Chronic lung allograft dysfunction phenotype and prognosis by machine learning CT analysis

Micheal C. McInnis[1], Jin Ma[2], Gauri Rani Karur[1], Christian Houbois[1,3], Liran Levy[4], Jan Havlin[4], Eyal Fuchs[4], Jussi Tikkanen[4], Chung-Wai Chow[4,5], Ella Huszti[2] and Tereza Martinu[4,5]

[1]Dept of Medical Imaging, University Health Network, University of Toronto, Toronto, ON, Canada. [2]Biostatistics Research Unit, University Health Network, University of Toronto, Toronto, ON, Canada. [3]Dept of Diagnostic and Interventional Radiology, University of Cologne, Cologne, Germany. [4]Toronto Lung Transplant Program, Ajmera Transplant Centre, University Health Network, Toronto, ON, Canada. [5]Division of Respirology, Dept of Medicine, University of Toronto, ON, Canada.

Corresponding author: Micheal C. McInnis (micheal.mcinnis@uhn.ca)

Shareable abstract (@ERSpublications)
Machine learning CT lung texture analysis and radiologist analysis both predict prognosis in chronic lung allograft dysfunction (CLAD), independent of CLAD phenotype. Machine learning can discriminate between CLAD phenotypes without expiratory CT. https://bit.ly/3G0Ro4M

Cite this article as: McInnis MC, Ma J, Karur GR, *et al*. Chronic lung allograft dysfunction phenotype and prognosis by machine learning CT analysis. *Eur Respir J* 2022; 60: 2101652 [DOI: 10.1183/13993003.01652-2021].

Received: 11 June 2021
Accepted: 23 Nov 2021

Abstract

Background Chronic lung allograft dysfunction (CLAD) is the principal cause of graft failure in lung transplant recipients and prognosis depends on CLAD phenotype. We used a machine learning computed tomography (CT) lung texture analysis tool at CLAD diagnosis for phenotyping and prognostication compared with radiologist scoring.

Methods This retrospective study included all adult first double lung transplant patients (January 2010–December 2015) with CLAD (censored December 2019) and inspiratory CT near CLAD diagnosis. The machine learning tool quantified ground-glass opacity, reticulation, hyperlucent lung and pulmonary vessel volume (PVV). Two radiologists scored for ground-glass opacity, reticulation, consolidation, pleural effusion, air trapping and bronchiectasis. Receiver operating characteristic curve analysis was used to evaluate the diagnostic performance of machine learning and radiologist for CLAD phenotype. Multivariable Cox proportional hazards regression analysis for allograft survival controlled for age, sex, native lung disease, cytomegalovirus serostatus and CLAD phenotype.

Results 88 patients were included (57 bronchiolitis obliterans syndrome (BOS), 20 restrictive allograft syndrome (RAS)/mixed and 11 unclassified/undefined) with CT a median 9.5 days from CLAD onset. Radiologist and machine learning parameters phenotyped RAS/mixed with PVV as the strongest indicator (area under the curve (AUC) 0.85). Machine learning hyperlucent lung phenotyped BOS using only inspiratory CT (AUC 0.76). Radiologist and machine learning parameters predicted graft failure in the multivariable analysis, best with PVV (hazard ratio 1.23, 95% CI 1.05–1.44; p=0.01).

Conclusions Machine learning discriminated between CLAD phenotypes on CT. Both radiologist and machine learning scoring were associated with graft failure, independent of CLAD phenotype. PVV, unique to machine learning, was the strongest in phenotyping and prognostication.

Introduction

Chronic lung allograft dysfunction (CLAD) is the principal cause of graft failure in lung transplant recipients. Bronchiolitis obliterans syndrome (BOS) is the most frequent CLAD phenotype and is characterised by progressive obstructive lung disease [1]. A restrictive phenotype, restrictive allograft syndrome (RAS), is now well recognised, and is associated with parenchymal fibrosis and a worse prognosis, but is less frequent than BOS [2]. In addition to these two phenotypes, the International Society of Heart and Lung Transplantation 2019 consensus statement added a mixed phenotype and an undefined phenotype [1]. Furthermore, a small but significant proportion of patients may be unclassifiable using the existing framework [3].

Imaging of CLAD is best documented for BOS [4]. On computed tomography (CT), BOS is characterised by geographic regions of air trapping that manifest as hyperlucent lung, accentuated by exhalation. Bronchial wall thickening occurs with bronchiectasis developing over time, often in association with mucus plugging and clustered nodules [5]. The typical appearance of RAS is consolidation and ground-glass opacity evolving into reticulation and traction bronchiectasis with pleural and parenchymal fibrosis, more commonly in the upper lung zones [5]. CT features have been associated with allograft survival in double lung transplant patients using a radiologist-scored approach [3, 6–8] and when using quantitative measures of lung density, volume, parametric response mapping and airway measures [9–15].

The CALIPER (Computer-Aided Lung Informatics for Pathology Evaluation and Rating) tool is a validated machine learning algorithm that performs automated, quantitative analysis of lung texture on CT [16]. It segments the lung from the chest and classifies each portion of the lung as either normal, hyperlucent, reticular, ground-glass or honeycomb texture. It has been principally used in the evaluation of idiopathic pulmonary fibrosis where it is superior to visual analysis in quantifying extent of disease and predicting survival [17–19]. The strongest predictor of survival is pulmonary vessel volume (PVV), a biomarker derived from the machine learning analysis with no direct visual correlate [20].

Although hyperlucent, ground-glass and reticular textures are typical features of diffuse lung disease in nontransplant patients, they are also reported features of CLAD. For this reason, we hypothesised that an existing machine learning analysis tool may be useful in the phenotyping of CLAD and in predicting graft survival. Indeed, quantitative assessment of chest CT has been identified as a key future direction in CLAD research [1] and we sought to address this need.

Materials and methods

Patient selection

This single-centre retrospective study has been approved by the University Health Network Research Ethics Board (REB 15-9531-AE). All adult first double lung transplant patients transplanted between 2010 and 2015 were selected. Exclusion criteria included single lung transplants, heart–lung transplants and re-transplants. We then excluded those who died without CLAD, those who were CLAD-free up to 31 December 2019 and those who had non-CLAD pulmonary function decline. Patients were required to have had a thin-section CT within 100 days after CLAD onset or, if unavailable, within 28 days before onset. Clinical characteristics were extracted from the medical record and date of CLAD onset was derived. CLAD phenotype and date of onset was assigned according to previously described methodology and considered the gold standard for the subsequent sensitivity analyses [3]. Graft failure was defined as death or re-transplantation.

We further collected virtual crossmatch status at the time of transplant (*i.e.* donor-specific antibodies (DSA) at transplant), the presence of DSA at CLAD onset based on the last test before onset or ⩽3 weeks after onset, as well as development of *de novo* DSA between transplant and CLAD onset. The pre-CLAD infection score (number of pre-CLAD bronchoalveolar lavage with significant pathogens divided by the number of all pre-CLAD bronchoalveolar lavage) and pre-CLAD acute cellular rejection score (sum of histological A grades divided by the number of all evaluable biopsies) were also calculated.

Radiological follow-up

Our routine lung transplant CT protocol is a low-dose scan acquired at end-inspiration and followed by a minimal-dose scan at end-expiration. Generally, a low-dose scan is performed at around one-third the dose of a regular dose scan and the minimal-dose scan is performed at one-tenth the dose of a regular dose scan. Surveillance CT scans are performed at 3-month intervals for the first year post-transplant and then at 18 and 24 months, with CT also being performed as needed when CLAD is suspected. All patients were scanned on one of three multislice CT units (Aquilion, Aquilion ONE or Aquilion PRIME; Canon Medical, Otawara, Japan) at Toronto General Hospital (Toronto, ON, Canada). For inclusion in the study, the CT scan had to be noncontrast and have a transverse image series in a mediastinal reconstruction algorithm with contiguous slices of thickness ⩽1 mm. Each scan was reviewed by a fellowship-trained thoracic radiologist (M.C.M., 5 years of experience) to assess for factors that may interfere with texture analysis, including respiratory motion artefacts and patient intubation.

Machine learning ("ML") analysis

Qualifying CT scans were analysed using the Lung Texture Analysis tool (Imbio, Minneapolis, MN, USA) based on CALIPER, with the development and training of the machine learning algorithm having been previously described [16]. This is a tool that performs a texture analysis in interstitial lung disease but was

not trained on our lung transplant cohort. In brief, the lung is segmented from the chest wall, and the large airways and central vasculature are removed. It then assigns a texture (normal lung (NL_{ML}), hyperlucent lung (HL_{ML}), ground-glass opacity (GGO_{ML}), reticular (RET_{ML}) or honeycombing) to each voxel and a colour overlay output is made available for review on the picture archiving and communications system (Coral, Toronto, ON, Canada) (figure 1). Quantitative data in the form of total lung capacity (CT_{TLC}), volumes for each assigned texture and pulmonary vessel volume (PVV_{ML}) are provided in a comma-separated values file output from a locally hosted server. A thoracic radiologist (M.C.M.) reviewed all CT scans with the output overlay present to exclude studies where segmentation failed (*i.e.* extrapulmonary structures comprised an estimated >1% of lung volume).

Radiologist ("RAD") analysis

CT scans were independently evaluated by two fellowship-trained cardiothoracic radiologists each with 5 years of experience (G.R.K. and C.H.), who were blinded to all clinical information, outcome and machine learning results. A semiquantitative scoring system was used as previously described by SUHLING *et al.* [6]. In brief, a CT image of the upper, mid and lower lung was selected at a predetermined interval (25th, 50th and 75th percentile image) and presented in lung windows (width 1500 HU, level −600 HU) and mediastinal windows (width 350 HU, level 40 HU). Each lung was evaluated for consolidation (CON_{RAD}), ground-glass opacity (GGO_{RAD}), reticulation (RET_{RAD}), traction bronchiectasis and pleural

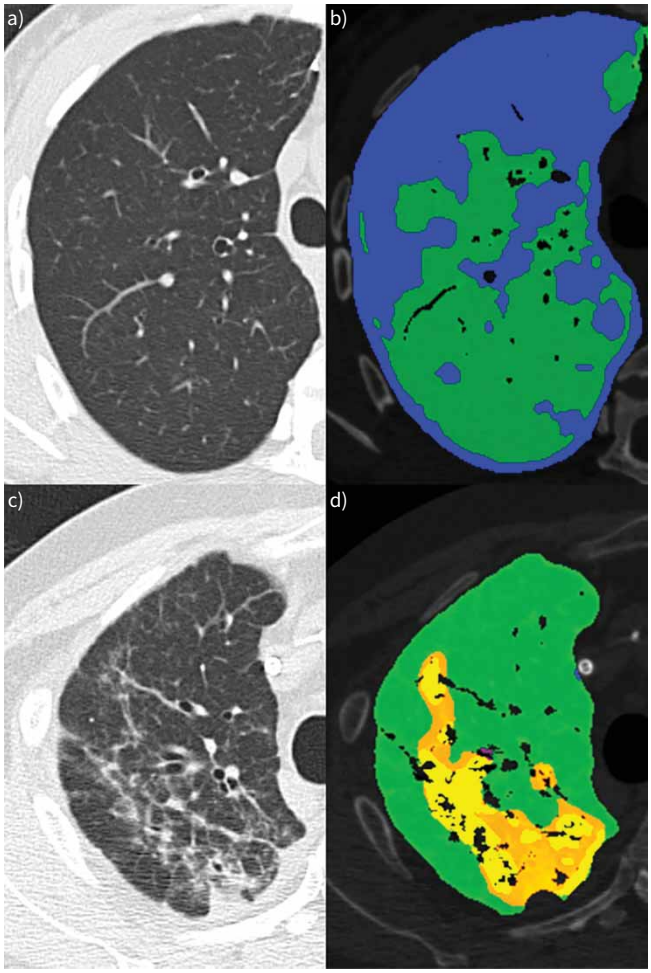

FIGURE 1 Inspiratory computed tomography (CT) and machine learning output in chronic lung allograft dysfunction. a, b) In a 36-year-old woman with bronchiolitis obliterans syndrome, a) conventional CT analysis demonstrates normal appearing lung that, on b) texture analysis, is a mix of normal (green) and hyperlucent (blue). c, d) In a 40-year-old man with a mixed phenotype, c) conventional CT analysis demonstrates ground-glass opacity and reticulation that, on d) texture analysis, is also classified as a mixture of ground-glass opacity (yellow) and reticulation (orange).

effusion using established definitions [21]. Each lung was also evaluated for air trapping (AT_{RAD}) using the same selected images paired with an expiratory image at the same level, when expiratory was performed. Each finding in each lung was scored on a grading of 0–2 (0: abnormality was not present; 1: abnormality involved <10% of the lung; 2: abnormality involved >10% of the lung). The radiologists were initially trained on a set of 18 images derived from six representative CLAD patients who were not included in this study.

Statistical analysis

Continuous variables were described using median (interquartile range (IQR)) or mean with standard deviation and categorical variables were described using number (percentage). Comparisons between two groups were made by the independent samples t-test or Mann–Whitney U-test for continuous variables and Fisher's exact test or the Chi-squared test for categorical variables. Variables analysed by a t-test were evaluated for normality using histograms or a Shapiro–Wilks test, when appropriate.

For radiologist scoring, agreement was assessed with intraclass correlation coefficients (ICCs) for GGO_{RAD}, RET_{RAD}, CON_{RAD} and AT_{RAD}, and with weighted κ for pleural effusion and bronchiectasis. The mean of the two radiologist's scores was used in analysis. For machine learning scoring, all variables were divided by CT_{TLC} to derive the proportion of total lung involved and are expressed as a percentage value. The relationship between AT_{RAD} and HL_{ML} was evaluated using the Pearson correlation coefficient. The relationship between CT_{TLC} and TLC from pulmonary function testing at the time of CLAD diagnosis was evaluated using the Pearson correlation coefficient.

The radiologist and machine learning scores were compared between CLAD phenotype groups using logistic regression models. RAS and mixed CLAD phenotypes were grouped to account for small numbers, as were unclassified and undefined phenotypes. Receiver operating characteristic (ROC) curve analysis was used to determine the diagnostic performance for CLAD phenotyping.

Cox proportional hazards regression analysis was used to assess the association of radiologist and machine learning scoring with graft failure, and proportional hazards assumption validity was tested. Univariable and multivariable Cox proportional hazards regression analyses were performed for each score adjusted for the following covariates selected *a priori* based on their known association with post-lung transplant outcomes: age (per 5-year intervals), male sex, native lung disease (chronic obstructive pulmonary disease, cystic fibrosis or interstitial lung disease), cytomegalovirus recipient–donor serostatus matching and CLAD phenotype (RAS/mixed or unclassified/undefined) [22]. For illustration purposes, Kaplan–Meier curves were generated using the tertiles of each quantitative and semiquantitative variable assessed. A p-value <0.05 was considered statistically significant. Analysis was conducted in R version 4.0 (R Foundation for Statistical Computing, Vienna, Austria).

Results

Patient cohort and CT acquisition

There were 211 patients with a diagnosis of CLAD, of whom 95 had an eligible CT (figure 2). Machine learning segmented the lung in 93% of cases. The final CLAD cohort therefore consisted of 88 patients: 57 BOS, 12 RAS, eight mixed, eight undefined and three unclassified. The only significant difference across CLAD phenotypes in clinical characteristics was a longer time from CLAD diagnosis to graft failure in patients with BOS (p=0.012) (table 1). Analysis of pre-CLAD immunological variables (infection score, rejection score and *de novo* DSA) as well DSA status at CLAD onset and virtual crossmatch status at transplant revealed no significant differences across CLAD phenotypes (supplementary table E1).

CT was performed on the day of CLAD onset in 15 patients, in the first 100 days after CLAD onset in 64 patients and in the 28 days preceding CLAD onset in nine patients, with an overall median interval of 9.5 days after CLAD diagnosis. CT_{TLC} was strongly correlated with TLC (r=0.897, p<0.001) (supplementary figure E1). Technical parameters of the CT examinations are provided in supplementary table E2.

Machine learning analysis

The proportion of HL_{ML} was highest in the BOS group (p<0.001). GGO_{ML}, RET_{ML} and PVV_{ML} were highest in the RAS/mixed group (p<0.001) (table 2). A visual representation of all machine learning output in the form of glyphs is presented in supplementary figures E2 and E3. No significant volume of honeycombing was identified (median (IQR) 0.022% (0.009–0.06%)) and it was therefore excluded from analyses.

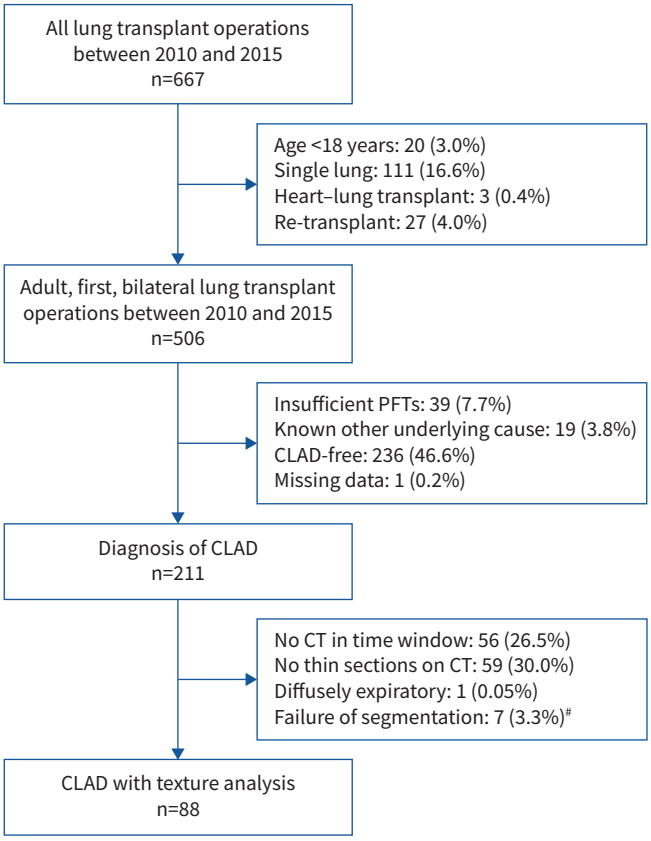

FIGURE 2 Flowchart of patient inclusion criteria. #: failure of segmentation related to classifying trachea (four cases), chest wall (two cases) or abdominal fat (one case) as lung. PFT: pulmonary function test; CLAD: chronic lung allograft dysfunction; CT: computed tomography.

ROC analysis demonstrated that GGO_{ML}, RET_{ML} and PVV_{ML} had strong diagnostic capability in identifying the RAS/mixed phenotype (area under the curve (AUC) 0.84–0.85), and this was strongest with PVV_{ML}. Using a cut-point of 3%, PVV_{ML} had a sensitivity and specificity of 90.0% and 70.1%, respectively (OR 2.08; p<0.001) (table 3). The ROC curves are provided in supplementary figure E4. Phenotyping of BOS with HL_{ML} was achieved using inspiratory CT (AUC 0.76) and did not utilise the expiratory CT to identify air trapping (table 3).

Radiologist scoring

Interrater reliability was good for GGO_{RAD}, RET_{RAD} and CON_{RAD} (ICC 0.89, 0.81 and 0.84, respectively). There was moderate interrater reliability for pleural effusion (weighted κ=0.60) and poor interrater reliability for bronchiectasis (weighted κ=0.42), likely related to the infrequency of these findings. All scored features were highest in the RAS/mixed group (p<0.001) except bronchiectasis (table 2). ROC analysis demonstrated strong diagnostic capability for GGO_{RAD}, RET_{RAD} and CON_{RAD} in identifying the RAS/mixed phenotype (table 3).

A subset of 50 patients had expiratory imaging available (34 BOS, 11 RAS/mixed and five undefined). Interrater reliability was moderate for AT_{RAD} (ICC 0.52) and among the lowest of the radiological findings. AT_{RAD} was significantly higher in the BOS group (median (IQR) 6 (3.5–8)) than in RAS/mixed (median (IQR) 2 (1–5)) and undefined (median (IQR) 2 (0.75–3); p=0.003), but was only weakly correlated with HL_{ML} (r=0.189, p=0.188).

Allograft survival analysis

Univariable analysis of the baseline clinical variables showed that age (HR 0.89; p=0.003) and RAS/mixed phenotype (HR 2.24; p=0.008) were associated with time to graft failure after CLAD diagnosis (supplementary table E3). BOS phenotype was not associated with graft failure (HR 0.83; p=0.674) and

TABLE 1 Patient characteristics by chronic lung allograft dysfunction (CLAD) phenotype

	BOS (n=57)	RAS/mixed (n=12/8)	Undefined/unclassified (n=8/3)	p-value
Age at transplant (years)	47.81±15.41	45.45±17.39	48.09±17.02	0.84
Sex				0.526
Female	31 (54.4)	8 (40.0)	6 (54.5)	
Male	26 (45.6)	12 (60.0)	5 (45.5)	
Native lung disease				0.45
ILD	20 (35.1)	10 (50.0)	1 (9.1)	
Cystic fibrosis	16 (28.1)	4 (20.0)	4 (36.4)	
COPD	14 (24.6)	3 (15.0)	4 (36.4)	
Other	7 (12.3)	3 (15.0)	2 (18.2)	
Time to CLAD onset (days)	700 (367–1171)	624 (425–1125)	804 (445–1370)	0.765
Graft failure				0.803
Death	27 (77.1)	13 (76.5)	6 (66.7)	
Re-transplant	8 (22.9)	4 (23.5)	3 (33.3)	
CMV serostatus				0.638
D−R−	8 (14.0)	1 (5.0)	1 (9.1)	
D+R−	15 (26.3)	4 (20.0)	4 (36.4)	
R+	34 (59.6)	15 (75.0)	6 (54.5)	
Time from CLAD onset to graft failure (days)	746 (445–1135)	370 (250–574)	520 (429–694)	0.012*

Data are presented as mean±SD, n (%) or median (interquartile range), unless otherwise stated. BOS: bronchiolitis obliterans syndrome; RAS: restrictive allograft syndrome; ILD: interstitial lung disease; COPD: chronic obstructive pulmonary disease; CMV: cytomegalovirus; D: donor; R: recipient. *: p<0.05.

was not included in the multivariable analysis. Univariable analysis of the radiologist and machine learning scores showed an association with graft failure for all variables assessed except NL_{ML} and HL_{ML} (table 4).

In multivariable analyses, radiologist scoring of GGO_{RAD}, RET_{RAD} and CON_{RAD} was independently associated with graft failure (table 4). Machine learning scoring of RET_{ML} and PVV_{ML} was also independently associated with graft failure, and GGO_{ML} approached statistical significance (HR 1.36, 95% CI 1.00–1.86; p=0.050) (table 4). NL_{ML} and HL_{ML} were not associated with graft failure.

Although the proportions of NL_{ML} and HL_{ML} were not associated with graft failure, in their multivariable analyses there was an independent association of RAS/mixed phenotype with graft failure (HR 2.73, 95% CI 1.42–5.28; p=0.00 and HR 2.43, 95% CI 1.21–4.85; p=0.01, respectively) and it was only in these two

TABLE 2 Radiologist and machine learning analysis of chest computed tomography scans tabulated by chronic lung allograft dysfunction phenotypes

	BOS (n=57)	RAS/mixed (n=12/8)	Undefined/unclassified (n=8/3)	p-value
Radiologist (score)				
Consolidation	0.50 (0.00–1.00)	3.00 (1.50–5.00)	0.50 (0.00–1.25)	<0.001*
Ground-glass opacity	4.00 (2.00–6.50)	9.00 (7.38–10.50)	4.00 (1.50–7.25)	<0.001*
Reticulation	2.50 (0.50–6.00)	8.00 (6.50–9.62)	3.00 (0.75–6.50)	<0.001*
Bronchiectasis	0.00 (0.00–0.50)	0.00 (0.00–0.62)	0.00 (0.00–0.00)	0.298
Pleural effusion	0.00 (0.00–0.00)	1.00 (0.50–1.88)	0.00 (0.00–0.50)	<0.001*
Machine learning (%)				
Normal	88.74 (77.14–94.51)	89.50 (81.59–93.24)	94.45 (79.59–96.21)	0.534
Hyperlucent	5.75 (1.00–19.88)	0.12 (0.05–1.51)	0.68 (0.23–5.81)	<0.001*
Ground-glass opacity	0.45 (0.17–1.70)	6.43 (1.67–13.08)	0.85 (0.22–2.67)	<0.001*
Reticular	0.53 (0.23–1.46)	2.35 (1.43–3.63)	1.15 (0.22–2.01)	<0.001*
PVV	2.27 (1.95–3.14)	4.06 (3.28–5.25)	2.29 (1.92–3.31)	<0.001*

Data are presented as median (interquartile range), unless otherwise stated. BOS: bronchiolitis obliterans syndrome; RAS: restrictive allograft syndrome; PVV: pulmonary vessel volume. *: p<0.05.

TABLE 3 Univariable logistic regression and receiver operating (ROC) curve analysis for the proportion of machine learning and radiologist variables in bronchiolitis obliterans syndrome phenotype for hyperlucent lung (HL_{ML}) and restrictive allograft dysfunction/mixed phenotype for all other variables

Texture	ROC curve analysis					Univariable logistic regression	
	Optimal cut-point	Sensitivity	Specificity	Accuracy	AUC	Odds ratio	p-value
Machine learning							
HL_{ML}	0.81%	0.79	0.71	0.76	0.763	1.11	0.006*
GGO_{ML}	0.79%	1.00	0.56	0.66	0.845	1.12	0.004*
RET_{ML}	0.93%	1.00	0.65	0.73	0.835	1.59	0.003*
PVV_{ML}	3.02%	0.90	0.71	0.75	0.851	2.08	<0.001*
Radiologist							
GGO_{RAD}	7.00	0.80	0.79	0.80	0.829	1.44	0.001*
RET_{RAD}	6.00	0.90	0.74	0.77	0.838	1.42	0.001*
CON_{RAD}	1.50	0.80	0.76	0.77	0.820	1.88	0.001*

AUC: area under the curve; GGO: ground-glass opacity; RET: reticulation; PVV: pulmonary vessel volume; CON: consolidation. *: p<0.05.

multivariable analyses that any independent association of CLAD phenotype with graft failure was identified. Details of the machine learning and radiologist multivariable analyses are provided in supplementary tables E4–E6.

The proportion of each radiological finding was stratified by tertiles and Kaplan–Meier curves were generated to facilitate visualisation of lung allograft survival after CLAD diagnosis (figure 3). The greatest separation of all tertiles was seen with RET_{RAD}, with 50% allograft survival at 439 days in the tertile with the most reticulation compared with 1134 and 1700 days in the middle and lowest tertile, respectively (p=0.002).

Because of the heterogeneity seen in supplementary figures E2 and E3 where some cases of BOS were observed to have ground-glass opacity and reticulation, whereas other cases in the undefined/unclassified grouping were observed to contain hyperlucent lung, we empirically regrouped cases into three basic imaging patterns: Group 1: inflammatory presentation (GGO_{ML} and RET_{ML} >1% of CT_{TLC}); Group 2: hyperlucent presentation (HL_{ML} >10% of CT_{TLC}); and Group 3: indeterminate presentation (fulfilling neither Group 1 nor Group 2 criteria). The revised glyph grouping is presented in supplementary figure E5 and Kaplan–Meier curves are presented in supplementary figure E6. Using this classification, the

TABLE 4 Univariable and multivariable analysis of automated and semiquantitative scoring for the highest tertile of each radiological abnormality for the diagnosis of graft failure (death or re-transplantation)

	Univariable analysis		Multivariable analysis	
	Hazard ratio (95% CI)	p-value	Hazard ratio (95% CI)	p-value
Machine learning				
NL_{ML}	0.93 (0.78–1.10)	0.37	0.96 (0.80–1.15)	0.65
HL_{ML}	0.92 (0.76–1.12)	0.40	0.90 (0.72–1.13)	0.37
GGO_{ML}	1.45 (1.12–1.88)	<0.001*	1.36 (1.00–1.86)	0.05
RET_{ML}	1.25 (1.11–1.39)	<0.001*	1.20 (1.05–1.37)	0.01*
PVV_{ML}	1.30 (1.14–1.48)	<0.001*	1.23 (1.05–1.44)	0.01*
Radiologist				
GGO_{RAD}	1.16 (1.08–1.25)	<0.001*	1.14 (1.05–1.23)	<0.001*
RET_{RAD}	1.18 (1.10–1.26)	<0.001*	1.17 (1.08–1.27)	<0.001*
CON_{RAD}	1.26 (1.12–1.41)	<0.001*	1.16 (1.01–1.34)	0.04*

NL: normal lung; HL: hyperlucent lung; GGO: ground-glass opacity; RET: reticulation. PVV: pulmonary vessel volume; CON: consolidation. The multivariable model was performed for each radiological variable separately, and is adjusted for sex, chronic lung allograft dysfunction phenotype, age, native lung disease and cytomegalovirus serostatus matching. *: p<0.05.

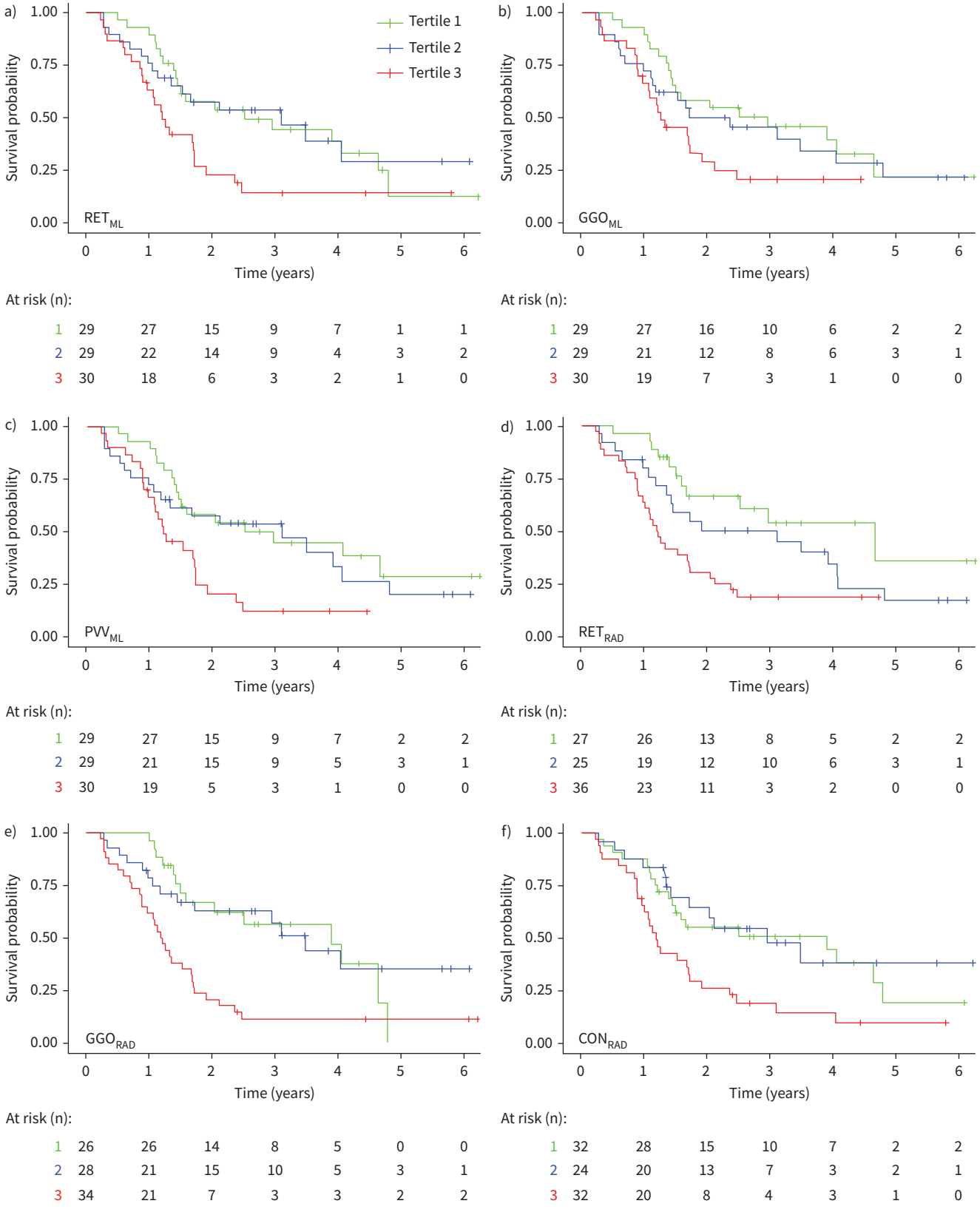

FIGURE 3 Kaplan–Meier curves demonstrating allograft survival probability over time for a) RET_{ML}, b) GGO_{ML}, c) PVV_{ML}, d) RET_{RAD}, e) GGO_{RAD} and f) CON_{RAD}. Tertiles of radiological abnormality are indicated (1=highest amount). RET: reticulation; GGO: ground-glass opacity; PVV: pulmonary vessel volume; CON: consolidation.

inflammatory group would be comprised of 16 RAS/mixed, five undefined/unclassified and 15 BOS. The indeterminate group would be 20 BOS, four RAS/mixed and four undefined/unclassified. The hyperlucent group would be 22 BOS and two undefined/unclassified.

Discussion

We found an excellent diagnostic performance of the machine learning tool in identifying the RAS and mixed CLAD phenotype. Furthermore, machine learning identified hyperlucent lung in patients with BOS, without necessitating expiratory CT. Both machine learning and radiologist scoring at the time of CLAD onset were of prognostic importance, independent of CLAD phenotype in our multivariable model, emphasising the importance of CT early in the evaluation of CLAD. PVV_{ML}, a biomarker unique to machine learning, had the strongest diagnostic and prognostic performance among all variables assessed.

We investigated PVV because it was found to be a strong predictor of mortality in diffuse lung disease [20, 23]. As a marker of total vessel volume in the lung, possible explanations for the significance of PVV in pulmonary fibrosis include redistribution of blood flow from abnormal lung to normal lung, the presence of arterial shunts or a dilatation of blood vessels related to increased negative pressure at inspiration in the context of lung stiffness [20]. It is also possible that perivascular inflammatory opacity is being classified as vessel volume by machine learning. Although PVV increases as CT_{TLC} decreases, a reduction in lung volumes alone does not fully account for the significance of PVV as we controlled for CT_{TLC} in our analyses. PVV was expressed as a proportion of CT_{TLC} and we also added CT_{TLC} into the PVV multivariable model.

In a previous study, histopathological examination of explanted lungs with end-stage restrictive CLAD revealed focal areas of capillary obliteration with a resultant decreased microvascular density, a finding that would suggest a lower PVV in advanced CLAD [24]. However, the same authors also describe areas of reactive microvascular proliferation among fibrosis and ectasia of small lymphatic vessels that, in our estimation, might account for an increased PVV. In the absence of a definite histological correlate at the onset of CLAD, the prognostic significance of increased PVV in CLAD could relate to increased negative intrathoracic pressure at end-inspiration, particularly in those patients with restriction and lung stiffness, and possibly the inclusion of perivascular inflammation in vessel segmentation, given that the analysis is not capable of reliably distinguishing between a true vessel *versus* a perivascular opacity. However, a further quantitative analysis of the small vessels in CLAD, both at onset and at end-stage, is warranted given the intriguing findings of microvascular damage in lungs with restrictive CLAD or RAS [24].

Machine learning-identified hyperlucent lung is a feature associated with BOS on fully inspiratory exams and this likely corresponds to air trapping in this cohort. Conventionally, radiologists assess for air trapping using paired inspiratory and expiratory imaging; using this approach to diagnose BOS, Bankier *et al.* [25] describe a threshold of 32% air trapping as being 83% sensitive, 89% specific and 88% accurate. Indeed, radiologist-assessed air trapping may be insensitive particularly if there is poor timing of the CT scan and limited expiratory effort (supplementary figure E7). More recently, Verleden *et al.* [26] used parametric response mapping alongside a matched stable cohort, and found a sensitivity of 62.5% and specificity of 93.8%. With machine learning, and among a cohort of only CLAD patients, we found a cut-off of just 0.81% hyperlucent lung to be 79% sensitive, 71% specific and 76% accurate. The lower specificity in our study is likely due to including patients with all forms of CLAD. Interestingly, patients with CLAD of a mixed and RAS phenotype may demonstrate lesions of obliterative bronchiolitis and indeed we found hyperlucent lung in all CLAD groups [27].

In the multivariable analysis, we found that NL_{ML} and HL_{ML} were not independently associated with graft failure, whereas CLAD phenotype defined clinically was. Despite this limitation of HL_{ML} as a marker of CLAD prognosis, the ability of machine learning to identify this important feature of CLAD on inspiratory CT warrants further investigation. Foregoing expiratory CT would offer a radiation dose reduction and improved efficiency in performing and reporting the CT studies.

Although the machine learning analysis presents unique insights in PVV and HL_{ML} quantification, the radiologist scoring also performed well in our study. The semiquantitative radiologist scoring system in this study was used in a similarly sized lung transplant cohort by Suhling *et al.* [6] in 2016 who found that, by combining CT findings of consolidation and reticulation with TLC, they could identify a subgroup of patients with restrictive CLAD and poorer survival. Their cohort differed substantially from ours in that one-third had severe traction bronchiectasis, a marker of advanced fibrosis, and CT was performed a median of 387 days after CLAD onset. In contrast, CT was performed a median of 9.5 days after CLAD onset in the current study and only one case had severe bronchiectasis. Furthermore, Suhling *et al.* [6]

found that ground-glass opacity was not associated with graft failure, whereas we found that it was, and this may also relate to CT timing if ground-glass opacity is attributable to acute inflammation near the onset of CLAD. Indeed, PHILIPPOT *et al.* [28] also applied the same scoring system to single lung transplant patients and found it useful in the early detection of RAS. Neither study controlled for CLAD phenotype in their multivariable analysis and our finding of opacities as independent predictors of graft failure suggests that a quantitative assessment of the lung opacities, achieved through machine learning, is of value.

Although we demonstrate a strong ability of machine learning to phenotype CLAD and its determination of graft survival, there are several other potential uses. The quantitative data provided by machine learning analysis may be of use in future clinical drug trials and treatment monitoring where there is a need for reproducible biomarkers of disease. With validation, the use of an automated machine learning tool in lung transplant studies would allow improved comparison of cohorts across centres. It should also be noted that machine learning is sensitive to fractional volumes of lung abnormality, the optimal cut-off for HL_{ML} in the current study being 0.81%. This finding raises the question of the ability of machine learning to detect lung texture abnormalities in CLAD at its earliest stage, before becoming clinically evident.

Limitations of this study include the single-centre and retrospective design, although a CALIPER-based machine learning analysis is compatible with a wide range of CT protocols. Furthermore, we were restricted in our ability to include a larger number of patients due to the absence of archived CT thin slices prior to 2013. Despite this, our cohort is among the largest with a comprehensive CT analysis in CLAD. Finally, this machine learning tool was not trained on lung transplant cases and does not capture all features of particular interest in lung transplant patients, such as pleural thickening or complications at the bronchial anastomoses. A future machine learning tool trained on and tailored for lung transplant patients might be of even greater value. Of course, the clinician and radiologist remain critically important in evaluating the CT and considering the findings in the appropriate clinical context.

In conclusion, machine learning strongly discriminated between CLAD phenotypes using automated analysis. Both radiologist and machine learning scoring were associated with graft failure, independent of the CLAD phenotype and without using expiratory CT. Integration of machine learning into clinical use may facilitate automated CT analysis for phenotyping and prognostication in a reliable and reproducible manner. PVV, a biomarker unique to machine learning, was best in CLAD phenotyping and prognostication, and warrants future investigation in this population.

Acknowledgements: The authors would like to thank Benjamin Renaud-Picard, Gregory Berra, Mitsuaki Kawashima and Akihiro Takahagi, as well as other members of the Toronto Lung Transplant Program (Toronto, ON, Canada), for helping with CLAD and CLAD phenotype adjudication. We are also grateful to Rasheed Ghany for managing the Toronto Lung Transplant Program Database of clinical data.

Author contributions: Conception and design: M.C. McInnis and T. Martinu. Acquisition of data: M.C. McInnis, G.R. Karur, C. Houbois, L. Levy, J. Havlin, E. Fuchs, J. Tikkanen and T. Martinu. Analysis and interpretation of data: M.C. McInnis, T. Martinu, J. Ma and E. Huszti. Drafting the manuscript and intellectual content: M.C. McInnis, J. Tikkanen, C-W. Chow and T. Martinu.

Conflict of interest: C. Houbois reports grants from German Research Foundation (project number 419344766), outside the submitted work. T. Martinu reports grants from the Canadian Institutes of Health Research, Cystic Fibrosis Foundation, Ontario Thoracic Society, National Institutes of Health and Sanofi; receipt of materials for experiments from APCBio; outside the submitted work. All other authors have nothing to disclose.

References

1 Verleden GM, Glanville AR, Lease ED, *et al.* Chronic lung allograft dysfunction: definition, diagnostic criteria, and approaches to treatment – a consensus report from the Pulmonary Council of the ISHLT. *J Heart Lung Transplant* 2019; 38: 493–503.
2 Sato M, Waddell TK, Wagnetz U, *et al.* Restrictive allograft syndrome (RAS): a novel form of chronic lung allograft dysfunction. *J Heart Lung Transplant* 2011; 30: 735–742.
3 Levy L, Huszti E, Renaud-Picard B, *et al.* Risk assessment of chronic lung allograft dysfunction phenotypes: validation and proposed refinement of the 2019 International Society for Heart and Lung Transplantation classification system. *J Heart Lung Transplant* 2020; 39: 761–770.
4 Worthy SA, Park CS, Kim JS, *et al.* Bronchiolitis obliterans after lung transplantation: high-resolution CT findings in 15 patients. *AJR Am J Roentgenol* 1997; 169: 673–677.

5 Hota P, Dass C, Kumaran M, *et al.* High-resolution CT findings of obstructive and restrictive phenotypes of chronic lung allograft dysfunction: more than just bronchiolitis obliterans syndrome. *AJR Am J Roentgenol* 2018; 211: W13–W21.

6 Suhling H, Dettmer S, Greer M, *et al.* Phenotyping chronic lung allograft dysfunction using body plethysmography and computed tomography. *Am J Transplant* 2016; 16: 3163–3170.

7 Dettmer S, Shin HO, Vogel-Claussen J, *et al.* CT at onset of chronic lung allograft dysfunction in lung transplant patients predicts development of the restrictive phenotype and survival. *Eur J Radiol* 2017; 94: 78–84.

8 Dubbeldam A, Barthels C, Coolen J, *et al.* Restrictive allograft syndrome after lung transplantation: new radiological insights. *Eur Radiol* 2017; 27: 2810–2817.

9 Horie M, Levy L, Houbois C, *et al.* Lung density analysis using quantitative chest CT for early prediction of chronic lung allograft dysfunction. *Transplantation* 2019; 103: 2645–2653.

10 Saito T, Horie M, Sato M, *et al.* Low-dose computed tomography volumetry for subtyping chronic lung allograft dysfunction. *J Heart Lung Transplant* 2016; 35: 59–66.

11 Horie M, Salazar P, Saito T, *et al.* Quantitative chest CT for subtyping chronic lung allograft dysfunction and its association with survival. *Clin Transplant* 2018; 32: e13233.

12 Belloli EA, Degtiar I, Wang X, *et al.* Parametric response mapping as an imaging biomarker in lung transplant recipients. *Am J Respir Crit Care Med* 2017; 195: 942–952.

13 Verleden SE, Vos R, Vandermeulen E, *et al.* Parametric response mapping of bronchiolitis obliterans syndrome progression after lung transplantation. *Am J Transplant* 2016; 16: 3262–3269.

14 Nascimento DZ, Watte G, Torres FS, *et al.* Utilization of quantitative computed tomography assessment to identify bronchiolitis obliterans syndrome after single lung transplantation. *Lung* 2021; 199: 29–35.

15 Barbosa EJM Jr, Lanclus M, Vos W, *et al.* Machine learning algorithms utilizing quantitative CT features may predict eventual onset of bronchiolitis obliterans syndrome after lung transplantation. *Acad Radiol* 2018; 25: 1201–1212.

16 Bartholmai BJ, Raghunath S, Karwoski RA, *et al.* Quantitative computed tomography imaging of interstitial lung diseases. *J Thorac Imaging* 2013; 28: 298–307.

17 Jacob J, Bartholmai BJ, Rajagopalan S, *et al.* Automated quantitative computed tomography *versus* visual computed tomography scoring in idiopathic pulmonary fibrosis: validation against pulmonary function. *J Thorac Imaging* 2016; 31: 304–311.

18 Maldonado F, Moua T, Rajagopalan S, *et al.* Automated quantification of radiological patterns predicts survival in idiopathic pulmonary fibrosis. *Eur Respir J* 2014; 43: 204–212.

19 Lee SM, Seo JB, Oh SY, *et al.* Prediction of survival by texture-based automated quantitative assessment of regional disease patterns on CT in idiopathic pulmonary fibrosis. *Eur Radiol* 2018; 28: 1293–1300.

20 Jacob J, Bartholmai BJ, Rajagopalan S, *et al.* Mortality prediction in idiopathic pulmonary fibrosis: evaluation of computer-based CT analysis with conventional severity measures. *Eur Respir J* 2017; 49: 1601011.

21 Hansell DM, Bankier AA, MacMahon H, *et al.* Fleischner Society: glossary of terms for thoracic imaging. *Radiology* 2008; 246: 697–722.

22 Chambers DC, Cherikh WS, Harhay MO, *et al.* The International Thoracic Organ Transplant Registry of the International Society for Heart and Lung Transplantation: thirty-sixth adult lung and heart-lung transplantation report – 2019; focus theme: donor and recipient size match. *J Heart Lung Transplant* 2019; 38: 1042–1055.

23 Jacob J, Hirani N, van Moorsel CHM, *et al.* Predicting outcomes in rheumatoid arthritis related interstitial lung disease. *Eur Respir J* 2019; 53: 1800869.

24 von der Thusen JH, Vandermeulen E, Vos R, *et al.* The histomorphological spectrum of restrictive chronic lung allograft dysfunction and implications for prognosis. *Mod Pathol* 2018; 31: 780–790.

25 Bankier AA, Van Muylem A, Knoop C, *et al.* Bronchiolitis obliterans syndrome in heart-lung transplant recipients: diagnosis with expiratory CT. *Radiology* 2001; 218: 533–539.

26 Verleden SE, Vos R, Vandermuelen E, *et al.* Parametric response mapping of bronchiolitis obliterans syndrome progression after lung transplantation. *Am J Transplant* 2016; 16: 3262–3269.

27 Ofek E, Sato M, Saito T, *et al.* Restrictive allograft syndrome post lung transplantation is characterized by pleuroparenchymal fibroelastosis. *Mod Pathol* 2013; 26: 350–356.

28 Philippot Q, Debray MP, Bun R, *et al.* Use of CT-SCAN score and volume measures to early identify restrictive allograft syndrome in single lung transplant recipients. *J Heart Lung Transplant* 2020; 39: 125–133.