Artificial intelligence-assisted analysis of CT abnormalities during COVID-19 recovery

Supplementary Material

Department of Radiology, Medical University of Innsbruck

2024-03-22

Supplementary Methods

Ethics

The CovILD study was conducted in accordance with the Declaration of Helsinki and the European Data policy. All participants gave written informed consent to participate and to process their data. The study data were processed, stored and analyzed in anonymized form. The study protocol was approved by the ethics committee of the Medical University of Innsbruck, Austria (approval number: 1103/2020). The study was registered at ClinicalTrials.gov (NCT04416100).

Study design and participants, handling of missing data

The longitudinal observation CovILD study aimed at investigation of symptom, cardiopulmonary and mental health recovery of COVID-19 patients. The details of the study design and the cohort are provided in our recent publications (1–4). In brief, n = 145 convalescent COVID-19 survivors were recruited among patients of the University Hospital of Innsbruck, St. Vinzenz Hospital in Zams and Karl-Landsteiner Rehabilitation Center in Münster (all in Austria) between March and June 2020. The inclusion criteria were age ≥ 18 years, SARS-CoV-2 positivity confirmed by PCR and presence of COVID-19 symptoms. All participants were infected with the wild-type form of SARS-CoV-2. The study visits were scheduled at two, three, six, and twelve months after COVID-19 diagnosis. In the current analysis, participant-matched longitudinal observations were utilized. The analysis inclusion criterion was availability of data from computed tomography (CT) of the chest and lung function testing (LFT). This criterion was fulfilled by 420 observations obtained from 140 CovILD study participants. For the patients included in the analysis, the set of explanatory variables concerning demographic and clinical background as well as the course and treatment of acute COVID-19 was complete.

The study design and analysis inclusion process are summarized in a flow chart in **Figure 1**. Numbers of observations at study follow-up examinations are presented in **Supplementary Table S1**. Analyzed variables are listed in **Supplementary Table S2**. Baseline demographic and clinical characteristic of the study cohort and characteristic of acute COVID-19 are provided in **Table 1**. Frequency of CT and LFT abnormalities, scoring of CT lesions and values of LFT readouts are listed in **Supplementary Tables S3** and **S4**. Frequencies of COVID-19 -related symptoms of relevance for lung function are presented in **Supplementary Table S5**.

Study procedures and data sources

Information on demography (age, sex), smoking status, routine medication, pre-existing conditions, as well as the course and treatment of acute COVID-19 were obtained in an interview and retrieved from electronic patient's record at the two-month follow-up visit. The standard follow-up visit protocol included a survey of symptoms and self-reported physical performance, examination and interview by a physician, determination of standard blood markers (hemoglobin, iron turnover, complete blood count, markers of inflammation and vascular pathology), trans-thoracic echocardiography, LFT, and CT of the chest. The details of the study procedures are provided in our recent publications (1–4). The complete list of study variables with their format and descriptions is provided in **Supplementary Table S2**.

Regarding the severity of acute COVID-19, the study participants were classified as 'ambulatory' (home isolated, WHO ordinal scale for clinical improvement: 1 - 2), 'hospitalized moderate' (hospitalized, without oxygen therapy or oxygen by mask or nasal prongs, WHO: 3 - 4), and 'hospitalized severe' (hospitalized, high flow oxygen or mechanical ventilation, WHO: 5 - 7).

Chest CT was done with a 128 slice multi-detector SOMATOM Definition Flash device (Siemens Healthineers, Erlangen, Germany) with a 128 × 0.6 mm collimation and spiral pitch factor of 1.1. Scans were obtained in craniocaudal direction without iodine contrast agent and in low-dose setting (100 kVp tube potential). Artificial intelligence (AI) supported analysis of the lung CT scans was done with the Syngo.via CT Pneumonia Analysis Software (Siemens Healthineers, Erlangen, Germany), which returned percentage of the lungs with opacity and high opacity. Evaluation of the CT scans by the radiologists was done in accordance with the Fleischner society guidelines as described before (1,3,5). In brief, the scans were examined for presence of ground glass opacity (GGO), reticulation, consolidations, and bronchiectasis. Severity of lung abnormalities was rated by the radiologist with the CT severity score (CTSS) described before (1,3). For computation of CTSS, radiological abnormalities were scored separately for each lobe with the following scheme:

- 0: no abnormalities
- 1: minimal, subtle GGO
- 2: mild, several GGO and subtle reticulations
- 3: moderate, multiple GGO, reticulation, small consolidation
- 4: severe, extensive GGO, consolidation, reticulation with distortion
- 5: massive findings, parenchymal destruction

CTSS was calculated as a sum of the scores over all lobes and ranged from 0 to 25 (1,3).

The following LFT parameters were recorded: diffusion capacity for carbon monoxide (DLCO), forced vital capacity (FVC), forced expiratory volume in one second (FEV1), total lung capacity (TLC), and FEV1 to FVC ratio (FEV1:FVC). Among them, DLCO, FVC, and FEV1, whose abnormalities were the most frequent in the CovILD cohort, were analyzed in the current report. These variables were expressed as percentage of the patient's individual age-, sex- and weight-specific reference. Insufficiency of DLCO, FVC, or FEV1 were defined as values below 80% of the reference (1,2,4).

The following longitudinally recorded symptoms of potential relevance for lung function were evaluated: dyspnea measured by Modified Medical Research Council scale (mMRC, presence of dyspnea defined as mMRC > 0), self-reported cough, and self-rated physical performance measured by Eastern Cooperative Oncology Group (ECOG, impaired physical performance defined as ECOG > 0) (4). Missing information on the symptoms or their intensity in the patient's symptom survey were interpreted as absence of the complaint.

The source study data set was fetched from the CovILD study data repository available as an R package to authorized users.

Software

The analysis was done with R version 4.2.3. Tabular data and code pipelines were handled by tools provided by the packages *tidyverse* (6), *rlang* (7), and *trafo*. Text data were handled with *stringi* (8). Parallelization was accomplished with *furrr* (9) and *doParallel* (10).

In exploratory data analysis, statistical hypothesis testing and analysis of correspondence and correlations, the packages *rstatix* (11), *rcompanion* (12), *MASS* (13), *clustTools*, and *ExDA* were used. Receiver-operating characteristic (ROC), confusion matrix analysis and inter-rater reliability analyses were done with *caret* (14), *OptimalCutpoints* (15) and *bootStat*.

For machine learning modeling of lung function testing outcomes, the packages *caret* (14) and *caretExtra* were utilized, which provide wrappers around the R implementations of the random forest algorithm (*ranger*) (16,17), neural network (*nnet*) (18), support vector machines (SVM, *kernlab*) (19,20), and gradient boosted machines (GBM, *gbm*) (21–23). Model diagnostic statistics and performance evaluation metrics were computed with the development package *caretExtra*. SHAP (Shapley additive explanations) as a metric of global variance importance in machine learning models was calculated with the package *kernelshap* (24–26).

Analysis results were visualized with *ggplot* (6) and *ExDA* (scatter, stack and box plots), *clustTools* (visualization of results of correspondence analysis), *ggvenn* (27) (Venn plots), *caretExtra* (plots for machine learning model diagnostic and evaluation), *plotROC* (28)

(ROC curves) and *shapviz* (26,29) (violin/scatter plots of absolute SHAP values). Result tables were created with *flextable* (30). Figures were generated with *cowplot* (31). Parts of the manuscript and supplementary material were written in the *rmarkdown* environment (32) with the *bookdown* package (33). Figures, tables, links and R expressions in the Rmarkdown document were managed with *figur*.

Descriptive statistic, statistical tests and effect sizes, correspondence analysis

If not stated otherwise, numeric values are presented as medians with interquartile ranges and ranges within the analysis data set or analysis strata. Qualitative variables are presented as percentages and counts of the category within the analysis data set or analysis strata. Normality of distribution of numeric variables upon identity, logarithm and square root transformation was checked with Shapiro-Wilk test (function `shapiro_test`, package *rstatix*). Since multiple numeric variables were not normally distributed, Mann-Whitney test or Kruskal-Wallis test were routinely used for comparison of independently distributed numeric variables between analysis groups. Differences in distribution of categorical variables between the study groups were assessed with χ^2 test. P values were adjusted for multiple testing with the false discovery rate method separately for each analysis task (e.g. comparison of explanatory variables between observations with and without LFT findings) (34). Effects with $p < 0.05$ were considered significant.

The following effect size metrics were used to assess differences between analysis groups and to evaluate performance of machine learning models (35–37):

- biseral r was used for two-group comparisons of numeric variables: < 0.3: small, 0.3 - 0.5: moderate, ≥ 0.5: large effect size
- η^2 was used for multi-group comparison of numeric variables: < 0.13: small, 0.13 - 0.26: moderate, ≥ 0.26: large effect size
- Cramer's V was used for comparisons of categorical variables: < 0.3: small, 0.3 - 0.5: moderate, ≥ 0.5: large effect size
- Spearman's ρ coefficient of correlation: < 0.3: small, 0.3 - 0.5: moderate, ≥ 0.5: large effect size
- pseudo-R^2 was used as a measure of explained variance of a model: < 0.13: small, 0.13 - 0.26: moderate, ≥ 0.26: large effect size
- Cohen's κ inter-rater reliability statistic: < 0.20: no effect, 0.2 - 0.4: fair, 0.4 - 0.6: moderate, 0.6 - 0.8: good, ≥ 0.8: excellent concordance

Statistical significance and effect sizes for comparisons and correlations of independent observations was assessed with the functions `compare_variables()` and `correlate_variables()` (package *ExDA*). Because of non-independent observations included in the analysis, pairwise correlations of numeric LFT or CT variables were investigated with blocked bootstrap Spearman's rank test described below. Analogically, for comparison of medians between two subsets of non-independent observations, blocked bootstrap test was utilized. Please refer to **Comparison of CT readouts between observations with and without LFT abnormality, correlation analysis** for details.

Co-occurrence of LFT abnormalities (DLCO < 80% reference, FVC < 80% of reference, and FEV1 < 780% of reference) was investigated by two-dimensional correspondence analysis (function `mca()`, package *MASS*) (13) and whose column factors were visualized in a scatter plot with functions from the *clustTools* package. Additionally, counts of observations with co-occurring LFT abnormalities were visualized in Venn plot (function `ggvenn()`, package *ggvenn*).

Multi-parameter modeling of lung function

Presence of reduced DLCO, FVC and FEV1 (each < 80% of reference), as well as values of DLCO, FVC and FEV1 expressed as percentages of the reference value were modeled with 37 explanatory variables including:

- baseline demographic and clinical characteristic (e.g. age, sex, smoking, comorbidities)
- explanatory variables referring to the course of acute COVID-19 (e.g. WHO ordinal scale for clinical improvement, hospitalization status and length, treatment)
- follow-up after COVID-19 diagnosis (two, three, six and twelve months)
- presence and rating of longitudinally recorded symptoms of relevance for lung function (dyspnea, cough, impaired physical performance)
- presence and severity of lung CT findings rated by the radiologist (e.g. GGO, consolidations, CTSS) and AI (opacity and high opacity)

To reduce the number of explanatory variables, some qualitative variables with rare categories were re-coded. This concerned:

- *pulmonary illness* which subsumed pre-existing chronic obstructive lung disease (COPD), bronchial asthma, interstitial lung disease, chronic lung diseases and other pulmonary conditions
- *anti-coagulants during COVID-19* which subsumed anti-coagulation and anti-platelet treatment during acute COVID-19

- *anti-infectives during COVID-19* which subsumed anti-infective and anti-macrolide treatment during acute COVID-19
- *smoking history*, where ex- and active smokers were grouped together

Severity of acute COVID-19 measured by the WHO ordinal scale of clinical improvement was re-coded as categorical explanatory variable. The response and explanatory variables are listed in **Supplementary Table S2**, the modeling strategy is schematically summarized in **Supplementary Figure S8**.

Multi-parameter models were constructed with four machine learning algorithms employing diverse mathematical principles: canonical random forest (implemented in R by the *ranger* package, caret's method name: 'ranger') (16,17), GBM (*gbm* package, caret's method name: 'gbm') (21–23), neural network with a single hidden layer (*nnet* package, caret's method name: 'nnet') (18), and SVM with radial kernel (*kernlab* package, caret's method name: 'svmRadial') (19,20). For random forest models, 1000 random trees per model were generated. Bernoulli and Gaussian loss functions were implemented in classification and regression GBM models, respectively. Optimal values of parameters controlling the algorithm's behavior such as number of observations in terminal nodes of tree models in the random forest models or cost penalty in the SVM models were found by maximizing the Youden's J statistic ($J = Sensitivity + Specificity - 1$, classification models of binary LFT outcomes) or by minimizing mean absolute error (MAE, regression models of numeric LFT outcomes) in 10-repeats 10-fold cross-validation. To account for the participant matching of the observations, a blocked cross-validation workflow was designed to keep all observations obtained from a particular participant either in the training or in the test portion of the cross-validation split. Construction of such 'participant-wise' blocked cross-validation folds was done with an in-house-developed script employing the `createMultiFolds()` function (package *caret*). For tuning and training of the machine learning models, the `train()` wrapper from the *caret* package was employed (14). The optimal values of the tuning parameters are listed in **Supplementary Table S6**.

Performance of the machine learning models was evaluated by comparison of the model predictions with the observed outcome in the genuine training data set and in 10-repeats 10-fold cross-validation with the observed outcomes. Performance of the classification models was assessed by Cohen's κ as a measure of concordance between the predicted and observed outcome (35,37), standard receiver operating characteristic metrics (AUC, overall accuracy, sensitivity, specificity) to assess the ability of the model to detect the pathological outcome (14), and with Brier score to investigate credibility of the predictions and model calibration (38,39) (**Table 2** and **Supplementary Table S7**). Performance of the regression models was assessed with mean absolute error (MAE) as a measure of fit quality, pseudo-R^2 as a metric of explained variance and Spearman's ρ for correlation between the predicted and observed outcome as as measure of overall model calibration (14) (**Table 3** and **Table 8**). Pseudo-R^2 was computed with the following formulas:

$$pseudoR^2 = 1 - \frac{MSE(y)}{Var(y)}$$

$$MSE(y) = \frac{1}{N}\sum_{i}^{N}(y_i - \hat{y}_i)^2$$

where $MSE(y)$ stands for mean squared error of the outcome variable y, $Var(y)$ is the variance of the outcome variable y, N is the total observation number, i represents the observation index, y_i is the observed outcome value for the i-th observation, and $\hat{y}_i$ is the predicted outcome value for the i-th observation. Performance statistics were retrieved from the `caret` models with the `summary()` method from the *caretExtra* package. Visualization of the model tuning and evaluation results was accomplished with the method `plot()` from the *caretExtra* package and in-house developed, project-specific functions.

In order to assess the model accuracy as a function of acute COVID-19 severity and time after diagnosis, we computed Cohen's κ (classification models) and MAE (regression models) of out-of-fold predictions for observations stratified by acute COVID-19 severity (mild, moderate, severe) and follow-up visit (2-, 3-, 6- and 12-month follow-up). The error values were subsequently visualized as heat maps.

Variable importance for multi-parameter models of lung function

Shapley additive explanations (SHAP) were used as a global importance statistic for the model's explanatory variables. In brief, SHAP measures contribution of the given variable to the model fit by comparing a goodness-of-fit statistic between models re-fitted with subsets of explanatory variables including the variable of interest and models re-fitted with subsets of explanatory variables without the variable of interest (24,25). Matrices of SHAP values for observations and variables were computed for the machine learning models of lung function outcomes with the function `kernelshap()` provided by the *kernelshap* package (29). As a background for SHAP calculation, a single observation data frame was used with numeric explanatory variables set to the 25th percentile of the values in the entire data set and with qualitative explanatory variables set to the baseline category (29). For mean absolute SHAP values of the most influential explanatory variables in the Random Forest, neural network, SVM, and GBM models, see: **Table 9**. Visualization of the mean SHAP values for the top most influential explanatory variables was accomplished with the `shapviz()`, `sv_importance()` (package *shapviz*) (26), and in-house developed functions.

Comparison of CT readouts between observations with and without LFT abnormality, correlation analysis

Because observations were non-independent, i.e. participant-matched, comparison of CTSS, opacity and high opacity of the lung between data points with and without LFT abnormalities could not be investigated by standard statistical tests (e.g. Mann-Whitney test) and required instead a special approach. Differences in median of the CT parameters between observations with and without a LFT abnormality were investigated with blocked bootstrap test with biseral r effect size statistic (H0: no difference in CT parameter, H1: CT parameter values are higher in observations with the LFT abnormality). The blocked bootstrap scheme involved re-sampling with repetition of the participants instead of single observations (B = 2000 re-samples). P values were obtained by counting re-samples in which the test hypothesis H1 was not met and dividing the count by the total re-sample number. The effect size statistic was computed as an arithmetic average of effect sizes in the re-samples.

A similar blocked bootstrap approach was applied in a correlation analysis of the CT (CTSS, opacity, high opacity) and LFT readouts (DLCO, FVC, FEV1, expressed as percentages of reference values). The H0 null hypothesis of such correlation test was no or positive correlation of a CT and LFT readout. The H1 hypothesis was the presence of negative correlation between the CT and LFT readout. Spearman's ρ correlation coefficients were computed for the bootstrap re-samples (B = 2000). The p values were obtained by counting re-samples with $\rho \geq 0$ and dividing the count by the total re-sample number. The ρ values were averaged and their confidence values were calculated as the 2.5 - 97.5 percentile range.

Bootstrap tests were performed with a functional interface implemented in the development package *bootStat*. The analysis results are listed in **Supplementary Table S10** and **11**.

Inter-rater reliability and receiver-operating characteristic analysis

The optimal cutoffs of human-determined CTSS, and AI-determined lung opacity and high opacity for detection of insufficient DLCO (< 80% of reference) were computed with the maximum of the Youden's J statistic ($J = Sensitivity + Specificity - 1$) in the entire analysis data set (function `optimal.cutpoints()`, package *OptimalCutpoints*) (15). Those optimal cutoffs were subsequently used to dichotomize the analyzed observations. Concordance between the $CTSS^{high}$, $opacity^{high}$ and $high\ opacity^{high}$ strata assignment and presence of LFT insufficiency of reference was assessed by Cohen's κ inter-rater reliability statistic (35,37) as well as sensitivity, specificity and AUC metrics. Those statistics were computed with the function `multiClassSummary()` from the *caret* package (14).

Boundaries of the statistic's 95% confidence intervals were obtained by blocked bootstrap with 2000 iterations and were defined by the 2.5 and 97.5 percentiles of the blocked bootstrap estimates. Bootstrapping was accomplished by a functional interface implemented in the development package *bootStat*. The analysis results are listed in **Table 4**.

Data and code availability

The entire R analysis pipeline is available as a GitHub repository. The CovILD data set will be made available on reasonable request to the corresponding author.

Supplementary Tables

Supplementary Table S1: Numbers of observations included in the analysis split by acute COVID-19 severity and the follow-up examination.

Follow-up, months	Cohort	Ambulatory, mild COVID-19	Hospitalized, moderate COVID-19	Hospitalized, severe COVID-19
2-month	120	27	71	22
3-month	124	29	66	29
6-month	85	10	51	24
2-month	91	19	46	26

Supplementary Table S2: Study variables. The table is available as a supplementary Excel file.

Supplementary Table S3: Chest computed tomography variables at consecutive follow-ups. Numeric variables are presented as medians with interquartile ranges (IQR) and ranges. Categorical variables are presented as percentages and counts within the complete observation set.

Severity subset	Variable[a]	2-month follow-up	3-month follow-up	6-month follow-up	12-month follow-up
CovILD cohort	any CT findings	71% (n = 85) complete: n = 120	57% (n = 71) complete: n = 124	64% (n = 54) complete: n = 85	54% (n = 49) complete: n = 91
	GGO	69% (n = 83) complete: n = 120	53% (n = 66) complete: n = 124	52% (n = 44) complete: n = 85	44% (n = 40) complete: n = 91
	reticulation	51% (n = 61) complete: n = 120	48% (n = 60) complete: n = 124	55% (n = 47) complete: n = 85	43% (n = 39) complete: n = 91
	consolidation	11% (n = 13) complete: n = 120	6.5% (n = 8) complete: n = 124	1.2% (n = 1) complete: n = 85	1.1% (n = 1) complete: n = 91
	bronchiectasis	10% (n = 12) complete: n = 120	5.6% (n = 7) complete: n = 124	8.2% (n = 7) complete: n = 85	8.8% (n = 8) complete: n = 91
	CTSS, points	5 [IQR: 0 - 13] range: 0 - 20 complete: n = 120	2 [IQR: 0 - 7.2] range: 0 - 18 complete: n = 124	2 [IQR: 0 - 5] range: 0 - 15 complete: n = 85	1 [IQR: 0 - 5] range: 0 - 15 complete: n = 91
	opacity, AI, % of lung	0.19 [IQR: 0 - 2.7] range: 0 - 37 complete: n = 120	0.074 [IQR: 0 - 0.86] range: 0 - 22 complete: n = 124	0.04 [IQR: 0 - 0.43] range: 0 - 12 complete: n = 85	0 [IQR: 0 - 0.12] range: 0 - 6.2 complete: n = 91
	high opacity, AI, % of lung	0.005 [IQR: 0 - 0.11] range: 0 - 3.1 complete: n = 120	0.002 [IQR: 0 - 0.053] range: 0 - 2.4 complete: n = 124	0 [IQR: 0 - 0.01] range: 0 - 0.46 complete: n = 85	0 [IQR: 0 - 0.01] range: 0 - 0.11 complete: n = 91
ambulatory	any CT findings	26% (n = 7) complete: n = 27	17% (n = 5) complete: n = 29	30% (n = 3) complete: n = 10	11% (n = 2) complete: n = 19
	GGO	22% (n = 6) complete: n = 27	14% (n = 4) complete: n = 29	10% (n = 1) complete: n = 10	5.3% (n = 1) complete: n = 19
	reticulation	7.4% (n = 2) complete: n = 27	6.9% (n = 2) complete: n = 29	30% (n = 3) complete: n = 10	11% (n = 2) complete: n = 19
	consolidation	3.7% (n = 1) complete: n = 27	3.4% (n = 1) complete: n = 29	0% (n = 0) complete: n = 10	0% (n = 0) complete: n = 19
	bronchiectasis	0% (n = 0) complete: n = 27	0% (n = 0) complete: n = 29	0% (n = 0) complete: n = 10	0% (n = 0) complete: n = 19

Severity subset	Variable[a]	2-month follow-up	3-month follow-up	6-month follow-up	12-month follow-up
	CTSS, points	0 [IQR: 0 - 1] range: 0 - 20 complete: n = 27	0 [IQR: 0 - 0] range: 0 - 9 complete: n = 29	0 [IQR: 0 - 2.2] range: 0 - 5 complete: n = 10	0 [IQR: 0 - 0] range: 0 - 3 complete: n = 19
	opacity, AI, % of lung	0 [IQR: 0 - 0.04] range: 0 - 14 complete: n = 27	0 [IQR: 0 - 0.04] range: 0 - 1.2 complete: n = 29	0 [IQR: 0 - 0.025] range: 0 - 0.17 complete: n = 10	0 [IQR: 0 - 0] range: 0 - 0.16 complete: n = 19
	high opacity, AI, % of lung	0 [IQR: 0 - 0.001] range: 0 - 1.9 complete: n = 27	0 [IQR: 0 - 0] range: 0 - 0.12 complete: n = 29	0 [IQR: 0 - 0] range: 0 - 0.03 complete: n = 10	0 [IQR: 0 - 0] range: 0 - 0.05 complete: n = 19
moderate	any CT findings	80% (n = 57) complete: n = 71	61% (n = 40) complete: n = 66	57% (n = 29) complete: n = 51	54% (n = 25) complete: n = 46
	GGO	79% (n = 56) complete: n = 71	56% (n = 37) complete: n = 66	49% (n = 25) complete: n = 51	41% (n = 19) complete: n = 46
	reticulation	55% (n = 39) complete: n = 71	52% (n = 34) complete: n = 66	47% (n = 24) complete: n = 51	43% (n = 20) complete: n = 46
	consolidation	9.9% (n = 7) complete: n = 71	9.1% (n = 6) complete: n = 66	0% (n = 0) complete: n = 51	0% (n = 0) complete: n = 46
	bronchiectasis	7% (n = 5) complete: n = 71	7.6% (n = 5) complete: n = 66	5.9% (n = 3) complete: n = 51	4.3% (n = 2) complete: n = 46
	CTSS, points	6 [IQR: 1.5 - 12] range: 0 - 20 complete: n = 71	2 [IQR: 0 - 5.8] range: 0 - 13 complete: n = 66	1 [IQR: 0 - 4.5] range: 0 - 13 complete: n = 51	1 [IQR: 0 - 2.8] range: 0 - 13 complete: n = 46
	opacity, AI, % of lung	0.21 [IQR: 0.0055 - 2] range: 0 - 22 complete: n = 71	0.057 [IQR: 0 - 0.59] range: 0 - 11 complete: n = 66	0.02 [IQR: 0 - 0.29] range: 0 - 12 complete: n = 51	0 [IQR: 0 - 0.048] range: 0 - 2 complete: n = 46
	high opacity, AI, % of lung	0.005 [IQR: 0 - 0.053] range: 0 - 3.1 complete: n = 71	5e-04 [IQR: 0 - 0.027] range: 0 - 2.4 complete: n = 66	0 [IQR: 0 - 0] range: 0 - 0.46 complete: n = 51	0 [IQR: 0 - 0] range: 0 - 0.11 complete: n = 46
severe	any CT findings	95% (n = 21) complete: n = 22	90% (n = 26) complete: n = 29	92% (n = 22) complete: n = 24	85% (n = 22) complete: n = 26
	GGO	95% (n = 21) complete: n = 22	86% (n = 25) complete: n = 29	75% (n = 18) complete: n = 24	77% (n = 20) complete: n = 26
	reticulation	91% (n = 20) complete: n = 22	83% (n = 24) complete: n = 29	83% (n = 20) complete: n = 24	65% (n = 17) complete: n = 26

Severity subset	Variable[a]	2-month follow-up	3-month follow-up	6-month follow-up	12-month follow-up
	consolidation	23% (n = 5) complete: n = 22	3.4% (n = 1) complete: n = 29	4.2% (n = 1) complete: n = 24	3.8% (n = 1) complete: n = 26
	bronchiectasis	32% (n = 7) complete: n = 22	6.9% (n = 2) complete: n = 29	17% (n = 4) complete: n = 24	23% (n = 6) complete: n = 26
	CTSS, points	14 [IQR: 10 - 15] range: 0 - 20 complete: n = 22	10 [IQR: 5 - 15] range: 0 - 18 complete: n = 29	6 [IQR: 2.8 - 10] range: 0 - 15 complete: n = 24	5 [IQR: 2 - 11] range: 0 - 15 complete: n = 26
	opacity, AI, % of lung	5.8 [IQR: 0.34 - 9.3] range: 0 - 37 complete: n = 22	1.9 [IQR: 0.39 - 6.8] range: 0 - 22 complete: n = 29	0.32 [IQR: 0.042 - 2] range: 0 - 6.7 complete: n = 24	0.28 [IQR: 0 - 0.67] range: 0 - 6.2 complete: n = 26
	high opacity, AI, % of lung	0.19 [IQR: 0.066 - 0.81] range: 0 - 2.1 complete: n = 22	0.06 [IQR: 0.011 - 0.26] range: 0 - 1 complete: n = 29	0 [IQR: 0 - 0.02] range: 0 - 0.23 complete: n = 24	0.01 [IQR: 0 - 0.02] range: 0 - 0.11 complete: n = 26

[a]any CT findings: any CT findings diagnosed by a human pathologist; GGO: ground glass opacity; CTSS: human-determined CT severity score; reticulation, consolidation, bronchiectasis: diagnosed by a human radiologist; opacity and high opacity, AI: opacity and high opacity determined by artificial intelligence.

Supplementary Table S4: Lung function testing variables at consecutive follow-ups. Numeric variables are presented as medians with interquartile ranges (IQR) and ranges. Categorical variables are presented as percentages and counts within the complete observation set.

Severity subset	Variable[a]	2-month follow-up	3-month follow-up	6-month follow-up	12-month follow-up
CovILD cohort	FVC, % of reference	88 [IQR: 80 - 97] range: 46 - 130 complete: n = 120	90 [IQR: 83 - 97] range: 54 - 120 complete: n = 124	91 [IQR: 84 - 98] range: 50 - 120 complete: n = 85	90 [IQR: 84 - 99] range: 55 - 120 complete: n = 91
	FVC < 80%	24% (n = 29) complete: n = 120	21% (n = 26) complete: n = 124	18% (n = 15) complete: n = 85	14% (n = 13) complete: n = 91
	FEV1, % of reference	92 [IQR: 82 - 100] range: 50 - 140 complete: n = 120	92 [IQR: 82 - 100] range: 30 - 130 complete: n = 124	94 [IQR: 84 - 100] range: 30 - 130 complete: n = 85	95 [IQR: 84 - 110] range: 60 - 140 complete: n = 91
	FEV1 < 80%	20% (n = 24) complete: n = 120	21% (n = 26) complete: n = 124	15% (n = 13) complete: n = 85	15% (n = 14) complete: n = 91
	DLCO, % of reference	91 [IQR: 77 - 110] range: 44 - 140 complete: n = 120	94 [IQR: 83 - 110] range: 31 - 140 complete: n = 124	95 [IQR: 83 - 110] range: 32 - 140 complete: n = 85	93 [IQR: 86 - 110] range: 48 - 130 complete: n = 91
	DLCO < 80%	29% (n = 35) complete: n = 120	22% (n = 27) complete: n = 124	21% (n = 18) complete: n = 85	15% (n = 14) complete: n = 91
ambulatory	FVC, % of reference	95 [IQR: 87 - 100] range: 46 - 110 complete: n = 27	94 [IQR: 87 - 100] range: 64 - 110 complete: n = 29	91 [IQR: 86 - 100] range: 64 - 110 complete: n = 10	95 [IQR: 85 - 100] range: 72 - 110 complete: n = 19
	FVC < 80%	11% (n = 3) complete: n = 27	17% (n = 5) complete: n = 29	10% (n = 1) complete: n = 10	21% (n = 4) complete: n = 19
	FEV1, % of reference	94 [IQR: 86 - 100] range: 50 - 110 complete: n = 27	98 [IQR: 87 - 100] range: 65 - 120 complete: n = 29	95 [IQR: 85 - 100] range: 70 - 110 complete: n = 10	96 [IQR: 88 - 100] range: 72 - 110 complete: n = 19
	FEV1 < 80%	11% (n = 3) complete: n = 27	14% (n = 4) complete: n = 29	10% (n = 1) complete: n = 10	16% (n = 3) complete: n = 19
	DLCO, % of reference	110 [IQR: 91 - 120] range: 52 - 140 complete: n = 27	100 [IQR: 96 - 110] range: 69 - 130 complete: n = 29	99 [IQR: 89 - 110] range: 77 - 130 complete: n = 10	98 [IQR: 92 - 110] range: 74 - 120 complete: n = 19
	DLCO < 80%	15% (n = 4) complete: n = 27	10% (n = 3) complete: n = 29	20% (n = 2) complete: n = 10	5.3% (n = 1) complete: n = 19
moderate	FVC, % of reference	87 [IQR: 80 - 96] range: 46 - 130	90 [IQR: 84 - 98] range: 54 - 110	96 [IQR: 84 - 99] range: 50 - 120	91 [IQR: 85 - 99] range: 55 - 120

Severity subset	Variable[a]	2-month follow-up	3-month follow-up	6-month follow-up	12-month follow-up
		complete: n = 71	complete: n = 66	complete: n = 51	complete: n = 46
	FVC < 80%	24% (n = 17) complete: n = 71	17% (n = 11) complete: n = 66	16% (n = 8) complete: n = 51	11% (n = 5) complete: n = 46
	FEV1, % of reference	92 [IQR: 83 - 100] range: 60 - 140 complete: n = 71	92 [IQR: 85 - 100] range: 30 - 120 complete: n = 66	97 [IQR: 85 - 110] range: 30 - 130 complete: n = 51	95 [IQR: 87 - 110] range: 60 - 140 complete: n = 46
	FEV1 < 80%	17% (n = 12) complete: n = 71	17% (n = 11) complete: n = 66	12% (n = 6) complete: n = 51	15% (n = 7) complete: n = 46
	DLCO, % of reference	93 [IQR: 79 - 100] range: 52 - 120 complete: n = 71	96 [IQR: 85 - 110] range: 31 - 140 complete: n = 66	100 [IQR: 89 - 110] range: 32 - 140 complete: n = 51	97 [IQR: 87 - 110] range: 54 - 130 complete: n = 46
	DLCO < 80%	25% (n = 18) complete: n = 71	20% (n = 13) complete: n = 66	18% (n = 9) complete: n = 51	15% (n = 7) complete: n = 46
severe	FVC, % of reference	83 [IQR: 71 - 91] range: 58 - 110 complete: n = 22	86 [IQR: 71 - 91] range: 54 - 120 complete: n = 29	87 [IQR: 80 - 91] range: 61 - 110 complete: n = 24	86 [IQR: 83 - 93] range: 58 - 120 complete: n = 26
	FVC < 80%	41% (n = 9) complete: n = 22	34% (n = 10) complete: n = 29	25% (n = 6) complete: n = 24	15% (n = 4) complete: n = 26
	FEV1, % of reference	83 [IQR: 75 - 97] range: 62 - 120 complete: n = 22	88 [IQR: 78 - 98] range: 60 - 130 complete: n = 29	88 [IQR: 83 - 95] range: 66 - 110 complete: n = 24	94 [IQR: 83 - 99] range: 63 - 120 complete: n = 26
	FEV1 < 80%	41% (n = 9) complete: n = 22	38% (n = 11) complete: n = 29	25% (n = 6) complete: n = 24	15% (n = 4) complete: n = 26
	DLCO, % of reference	76 [IQR: 58 - 87] range: 44 - 130 complete: n = 22	84 [IQR: 72 - 90] range: 49 - 110 complete: n = 29	85 [IQR: 71 - 93] range: 50 - 110 complete: n = 24	89 [IQR: 83 - 97] range: 48 - 120 complete: n = 26
	DLCO < 80%	59% (n = 13) complete: n = 22	38% (n = 11) complete: n = 29	29% (n = 7) complete: n = 24	23% (n = 6) complete: n = 26

[a]LFT: lung function testing; FVC: forced vital capacity; FEV1: forced expiratory volume in one second; DLCO: diffusion capacity for carbon monoxide.

Supplementary Table S5: Presence and rating of symptoms of relevance for lung function at consecutive follow-ups. Numeric variables are presented as medians with interquartile ranges (IQR) and ranges. Categorical variables are presented as percentages and counts within the complete observation set.

Severity subset	Variable[a]	2-month follow-up	3-month follow-up	6-month follow-up	12-month follow-up
CovILD cohort	dyspnea rating, mMRC, points	0 [IQR: 0 - 1] range: 0 - 4 complete: n = 120	0 [IQR: 0 - 1] range: 0 - 3 complete: n = 124	0 [IQR: 0 - 1] range: 0 - 4 complete: n = 85	0 [IQR: 0 - 0] range: 0 - 4 complete: n = 91
	dyspnea, mMRC > 0	45% (n = 54) complete: n = 120	34% (n = 42) complete: n = 124	27% (n = 23) complete: n = 85	22% (n = 20) complete: n = 91
	cough	18% (n = 22) complete: n = 120	17% (n = 21) complete: n = 124	13% (n = 11) complete: n = 85	15% (n = 14) complete: n = 91
	physical performance, ECOG, points	1 [IQR: 0 - 1] range: 0 - 4 complete: n = 120	0.5 [IQR: 0 - 1] range: 0 - 4 complete: n = 124	0 [IQR: 0 - 1] range: 0 - 1 complete: n = 85	0 [IQR: 0 - 1] range: 0 - 2 complete: n = 91
	impaired physical performance, ECOG > 0	53% (n = 64) complete: n = 120	50% (n = 62) complete: n = 124	33% (n = 28) complete: n = 85	34% (n = 31) complete: n = 91
ambulatory	dyspnea rating, mMRC, points	0 [IQR: 0 - 1.5] range: 0 - 4 complete: n = 27	0 [IQR: 0 - 1] range: 0 - 3 complete: n = 29	0 [IQR: 0 - 1] range: 0 - 1 complete: n = 10	0 [IQR: 0 - 0] range: 0 - 1 complete: n = 19
	dyspnea, mMRC > 0	44% (n = 12) complete: n = 27	38% (n = 11) complete: n = 29	40% (n = 4) complete: n = 10	21% (n = 4) complete: n = 19
	cough	22% (n = 6) complete: n = 27	24% (n = 7) complete: n = 29	20% (n = 2) complete: n = 10	21% (n = 4) complete: n = 19
	physical performance, ECOG, points	1 [IQR: 0 - 1] range: 0 - 2 complete: n = 27	1 [IQR: 0 - 1] range: 0 - 4 complete: n = 29	0 [IQR: 0 - 1] range: 0 - 1 complete: n = 10	0 [IQR: 0 - 1] range: 0 - 1 complete: n = 19
	impaired physical performance, ECOG > 0	59% (n = 16) complete: n = 27	59% (n = 17) complete: n = 29	40% (n = 4) complete: n = 10	37% (n = 7) complete: n = 19
moderate	dyspnea rating, mMRC, points	0 [IQR: 0 - 1] range: 0 - 4 complete: n = 71	0 [IQR: 0 - 1] range: 0 - 3 complete: n = 66	0 [IQR: 0 - 0.5] range: 0 - 3 complete: n = 51	0 [IQR: 0 - 0] range: 0 - 4 complete: n = 46
	dyspnea, mMRC > 0	38% (n = 27) complete: n = 71	36% (n = 24) complete: n = 66	25% (n = 13) complete: n = 51	20% (n = 9) complete: n = 46

Severity subset	Variable[a]	2-month follow-up	3-month follow-up	6-month follow-up	12-month follow-up
	cough	17% (n = 12) complete: n = 71	15% (n = 10) complete: n = 66	12% (n = 6) complete: n = 51	15% (n = 7) complete: n = 46
	physical performance, ECOG, points	0 [IQR: 0 - 1] range: 0 - 3 complete: n = 71	0 [IQR: 0 - 1] range: 0 - 1 complete: n = 66	0 [IQR: 0 - 0.5] range: 0 - 1 complete: n = 51	0 [IQR: 0 - 1] range: 0 - 1 complete: n = 46
	impaired physical performance, ECOG > 0	44% (n = 31) complete: n = 71	41% (n = 27) complete: n = 66	25% (n = 13) complete: n = 51	28% (n = 13) complete: n = 46
severe	dyspnea rating, mMRC, points	1 [IQR: 0 - 1] range: 0 - 4 complete: n = 22	0 [IQR: 0 - 0] range: 0 - 3 complete: n = 29	0 [IQR: 0 - 0.25] range: 0 - 4 complete: n = 24	0 [IQR: 0 - 0.75] range: 0 - 1 complete: n = 26
	dyspnea, mMRC > 0	68% (n = 15) complete: n = 22	24% (n = 7) complete: n = 29	25% (n = 6) complete: n = 24	27% (n = 7) complete: n = 26
	cough	18% (n = 4) complete: n = 22	14% (n = 4) complete: n = 29	12% (n = 3) complete: n = 24	12% (n = 3) complete: n = 26
	physical performance, ECOG, points	1 [IQR: 1 - 1] range: 0 - 4 complete: n = 22	1 [IQR: 0 - 1] range: 0 - 2 complete: n = 29	0 [IQR: 0 - 1] range: 0 - 1 complete: n = 24	0 [IQR: 0 - 1] range: 0 - 2 complete: n = 26
	impaired physical performance, ECOG > 0	77% (n = 17) complete: n = 22	62% (n = 18) complete: n = 29	46% (n = 11) complete: n = 24	42% (n = 11) complete: n = 26

[a]mMRC: modified Medical Research Council scale; ECOG: Eastern Cooperative Oncology Group performance scale.

Supplementary Table S6: Selection of machine learning algorithm parameters by cross-validation-mediated tuning.

Response[a]	Algorithm	Parameters
DLCO < 80%	Random Forest	mtry = 14 splitrule = gini min.node.size = 3
	Neural network	size = 19 decay = 0.01
	Support vector machines, radial kernel	sigma = 0.01529346 C = 0.2
	Gradient boosted machines	n.trees = 100 interaction.depth = 4 shrinkage = 0.1 n.minobsinnode = 10
FVC < 80%	Random Forest	mtry = 8 splitrule = gini min.node.size = 3
	Neural network	size = 13 decay = 1e-04
	Support vector machines, radial kernel	sigma = 0.01529346 C = 0.9
	Gradient boosted machines	n.trees = 150 interaction.depth = 3 shrinkage = 0.1 n.minobsinnode = 15
FEV1 < 80%	Random Forest	mtry = 14 splitrule = gini min.node.size = 1
	Neural network	size = 11 decay = 1e-04
	Support vector machines, radial kernel	sigma = 0.01529346 C = 0.1
	Gradient boosted machines	n.trees = 200 interaction.depth = 1 shrinkage = 0.1 n.minobsinnode = 5
DLCO	Random Forest	mtry = 6

Response[a]	Algorithm	Parameters
		splitrule = variance min.node.size = 7
	Neural network	size = 13 decay = 1e-04
	Support vector machines, radial kernel	sigma = 0.01691549 C = 0.2
	Gradient boosted machines	n.trees = 100 interaction.depth = 1 shrinkage = 0.1 n.minobsinnode = 5
FVC	Random Forest	mtry = 2 splitrule = variance min.node.size = 5
	Neural network	size = 1 decay = 1e-04
	Support vector machines, radial kernel	sigma = 0.01691549 C = 0.1
	Gradient boosted machines	n.trees = 50 interaction.depth = 1 shrinkage = 0.1 n.minobsinnode = 5
FEV1	Random Forest	mtry = 2 splitrule = variance min.node.size = 3
	Neural network	size = 1 decay = 1e-05
	Support vector machines, radial kernel	sigma = 0.01691549 C = 0.1
	Gradient boosted machines	n.trees = 50 interaction.depth = 1 shrinkage = 0.1 n.minobsinnode = 5

[a]LFT: lung function testing; DLCO: diffusion capacity for carbon monoxide; FVC: forced vital capacity; FEV1: forced expiratory volume in one second.

Supplementary Table S7: Performance of binary machine learning classifiers at predicting lung function testing (LFT) abnormalities in the entire data set.

Response[a]	Algorithm[b]	Overall accuracy[c]	κ[d]	Brier score	AUC[e]	Sensitivity	Specificity
DLCO < 80%	Random Forest	1.00	1.00	0.014000	1.00	1.00	1.00
	Neural network	1.00	1.00	0.000043	1.00	1.00	1.00
	SVM radial	0.90	0.72	0.066000	0.96	0.77	0.94
	GBM	0.98	0.94	0.019000	1.00	0.94	0.99
FVC < 80%	Random Forest	0.99	0.96	0.024000	1.00	0.95	1.00
	Neural network	0.92	0.76	0.053000	0.95	0.82	0.95
	SVM radial	0.92	0.71	0.059000	0.97	0.63	0.99
	GBM	0.95	0.82	0.048000	0.98	0.78	0.99
FEV1 < 80%	Random Forest	1.00	1.00	0.016000	1.00	1.00	1.00
	Neural network	0.93	0.78	0.049000	0.96	0.86	0.95
	SVM radial	0.92	0.67	0.061000	0.98	0.56	1.00
	GBM	0.90	0.57	0.084000	0.91	0.47	0.99

[a]LFT: lung function testing, DLCO: diffusion capacity for carbon monoxide, FVC: forced vital capacity; FEV1: forced expiratory volume in one second.

[b]SVM: support vector machines with radial kernel; GBM: gradient boosted machines.

[c]Ratio of correct predictions to the total observation number.

[d]Cohen κ statistic of inter-rater reliability between the predicted and observed outcome.

[e]AUC: are under the receiver-operating characteristic curve.

Supplementary Table S8: Performance of regression machine learning models at predicting values of lung function testing parameters in the entire data set.

Response[a]	Algorithm[b]	Data set[c]	pseudo-R^{2}[d]	MAE[e]	ρ
DLCO	Random Forest	training	0.8700	5.1	0.95
	Neural network	training	0.3600	12.0	0.54
	SVM radial	training	0.5500	9.4	0.77
	GBM	training	0.6200	9.3	0.74
FVC	Random Forest	training	0.6900	5.8	0.91
	Neural network	training	0.0024	10.0	
	SVM radial	training	0.2200	8.9	0.57
	GBM	training	0.2800	8.9	0.50
FEV1	Random Forest	training	0.7400	5.9	0.94
	Neural network	training	0.0024	12.0	
	SVM radial	training	0.2100	9.8	0.59
	GBM	training	0.3000	9.9	0.54

[a]DLCO: diffusion capacity for carbon monoxide, FVC: forced vital capacity; FEV1: forced expiratory volume in one second.

[b]SVM: support vector machines with radial kernel; GBM: gradient boosted machines.

[c]Defined as 1 - ratio of mean squared error and variance.

[d]MAE: mean absolute error.

[e]P: Spearman coefficient of correlation between the predicted and observed response values.

Supplementary Table S9: Mean values of SHAP (Shapley additive explanations) variable importance statistic for the models of reduced diffusion capacity for carbon monoxide and of diffusion capacity for carbon monoxide < 80% of reference. The table is available as a supplementary Excel file.

Supplementary Table S10: Differences in chest computed tomography severity score (CTSS), and AI-determined lung opacity and high opacity in CovILD study participants with and without lung function testing (LFT) abnormalities. Numeric variables are presented as medians with interquartile ranges (IQR) and ranges.

LFT abnormality[a]	Variable[b]	Abnormality absent	Abnormality present	Significance[c]	Effect size[c]
DLCO < 80%	Observations, n	326	94		
	CTSS, points	1 [IQR: 0 - 5] range: 0 - 20 complete: n = 326	10 [IQR: 4 - 15] range: 0 - 20 complete: n = 94	p < 0.001	r = 0.57
	opacity, AI, % of lung	0.01 [IQR: 0 - 0.31] range: 0 - 22 complete: n = 326	1.3 [IQR: 0.19 - 5.3] range: 0 - 37 complete: n = 94	p < 0.001	r = 0.63
	high opacity, AI, % of lung	0 [IQR: 0 - 0.01] range: 0 - 2.4 complete: n = 326	0.063 [IQR: 0.0075 - 0.38] range: 0 - 3.1 complete: n = 94	p < 0.001	r = 0.58
FVC < 80%	Observations, n	337	83		
	CTSS, points	2 [IQR: 0 - 7] range: 0 - 20 complete: n = 337	5 [IQR: 0 - 12] range: 0 - 20 complete: n = 83	ns (p = 0.092)	r = 0.21
	opacity, AI, % of lung	0.03 [IQR: 0 - 0.53] range: 0 - 31 complete: n = 337	0.43 [IQR: 5e-04 - 2.7] range: 0 - 37 complete: n = 83	p = 0.0051	r = 0.29
	high opacity, AI, % of lung	0 [IQR: 0 - 0.017] range: 0 - 2 complete: n = 337	0.01 [IQR: 0 - 0.19] range: 0 - 3.1 complete: n = 83	ns (p = 0.062)	r = 0.3
FEV1 < 80%	Observations, n	343	77		
	CTSS, points	2 [IQR: 0 - 7] range: 0 - 20 complete: n = 343	4 [IQR: 0 - 12] range: 0 - 20 complete: n = 77	ns (p = 0.11)	r = 0.2
	opacity, AI, % of lung	0.03 [IQR: 0 - 0.55] range: 0 - 31 complete: n = 343	0.33 [IQR: 0.001 - 2.5] range: 0 - 37 complete: n = 77	p = 0.009	r = 0.27
	high opacity, AI, % of lung	0 [IQR: 0 - 0.017] range: 0 - 3.1 complete: n = 343	0.016 [IQR: 0 - 0.2] range: 0 - 2.5 complete: n = 77	ns (p = 0.062)	r = 0.3

LFT abnormality[a]	Variable[b]	Abnormality absent	Abnormality present	Significance[c]	Effect size[c]

[a]LFT: lung function testing; DLCO: diffusion capacity for carbon monoxide; FVC: forced vital capacity; FEV1: forced expiratory volume in one second.

[b]CTSS: CT severity score determined by a human radiologist, sum for all lobes; opacity and high opacity, AI: opacity and high opacity of the lungs determined by artificial intelligence.

[c]Blocked bootstrap test with biseral r effect size statistic. P values corrected for multiple testing with the false discovery rate method.

Supplementary Table S11: Correlation of LFT variables with chest computed tomography severity score, and AI-determined opacity and high opacity.

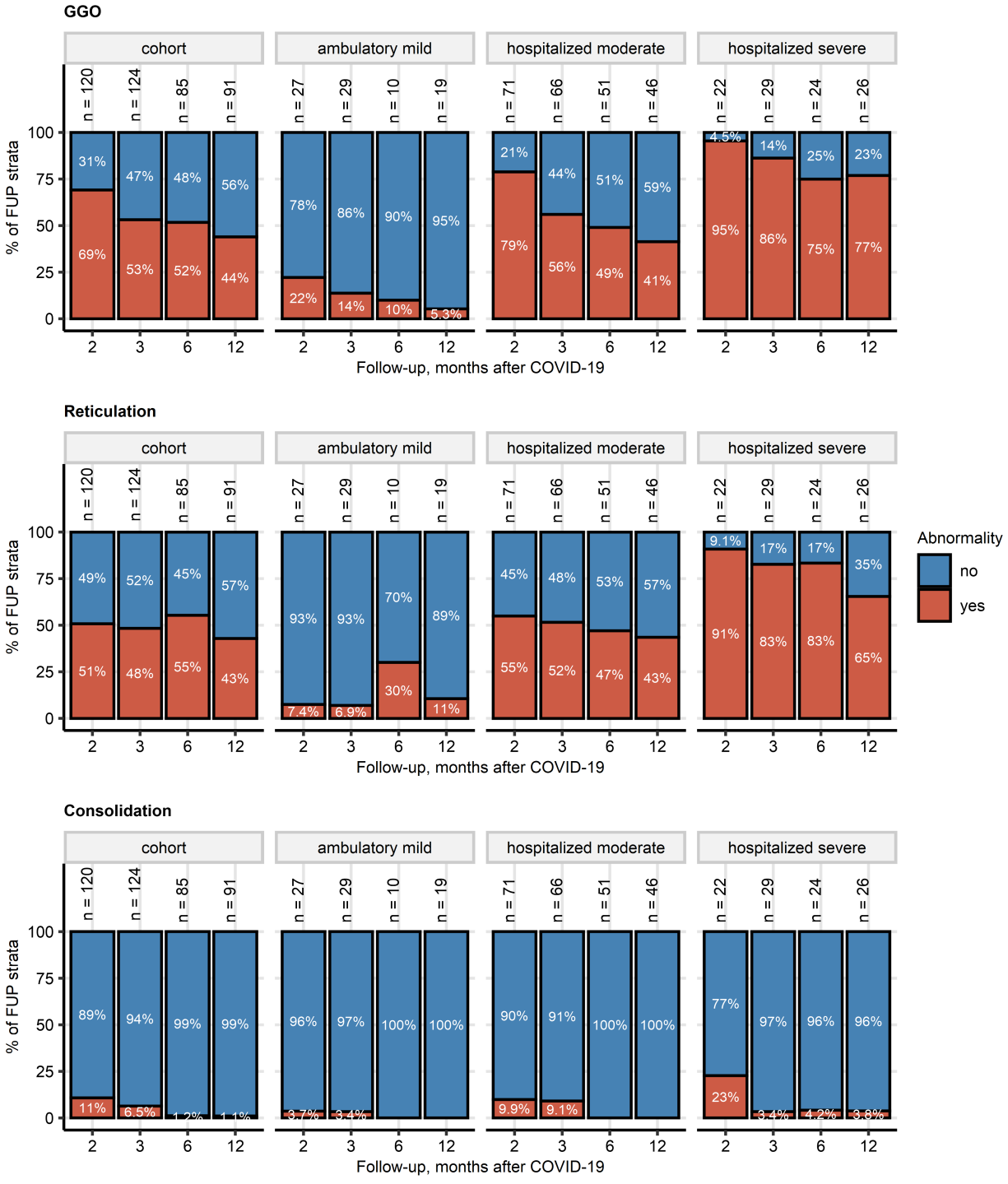
LFT variable[a]	CT variable[b]	N	Correlation coeffcient[c]	Significance[c]
DLCO	CTSS	420	ρ = -0.44 [-0.56 - -0.3]	p < 0.001
	opacity, AI	420	ρ = -0.46 [-0.57 - -0.33]	p < 0.001
	high opacity, AI	420	ρ = -0.46 [-0.56 - -0.36]	p < 0.001
FVC	CTSS	420	ρ = -0.21 [-0.34 - -0.074]	p = 0.0023
	opacity, AI	420	ρ = -0.27 [-0.39 - -0.13]	p < 0.001
	high opacity, AI	420	ρ = -0.3 [-0.42 - -0.18]	p < 0.001
FEV1	CTSS	420	ρ = -0.14 [-0.28 - 0.005]	p = 0.029
	opacity, AI	420	ρ = -0.21 [-0.34 - -0.067]	p = 0.0019
	high opacity, AI	420	ρ = -0.27 [-0.39 - -0.14]	p < 0.001

[a]DLCO: diffusion capacity for carbon monoxide; FVC: forced vital capacity; FEV1: forced expiratory volume in one second.

[b]CTSS: CT severity score determined by a human radiologist, sum for al lobes; opacity and high opacity, AI: opacity and high opacity of the lungs determined by artificial intelligence.

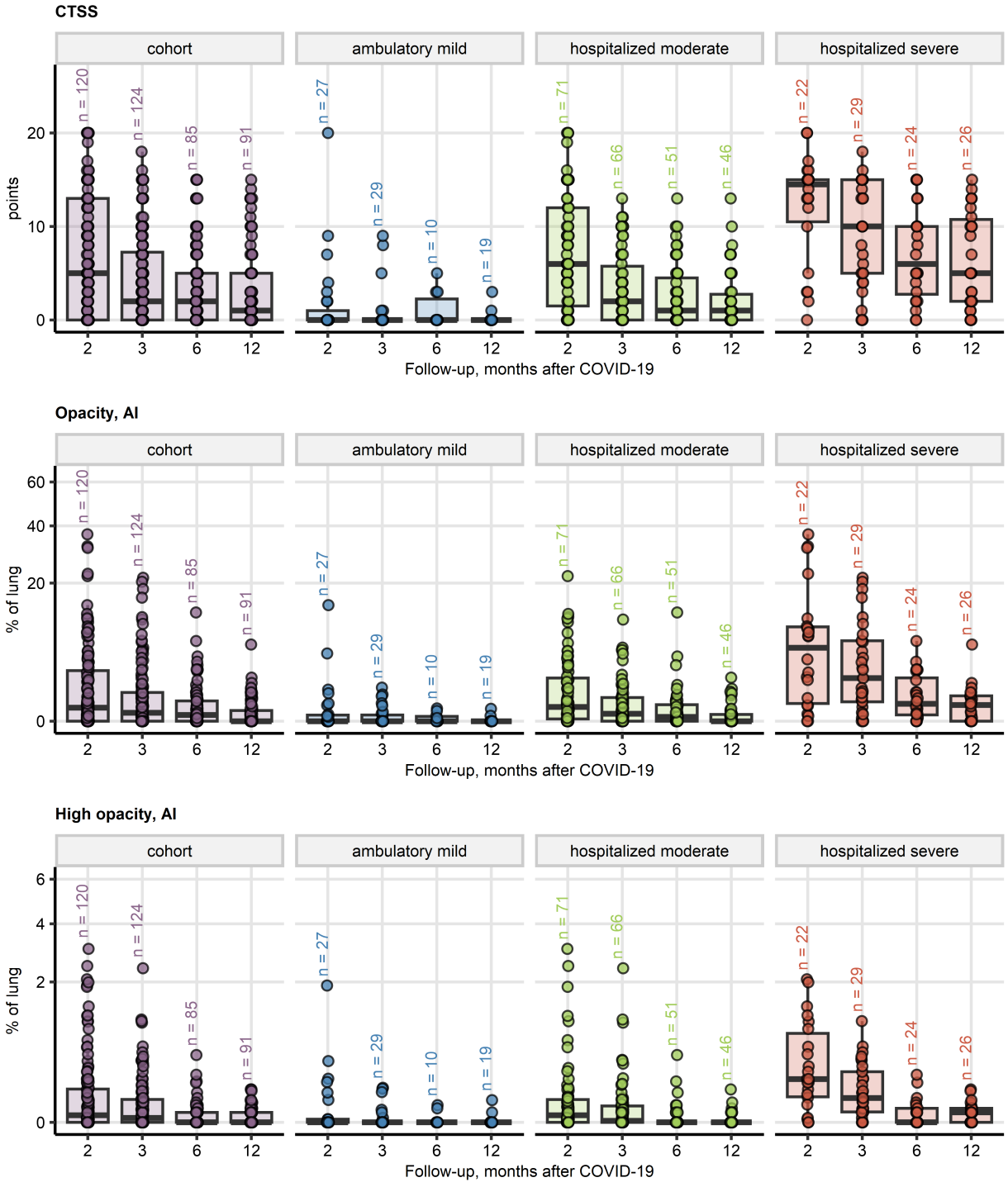
[c]Blocked bootstrap Spearman correlation test; ρ correlation coefficients with 95% confidence intervals. P values were corrected for multiple testing with the false discovery rate method.

Supplementary Figures



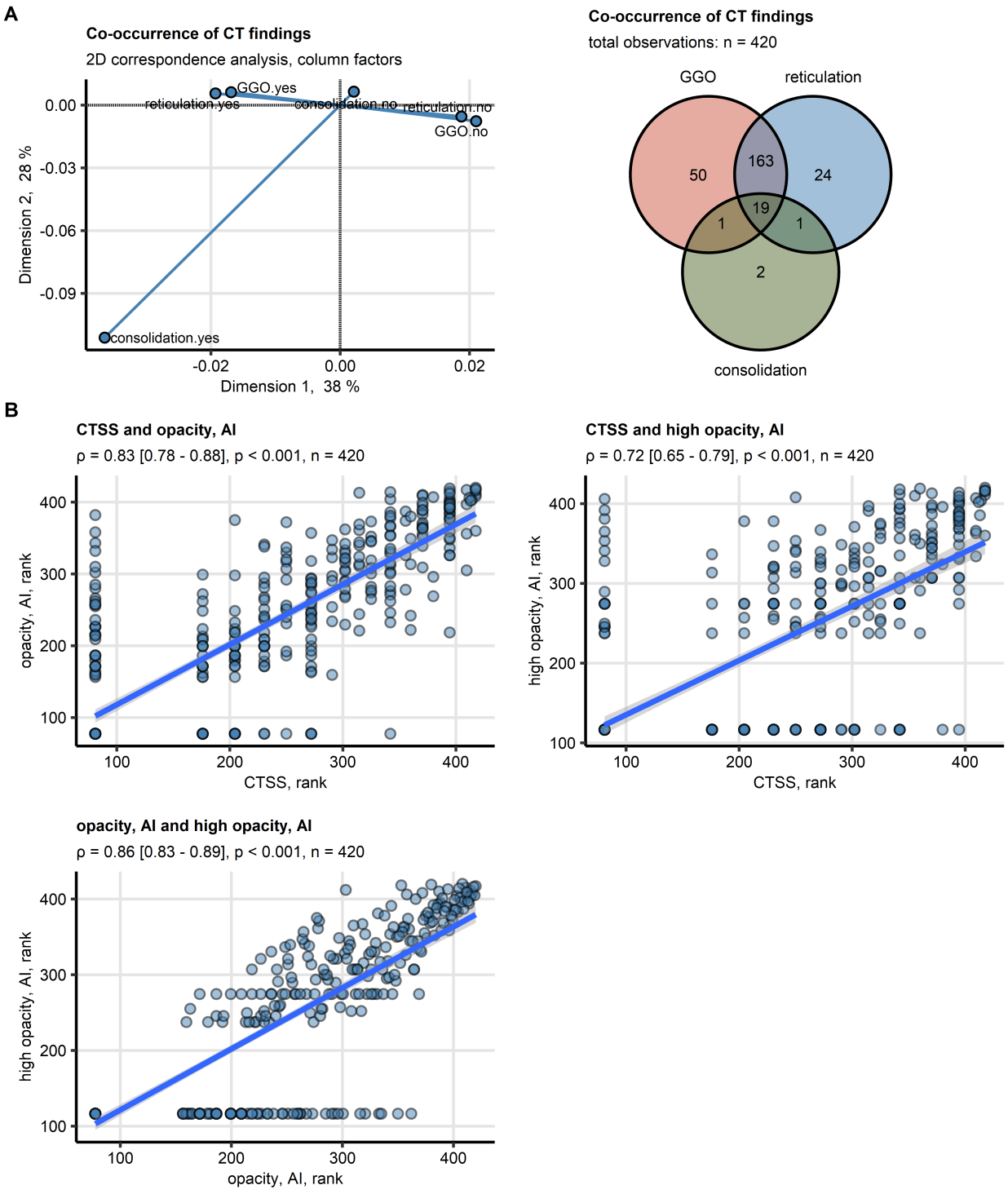
Supplementary Figure S1. Frequency of the most common abnormalities in computed tomography of the chest at the consecutive follow-ups in participants stratified by severity of acute COVID-19.

Frequencies of any chest computed tomography (CT) abnormalities, ground glass opacity (GGO), and reticulations identified by a human radiologist at the consecutive follow-ups after COVID-19 in the CovILD study patients stratified by the severity of acute COVID-19 were presented in stack plots. Numbers of complete observations at the follow-up examinations are displayed above the data bars.



Supplementary Figure S2. Time course of numeric computed tomography readouts at the consecutive follow-ups in participants stratified by severity of acute COVID-19.

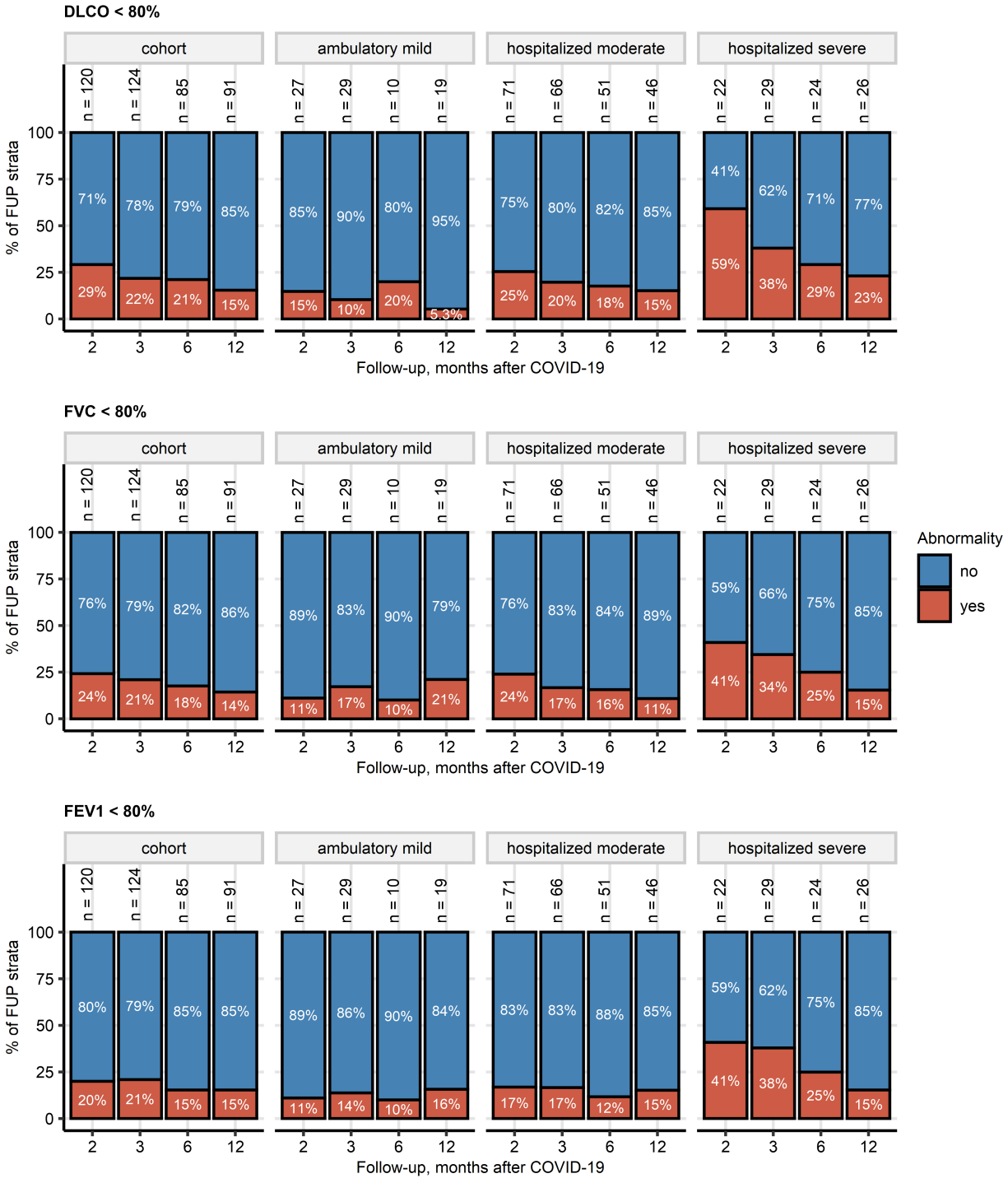
Severity of chest computed tomography (CT) abnormalities was assessed by a human-determined CT severity score (CTSS), and AI-determined percentages of the lung tissue with opacity and high opacity. Time courses of these parameters are visualized in box plots. Boxes represent medians with interquartile ranges, whiskers span over 150% of the interquartile range. Single observations are depicted as points. Numbers of complete observations at the follow-up examinations are displayed above the data points.



Supplementary Figure S3. Co-occurrence of abnormalities of chest computed tomography and correlation of computed tomography readouts in COVID-19 convalescents.

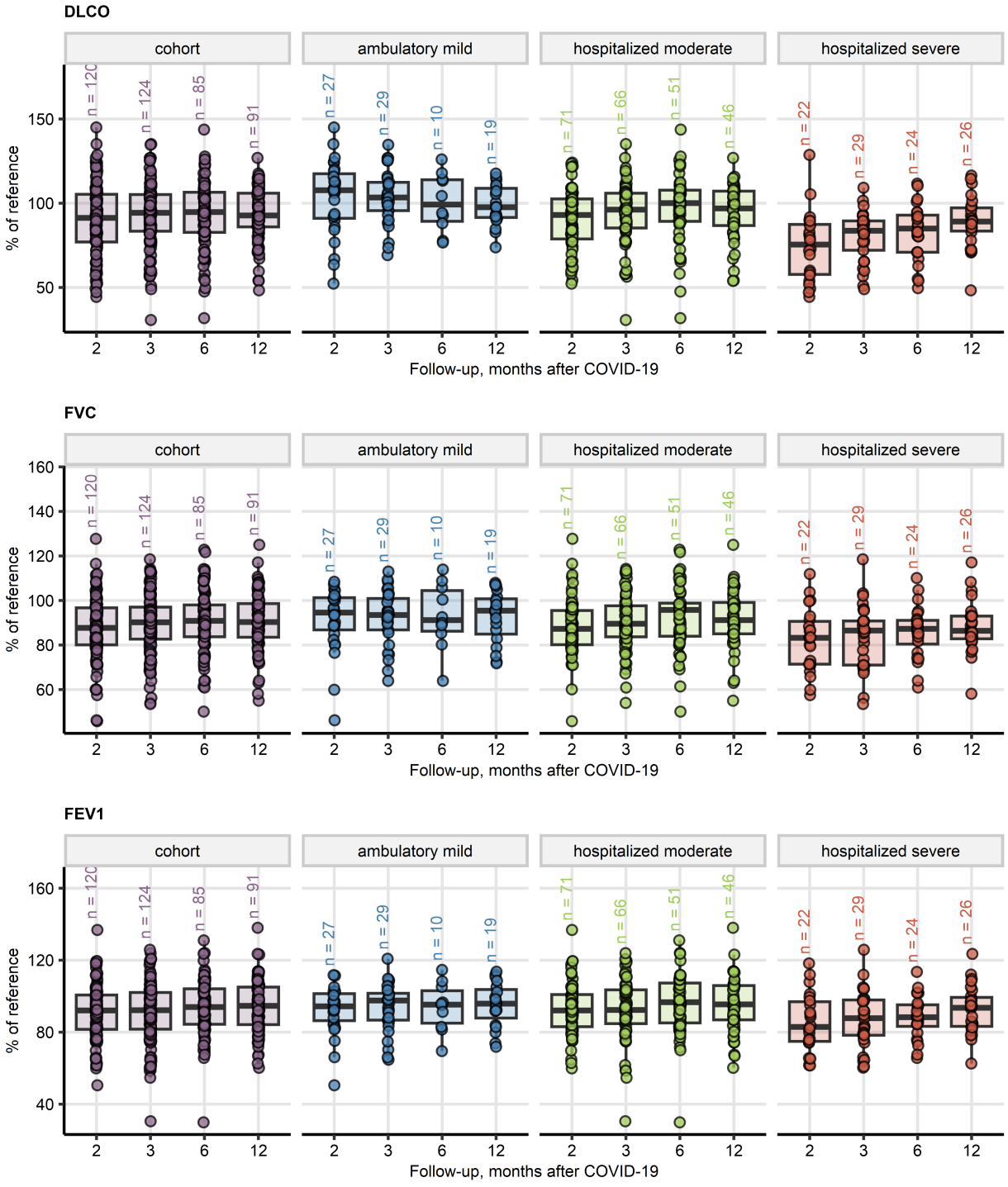
(A) Co-occurrence of ground glass opacity (GGO), reticulation, and consolidation diagnosed by a human radiologist was investigated by two-dimensional correspondence analysis. Left: plots of the column factors, proximity of points reflects co-occurrence of normal or pathological values of lung function testing. Right: Venn plot of frequencies of GGO, reticulation, and consolidation; the total observation number is displayed in the plot caption.

(B) Correlation between human-determined (CTSS: computed tomography severity score, sum of points form the entire lung) and artificial intelligence determined (AI, opacity and high opacity, percentage of the lung) readouts of severity of structural lung lesions was assessed by blocked bootstrap Spearman's test and visualized in scatter plots of observation ranks. Each point represents a single observation, linear trends with standard errors are depicted as blue lines with gray ribbons. Values of correlation coefficients with 95% confidence intervals and p values are displayed in the plot captions.



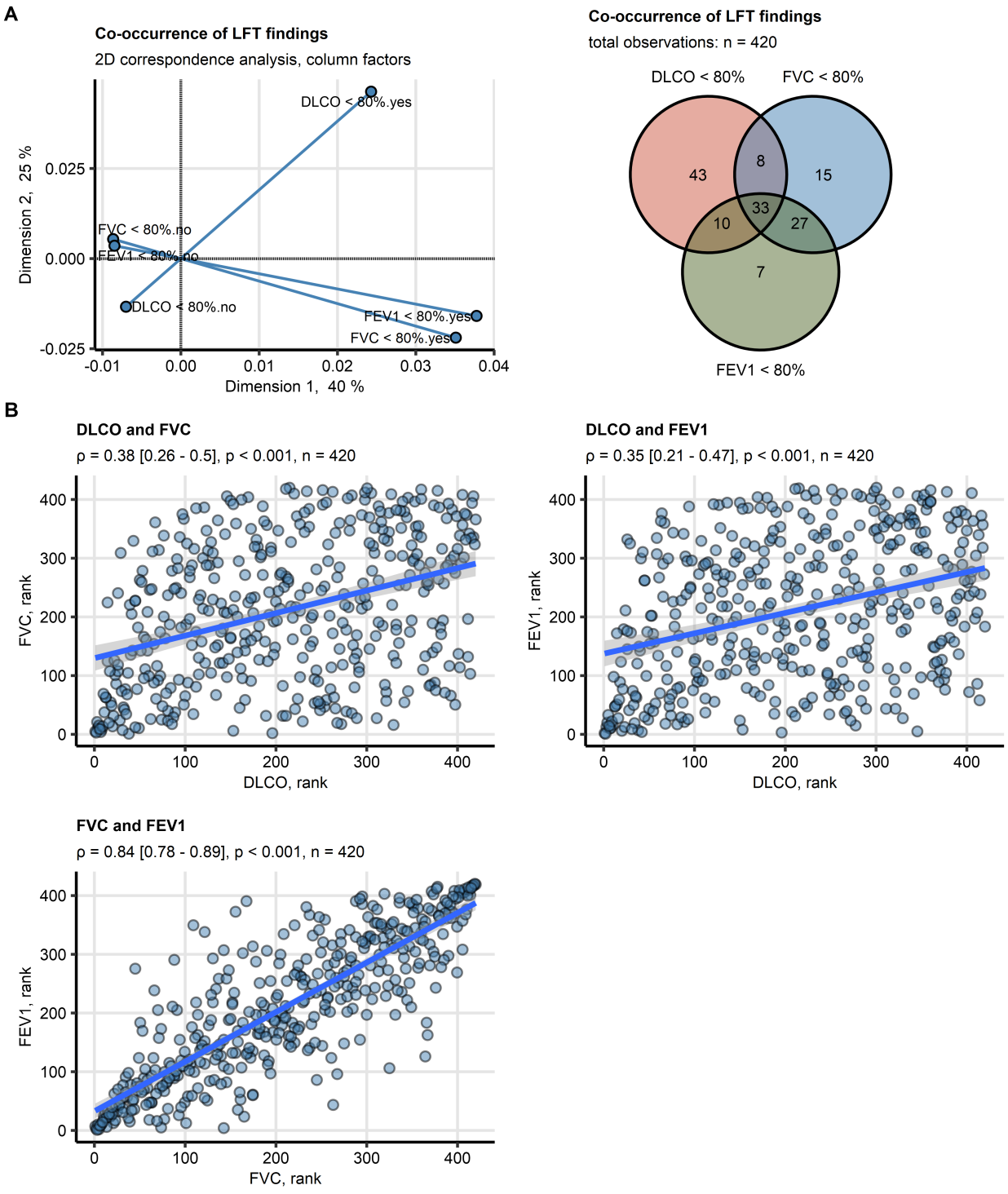
Supplementary Figure S4. Frequency of lung function testing abnormalities at the consecutive follow-ups in participants stratified by severity of acute COVID-19.

Insufficient diffusion capacity for carbon monoxide (DLCO), forced vital capacity (FVC), and forced expiratory volume in one second (FEV1) was defined as values < 80% of the individual's reference. Frequencies of these abnormalities in in the CovILD study patients stratified by the severity of acute COVID-19 were presented in stack plots. Numbers of complete observations at the follow-up examinations are displayed above the data bars.



Supplementary Figure S5. Time course of numeric lung function testing readouts at the consecutive follow-ups in participants stratified by severity of acute COVID-19.

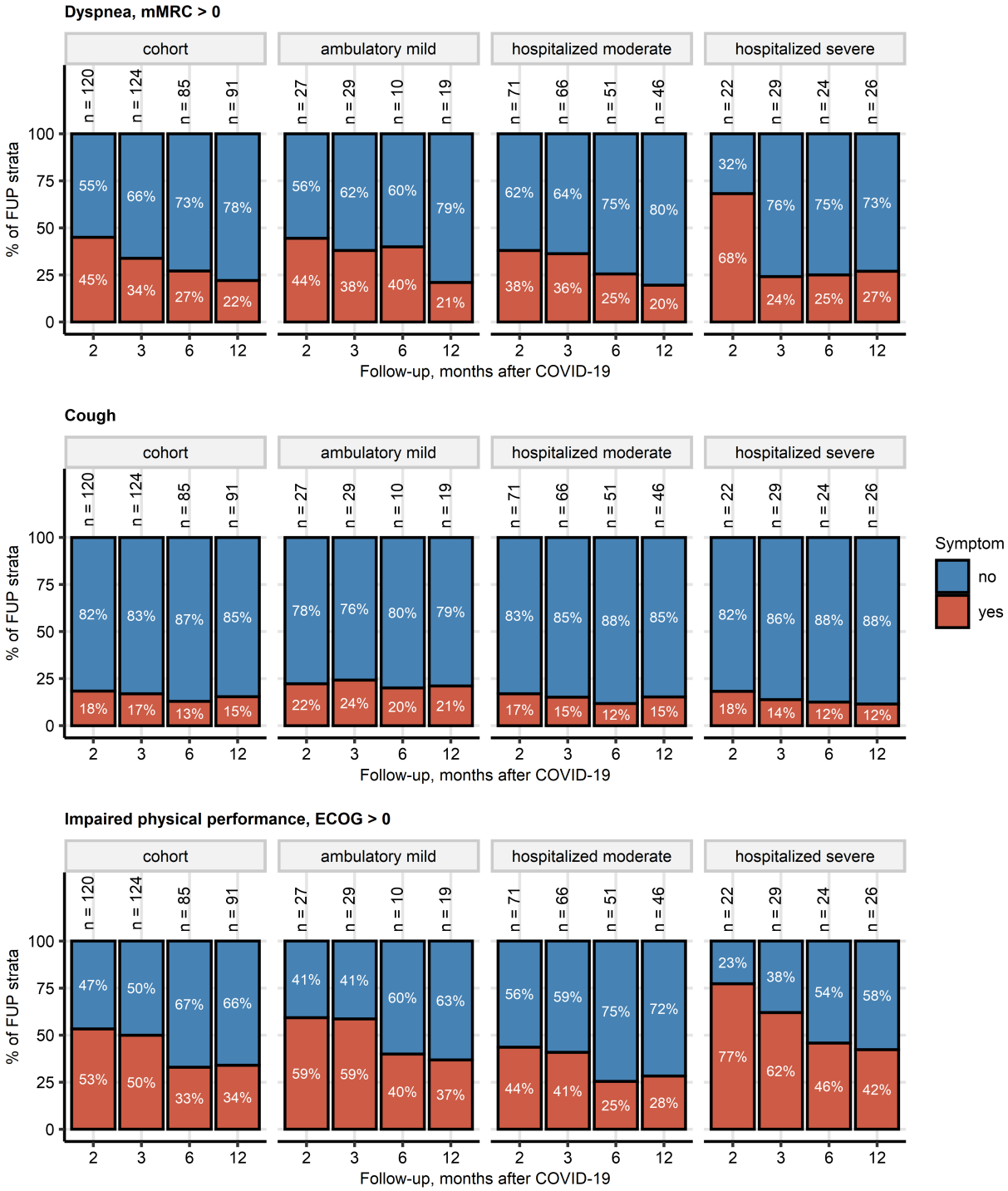
Diffusion capacity for carbon monoxide (DLCO), forced vital capacity (FVC), forced expiratory volume in one second (FEV1) were recorded a percentages of the patient's reference values. Time courses of these parameters are visualized in box plots. Boxes represent medians with interquartile ranges, whiskers span over 150% of the interquartile range. Single observations are depicted as points. Numbers of complete observations at the follow-up examinations are displayed above the data points.



Supplementary Figure S6. Co-occurrence of abnormalities of lung function testing and correlation of lung function readouts in COVID-19 convalescents.

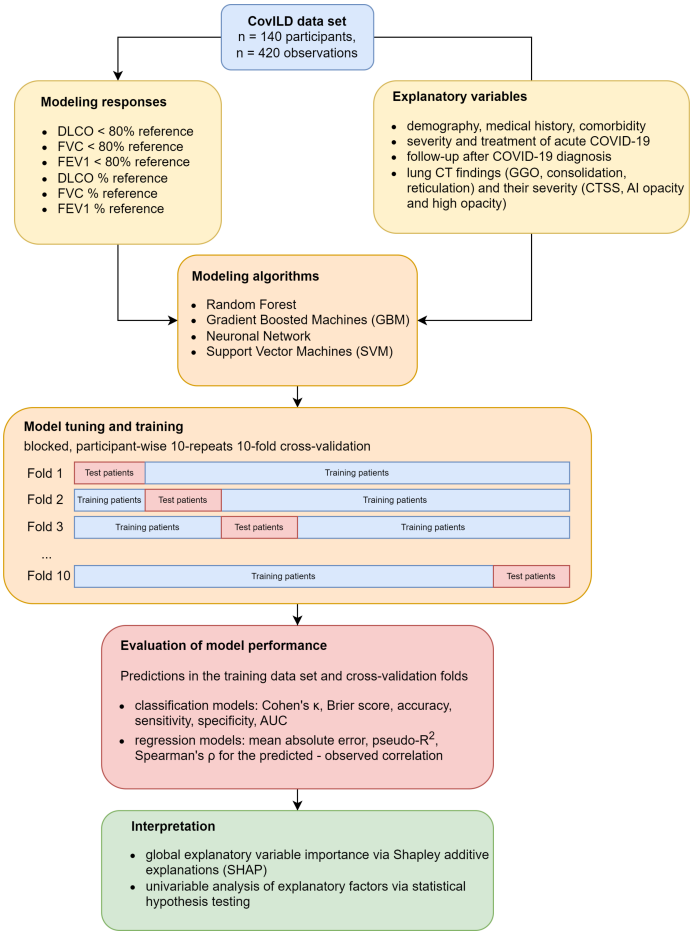
(A) Co-occurrence of insufficient diffusion capacity for carbon monoxide (DLCO < 80% of reference value), insufficient forced vital capacity (FVC < 80% of reference), and insufficient forced expiratory volume in one second (FEV1 < 80% of reference) was investigated by two-dimensional correspondence analysis. Left: plots of the column factors, proximity of points reflects co-occurrence of normal or pathological values of lung function testing. Right: Venn plot of frequencies of insufficient DLCO, FVC, and FEV1; the total observation number is displayed in the plot caption.

(B) Correlation between the lung function testing parameters was assessed by blocked bootstrap Spearman's test and visualized in scatter plots of observation ranks. Each point represents a single observation, linear trends with standard errors are depicted as blue lines with gray ribbons. Values of correlation coefficients with 95% confidence intervals and p values are displayed in the plot captions.



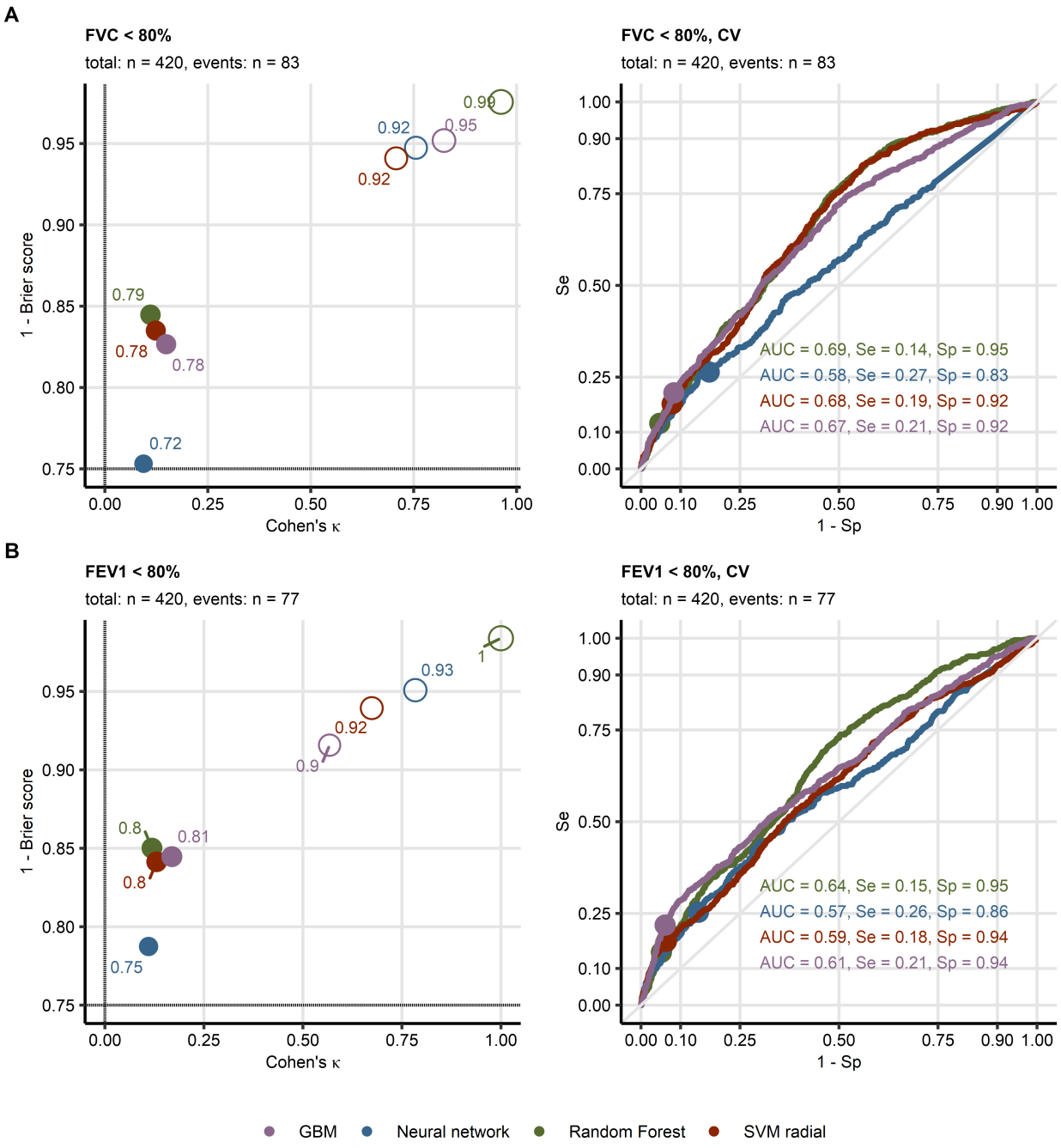
Supplementary Figure S7. Frequency of persistent symptoms of relevance for lung function at the consecutive follow-ups in participants stratified by severity of acute COVID-19.

Intensity of self-reported dyspnea was assessed with the modified Medical Research Council scale (mMRC) and presence of dyspnea was defined as mMRC > 0. Presence of cough was surveyed by a single 'yes'/'no' item. Impairment of physical performance was assessed with the Eastern Cooperative Oncology Group scale (ECOG, high values indicate more profound impairment) and impaired physical performance was defined as ECOG > 0. Frequency of dyspnea, cough and impairment of physical performance at the consecutive follow-ups after COVID-19 in the CovILD cohort stratified by the severity of acute COVID-19 was presented in stacked bar plots. Numbers of complete observations at the follow-up examinations are displayed above the data bars.



Supplementary Figure S8. Modeling strategy.

DLCO: diffusion capacity for carbon dioxide; FVC: forced vital capacity; FEV1: forced expiratory volume in 1 second; CT: computed tomography; GGO: ground glass opacity; CTSS: human-determined CT severity score; AI opacity and high opacity: artificial intelligence-determined opacity and high opacity of the lung; AUC: area under the receiver-operating characteristic curve.

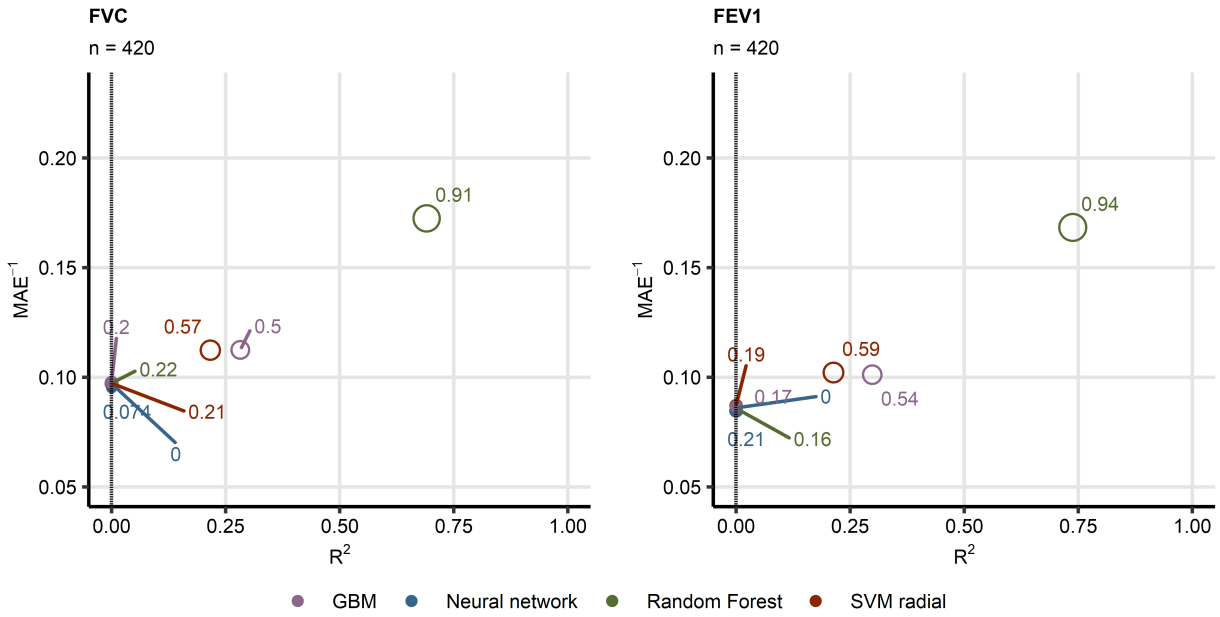


Supplementary Figure S9. Evaluation of performance of machine learning classification models at prediction of insufficient forced vital capacity and forced expiratory volume in one second.

Machine learning classification models of insufficient forced vital capacity (A, FVC < 80% of reference value: n = 83, total observations: n = 420) and of insufficient forced expiratory

volume in one second (B, FEV1 < 80% of reference value: n = 77, total observations: n = 420) employing computed tomography readouts, demographic and clinical explanatory variables were trained. The model performance was evaluated in the entire data set and 10-repeats 10-fold cross-validation with overall accuracy metric, Cohen's κ as a measure of concordance between predicted and observed outcome, and Brier score as a measure of model's calibration. Left: numeric performance measures of the models (open circles: the entire data set, filled circles: cross-validation); point sizes and point labels represent overall model accuracy, the dashed lines visualize values of Cohen's κ and Brier score expected for prediction of insufficient DLCO be chance. Right: receiver-operating characteristic curves for predictions in cross-validation folds, numeric statistics are displayed in the plot. Numbers of complete observations and observations with the particular LFT insufficiency ('events') are displayed in the plot captions.

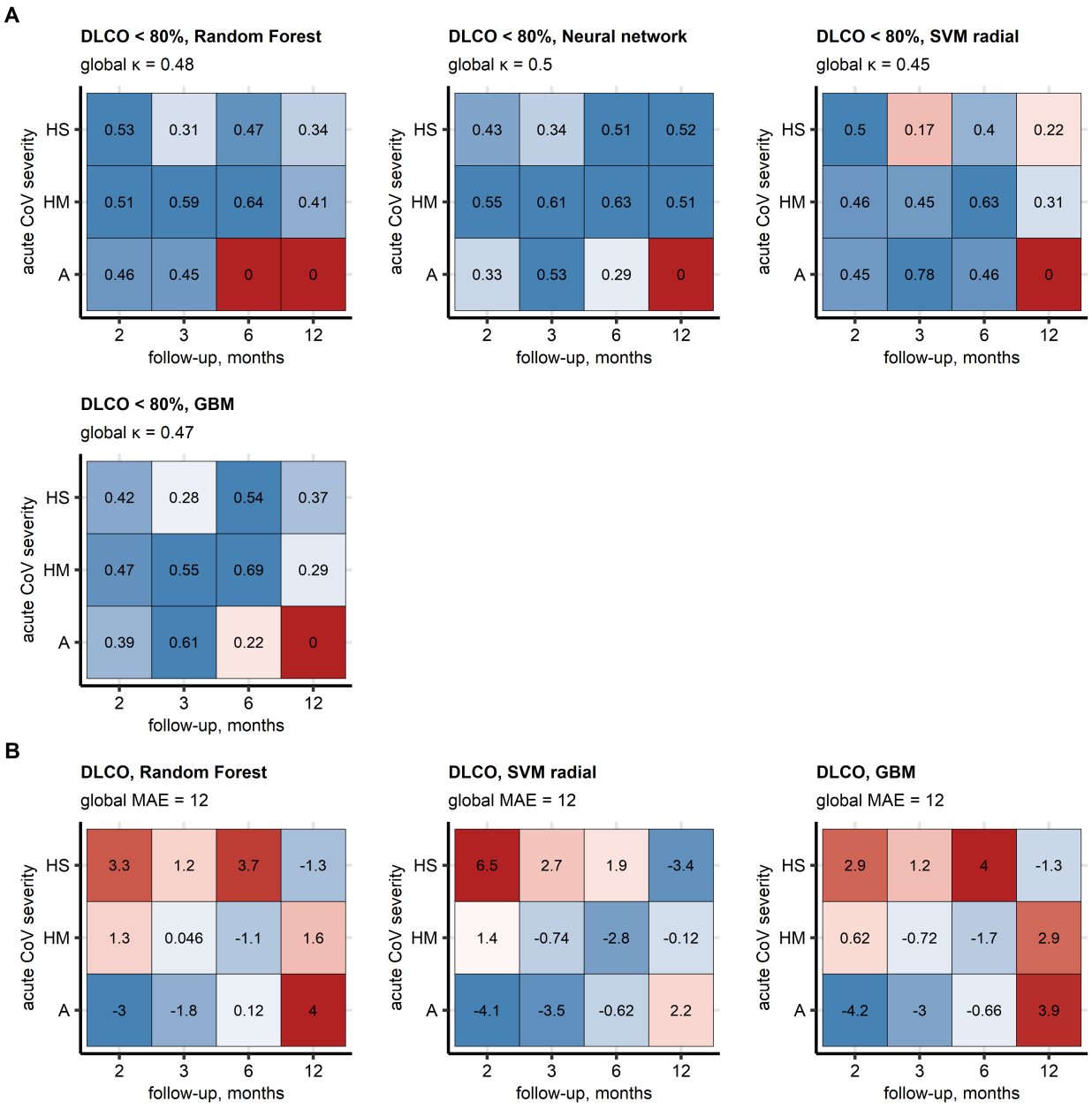
FVC: forced vital capacity; FEV1: forced expiratory volume in one second, CV: cross-validation; AUC: are under the receiver-operating characteristic curve; Se: sensitivity; Sp: specificity; GBM: gradient boosted machines; SVM radial: support vector machines with radial kernel.



Supplementary Figure S10. Evaluation of performance of machine learning regression models at prediction of forced vital capacity and forced expiratory volume in one second.

Machine learning regression models of forced vital capacity (FVC, total observations: n = 420) and of forced expiratory volume in one second (FEV1, total observations: n = 420) employing computed tomography readouts, demographic and clinical explanatory variables were trained. Their performance was evaluated in th entire data set and 10-repeats 10-fold cross-validation with R^2 as a measure of explained variation, mean absolute error, and ρ Spearman's coefficient of correlation between the predicted and observed values. Numeric performance measures of the model are presented in bubble plots (open circles: the entire data set, filled circles: cross-validation); point sizes and point labels represent values of ρ correlation coefficient, the dashed line visualizes R^2 value expected for a meaningless model. Numbers of complete observations are shown in the plot captions.

FVC: forced vital capacity; FEV1: forced expiratory volume in one second, CV: cross-validation; MAE: mean absolute error; GBM: gradient boosted machines; SVM radial: support vector machines with radial kernel.

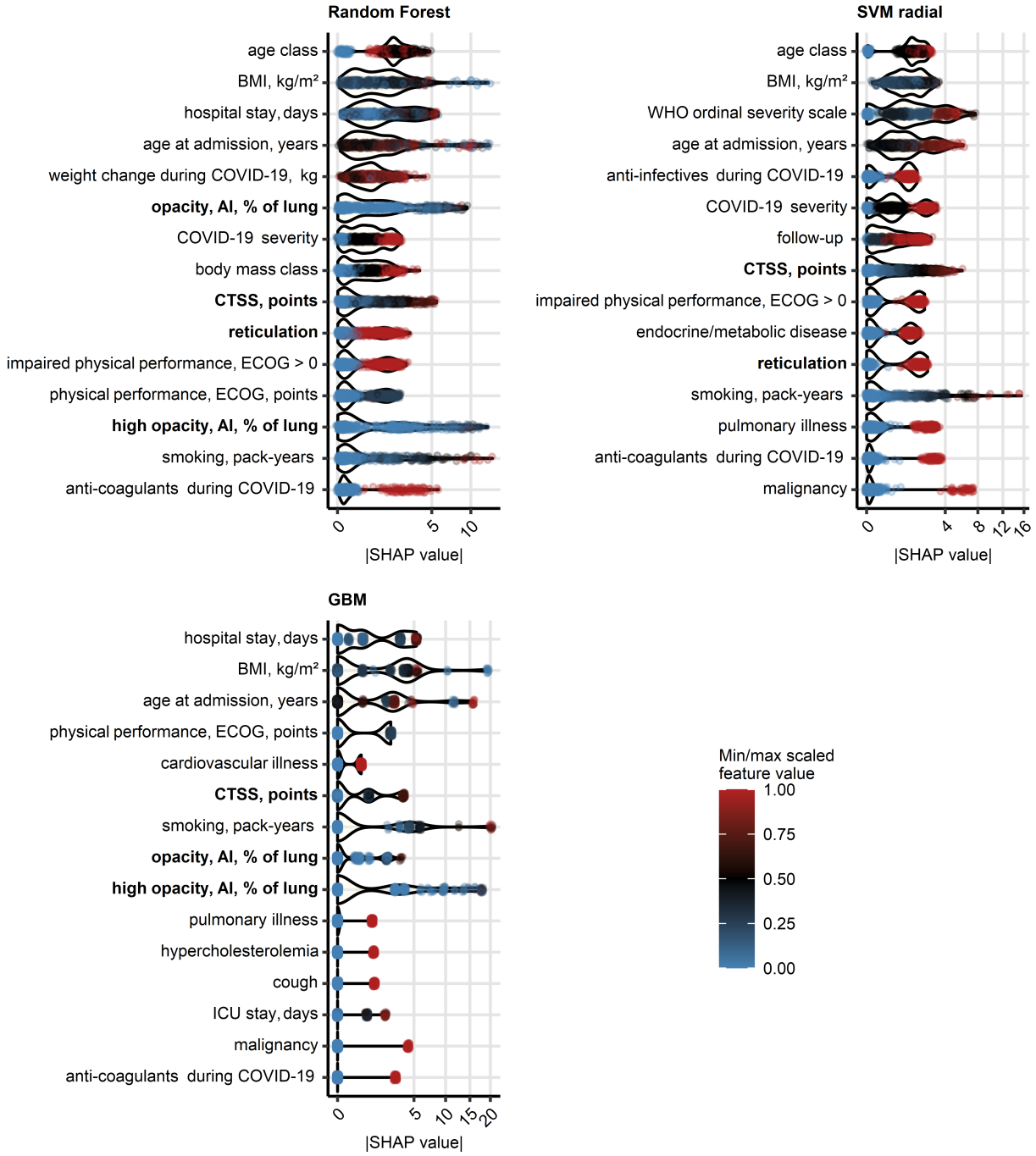


Supplementary Figure S11. Cohen's κ and mean error of the machine learning models of insufficiency and percentage of reference of diffusion capacity for carbon monoxide in observations stratified by acute COVID-19 severity and follow-up visit.

Classification machine learning models of insufficient diffusion capacity for carbon monoxide (A, DLCO < 80% of reference value: n = 94, total observations: n = 420) and regression models of diffusion capacity for carbon monoxide (DLCO, % of reference value, total observations: n = 420) were trained and evaluated as presented in Figure 2. Cohen's κ (A) and mean errors (B) of cross-validated model predictions for observations stratified by severity of acute COVID-19

and the consecutive follow-up visits were investigated. Cohen's κ and error values are visualized in heat maps; tiles are labeled with the error values. Global κ and mean absolute model errors are indicated in the plot captions.

DLCO: diffusion capacity for carbon monoxide; SVM radial: support vector machines with radial kernel; GBM: gradient boosted machines; MAE: mean absolute error; A: ambulatory COVID-19; HM: hospitalized, moderate COVID-19; HS: hospitalized, severe COVID-19..



Supplementary Figure S12. Explanatory variable importance for models of diffusion capacity for carbon monoxide measured by Shapley additive explanations.

Importance of explanatory variables for the machine learning models of diffusion capacity for carbon monoxide (% of reference, Figure 2) was investigated by Shapley additive explanations (SHAP). Absolute SHAP values for explanatory variables with the 15 largest

mean SHAP values are presented in violin plots. Points represent single observations, point colors code for minimum/maximum scaled value of the explanatory variable. Explanatory variables obtained via computed tomography are highlighted with bold font in the Y axes.

CT: computed tomography; DLCO: diffusion capacity for carbon monoxide; BMI: body mass index; opacity and high opacity, AI: opacity and high opacity of the lung determined by artificial intelligence; CTSS: human-determined CT severity score, sum for all lobe; ECOG: Eastern Cooperative Oncology Group physical performance score; mMRC: modified Medical Research Council dyspnea scale; ICU: intensive care unit.

References

1. Sonnweber T, Sahanic S, Pizzini A, Luger A, Schwabl C, Sonnweber B, Kurz K, Koppelstätter S, Haschka D, Petzer V, et al. Cardiopulmonary recovery after COVID-19: An observational prospective multicentre trial. *European Respiratory Journal* (2021) 57: doi: 10.1183/13993003.03481-2020

2. Sonnweber T, Tymoszuk P, Sahanic S, Boehm A, Pizzini A, Luger A, Schwabl C, Nairz M, Grubwieser P, Kurz K, et al. Investigating phenotypes of pulmonary COVID-19 recovery: A longitudinal observational prospective multicenter trial. *eLife* (2022) 11: doi: 10.7554/ELIFE.72500

3. Luger AK, Sonnweber T, Gruber L, Schwabl C, Cima K, Tymoszuk P, Gerstner AK, Pizzini A, Sahanic S, Boehm A, et al. Chest CT of Lung Injury 1 Year after COVID-19 Pneumonia: The CovILD Study. *Radiology* (2022) 304:462–470. doi: 10.1148/radiol.211670

4. Sahanic S, Tymoszuk P, Luger AK, Hüfner K, Boehm A, Pizzini A, Schwabl C, Koppelstätter S, Kurz K, Asshoff M, et al. COVID-19 and its continuing burden after 12 months: a longitudinal observational prospective multicentre trial. *ERJ open research* (2023) 9:00317–2022. doi: 10.1183/23120541.00317-2022

5. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Müller NL, Remy J. Fleischner Society: Glossary of terms for thoracic imaging. (2008) 246:697–722. doi: 10.1148/radiol.2462070712

6. Wickham Hadley. *ggplot2: Elegant Graphics for Data Analysis*. 1st ed. New York: Springer-Verlag (2016). https://ggplot2.tidyverse.org

7. Henry L, Wickham Hadley. rlang: Functions for Base Types and Core R and 'Tidyverse' Features. (2022) https://cran.r-project.org/web/packages/rlang/index.html

8. Gagolewski M, Tartanus B. Package 'stringi'. (2021) https://cran.r-project.org/web/packages/stringi/index.html http://cran.ism.ac.jp/web/packages/stringi/stringi.pdf

9. Vaughan D, Dancho M, RStudio. furrr: Apply Mapping Functions in Parallel using Futures. (2022) https://cran.r-project.org/package=furrr

10. Folashade D, Microsoft Corporation, Weston S, Tenenbaum D. doParallel: Foreach Parallel Adaptor for the 'parallel' Package. (2022) https://cran.r-project.org/web/packages/doParallel/index.html

11. Kassambara A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. (2021) https://cran.r-project.org/package=rstatix

12. Mangiafico S. rcompanion: Functions to Support Extension Education Program Evaluation. (2022) https://cran.r-project.org/package=rcompanion

13. Ripley B. MASS: Support Functions and Datasets for Venables and Ripley's MASS. (2022) https://cran.r-project.org/package=MASS

14. Kuhn M. Building predictive models in R using the caret package. *Journal of Statistical Software* (2008) 28:1–26. doi: 10.18637/jss.v028.i05

15. López-Ratón M, Rodríguez-Álvarez MX, Cadarso-Suárez C, Gude-Sampedro F. Optimalcutpoints: An R package for selecting optimal cutpoints in diagnostic tests. *Journal of Statistical Software* (2014) 61:1–36. doi: 10.18637/jss.v061.i08

16. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* (2017) 77:1–17. doi: 10.18637/JSS.V077.I01

17. Breiman L. Random forests. *Machine Learning* (2001) 45:5–32. doi: 10.1023/A:1010933404324

18. Ripley BD. *Pattern recognition and neural networks*. Cambridge University Press (2014). doi: 10.1017/CBO9780511812651

19. Weston J, Watkins C. Multi-Class Support Vector Machines. (1998)

20. Karatzoglou A, Hornik K, Smola A, Zeileis A. kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistical Software* (2004) 11:1–20. doi: 10.18637/JSS.V011.I09

21. Greenwell B, Boehmke B, Cunningham J, Developers G. gbm: Generalized Boosted Regression Models. (2022) https://cran.r-project.org/package=gbm

22. Friedman JH. Stochastic gradient boosting. *Computational Statistics & Data Analysis* (2002) 38:367–378. doi: 10.1016/S0167-9473(01)00065-2

23. Friedman JH. Greedy function approximation: A gradient boosting machine. *https://doiorg/101214/aos/1013203451* (2001) 29:1189–1232. doi: 10.1214/AOS/1013203451

24. Covert I, Lee SI. Improving KernelSHAP: Practical Shapley Value Estimation via Linear Regression. *Proceedings of Machine Learning Research* (2020) 130:3457–3465. https://arxiv.org/abs/2012.01536v3

25. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* (2017) 2017-Decem:4766–4775. https://arxiv.org/abs/1705.07874v2

26. Mayer M, Stando A. shapviz: SHAP Visualizations. (2023) https://cran.r-project.org/web/packages/shapviz/index.html

27. Yan L. ggvenn: Draw Venn Diagram by 'ggplot2'. (2021) https://cran.r-project.org/package=ggvenn

28. Sachs MC. Plotroc: A tool for plotting ROC curves. *Journal of Statistical Software* (2017) 79:1–19. doi: 10.18637/jss.v079.c02

29. Mayer M, Watson D, Biecek P. kernelshap: Kernel SHAP. (2023) https://cran.r-project.org/web/packages/kernelshap/index.html

30. Gohel D. flextable: Functions for Tabular Reporting. (2022) https://cran.r-project.org/web/packages/flextable/index.html

31. Wilke CO. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. 1st ed. Sebastopol: O'Reilly Media (2019).

32. Allaire J, Xie Y, McPherson J, Luraschi J, Ushey K, Atkins A, Wickham H, Cheng J. rmarkdown: Dynamic Documents for R. (2022) https://cran.r-project.org/web/packages/rmarkdown/index.html

33. Xie Y. *Bookdown: Authoring books and technical documents with R Markdown*. (2016). doi: 10.1201/9781315204963

34. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* (1995) 57:289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

35. McHugh ML. Interrater reliability: the kappa statistic. *Biochemia Medica* (2012) 22:276. doi: 10.11613/bm.2012.031

36. Cohen J. Statistical Power Analysis for the Behavioral Sciences. *Statistical Power Analysis for the Behavioral Sciences* (2013) doi: 10.4324/9780203771587

37. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* (1960) 20:37–46. doi: 10.1177/001316446002000104

38. Brier GW. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. *Monthly Weather Review* (1950) 78:1–3. doi: 10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2

39. Goldstein-Greenwood J. A Brief on Brier Scores | UVA Library. (2021) https://library.virginia.edu/data/articles/a-brief-on-brier-scores [Accessed September 5, 2023]