Lung Capacity Prediction in Patients with Pulmonary Fibrosis

ABSTRACT

Respiratory failure is the 4th leading cause of death world-wide. After 30 years of age respiratory function declines, seemingly inexorably and exposure to cigarette smoke, infectious agents, atmospheric pollution and genetic predisposition are thought to play important roles in accelerating this process by impinging on the ability of stem and progenitor cells to repair the respiratory epithelium.

Pulmonary fibrosis is a chronic progressive lung disease characterized by scarring of the lung parenchyma. As the disease progresses, lung's ability to oxygenate the blood decreases eventually leading to the death of the patient. Currently, there are no valid tools to predict the prognosis of patients with pulmonary fibrosis.

The goal of this study was to develop a deep learning-based algorithm that, based on patient general information and chest CT images predicts how lung capacity will change over time. An ensemble approach was utilized in which chest images were processed through a convolutional network, while medical features of the patient were analyzed with a fully connected neural network. Numerous approaches were tested, and the best performing models employed transfer learning from the pretrained Vgg16 model.

DATA OVERVIEW AND EDA

Data overview

Dataset is available on Kaggle (https://www.kaggle.com/c/osic-pulmonary-fibrosis-progression) and it was created by combining medical information data of 176 patients from several private and public hospitals. For two patients, CT images were not retrievable and thus they were discarded. For each of the remaining 174 patient the following data was available:

- **Chest CT images.** Every patient was imaged to confirm the diagnosis of pulmonary fibrosis. The number of images per patient varied from a minimum of 12 to a maximum of 1018.

- **Patient clinical information.** This included the sex, age, smoking habit (currently smoking, never smoke or ex-smoker), lung capacity measurements (FVC) and when they were taken relative to the date MRI imaging was performed. FVC stands for Forced Lung Capacity and is the amount of air expressed in milliliters a person can forcefully and quickly exhale after taking a deep breath.

Data wrangling and features extraction

A) MRI images. Since imaging was performed in different and independent hospitals with their own machines following their own standard procedures, CT images differed from each other in:

- Presence of a gray bounding box around the CT image. For some patients, chest images were surrounded by a gray bounding box, which carried no meaningful information and was, therefore, removed (Exhibit 1 center panel).
- Image resolution. Even though 512x512 is the standard CT images resolution, some images were taken at higher resolution (e.g. 768x768) (Exhibit 1 right panel). All images were resized to standard format of 512x512.

- Image contrast. The contrast of images varied from patient to patient (Exhibit 1) and sometimes even within the same patient. A well contrasted image was selected (Exhibit 1 left panel) as reference and the contrast of all the remaining images was normalized on that image.

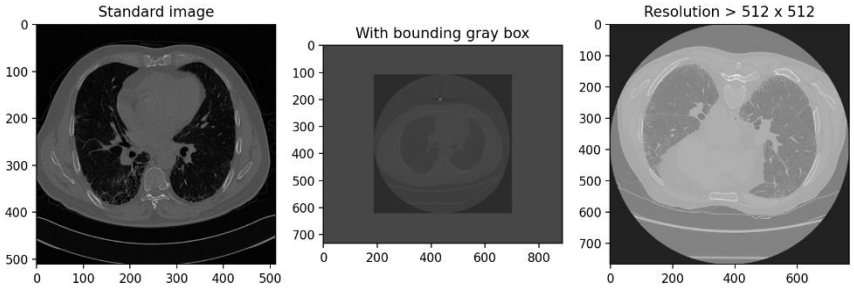

Exhibit 1. CT images formats

- Number of images per patient. Images were provided as sequence and as mentioned above the number of images per patient varied from 12 to a maximum of 1018. Furthermore, some images did not show the lung of the patients but instead displayed the gut, the stomach, the neck or the head of the patient. The first and last images displaying the lung were manually identified and stored in a dictionary for later use.
- Imaging direction. Some patients were imaged feet-first others head-first. The information was retrieved from the DCM files and stored in a dictionary for later use.
- Patient arrangement on the scanning bed. Patients can be arranged on the imaging bed in mainly 4 ways: laying on their back, chest, right side and left side. In the dataset provided, patients were imaged while laying either on their back or chest. Patient arrangement on the bed was extracted from the DCM file and stored in a dictionary for later use.

After images were normalized as mentioned above step the lung tissue was carved out from the image. The portion of the CT image showing the lung was carved out from the surrounding tissues (ribs, muscle, bone, heart and etc.) (Exhibit 2). Since pulmonary fibrosis spread inside the lung, it is safe to assume that predictions made on just the lung tissue would be more accurate than prediction made on all the image. Furthermore, removal of the black background, around the carved images reduces the input size of the deep learning algorithm, leading to faster training and testing.

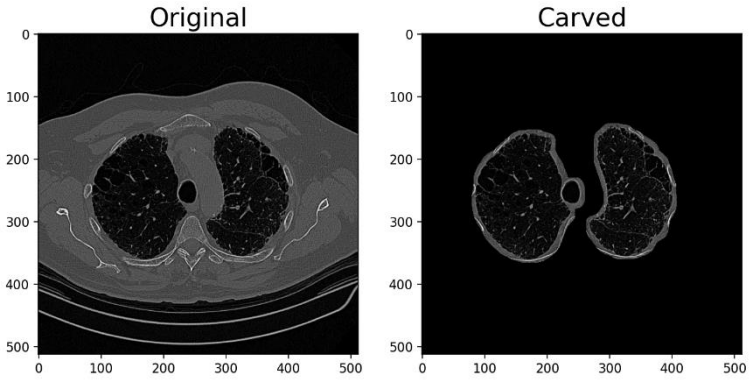

Exhibit 2. Original and lung carved CT images

B) Patient medical information. This dataset did not have missing values but had some outliers in the FVC measurements. For each patient, 6 to 10 lung capacity measurements were provided. We defined outliers those FVC values that were 2.5 standard deviation away from the mean. Outliers were removed unless they were either the first or the last measurement. We assumed that patient's lung capacity could drastically drop

on the last measurement (e.i. the patient is terminal) or after the first measurement (e.i. before any treatment is initiated). Exhibit 3 exemplifies these 3 scenarios.

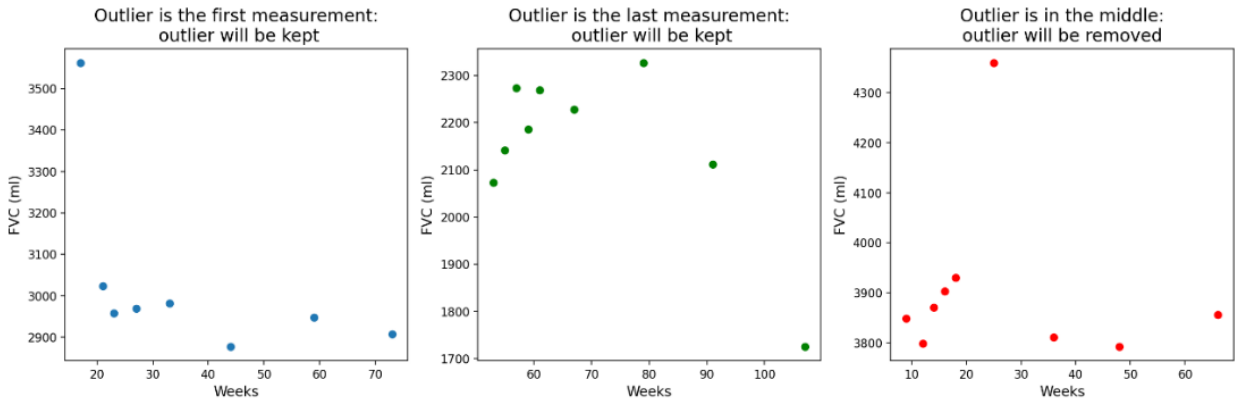

Exhibit 3. Outliers in the FVC measurements and how they were handled

After removal of the outliers, each patient FVC measurements were modeled with a linear regressor and its slope was stored. The slope of the line best approximating the FVC value is the dependent continuous variable the machine learning models will predict.

Explorative data analysis

The number of images showing the lung per patient varied for a minimum of 9 to a max of 227 (Exhibit 4). For each patient, we, thus, selected nine images: the closest one to the intestine, the closest one to the neck and 7 equally spaced images between these two. This approach guaranteed similar representation of the lung in each patient. The 9 images (carved and non carved) were collaged onto one 3x3 grid.

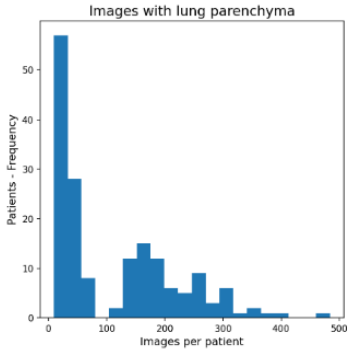

Exhibit 4. Images with lung per patient

Exhibit 5 left panel shows that the age distribution of the patients is normally distributed, with a mean of 67.3 years. Most of the patients (79%) are males (Exhibit 5, center panel) and ex-smokers (67%) (Exhibit 5, right panel). This information is in line with the archetypal patient with lung fibrosis: old male with history of smoking.

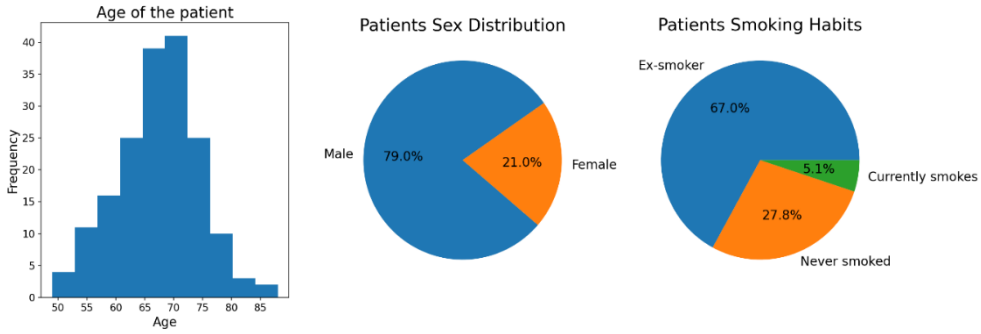

Exhibit 5. Patient medical information: age, sex and smoking habits

Lung capacity is typically proportional to the size of the chest and body of the patient. This fact is illustrated in Exhibit 6 where FVC measurements are plotted against the age of the patient and group by sex. Males have higher FVC values than females within the same age, as they normally have bigger bodies and chests.

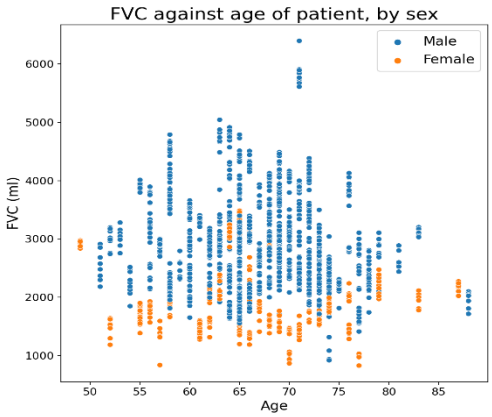

Exhibit 6. Patient lung capacity against age, by sex

The left panel of Exhibit 7 depicts how the patient lung capacity changed over the period of the study. For most of the patients FVC values dropped with time. This is expected since respiratory function normally starts declining at 30 years of age and all the patients in this cohort had lung fibrosis. This is confirmed by the observation that most of the slopes of the line that approximated the FVC values, is negative (Exhibit 7 right panel). However, lung capacity for a small group of patients increased over time. This could be explained by bad measurements or resolution non-chronic respiratory diseases (e.g. cold, pneumonia) which negatively affected the first few measurements.

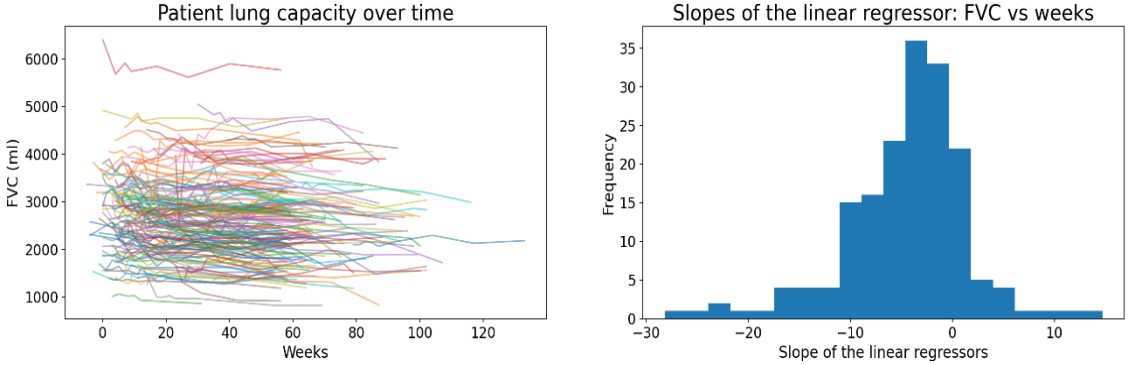

Exhibit 7. Patient lung capacity change over the study period

Finally, area of the lung covered by the left and right lungs was calculated and plotted as ratio over the area covered by the rest of the body (Exhibit 8). Since the most of heart is localized in the left half side of the chest, right lungs are wider and thus have higher transverse area.

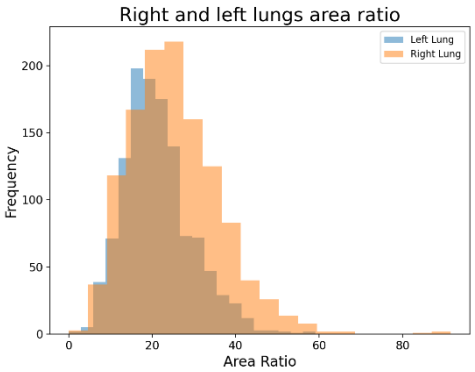

Exhibit 8. Area covered by right and left lung in CT images

MACHINE LEARNING MODELING

Prediction of the patient's lung function can be achieved with a deep learning regressor that, given medical information and chest CT images of the patient, predicts the rate at which the patient lung capacity changes over time. Since at least one FVC measurement is always available, as it is mandatory for diagnostic purposes, the intercept of the line approximating the FVC measurements can be readily calculated and future predictions made.

Several standalone and ensemble models were generated and tested:

- **Fully connected network (FCN)** with patient medical information as inputs. This model leveraged age, sex, smoking status and lungs area ratio data to predict the lung function decay rate.
- **Convolutional neural network** using patient chest CT images as input. Three separate approaches were employed:
  A) Standard approach: Carved and uncarved images, with original and normalized contrast were fed to convolutional neural network with two fully connected layers. In this standard approach several convolutional neural network architectures were tested and their parameters (number of filters, kernel size, etc.) tuned until prediction error plateaued.

  B) Transfer learning. State of the art convolutional neural network models such as Vgg16 and Resnet have been trained on thousands or even millions of images in order. The first few layers of this trained models can be utilized as features extractors in other visual machine learning models. Herein, the first two layers of the trained Vgg16 were extracted and their weights frozen. The network was completed by adding 7 trainable layers (5 convolutional and 2 fully connected layers). Vgg16 is trained on color images, so the gray color CT images were converted to RGB images and downsized to 2/3 (1024x1024 pixels) to reduce computer memory usage.

  C) Autoencoders. Autoencoders have a bottleneck which forces data to be expressed in a compressed version. This approach allows the data to be represented in a compact version with little loss of information in the original data. The 3x3 grid (1536x1536 pixels) with 9 non carved images were processed through four different autoencoders with the bottleneck "feature" layer of dimension 48x48, 96x96, 192x192 and 384x384. Exhibit 9 shows the encoding/decoding loss for each autoencoder. An example of an

unprocessed image, its compressed representation and its reconstructed form is reported for each of the autoencoder in the *Appendix*. The 384x384 and 192x192 autoencoders produced the lowest loss and the best image reconstruction fidelity among the 4 autoencoders. The compressed ('encoded') representations of CT images were used as input to a convolutional neural network whose best architecture and parameters were then identified following the approach mentioned above.

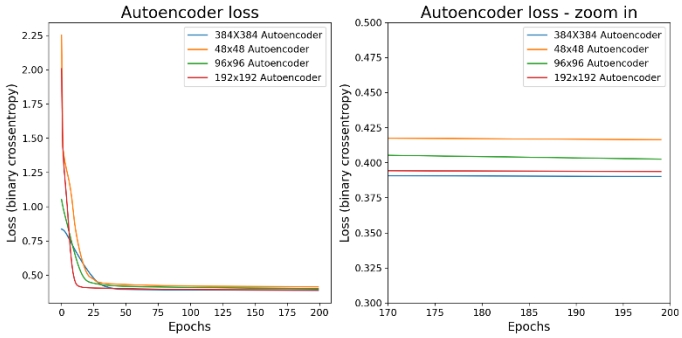

Exhibit 9. Loss of each of the four trained autoencoders.

- **Ensemble models.** Ensemble models usually produce more accurate predictions than standalone models. Thus, predictions made by each **convolutional neural network based model** were combined and averaged with the prediction made by fully connected neural network. This approach generated models that took into account all the information available for the patient.

Slopes of the line approximating the FVC measurements were predicted using the models described above and the first FVC value (baseline value), was used to calculate the intersect of the line. The remaining FVC measurements were then predicted and compared with the observed values. Table 1 summarizes the information of the trained models and their performance.

Model	Input	FVC prediction error: (MAPE)	FVC prediction error when ensembled with NN (MAPE)
NN	Numerical and categorical medical data	7.19%	N/A
CNN-1a	1536x1536 px, non-carved 3x3 images	Original contrast: 7.58% Normalized contrast: 7.33%	Original contrast: 7.33% Normalized contrast: 7.23%
CNN-1b	1120x1460 px carved 3x3 images	Original contrast: 7.34% Normalized contrast: 7.34%	Original contrast: 7.25% Normalized contrast: 7.24%
CNN-2 Encoder	1536x1536 px, non-carved 3x3 images	Enc 48x48, Norm. contrast: 7.24% Enc 96x96, Norm. contrast: 7.34% Enc 192x192, Norm. contrast: 7.44% Enc 384x384, Norm. contrast: 7.32%	Enc 48x48: 7.23% Enc 96x96: 7.22% Enc 192x192: 7.26% Enc 384x384: 7.22%
CNN-3 Transfer Learning	1024x1024 px RGB images 3x3 non-carved	Normalized contrast: 7.06%	Normalized contrast: 7.11%**

Table 1. Model performance summary

The key takeaways are:

- Both patient medical data and chest CT images hold key insights for the prediction of future patient's lung capacity.
- Normalizing the contrast on the images consistently improve the model predictive power. Since the process of adjusting the contrast of every image of the dataset requires time and substantial computing power, we confirmed that this step truly leads to an improvement of the model performance.
- Carving the lung out of the CT image does not consistently improve the model performance. Identifying and slicing the area of the image representing the lung need a noteworthy amount of time, and computer memory, but it does not reduce the prediction of the model.
- Transfer learning is the winning strategy. The extremely well-trained initial layers of the Vgg16 model helped to generate the best standalone model.

Even though CNN1a and CNN1b models were not the best performing we created the feature map for some kernel of every convolutional layers of the model to investigate what the model "sees" in the CT images. Exhibit 10 shows four feature maps of the last convolutional layer. Even though it is not immediately clear, the kernels appear to detect string-like patterns which resemble the shape of the scar tissue in the fibrotic lung.

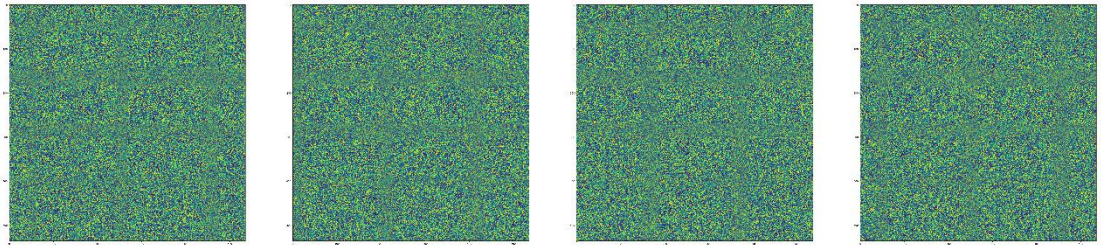

Exhibit 10. Error of the model analysis.

The results from the ensemble model with transfer learning (**) were further analyzed. The left panel of Exhibit 11 left panel shows the mean absolute percentage error (MAPE) for each patient. For three patients, MAPE is above 15%, with one of them approaching the 25%. The error of the model is not correlated with the mean of the observed FVC measurements (Exhibit 11 center panel). However, the longer the patient was enrolled in the study for, the higher the error of the model (Exhibit 11 left panel). This finding is expected because the longer the patient is under observation the higher the chance his health status could suddenly change leading the lung capacity to deviated from the predictions.

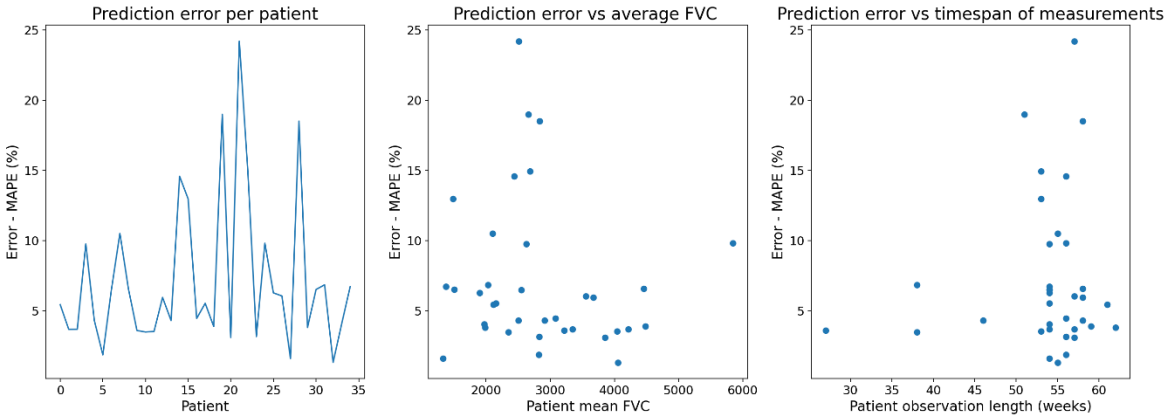

Exhibit 11. Error of the model analysis.

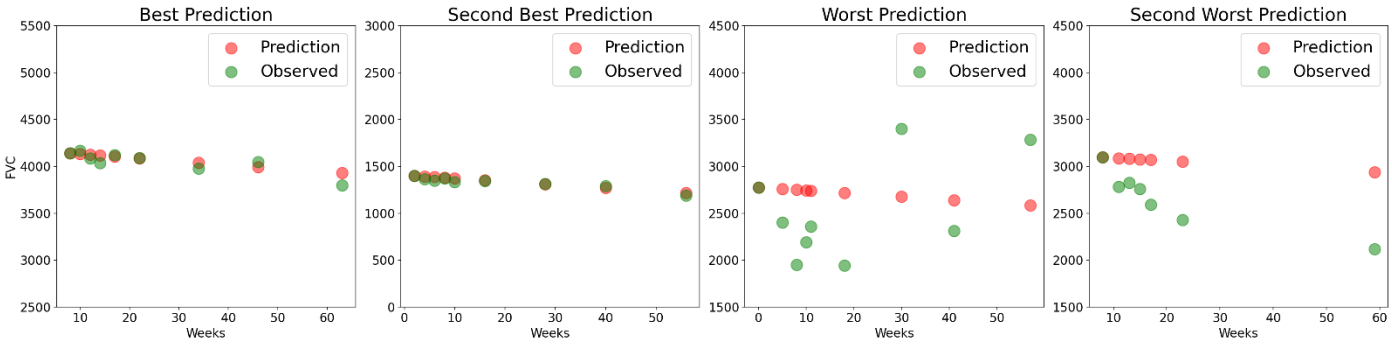

Exhibit 12. Best and worst predictions

The best and worst two predictions were further analyzed (Exhibit 12). For the four plots, the same range for the dependent variable was used to ease visual comparison. When the lung capacity of the patient drops linearly overtime, the model performs well. However, it fails to predict unexpected and drastic drops of the lung function.

Conclusion and future directions.

Both patient medical information and chest CT images hold key insights for the prediction of future patient's lung capacity. Out of all the approaches tested, transfer learning from the pretrained Vgg16 model is the winning approach. However, the model performs well when lung capacity decays linearly, but underperforms otherwise. In the clinical setting, one way to correct for this issue, is to utilize the last FVC measurement (or the average of last n measurements) as baseline value for the following predictions. This approach will allow the model to better adapt to unexpected drastic drops of the patient lung capacity.

The model could be improved by including more information about the patient. For instance, there is no information on the pharmacological treatment (and when it was initiated), presence of comorbidities (such as diabetes, cardiovascular disease, other chronic diseases) and other general information (blood pressure, weight, body mass index). Furthermore, more images could help improving model performance. If the number of images is high enough, the whole 3D volume of the lung could be analyzed producing the most comprehensive strategy for this clinical problem.

In conclusion, with a mean absolute percentage error ranging from 7.06% to 7.11% the transfer learning-based deep learning produces a satisfying result. However, since the worst predictions are obtained for those patients with drastically drops in lung function, the implementation of this model should be carefully well-thought-out and should not substitute the assessment of the patient's physician.

Appendix

A) Encoded representation and reconstructed image for the 4 autoencoders