Supplemental Material

Chronic lung lesions in COVID-19 survivors: predictive clinical model

Carlos R R Carvalho, Rodrigo C Chate, Marcio VY Sawamura, Michelle L Garcia, Celina A Lamas, Diego AC Cardenas, Daniel M Lima, Paula G Scudeller, João M Salge, Cesar H Nomura, Marco A Gutierrez, HCFMUSP Covid-19 Study Group.

Contents

Supplemental Methods

Datasets

The SIIM-RSNA dataset contains 6,334 posterior-anterior radiographic images from 6,054 patients obtained from the public dataset Machine Learning Challenge on COVID-19 Pneumonia Detection and Localization.[1] Specialists classified images as “negative for pneumonia” or “COVID-19 pneumonia”. A total of 6,030 images were selected and randomly distributed in training and validation sets (1,276 negative and 3,711 positive findings) and a test set (400 negative and 643 positive findings).

The Institute of Radiology (InRad) dataset contains chest X-Ray (CXR) and chest computed tomographic (CT) images of 257 patients. The CXR images were classified as normal (n=145) or with findings related to COVID-19 (n=112) and randomly distributed in training and validation sets (214 patients) and a test set (n=43). Images were obtained from the InRad of the Hospital das Clínicas, Faculdade de Medicina, Universidade de São Paulo (HCFMUSP).

Because of differences in dataset sizes, a data augmentation technique was adopted using random transformations, including rotation (0–15 degrees), horizontal mirroring, and random changes in intensity and contrast (0–5%).

Classification of chest radiography images

A deep-learning (DL) approach using a convolutional neural network (CNN) based on an EfficientNetB7 architecture was used.[2] The network classification layer was replaced by a global average pooling operation, followed by batch normalization and the adoption of a dense layer with one neuron and sigmoid activation function. Each training iteration was run for 40 epochs with an Adam optimizer at a learning rate of 0.0001. All images were resized to 600 x 600 pixels.

The CNN was trained using the SIIM-RSNA dataset to detect radiographic patterns of COVID-19 pneumonia. Training was initiated in EfficientNetB7 using weights after pre-training with the ImageNet dataset.[3]

A five-fold cross-validation strategy was adopted for the training and validation sets. The training weights obtained for each fold were used with the

test set of the SIIM-SNA to evaluate classification accuracy (Table 1). The fold with the best area under the receiver operating characteristic curve (AUC), in this case, fold 1 with AUC of 0.89, defines the final weights of the CNN.

Table 1. Classification of the test set of the SIIM-RSNA dataset as negative (normal) or positive (patterns of COVID-19 pneumonia).

Dataset	5-fold	Acc	Prec	Sensitivity	Specificity	F1-score	AUC
SIIM-RSNA	0	0.80	0.85	0.82	0.76	0.83	0.88
	1	0.80	0.85	0.82	0.77	0.84	**0.89**
	2	0.78	0.77	0.92	0.56	0.84	0.87
	3	0.76	0.74	0.93	0.48	0.83	0.86
	4	0.76	0.74	0.93	0.48	0.83	0.86

Area under the receiver operating characteristic curve (AUC); Accuracy (Acc); Precision (Prec).

For the InRad dataset, the CNN was initialized with the final weights defined in the training set of SIIM-RSNA. After initialization, the CNN was retrained to classify images as normal or with findings related to COVID-19.

The InRad dataset was divided into six-folds during the retraining, five folds for training and validation, and one-fold for test. To avoid bias, the test fold was selected to run all six folds available and, for each test fold selected, a five-fold cross-validation strategy was applied in the remaining training and validation folds (Table 2).

Table 2. Classification using six test folds of the InRad database.

Dataset	Test fold	Acc	Prec	Sensitivity	Specificity	F1-score	AUC
InRad	0	0.79±0.01	0.74±0.04	0.82±0.07	0.77±0.06	0.78±0.02	0.86±0.02
	1	0.69±0.02	0.62±0.03	0.84±0.06	0.57±0.07	0.71±0.02	0.75±0.01
	2	0.67±0.05	0.60±0.06	0.81±0.08	0.57±0.13	0.68±0.02	0.76±0.02
	3	0.77±0.04	0.71±0.07	0.80±0.04	0.74±0.10	0.75±0.03	0.80±0.02
	4	0.82±0.05	0.77±0.11	0.89±0.10	0.78±0.14	0.81±0.03	0.89±0.04
	5	0.71±0.04	0.62±0.04	0.90±0.02	0.58±0.08	0.73±0.03	0.80±0.02

Data represent the mean and standard deviation after five-fold cross validation. Area under the receiver operating characteristic curve (AUC); Accuracy (Acc); Precision (Prec).

Detection of chronic lung lesions on computed tomography images

Three machine learning models were developed based on the clinical data, including the modified Medical Research Council dyspnea scale (mMRC), oximetry (SpO_2) and spirometry (forced vital capacity, FVC), and five radiographic

probabilities (p_{CXR0} top_{CXR4}) with findings related to COVID-19 (p_{CXRn}=1) and normal (p_{CXRn}=0), which were obtained from the previous step (Table 2). As output, the models predict the value of a binary variable (p_{CT}) related to the presence of chronic lung lesions on CT images, with p_{CT}=1 for a CT score ≥ 7 (n=129) and p_{CT}=0 for a CT score < 7 (n=128) (Figure 1).

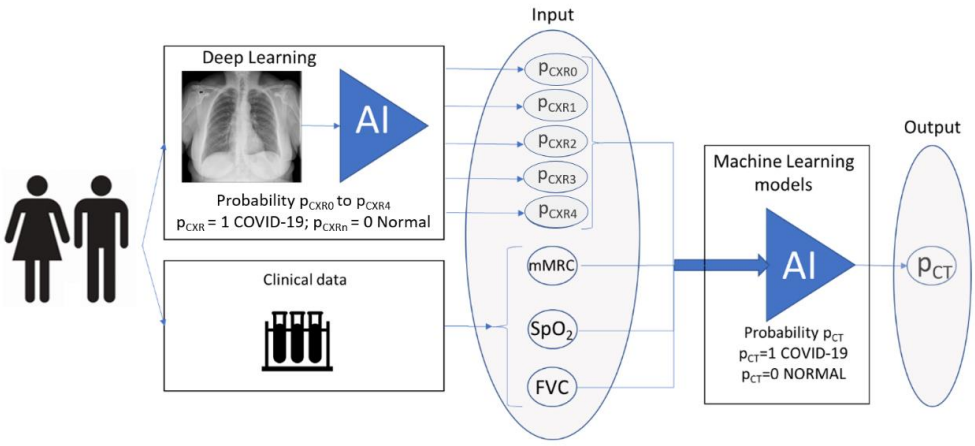


Figure 1. Machine learning-based model. Data on the modified Medical Research Council (mMRC) dyspnea scale, oximetry (SpO2), and spirometry (forced vital capacity [FVC]), and radiographic probabilities (p_{CXR0} to p_{CXR4}) with findings related to COVID-19 (p_{CXRn}=1) and normal (p_{CXRn}=0) were used as input variables, and the presence of lung lesions due to COVID-19 (p_{CT}) was used as output. AI, artificial intelligence. CT, computed tomography.

The first model was a logistic regression (LR) model with L2 regularization to prevent overfitting,[4] whereas the second model was a random forest model with 100 trees (RF-100), Gini criterion, minimum of two samples for splitting, minimum of one sample in leaves, and bootstrap.[4] The third model was a random forest model with parameters as described above, except for the limit of 10 trees and maximum depth h_max=6 (RF-10).[4] The performance of the machine-learning models was evaluated based on sensitivity, specificity, AUC, and F1-score.

Three combinations of input variables were evaluated: 1) clinical variables (mMRC, SpO_2, and FVC); 2) CXR; and 3) clinical variables (mMRC, SpO_2, FVC) and CXR.

For each model, a five-fold cross-validation strategy was adopted for the training and validation sets. The performance of the LR model was better when a combination of all variables (clinical variables and CXR) was used. The following metrics expressed in terms of mean ± standard deviation and 95% Confidence Interval (CI) were considered: sensitivity, 0.85±0.08 (95% CI [0.77, 0.94]); specificity, 0.70±0.14 (95% CI [0.55, 0.85]); F1-score, 0.79±0.06 (95% CI [0.73, 0.85]); and AUC, 0.80±0.07(95% CI [0.72, 0.87]) (Table 3).

Table 3. Predictive performance of three multivariate models using three datasets.

Groups of variables	Method	Sensitivity	Specificity	F1-score	AUC
1 SpO_2, mMRC score, and FVC	LR	0.87±0.16	0.42±0.33	0.71±0.03	0.68±0.10
	RF-10	0.88±0.15	0.37±0.32	0.71±0.03	0.66±0.08
	RF-100	0.82±0.12	0.44±0.13	0.69±0.08	0.62±0.12
2 CXR	LR	0.88±0.05	0.52±0.14	0.75±0.04	0.78±0.05
	RF-10	0.91±0.08	0.41±0.18	0.73±0.04	0.73±0.06
	RF-100	0.94±0.07	0.33±0.19	0.72±0.03	0.72±0.03
3 SpO_2, mMRC score, FVC and CRX	LR	**0.85±0.08**	**0.70±0.14**	**0.79±0.06**	**0.80±0.07**
	RF-10	0.85±0.09	0.61±0.22	0.76±0.04	0.76±0.08
	RF-100	0.89±0.06	0.49±0.17	0.75±0.04	0.76±0.07

Values are presented as the mean ± standard deviation after five-fold cross validation for each test fold. Area under the receiver operating characteristic curve (AUC); Accuracy (Acc); Chest X-Ray (CRX); Forced vital capacity (FVC); Logistic Regression (LR); modified Medical Research Council dyspnea scale (mMRC); Precision (Prec); Random forest (RF).

The LR model is represented by the following function:

$$p_{CT} = \sigma\left(\beta_1 FVC^* + \beta_2\, mMRC^* + \beta_3 S_p O_2 + \beta_4 p_{CXR0} + \beta_5 p_{CXR1} + \beta_6 p_{CXR2} + \beta_7 p_{CXR3} + \beta_8 p_{CXR4}\right)$$

$$\beta_1 = -0.3705 \;\; \beta_2 = -2.2807 \;\; \beta_3 = -0.745 \;\; \beta_4 = 1.1257$$

$$\beta_5 = 1.4960 \;\; \beta_6 = 1.0761 \;\; \beta_7 = 0.7328 \;\; \beta_8 = -0.7613$$

where p_{CT} is the probability of the presence of abnormalities on CT images, σ is the sigmoid function to restrict p_{CT} between 0 and 1, $FVC^* = \frac{FVC_{Resting}}{2FVC_{min}}$, $mMRC^* = \frac{mMRC}{4}$, and p_{CXR0} to p_{CXR4} are the probabilities that the CXR image has findings related to sequelae from COVID-19, obtained in each fold (0 to 4) during a 5-folds cross validation. Table 4 shows the estimates for the logistic regression function.

Table 4. Estimates of the logistic regression function.

Variable	Estimated regression coefficient (β)	Estimated Standard Error	*p*-value	95% CI for regression coefficient (β)		Estimated odds ratios
FVC^*	-0.3705	0.3210	0.248	-0.9990	0.2580	0.6904
$mMRC^*$	-2.2807	0.3020	<0.001	-2.8730	-1.6890	0.1022
S_pO_2	-0.7450	0.2320	0.001	-1.2010	-0.2890	0.4747
p_{CXR0}	1.1257	0.4150	0.007	0.3120	1.9400	3.0824
p_{CXR1}	1.4960	0.4160	<0.001	0.6810	2.3110	4.4638
p_{CXR2}	1.0761	0.3390	0.002	0.4120	1.7410	2.9332
p_{CXR3}	0.7328	0.3380	0.030	0.0710	1.3950	2.0809
p_{CXR4}	-0.7613	0.4580	0.096	-1.6590	0.1360	0.4671

Forced vital capacity (FVC); modified Medical Research Council dyspnea scale (mMRC); radiographic probabilities (Pcxr0 to Pcxr4).

Also, we included demographic and anthropometric variables on the logistic regression prediction model, performing experiments using six different combinations of variables (age, gender, body mass index [BMI], SpO2, mMRC score, FVC and CXR). The performance of each combination is reported in the Table 5. The model performance with the inclusion of demographic or anthropometric variables did not result in significant improvement. According to our experiments, the combination of SpO2, mMRC score, FVC and CXR presented the best performance.

Table 5. Performance of the predictive model using six combinations of variables (N=257).

Groups of variables	Sensitivity	Specificity	F1-score	AUC
1 **Age, gender, and BMI**	0.87±0.09	0.40±0.27	0.71±0.03	0.64±0.09
2 **SpO_2, mMRC score, and FVC**	0.87±0.16	0.42±0.33	0.71±0.03	0.68±0.10
3 **Age, Gender, BMI, SpO2, mMRC score, and FVC**	0.95±0.05	0.37±0.30	0.75±0.06	0.71±0.10
4 **CXR**	0.88±0.05	0.52±0.14	0.75±0.04	0.78±0.05
5 **Age, Gender, BMI, SpO_2, mMRC score, FVC, and CXR**	0.87±0.08	0.65±0.16	0.79±0.06	0.79±0.06
6 **SpO_2, mMRC score, FVC, and CXR**	0.85±0.08	0.70±0.14	0.79±0.06	0.80±0.07

Values are presented as the mean ± standard deviation after five-fold cross validation for each test fold. Area under the receiver operating characteristic curve (AUC); Body Mass Index (BMI); Chest X-Ray (CRX); Forced vital capacity (FVC); modified Medical Research Council dyspnoea scale (mMRC).

Dataset and normalization of clinical data

A total of 257 patients with data on the mMRC dyspnea scale, oximetry, spirometry, CRX, and chest CT were selected to predict pulmonary changes. Of the 257 patients, 128 had no significant CT changes (scores < 7). A CT score of 7 was used as the cutoff value by maximizing F1 scores and AUC (Figure 2).

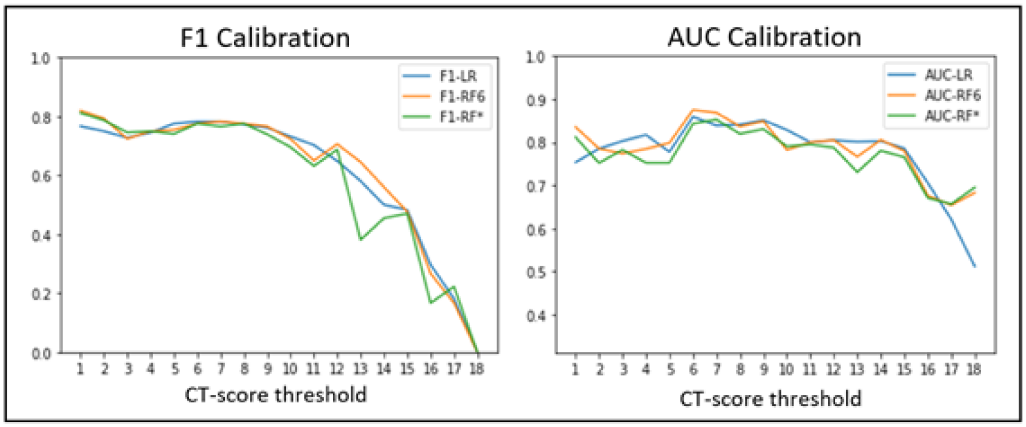
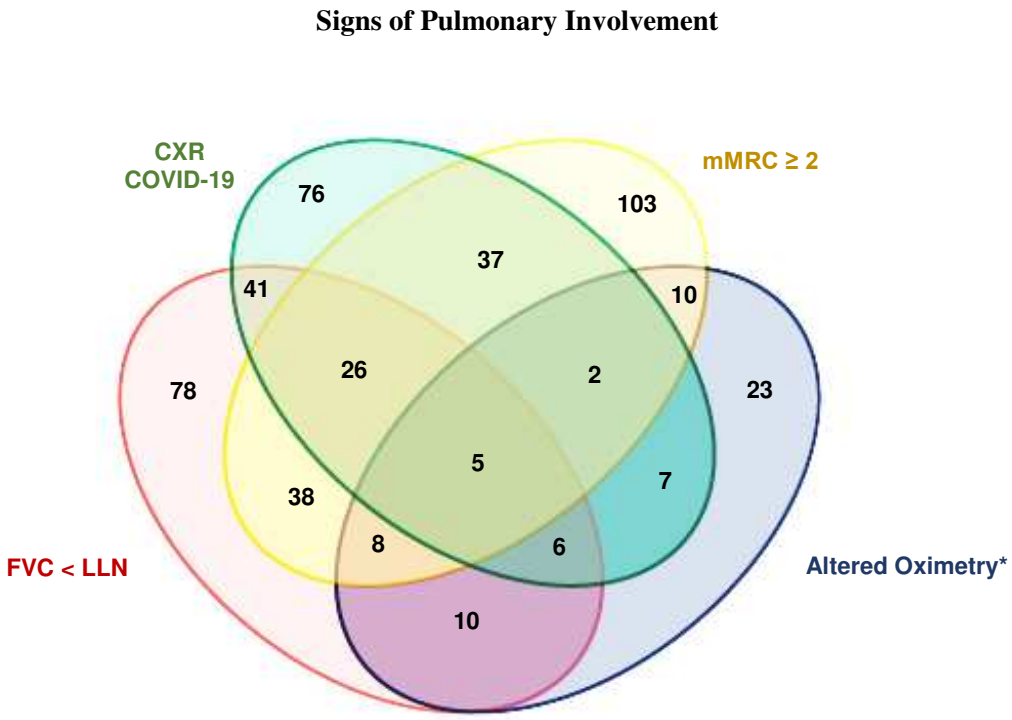
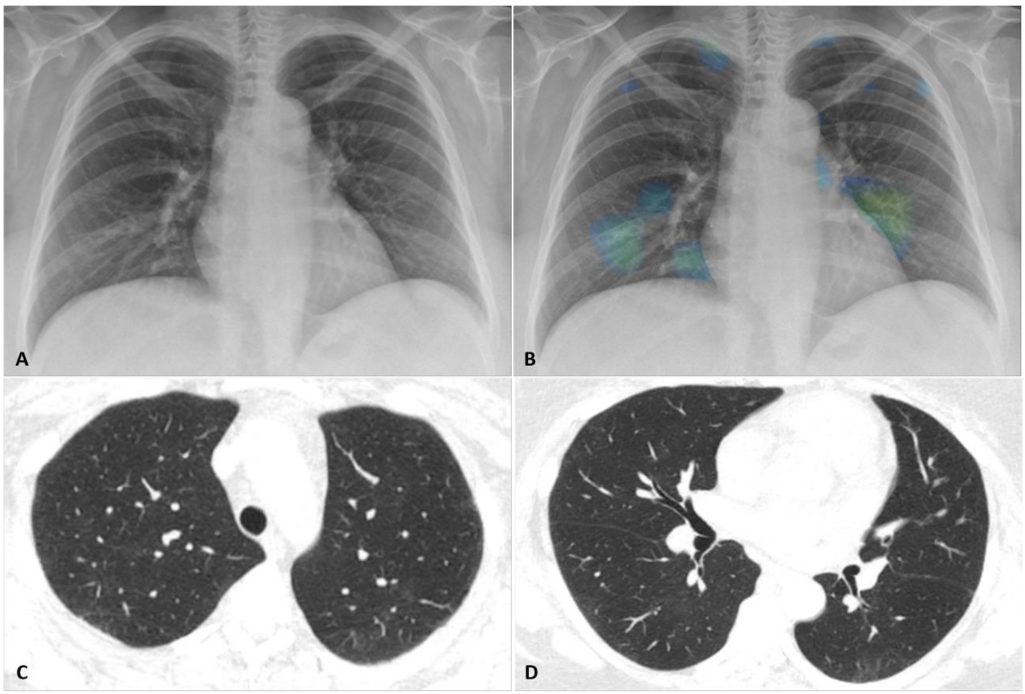


Figure 2. Computed tomography scores based on the F1-score and AUC values.

Clinical variables were normalized by dividing the mMRC values by 4 (resulting in values between 0 and 1) and the $FVC_{Resting}$ by twice the FVC_{min} (resulting in a minimum value of 0.257 and a maximum value of 0.847).



Supplemental Figure S1. Diagram showing the overlap in the changes of parameters used as pulmonary criteria to refer patients for thorax computed tomography. Values are expressed as the number of patients showing the correspondent alterations. CXR, chest X-Ray; FVC, forced vital capacity; LLN, lower limit of normal; mMRC, modified Medical Research Council dyspnea scale. *Resting SpO_2 ≤ 90% or a decrease in SpO_2 of ≥ 4% during the 1 min sit-and-stand test.



Supplemental Figure 2. Representative scan of a patient in her late 40s showing resolving ground glass abnormality after moderate COVID-19. (A) PA chest radiograph obtained 8 months after admission was considered normal by radiologists. (B) The same radiograph analyzed by the AI algorithm with heat map. Small focal abnormalities in the apical and paracardiac regions of the lungs are highlighted in green and blue. (C, D) Chest CT obtained 11 months after admission shows mild residual ground glass abnormality in the periphery of the upper lobes and left lower lobe. The patient complained of dyspnea (mMRC=3) but had normal lung function (FVC=3.81 L/91% pred) and normal oximetry (99%).

Supplemental Table S1. Demographic and clinical characteristics of the cohort of post-COVID-19 patients in this study (N=749).

Variables	Values
Age (years)	56.1 (44.4–65.1)
Male sex	399 (53.3)
BMI (kg/m^2)	30.8 (27.7–35.6) {746}
Comorbidities	
Hypertension	425 (56.7)
Smokers	285/743 (38.4)
Diabetes	261 (34.8)
COPD	55 (7.3)
Admission	
ICU	445 (59.4)
Length of ICU stay (days)	10 (6–18) {445}
IMV	304/445 (68.3)
Vital signs	
Body temperature (°C)	36.1 (35.6–36.0) {748}
Systolic blood pressure (mmHg)	124 (116–135) {743}
Diastolic blood pressure (mmHg)	77 (70–84) {743}
Heart rate (bpm)	73 (67–83) {747}
Respiratory rate (rpm)	20 (18–2) {736}
Oxygen saturation (%)	97 (95.2–98) {746}

Values are presented as median (IQR), median (IQR) {n}, n (%), or n/N (%). COPD, chronic obstructive pulmonary disease; BMI, body mass index; ICU, intensive care unit. IMV, invasive mechanical ventilation.

Supplemental Table S2. Demographic and clinical characteristics of patients with and without pulmonary involvement (N=749).

Variables	Pulmonary involvement (n=470)	No pulmonary involvement (n=279)	*p*-value
Age (years)	57.9 (45.7–65.8)	53.9 (42.5–63.7)	0.000
Male sex	228 (48.5)	171 (61.3)	0.001
BMI (kg/m^2)	31.2 (27.7–35.9) {469}	30.5 (27.6–35.2) {277}	0.111
Comorbidities			
Hypertension	287 (61.1)	138 (49.5)	0.000
Smokers	188/468 (40.2)	97/275 (35.3)	0.104
Diabetes	179 (38.1)	82 (29.4)	0.009
COPD	42 (8.9)	13 (4.7)	0.044
Admission			
ICU	317 (67.4)	128 (45.9)	0.000
Length of ICU stay (days)	11 (6–20) {317}	8 (4–14) {128}	0.000
IMV	222/317 (70)	82/128 (64.1)	0.260

Values are presented as median (IQR), median (IQR) {n}, n (%), or n/N (%). COPD, chronic obstructive pulmonary disease; BMI, body mass index; ICU, intensive care unit. IMV, invasive mechanical ventilation.

Supplemental Table S3. Demographic and clinical characteristics of COVID-19 patients with signs of pulmonary involvement (N=470).

Variables	Patients with signs of pulmonary involvement		*p*-value
	Those who underwent CT (n=348)	Those who did not undergo CT (n=122)	
Age (years)	57.8 (45.7–65.8)	58.1 (45.3–65.8)	0.490
Male sex	163 (46.8)	65 (53.3)	0.392
BMI (kg/m^2)	31.6 (28.0–36.0)	30.3 (27.0–35.9) {121}	0.041
Comorbidities			
Hypertension	215 (61.8)	72 (59)	0.469
Smokers	139/347 (40.1)	49/121 (40.5)	0.762
Diabetes	142 (40.8)	37 (30.3)	0.999
COPD	32 (9.2)	10 (8.2)	0.826
Admission			
ICU	237 (68.1)	80 (65.6)	0.999
Length of ICU stay (days)	11 (6–20) {237}	10 (4.7–19) {80}	0.913
IMV	174/237 (73.4%)	48/80 (60%)	0.034

Values are presented as median (IQR), median (IQR) {n}, n (%), or n/N (%). COPD, chronic obstructive pulmonary disease; BMI, body mass index; ICU, intensive care unit. IMV, invasive mechanical ventilation.

Supplemental Table S4. Chest computed tomography (CT) features in COVID-19 patients with CT score ≥ 7 (N=156).

Variables	CT changes
CT score ≥ 7	156/328 (47.6)
Characteristics (n=156)	
Ground-glass opacities	153 (98.1)
Parenchymal bands	143 (91.7)
Reticulations	134 (85.9)
Traction bronchiectasis	92 (59)
Architectural distortion	73 (46.8)
Perilobular opacities	50 (32.1)
Bronchial wall thickening	38 (24.4)
Mosaic attenuation pattern	32 (20.5)
Consolidations	3 (1.9)
Pneumatocele	2 (1.3)
Honeycombing	-

Of the 328 patients who underwent CT scan, 47.6% had a CT score ≥ 7. Values are n/N (%) or n (%).

Supplemental Table S5. Computed tomography changes 6 to 11 months after hospitalization due to COVID-19 (N=328).

Characteristics	Total cohort (N=328)	ICU Patients (N=222)	Ward Patients (N=106)
Ground-glass opacities	251 (76.5)	197 (86.6)	54 (51.3)
Parenchymal bands	209 (63.7)	169 (76.5)	40 (41)
Reticulations	169 (51.5)	145 (66.5)	24 (23.1)
Traction bronchiectasis	98 (29.9)	91 (44.1)	7 (7.7)
Architectural distortion	78 (23.8)	73 (35.8)	5 (6.4)
Bronchial wall thickening	89 (27.1)	60 (27.4)	29 (25.6)
Mosaic attenuation pattern	58 (17.7)	46 (20.1)	12 (11.5)
Perilobular opacities	50 (14)	47 (24.6)	3 (2.6)
Consolidation	3 (0.9)	3 (1.7)	-
Pneumatocele	2 (0.6)	2 (1.1)	-
Honeycombing	-	-	-

Values are presented as n (%).

Supplemental Table S6. Demographic and clinical characteristics of COVID-19 patients with pulmonary involvement stratified by inclusion in prediction analysis of pulmonary changes (N=328).

Variables	Patients with Pulmonary Changes		*p*-value
	Included Patients (N=257)	Excluded Patients (N=91)	
Age (years)	56.5 (45.7–64.4)	60.5 (46.9–69.9)	0.011
Male sex	113 (44)	50 (54.9)	0.068
BMI (kg/m^2)	32 (28.8–36.8)	30.6 (26.8–35.4)	0.054
Comorbidities			
Hypertension	151 (58.7)	64 (70.3)	0.060
Smokers	97/256 (37.9)	42 (46.1)	0.173
Diabetes	103 (40.1)	39 (42.9)	0.710
COPD	20 (7.8)	12 (13.2)	0.141
Admission			
ICU	179 (69.6)	58 (63.7)	0.359
Length of ICU stay (days)	12 (6–20.5) {179}	9.5 (6.2–19.7) {58}	0.209
IMV	140 (54.7)	35 (38.6)	0.010

Values are presented as median (IQR), median (IQR) {n}, n (%), or n/N (%). BMI, body mass index; COPD, chronic obstructive pulmonary disease; ICU, intensive care unit. IMV, invasive mechanical ventilation.

Supplemental References

1. Stephens K. SIIM, FISABIO, and RSNA Host Machine Learning Challenge for COVID-19 Detection and Localization. . *AXIS Imaging News* . 2021.

2. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning* 2019. p. 6105-14.

3. Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 2015; **115**: 211-52.

4. Vittinghoff E, Glidden DV, Shiboski SC, et al. Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models. 2nd ed. New York: Springer-Verlag, 2012:1272.