# ONLINE SUPPLEMENTAL MATERIAL

**Artificial intelligence in CT for quantifying lung changes in the era of CFTR modulators**

**Authors**

Gael Dournes MD-PhD[1,2*], Chase S. Hall MD[3]*, Matthew M. Willmering PhD[4], Alan S. Brody MD[4], Julie Macey MD[2], Stephanie Bui MD[5], Baudouin Denis-De-Senneville PhD[6], Patrick Berger MD-PhD[1,2], François Laurent MD[1,2], Ilyes Benlala MD-PhD[1,2], Jason C. Woods PhD[4,7]

* indicates that both authors contributed the same to the study

## SUPPLEMENTAL METHODS

### Supplemental Method E1. Artificial Intelligence Training Framework

Supplemental Table E1 describes the population characteristics of the 78 cystic fibrosis (CF) patients whose computed tomography (CT) examination was entered in the artificial intelligence (AI) Training dataset. There was a wide range of ages, from 4- to 51-year-old, and a wide range of disease severity, as assessed by forced expiratory volume in 1-second percentage predicted (FEV1%) at pulmonary function test (PFT), from 31 to 114%.

Three CF reference centers from two Institutions were involved: the Adult's Hospital of Haut Leveque (Pessac, France; Site1), the Children's Hospital of Pellegrin (Bordeaux, France; Site2), and Cincinnati Children Hospital Medical Center (Ohio, United States of America; Site3). All three sites correspond to geographically distinct CF reference centers, notably with their medical team and their own CT machines[1]. CT and PFT were performed the same as part of the annual evaluation.

Pulmonary function tests were completed by using a bodyplethysmography devices (site1: Medisoft, Belgium; site2: Jaeger, Germany; site3: SensorMedics, USA). The examinations were performed according to the joint ATS/ERS taskforce guidelines [2], and a daily calibration of devices was routinely performed. Reference values were determined according to Quanjer *et al.* in site 1 and 2[3], and according to Wang *et al*. in Site 3[4]. This evaluation requires the cooperation of the patients, which is not always possible notably in children under the age of 6-year-old[5].

Supplemental Table E2 describes the CT characteristics. There were seven different machine models from 2 major manufacturers over the three sites, namely General Electric (GE) GE Lightspeed 16®, GE LightSpeed VGT®, GE Revolution®, Siemens Somatom Emotion®, Siemens Somatom Sensation 16®, Siemens Somatom Definition 64®, and Siemens Somatom Force®. The matrix was 512*512, the dose-length product ranged from 8 to 260 mGy.cm and the slice thickness from 1 to 1.25 mm.

All patients were thoroughly coached in breathing techniques before each CT scan and CT at full inspiration and reconstructed with standard kernels were used. This methodology choice deserves some comments. A previous study has shown that standard kernel CT noise texture is similar between manufacturers[6] and avoids the high level of noise-induced by "sharp" filters[6]. Second, AI was trained by using inspiratory CT images only. Expiratory CT requires additional radiation exposure and, despite advances in CT reduction of radiation doses, this is not practiced in all CF centers[7–11]. Moreover, inspiratory images are more easily obtained than expiratory images[12], improving reliability and allowing younger patients to provide the necessary cooperation. Importantly, using only inspiratory images decreases radiation exposure by 50%.

*Methodology used for labeling of CT slices*

The annotation of CT slices was done in consensus between three observers of 6, 12, and 25 years of experience in thoracic imaging, who are part of a CF reference center which belongs to the European Cystic Fibrosis Society Clinical Trial Network, and with published expertise in CF scoring of lung CT and MRI[13–17].

Manual segmentation of labels was performed by using the 3D Slicer software 4.11, an open-source software. CT images were displayed with parenchymal window width and level (width, 1500 Hounsfield Unit; level -450 Hounsfield Unit)[18]. Five labels were created to represent five main hallmarks of structural alterations of CF: bronchiectasis, peribronchial thickening, bronchial mucus plugs, bronchiolar mucus plugs with the "tree-in-bud" pattern, and collapse/consolidation[19]. In this study, bronchiectasis refers to the mucus-free airway lumen dilatation, and the bronchial mucus plug was scored when a secretion filled the bronchial lumen entirely. A sixth label was also created, which corresponds to the lung parenchyma, as the total lung minus the sum of other abnormal labels. Bulla or sacculation was also not part of the analysis, the former being a rare abnormality[20] and the second without a definition[19]. One could discuss that bronchiectasis was meant for mucus-free

bronchial lumen dilatation herein. There is not a single definition of bronchiectasis[21]. However, the multilabel method allows flexible evaluations and could enable customized combinations, such as a mix of the airway lumen, airway wall, and mucus alterations, as proposed earlier[22]. In this study, a detailed description of each feature was provided, and we did not attempt to perform such combinations. The pipeline to reach a consensus CT evaluation is illustrated in Figure E1. One observer with 12 years of experience in thoracic imaging and published expertise in CT scoring of CF made the annotations on a slice-by-slice analysis over a full CT acquisition. After recognizing a specific label, the observer had to delineate their shape and extent. Multiplanar reformations and scrolling of CT slices were allowed to identify target structural alterations better. Two independent observers of 6 and 25 years of experience in thoracic imaging had to visually check the segmentations at the segmental level. A segment was considered false-negative if a specific label was missing in a lung segment. Conversely, a false-positive was scored when a label was incorrectly present in a lung segment. Moreover, a visual agreement of more than 80% in the visible spatial extent of true-positive findings was necessary. The threshold of 80% was arbitrary, to take into account the human interobserver reproducibility. The true-negative results from the surrounding lung parenchyma were not considered for visual consensus analysis.

If at least one segment was scored as incorrectly labeled by one observer due to false-positive and/or false-negative labeling or an agreement in the spatial extent of true-positive labels <80%, the CT examination was returned for edits. The process was continued until all observers agree that no false-positive or false-negative lung segments were present in the multilabel segmentation. The visual extent of true-positive matched all three observers by more than 80%. Thus, the CT multilabel segmentation was considered a consensus CT segmentation and entered in the AI framework as "ground-truth" (GT). The mean time to reach a first CT multilabel segmentation was 10 hours (including all labels). The mean time to achieve a consensus CT segmentation was six additional hours, depending on the number of structural lung alterations.

All ground-truth labels were performed randomly, blinded to any other data, and before any AI labeling.

*Description of the AI pipeline*

Convolutional neural networking training was performed on Lambda Labs computer running Ubuntu with ten core I9-9820X processor, 128GB memory, Titan RTX GPU with 24GB

memory. We allocated 23 530 axial CT slices from 78 CF patients' CT scans to create the image analysis pipeline. As mentioned above, they were annotated by the consensus of three expert radiologists as training data. The multilabel segmentation included five classes representing five main hallmarks of structural alteration in the cystic fibrosis lung and a sixth class to characterize the surrounding lung parenchyma. Then, each CT slice was scaled to a value between 0-1. To improve the method's generalizability, we used the Vicinal Risk Minimization principle to train similar but different training data examples through data augmentation[23]. The accompanying segmentation was used to create heuristic data augmentation by applying a deterministic sequence of transformation functions. In our implementation, ten new image/segmentation combinations were obtained by applying affine transformations, including random combinations of shearing, scaling, rotation, and translation. Data augmentation was performed using Keras image data preprocessing tools (available at https://keras.io/api/preprocessing/image/). After augmentation, there were 258 830 unique 2D-CT image and semantic segmentation pairs (1 original plus ten augmented) for neural network training. To further improve generalizability, random pairs of the image/segmentation data were selected to undergo Mixup augmentation[24]. Another 30 000 Mixup image/segmentation pairs were created and made available for neural network training.

A total of 288 830 CT slices data were pooled together, shuffled regardless of the CT scan they were originally coming from, and then split randomly 80%/20% as training and validation datasets for neural network optimization. Three two-dimensional (2D) convolutional neural network (CNN) architectures were trained based on the popular U-Net model with different backbone architectures. These included:

1) InceptionResNetv2 (Model 1) is a convolutional neural architecture that builds on the Inception family of architectures but incorporates residual connections, replacing the filter concatenation stage of the Inception architecture[25];

2) ResNet50 (Model 2) is a convolutional neural network that is 50 layers deep and uses residual learning[26];

3) the classic U-net (Model 3) is a convolutional neural network, where the main principle is to supplement a usual contracting network by successive layers, where polling operators are replaced by upsampling operators[27].

These models were chosen for two main reasons. First, the three of them are known to have made such significant contributions to the field of imaging segmentation that they have

become widely considered as current standards[28]. Thus, they are commonly used as building blocks for many segmentation architectures[29]. Second, their backbone architectures are different; thus, their segmentation result is not expected to be entirely similar, allowing a Majority Vote ensemble of different classifiers[30].

The optimizer algorithm selected was Adam, a replacement optimization algorithm for stochastic gradient descent for training deep learning models[31]. The loss function was combined with categorical cross-entropy and Dice[32] by taking into account the overall performance of the six labels. The Input shape was (512x512x1), and the Output shape was (512x512x7). The batch size was 3, and 15 epochs were performed.

Finally, to improve segmentation consistency, a majority vote[33] of the three outputs was performed at each pixel to determine the final semantic multilabel segmentation using ANTs (https://github.com/ANTsX/ANTs). The rationale is as follows:

The rationale is as follows:

- Assume *n* independent classifiers with an error rate $\epsilon$.
- Assume a binary classification task (yes/no)
- Assume the error rate of each independent classifier is better than random guessing (*i.e.*, $\epsilon$ is lower than 0.5 for each binary classification)

  Let $X_k (1 \leq k \leq n)$be a Bernouilli variable: $X_k = 0$if the classifier $k$ makes a good prediction (this happens with a probability 1-$\epsilon$) and $X_k = 1$if the classifier $k$ makes a wrong prediction (this happens with a probability $\epsilon$).

  Let $X = \sum_{k=1}^{n} X_k$be the number of classifier that make a wrong prediction. $X$is a Binomial variable and we have:

$$P(X = k) = \binom{n}{k} \epsilon^k (1-\epsilon)^{n-k}$$

  Therefore, the probability that we make a wrong prediction via the ensemble on *n* classifier is equal to:

$$P\left(X > \frac{n}{2}\right) = \sum_{k=\lceil n/2 \rceil}^{n} \binom{n}{k} \epsilon^k (1-\epsilon)^{n-k}$$

  As a consequence, by making the assumptions mentioned above, it is expected that the majority voting error of an ensemble of *n* independent classifiers converges toward 0 as long as the number *n* of classifiers increase.

That being said, we have used three models, mainly because we have used a “hard voting” system instead of “soft voting” system. Indeed, we have not weighted the prediction made per each model. Thus, using a hard voting system, any pair number of models would lead to the possibility of equality between classifiers, and thus, unlabeled pixels. In this implementation, we have chosen to assign a label to all pixels.

However, other methods of voting systems could be implemented and tested in next studies or other groups, for instance soft voting systems or a number of models higher than three. However, one could also expect that the time required to get the final results will be necessarily much higher by using more than three models.

*Pilot evaluation of the manual segmentations chosen as Ground Truth*

Supplemental Table E3 shows the result of a pilot statistical analysis performed in the Training data set. It shows that all labels from the consensus CT segmentations significantly correlated to other well-established biomarkers of the lung disease severity, notably FEV1% at PFT, and a modified CT Brody system at CT (Supplemental Table E4)[34], with all p-value from all labels being $\leq 0.001$.

A modified version of Brody and colleagues' original scoring system (24) was necessary since expiratory CT was not performed in two of the three CF reference centers. Thus, the feature of air trapping was not available for analysis as a sub-score and was not part of the visual CT scoring evaluation. In the Training dataset, the visual modified Brody score of anonymized CTs was established by Obs3, blinded to any other data.

**Supplemental Method E2. Test Cohort evaluation**

All CF patients from the external Test cohort were not part of the Training cohort. All manual ground-truth CT labels in the Test dataset were done using the same method as in the Training dataset and before any AI segmentation.

Since the AI-driven quantification was performed using 2D-CNNs, an evaluation of the similarity between AI-driven segmentation and GT labels was planned via a pixel-by-pixel 2D-similarity assessment over 11345 CT slices of 36 patients CTs, after anonymization, blinded from any other data. For this, all 2D-axial CT slices were shuffled randomly

altogether before being segmented by the 2D-CNNs. True-positive (TP), true-negative (TN), false-positive (FP), and false-negative results (FN) were counted and summed over the full dataset of 11345 CT slices to calculate the balanced accuracy, Sorensen-Dice coefficient, recall, and precision, as reported earlier[35].

The standard formula of calculation were as follows:

Dice = 2*TP / (2*TP + FP + FN)

Recall = TP / (TP + FN)

Precision = TP / (TP + FP)

True negative rate = TN / (TN + FP)

Balanced accuracy = (Recall + True negative rate) / 2

However, in the specific field of lung CT of CF airways, the similarity metric evaluations have to deal with specific issues as follows:

- The study deals with 2D algorithms. Thus, the unit of measurement is the pixelwise similarity, since there is no 3D information in the code of the CNNs, neither for Training nor for Test purposes.

- The full CT examinations of all patients were used, to enable an extensive overview of the model performance over a full CT scan, without any pre-selection of some CT slices. This includes both tasks of detecting and ruling out the disease presence or absence.

- However, it is known that there is a vast heterogeneity in the regional distribution of structural abnormalities. Therefore, the five structural abnormalities were heterogeneously present/absent across the stack of CT slices.

- Moreover, each label were not segmented by 6 different AI algorithm, but the same AI algorithm in a multilabeling fashion. Thus, six labels (including the normal lung parenchyma) plus the background image (extrapulmonary pixels) were considered to perform probability maps per each CT slice by the same AI algorithm, before allocating a single label per each pixel of the CT slice image.

- In addition, it is known that the similarity metrics cannot be considered similarly when dealing with small or large structures. Owing to the known vast heterogeneity in size and shape of structural abnormalities, the metrics would not have the same meaning from one CT slice to another[36].

Therefore, the heterogeneity of distribution of lesions does not allow to provide the results as a mean per CT slice with standard deviation. Notably, the heterogenous distribution of lesions would inevitably lead to a substantial amount of 0 divisions in the calculations, thus a mathematical impossibility to calculate the metrics. In addition, the heterogeneity in size and shape of the structural abnormalities would also lead to mix similarity results that would not have the same meaning from one evaluation to another. Thus, such approach would also lead to inconsistent and uninterpretable results[36].

This is why we have performed the similarity evaluation by using a spatial overlap calculation over the full set of CT slices[35]. By doing so, one could remark that the uncertainty of the result is expected to be negligible, since it is performed over 512x512 pixels per CT slice, over 11435 CT slices herein.

Indeed, the mathematical formula of the 95% confidence interval would be:

$P = p \pm 1.96 \sqrt{[p(1-p)/n]}$ where P is the maximum or minimum limit of the 95% confidence interval of a ratio, p the measured ratio, and n is the number of evaluations (herein the number of pixels).

Thus, we have assumed that the 95% confidence interval of the pixelwise similarity metrics are negligible.

Finally, the Total Abnormal Lung's similarity result was calculated as the mean of the five label results, related to bronchiectasis, peribronchial thickening, bronchial mucus, bronchiolar mucus, and collapse/consolidation measurements.

Then, the shuffled CT slices were re-assigned to their initial CT examination, and the volume of labels was calculated per each CT scan according to the volume of positive findings of each label, and expressed as a volume in milliliters. One could remark that these volumetric measurements are original, as compared the standard cross-sectional measurements of airways, which represents the plain area of a single cross-section along a bronchial path[14].

The Total Abnormal Volume was defined as the sum of the five structural alteration volumes per CT scan. The Total Lung Volume was defined as the sum (Total Abnormal Volume + Lung Parenchyma Volume).

To take into account variations in lung volumes, notably between children and adults, or related to lung growth over time in children and teenagers, normalization was performed as follows: Normalized Volume of Label(y) = [Volume of Label(y) / Total Lung Volume] x $10^4$. The factor $10^4$ was done to take into account the expected magnitude of volume difference between the normal central airway tree at the segmental level and the lung volume[37].

**Supplemental Method E3. Visual CT scorings.**

As mentioned above, we used a modified Brody score on CT[34] (Supplemental Table E4).

Two separate sessions were done: the first session was dedicated to CTs of the Test cohort, and the second session was dedicated to CTs of the Clinical Validation cohort.

Per each session, anonymized CTs of a given cohort were analyzed randomly by Obs1 and Obs2, independently and blinded to any other data. The mean of both evaluations was kept for further analysis. The time required to perform a CT Brody score ranges between 15 to 20 minutes.

**Supplemental Method E4. Reproducibility and repeatability of evaluations.**

To assess the reproducibility of AI evaluations, the 140 CTs of the Clinical Validation cohort were runned on two different computers:

- An “advanced” computer, with the following characteristics: Lambda Labs computer running Ubuntu with ten core I9-9820X processor, 128GB memory, Titan RTX GPU with 24GB memory
- A “standard” computer, with the following characteristics: Dell computer running Windows 10 with I7-6700 processor, 32 GB memory, GeForce GT 730 with 2 GB memory.

The repeatability of AI evaluation was also assessed by repeating twice the 140 CTs by using the advanced computer.

Moreover, a random subset of 8 patients' CTs (e-Table5) was segmented independently by Observer 1 and 2 with 6 and 12 years of experience, respectively, to assess the manual interobserver reproducibility. The same dataset was manually segmented a second time by Observer 2, 6 months apart from the first evaluation, to assess the intra-observer repeatability. Observer 1 and 2 were the same observers than those who were part of the Training evaluations.

## SUPPLEMENTAL REFERENCES


1. Muco CFTR. Centres de référence de lutte contre la mucoviscidose. https://muco-cftr.fr/index.php/fr/la-filiere/la-filiere-muco-cftr/les-acteurs-de-la-filiere/8-la-filiere-muco-cftr/164-listes-des-centres-mucoviscidose. Last accessed March 01, 2021.

2. Miller MR, Hankinson J, Brusasco V, Burgos F, Casaburi R, Coates A, Crapo R, Enright P, van der Grinten CPM, Gustafsson P, Jensen R, Johnson DC, MacIntyre N, McKay R, Navajas D, Pedersen OF, Pellegrino R, Viegi G, Wanger J, ATS/ERS Task Force. Standardisation of spirometry. *Eur Respir J* 2005; 26: 319–338.

3. Quanjer PH, Stanojevic S, Cole TJ, Baur X, Hall GL, Culver BH, Enright PL, Hankinson JL, Ip MSM, Zheng J, Stocks J, the ERS Global Lung Function Initiative. Multi-ethnic reference values for spirometry for the 3–95-yr age range: the global lung function 2012 equations. *Eur Respir J* 2012; 40: 1324–1343.

4. Wang X, Dockery DW, Wypij D, Fay ME, Ferris BG. Pulmonary function between 6 and 18 years of age. *Pediatr Pulmonol* 1993; 15: 75–88.

5. Marostica PJC, Weist AD, Eigen H, Angelicchio C, Christoph K, Savage J, Grant D, Tepper RS. Spirometry in 3- to 6-Year-Old Children with Cystic Fibrosis. *Am J Respir Crit Care Med* 2002; 166: 67–71.

6. Solomon JB, Christianson O, Samei E. Quantitative comparison of noise texture across CT scanners from different manufacturers: Quantitative comparison of noise texture across CT scanners. *Med Phys* 2012; 39: 6048–6055.

7. Ronan NJ, Einarsson GG, Twomey M, Mooney D, Mullane D, NiChroinin M, O'Callaghan G, Shanahan F, Murphy DM, O'Connor OJ, Shortt CA, Tunney MM, Eustace JA, Maher MM, Elborn JS, Plant BJ. CORK Study in Cystic Fibrosis: Sustained Improvements in Ultra-Low-Dose Chest CT Scores After CFTR Modulation With Ivacaftor. *Chest* 2018; 153: 395–403.

8. Chassagnon G, Martin C, Burgel P-R, Hubert D, Fajac I, Paragios N, Zacharaki EI, Legmann P, Coste J, Revel M-P. An automated computed tomography score for the cystic fibrosis lung. *Eur Radiol* 2018; 28: 5111–5120.

9. Delacoste J, Feliciano H, Yerly J, Dunet V, Beigelman-Aubry C, Ginami G, van Heeswijk RB, Piccini D, Stuber M, Sauty A. A black-blood ultra-short echo time (UTE) sequence for 3D isotropic resolution imaging of the lungs. *Magn Reson Med* 2019; 81: 3808–3818.

10. Bhalla M, Turcios N, Aponte V, Jenkins M, Leitman BS, McCauley DI, Naidich DP. Cystic fibrosis: scoring system with thin-section CT. *Radiology* 1991; 179: 783–788.

11. Helbich TH, Heinz-Peer G, Eichler I, Wunderbaldinger P, Götz M, Wojnarowski C, Brasch RC, Herold CJ. Cystic fibrosis: CT assessment of lung involvement in children and adults. *Radiology* 1999; 213: 537–544.

12. Lucaya J, García-Peña P, Herrera L, Enríquez G, Piqueras J. Expiratory Chest CT in Children. *Am J Roentgenol* 2000; 174: 235–241.

13. Dournes G, Berger P, Refait J, Macey J, Bui S, Delhaes L, Montaudon M, Corneloup O, Chateil J-F, Marthan R, Fayon M, Laurent F. Allergic Bronchopulmonary Aspergillosis in Cystic Fibrosis: MR Imaging of Airway Mucus Contrasts as a Tool for Diagnosis. *Radiology* 2017; 285: 261–269.

14. Montaudon M, Berger P, Cangini-Sacher A, de Dietrich G, Tunon-de-Lara JM, Marthan R, Laurent F. Bronchial measurement with three-dimensional quantitative thin-section CT in patients with cystic fibrosis. *Radiology* 2007; 242: 573–581.

15. Dournes G, Menut F, Macey J, Fayon M, Chateil J-F, Salel M, Corneloup O, Montaudon M, Berger P, Laurent F. Lung morphology assessment of cystic fibrosis using MRI with ultra-short echo time at submillimeter spatial resolution. *Eur Radiol* 2016; 26: 3811–3820.

16. Refait J, Macey J, Bui S, Fayon M, Berger P, Delhaes L, Laurent F, Dournes G. CT evaluation of hyperattenuating mucus to diagnose allergic bronchopulmonary aspergillosis in the special condition of cystic fibrosis. *J Cyst Fibros* 2019;18(4):e31-e36.

17. Benlala I, Point S, Leung C, Berger P, Woods JC, Raherison C, Laurent F, Macey J, Dournes G. Volumetric quantification of lung MR signal intensities using ultrashort TE as an automated score in cystic fibrosis. *Eur Radiol* 2020;30(10):5479-5488.

18. Lederlin M, Laurent F, Portron Y, Ozier A, Cochet H, Berger P, Montaudon M. CT Attenuation of the Bronchial Wall in Patients With Asthma: Comparison With Geometric Parameters and Correlation With Function and Histologic Characteristics. *Am J Roentgenol* 2012; 199: 1226–1233.

19. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Müller NL, Remy J. Fleischner Society: Glossary of Terms for Thoracic Imaging. *Radiology* 2008; 246: 697–722.

20. Brody AS, Sucharew H, Campbell JD, Millard SP, Molina PL, Klein JS, Quan J. Computed tomography correlates with pulmonary exacerbations in children with cystic fibrosis. *Am J Respir Crit Care Med* 2005; 172: 1128–1132.

21. Tiddens HAWM, Meerburg JJ, van der Eerden MM, Ciet P. The radiological diagnosis of bronchiectasis: what's in a name? *Eur Respir Rev* 2020; 29(156):190120.

22. Eichinger M, Optazaite D-E, Kopp-Schneider A, Hintze C, Biederer J, Niemann A, Mall MA, Wielpütz MO, Kauczor H-U, Puderbach M. Morphologic and functional scoring of cystic fibrosis lung disease using MRI. *Eur J Radiol* 2012; 81: 1321–1329.

23. Cao Y, Rockett PI. The use of vicinal-risk minimization for training decision trees. *Appl Soft Comput* 2015; 31: 185–195.

24. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. mixup: Beyond Empirical Risk Minimization. http://arxiv.org/abs/1710.09412. Last accessed March 01, 2021.

25. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. http://arxiv.org/abs/1602.07261. Last accessed March 01, 2021.

26. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. http://arxiv.org/abs/1512.03385. Last accessed March 01, 2021.

23. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. http://arxiv.org/abs/1505.04597. Last accessed March 01, 2021.

28. Khan A, Sohail A, Zahoora U, Qureshi AS. A survey of the recent architectures of deep convolutional neural networks. *Artif Intell Rev* 2020; 53: 5455–5516.

29. Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Martinez-Gonzalez P, Garcia-Rodriguez J. A survey on deep learning techniques for image and video semantic segmentation. *Appl Soft Comput* 2018; 70: 41–65.

30. Atallah R, Al-Mousa A. Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method. https://ieeexplore.ieee.org/document/8923053. Last accessed May 31, 2021.

31. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. http://arxiv.org/abs/1412.6980. Last accessed March 01, 2021.

32. Marques F, de Bruijne M, Dubost F, Tiddens HAW, Kemner-van de Corput M. Quantification of lung abnormalities in cystic fibrosis using deep networks. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10574/2292188/Quantification-of-lung-abnormalities-in-cystic-fibrosis-using-deep-networks/10.1117/12.2292188.full. Last accessed March 01, 2021.

33. Chassagnon G, Vakalopoulou M, Battistella E, Christodoulidis S, Hoang-Thi T-N, Dangeard S, Deutsch E, Andre F, Guillo E, Halm N, El Hajj S, Bompard F, Neveu S, Hani C, Saab I, Campredon A, Koulakian H, Bennani S, Freche G, Barat M, Lombard A, Fournier L, Monnier H, Grand T, Gregory J, Nguyen Y, Khalil A, Mahdjoub E, Brillet P-Y, Tran Ba S, et al. AI-driven quantification, staging and outcome prediction of COVID-19 pneumonia. *Med Image Anal* 2021; 67: 101860.

34. Brody AS, Klein JS, Molina PL, Quan J, Bean JA, Wilmott RW. High-resolution computed tomography in young patients with cystic fibrosis: distribution of abnormalities and correlation with pulmonary function tests. *J Pediatr* 2004; 145: 32–38.

35. Zhou X, Ito T, Takayama R, Wang S, Hara T, Fujita H. Three-Dimensional CT Image Segmentation by Combining 2D Fully Convolutional Network with 3D Majority Voting. http://link.springer.com/10.1007/978-3-319-46976-8_12.Last accessed May 31, 2021.

36. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging* 2015; 15: 29.

37. Gupta S, Hartley R, Khan UT, Singapuri A, Hargadon B, Monteiro W, Pavord ID, Sousa AR, Marshall RP, Subramanian D, Parr D, Entwisle JJ, Siddiqui S, Raj V, Brightling CE. Quantitative computed tomography-derived clusters: redefining airway remodeling in asthmatic patients. *J Allergy Clin Immunol* 2014; 133: 729-738.e18.

## SUPPLEMENTAL TABLES

**Table E1. Characteristics of 78 cystic fibrosis patients in the Training dataset**

| | | Training dataset |
|---|---|---|
| **Age** | Years | 21 (4-51) |
| **Gender** | Male/Female | 36/42 |
| **Body mass index** | $kg.m^{-2}$ | 19 (12-28) |
| **Pulmonary function tests** | FEV1% | 74 (31-114) |
| | 100xFEV1/FVC | 73 (38-92) |
| | 100xRV/TLC | 41 (17-106) |
| **Visual CT score** | mBrody score | 40 (0-151) |

Data are median with (minimum-maximum) range of values

Legends: FEV1=forced expiratory volume in 1 second; FVC=forced vital capacity; RV=residual volume; TLC=total lung capacity; %=percentage predicted; mBrody=modified Brody score.

**Table E2. Characteristics of CT scans**

| Dataset | Machine model | Kernel | Reconstruction | DLP (mGy.cm) | kV | mAs | Slice thickness (mm) |
|---|---|---|---|---|---|---|---|
| **Training** | Somatom Sensation 16® (Site1, n=7 ; Site2, n=9)<br>Somatom Definition 64® (Site1, n=8 ; Site2, n=10)<br>Somatom Force® (Site1, n=9)<br>Somatom Emotion® (Site3, n=4)<br>GE LightSpeed 16® (Site3, n=9)<br>GE LightSpeed VGT® (Site3, n=6)<br>GE Revolution® (Site2, n=16) | STD (n=31)<br>B40s (n=20)<br>Br40 (n=7)<br>I30f (n=20) | FBP (n=42)<br>ASiR (n=16)<br>SAFIRE (n=20) | (8-260) | (100-140) | (5-40) | (1-1.25) |
| **Test** | Somatom Sensation 16® (Site1, n=2 ; Site2, n=5)<br>Somatom Definition 64® (Site1, n=3 ; Site2, n=6)<br>Somatom Force® (Site1, n=5)<br>Somatom Emotion (Site3, n=1)<br>GE LightSpeed® 16 (Site3, n=2)<br>GE LightSpeed VGT® (Site3, n=2)<br>GE Revolution® (Site2, n=10) | STD (n=14)<br>B40s (n=8)<br>Br40 (n=5)<br>I30F (n=9) | FBP (n=12)<br>ASiR (n=10)<br>SAFIRE (n=14) | (9-210) | (100-140) | (5-40) | (1-1.25) |
| **Clinical Validation** | Somatom Sensation 16® (Site1, n=28 ; Site2, n=35)<br>Somatom Definition 64® (Site1, n=37 ; Site2, n=40) | B40s (n=53)<br>Br40 (n=22)<br>I30F (n=65) | FBP (n=75)<br>SAFIRE (n=65) | (12-64) | 110 | (5-54) | 1 |

Legend: Site1=Adult Hospital of Haut Levêque (Pessac, France); Site2=Children Hospital of Pellegrin (Bordeaux, France); Site3=Cincinnati Children Hospital Medical Center (Ohio, United States of America); GE=General Electric®; STD=standard kernel; FBP=filtered-back projection; ASiR=adaptive statistical iterative reconstruction; SAFIRE=sinogram affirmed iterative reconstruction; kV=kilovoltage, mAs=milliampere second; DLP=dose length product; for kV, mAs and pixel size, data between parentheses are the (minimum-maximum) range of values.

**Table E3. Correlation between structural abnormality volumes, lung function, and structural severity in the Training dataset.**

| Normalized volumes | Manual segmentation | | | |
|---|---|---|---|---|
| | FEV1% | | mBrody score | |
| | rho | p-value | rho | p-value |
| **Bronchiectasis** | -0.45 | **0.001** | 0.72 | **<0.001** |
| **Peribronchial thickening** | -0.49 | **<0.001** | 0.70 | **<0.001** |
| **Bronchial mucus plug** | -0.64 | **<0.001** | 0.67 | **<0.001** |
| **Bronchiolar mucus plug** | -0.46 | **<0.001** | 0.69 | **<0.001** |
| **Collapse/Consolidation** | -0.35 | **0.001** | 0.39 | **<0.001** |
| **Total Abnormal Volume** | -0.60 | **<0.001** | 0.79 | **<0.001** |

Note: Data are Spearman's rho coefficient of correlation. The Total Abnormal Volume corresponds to the sum of five structural alteration volumes. Normalized volumes were obtained by dividing a given structural alteration volume by the corresponding total lung volume.

Legends: FEV1%=forced expiratory volume in 1-second percentage predicted; mBrody=modified Brody score

**Table E4. Brody HRCT score** (reproduced from the original publication by *A. S. Brody et al. J Pediatr 2004).*

| Parameter | Calculation |
|---|---|
| **Bronchiectasis score (0-12)** | (Extent of bronchiectasis in central lung + Extent of bronchiectasis in peripheral lung) x Average bronchiectasis size multiplier [0.5 = 0; 1 = 1; 1.5 = 1.25; 2.0 = 1.5; 2.5 =1.75; 3 = 2]<br>where<br>Average bronchiectasis size = ( Size of largest dilated bronchus + Average size of dilated bronchus )/2 |
| **Mucus plugging score (0-6)** | The extent of mucous plugging in central lung + Extent of mucous plugging in peripheral lung |
| **Peribronchial thickening score (0-9)** | (Extent of peribronchial thickening in central lung + Extent of peribronchial thickening in peripheral lung) x Severity of peribronchial thickening [1 = mild; 1.25 = moderate; 1.5 = severe] |
| **Parenchyma score (0-9)** | The extent of dense parenchymal opacity + Extent of ground-glass opacity + Extent of cysts or bullae |
| **Air trapping score (0-4.5)** | Extent of air trapping x Appearance of air trapping [1 = subsegmental; 1.5 = segmental or larger] |

Finding extent scoring: absent (0), 1/3 of the lobe (1), 1/3 to 2/3 of the lobe (2), more than 2/3 of the lobe (3)
Bronchiectasis Severity: less than 2X adjacent vessel (1), 2x to 3x adjacent vessel (2), more than 3X adjacent vessel (3)

**Parameters' definitions**

1.Bronchiectasis: one or more of the following criteria: a broncho arterial ratio >1, a non-tapering bronchus, a bronchus within 1 cm of the costal pleura, or a bronchus abutting the mediastinal pleura

2. Peribronchial thickening: bronchial wall thickness >2 mm in the hila, 1 mm in the central portion of the lung, and 0.5 mm in the peripheral lung

3. Mucus plugging: Central mucous plugging was defined as an opacity filling a defined bronchus, and peripheral mucous plugging was defined as the presence of either dilated mucous-filled bronchi or peripheral thin branching structures or centrilobular nodules in the peripheral lung

4. Air trapping: areas of the lung on the expiratory images that remained similar in attenuation to the appearance on inspiratory images

***Note: in this study, we used a modified version of the scoring system, and the feature of air trapping was not scored. Indeed, in this retrospective study, expiratory CT was not performed in 2/3 sites.***

**Table E5. Characteristics of 8 cystic fibrosis patients of the Clinical Validation cohort for interobserver manual similarity assessments.**

| | | **N=8** |
|---|---|---|
| **Age** | Years | 12 (6-42) |
| **Gender** | Male/Female | 3/5 |
| **Body mass index** | $kg.m^{-2}$ | 17 (13-21) |
| **Pulmonary function tests** | FEV1% | 68 (38-95) |
| | 100xFEV1/FVC | 77 (51-101) |
| | 100xRV/TLC | 42 (24-85) |
| | mBrody score | 115 (0-152) |

Data are median with (minimum-maximum) range of values

Legends: FEV1=forced expiratory volume in 1 second; FVC=forced vital capacity; RV=residual volume; TLC=total lung capacity; %=percentage predicted; mBrody=modified Brody score.

**Table E6. Background therapeutic management in the Clinical Validation cohort.**

| | | Clinical Validation Cohort n=70 | |
|---|---|---|---|
| | | n=10 patients with lumacaftor/ivacaftor | n=60 patients without lumacaftor/ivacaftor |
| **Inhaled treatment** | Antibiotics | 3 | 24 |
| | LABA | 4 | 15 |
| | Corticosteroid | 4 | 15 |
| | Mucolytic | 7 | 43 |
| **Oral treatment** | Antibiotics | 0 | 5 |
| | Corticosteroids | 0 | 0 |
| | Antifungal | 0 | 4 |
| **Intravenous treatment** | Antibiotics | 0 | 5 |
| | Corticosteroids | 0 | 0 |
| | Antifungal | 0 | 0 |

Data are the absolute number of patients with a given chronic treatment.

Legends: LABA=long-acting beta-agonist.

**Table E7. Description of the volume of the six labels in the Test cohort per each CT slice, in milliliters.**

| Labels | AI segmentation | | | | | Manual segmentation | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Median | IQR | 95% CI | Minimum | Maximum | Median | IQR | 95% CI | Minimum | Maximum |
| **Bronchiectasis** | 0.0005 | 0-0.06 | 0-0.04 | 0 | 2.4 | 0.0005 | 0-0.08 | 0-0.05 | 0 | 3 |
| **Peribronchial thickening** | 0.001 | 0-0.01 | 0-0.5 | 0 | 2.4 | 0 | 0-0.01 | 0-0.7 | 0 | 2.5 |
| **Central mucus** | 0.0005 | 0-0.05 | 0-0.4 | 0 | 1.3 | 0 | 0-0.07 | 0-0.4 | 0 | 1.5 |
| **Peripheral mucus** | 0.001 | 0-0.03 | 0-0.1 | 0 | 1.9 | 0.001 | 0-0.09 | 0-0.5 | 0 | 2 |
| **Collapse consolidation** | 0 | 0-0.08 | 0-0.2 | 0 | 4.3 | 0 | 0-0.09 | 0-0.4 | 0 | 4.4 |
| **Lung parenchyma** | 21.3 | 0-34.6 | 0-41.9 | 0 | 50.8 | 21.2 | 0-34.2 | 0-41.9 | 0 | 50 |

Note: data corresponds to the volume per each CT slice, and expressed in milliliters.

The summary characteristics were calculated from 11435 CT slices of 36 CF patients' CT

Legend: AI=artificial intelligence; IQR=interquartile range; CI=confidence interval

**Table E8. Performance of three convolutional neural networks in the Test dataset.**

| Overall pixelwise similarity in 11435 axial CT slices | | Bronchiectasis | Peribronchial Thickening | Bronchial mucus plug | Bronchiolar mucus plug | Collapse /consolidation | Total Abnormal Lung |
|---|---|---|---|---|---|---|---|
| **InceptionResNetv2** | **DICE** | 0.85 | 0.68 | 0.79 | 0.46 | 0.74 | 0.70 |
| | **Precision** | 0.89 | 0.71 | 0.82 | 0.61 | 0.83 | 0.77 |
| | **Recall** | 0.81 | 0.66 | 0.76 | 0.37 | 0.67 | 0.65 |
| | **Balanced Accuracy** | 0.90 | 0.83 | 0.88 | 0.68 | 0.85 | 0.82 |
| **ResNet50** | **DICE** | 0.83 | 0.67 | 0.77 | 0.48 | 0.70 | 0.69 |
| | **Precision** | 0.89 | 0.76 | 0.80 | 0.65 | 0.74 | 0.76 |
| | **Recall** | 0.79 | 0.60 | 0.74 | 0.38 | 0.65 | 0.63 |
| | **Balanced Accuracy** | 0.89 | 0.80 | 0.87 | 0.69 | 0.85 | 0.82 |
| **U-net** | **DICE** | 0.82 | 0.65 | 0.75 | 0.45 | 0.72 | 0.68 |
| | **Precision** | 0.88 | 0.77 | 0.82 | 0.74 | 0.80 | 0.79 |
| | **Recall** | 0.77 | 0.56 | 0.69 | 0.32 | 0.66 | 0.60 |
| | **Balanced Accuracy** | 0.89 | 0.78 | 0.84 | 0.66 | 0.85 | 0.80 |
| **Majority Vote** | **DICE** | 0.84 | 0.69 | 0.79 | 0.49 | 0.75 | 0.71 |
| | **Precision** | 0.90 | 0.78 | 0.87 | 0.78 | 0.86 | 0.84 |
| | **Recall** | 0.79 | 0.61 | 0.73 | 0.37 | 0.66 | 0.63 |
| | **Balanced Accuracy** | 0.90 | 0.81 | 0.87 | 0.68 | 0.85 | 0.82 |

Note: Owing to the large number of pixels over 11435 CT slices , the confidence interval of measurements was considered as negligible.
The Total Abnormal Lung values correspond to the average of the five structural alterations results.

**Table E9. Longitudinal evaluation of CF patients at initial evaluation and at follow-up, with or without lumacaftor/ivacaftor treatment.**

| Clinical Validation cohort | | CF patients with lumacaftor/ivacaftor (n=10) | | CF patients without lumacaftor/ivacaftor (n=60) | |
|---|---|---|---|---|---|
| | | Median difference | 95%CI of median difference | Median difference | 95% CI of median difference |
| **Normalized AI volumes** | Bronchiectasis | -0.2 | [-7; 4.5] | 3.1 | [1; 5.6] |
| | Peribronchial thickening | -6.4 | [-22; -2.2] | 3.3 | [0.1; 9.9] |
| | Bronchial mucus plug | -2.5 | [-19; -0.2] | -0.3 | [-2.4; 0.8] |
| | Bronchiolar mucus plug | -4.1 | [-44; -0.3] | -0.01 | [-0.7; 1.2] |
| | Collapse/Consolidation | -1.4 | [-72; 0.01] | 0.1 | [-1; 0.8] |
| | Total Abnormal Volume | -51 | [-146; -4.2] | 3.6 | [-6.6; 8.7] |
| **PFT** | FEV1% | 5.5 | [-1; 19] | -1.5 | [-4; 0] |
| **Visual CT score** | mBrody score | -2.5 | [-30; 0] | 5 | [0; 5] |

Note: The Total Abnormal Volume corresponds to the sum of the five structural alterations volumes per CT scan.

Legends: AI=artificial intelligence; PFT=pulmonary function test; FEV1%=forced expiratory volume in 1 second percentage predicted; mBrody score=modified Brody score

**Table E10. Paired comparisons of raw AI-driven label volumes in CF patients with and without lumacaftor/ivacaftor treatment**

| Clinical Validation cohort | | | CF patients with lumacaftor/ivacaftor (n=10) | | | CF patients without lumacaftor/ivacaftor (n=60) | | |
|---|---|---|---|---|---|---|---|---|
| | | | **M0** | **M12** | **p-value** | **M0** | **M24** | **p-value** |
| **Raw AI volumes (ml)** | **Bronchiectasis** | Median | 6.8 | 5.8 | 0.88 | 5.8 | 8.6 | **0.005** |
| | | Range | (0-75) | (0-82) | | (0-144) | (0.1-146) | |
| | **Peribronchial thickening** | Median | 6.3 | 3.9 | **0.005** | 6.8 | 11.5 | **0.003** |
| | | Range | (1-18) | (0-11) | | (0-84) | (0.2-99) | |
| | **Bronchial mucus plug** | Median | 2.3 | 2.0 | **0.005** | 3.0 | 2.7 | 0.96 |
| | | Range | (0.08-20) | (0.01-13) | | (0-110) | (0.1-58) | |
| | **Bronchiolar mucus plug** | Median | 8.3 | 3.2 | **0.006** | 1.7 | 2.8 | 0.52 |
| | | Range | (0.1-36) | (0.01-25) | | (0-32) | (0-49) | |
| | **Collapse/Consolidation** | Median | 3.0 | 1.5 | **0.01** | 1.5 | 1.3 | 0.68 |
| | | Range | (0-80) | (0-17) | | (0-55) | (0.9-46) | |
| | **Total Abnormal Volume** | Median | 56.0 | 20.4 | **0.005** | 21.4 | 29.5 | 0.17 |
| | | Range | (1.0-294) | (0.8-100) | | (0-276) | (0.8-249) | |
| | **Lung Parenchyma** | Median | 3250 | 3457 | **0.04** | 3549 | 3929 | **0.001** |
| | | Range | (2326-6494) | (2328-6494) | | (1009-7405) | (1389-7455) | |

Note: Data are medians, with (minimum-maximum) range of values. The Total Abnormal volume corresponds to the sum of the five structural alterations volumes per CT scan.

Legends: M0=initial evaluation; M12=second evaluation at one year; M24=second evaluation at two years.

**Table E11. Characteristics of CT scans in the follow-up of 140 CF.**

| **Clinical Validation group** | | **CF with lumacaftor/ivacaftor** (n=10) | | **CF without lumacaftor/ivacaftor** (n=60) | |
|---|---|---|---|---|---|
| | | **M0** | **M12** | **M0** | **M24** |
| **Machine brand** | | Somatom Definition 64® (n=10) | Somatom Definition 64® (n=10) | Somatom Definition 64® (n=33)<br>Somatom Sensation 16® (n=27) | Somatom Definition 64® (n=34)<br>Somatom Sensation 16® (n=26) |
| **Kernel** | | I30f (n=10) | I30f (n=10) | I30f (n=22)<br>Br40 (n=11)<br>B40s (n=27) | I30f (n=23)<br>Br40 (n=11)<br>B40s (n=26) |
| **Reconstruction** | | SAFIRE (n=10) | SAFIRE (n=10) | FBP (n=38)<br>SAFIRE (n=22) | FBP (n=37)<br>SAFIRE (n=23) |
| **DLP** | mGy.cm (minimum-maximum) | (12-17) | (12-18) | (12-53) | (13-64) |
| **kV** | | 110 | 110 | 110 | 110 |
| **mAs** | **Dose modulation* (yes/no)** | 10/0 | 10/0 | 38/32 | 39/31 |
| | **If yes, reference values** (minimum-maximum) | (5-10) | (5-10) | (5-10) | (5-10) |
| | **If no, fixed value** (minimum-maximum) | NA | NA | (35-54) | (35-54) |
| **Slice thickness** | **(mm)** | 1 | 1 | 1 | 1 |

*Note: the dose modulation system was CareDose4D®.

Legends: FBP=filtered-back projection; SAFIRE=sinogram affirmed iterative reconstruction; kV=kilovoltage, mAs=milliampere second; DLP=dose length product

**Table E12. Reproducibility and repeatability of AI and manual interobserver similarity in the Clinical Validation cohort**

| **2D pixelwise similarity**<br>n=42280 CT slices in 140 CTs | **$AI_1$ vs. $AI_2$**<br>**Dice** | **$AI_1$ vs. $AI_1$**<br>**Dice** |
|---|---|---|
| **Bronchiectasis** | >0.99 | >0.99 |
| **Peribronchial thickening** | >0.99 | >0.99 |
| **Bronchial mucus plug** | >0.99 | >0.99 |
| **Bronchiolar mucus plug** | >0.99 | >0.99 |
| **Collapse/Consolidation** | >0.99 | >0.99 |
| **Total Abnormal Lung** | >0.99 | >0.99 |
| **Lung Parenchyma** | >0.99 | >0.99 |
| | **$Manual_1$ vs. $Manual_2$** | **$Manual_1$ vs. $Manual_1$** |
| n=2850 CT slices in 8 CTs | **Dice** | **Dice** |
| **Bronchiectasis** | 0.86 | 0.84 |
| **Peribronchial thickening** | 0.70 | 0.73 |
| **Bronchial mucus plug** | 0.72 | 0.73 |
| **Bronchiolar mucus plug** | 0.62 | 0.65 |
| **Collapse/Consolidation** | 0.73 | 0.77 |
| **Total Abnormal Lung** | 0.72 | 0.74 |
| **Lung Parenchyma** | 0.99 | 0.99 |

Note: the Total Abnormal Lung corresponds to the average of the five structural alterations similarity results.

Legends: $AI_1$=artificial intelligence-driven measurement performed on an advanced computer device; $AI_2$=artificial intelligence-driven measurement performed on a standard computer device; Manualx=segmentation performed by Observer x

**SUPPLEMENTAL FIGURES**

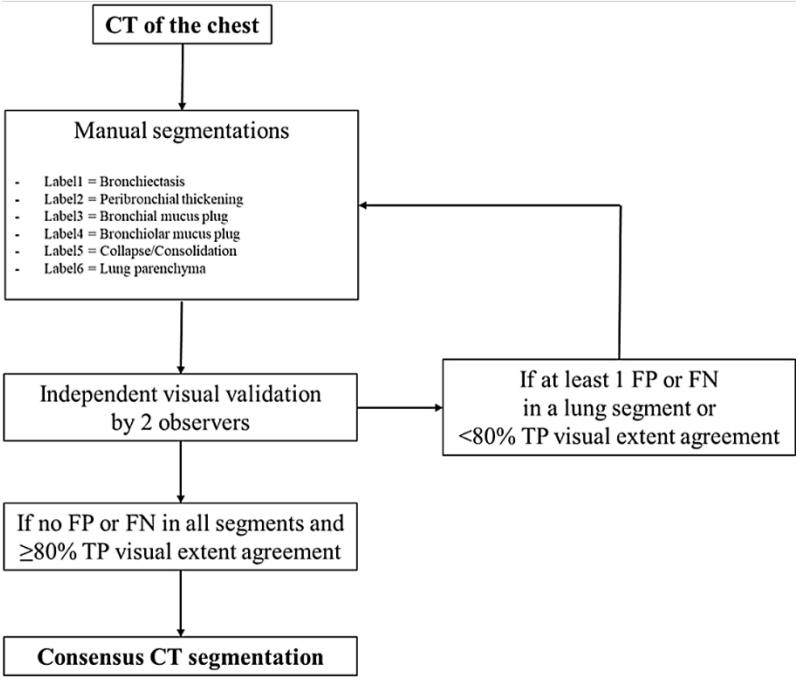



**Figure E1**. Flow chart of the method to produce consensus CT semantic segmentation for Training. The segmentations were visually checked at the segmental level. TP=true positive; FP=false positive; FN=false negative.

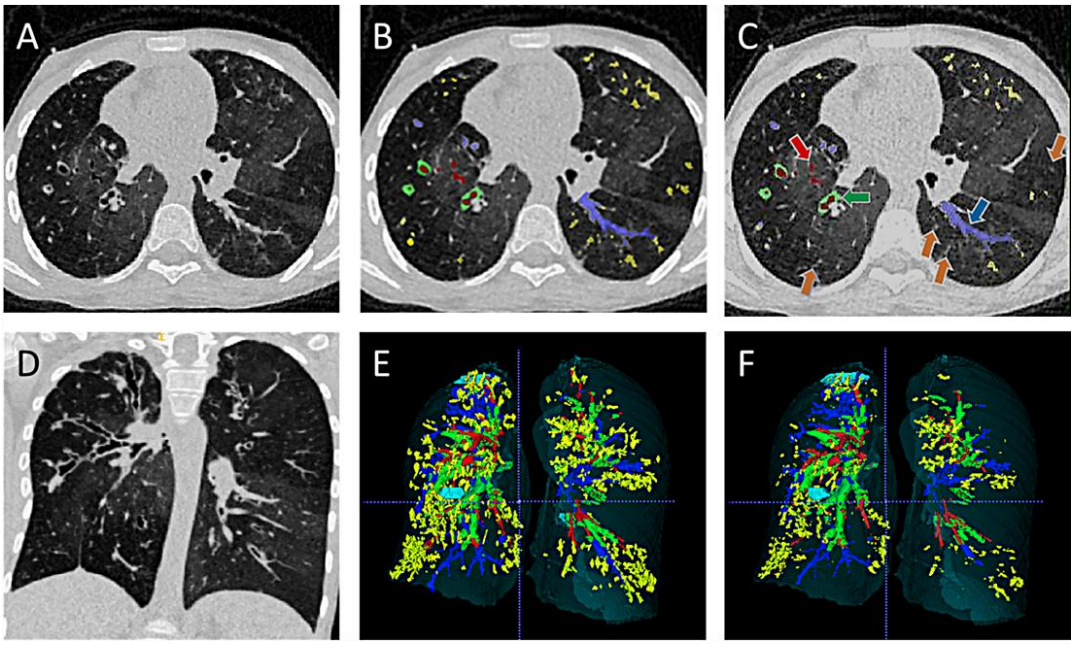
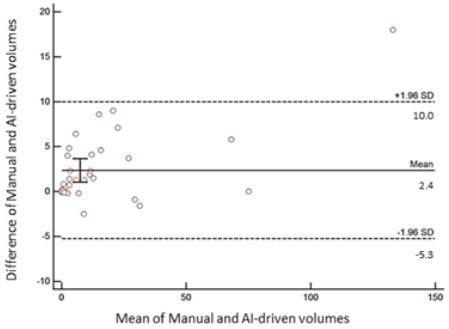
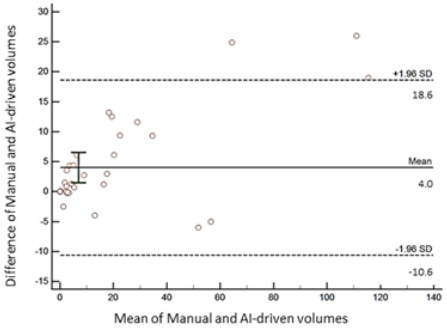



**Figure E2**. Axial (A) and coronal reformations (D) of a lung CT scan acquired in a 15-year-old female with cystic fibrosis. Manual (B, E) and AI-driven (C, F) semantic multilabel segmentation are shown and displayed in corresponding axial CT slice (B, C) and volume rendering in coronal view (E, F). In panels B, C, E, F, red arrow and red labels indicate mucus-free bronchial lumen dilatations, green arrow, and green labels show peribronchial thickening, blue arrow, and blue labels indicate central bronchoceles. Bronchiolar mucus plugs were labeled in yellow color, and orange arrows show some AI's false-negative results of this feature (C). In panels E and F, cyan labels indicate consolidations. Note the heterogeneity of structural alterations and their regional distribution within the same lung CT volume. In this patient, the mean Dice coefficient of similarity between manual and AI-driven segmentation was equal to 0.70.

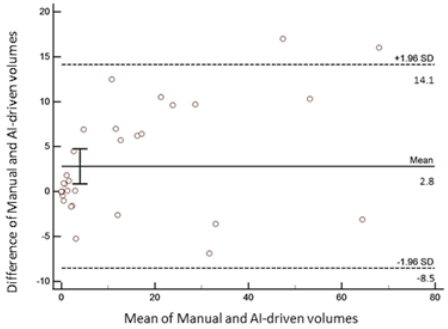
## Bronchiectasis



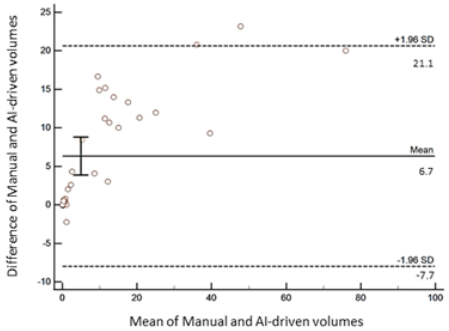

## Peribronchial thickening



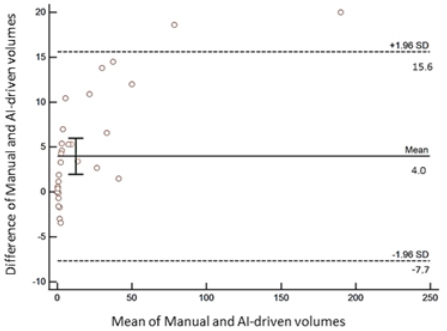

## Bronchial mucus




## Bronchiolar mucus




## Collapse/Consolidation




## Total abnormal volume

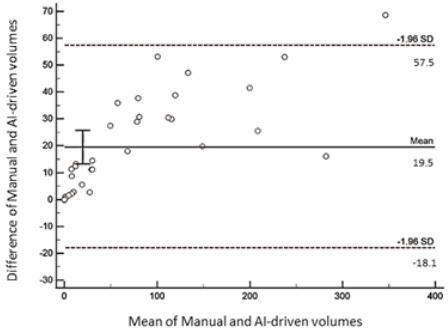


**Figure E3.** Bland-Altman analyses of manual versus AI-driven label volumes in the Test cohort (n=36), expressed in milliliters. The plain lines represent the mean difference and the bars their 95% confidence interval; the dashed lines represent the limits of agreement.

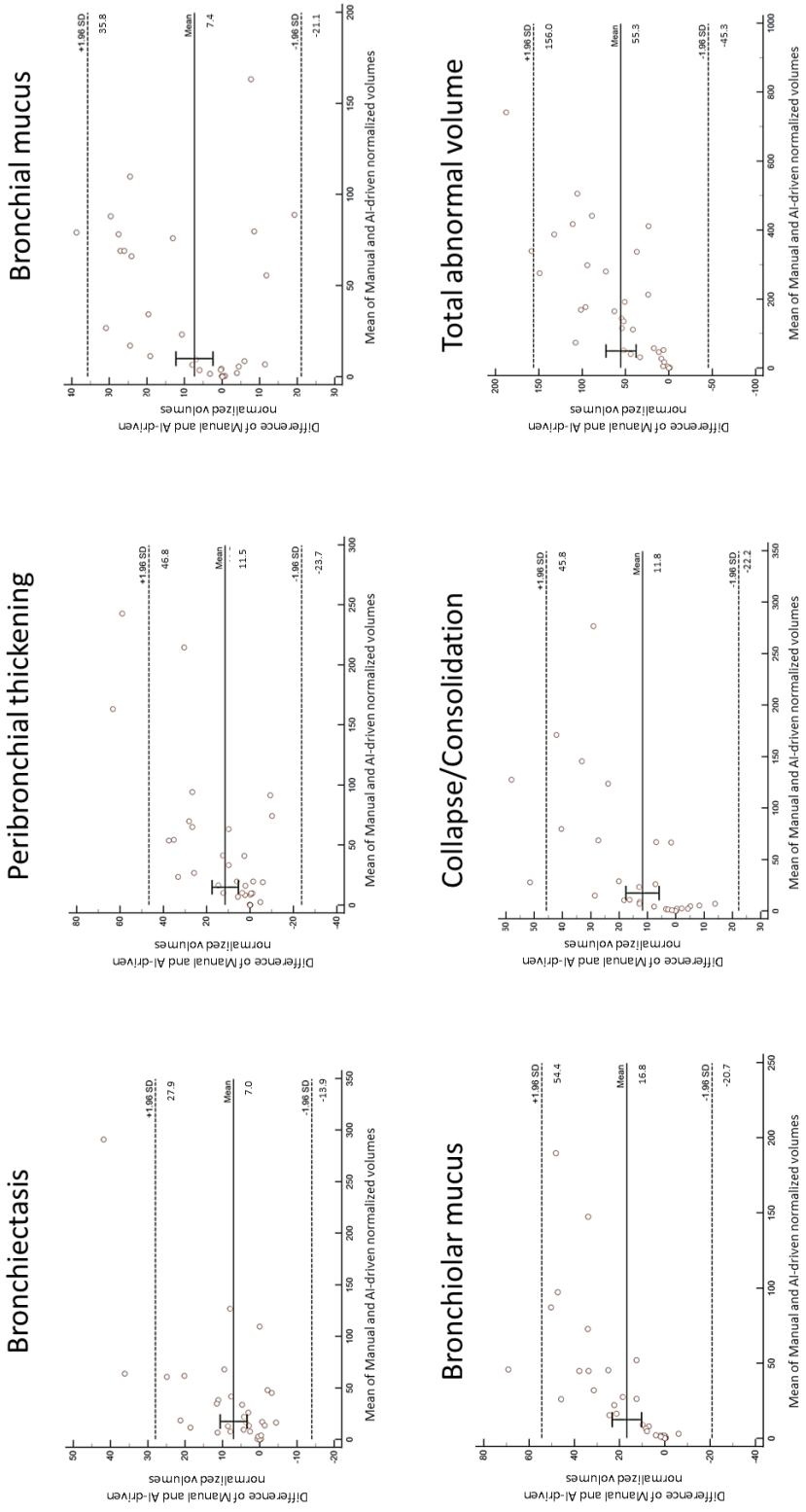


**Figure E4.** Bland-Altman analyses of manual versus AI-driven normalized volumes in the Test cohort (n=36). The plain lines represent the mean difference and the bars their 95% confidence interval; the dashed lines represent the limits of agreement.

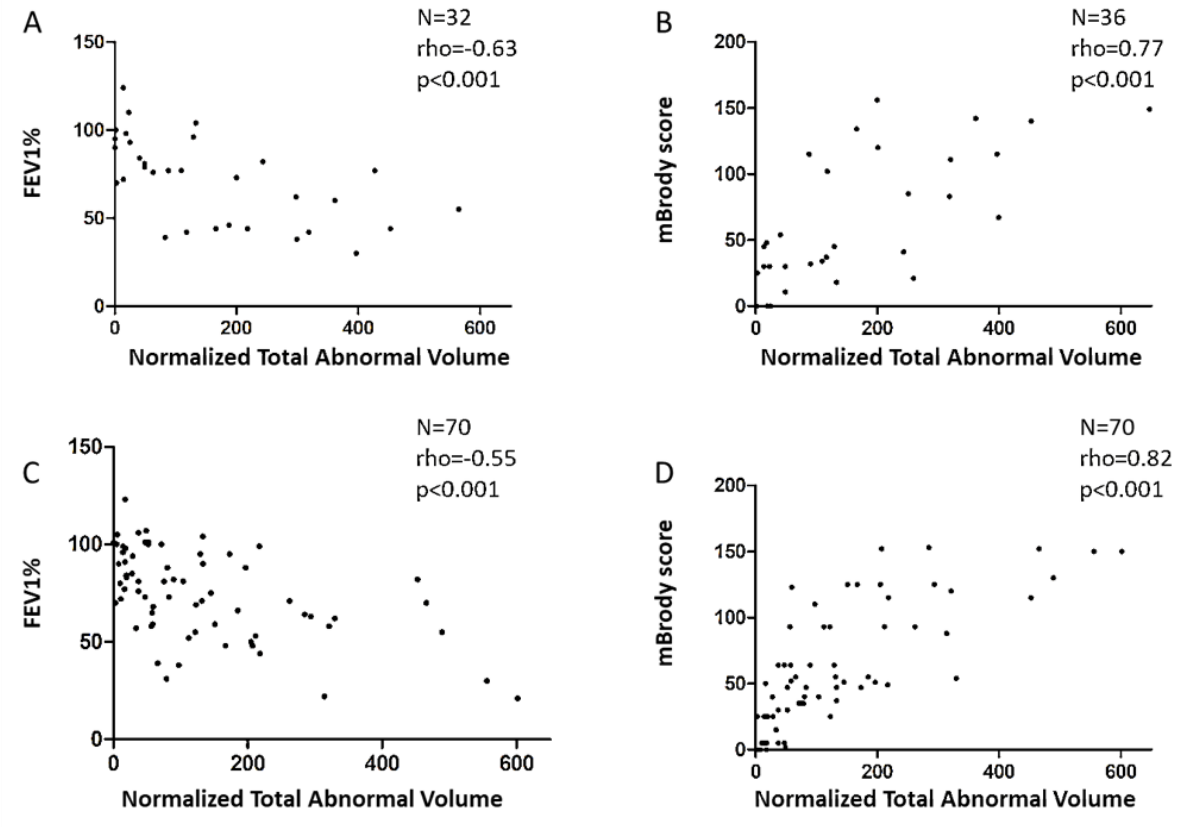



**Figure E5.** Spearman's correlations between AI-driven measurement of normalized total abnormal volume and CF disease severity, as assessed by the forced expiratory volume in 1-second percentage predicted (FEV1%; A and C) and the modified Brody score (mBrody; B and D). Results are shown for both the Test (A, B) and the Clinical Validation (C, D) cohorts.

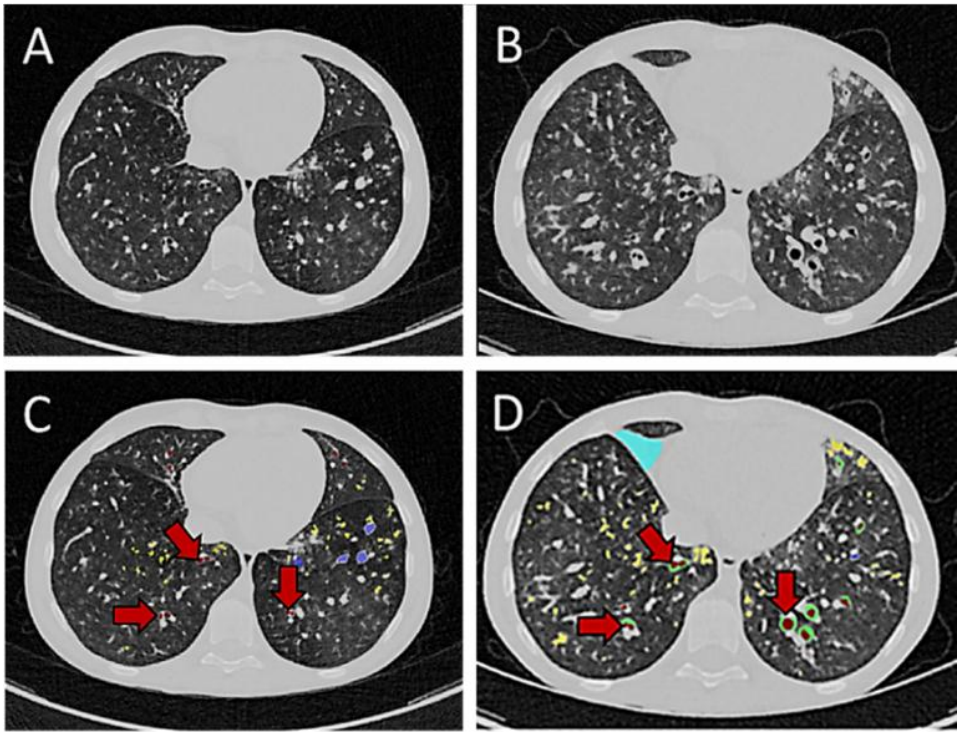



**Figure E6.** Comparison of AI-driven semantic labeling in the Clinical Validation cohort, at initial evaluation (A, C) and after two years (B, D) of standard management, in a 15-year-old male with cystic fibrosis. Axial CT slices (A, B) are shown, with AI-driven semantic labeling displayed in the corresponding axial slice (C, D). Mucus-free bronchial lumen dilatations are labeled in red color, peribronchial thickening in green color, bronchial mucus plugs in blue color, bronchiolar mucus plugs in yellow color, and atelectasis in cyan color. In panels (C, D), red arrows show an increase in bronchial dilatations and peribronchial thickening over time.