HAL
open science

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

Gael Dournes, Chase S Hall, Matthew M Willmering, Alan S Brody, Julie Macey, Stephanie Bui, Baudouin Denis de Senneville, Patrick Berger, François Laurent, Ilyes Benlala, et al.

To cite this version:

Gael Dournes, Chase S Hall, Matthew M Willmering, Alan S Brody, Julie Macey, et al.. Artificial intelligence in CT for quantifying lung changes in the era of CFTR modulators. European Respiratory Journal, 2022, 59, pp.2100844. 10.1183/13993003.00844-2021 . hal-03453536

HAL Id: hal-03453536

https://hal.science/hal-03453536v1

Submitted on 8 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Title. 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

Affiliations

[1] Univ. Bordeaux, INSERM, Centre de Recherche Cardio-Thoracique de Bordeaux, U1045, CIC 1401, F-33000 Bordeaux, France

[2] CHU Bordeaux, Service d'Imagerie Thoracique et Cardiovasculaire, Service des Maladies Respiratoires, Service d'Exploration Fonctionnelle Respiratoire, CIC 1401, F-33600 Pessac, France

[3] Division of Pulmonary, Critical Care and Sleep Medicine, Department of Internal Medicine, University of Kansas School of Medicine, Kansas City, Kansas, United States of America

[4] Center for Pulmonary Imaging Research, Division of Pulmonary Medicine and Department of Radiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, United States of America

[5] Bordeaux University Hospital, Hôpital Pellegrin-Enfants, paediatric Cystic Fibrosis Reference Center (CRCM), Centre d'Investigation Clinique (CIC 1401), F-33000 Bordeaux, France

[6] Univ. Bordeaux, Mathematical Institute of Bordeaux (IMB), UMR CNRS 5251, FR-33405 Talence, France

[7] Department of Pediatrics, College of Medicine, University of Cincinnati, Cincinnati, Ohio, United States of America

Contact author : Dr Gaël Dournes, Centre de Recherche Cardio-thoracique de Bordeaux, INSERM U1045, Université de Bordeaux, 146 rue Léo Saignat, 33076 Bordeaux Cedex, France. Tel: +33 5 57 57 46 02 Fax: +33 5 57 57 16 95. E-mail: gael.dournes@chu-bordeaux.fr

Take home message

Artificial intelligence allows a fully automated volumetric scoring of lung structural abnormalities in cystic fibrosis using computed tomography. It could be used as a robust quantitative outcome to assess disease changes in the era of CFTR modulators.

Abstract

Rationale. Chest computed tomography (CT) remains the imaging standard for demonstrating cystic fibrosis airway structural disease *in vivo*. However, visual scorings as an outcome measure are time-consuming, require training, and lack high reproducibility. **Objective.** To validate a fully automated artificial intelligence-driven scoring of cystic fibrosis lung disease severity.

Methods. Data were retrospectively collected in three cystic fibrosis reference centers, between 2008 and 2020, in 184 patients 4 to 54-years-old. An algorithm using three two-dimensional convolutional neural networks was trained with 78 patients' CTs (23530 CT slices) for the semantic labeling of bronchiectasis, peribronchial thickening, bronchial mucus, bronchiolar mucus, and collapse/consolidation. 36 patients' CTs (11435 CT slices) were used for testing versus ground-truth labels. The method's clinical validity was assessed in an independent group of 70 patients with or without lumacaftor/ivacaftor treatment (n=10 and 60, respectively) with repeat examinations. Similarity and reproducibility were assessed using Dice coefficient, correlations using Spearman test, and paired comparisons using Wilcoxon rank test.

Measurement and main results. The overall pixelwise similarity of artificial intelligence-driven versus ground-truth labels was good (Dice coefficient=0.71 [95% confidence interval: 0.70-0.71]). All artificial intelligence-driven volumetric quantifications had moderate to very good correlations to a visual imaging scoring ($p<0.001$) and fair to good correlations to FEV1% at pulmonary function test ($p<0.001$). Significant decreases in peribronchial thickening ($p=0.005$), bronchial mucus ($p=0.005$), bronchiolar mucus ($p=0.007$) volumes were measured in patients with lumacaftor/ivacaftor. Conversely, bronchiectasis ($p=0.002$) and peribronchial thickening ($p=0.008$) volumes increased in patients without lumacaftor/ivacaftor. The reproducibility was almost perfect (Dice>0.99).

Conclusion. Artificial intelligence allows a fully automated volumetric quantification of cystic fibrosis-related modifications over an entire lung. The novel scoring system could provide a robust disease outcome in the era of effective CFTR modulator therapy.

Revised Abstract word count: 281

Revised Main manuscript word count: 3560

INTRODUCTION

Cystic fibrosis (CF) is one of the most common life-shortening genetic disorders in Caucasians, and lung disease remains the most common cause of mortality and death[1]. Recently, the need for developing robust biomarkers has been emphasized, as novel treatments have emerged[2]. Relying on standard outcomes such as the forced expiratory volume in 1 second (FEV1) at pulmonary function tests (PFT) may become increasingly difficult to demonstrate a new treatment efficacy[3, 4]. In this setting, computed tomography (CT) is the primary imaging tool for assessing lung morphology in the clinical care of patients with CF. Owing to its excellent spatial resolution, CT allows the accurate identification of structural abnormalities, whose evaluation as an outcome measure for CF research has long been advocated[5–7]. Multiple scoring methods have been used to quantify the lung disease severity[8–10]. All of them share important drawbacks, including the need for trained expert readers and a time-consuming scoring process[11]. One of the most widely used scoring systems is the Brody score, but this has been found to have limited reproducibility[12] and poor sensitivity[13] to mild disease variation. Thus, longitudinal evaluations may be impaired, when the same consensus of experts is not continued over years[12]. However, there is a worldwide shortage of radiologists[14]. Recently, the PRAGMA-CF score was developed specifically for use in infants and young children[15]. This system is more reproducible and more sensitive to lung disease progression than the Brody score[13, 15]. However, the use of a grid system limits the PRAGMA-CF method, and only a single finding is scored when more than one occurs in the same grid square, prioritizing bronchiectasis and air trapping. Because of these discarded data, the system cannot assess the extent of other abnormalities, including mucous plugging. This system is also limited by a laborious scoring process requiring an expert trained in the system, taking approximately 20 minutes to score a subset of 10 CT slices per CT scan. Automated analyses may have the potential to address these limitations.

DeBoer and colleagues published a computer-based system that counted visible airways and demonstrated a good correlation with lung function and neutrophil elastase activity[16]. Despite this encouraging result, further developments awaited the advent of artificial intelligence (AI) with deep learning algorithms[17-19]. Deep learning currently represents the most advanced machine learning technique, allowing the creation of models that perform as well or even better than humans[20] while reproducing the human visual perception system. A CT few studies have reported testing AI to detect abnormal airways in CF[21–23]. As compared to chest radiograph[24], chest CT of CF involves additional challenges, due to the higher spatial resolution and the large heterogeneity in distribution and size of structural abnormalities over hundred of CT slices. Thus, previous CT models had difficulties in discriminating structural abnormalities from normal lung parenchyma. Also, no quantification of the disease extent was reported, and there was no attempt to correlate the findings to the clinical disease status.

We hypothesized that an AI-driven semantic quantification of lung structural alterations is feasible in CF and could build an automated scoring system. Clinical validation expects that a biomarker reflects the clinical severity, correlates to a known outcome, and may improve with an effective therapy[7, 25]. Thus, the objective was to develop an algorithm enabling recognition of five structural alterations hallmarks on CT slices. Then, we aimed to assess the clinical validity of the quantitative scoring method by correlating to the patient's disease severity, as assessed by the CT Brody score. Other secondary objectives to support the clinical validity were to correlate to PFT, assess variations in patients with and without lumacaftor/ivacaftor, and evaluate the reproducibility.

Material and Methods

Study Design

CT scans and patient data were collected from patients with a diagnosis of CF confirmed by genetic and/or sweat chloride test[26] at three CF reference centers from two Institutions: 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). Clinical evaluation, PFT[27], and non-contrast-enhanced CT had to be performed as part of annual clinical care the same day[28]. The Institutional Review Boards approved the study after waiver of written informed consent (registration number NCT04760548). To assess the main outcome, a minimum of 36 patients was calculated to assess correlations of more than 0.45 with the CT Brody score, with a power of 0.8 and a risk alpha of 0.05[8].

Anonymized data from consecutive CF patients were collected from 2017 to 2020 in site1 (n=43) and site2 (n=47), and from 2008 to 2010 in Site3 (n=24). The data collection periods enabled a wide range of CT machines, from relatively old to the newest ones, without exclusion. Stratified randomization based on the CT scanner models was used to split the original CT dataset into two non-overlapping groups[29], *i.e.,* Training (n=23530 CT slices from 78 patients; Supplemental Table E1) and Test (n=11435 CT slices from 36 patients). There were seven CT models from two major manufacturers (Supplemental Table E2), and there was no overlap in CF patients between the two cohorts.

To create a final independent clinical validation cohort (n=23940 CT slices from 70 patients), including a longitudinal analysis, data from consecutive CF patients were retrieved from 2014 to 2016 at Site1 (n=32) and Site2 (n=38). Patients were not included if imaging was performed <4 weeks from an acute exacerbation[30] or if the participant had previously been included in the Training or Test cohorts (n=0). Ten of the patients had initiated lumacaftor/ivacaftor treatment, which was introduced for clinical use during 2016 in the

authors' country, and repeated examinations at one year. The remaining sixty patients repeated examinations at two years (Figure 1).

AI Training Framework

Briefly, five labels were manually completed in the CT axial plane by consensus of three thoracic radiologists on inspiratory CTs with standard kernels[31] and designated as ground-truth (GT) (Supplemental Method E1; Figure E1). They represented five lung structural alterations hallmarks[32]: bronchiectasis, peribronchial thickening, bronchial mucus; bronchiolar mucus; and collapse/consolidation. A sixth label identified the surrounding lung parenchyma. Three 2D-convolutional neural networks (CNNs)[33–35] were trained utilizing the Training cohort after data augmentation[36, 37]. A majority vote[38] was performed to complete a final AI-driven multi-label segmentation (Supplemental Method E1; Table E1-E2-E3).

Evaluation of AI semantic similarity and agreement

In the Test cohort, 36 CT scans (11435 axial CT slices) were shuffled randomly before being segmented by the 2D-CNNs to assess the 2D-similarity between AI-driven and GT test labels. Then, the shuffled CT slices were re-assigned to their initial study to calculate each CT scan's label volume, and a 3D-agreement was assessed.

Evaluation of AI clinical validity

Patients' management was performed according to a standard of care[39]. AI-driven volumes were normalized to determine correlations to PFT and a modified CT Brody score[8]

(Supplemental Method E2-E3; Table E4). Longitudinal evaluations were performed in the Clinical Validation cohort by using paired-comparison analyses.

Reproducibility and repeatability

In the Clinical Validation cohort, AI-driven measurements were performed twice, by using an advanced computer system and then a standard computer system (Supplemental Method E4). A random subset of 8 patients' CTs (Supplemental Table E5) was also manually segmented twice by an observer, 6 months apart, and then independently by a second observer, blinded to any other data and the other observer labels.

Statistical Analysis

Statistical analyses were performed using the MedCalc® software (Ostend, Belgium) and graphs by using the Prism® (San Diego, USA) softwares. As a first attempt study, no assumption on the distribution of AI quantitative parameters was possible *a priori*. Thus, non-parametric statistical tests were used. Data were expressed as medians with minimum-to-maximum range. Similarity was assessed by calculating the overall pixelwise balanced accuracy, Sorensen-Dice similarity coefficient (Dice), precision, and recall[40] (Supplemental Method E2). Agreement were assessed by Kendall's tau correlation and Bland-Altman analysis[41], respectively. The bias was further assessed by using a Passing Bablock regression. Spearman's rho coefficient assessed correlations, and comparison of paired-medians was made by Wilcoxon-rank test. Correlation coefficients were classified as null (=0) to almost perfect (≥0.95)[42]. Comparison of correlation coefficients was performed according to Hinkle and colleagues[43]. A Bonferoni correction was not deemed necessary,

since all tests were used to address planned hypotheses[44] and a p-value≤0.05 was considered significant.

Results

Study populations

Clinical, functional, and CT characteristics of CF study cohorts are summarized in Table 1. Taken together, the median age was 13.5, ranging from 4 to 54-year-old; the ratio of male/female was balanced; 49% of CF patients were homozygous for the DeltaF508 mutation, and 28% had a chronic infection by *Pseudomonas aeruginosa.* The CT Brody scores ranged from 0 to 156 and took 15 to 20 minutes per examination. 19 out of 36 CF patients, and 44 out of 70 CF patients had a measurement of FEV1%>70 at baseline in the Test cohort and Clinical Validation cohort, respectively. Table E6 provides additional information on the background therapeutic management in the Clinical Validation cohort. Notably, none of patients with lumacaftor/ivacaftor had either oral or intravenous anti-infectious chronic therapy.

Similarity and Concordance of Test Cohort.

In the Test cohort, the volume of structural abnormalities per CT slice is given in Table E7. The highest pixelwise similarity between AI-driven and ground-truth labels was found for bronchiectasis (Dice=0.86) and was lowest for peripheral mucus plugs (Dice=0.49). The average labeling results were a Dice of 0.71, a balanced accuracy of 0.82, a recall of 0.63, and a precision of 0.84 (Table 2; Figure E2). The majority vote reconstruction had higher precision and Dice than its three CNNs components (Supplemental Table E8).

The 3D agreement between AI-driven volume calculations and GT labels was good to almost perfect, demonstrating τ ranging from 0.79 to 0.93 (Figure 2). Bland-Altman analysis showed that the AI-driven segmentation tended to systematically under-label abnormalities compared to GT; however, the mean difference was small and less than 7ml for each structural abnormality label (Supplemental Figure E3). Regarding the Total abnormal volume, the systematic underestimation had a linear pattern, as assessed by the Passing-Bablock regression (slope=1.29; Intercept=0.63). Moreover, recognition of the surrounding lung parenchyma showed almost perfect pixel-wise similarity (Dice=0.99) and volume agreement (τ =0.99).

Correlations with CF Severity

In the Test cohort, PFT was not performed in four children, who were all less than 6-years-old. There was a significant correlation between all AI-driven normalized volume labels and FEV1% (n=32; $p \leq 0.04$) and to the modified Brody score (n=36; $p \leq 0.001$) (Table 3). The correlation coefficients were similar to those of the corresponding GT labels ($p \geq 0.47$). Similarly, in an independent clinical validation cohort, all AI-driven normalized volume labels significantly correlated to PFT and visual CT scoring at both initial (n=70; $p<0.001$) and follow-up evaluations (n=70; $p<0.001$) (Table 3; Supplemental Figure E4). Three examples of lung AI-driven semantic labeling from the Clinical Validation cohort are shown (Figure 3).

Paired comparisons in patients with and without lumacaftor/ivacaftor

In the Clinical Validation cohort, patients who underwent treatment with lumafactor/ivacaftor treatment (n=10) had a significant reduction in normalized volumes at one year, notably

peribronchial thickening (median difference:-6.4 [95% confidence interval: -22; -2.2]; p=0.005), bronchial mucus plugs (-2.5 [-19; -0.2]; p=0.005), bronchiolar mucus plugs (-4.1 [-44; -0.3]; p=0.007), and the Total Abnormal Volume (-51 [-146; -4.2]; p=0.005), but not bronchiectasis ((-0.2 [-7; 4.5]; p=0.59) (Table 4; Table E9; Figure 4). Four out of these ten CF patients had an FEV1% of more than 80%. The ten patients with lumacaftor/ivacaftor had a standardized CT acquisition at two-time points, with no change in machine manufacturer or CT protocol (Supplemental Tables E9-E10).

Conversely, patients without CFTR modulator treatment (n=60) increased both bronchiectasis (3.1 [1; 56]; p=0.002) and peribronchial thickening volumes (3.3 [0.1; 9.9]; p=0.008) at two years of routine follow-up (Table 4; Supplemental Figure E5).

Reproducibility and repeatability

As a fully automated measurement, AI-driven quantitative measurements had an almost perfect reproducibility and repeatability when completed twice in 140 CTs (42280 axial CT slices; Dice>0.99; Supplemental Table E12). The median time to reach AI-driven labeling was 2 and 10.5 minutes per CT scan by using the advanced machine and standard computer machines, respectively. The similarity between two independent manual segmentations was also assessed in a subset of 8 CTs (2580 axial CT slices). These additional evaluations showed an average Dice coefficient of 0.74 and 0.72, respectively, although with a median time of 540 minutes per observer per CT scan.

Discussion

The study demonstrates that AI-driven quantitative measurement of lung structural abnormalities on CT scanning in CF is feasible and can provide clinically important

information in a broad range of patients using a wide range of CT scanners and CT techniques. The system showed good similarity and very good concordance with “ground truth” identification of expert observers' abnormalities, but dramatically quicker, with high reproducibility. Volumetric measurements showed a strong correlation to PFT and a well-validated visual CT score at several timepoints. The automated quantifications were found to sensitively detect longitudinal changes, either a reduction in CF patients with lumacaftor/ivacaftor treatment or an increase during the natural course of the disease. As a fully automated outcome measurement, the reproducibility was almost perfect.

This AI-based CT analysis differs from previously reported AI studies in CF [21–23] in several important ways. First, labeling was performed based on the lung volume occupied by all abnormalities separately, without pre-determined limits. Previous AI studies used a patch-based approach, consisting of dividing a lung CT slice into a grid of several centimeter squares. Several difficulties related to this AI approach have been outlined. When multiple abnormalities co-exist within a single patch, a hierarchical classification is needed to assign a single label to the entire patch. If an abnormality represents less than 50% of a patch, it is labeled a normal lung and *vice versa*. This can lead to confusion between AI labels and, in the previous literature, the Dice coefficients were reported to be 0.33, at best[21]. Moreover, we used 2D-CNNs to train the model in a slice-by-slice fashion. This study used 78 patients’ CTs (23530 unique axial CT slices augmented to 288830 CT slices), in patients with a median age of 21 (from 4 to 51-years-old). Previous reports were performed using initial datasets of less than 2000 CT slices[21–23]. An important factor in the quality of AI implementation is the dataset's size to develop the system[45]. Also, three 2D-CNNs were trained, and the group of CNNs used a “majority rule” design. This allowed prioritizing detection precision to minimize false-positive identifications.

In addition, the evaluation was done in an external Test cohort of 36 patients (11435 CT slices), where all CT slices composing each CT scan were present. Previous studies were performed on a subset of CT slices[21] or in the same dataset used to tune the AI model[22]. Thus, a singularity of this CF study is to provide an extensive overview of the method's performance to segment an entire lung CT scan. This allows the model to be used without any manual interventions, such as cropping the lung areas or pre-selecting CT slices. The highest Dice coefficient of similarity was found for bronchiectasis and the lowest for bronchiolar mucus plugs. However, when structures are very small, it is known that Dice is not an appropriate method of evaluation[40] since any pixel of disagreement may dramatically reduce the Dice coefficient even when there is large-scale agreement. In addition, in 2D, both normal small vessels and bronchiolar mucus plugs may share similar aspects cross-sectionnally, in the form of multiple millimeter spots[32]. However, the precision was good, and the AI segmentation was able to clinically correlate to the disease severity and to detect a variation in patients with CFTR modulator treatment. Moreover, this metric discards the true-negative results. This is why we used τ correlations to evaluate the disease extent at the patient level, and all reconstructed label volumes showed a very good 3D agreement with ground-truth volumes over the full range of disease severity. Of note, the majority of the patients involved in this study had mild disease, according to PFT[46].

Clinical validation is necessary for any biomarker to be utilized for research or clinical practice. This study's retrospective and multisite design provided the opportunity for a “real world” evaluation. The study cohorts had patient characteristics similar to that of CF patients' larger populations, such as genetic, functional, and microbiological data[47]. First, the volumetric multilabel quantifications were correlated to the lung disease severity, assessed by PFT and a well-validated human-based CT score. Correlations were demonstrated in a Test

cohort of 36 patients and an independent clinical validation cohort of 70 patients, at two time points. At cross-sectional analysis, the correlates' strength appeared similar in all cohorts, despite some heterogeneity in CT models and acquisition techniques[31], supporting the model's generalizability. Second, in patients with CFTR modulators, the fully automated AI-driven measurements demonstrated a significant volume reduction in peribronchial thickening, bronchial mucus, bronchiolar mucus, and collapse/consolidation. Conversely, PFT and a visual CT score only approached significance in a small sample of 10 patients—4 of whom had normal FEV1% at baseline. As for previous CF studies using CT, we did not directly compare to a control group[48, 49]. However, the natural disease course would be stable or worsening[50], while a similar reduction in peribronchial thickening and mucus plugs in patients with CFTR modulators has been reported[48, 49]. In fact, such significant reduction has never before been reported out of a result of therapeutics.

In patients without CFTR modulators, the increase in bronchiectasis is also consistent with the known contribution of bronchiectasis to disease progression[9, 15]. This is also critical information, allowing clinicians to monitor disease progression and supporting treatment plans[51]. Besides, the lack of reversibility of bronchiectasis under novel CF treatments agrees with the literature[48, 49]. Therefore, our findings' clinical validity is consistently supported by international literature, both cross-sectionally and longitudinally. The results are promising and pave the way for future prospective trials, since AI may facilitate large-scale studies at reduced cost. In this setting, chest CT standardization is likely further beneficial to optimize the method's sensitivity[52].

The AI-based system offers multiple advantages over expert reader scoring systems. The time required to get AI-driven quantifications in each full-lung CT exam was only 2 minutes, compared to many hours for a human performing similar whole-lung analysis.

Simultaneously, the reproducibility and repeatability of the clinically relevant assessments was almost perfect. This level of reproducibility is unique as compared to reader scoring systems. The Digital Imaging and Communications in Medicine (DICOM) standard used by all imaging equipment manufacturers allows for easy CT data transfer in a research context. In CF centers, anyone with basic computer skills can be quickly trained to perform the analysis. This simplifies the use of CT scoring in both clinical use and for drug development. In addition to these advantages over expert reader scoring systems, these initial results suggest that peribronchial thickening, which has not been a useful parameter in expert reader systems, may be an important measure using an AI-based system. Peribronchial thickening is one of the most difficult findings to assess visually[12], and it is considered at the third rank of priority in hierarchical systems[21]. However, a reliable assessment of reversible structural alterations has become important in the contemporary context of novel CF treatments[48]. With the increasing interest in sensitively assessing rapidly reversible changes, this may dramatically improve CT scanning's ability to serve as a short-term outcome surrogate. Nevertheless, both visual and automated quantitative measurements are complementary practices, in the authors' opinion[53].

Several limitations of the study could be pointed out. First, as an initial description, the study was retrospective; a multicenter, prospective study would be a useful comparison. Second, expiratory CTs were not performed in two of the three sites, so the functional evaluation of air trapping was not possible as a sub-score. Indeed, AI was trained in a multilabeling fashion, to detect structural abnormalities only. In addition, expiratory CT requires additional radiation exposure and, despite advances in the reduction of radiation doses, this is not practiced in all CF centers[48]. Third, the labelings were not made in a pure 3D fashion. However, as for human assessment, information from adjacent CT slices could improve the performance.

Further evaluation using alternative strategies such as multiplane consensus labeling[54] or 3D algorithm[45] would be worth evaluating, albeit with a heavy computational burden. Also, we have not created a label to segment the lung vessels[55]. An evaluation of the spatial evaluation of the lesions, such as central and peripheral lung, could be also interesting to address in a next implementation. Fourth, two major manufacturers were present in the datasets, and additional evaluation with other manufacturers would be of added value. Fifth, the lower limit of age was four years. Conversely, the semi-automated PRAGMA-CF method has been well-validated in infants and younger children[15]. Sixth, the lowest CT dose in the study was 8 mGy.cm. Although an ultra low dose CT may correspond a CT dose lower than 20 mGy.cm[56],this AI has not yet been trained at lower radiation doses. Finally, a translation to recently developed radiation-free MRI acquisitions at high resolution could be envisioned[53, 57].

To conclude, we have demonstrated for the first time that an automated AI-driven quantitative scoring of structural abnormality is feasible and robust in CF, using non-contrast-enhanced CT. The multilabel scoring method demonstrated clinical validity for a reproducible and precise evaluation of an entire CF lung, with quantifiable changes over time. Moreover, it could provide an outcome for a sensitive therapeutic response detection in the current context of highly effective CFTR modulator therapy.

Acknowledgment

The authors thank David D Roach, PhD. and Xavier Pineau for technical support. The study was completed in the context of Laboratory of Excellence TRAIL, ANR-10-LABX-57. Dr. Gael Dournes received academic funding from the IdEx, ANR-10-IDEX-03-02, and the French Society of Radiology (grant Alain Rahmouni 2019-2020).

References

1. Bell SC, Mall MA, Gutierrez H, Macek M, Madge S, Davies JC, Burgel P-R, Tullis E, Castaños C, Castellani C, Byrnes CA, Cathcart F, Chotirmall SH, Cosgriff R, Eichler I, Fajac I, Goss CH, Drevinek P, Farrell PM, Gravelle AM, Havermans T, Mayer-Hamblett N, Kashirskaya N, Kerem E, Mathew JL, McKone EF, Naehrlich L, Nasr SZ, Oates GR, O'Neill C, et al. The future of cystic fibrosis care: a global perspective. *Lancet Respir Med* 2020; 8: 65–124.

2. Elborn JS, Ramsey BW, Boyle MP, Konstan MW, Huang X, Marigowda G, Waltz D, Wainwright CE, VX-809 TRAFFIC and TRANSPORT study groups. Efficacy and safety of lumacaftor/ivacaftor combination therapy in patients with cystic fibrosis homozygous for Phe508del CFTR by pulmonary function subgroup: a pooled analysis. *Lancet Respir Med* 2016; 4: 617–626.

3. Heltshe SL, Cogen J, Ramos KJ, Goss CH. Cystic Fibrosis: The Dawn of a New Therapeutic Era. *Am J Respir Crit Care Med* 2017; 195: 979–984.

4. Graeber SY, Boutin S, Wielpütz MO, Joachim C, Frey DL, Wege S, Sommerburg O, Kauczor H-U, Stahl M, Dalpke AH, Mall MA. Effects of Lumacaftor-Ivacaftor on Lung Clearance Index, Magnetic Resonance Imaging and Airway Microbiome in Phe508del Homozygous Patients with Cystic Fibrosis. *Ann Am Thorac Soc* 2021;18(6): 971-980.

5. Tiddens HAWM. Chest computed tomography scans should be considered as a routine investigation in cystic fibrosis. *Paediatr Respir Rev* 2006; 7: 202–208.

6. Brody AS, Guillerman RP. Don't let radiation scare trump patient care: 10 ways you can harm your patients by fear of radiation-induced cancer from diagnostic imaging. *Thorax* 2014; 69: 782–784.

7. Ramsey BW. Use of lung imaging studies as outcome measures for development of new therapies in cystic fibrosis. *Proc Am Thorac Soc* 2007; 4: 359–363.

8. 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.

9. 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.

10. 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.

11. Calder AD, Bush A, Brody AS, Owens CM. Scoring of chest CT in children with cystic fibrosis: state of the art. *Pediatr Radiol* 2014; 44: 1496–1506.

12. Brody AS, Kosorok MR, Li Z, Broderick LS, Foster JL, Laxova A, Bandla H, Farrell PM. Reproducibility of a scoring system for computed tomography scanning in cystic fibrosis. *J Thorac Imaging* 2006; 21: 14–21.

13. Tiddens HAWM, Andrinopoulou E-R, McIntosh J, Elborn JS, Kerem E, Bouma N, Bosch J, Kemner-van de Corput M. Chest computed tomography outcomes in a randomized clinical trial in cystic fibrosis: Lessons learned from the first ataluren phase 3 study. Yammine S, editor. *PLoS ONE* 2020; 15: e0240898.

14. RSNA. International Radiological Societies Tacle Radiologist Shortage. https://www.rsna.org/news/2020/february/international-radiology-societies-and-shortage. Date last updated: February 19, 2020. Date last accessed: June 1st, 2021.

15. Rosenow T, Oudraad MCJ, Murray CP, Turkovic L, Kuo W, de Bruijne M, Ranganathan SC, Tiddens HAWM, Stick SM, Australian Respiratory Early Surveillance Team for Cystic Fibrosis (AREST CF). PRAGMA-CF. A Quantitative Structural Lung Disease Computed Tomography Outcome in Young Children with Cystic Fibrosis. *Am J Respir Crit Care Med* 2015; 191: 1158–1165.

16. DeBoer EM, Swiercz W, Heltshe SL, Anthony MM, Szefler P, Klein R, Strain J, Brody AS, Sagel SD. Automated CT scan scores of bronchiectasis and air trapping in cystic fibrosis. *Chest* 2014; 145: 593–603.

17. Lee SM, Seo JB, Yun J, Cho Y-H, Vogel-Claussen J, Schiebler ML, Gefter WB, van Beek EJR, Goo JM, Lee KS, Hatabu H, Gee J, Kim N. Deep Learning Applications in Chest Radiography and Computed Tomography: Current State of the Art. *J Thorac Imaging* 2019; 34: 75–85.

18. Chassagnon G, Vakalopolou M, Paragios N, Revel M-P. Deep learning: definition and perspectives for thoracic imaging. *Eur Radiol* 2020; 30: 2021–2030.

19. Mongan J, Moy L, Kahn CE. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiol Artif Intell* 2020; 2: e200029.

20. Hwang EJ, Park S, Jin K-N, Kim JI, Choi SY, Lee JH, Goo JM, Aum J, Yim J-J, Cohen JG, Ferretti GR, Park CM, for the DLAD Development and Evaluation Group. Development and Validation of a Deep Learning–Based Automated Detection

Algorithm for Major Thoracic Diseases on Chest Radiographs. *JAMA Netw Open* 2019; 2: e191095.

21. 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.

22. Nezamabadi K, Naseri Z, Moghaddam HA, Modarresi M, Pak N, Mahdizade M. Lung HRCT pattern classification for cystic fibrosis using convolutional neural network. *SIViP* 2019; 13: 1225–1232.

23. Ciompi F, Palaioroutas A, Loeve M, Pujol O, Radeva P, Tiddens H, Bruijne M. Lung tissue classification in severe advanced cystic fibrosis from CT scans. In: Beichel R, editor. *The Fourth International Workshop on Pulmonary Image Analysis.* Toronto, Canada; 2011. p. 57-68.

24. Zucker EJ, Barnes ZA, Lungren MP, Shpanskaya Y, Seekins JM, Halabi SS, Larson DB. Deep learning to automate Brasfield chest radiographic scoring for cystic fibrosis. *J Cyst Fibros* 2020; 19: 131–138.

25. Temple RJ. A regulatory authority's opinion about surrogate endpoints. In: Nimmo WS, Tucker GT, editors. Clinical measurement in drug evaluation. New York: Wiley; 1995. pp. 3–22.

26. Castellani C, Duff AJA, Bell SC, Heijerman HGM, Munck A, Ratjen F, Sermet-Gaudelus I, Southern KW, Barben J, Flume PA, Hodková P, Kashirskaya N, Kirszenbaum MN, Madge S, Oxley H, Plant B, Schwarzenberg SJ, Smyth AR, Taccetti G, Wagner TOF, Wolfe SP, Drevinek P. ECFS best practice guidelines: the 2018 revision.*J Cyst Fibros* 2018; 17: 153–178.

27. 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.

28. Smyth AR, Bell SC, Bojcin S, Bryon M, Duff A, Flume P, Kashirskaya N, Munck A, Ratjen F, Schwarzenberg SJ, Sermet-Gaudelus I, Southern KW, Taccetti G, Ullrich G, Wolfe S. European Cystic Fibrosis Society Standards of Care: Best Practice guidelines. *J Cyst Fibros* 2014; 13: S23–S42.

29. 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.

30. Fuchs HJ, Borowitz DS, Christiansen DH, Morris EM, Nash ML, Ramsey BW, Rosenstein BJ, Smith AL, Wohl ME. Effect of aerosolized recombinant human DNase

on exacerbations of respiratory symptoms and on pulmonary function in patients with cystic fibrosis. The Pulmozyme Study Group. *N Engl J Med* 1994; 331: 637–642.

31. 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.

32. 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.

33. 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.

34. 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.

35. 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.

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

37. 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.

38. Tam S, Boukadoum M, Campeau-Lecours A, Gosselin B. A Fully Embedded Adaptive Real-Time Hand Gesture Classifier Leveraging HD-sEMG and Deep Learning. *IEEE Trans Biomed Circuits Syst* 2020; 14: 232–243.

39. Haute Autorité de Santé, France. Protocole National de Diagnostic et de Soin. https://www.has-sante.fr/portail/upload/docs/application/pdf/2017-09/pnds_2017_vf1.pdf. . Last accessed March 01, 2021.

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

41. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999; 8: 135–160.

42. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–174.

43. Hinkle DE, Wiersma W, Jurs SG. Applied statistics for the behavioral sciences. 5th ed. Boston: Houghton Mifflin; 2003.

45. Armstrong RA. When to use the Bonferroni correction. *Ophthalmic Physiol Opt* 2014; 34: 502–508.

45. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; 521: 436–444.

46. Pellegrino R, Viegi G, Brusasco R, Crapo O, Burgos F, Casaburi R, Coates A, van der Grinten CPM, Gustafsson P, Hankinson J, Jensen R, Johnson DC, MacIntyre N, McKay R, Miller MR, Navajas D, Pedersen OF, Wanger J. Interpretative strategies for lung function tests. *Eur Respir J* 2005; 26: 948–968.

47. European Cystic Fibrosis Society (ECFS). ECFS patient registry. https://www.ecfs.eu/sites/default/files/general-content-files/working-groups/ecfs-patient-registry/ECFSPR_Report_2018_v1.4.pdf. Last accessed: March 01, 2021.

48. 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.

49. Chassagnon G, Hubert D, Fajac I, Burgel P-R, Revel M-P, investigators. Long-term computed tomographic changes in cystic fibrosis patients treated with ivacaftor. *Eur Respir J* 2016; 48: 249–252.

50. Brody AS, Tiddens HAWM, Castile RG, Coxson HO, de Jong PA, Goldin J, Huda W, Long FR, McNitt-Gray M, Rock M, Robinson TE, Sagel SD, CT Scanning in Cystic Fibrosis Special Interest Group. Computed tomography in the evaluation of cystic fibrosis lung disease. *Am J Respir Crit Care Med* 2005; 172: 1246–1252.

51. Tepper LA, Caudri D, Utens EMWJ, van der Wiel EC, Quittner AL, Tiddens HAWM. Tracking CF disease progression with CT and respiratory symptoms in a cohort of children aged 6-19 years. *Pediatr Pulmonol* 2014; 49: 1182–1189.

52. Kuo W, Kemner-van de Corput MPC, Perez-Rovira A, de Bruijne M, Fajac I, Tiddens HAWM, van Straten M, ECFS-CTN/SCIFI CF study group. Multicentre chest computed tomography standardisation in children and adolescents with cystic fibrosis: the way forward. *Eur Respir J* 2016; 47: 1706–1717.

53. Dournes G, Walkup LL, Benlala I, Willmering MM, Macey J, Bui S, Laurent F, Woods JC. The clinical use of lung MRI in cystic fibrosis: what, now, how? *Chest* 2020; : S0012369220354532.

54. Zha W, Fain SB, Schiebler ML, Evans MD, Nagle SK, Liu F. Deep convolutional neural networks with multiplane consensus labeling for lung function quantification using UTE proton MRI. *J Magn Reson Imaging* 2019; 50: 1169–1181.

55. Kuo W, Soffers T, Andrinopoulou E-R, Rosenow T, Ranganathan S, Turkovic L, Stick SM, Tiddens HAWM, AREST CF. Quantitative assessment of airway dimensions in young children with cystic fibrosis lung disease using chest computed tomography. *Pediatr Pulmonol* 2017; 52: 1414–1423.

56. Ludes C, Schaal M, Labani A, Jeung M-Y, Roy C, Ohana M. [Ultra-low dose chest CT: The end of chest radiograph?]. *Presse Med* 2016; 45: 291–301.

57. Peng Y, Chen S, Qin A, Chen M, Gao X, Liu Y, Miao J, Gu H, Zhao C, Deng X, Qi Z. Magnetic resonance-based synthetic computed tomography images generated using

generative adversarial networks for nasopharyngeal carcinoma radiotherapy treatment planning. *Radiother Oncol* 2020; 150: 217–224.

Tables

Table 1. Patient characteristics at initial evaluation

		Test Cohort (n=36)	**Clinical Validation Cohort (n=70)**	
			n=10 patients with lumacaftor/ivacaftor	n=60 patients without lumacaftor/ivacaftor
Age	Years	13 (4-54)	13.5 (12-37)	15 (6-48)
Gender	Male/Female	16/20	5/5	28/32
Body mass index	$kg.m^{-2}$	18 (13-27)	17 (12-23)	19.5 (13-33)
Genetic mutation	DeltaF508 homozygous/heterozygous	18/14	10/0	24/36
Pancreatic insufficiency*	Yes/no	21/11	9/1	45/15
Diabetes Mellitus*	Yes/no	0/32	0/10	3/57
Hepatobiliary disease*	Yes/no	3/29	2/8	8/52
PFT*	FEV1%	77 (30-124)	71 (44-104)	76 (22-123)
	100xFEV1/FVC	75 (46-109)	68 (62-89)	73 (49-97)
	100xRV/TLC	31.8 (14-110)	31.4 (13-72)	32 (19-86)
Visual CT score	modified Brody score	39 (0-156)	45.5 (5-152)	48 (0-153)
Chronic colonization*	*Pseudomonas aeruginosa (yes/no)*	11/21	3/7	16/44

Staphylococcus aureus (yes/no)	12/20	4/6	29/31

Data are median with (minimum-maximum) range during the initial evaluation of CF patients.

*In the Test cohort, data were missing in 4 CF patients.

Legends: CF=cystic fibrosis; PFT=pulmonary function tests; FEV1=forced expiratory volume in 1 second; FVC=forced vital capacity; RV=residual volume; TLC=total lung capacity; %=percentage predicted; mBrody score=modified Brody score

Table 2. Semantic evaluation of 2D pixel similarity between AI-driven and ground-truth labels in the Test cohort.

Overall pixelwise similarity N=11435 axial CT slices	**Balanced Accuracy**	**Dice**	**Recall**	**Precision**
Bronchiectasis	0.91	0.86	0.79	0.90
Peribronchial thickening	0.81	0.69	0.61	0.78
Bronchial mucus plug	0.87	0.79	0.73	0.87
Bronchiolar mucus plug	0.68	0.49	0.37	0.78
Collapse/Consolidation	0.85	0.75	0.66	0.86
Total Abnormal Lung	0.82	0.71	0.63	0.84
Lung Parenchyma	0.99	0.99	0.99	0.99

Note: Owing to the number of 11435 CT slices, the confidence interval of the pixelwise similarity measurements was considered negligible. The Total Abnormal Lung corresponds to the average of five structural alteration similarity results.

Table 3. Correlations of normalized label volumes with pulmonary function test and visual CT scoring in the Test cohort and Clinical Validation cohort.

Test cohort	**AI-driven labeling**				**Manual GT labeling**			
	FEV1% (n=32)		**mBrody score** (n=36)		**FEV1%** (n=32)		**mBrody score** (n=36)	
	rho	p-value	rho	p-value	rho	p-value	rho	p-value
Bronchiectasis	-0.54	**0.001**	0.72	**<0.001**	-0.50	**0.003**	0.70	**<0.001**
Peribronchial thickening	-0.49	**0.004**	0.81	**<0.001**	-0.49	**0.004**	0.74	**<0.001**
Bronchial mucus plug	-0.69	**<0.001**	0.77	**<0.001**	-0.61	**<0.001**	0.73	**<0.001**
Bronchiolar mucus plug	-0.36	**0.04**	0.49	**0.002**	-0.39	**0.02**	0.58	**<0.001**
Collapse/Consolidation	-0.48	**0.004**	0.45	**0.006**	-0.50	**0.003**	0.45	**0.006**
Total Abnormal Volume	-0.63	**<0.001**	0.77	**<0.001**	-0.63	**<0.001**	0.78	**<0.001**

Independent Clinical Validation cohort	**AI-driven labeling at initial evaluation**				**AI-driven labeling at follow-up evaluation**			
	FEV1% (n=70)		**mBrody score** (n=70)		**FEV1%** (n=70)		**mBrody score** (n=70)	
	rho	p-value	rho	p-value	rho	p-value	rho	p-value
Bronchiectasis	-0.46	**<0.001**	0.76	**<0.001**	-0.54	**<0.001**	0.69	**<0.001**
Peribronchial thickening	-0.50	**<0.001**	0.74	**<0.001**	-0.56	**<0.001**	0.67	**<0.001**
Bronchial mucus plug	-0.59	**<0.001**	0.72	**<0.001**	-0.61	**<0.001**	0.70	**<0.001**
Bronchiolar mucus plug	-0.48	**<0.001**	0.65	**<0.001**	-0.54	**0.001**	0.59	**<0.001**
Collapse/Consolidation	-0.43	**<0.001**	0.54	**<0.001**	-0.59	**<0.001**	0.64	**<0.001**
Total Abnormal Volume	-0.55	**<0.001**	0.82	**<0.001**	-0.68	**<0.001**	0.80	**<0.001**

Table 4. Paired-comparisons in CF at initial evaluation and at follow-up, with or without lumacaftor/ivacaftor treatment.

Note: data are Spearman's rho correlation coefficients. The follow-up evaluation was performed at 1 year in CF patients with lumacaftor/ivacaftor treatment (n=10) and 2 years in CF patients without lumacaftor/ivacaftor treatment (n=60).

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

Legends: AI=artificial intelligence; GT=ground truth; CF=cystic fibrosis; FEV1%=forced expiratory volume in 1 second; mBrody=modified Brody score

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**
Normalized AI volumes	Bronchiectasis	Median	21.1	14.1	0.59	15.5	20.4	**0.002**
		Range	(0-115)	(0-126)		(0-315)	(0.3-357)	
	Peribronchial thickening	Median	18.4	12.1	**0.005**	23	25.7	**0.008**
		Range	(0.1-57)	(0-33)		(0-188)	(0.6-220)	
	Bronchial mucus plug	Median	4.1	3.6	**0.005**	10.3	6.8	0.64
		Range	(0.2- 67)	(0-56)		(0-186)	(0.3-148)	
	Bronchiolar mucus plug	Median	29.0	6.0	**0.007**	5.0	6.3	0.69
		Range	(0.3-105)	(0-67)		(0-94)	(0.0-136)	
	Collapse/Consolidation	Median	9.7	2.9	**0.02**	4.1	3.4	0.41
		Range	(0-123)	(0-56)		(0-102)	(0.3-114)	
	Total Abnormal Volume	Median	170.0	51.5	**0.005**	73.3	80.2	0.46
		Range	(2.9-452)	(2.4-213)		(0-601)	(2.6-663)	
PFT	FEV1%	Median	71	82.5	0.058	78.5	73.5	0.08
		Range	(44-104)	(44-118)		(21-123)	(22-128)	
Visual CT score	mBrody score	Median	45.5	35	0.06	48	52	0.18
		Range	(5-152)	(5-125)		(0-153)	(2-155)	

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 1 year; M24=second evaluation at 2 years; PFT=pulmonary function test; FEV1%=forced expiratory volume in 1 second percentage predicted; mBrody score=modified Brody score

Figures Legends

Figure 1. Study flow chart. CF=cystic fibrosis; CT=computed tomography.

Figure 2. Concordance between AI-driven and manual volumes in the Test cohort. The volume values are given in milliliters. Black circles (•) represent data acquired with a CT machine from GE® manufacturer, and white triangles (Δ) represent data obtained with a CT machine from Siemens® manufacturer. The black diagonal lines indicate the lines of equality. CCC=concordance correlation coefficient.

Figure 3. Examples of AI-driven semantic labeling from the Clinical Validation cohort, in three patients with cystic fibrosis, a 14-year-old male (A, B, C), a 13-year-old male (D, E, F), and a 32-year-old female (G, H, I). CF patients had an increased level of disease severity from top to bottom, as assessed by the forced expiratory volume in 1 second (FEV1%). The left column shows axial CT slices (A, D, G). The middle column shows the corresponding AI-driven semantic labeling (B, E, H). By integrating all individual 2D labelings over the entire CT scans, 3D reconstructions were allowed and displayed in coronal view (C, F, I). In panels E, F, H, I, blue arrows and blue labels highlight areas of central mucus plugs; red arrow and red labels show mucus-free lumen dilatations; green arrow and green labels show peribronchial thickening; cyan arrow and cyan labels show a consolidation.

Figure 4. Comparison of AI-driven semantic labeling in the Clinical Validation cohort, before (A, C, E) and after (B, D, F) 1 year of treatment by lumacaftor/ivacaftor in a 15-year-old male with cystic fibrosis. 2D axial CT slices (A, B) are shown, with AI-driven semantic labeling displayed in the corresponding axial slice (C, D). By integrating all individual 2D labelings over the entire CT scan, a 3D reconstruction is allowed and displayed in coronal view (E, F). Bronchiolar mucus plugs with the “tree-in-bud” pattern are labeled in yellow color, mucus-free bronchial lumen dilatation in red color, and peribronchial thickening in green color (C, D, E, F). Yellow arrows emphasize the example of bronchiolar mucus plugs (C) that were not visible at one year (D). There was a reduction in total abnormal volume (TAV) before and after treatment (E, F). FEV1% was normal at baseline (104%) and remained stable at follow-up (103%).

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 ϵ .
- Assume a binary classification task (yes/no)
- Assume the error rate of each independent classifier is better than random guessing (*i.e.*, ϵ 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-ϵ) and $X_k = 1$if the classifier k makes a wrong prediction (this happens with a probability ϵ).

  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 <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 certainty of the result is expected to be high, since it is performed 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 measured ratio, and n is the number of evaluations (herein n=11435).

For instance, for a similarity metric of 0.71, the 95% confidence interval would be [0.70-0.71], and this narrow interval would be similar for all metrics. In this evaluation, we have assumed that the 95% confidence interval of the similarity metrics would be 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) Somatom Definition 64® (Site1, n=8 ; Site2, n=10) Somatom Force® (Site1, n=9) Somatom Emotion® (Site3, n=4) GE LightSpeed 16® (Site3, n=9) GE LightSpeed VGT® (Site3, n=6) GE Revolution® (Site2, n=16)	STD (n=31) B40s (n=20) Br40 (n=7) I30f (n=20)	FBP (n=42) ASiR (n=16) SAFIRE (n=20)	(8-260)	(100-140)	(5-40)	(1-1.25)
Test	Somatom Sensation 16® (Site1, n=2 ; Site2, n=5) Somatom Definition 64® (Site1, n=3 ; Site2, n=6) Somatom Force® (Site1, n=5) Somatom Emotion (Site3, n=1) GE LightSpeed® 16 (Site3, n=2) GE LightSpeed VGT® (Site3, n=2) GE Revolution® (Site2, n=10)	STD (n=14) B40s (n=8) Br40 (n=5) I30F (n=9)	FBP (n=12) ASiR (n=10) SAFIRE (n=14)	(9-210)	(100-140)	(5-40)	(1-1.25)
Clinical Validation	Somatom Sensation 16® (Site1, n=28 ; Site2, n=35) Somatom Definition 64® (Site1, n=37 ; Site2, n=40)	B40s (n=53) Br40 (n=22) I30F (n=65)	FBP (n=75) 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] where 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 over 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 number of 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) Somatom Sensation 16® (n=27)	Somatom Definition 64® (n=34) Somatom Sensation 16® (n=26)
Kernel		I30f (n=10)	I30f (n=10)	I30f (n=22) Br40 (n=11) B40s (n=27)	I30f (n=23) Br40 (n=11) B40s (n=26)
Reconstruction		SAFIRE (n=10)	SAFIRE (n=10)	FBP (n=38) SAFIRE (n=22)	FBP (n=37) 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 repeatibility of AI and manual interobserver similarity in the Clinical Validation cohort

2D pixelwise similarity n=42280 CT slices in 140 CTs	**AI_1 vs. AI_2** **Dice**	**AI_1 vs. AI_1** **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

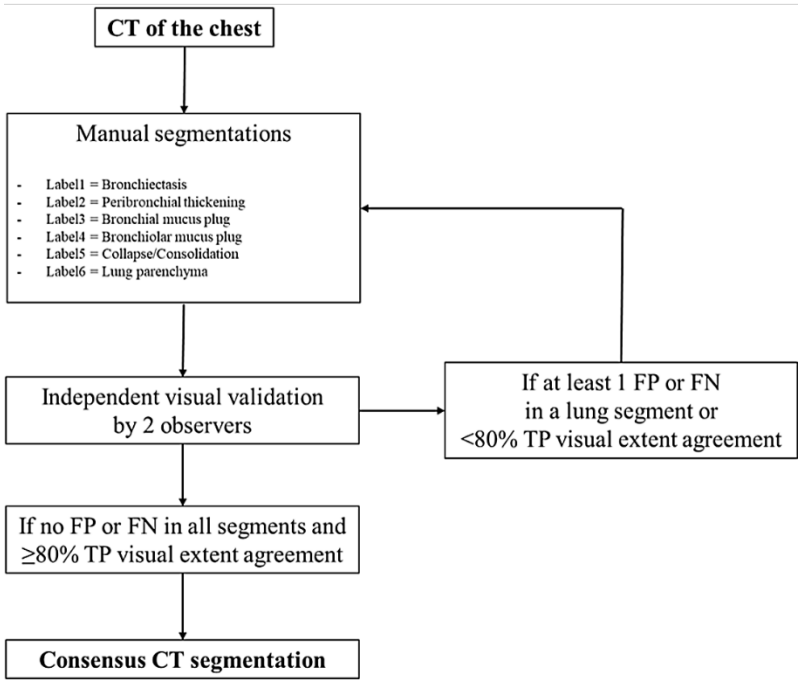

Figure E1. Flow chart of the method to produce consensus CT semantic segmentation for Training. The segmentations were visually checked at the segmental level. TP=true positive; FP=false positive; FN=false negative.

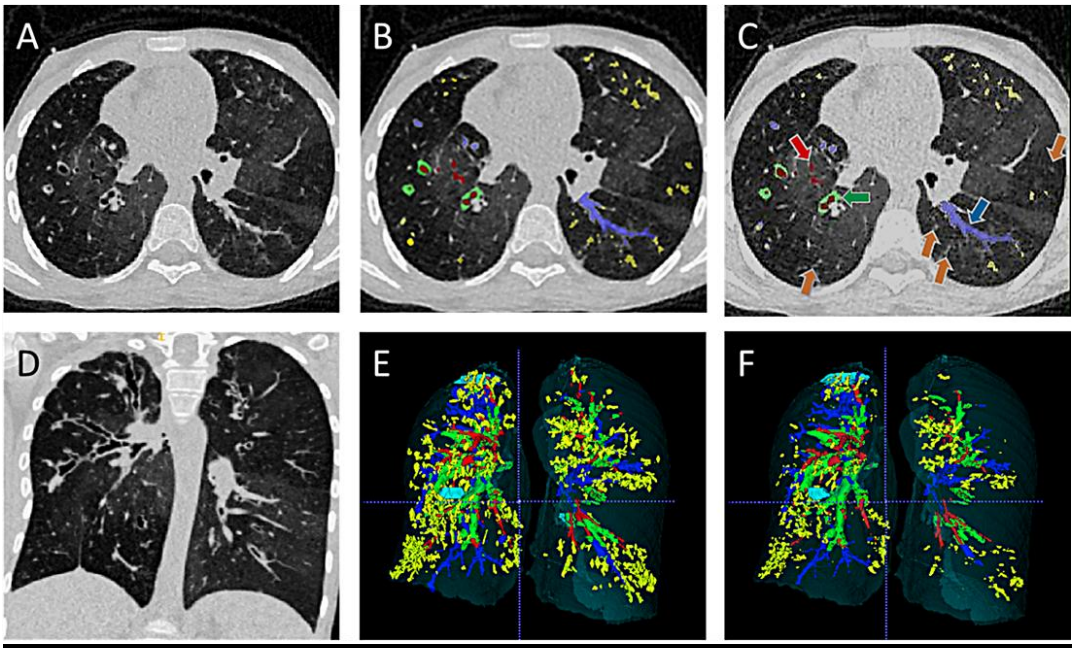

Figure E2. Axial (A) and coronal reformations (D) of a lung CT scan acquired in a 15-year-old female with cystic fibrosis. Manual (B, E) and AI-driven (C, F) semantic multilabel segmentation are shown and displayed in corresponding axial CT slice (B, C) and volume rendering in coronal view (E, F). In 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.

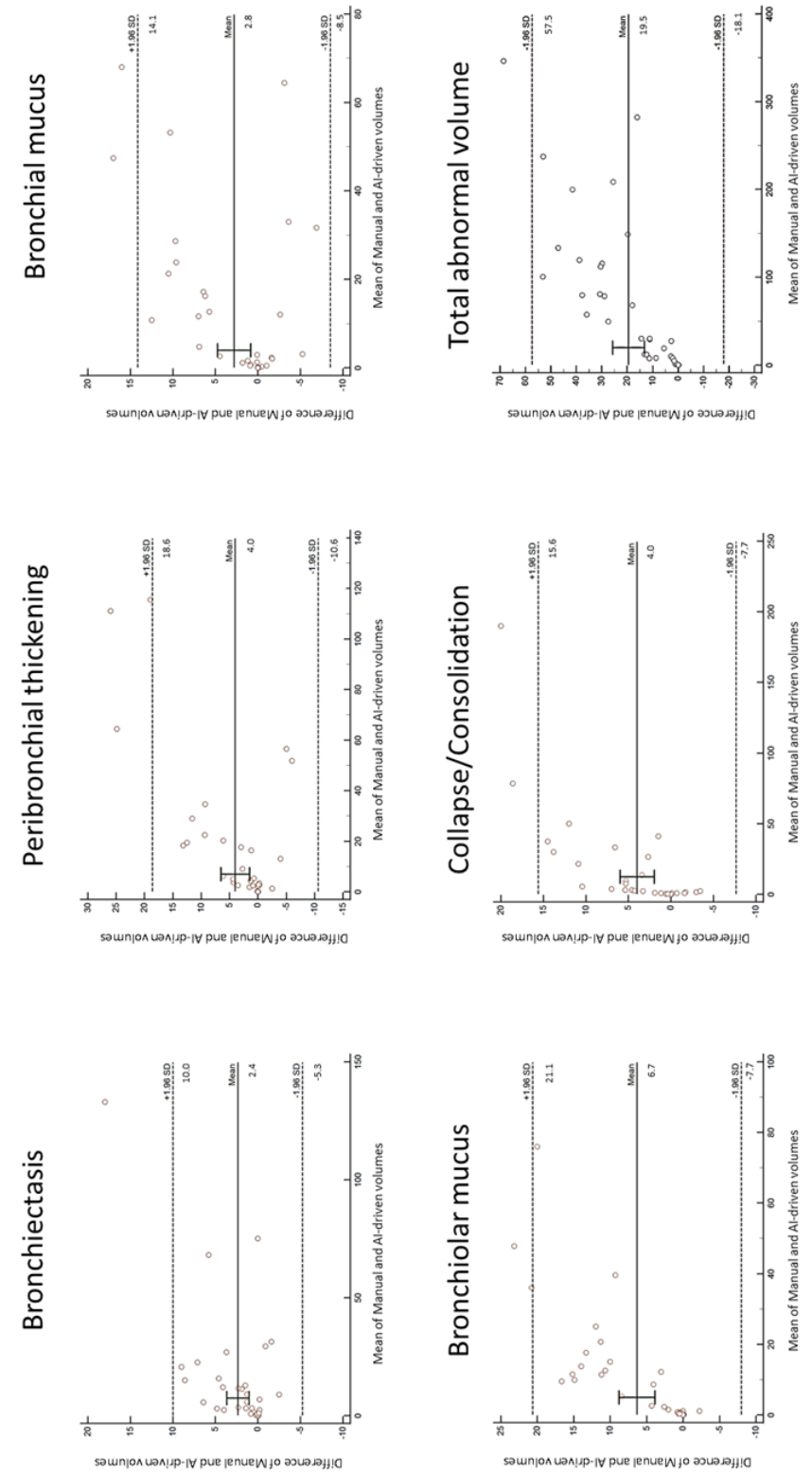
Bronchial mucus
+1.96 SD
14.1
Mean
2.8
-1.96 SD
-8.5
Difference of Manual and AI-driven volumes
Mean of Manual and AI-driven volumes
Total abnormal volume
+1.96 SD
57.5
Mean
19.5
-1.96 SD
-18.1
Difference of Manual and AI-driven volumes
Mean of Manual and AI-driven volumes
Peribronchial thickening
+1.96 SD
18.6
Mean
4.0
-1.96 SD
-10.6
Difference of Manual and AI-driven volumes
Mean of Manual and AI-driven volumes
Collapse/Consolidation
+1.96 SD
15.6
Mean
4.0
-1.96 SD
-7.7
Difference of Manual and AI-driven volumes
Mean of Manual and AI-driven volumes
Bronchiectasis
+1.96 SD
10.0
Mean
2.4
-1.96 SD
-5.3
Difference of Manual and AI-driven volumes
Mean of Manual and AI-driven volumes
Bronchiolar mucus
+1.96 SD
21.1
Mean
6.7
-1.96 SD
-7.7
Difference of Manual and AI-driven volumes
Mean of Manual and AI-driven volumes

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.

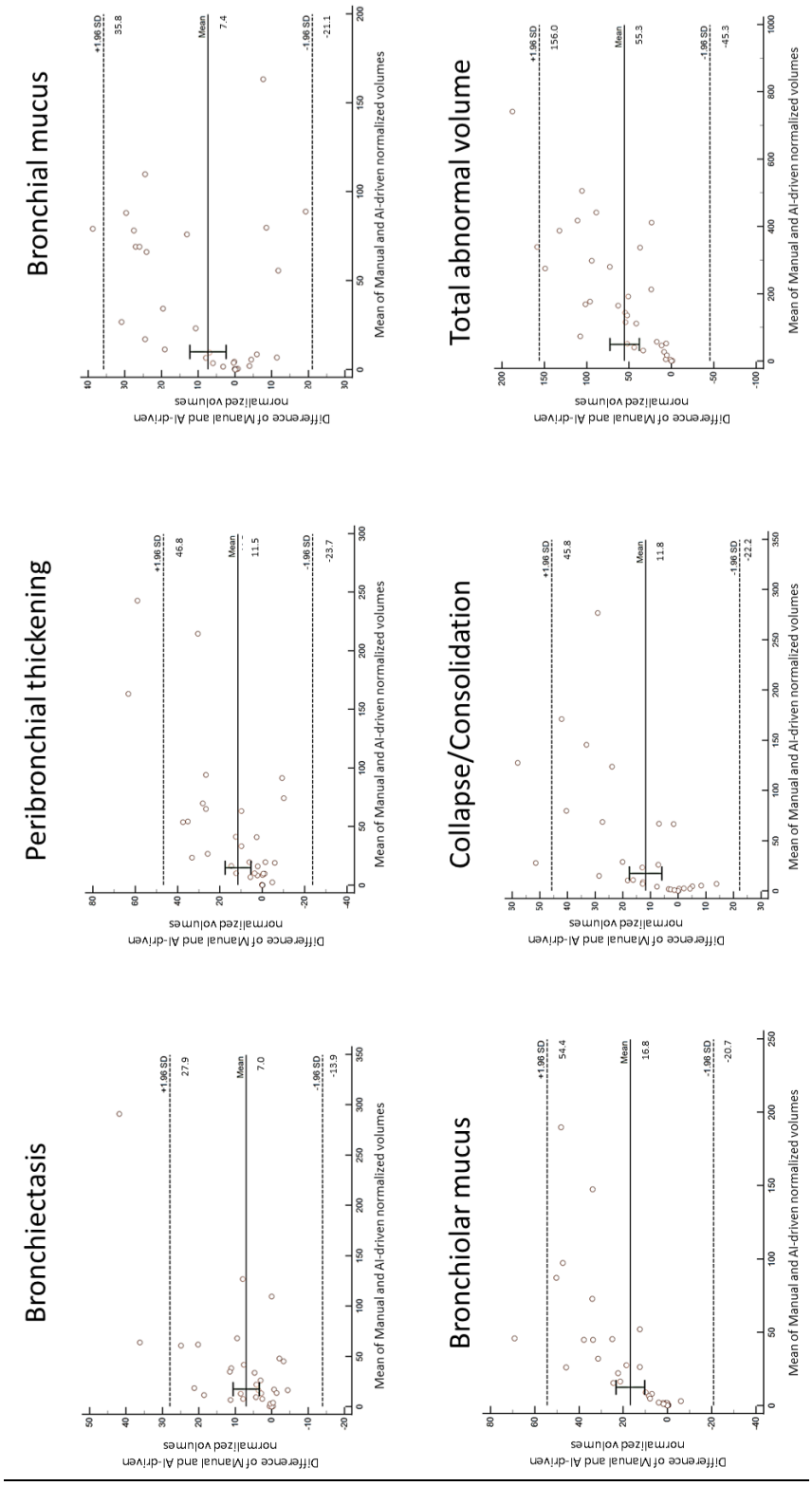

Figure E4. Bland-Altman analyses of manual versus AI-driven normalized volumes in the Test cohort (n=36). The plain lines represent the mean difference and the bars their 95% confidence interval; the dashed lines represent the limits of agreement.

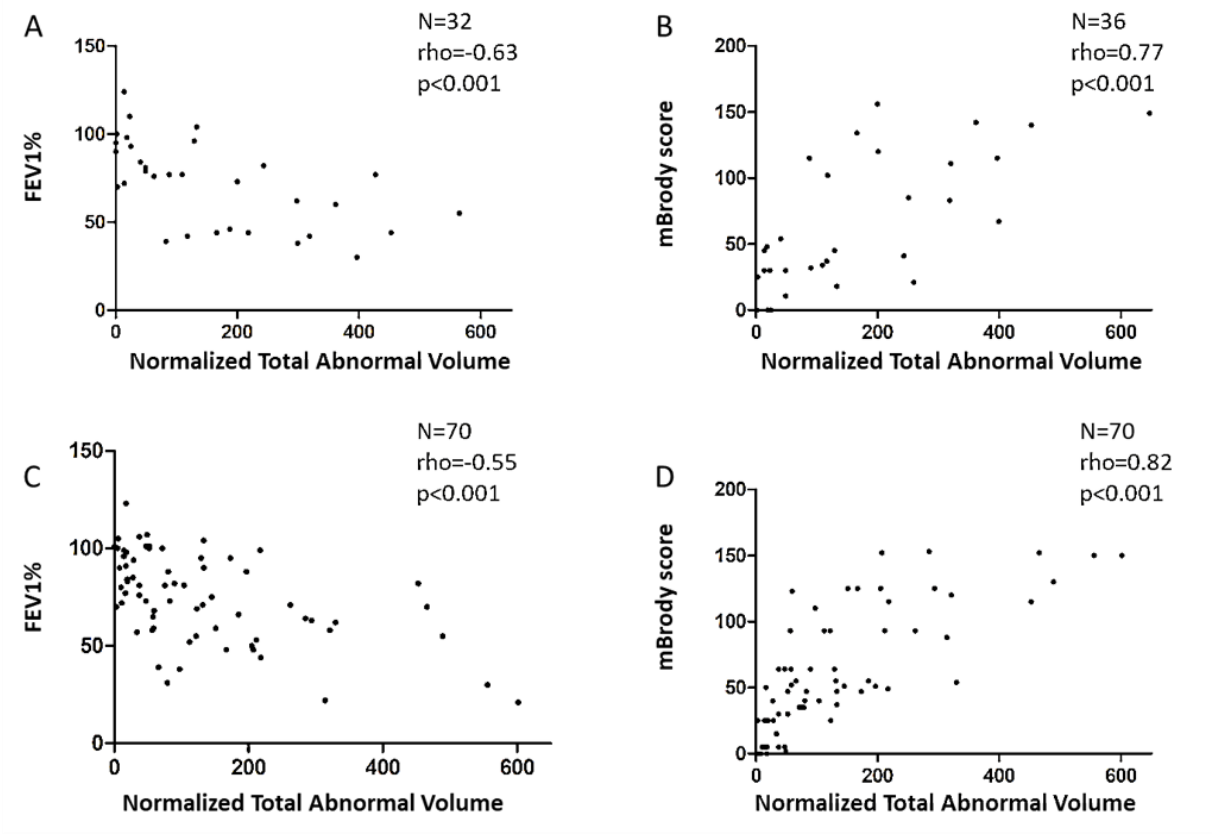

Figure E5. Spearman's correlations between AI-driven measurement of normalized total abnormal volume and CF disease severity, as assessed by the forced expiratory volume in 1-second percentage predicted (FEV1%; A and C) and the modified Brody score (mBrody; B and D). Results are shown for both the Test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

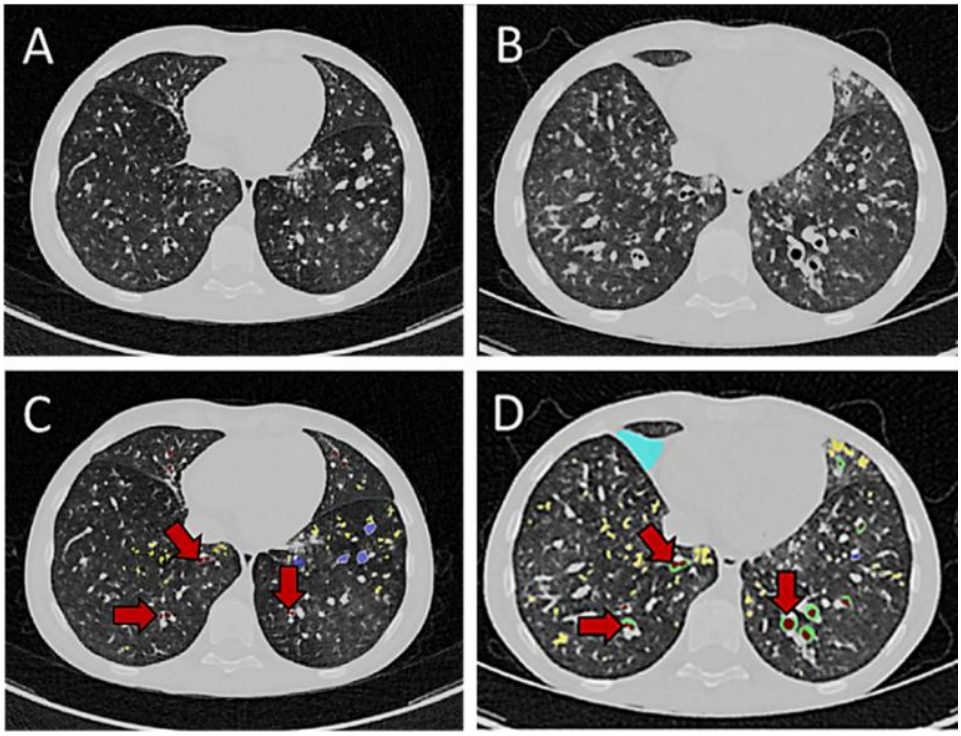

Figure E6. Comparison of AI-driven semantic labeling in the Clinical Validation cohort, at initial evaluation (A, C) and after two years (B, D) of standard management, in a 15-year-old male with cystic fibrosis. Axial CT slices (A, B) are shown, with AI-driven semantic labeling displayed in the corresponding 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.