortaredona S, Melenotte C, Amrane S, Eldin C, La Scola B, et al. SARS-CoV-2 persistent viral shedding in the context of hydroxychloroquine-azithromycin treatment. Viruses. 2021;13:890. https://doi.org/10.3390/v13050890.
33. Hoseinnezhad T, Soltani N, Ziarati S, Behboudi E, Mousavi MJ. The role of HLA genetic variants in COVID-19 susceptibility, severity, and mortality: A global review. J Clin Lab Anal. 2024;38: e25005. https://doi.org/10.1002/jcla.25005.
34. Behboudi E, Nooreddin Faraji S, Daryabor G. Mohammad Ali Hashemi S, Asadi M, Edalat F, et al, SARS-CoV-2 mechanisms of cell tropism in various organs considering host factors. Heliyon. 2024;10:e26577. https://doi.org/10.1016/j.heliyon.2024.e26577.
35. Ayatollahi AA, Aghcheli B, Amini A, Nikbakht H, Ghassemzadehpirsala P, Behboudi E, et al. Association between blood groups and COVID-19 outcome in Iranian patients. Future Virol. 2021:https://doi.org/10.2217/fvl-2021-0090. https://doi.org/10.2217/fvl-2021-0090.
36. Kim SJ, Kim T, Choi H, Shin TR, Kim HI, Jang SH, et al. Respiratory pathogen and clinical features of hospitalized patients in acute exacerbation of chronic obstructive pulmonary disease after COVID 19 pandemic. Sci Rep. 2024;7(14):10462. https://doi.org/10.1038/s41598-024-61360-4.

37. Xiong R, Zhao Z, Lu H, Ma Y, Zeng H, Chen Y. Asthma patients benefit more than chronic obstructive pulmonary disease patients in the coronavirus disease 2019 pandemic. Front Med (Lausanne). 2021;8: 709006. https://doi.org/10.3389/fmed.2021.709006.
38. Kurai D, Saraya T, Ishii H, Takizawa H. Virus-induced exacerbations in asthma and COPD. Front Microbiol. 2013;4:293. https://doi.org/10.3389/fmicb.2013.00293.
39. Zhang JJ, Dong X, Cao YY, Yuan YD, Yang YB, Yan YQ, et al. Clinical characteristics of 140 patients infected with SARS-CoV-2 in Wuhan. China In: Allergy. 2020;75:1730–41. https://doi.org/10.1111/all.14238.
40. Kuba K, Imai Y, Rao S, Gao H, Guo F, Guan B, et al. A crucial role of angiotensin converting enzyme 2 (ACE2) in SARS coronavirus-induced lung injury. Nat Med. 2005;11:875–9. https://doi.org/10.1038/nm1267.
41. Kuba K, Imai Y, Rao S, Jiang C, Penninger JM. Lessons from SARS: control of acute lung failure by the SARS receptor ACE2. J Mol Med (Berl). 2006;84:814–20. https://doi.org/10.1007/s00109-006-0094-9.
42. Wu SC, Arthur CM, Jan HM, Garcia-Beltran WF, Patel KR, Rathgeber MF, et al. Blood group A enhances SARS-CoV-2 infection. Blood. 2023;142:742–7. https://doi.org/10.1182/blood.2022018903.
43. Deravi N, Fathi M, Vakili K, Yaghoobpoor S, Pirzadeh M, Mokhtari M, et al. SARS-CoV-2 infection in patients with diabetes mellitus and hypertension: a systematic review. Rev Cardiovasc Med. 2020;21:385–97. https://doi.org/10.31083/j.rcm.2020.03.78.
44. Wünsch K, Anastasiou OE, Alt M, Brochhagen L, Cherneha M, Thümmler L, et al. COVID-19 in elderly, immunocompromised or diabetic patients-from immune monitoring to clinical management in the hospital. Viruses. 2022;14:746. https://doi.org/10.3390/v14040746.
45. Prendecki M, Thomson T, Clarke CL, Martin P, Gleeson S, De Aguiar RC, et al. Immunological responses to SARS-CoV-2 vaccines in kidney transplant recipients. Lancet. 2021;398:1482–4. https://doi.org/10.1016/S0140-6736(21)02096-1.
46. Danziger-Isakov L, Blumberg EA, Manuel O, Sester M. Impact of COVID-19 in solid organ transplant recipients. Am J Transplant. 2021;21:925–37. https://doi.org/10.1111/ajt.16449.
47. Lee JC, Mehdizadeh S, Smith J, Young A, Mufazalov IA, Mowery CT, et al. Regulatory T cell control of systemic immunity and immunotherapy response in liver metastasis. Sci Immunol. 2020;5:eaba0759. https://doi.org/10.1126/sciimmunol.aba0759.
48. Emmanouilidou-Fotoulaki E, Karava V, Dotis J, Kondou A, Printza N. Immunologic response to SARS-CoV-2 vaccination in pediatric kidney transplant recipients: a systematic review and meta-analysis. Vaccines (Basel). 2023;11:1080. https://doi.org/10.1016/10.3390/vaccines11061080.
49. Tarke A, Coelho CH, Zhang Z, Dan JM, Yu ED, Methot N, et al. SARS-CoV-2 vaccination induces immunological T cell memory able to cross-recognize variants from Alpha to Omicron. Cell. 2022;185:847-59.e11. https://doi.org/10.1016/j.cell.2022.01.015.
50. Boshier FAT, Pang J, Penner J, Parker M, Alders N, Bamford A, et al. Evolution of viral variants in remdesivir-treated and untreated SARS-CoV-2-infected pediatrics patients. J Med Virol. 2022;94:161–72. https://doi.org/10.1002/jmv.27285.
51. Liu Z, Liang Q, Ren Y, Guo C, Ge X, Wang L, et al. Immunosenescence: molecular mechanisms and diseases. Signal Transduct Target Ther. 2023;8:200. https://doi.org/10.1038/s41392-023-01451-2.
52. Kazer SW, Aicher TP, Muema DM, Carroll SL, Ordovas-Montanes J, Miao VN, et al. Integrated single-cell analysis of multicellular immune dynamics during hyperacute HIV-1 infection. Nat Med. 2020;26:511–8. https://doi.org/10.1038/s41591-020-0799-2.
53. Laracy JC, Kamboj M, Vardhana SA. Long and persistent COVID-19 in patients with hematologic malignancies: from bench to bedside. Curr Opin Infect Dis. 2022;35:271–9. https://doi.org/10.1097/QCO.0000000000000841.
54. Chan M, Linn MMN, O'Hagan T, Guerra-Assunção JA, Lackenby A, Workman S, et al. Persistent SARS-CoV-2 PCR positivity despite anti-viral treatment in immunodeficient patients. J Clin Immunol. 2023;43:1083–92. https://doi.org/10.1007/s10875-023-01504-9.
55. Bange EM, Han NA, Wileyto P, Kim JY, Gouma S, Robinson J, et al. $CD8^+$ T cells contribute to survival in patients with COVID-19 and hematologic cancer. Nat Med. 2021;27:1280–9. https://doi.org/10.1038/s41591-021-01386-7.
56. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 2020;395:565–74. https://doi.org/10.1016/S0140-6736(20)30251-8.
57. National Clinical Laboratory Center for Quality Control in Health Industry. Discussion on "false negative" results of nucleic acid testing for SARS-CoV-2; 2020. https://www.sohu.com/a/373315225_100202861. Accessed 13 Mar 2023.
58. Peng Y, Falin S, Guijie W. Two cases of improved positive results of nucleic acid testing for SARS-CoV-2 through atomization-induced expectoration. Chin J Tuberc Respir Dis. 2020;43:E018. https://doi.org/10.3760/cma.j.issn.1001-0939.202.
59. Nussenblatt V, Roder AE, Das S, de Wit E, Youn JH, Banakis S, et al. Yearlong COVID-19 infection reveals within-host evolution of SARS-CoV-2 in a patient with B-Cell depletion. J Infect Dis. 2022;225:1118–23. https://doi.org/10.1093/infdis/jiab622.
60. Ciotti JR, Valtcheva MV, Cross AH. Effects of MS disease-modifying therapies on responses to vaccinations: a review. Mult Scler Relat Disord. 2020;45: 102439. https://doi.org/10.1016/j.msard.2020.102439.
61. Privratsky JR, Ide S, Chen Y, Kitai H, Ren J, Fradin H, et al. A macrophage-endothelial immunoregulatory axis ameliorates septic acute kidney injury. Kidney Int. 2023;103:514–28. https://doi.org/10.1016/j.kint.2022.10.008.
62. Brodin P. Immune determinants of COVID-19 disease presentation and severity. Nat Med. 2021;27:28–33. https://doi.org/10.1038/s41591-020-01202-8.
63. Kompaniyets L, Pennington AF, Goodman AB, Rosenblum HG, Belay B, Ko JY, et al. Underlying medical conditions and severe illness among 540,667 adults hospitalized with COVID-19, March 2020-March 2021. Prev Chronic Dis. 2021;18:E66. https://doi.org/10.5888/pcd18.210123.
64. Faraone JN, Qu P, Goodarzi N, Zheng YM, Carlin C, Saif LJ, et al. Immune evasion and membrane fusion of SARS-CoV-2 XBB subvariants EG.5.1 and XBB.2.3. Emerg Microbes Infect. 2023;12:2270069. https://doi.org/10.1080/22221751.2023.2270069.
65. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of Covid-19: systematic review and critical appraisal. BMJ. 2020;369: m1328. https://doi.org/10.1136/bmj.m1328.
66. Gao YD, Ding M, Dong X, Zhang JJ, Kursat Azkur A, Azkur D, et al. Risk factors for severe and critically ill COVID-19 patients: a review. Allergy. 2021;76:428–55. https://doi.org/10.1111/all.14657.
67. Nakagama Y, Candray K, Kaku N, Komase Y, Rodriguez-Funes MV, Dominguez R, et al. Antibody avidity maturation following recovery from infection or the booster vaccination grants breadth of SARS-CoV-2 neutralizing capacity. J Infect Dis. 2023;227:780–7. https://doi.org/10.1093/infdis/jiac492.
68. Tian D, Nie W, Sun Y, Ye Q. The epidemiological features of the SARS-CoV-2 omicron subvariant BA.5 and its evasion of the neutralizing activity of vaccination and prior infection. Vaccines (Basel). 2022;10:1699. https://doi.org/10.3390/vaccines10101699.
69. Phan HV, Tsitsiklis A, Maguire CP, Haddad EK, Becker PM, Kim-Schulze S, et al. Host-microbe multiomic profiling reveals age-dependent immune dysregulation associated with COVID-19 immunopathology. Sci Transl Med. 2024;16:eadj5154. https://doi.org/10.1126/scitranslmed.adj5154.
70. Dagan N, Barda N, Kepten E, Miron O, Perchik S, Katz MA, et al. BNT162b2 mRNA Covid-19 vaccine in a nationwide mass vaccination setting. N Engl J Med. 2021;384:1412–23. https://doi.org/10.1056/NEJMoa2101765.


## Publisher's Note