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

Manuscript

Department of Radiology, Medical University of Innsbruck

2025-03-05

# Abstract

# Introduction

# 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](https://classic.clinicaltrials.gov/ct2/show/NCT04416100) (NCT04416100).

## Study design and participants

Details on the study design and the cohort are provided in our recent publications (1–4). In brief, the longitudinal observation CovILD study aimed at investigation of symptom, cardiopulmonary and mental health recovery in a cohort of 145 COVID-19 survivors at two, three, six, and twelve months after COVID-19 diagnosis. The participants were recruited between March and June 2020 among patients of three clinical centers in Tyrol, Austria (Medical University of Innsbruck, St. Vinzenz Hospital in Zams, and Karl-Landsteiner Rehabilitation Center in Münster). The study 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. Herein, n = 420 longitudinal observations from 140 participants were analyzed, with complete computed tomography (CT) of the chest and lung function testing (LFT) as analysis inclusion criteria (**Figure 1**, **Supplementary Table S1**).

## Procedures

Details of study procedures and variables are provided in **Supplementary Methods**, **Supplementary Table S2**, and in our recent publications (1–4). In brief, baseline demographic and clinical data, and data concerning acute COVID-19 course were gathered 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.

Acute COVID-19 severity was 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 scans were evaluated by experienced radiologists blinded to the study design according to the Fleischner society guidelines (5) for presence of ground glass opacity (GGO), reticulation, consolidations, and bronchiectasis. Severity of lung lesions was scored separately for each lobe with a 0 - 5 scale and CT severity score (CTSS) was calculated as a sum of the scores for the entire organ, as described before (2,3). Additionally, percentages of the lung with opacity and high opacity were determined for the chest CT scans with an artificial intelligence (AI) based software (Syngo.via CT Pneumonia Analysis Software, Siemens Healthineers, Erlangen, Germany). In the current report, the following LFT parameters were analyzed: diffusion capacity for carbon monoxide (DLCO), forced vital capacity (FVC), and forced expiratory volume in one second (FEV1). The LFT variables were expressed as percentages of the patient’s reference value. Insufficiency was assumed for the LFT parameter value < 80% of reference (1,2,4). The following self-reported COVID-19 related symptoms of potential relevance for LFT results were analyzed: dyspnea (modified Medical Research Council [mMRC] scale > 0), cough (yes/no item), impaired physical performance (Eastern Cooperative Oncology Group [ECOG] scale > 0).

## Analysis endpoints

The primary analysis endpoint was construction and evaluation of multi-parameter models of the most common LFT abnormalities (each of DLCO < 80%, FVC < 80%, FEV1 < 80% of the patient’s reference) and of numeric LFT readouts (DLCO, FVC, FEV1) during COVID-19 convalescence. Such models employed common machine learning algorithms with a wide range of demographic and clinical features, as well as human-determined CT abnormalities, and AI-measured lung opacity and high opacity as explanatory variables. The secondary analysis endpoint was prediction of LFT abnormalities and numeric LFT readouts during COVID-19 recovery by human-determined CTSS and AI-measured lung opacity and high opacity. This endpoint was addressed by statistical hypothesis testing, correlation and receiver-operating characteristic (ROC) analysis.

## Statistical analysis

Details of statistical analysis are provided in **Supplementary Methods**.

Statistical analysis was performed with R version 4.2.3 (R Foundation for Statistical Computing). Numeric variables were presented as medians with interquartile ranges and ranges. Qualitative variables were presented as percentages and counts of the categories within the complete observations set. Differences in independently distributed numeric variables were analyzed by Mann-Whitney and Kruskal-Wallis test with, respectively, biseral r and effect size statistic. Statistical significance for differences in distribution of qualitative variables was determined by test with Cramer’s V effect size statistic. Co-occurrence of each of LFT findings, CT abnormalities, and symptoms was investigated by two-dimensional correspondence analysis (6).

Differences in medians of non-independently distributed, participant matched numeric variables between observations with and without LFT and CT abnormalities were assessed by a blocked bootstrap test with blocks defined by the participant’s identifier, and effect size measured by biseral r effect size statistic (7). Correlations of non-independently distributed CT and LFT readouts were assessed by blocked bootstrap Spearman’s rank test. Cutoffs of CTSS, opacity and high opacity for detection of LFT abnormalities were found by maximizing the Youden’s J statistic. ROC analysis statistics (area under the curve [AUC], sensitivity, specificity, Cohen’s ) for those optimal cutoffs were computed and their 95% confidence intervals were obtained by blocked bootstrap (8,9).

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. The explanatory variables included demographic features (e.g. age, sex, body mass index, smoking, comorbidity), characteristic of acute COVID-19 (severity, medication) and recovery (e.g. weight loss, symptoms of relevance for lung function, time after diagnosis), and presence of human- and AI-rated structural lung abnormalities in CT scans (e.g. GGO, CTSS, opacity and high opacity). The modeling responses and explanatory variables are listed in **Supplementary Table S2**. The models were constructed with four common machine learning algorithms: Random Forest (10,11), gradient boosted machines (GBM) (12–14), neural network with a single hidden layer (15), and support vector machines (SVM) with radial kernel (16,17). Selection of the optimal values of parameters controlling model behavior such as number of random trees, neurons in the hidden layer, or cost parameter was motivated by the maximum of Youden’s J statistic (classification models of LFT abnormalities) or minimum mean absolute error (MAE, regression models of LFT readouts) in 10-repeats 10-fold cross-validation (18). Because of the presence of participant-matched observation, blocked cross-validation design was used both in the model selection and model evaluation, with blocks defined by participant’s identifier. Model predictions were evaluated both in the training data and blocked 10-repeats 10-fold cross-validation. Concordance between the model-predicted and observed outcomes for classification models was assessed by Cohen’s inter-rater reliability statistic (8,19). Accuracy, AUC, specificity and sensitivity of the classification model were investigated by ROC. Calibration of the classification models was assessed by Brier scores (20). Fraction of explained variance in predictions of the regression models was measured by pseudo-R2, the regression model error was expressed as MAE. Spearman’s coefficients of correlation between the predicted and observed values were used to gauge calibration of the regression models. Over- and under-fitting was assessed by learning curves (21). Importance of explanatory variables was estimated by absolute values of SHAP statistics (Shapley additive explanations) (22–24). Co-linearity of the most influential explanatory variables was assessed by weighted correlation graph, whose edge weights corresponded to Kendall’s coefficients of pairwise correlations (25).

# Results

## Baseline Characteristic of the cohort and course of acute COVID-19

The longitudinal observation CovILD study cohort consisted of n = 145 convalescents of symptomatic SARS-CoV-2 infection initially enrolled during the fist outbreak of the pandemic (March - June 2020) in three clinical centers in western Austria (1–4). Out of them, n = 140 participants with n = 420 observations with complete CT and LFT data obtained at follow-up visits at the two, three, six and twelve months after COVID-19 diagnosis were included in the current analysis. Please note, that per participant, up to four observations were available. Due to this participant-matching, the observations were not independent, which had consequences for our analysis strategy (**Figure 1**). The majority of analyzed data points were obtained at the two- (n = 120) and three-month follow-up (n = 124, **Supplementary Table S1** - **S2**).

The analyzed participants were grouped according to the severity of acute COVID-19 as convalescents of ambulatory COVID-19 (25% of the cohort, home-isolated during infection, WHO ordinal scale for clinical improvement: 1 - 2), hospitalized moderate COVID-19 (52%, without oxygen or with low-flow oxygen, WHO ordinal scale for clinical improvement: 3 - 4), and hospitalized severe COVID-19 patients (23%, high-flow oxygen or mechanical ventilation, WHO ordinal scale for clinical improvement: 5 - 7, **Table 1**). In the analyzed data set, 56% of participants were male. The percentage of males was significantly higher in moderate and severe COVID as compared with ambulatory COVID-19 individuals (p < 0.001, effect size: V = 0.34, moderate). The median age at COVID-19 diagnosis was 56 years, ambulatory COVID-19 patients were significantly younger than hospitalized COVID-19 survivors (p < 0.001, effect size: η² = 0.17, moderate). Ex- or current smokers made up to 38.9% of the analyzed cohort. There were no significant differences in participants with smoking history between the acute COVID-19 severity subsets. Yet, the pack-year number was the highest in hospitalized moderate COVID-19 convalescents (p = 0.0061, effect size: η² = 0.072, small). Based on body mass index (BMI) > 25 kg/m2, 60% of participants were overweight or obese; there were no significant differences in percentages of obese or overweight participants between the acute COVID-19 severity subsets. Endocrine or metabolic conditions (44% of analyzed participants), cardiovascular diseases (39%), hypertension (30%), hypercholesterolemia (19%), and type II diabetes (17%) were the most common comorbidities at COVID-19 diagnosis. Frequencies of endocrine or metabolic comorbidities (p = 0.021, effect size: V = 0.26, small), cardiovascular conditions (p < 0.001, effect size: V = 0.38, moderate), hypercholesterolemia (p = 0.04, effect size: V = 0.23, small), and diabetes (p = 0.018, effect size: V = 0.26, small) were significantly more common in hospitalized than in ambulatory COVID-19 patients (**Table 1**).

The median length of hospital stay for analyzed participants was 8 days and was significantly longer for severe that moderate COVID-19 patients (p < 0.001, effect size: η² = 0.8, large). Antibiotic anti-infective treatment was applied to 54% of participants, this percentage was the highest in severe COVID-19 patients (p < 0.001, effect size: V = 0.53, large). Macrolides and anti-platelet drugs were administered to 19% and 15% of participants, respectively. Frequency of this treatment was significantly higher in hospitalized than in ambulatory patients (macrolides: p = 0.028, effect size: V = 0.24, small; anti-platelet: p = 0.028, effect size: V = 0.24, small). Systematic steroid administration was not considered as a treatment of moderate or severe COVID-19 during the first European outbreak of SARS-CoV-2 in 2020. Hence none of the study participants received corticosteroid therapy at the peak of acute COVID-19. Corticosteroids were administered to 16% of analyzed participants with persistent pneumonia beginning from the third week after diagnosis at the discretion of the physician. Such sub-acute corticosteroid treatment was the most frequent in severe COVID-19 convalescents (p = 0.0013, effect size: V = 0.33, moderate). The median weight loss during acute COVID-19 was 5 kg and was significantly higher in severe as compared with ambulatory or moderate COVID-19 patients (p < 0.001, effect size: η² = 0.33, large, **Table 1**).

## Pulmonary recovery

Frequency of any radiological lung abnormalities, GGO, reticulation, consolidation, and bronchiectasis diagnosed by a human radiologist was the highest in severe COVID-19 survivors followed by moderate COVID-19 patients. GGO and reticulation were the most frequent lesion types, independently of the follow-up and acute disease severity. Resolution of radiological lung findings. Consolidation and bronchiectasis lesions were detected primarily in moderate and severe COVID-19 survivors at the two-month follow-up examination. The increased frequency and severity of structural lung lesions in hospitalized patients as well as gradual resolution of the findings was also evident from analysis of the time course of human-determined CTSS, and AI-measured lung opacity and high opacity (**Supplementary Figures S1** and **S2**, **Supplementary Table S3**). As demonstrated by correspondence analysis, in the majority of observations with GGO, reticulations were also present (182 of 233 observations with GGO). consolidations co-occurred with GGO lesions only in a minute fractions of observations (20 of 233 observations with GGO), most frequently at the two-month follow-up (**Supplementary Figure S3A**). A high degree of correlation was observed between human-determined CTSS, and AI-determined lung opacity and high opacity (Spearman’s : 0.72 to 0.86, **Supplementary Figure S3B**).

Insufficient DLCO, FVC, and FEV1 defined as values below < 80% of the patient’s reference value were the most common abnormalities of lung function. As with the CT findings, their frequency was the highest in severe COVID-19 survivors followed by moderate COVID-19 patients, independently of the follow-up time point. Additionally, lung function recovery reflected by decreasing rates of insufficient DLCO and FVC, and by an improvement of DLCO and FVC values expressed as percentages of the patient’s reference was the most evident in severe COVID-19 survivors (**Supplementary Figures S4** and **S5**, **Supplementary Table S4**). In most of observations with insufficient FVC, insufficient FEV1 was found (60 of 83 observations with reduced FVC). By contrast, in approximately half of observations with insufficient DLCO reduced FVC or FEV1 was diagnosed (51 of 94 observations with insufficient DLCO, **Supplementary Figure S6A**). Accordingly, numeric values of FVC and FEV1 were strongly correlated (Spearman’s = 0.84), while correlation of DLCO with FVC or FEV1 was only moderate (Spearman’s : 0.35 to 0.39, **Supplementary Figure S6B**). This may suggest that reduced DLCO in COVID-19 survivors reflects a type of functional lung pathology that is distinct from abnormalities measured by FVC or FEV1.

Impaired physical performance defined by ECOG > 0 and exertional dyspnea defined by mMRC > 0 were the most frequent persistent COVID-19-related symptoms of potential relevance for pulmonary function, and their rates declined during convalescence in all acute COVID-19 severity subsets.  
Notably, we could not observe any substantial differences in frequency of physical performance impairment, dyspnea, and persistent cough between severe, moderate and ambulatory COVID-19 convalescents (**Supplementary Figure S7**, **Supplementary Table S5**).

## Modeling of lung function impairment in COVID-19 convalescents

Prediction of lung function deficits following COVID-19 with demographic, clinical and biochemical explanatory factors and characteristic of acute SARS-CoV-2 infection suffers from low accuracy (4). Additionally, relevance of residual structural lung lesions and severity of lung damage in COVD-19 convalescents for pulmonary function is not entirely clear (1,3). Herein, we sought to model the major LFT abnormalities, i.e. insufficient DLCO, FVC, and FEV1, and numeric readouts, i.e DLCO, FVC, FEV1 in the CovILD cohort with a broad panel of 37 independent variables including baseline demographic and clinical characteristic (e.g. gender, age, smoking history, comorbidity), characteristic of acute COVID-19 (e.g. severity, medication, hospitalization) and convalescence (follow-up time point, persistent symptoms), as well as presence and severity CT abnormalities (e.g. GGO, CTSS, AI-determined opacity and high opacity, **Supplementary Table S2**). To this end, we employed four popular machine learning algorithms, Random Forest (11), gradient boosted machines (GBM) (13,14), neural network with a single hidden layer (15), and support vector machines (SVM) (16). Because the CovILD data set consisted of non-independent, participant-matched observations, blocked cross-validation with participant’s identifier serving as the grouping variable was used for model optimization (tuning, **Supplementary Table S6**). Predictions of the models were finally evaluated by comparison with the observed outcome in the training data and the blocked cross-validation folds (18) (**Supplementary Figure S8**, **Table 2** and **3**, **Supplementary Tables S7** and **S8**).

Among the investigated LFT abnormalities, solely insufficient DLCO defined as values < 80% of the patient’s reference could be modeled with satisfactory accuracy in the cross-validation setting independently of the machine learning algorithm (overall accuracy: 0.82 to 0.85, Cohen’s : 0.45 to 0.5, AUC: 0.87 to 0.9). Low Brier score values for the models of insufficient DLCO indicated good overall calibration (Brier score: 0.11 to 0.14, **Figure 2A**). In turn, models of insufficient FVC and FEV1 exhibited poor predictive performance in the cross-validation setting (overall accuracy: 0.72 to 0.81, Cohen’s : 0.094 to 0.17, AUC: 0.57 to 0.69, **Supplementary Figure S9**).

As in case of the LFT abnormalities, solely DLCO expressed as percentages of the patient’s reference was modeled with sufficiently low error and substantial explanatory power in cross-validation with the Random Forest, GBM, and SVM algorithms (MAE: 11.6 to 12.5, pseudo-R2: 0.26 to 0.34). Those models were also characterized by good calibration assessed by Spearman’s coefficients of correlation between the predicted and observed DLCO (Spearman’s : 0.55 to 0.59). By contrast, the neural network model of DLCO suffered from large error and low fraction of explained variance in the cross-validation setting (MAE = 13.8, pseudo-R2 = 0.043, **Figure 2B**). Analogically, no meaningful models of FVC or FEV1 could be developed with any of the machine learning algorithms (cross-validation, R2: -0.086 to -0.03, **Supplementary Figure S10**).

Concerning the models of insufficient DLCO, the highest concordance between the predicted and observed DLCO < 80% of the patient’s reference was achieved for moderate COVID-19 convalescents at the two- to six-month follow-up examinations (all algorithms, : 0.45, 0.69). The accuracy was in turn the poorest for ambulatory patients at the six- and twelve-month follow-up (all algorithms, : 0, 0.46, **Supplementary Figure S11A**). The Random Forest, SVM and GBM models of DLCO expressed as percentage of the patient’s reference exhibited the lowest average errors for moderate COVID-19 survivors (mean error: -2.8 to 2.9). In turn, DLCO values were systematically overestimated for severe COVID-19 patients at the two- to six-month follow-up visits (mean error: 1.2 to 6.5) and underestimated for ambulatory COVID-19 convalescents at the same time points (mean error: -4.2 to 0.12, **Supplementary Figure S11B**). The best performance in the hospitalized moderate subset is likely attributed to the large number of observations and frequent DLCO deficits in this group of participants available for training of the models (n = 47 observations of DLCO < 80% in 234 observations from moderate COVID-19 patients). Conversely, the numbers of observations from ambulatory COVID-19 patients and numbers of observations DLCO insufficiency in this group was low (10 observations of DLCO < 80% in 85 observations from ambulatory patients) which resulted in relatively poor performance of the machine learning models.

To assess under- and over-fitting, we resorted to analyses of learning curves of the models of DLCO and reduced DLCO. With this approach, we evaluated performance of the models re-trained in subsets of the modeling data set of varying sizes. The performance evaluation was done for the training subsets, test subsets (one-forth of observations not used for the model training), and 10-repeats 10-fold cross-validation (21). The learning curves for the best-performing GBM models are presented in **Supplementary Figure S12**, for the remaining algorithms comparable results were obtained (not shown). As inferred from substantial differences of accuracy and Cohen’s between the training, test, and cross-validation subsets even for the largest sizes of the training data, models of insufficient DLCO suffer from over-fitting, i.e. poor generalizability for unseen data. The generalizability of the models of DLCO expressed as percentage of the patient#s reference was substantially better.

As investigated by absolute values of SHAP variable importance metrics (22,23), human- and AI-derived ratings of structural lung damage, CTSS, opacity and high opacity, belonged to the most influential explanatory variables of the models of DLCO insufficiency and the meaningful models of DLCO. Other influential explanatory variables were well characterized risk factors of severe COVID-19 (age, male sex, body mass index, pre-existing malignancy and cardiovascular disease), readouts of acute COVID-19 severity (severity class, WHO ordinal severity scale, hospitalization length, ICU treatment, anti-coagulant treatment, and weight change), smoking intensity (pack-years), as well as rating of physical performance impairment (ECOG) and exertional dyspnea (mMRC) (**Figure 3**, **Supplementary Figure S13**, **Supplementary Table S9**).

***placeholder for colinearity of the explanatory varibales***

## Rating of structural lung damage by human radiologist and AI, and lung function deficiency

Because CTSS, lung opacity, and high opacity were identified to be crucial for prediction of DLCO insufficiency and DLCO by machine learning, we explored the association of the human- and AI-determined readouts of lung damage with DLCO in more detail.

Human-determined CTSS (difference of medians: 9 points, p < 0.001, effect size: r = 0.57), AI-determined opacity (difference of medians: 1.3% of the lung, p < 0.001, effect size: r = 0.63) and high opacity (difference of medians: 0.063, p < 0.001% of the lung, effect size: r = 0.58) were significantly higher in observations with DLCO < 80% of the patient’s reference than in the remaining data points. The effect size of those differences was large (**Figure 4A**). AI-measured opacity of the lung, but not CTSS or high opacity, was found to be significantly higher in observations with insufficient FVC (difference of medians: 0.4, p = 0.0051, effect size: r = 0.29) and insufficient FEV1 (difference of medians: 0.3, p = 0.009, effect size: r = 0.27); sizes of those effects were, however, small (**Supplementary Table S12**). Furthermore, CTSS, opacity and high opacity correlated negatively with moderate effect size with DLCO expressed as percentage of the patient’s reference (p < 0.001, Spearman’s : -0.46 to -0.44). Interestingly, the CT readouts of structural lung damage also correlated significantly with FVC (Spearman’s : -0.3 to -0.21) and FEV1 (Spearman’s : -0.27 to -0.14); effect size of those associations was small (**Supplementary Table S13**).

Finally, in a univariable ROC analysis, human-determined CTSS (AUC = 0.78, 95% CI: 0.73 to 0.84), and AI-determined lung opacity (AUC = 0.81, 95% CI: 0.76 to 0.86) and high opacity (AUC = 0.79, 95% CI: 0.73 to 0.84) were identified as standalone markers of insufficient DLCO in COVID-19 patients. The optimal cutoffs of CTSS, opacity, and high opacity for detection of DLCO < 80% of the patient’s reference were, respectively, 4 points, 0.12% of the lung, and 0.002% of the lung, and allowed for identification of insufficient DLCO with moderate sensitivity and specificity (sensitivity: 0.78 to 0.81, specificity: 0.68 to 0.69, **Figure 4B**). In case of the AI measures of lung opacity and high opacity, the low values of the optimal cutoffs let us infer that event low-grade radiological abnormalities can be associated with clinically relevant functional lung impairment. Yet, concordance between the predicted observed DLCO insufficiency assessed by Cohen’s for CTSS, opacity, and high opacity as standalone markers (blocked bootstrap : 0.34 to 0.38, **Table 4**) was substantially lower that in the multi-parameter machine learning models (cross-validated : 0.45 to 0.5). This underlines the importance of other CT-unrelated explanatory factors such acute disease course for reliable prediction of functional lung deficits in COVID-19 convalescents.

# Discussion

# Limitations

Our study has limitations. First, the overall patient and observation number was low, in particular for ambulatory COVID-19 convalescents. Analogically, the study cohort was enriched in hospitalized individuals with in constitute a minute fraction of COVID-19 patients in the real world setting. Second, complete sets of longitudinal CT and LFT measurements at the two-, three-, six- and twelve-month follow-up examinations were available solely for 55 patients. Especially the number of observations obtained with ambulatory and moderate COVID-19 convalescents at the six- and twelve-month follow-up were substantially lower as compared with earlier time points. The incompleteness of the longitudinal data may have hence compromised performance of the machine learning models in particular for ambulatory COVID-19 cases and at the later time points. Third, because of the limited number of participants and observations, we abstained from definition of a test subset of the data used solely for bias-free model evaluation (hold-out strategy). Instead, both model selection and evaluation was done with blocked repeated cross-validation, which may have overestimated performance of the models. Hence, external validation of our findings in an independent cohort is recommended. Fourth, as inferred from the analysis of learning curves, the models of insufficient DLCO suffered for substantial over-fitting. While this problem can be partially traced back to an unsharp distinction between sufficient and reduced DLCO with the 80% reference cutoff, a modeling approach employing regularized machine learning algorithm like XGBoost or regularized neural networks may further improve quality of predictions. Furthermore, most of the highly influential explanatory variables for predictions of DLCO and reduced DLCO, and in particular the CT readouts of lung damage were strongly inter-correlated, which raises questions about redundancy of explanatory variables. Finally, the analyzed cohort was recruited in the initial phase of the pandemic and consisted of individuals infected with the wild-type variant of the SARS-CoV-2 virus. For this reason, it is not completely clear, how our findings translate to the recent variants of the pathogen and how the pulmonary recovery is affected by anti-SARS-CoV-2 immunity, improved treatment and care. However, it is feasible, that the CT severity readouts, human-determined CTSS as well as AI-determined opacity and high opacity, are equally applicable in the post-pandemic setting as standalone markers of functional lung impairment during recovery from COVID-19 and other respiratory infections.

# Data and code availability

The entire R analysis pipeline is available as a [GitHub repository](https://github.com/PiotrTymoszuk/CovILD_AI). The [CovILD data set](https://github.com/PiotrTymoszuk/covILD) will be made available on reasonable request to the corresponding author.

# Tables

Table 1: Baseline characteristics and COVID-19 course in the CovILD cohort. 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.

| **Variablea** | **Cohort** | **Ambulatory COVID-19** | **Moderate COVID-19** | **Severe COVID-19** | **Significanceb** | **Effect sizeb** |
| --- | --- | --- | --- | --- | --- | --- |
| Participants, n | 140 | 35 | 73 | 32 |  |  |
| sex | female: 44% (n = 61) male: 56% (n = 79) | female: 71% (n = 25) male: 29% (n = 10) | female: 38% (n = 28) male: 62% (n = 45) | female: 25% (n = 8) male: 75% (n = 24) | p < 0.001 | V = 0.34 |
| age at admission, years | 56 [IQR: 49 - 68] range: 19 - 87 | 48 [IQR: 38 - 55] range: 19 - 83 | 61 [IQR: 52 - 73] range: 27 - 87 | 57 [IQR: 53 - 64] range: 44 - 79 | p < 0.001 | η² = 0.17 |
| age class | young adult: 14% (n = 19) middle aged: 59% (n = 83) elderly: 27% (n = 38) | young adult: 34% (n = 12) middle aged: 60% (n = 21) elderly: 5.7% (n = 2) | young adult: 9.6% (n = 7) middle aged: 51% (n = 37) elderly: 40% (n = 29) | young adult: 0% (n = 0) middle aged: 78% (n = 25) elderly: 22% (n = 7) | p < 0.001 | V = 0.33 |
| smoking history | never smoker: 61% (n = 86) ex-smoker: 36% (n = 50) active smoker: 2.9% (n = 4) | never smoker: 77% (n = 27) ex-smoker: 20% (n = 7) active smoker: 2.9% (n = 1) | never smoker: 51% (n = 37) ex-smoker: 45% (n = 33) active smoker: 4.1% (n = 3) | never smoker: 69% (n = 22) ex-smoker: 31% (n = 10) active smoker: 0% (n = 0) | ns (p = 0.1) | V = 0.18 |
| smoking, pack-years | 0 [IQR: 0 - 10] range: 0 - 80 | 0 [IQR: 0 - 0] range: 0 - 30 | 2 [IQR: 0 - 20] range: 0 - 80 | 0 [IQR: 0 - 7] range: 0 - 60 | p = 0.0061 | η² = 0.072 |
| BMI, kg/m² | 26 [IQR: 23 - 29] range: 18 - 47 | 25 [IQR: 21 - 28] range: 19 - 37 | 26 [IQR: 24 - 29] range: 18 - 47 | 26 [IQR: 24 - 28] range: 19 - 34 | ns (p = 0.24) | η² = 0.0095 |
| body mass class | normal: 41% (n = 57) overweight: 41% (n = 57) obesity: 19% (n = 26) | normal: 54% (n = 19) overweight: 31% (n = 11) obesity: 14% (n = 5) | normal: 33% (n = 24) overweight: 48% (n = 35) obesity: 19% (n = 14) | normal: 44% (n = 14) overweight: 34% (n = 11) obesity: 22% (n = 7) | ns (p = 0.32) | V = 0.14 |
| cardiovascular illness | 39% (n = 54) | 8.6% (n = 3) | 44% (n = 32) | 59% (n = 19) | p < 0.001 | V = 0.38 |
| hypertension | 30% (n = 42) | 8.6% (n = 3) | 32% (n = 23) | 50% (n = 16) | p = 0.0027 | V = 0.31 |
| pulmonary illness | 18% (n = 25) | 17% (n = 6) | 19% (n = 14) | 16% (n = 5) | ns (p = 0.9) | V = 0.039 |
| COPD | 5% (n = 7) | 5.7% (n = 2) | 4.1% (n = 3) | 6.2% (n = 2) | ns (p = 0.9) | V = 0.043 |
| asthma | 6.4% (n = 9) | 8.6% (n = 3) | 4.1% (n = 3) | 9.4% (n = 3) | ns (p = 0.57) | V = 0.099 |
| interstitial lung disease | 0.71% (n = 1) | 0% (n = 0) | 1.4% (n = 1) | 0% (n = 0) | ns (p = 0.67) | V = 0.081 |
| endocrine/metabolic disease | 44% (n = 62) | 23% (n = 8) | 49% (n = 36) | 56% (n = 18) | p = 0.021 | V = 0.26 |
| hypercholesterolemia | 19% (n = 27) | 5.7% (n = 2) | 27% (n = 20) | 16% (n = 5) | p = 0.04 | V = 0.23 |
| type II diabetes mellitus | 17% (n = 24) | 2.9% (n = 1) | 18% (n = 13) | 31% (n = 10) | p = 0.018 | V = 0.26 |
| CKD | 6.4% (n = 9) | 0% (n = 0) | 5.5% (n = 4) | 16% (n = 5) | p = 0.048 | V = 0.22 |
| gastrointestinal disease | 13% (n = 18) | 2.9% (n = 1) | 18% (n = 13) | 12% (n = 4) | ns (p = 0.13) | V = 0.18 |
| chronic lung disease | 5.7% (n = 8) | 2.9% (n = 1) | 5.5% (n = 4) | 9.4% (n = 3) | ns (p = 0.57) | V = 0.098 |
| malignancy | 11% (n = 16) | 5.7% (n = 2) | 15% (n = 11) | 9.4% (n = 3) | ns (p = 0.39) | V = 0.13 |
| immune deficiency | 5.7% (n = 8) | 2.9% (n = 1) | 1.4% (n = 1) | 19% (n = 6) | p = 0.0034 | V = 0.31 |
| WHO ordinal severity scale | 3 [IQR: 2.8 - 4] range: 2 - 7 | 2 [IQR: 2 - 2] range: 2 - 2 | 3 [IQR: 3 - 4] range: 3 - 4 | 6 [IQR: 6 - 6] range: 5 - 7 | p < 0.001 | η² = 0.88 |
| hospital stay, days | 8 [IQR: 0.75 - 16] range: 0 - 69 | 0 [IQR: 0 - 0] range: 0 - 0 | 8 [IQR: 5 - 11] range: 1 - 33 | 30 [IQR: 24 - 44] range: 11 - 69 | p < 0.001 | η² = 0.8 |
| ICU stay, days | 0 [IQR: 0 - 0] range: 0 - 46 | 0 [IQR: 0 - 0] range: 0 - 0 | 0 [IQR: 0 - 0] range: 0 - 0 | 16 [IQR: 10 - 27] range: 0 - 46 | p < 0.001 | η² = 0.94 |
| steroids during COVID-19 | 16% (n = 23) | 2.9% (n = 1) | 14% (n = 10) | 38% (n = 12) | p = 0.0013 | V = 0.33 |
| anti-infectives during COVID-19 | 54% (n = 76) | 11% (n = 4) | 62% (n = 45) | 84% (n = 27) | p < 0.001 | V = 0.53 |
| macrolides during COVID-19 | 19% (n = 27) | 2.9% (n = 1) | 26% (n = 19) | 22% (n = 7) | p = 0.028 | V = 0.24 |
| anti-platelet drugs during COVID-19 | 15% (n = 21) | 2.9% (n = 1) | 15% (n = 11) | 28% (n = 9) | p = 0.028 | V = 0.24 |
| anti-coagulants during COVID-19 | 5% (n = 7) | 2.9% (n = 1) | 2.7% (n = 2) | 12% (n = 4) | ns (p = 0.12) | V = 0.19 |
| immunosuppressive drugs during COVID-19 | 4.3% (n = 6) | 0% (n = 0) | 4.1% (n = 3) | 9.4% (n = 3) | ns (p = 0.22) | V = 0.16 |
| weight change during COVID-19, kg | -5 [IQR: -8.2 - 0] range: -25 - 0 | 0 [IQR: -4.5 - 0] range: -12 - 0 | -5 [IQR: -7 - -1] range: -20 - 0 | -11 [IQR: -15 - -7.8] range: -25 - 0 | p < 0.001 | η² = 0.33 |
| aage class, young adult: 16 - 40: young adult; 41 - 65: middle aged; 66 and older: elderly; BMI: body mass index; body mass class, normal: BMI ≤ 25, overweight: 25 - 30, obesity: > 30 kg/m²; COPD: chronic obstructive pulmonary disease; CKD: chronic kidney disease; ICU: intensive care unit; steroids during COVID-19: in cases of non-improving pneumonia, ≥ 3 weeks after diagnosis. | | | | | | |
| bNumeric variables: Kruskal-Wallis test with η² effect size statistic. Categorical variables: χ² test with Cramer V effect size statistic. P values corrected for multiple testing with the false discovery rate method. | | | | | | |

Table 2: Cross-validated performance of binary machine learning classifiers at predicting lung function testing (LFT) abnormalities.

| **Responsea** | **Algorithmb** | **Overall accuracyc** | **κd** | **Brier score** | **AUCe** | **Sensitivity** | **Specificity** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| DLCO < 80% | Random Forest | 0.85 | 0.480 | 0.11 | 0.90 | 0.53 | 0.94 |
| Neural network | 0.85 | 0.500 | 0.14 | 0.88 | 0.60 | 0.91 |
| SVM radial | 0.82 | 0.450 | 0.13 | 0.87 | 0.58 | 0.89 |
| GBM | 0.84 | 0.470 | 0.12 | 0.90 | 0.53 | 0.93 |
| FVC < 80% | Random Forest | 0.79 | 0.110 | 0.16 | 0.69 | 0.14 | 0.95 |
| Neural network | 0.72 | 0.094 | 0.25 | 0.58 | 0.27 | 0.83 |
| SVM radial | 0.78 | 0.120 | 0.16 | 0.68 | 0.19 | 0.92 |
| GBM | 0.78 | 0.150 | 0.17 | 0.67 | 0.21 | 0.92 |
| FEV1 < 80% | Random Forest | 0.80 | 0.120 | 0.15 | 0.64 | 0.15 | 0.95 |
| Neural network | 0.75 | 0.110 | 0.21 | 0.57 | 0.26 | 0.86 |
| SVM radial | 0.80 | 0.130 | 0.16 | 0.59 | 0.18 | 0.94 |
| GBM | 0.81 | 0.170 | 0.16 | 0.61 | 0.21 | 0.94 |
| aLFT: lung function testing, DLCO: diffusion capacity for CO, FVC: forced vital capacity; FEV1: forced expiratory volume in one second. | | | | | | | |
| bSVM: support vector machines with radial kernel; GBM: gradient boosted machines. | | | | | | | |
| cRatio of correct predictions to the total observation number. | | | | | | | |
| dCohen κ statistic of inter-rater reliability between the predicted and observed outcome. | | | | | | | |
| eAUC: are under the receiver-operating characteristic curve. | | | | | | | |

Table 3: Cross-validated performance of regression machine learning models at predicting values of lung function testing parameters.

| **Responsea** | **Algorithmb** | **pseudo-R²c** | **MAEd** | **ρe** |
| --- | --- | --- | --- | --- |
| DLCO | Random Forest | 0.300 | 12 | 0.570 |
| Neural network | 0.043 | 14 | 0.450 |
| SVM radial | 0.260 | 12 | 0.550 |
| GBM | 0.340 | 12 | 0.590 |
| FVC | Random Forest | -0.030 | 10 | 0.220 |
| Neural network | -0.079 | 11 | 0.074 |
| SVM radial | -0.031 | 10 | 0.210 |
| GBM | -0.040 | 10 | 0.200 |
| FEV1 | Random Forest | -0.045 | 12 | 0.160 |
| Neural network | -0.086 | 12 | 0.210 |
| SVM radial | -0.039 | 11 | 0.190 |
| GBM | -0.047 | 12 | 0.170 |
| aDLCO: diffusion capacity for CO, FVC: forced vital capacity; FEV1: forced expiratory volume in one second. | | | | |
| bSVM: support vector machines with radial kernel; GBM: gradient boosted machines. | | | | |
| cDefined as 1 - ratio of mean squared error and variance. | | | | |
| dMAE: mean absolute error. | | | | |
| eΡ: Spearman coefficient of correlation between the predicted and observed response values. | | | | |

Table 4: Detection of reduced diffusion capacity for CO DLCO < 80% reference value) by single CT-derived parameters: AI-determined opacity and high opacity, and human-determined CT severity score.

| **CT variablea** | **Cutoffb** | **Statisticc** | **Value, 95% CI** |
| --- | --- | --- | --- |
| CTSS |  | AUC | 0.78 [0.727 - 0.84] |
| 4.000 | κ | 0.34 [0.23 - 0.45] |
| 4.000 | Sensitivity | 0.78 [0.64 - 0.89] |
| 4.000 | Specificity | 0.68 [0.6 - 0.75] |
| high opacity, AI |  | AUC | 0.79 [0.734 - 0.84] |
| 0.002 | κ | 0.37 [0.26 - 0.47] |
| 0.002 | Sensitivity | 0.8 [0.7 - 0.89] |
| 0.002 | Specificity | 0.68 [0.62 - 0.75] |
| opacity, AI |  | AUC | 0.81 [0.763 - 0.86] |
| 0.120 | κ | 0.38 [0.27 - 0.48] |
| 0.120 | Sensitivity | 0.81 [0.72 - 0.89] |
| 0.120 | Specificity | 0.69 [0.62 - 0.75] |
| aCTSS: human-determined CT severity score, sum for al lung lobes; high opacity and opacity, AI: percentage of the lungs with high opacity and opacity determined by artificial intelligence. | | | |
| bCutoff of the CT variable corresponding to the maximum of Jouden Y statistic. | | | |
| cAUC: area under the curve of receiver-operating characteristic; κ: Cohen κ statistic of inter-rater reliability between the predicted and observed outcome, computed for the CT variable cutoff; sensitivity and specificity: sensitivity and specificity computed at the CT variable cutoff. | | | |

# Figures

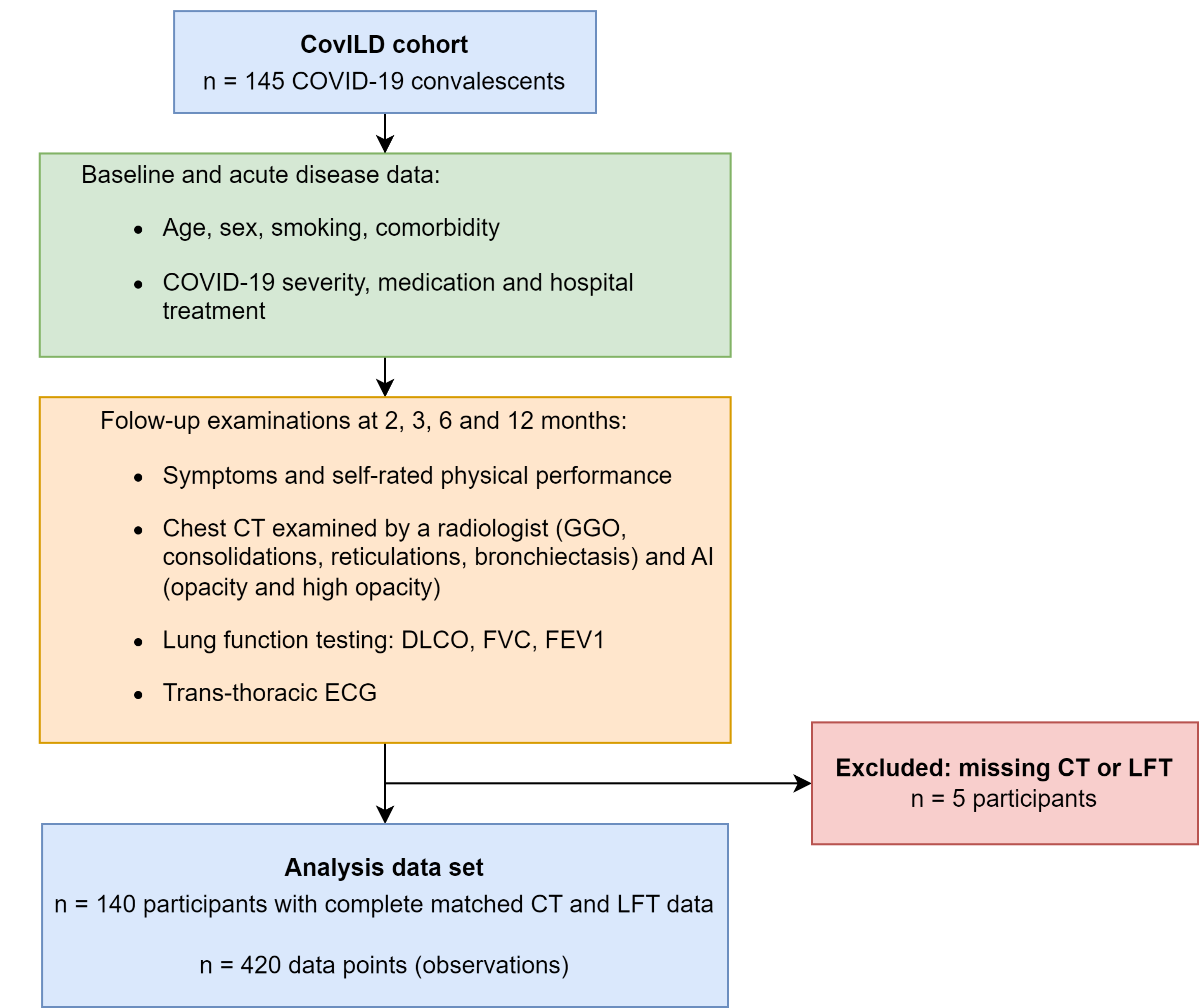


Figure 1: Analysis inclusion scheme.

**Figure 1. Analysis inclusion scheme.**

*The analysis inclusion criterion for participants of the longitudinal observation CovILD study was completeness of visit-matched computed tomography and lung function testing results.*

*CT: compute tomography; LFT: lung function testing; GGO: ground glass opacity; AI: artificial intelligence; DLCO: diffusion capacity for carbon monoxide (CO); FVC: forced vital capacity; FEV1: forced expiratory volume in one second.*

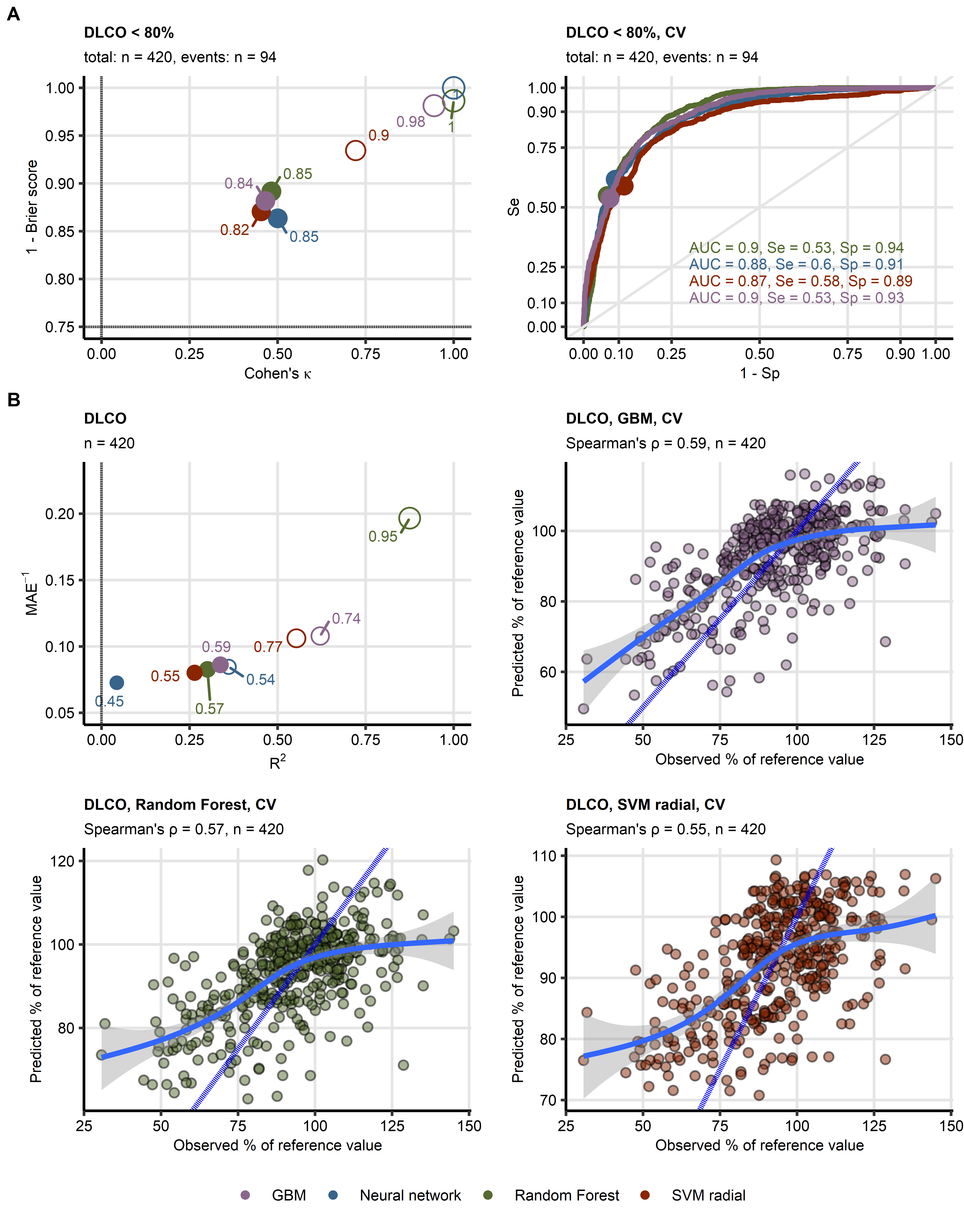


Figure 2: Evaluation of performance of machine learning models of diffusion capacity for CO during COVID-19 convalescence.

**Figure 2. Evaluation of performance of machine learning models of diffusion capacity for CO during COVID-19 convalescence.**

*(A) Four machine learning classification models of insufficient diffusion capacity for carbon monoxide (< 80% of reference value: n = 94, total observations: n = 420) employing time after COVID-19 diagnosis, computed tomography readouts, demographic and clinical explanatory variables were trained. Their performance was evaluated in the entire data set and 10-repeats 10-fold cross-validation with overall accuracy metric, Cohen’s 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 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 witj DLCO insufficiency (‘events’) are displayed in the plot captions.*

*(B) Four machine learning regression models of diffusion capacity for carbon monoxide (percentage of reference values, total observations: n = 420) employing time after COVID-19 diagnosis, 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 as a measure of explained variation, mean absolute error, and Spearman’s coefficient of correlation between the predicted and observed values. Bubble plot: numeric performance measures of the model (open circles: the entire data set, filled circles: cross-validation); point sizes and point labels represent values of correlation coefficient, the dashed line visualizes value expected for a meaningless model. Scatter plots: observed and predicted values of diffusion capacity for carbon monoxide in cross-validation folds; the blue dashed lines with slope 1 and intercept 0 represent absolutely accurate predictions, general additive model trends with standard errors are visualized as the solid blue lines with gray ribbons. Numbers of complete observations and Spearman’s coefficients of correlation between the predicted and observed values are displayed in the plot captions.*

*DLCO: diffusion capacity for carbon monoxide, CV: cross-validation; AUC: are under the receiver-operating characteristic curve; Se: sensitivity; Sp: specificity; MAE: mean absolute error; GBM: gradient boosted machines; SVM radial: support vector machines with radial kernel.*

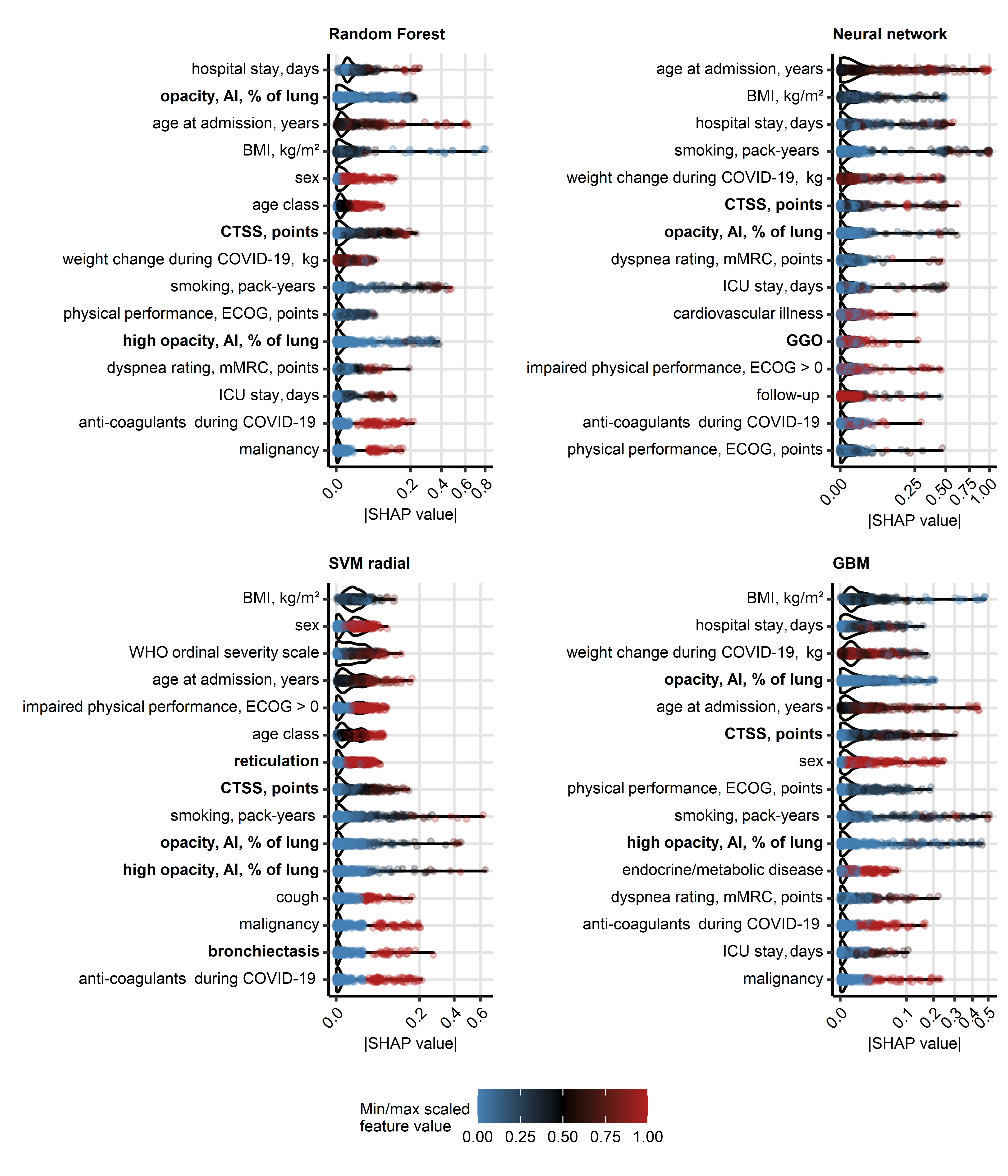


Figure 3: Explanatory variable importance for models of insufficient diffusion capacity for carbon monoxide measured by Shapley additive explanations.

**Figure 3. Explanatory variable importance for models of insufficient diffusion capacity for carbon monoxide measured by Shapley additive explanations.**

*Importance of explanatory variables for the machine learning models of insufficient diffusion capacity for carbon monoxide (< 80% 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; opacity and high opacity, AI: opacity and high opacity of the lung determined by artificial intelligence; BMI: body mass index; CTSS: human-determined CT severity score, sum for all lobes; ECOG: Eastern Cooperative Oncology Group physical performance score; mMRC: modified Medical Research Council dyspnea scale; ICU: intensive care unit.*

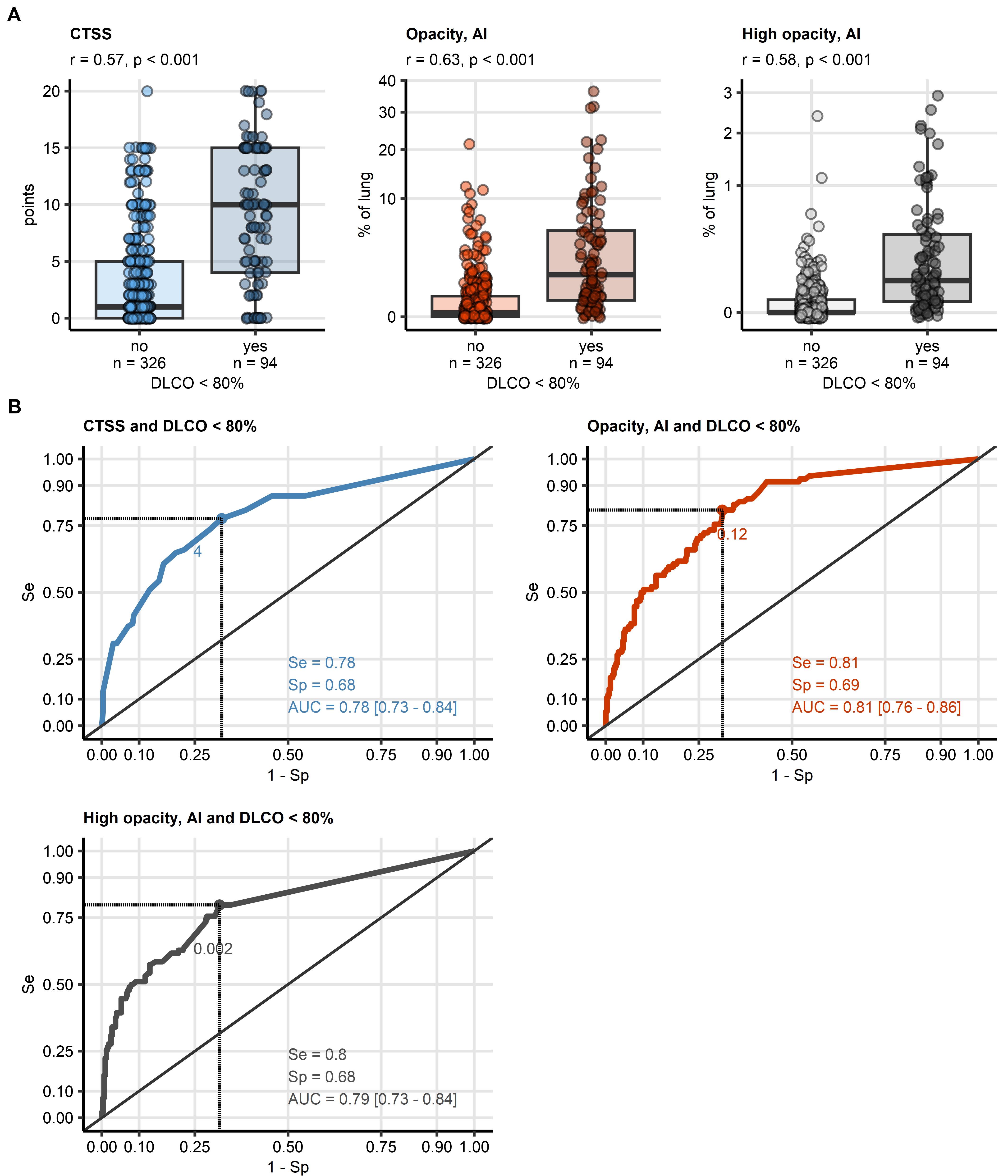


Figure 4: Detection of DLCO insufficiency by human- and artificial intelligence-determined CT readouts of severity of structural lung damage.

**Figure 4. Detection of DLCO insufficiency by human- and artificial intelligence-determined CT readouts of severity of structural lung damage.**

*Human- and artificial intelligence-determined computed tomography readouts of structural lung damage were identified as influential explanatory variables at prediction of insufficiency of diffusion capacity for carbon monoxide (< 80%) by machine learning.*

*(A) Values of the radiological readouts of lung damage severity were compared between data points with and without insufficient diffusion capacity for carbon monoxide by blocked bootstrap test with r effect size statistic. Median values with interquartile ranges are depicted as boxes, whiskers span over 150% of the interquartile ranges, single observations are visualized as points. Effect sizes and p values are displayed in the plot captions. Numbers of observations are indicated in the X axes.*

*(B) Quality of detection of insufficient diffusion capacity for carbon monoxide with the radiological readouts of lung damage severity was assessed by receiver-operating characteristic (ROC) analysis. ROC curves are shown, the optimal cutoffs of the severity readouts determined by Youden’s criterion are represented by points with numbers. Sensitivity, specificity at the optimal cutoff, and area under the curve statistic with 95% confidence interval are displayed in the plots.*

*CT: computed tomography; DLCO: diffusion capacity for carbon monoxide; CTSS: human-determined CT severity score, sum for all lobes; opacity and high opacity, AI: opacity and high opacity of the lung determined by artificial intelligence; AUC: are under the curve of receiver-operating characteristic; Se: sensitivity; Sp: specificity.*

# References

1. 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](https://doi.org/10.1183/23120541.00317-2022)

2. 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](https://doi.org/10.1183/13993003.03481-2020)

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](https://doi.org/10.1148/radiol.211670)

4. 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](https://doi.org/10.7554/ELIFE.72500)

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](https://doi.org/10.1148/radiol.2462070712)

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

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

8. McHugh ML. Interrater reliability: the kappa statistic. *Biochemia Medica* (2012) 22:276. doi: [10.11613/bm.2012.031](https://doi.org/10.11613/bm.2012.031)

9. 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](https://doi.org/10.18637/jss.v061.i08)

10. 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](https://doi.org/10.18637/JSS.V077.I01)

11. Breiman L. Random forests. *Machine Learning* (2001) 45:5–32. doi: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324)

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

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

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

15. Ripley BD. *Pattern recognition and neural networks*. Cambridge University Press (2014). doi: [10.1017/CBO9780511812651](https://doi.org/10.1017/CBO9780511812651)

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

17. 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](https://doi.org/10.18637/JSS.V011.I09)

18. 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](https://doi.org/10.18637/jss.v028.i05)

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

20. 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](https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2)

21. Viering T, Loog M. The Shape of Learning Curves: A Review. *IEEE transactions on pattern analysis and machine intelligence* (2023) 45:7799–7819. doi: [10.1109/TPAMI.2022.3220744](https://doi.org/10.1109/TPAMI.2022.3220744)

22. 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>

23. 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>

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

25. Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal* (2006) Complex Sy:1695. <https://igraph.org>