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

Rebuttal Letter

Department of Radiology, Medical University of Innsbruck

2025-03-09

# Reviewer 1

## Point 1

### Issue

Add results in the abstract as well.

### Response

*Anna and Gerlig: I think, we can stress that we were able to model DLCO but not other LFT responses and include some key characteristics of the machine learning models of DLCO < 80% and DLCO. My proposal for the response below.*

As suggested, we state now in Abstract, that we were able to establish meaningful models of DLCO and DLCO < 80% but not for other LFT responses. We also included key cross-validated performance statistics of meaningful machine learning models at predicting DLCO < 80% (accuracy: 0.82 - 0.85, Cohen’s : 0.45 - 0.5, AUC: 0.87 - 0.9) and DLCO expressed as percentage of the patient’s reference (mean absolute error [MAE]: 11.6 - 12.5%, : 0.26 - 0.34). Please note that the later performance statistics of DLCO as % of the patient’s reference, we provide the performance metric range for the random forest, GBM, and SVM algorithm, as the neural network model yielded a poor fit (MAE: 14%, : 0.043).

## Point 2

### Issue

The introduction is not sufficient.

### Response

*Anna and Gerlig: please answer, thanks!*

## Point 3

### Issue

The literature review part needs to be added.

### Response

*Anna and Gerlig: as above. Below some papersa I found interesting.* [*https://pubs.rsna.org/doi/full/10.1148/radiol.222888*](https://pubs.rsna.org/doi/full/10.1148/radiol.222888)*: no modeling but reduced DLCO in COVID-19 survivors without or delayed radiological resolution.* [*https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16915*](https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16915)*: a deep learning approach to modeling LFT, including DLCO with CT data in large data sets (s8K+ for training) of CT scans, allows for DLO predictions for single lobes; please have a look at Figure 3, where they report correlations between the observed and predicted DLCO - they are not so much better than our best GBM model of DLCO.* [*https://pubs.rsna.org/doi/full/10.1148/radiol.221488*](https://pubs.rsna.org/doi/full/10.1148/radiol.221488)*: they succeeded at predictions of FEV1, FVC, and FEV1:FVC using deep learning (neuronal network) and a large data set of CT images.* [*https://pmc.ncbi.nlm.nih.gov/articles/PMC8264499/*](https://pmc.ncbi.nlm.nih.gov/articles/PMC8264499/)*: a nice review with a short description of algorithms.*

## Point 4

### Issue

After adding the literature review, add a summary table and clearly mention this study’s research gaps and contributions.

### Response

*Anna and Gerlig, here some my thoughts: As I see our results as a bigger picture: we were limited by the number of patients and by the fact that we worked with processed CT data: scores and values of opacity/high opacity, and human-identified findings - for this limited information, we established quite nice models of DLCO and reduced DLCO < 80%.* *There are already some papers on machine learning and lung function.* *What looks pretty novel to me: we are among the first groups who used not only the CT information for modeling but also clinical and demographic data.* *We also integrate information derived from CT by a software and a human radiologist, which is also a quite novel approach.*

## Point 5

### Issue

More details, such as features, are required in the study data.

### Response

Our manuscript is accompanied by a thorough information on patients, procedures, and study variables provided in **Supplementary Methods**. Please refer to **Supplementary Table S2** for a detailed listing of modeling response and explanatory variables with their format and descriptions.

## Point 6

### Issue

The paper lacks details about hyper-parameters for machine learning models. For example, what was the C and class weight value for the SVM model?

### Response

Thanks for this point. We agree that the information on hyper-parameters in the initial submission was hard to understand for many clinical readers and too specific. In the revised manuscript, we provide short descriptions of the hyper-parameters in **Supplementary Methods/Multi-parameter modeling of lung function**. These descriptions were also added as footnotes of **Supplementary Table S6**.

## Point 7

### Issue

Have you considered the GridSearchCV for model hyper-parameters or set it manually?

### Response

Thanks for this question. The hyper-parameters were found by searching grids of values with the maximum Youden’s J statistic for the classification models, and minimum mean absolute error (MAE) for the regression models in 10-repeats 10-fold patient-blocked cross-validation as the selection criteria. This tuning procedure was executed by calling train() function from *caret* package (1), which, together with our open-source package [*caretExtra*](https://github.com/PiotrTymoszuk/caretExtra), belongs to our standard toolbox for explainable machine learning. This information is provided in **Supplementary Methods/Multi-parameter modeling of lung function**.

## Point 8

### Issue

Comparison with state-of-the-art studies is missing.

### Response

*Anna and Gerlig: you may consider examples of papers above.*

## Point 9

### Issue

The Practical Implication of the study is missing.

### Response

*Anna and Gerlig: please respod, thanks!*

## Point 10

### Issue

Conclusion also not well presented.

### Response

*Anna and Gerlig: I have few thoughts on conclusions:)* *First of all: we demonstrate the feasibility of a machine learning approach combining CT readouts with demographic and clinical information on acute COVID-19 and convalescence at predicting DLCO.* *Second: this modeling strategy merges information derived from CT images by a software and a human radiologist.* *Third: while we constructed meaningful models of DLCO, our machine learning models failed to predict FEV1 and FVC. This lets us reason that DLCO but not FEV1 or FVC is the prime LFT readout of post-COVID-19 structural lung damage. By univariable analyses we show that DLCO is tightly associated with CT readouts of severity of lung damage such as CTSS, opacity, or high opacity.*

# Reviewer 2

## Point 11

### Issue

The introduction does not sufficiently contextualize how other published multi-parameter or machine learning approaches have succeeded or failed. What specific gaps remain? A more detailed table is needed to perform the literature review.

### Response

*Anna and Gerlig: please respond. From my quite cursory search of the literature, I’ve an impression that there are indeed reports with more or less successful trials of prediction of LFT based on CT data - mostly raw images.* *In turn, I was not able to find a single modeling papers, where someone tried to combine the demographic, clinical, and rariological information to predict LFT outcomes.* *One can say: that the clinic and demography is irrelevant, once you have CT data.* *To me it is not, our combination approach let us understand the clinical background and identify some broad risk hactors such as patient’s sex and comorbidities.*

## Point 12

### Issue

What is new or improved about your multi-parameter modeling approach compared to existing frameworks?

### Response

*Anna and Gerlig: see my thoughts to Point 11.*

## Point 13

### Issue

More details about selection and potential biases (e.g., dropouts, missing visits) are needed for the data set.

### Response

We appreciate this point. While the CovILD cohort has been already extensively described in our previous papers (2–5), we provided in the revised manuscript information on the drop-ou rates at particular time points as compared with the initial collective of n = 145 patients. We also references numbers and percentages for the categorical modeling responses (DLCO < 80%, FEV1 < 80%, and FVC < 80%).

## Point 14

### Issue

It remains unclear how each scoring system aligns, differs, or complements the other.

### Response

*Anna and Gerlig: I’d suggest 1 - 2 sentences in Discussion with emphasis that CTSS is a human-derived discretized severity score averaged over the entire lung, and opacity and high opacity are continuous readouts of lesion extent computed by the software. CTSS is expected to be more coarse the opacity or high opacity and cannot accurately capture severe, highly localized cases such as severe damage restricted to only one lobe (it is scored with 5, the same as minimal damage in all lobes). In our data set, CTSS, opacity, and high opacity are strongly inter-correlated.*

## Point 15

### Issue

Overemphasis on “satisfactory accuracy” without deeper clinical context.

### Response

We apologize for this vague term. In the revised manuscript, avoid suggestive phrases on model performance and stick to teh performance metrics and evaluations.

## Point 16

### Issue

The neural network model performed poorly for DLCO.

### Response

In **Review Figure 1** we present correlations between the observed DLCO expressed as percentages of the patient’s reference and the out-of-fold predictions of DLCO in participant-blocked 10-repeats 10-fold cross-validation for all four machine learning algorithm. To facilitate the comparison, we set ranges of the X and Y axes to the range of observed DLCO values - we applied this transformation to the respective scatter plots in **Figure 2** as well. From this visualization it is that predictions by the neural network algorithm were shrunk to the 75% - 100% range and the neural network algorithm failed to predict cases with very low and very high values of DLCO. We can only speculate on the reasons for this behavior, yet, inclusion of more observations with at the lowest and the highest tail of DLCO is expected to improve performance of not only the neural net but also of the remaining models.

## Point 17

### Issue

The SHAP analysis identifies “top 15” predictors. However, the manuscript does not address correlations among those predictors or discuss whether certain features might be redundant.

### Response

Thanks for this is a very good point. In the initial manuscript, we presented results of exploratory analysis of the modeling data set including the overlap between radiological lung findings, and correlations between opacity, high opacity and CTSS (**Supplementary Figure S3**) - which were also identified as highly influential explanatory factors for prediction of DLCOP < 80% and DLCO (**Figure 3** and **Supplementary Figure S14**). In the revised manuscript, we specifically addressed associations between the most influential explanatory variables with an analysis of correlations graph (6,7) (**Supplementary Figure S15**). The graph edges were defined by pairwise correlations between the explanatory factors with Kendall’s 0.3 and weighted by values. For computation of the correlations, ordinal factors such as presence of GGS of severity of acute COVID-19 were transformed to integers. This analysis supports the findings of the initial manuscript version that the CT readouts: CTSS, opacity, high opacity, GGO and reticulations are tightly inter-correlated. Another cluster of tightly associated features was formed by readouts of severity of ACUTE COVID-19 such as acute COVID-19 severity, hospital and ICU stay, and anti-infective treatment during acute COVID-19. Another community of the influential features was made up by known risk factors of severe COVID-19 course (age, cardiovascular disease, hypercholesterolemia, and endocrine/metabolic disease). This leads us to the conclusion that there was indeed some redundancy between the modeling variables and a potential for pruning e.g. by regularization. These new results were also briefly discussed.

## Point 18

### Issue

With “opacity cutoff: 0.12%” or “high opacity cutoff: 0.002%” appear extremely small. The clinical significance of such minimal changes in lung volume being “abnormal” is unclear—are these thresholds truly meaningful in practice?

### Response

*Anna and Gerlig: please respond.*

## Point 19

### Issue

The paper concedes “model over-fitting” could be an issue. However, it lacks any systematic attempt (e.g., a learning-curve analysis or thorough regularization strategy) to demonstrate that over-fitting is minimal or under control.

### Response

Thanks for raising this important issue. As suggested, we tacked it by a learning curve analysis (8) for the predictions of DLCO < 80% and DLCO by the best performing GBMm models presented in **Supplementary Figure S13**. The over-fitting was particularly evident for the predictions of DLCO < 80% as visualized by a slow convergence of curves of accuracy and Cohen’s for training and test subsets of increasing sizes derived from the modeling data and cross-validation. By contrast, we can argument that over-fitting of for predictions of DLCO expressed as percentages of the patient’s reference was under control. We list over-fitting explicitly as a limitation in the revised manuscript and propose regularized algorithms such as XGBoost as a possible remedy. We we will certainly consider such algorithms in future studies.

In order to investigate the over-fitting behavior of the models of DLCO < 80%, we resorted to an analysis of the cross-validated model’s performance as a function of increasing DLCO percentages. As shown in **Supplementary Figure S12**, we found out that the models made uncertain/erroneous predictions - and hence generalized poorly - for observations with DLCO ranging from 70% - 80% of the patient’s reference. This phenomenon can be explained by the observation that the 80% DLCO cutoff - used in multiple COVID-19 papers (4,9) did not yielded two clearly separated groups of observations in our modeling data set and may not correspond to differences in demographic and clinical background. In addition, we believe that expansion observations with DLCO < 80%, either be increasing the cohort size or augmentation of the data set, e.g. by SMOTE (10), may at least partly correct this behavior.

# Review Figures

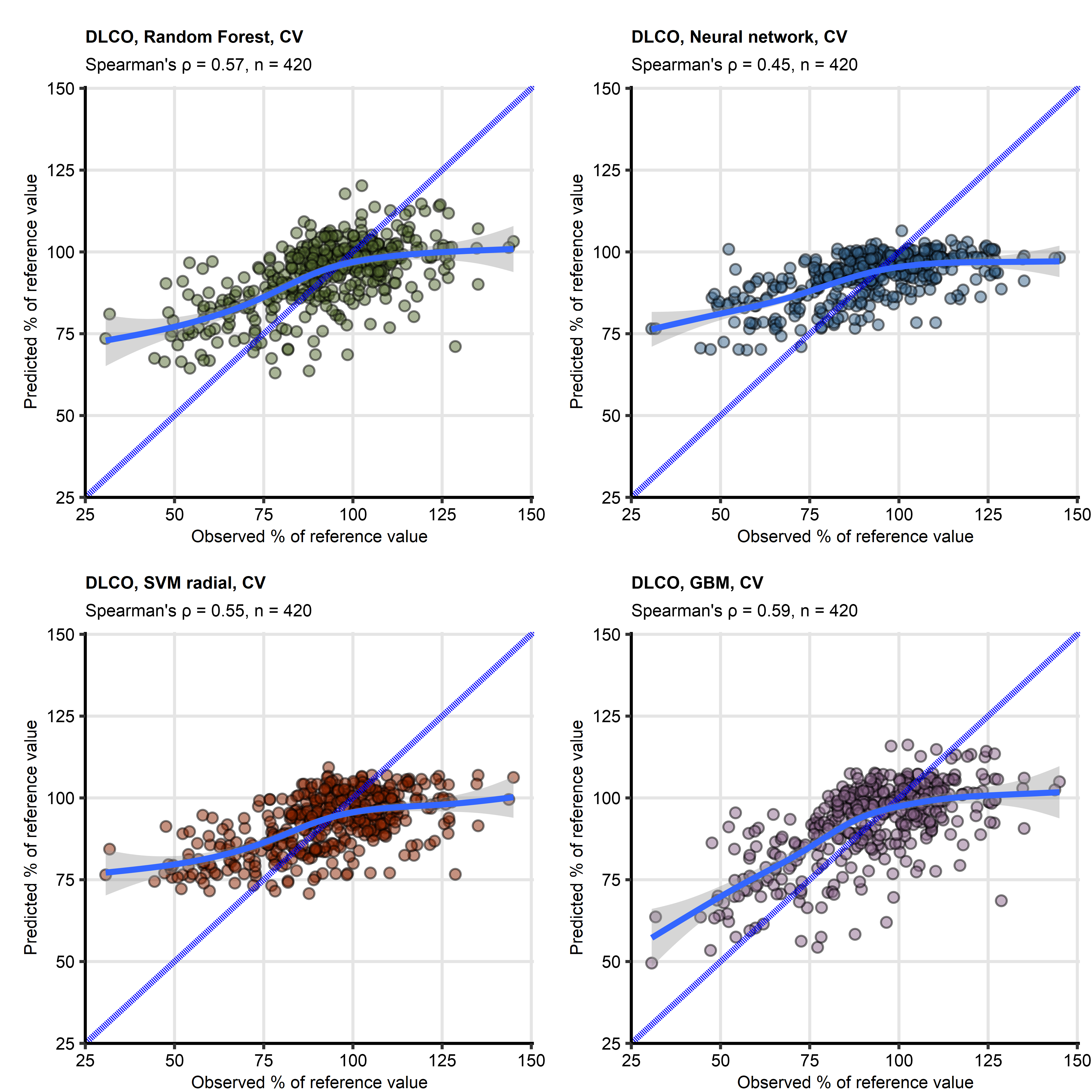


Figure 1: Correlation of predicted and observed diffusion capacity for carbon monoxide during COVID-19 convalescence

**Review Figure 1. Correlation of predicted and observed diffusion capacity for carbon monoxide during COVID-19 convalescence**

*Machine learning models of diffusion capacity for carbon monoxide (DLCO) expressed as percentage of the patient’s reference were developed with the random forest, gradient boosted machines (GBM), support vector machines with radial kernel (SVM), and single hidden layer neural network as presented in Figure 2 of the manuscript.* *Correlations between the predicted and observed DLCO´values were metered by Spearman’s correlation coefficient .* *The observed DLCO values and out-of-fold DLCO predictions (participant-blocked 10-repeats 10-fold cross-validation [CV]) are visualized with scatter plots. Each point represents a single observation. 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 coefficients are displayed in the plot captions. Ranges of displayed DLCO values were set to the range of observed DLCO.*

# References

1. 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)

2. 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)

3. 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)

4. 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)

5. 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)

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

7. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology* (2005) 4: doi: [10.2202/1544-6115.1128/MACHINEREADABLECITATION/RIS](https://doi.org/10.2202/1544-6115.1128/MACHINEREADABLECITATION/RIS)

8. 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)

9. Zhang H, Li X, Huang L, Gu X, Wang Y, Liu M, Liu Z, Zhang X, Yu Z, Wang Y, et al. Lung-function trajectories in COVID-19 survivors after discharge: A two-year longitudinal cohort study. *eClinicalMedicine* (2022) 54:101668. doi: [10.1016/J.ECLINM.2022.101668](https://doi.org/10.1016/J.ECLINM.2022.101668)

10. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* (2002) 16:321–357. doi: [10.1613/JAIR.953](https://doi.org/10.1613/JAIR.953)