AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

Disease Progression Modelling in Chronic Obstructive Pulmonary Disease (COPD)

Alexandra L. Young*[,1,2,3], Felix J.S. Bragman*[,1,4], Bojidar Rangelov[1], MeiLan Han[5+], Craig J.Galbán[6], David A. Lynch[7], David J. Hawkes[1], Daniel C. Alexander[1,2] and John R. Hurst[~,8]; for the COPDGene Investigators

1. Centre for Medical Image Computing, Department of Medical Physics and Biomedical Engineering, University College London, London, United Kingdom

2. Department of Computer Science, University College London, London, United Kingdom

3. Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

4. Artificial Medical Intelligence Group, School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom

5. University of Michigan, Pulmonary & Critical Care, Ann Arbor, Michigan, United States

6. Center for Molecular Imaging, Michigan, Michigan, United States

7. Department of Radiology, National Jewish Health, Denver, Colorado, United States

8. UCL Respiratory, University College London, London, United Kingdom

* Joint first authors.

~Corresponding author: j.hurst@ucl.ac.uk

+ Associate Editor, AJRCCM (participation complies with American Thoracic Society requirements for recusal from review and decisions for authored works).

Author Contributions: All authors meet criteria for authorship as recommended by the International Committee of Medical Journal Editors. AY, FB, DH, DA and JH designed the study. AY and FB performed the modelling and statistical analysis and wrote the initial manuscript. COPDGene Investigators including DL, MH and CG assisted with collection and analysis of COPDGene data. All

authors contributed to the production of the final manuscript with revision for important intellectual content.

Support: FB was supported by the EPSRC under Grant EP/H046410/1 and EP/K502959/1. FB and DH were supported under a UCLH NIHR RCF Senior Investigator Award under Grant RCF107/DH/2014. AY is supported by an EPSRC Doctoral Prize Fellowship. BR is supported by the EPSRC Centre For Doctoral Training in Medical Imaging with grant EP/L016478/1 and by an industrial CASE studentship with funding from GlaxoSmithKline Research and Development, agreement number BIDS3000032413. DA was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 666992 and EPSRC grants M020533, M006093, J020990. This work was supported by the NIHR UCLH Biomedical Research Centre.

The COPDGene Study was supported by Award Number U01 HL089897 and Award Number U01 HL089856 from the National Heart, Lung, and Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health. The COPDGene project is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Novartis, Pfizer, Siemens and Sunovion.

Running-head: *Disease Progression Modelling in COPD*

Subject category number: 9.9 COPD: General

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

At a Glance Commentary:

Scientific Knowledge on the Subject: *COPD progresses over decades so little is known about longitudinal changes in individual patients, and whether there are different patterns of disease progression in different patient subgroups.*

What this Study Adds to the Field: *Computational modelling of CT biomarkers suggests there are two patterns of disease progression in COPD. These disease progression patterns or 'subtypes' can be used to stratify individuals into two groups with distinct clinical characteristics, and to stage individuals along their disease time-course. Early stages of both subtypes are identifiable in a proportion of 'healthy smokers' providing a biomarker of early COPD.*

This article has an online data supplement, which is accessible from the issue's table of contents online at www.atsjournals.org

Abstract

Rationale: The decades-long progression of Chronic Obstructive Pulmonary Disease (COPD) renders identifying different trajectories of disease progression challenging.

Objectives: To identify subtypes of COPD patients with distinct longitudinal progression patterns using a novel machine-learning tool called "Subtype and Stage Inference (SuStaIn)", and to evaluate the utility of SuStaIn for patient stratification in COPD.

Methods: We applied SuStaIn to cross-sectional CT imaging markers in 3698 GOLD1-4 patients and 3479 controls from the COPDGene study to identify COPD patient subtypes. We confirmed the identified subtypes and progression patterns using ECLIPSE data. We assessed the utility of SuStaIn for patient stratification by comparing SuStaIn subtypes and stages at baseline with longitudinal follow-up data.

Measurements and Main Results: We identified two trajectories of disease progression in COPD: a "Tissue→Airway" subtype (n=2354, 70.4%) in which small airway dysfunction and emphysema precede large-airway wall abnormalities, and an "Airway→Tissue" subtype (n=988, 29.6%) in which large-airway wall abnormalities precede emphysema and small airway dysfunction. Subtypes were reproducible in ECLIPSE. Baseline stage in both subtypes correlated with future FEV_1/FVC decline (r=-0.16 (p<0.001) in the Tissue→Airway group; r=-0.14 (p=0.011) in the Airway→Tissue group). SuStaIn placed 30% of smokers with normal lung function at non-baseline stages suggesting imaging changes consistent with early COPD. Individuals with early changes were 2.5 times more likely to meet COPD diagnostic criteria at follow-up.

Conclusions: We demonstrate two distinct patterns of disease progression in COPD using SuStaIn, likely representing different endotypes. One-third of healthy smokers have detectable imaging changes, suggesting a new biomarker of 'early COPD'.

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

Keywords: *Clustering; Disease staging; CT imaging; Emphysema; Bronchitis; Pulmonary Disease, Chronic Obstructive*

Introduction

Chronic Obstructive Pulmonary Disease (COPD) can be characterised as the consequence of a genetically susceptible individual being exposed to sufficient environmental exposures (1). The pulmonary components are heterogeneous (2) and include emphysema, small airway loss and obstruction, and larger airway inflammation. COPD progresses over decades and often remains subclinical until the later development of symptoms or exacerbations. Slow progression and heterogeneous manifestations make it challenging to construct long-term models of disease progression, as most studies collect only cross-sectional or short-term longitudinal data.

Incomplete understanding of disease progression and heterogeneity in COPD has consequences for clinical practice and drug development. First, we are currently unable to identify early stages of disease in 'healthy smokers', preventing interventions in 'early COPD' where disease-modifying treatments may be most effective. Second, clinically relevant populations with severe airflow obstruction may have arrived at this point through different early mechanisms ('endotypes'), which may therefore have been amenable to different interventions (2).

Quantitative imaging of the lung through Computed Tomography offers the opportunity to better evaluate the complex relationship between structure and function in COPD. Specifically, airway wall geometry informs on chronic bronchitis whilst emphysematous tissue destruction and gas trapping due to small airways obstruction and destruction can be quantified using density thresholds. Whilst this facilitates direct disease quantification, understanding the progression and heterogeneity of pathology detected by imaging measures has remained limited (3).

Previous imaging studies attempting to disentangle the heterogeneity of COPD have used clustering techniques (4, 5), probabilistic modelling (6, 7) or dimensionality reduction (8, 9). Clustering does not naturally group individuals on the same trajectory, since patients at early and late stages of a cluster may look very different. Thus, these approaches confound disease subtypes with stage (see

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

Glossary), preventing the identification of specific phenotypes independently of temporal progression. The ability to identify disease subtypes independently of disease stage has been a long-standing unmet need.

Significant progress in the understanding of neurodegenerative diseases has been made using techniques collectively called 'Disease Progression Modelling', which reconstruct the long-term temporal progression of disease from cross-sectional data via unsupervised learning (10–15). Subtype and Stage Inference (SuStaIn) (16) is a recent innovation arising from the study of dementia that integrates clustering and disease progression modelling, offering new ability to disentangle the heterogeneity of disease subtypes from assessment of disease stages. SuStaIn identifies subgroups of individuals (disease subtypes) with distinct progression patterns, while simultaneously reconstructing the trajectory (stage progression) of each subtype. Such data-driven progression models have not previously been applied in the field of respiratory medicine, and offer a major opportunity to explain disease heterogeneity, and enhance precision medicine in conditions of long natural history such as COPD.

Some of the results of this study have been previously reported in the form of an abstract (17).

Method

This is an abbreviated version of the Method, please see the Online Supplement for further detail about each analysis step.

Definitions and Overview

Key terminology is defined in the Glossary.

Model development used CT data from COPDGene Phase 1 (18), comprising a cross-sectional dataset of baseline measurements from 3479 smoking controls and 3698 COPD patients. We repeated the SuStaIn algorithm using baseline data from 303 smoking controls and 1809 COPD

patients in the ECLIPSE study (19) to verify consistency of SuStaIn output in an independent data set. We evaluated longitudinal progression using follow-up (Phase 2) COPDGene scans and data to verify the SuStaIn subtype progression patterns (reconstructed from cross-sectional data) against true longitudinal progression of individual subjects. This included 1929 COPD subjects and 2158 controls who had all imaging biomarkers available from the initial scan together with measures of lung function from both time points, and a second dataset of 1675 COPD subjects and 1939 controls who had all imaging biomarkers available at both phases of the COPDGene study.

GLOSSARY:
SUBTYPE – a group of subjects who share a particular trajectory of biomarker evolution.
STAGE – the position on a subtype trajectory of an individual subject at a specific time. In SuStaIn this represents the degree of abnormality in imaging biomarkers and a change in stage occurs when an imaging biomarker becomes more abnormal relative to a control population.
DISEASE PROGRESSION – change in stage with time as the natural history of the condition unfolds. We use the term in two distinct contexts: 1. GROUP (SUBTYPE) LEVEL: referring to the sequence of changes that the typical patient undergoes from start to finish. 2. INDIVIDUAL LEVEL: change in stage or severity of an individual subject as biomarkers become increasingly abnormal.

Imaging features

A set of four imaging features were derived in COPDGene: 1) emphysema, obtained using parametric response mapping (PRM) (20), 2) functional small airways disease (fSAD) obtained from PRM, 3) Pi10 square root wall area (SRWA) (21) and 4) segmental airway wall thickness. CT analysis to obtain the imaging features was performed using Thirona lung quantification software (Thirona, Netherlands, http://www.thirona.eu) (18). There were only two imaging features available in the ECLIPSE study: emphysema and Pi10 SRWA, obtained using VIDA software (22).

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

Disease progression modelling

Given a cross-sectional data set, SuStaIn simultaneously identifies a set of disease subtypes, each defined by a distinct trajectory of biomarker evolution with a probabilistic assignment of each subject to a subtype and stage along the corresponding trajectory. The trajectory of each subtype is described as a linear z-score model (15), consisting of a series of stages in which each stage corresponds to a biomarker reaching a particular z-score relative to a control group. The optimal number of subtypes is determined using information criterion (a statistical technique that balances model complexity with model accuracy). This provides a population-level disease progression model which can be used to assign individuals to subtypes and stages probabilistically. A conceptual overview is provided as Supplementary Figure 1 (16).

Identification of COPD subtypes

We applied the SuStaIn algorithm (16) to COPD GOLD1-4 patients from the COPDGene dataset. As SuStaIn requires monotonic measurements (biomarkers that change over time in one direction only, see Discussion), we replaced fSAD, which may convert to emphysema at later stages of COPD (20), with a combined measure we term 'overall tissue damage'. This was computed as the sum of fSAD and emphysema (and thus is similar to a measure of air trapping). As SuStaIn requires input features expressed as z-scores relative to a control population, we transformed each dataset into z-scores relative to the smoking controls in COPDGene. Prior to performing the z-score transformation, imaging measures were log transformed to improve normality.

Independent evaluation of COPD subtypes

To evaluate the subtypes in an independent dataset we repeated our analysis in COPD GOLD1-4 patients from ECLIPSE using the subset of CT metrics available from inspiratory scans, and the corresponding ECLIPSE smoking controls to perform z-score transformation. As ECLIPSE only has inspiratory scans we re-fitted the SuStaIn algorithm to a COPDGene cross-sectional dataset consisting of baseline measurements from 4102 smoking controls and 4152 COPD patients with

inspiratory measurements available for emphysema and Pi10 SRWA. We refer to these data as the 'Inspiratory COPDGene' dataset.

Subtyping and staging

We used the SuStaIn model (i.e. the subtype progression patterns identified using the SuStaIn algorithm) to automatically assign individuals to their most probable subtype and stage. We did this for all COPDGene COPD patients and control subjects at each of the two visits. We repeated the same process of assigning individuals to SuStaIn subtypes and stages in the ECLIPSE and Inspiratory COPDGene datasets. We further assigned individuals from COPDGene Phase 2 to subtypes and stages using the same procedure described above, identifying the subtypes and stages from the subtype progression patterns estimated using the COPDGene Phase 1 dataset.

Statistical analysis

Clinical characteristics of the subtypes

We compared the clinical characteristics of individuals assigned to each subtype using two sample t-tests for continuous variables, chi-squared tests for categorical variables, and Mann-Whitney U-tests for frequency data.

Relationship between SuStaIn stage and lung function

We verified that SuStaIn stage could be used as a measure of disease severity in COPD by examining whether SuStaIn stage correlated with spirometric impairment as assessed by FEV_1/FVC and FEV_1%predicted. We further evaluated whether a higher SuStaIn stage could be used as an indicator of future lung function decline (disease progression at an individual level) by assessing whether baseline SuStaIn stage was correlated with change in lung function between baseline and follow-up.

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

Longitudinal consistency of subtype and stage

Over time we would expect that subtype remains consistent but that stage will progress. We assessed whether the SuStaIn subtypes remained consistent at five-year follow-up, quantifying consistency as the percentage of individuals in which the subtype assignment remained the same. We assessed whether individuals progressed in SuStaIn stage between baseline and follow-up by comparing the distribution of SuStaIn stages at baseline and follow-up in GOLD1-2 and GOLD3-4 patients using two sample t-tests.

Analysis of smoking controls

We repeated the above analyses in the COPDGene smoking control group to test whether SuStaIn subtype and stage might be useful for identification of otherwise healthy individuals at risk of developing COPD.

Results

Subject Characteristics

The baseline data of the COPDGene study participants used to develop the model are reported in Table 1. The control population (n=3479) was used to derive the z-scores, whilst the GOLD1-4 patients (n=3698) were used to produce the subtypes.

1. Cross-Sectional Analyses in COPD

COPD Subtypes

SuStaIn identified two distinct COPD progression patterns or 'subtypes' (Figure 1). We have termed these "Tissue→Airway" and "Airway→Tissue". In the Tissue→Airway group (n=2354, 70.4%), functional small airways disease and emphysema are the earliest disease stages. Only subsequent to this do pathological alterations in larger airways become apparent. In the Airway→Tissue subgroup (n=988, 29.6%), the earliest stages comprise abnormalities in larger airways, followed by functional

small airways disease and emphysema. These subtypes were reproducible in the ECLIPSE study (Supplementary Figure 2 and Supplementary Results).

Clinical characteristics of the COPD Subtypes

We next investigated differences in the clinical characteristics of patients between the two subtypes (Table 2). There was a smaller proportion of men in the Tissue→Airway compared to the Airway→Tissue subtype (52.3% versus 66.5%, p<0.001). Patients in the Tissue→Airway group had a significantly lower BMI than those in the Airway→Tissue group (26.65 versus 30.54 kgm^{-2}, p<0.001) and a lower prevalence of chronic bronchitis (25.1% versus 31.8%, p<0.001). Detailed relationships between subtype, stage, breathlessness and exacerbations are reported in Supplementary Table 3. Patients in the Tissue→Airway group had marginally more severe spirometric impairment (FEV_1 % predicted 53.63% versus 58.64%, p<0.001; FEV_1/FVC ratio 0.49 versus 0.56, p<0.001). The clinical characteristics of the SuStaIn subtypes were broadly replicable in the ECLIPSE dataset (Supplementary Tables 4 and 5 and Supplementary Results).

Relationship of COPD subtype stage with baseline lung function

We investigated whether SuStaIn stage could be used as a measure of disease severity in COPD by examining correlations between SuStaIn stage and baseline spirometry. We found a significant correlation between SuStaIn stage and FEV_1/FVC (Figure 2A) and FEV_1 % predicted (Supplementary Figure 3A). The relationship was stronger in the Tissue→Airway group: SuStaIn stage correlation with FEV_1/FVC and FEV_1 %predicted r= -0.63 (p<0.001) and r=-0.66 (p<0.001) respectively. In the Airway→Tissue group, the correlation coefficients were -0.58 (p<0.001) for FEV_1/FVC and -0.51 (p<0.001) for FEV_1 %predicted. The relationship between SuStaIn stage and baseline lung function was non-linear in the Tissue→Airway subtype and linear in the Airway→Tissue subtype (Supplementary Results). The correlations between baseline lung function and SuStaIn stage were replicable in the ECLIPSE dataset (Supplementary Figures 4 and 5 and Supplementary Results).

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

2. Longitudinal Analyses in COPD

Relationship of SuStaIn stage with longitudinal decline in lung function

We tested whether baseline SuStaIn stage correlated with future decline in lung function in the subset of individuals with spirometry available at both time points (patient characteristics reported in Supplementary Table 6). Earlier SuStaIn stages were associated with more rapid future, measured individual level progression of FEV_1/FVC ratio and FEV_1%predicted. Considering the annualised change in spirometry after five-years follow-up in GOLD1-2 patients (Figure 2B for FEV_1/FVC ratio and Supplementary Figure 3B for FEV_1 %predicted), we found that baseline SuStaIn stage correlated with rate of decline in FEV_1/FVC and FEV_1 %predicted in both subtypes: r=-0.16 (p<0.001) and r=-0.14 (p=0.011) for baseline SuStaIn stage and change in FEV_1/FVC in the Tissue→Airway and Airway→Tissue groups respectively; and r=-0.20 (p<0.001) and r=-0.14 (p=0.011) between baseline SuStaIn stage and change in FEV_1 %predicted. In GOLD3-4 patients assigned to the Tissue→Airway subtype there was no significant correlation between baseline SuStaIn stage and change in FEV_1/FVC (r=-0.001, p=0.98) or FEV_1 %predicted (r=-0.019, p=0.69). In GOLD3-4 patients assigned to the Airway→Tissue subtype there was a significant correlation between baseline SuStaIn stage and change in FEV_1/FVC (r=-0.23, p=0.005), but this was not reflected in the FEV_1 %predicted measure (r=-0.15, p=0.069).

Stability of COPD Subtype, and Progression of COPD Stage over time

SuStaIn assumes that individuals belong to a single disease subtype, progressing only in stage with time. We verified that the SuStaIn subtypes remained the same at five-year follow-up using a longitudinal validation dataset consisting of COPDGene individuals who had all imaging biomarkers available at both phases (Supplementary Results and Supplementary Tables 7-9). The assignment to Tissue→Airway and Airway→Tissue subtypes remained consistent in 1283/1472 (87%) individuals. SuStaIn stages showed a strong correlation between baseline and follow-up, but individuals tended to progress in stage within each subtype (Supplementary Results and Supplementary Figure 6) giving

confidence that the model is a good representation of disease. Individual stage progression was more rapid in GOLD1-2 patients than GOLD3-4 patients (Supplementary Results), supporting the clinically important hypothesis that disease activity is greatest earlier in disease, whilst spirometrically more severe disease may be considered less active.

3. Analyses in Control Smokers without COPD

Early detection of individuals at risk for COPD in the control population

We hypothesised that a subset of the smoking control population would exhibit features of early COPD SuStaIn stages despite spirometry within the normal range. The majority of control patients were staged at SuStaIn stage 0 (n=2457, 71%). By considering control subjects assigned a stage >0, we were able to identify a group of control subjects (29%) with imaging abnormalities. There were 641 control subjects (18% of the control population) in the Tissue→Airway subtype and 381 subjects (11% of the control population) in the Airway→Tissue subtype. Moreover, within each respective subtype, there were 37 (6%) and 40 (10%) individuals at SuStaIn stages ≥3.

Relationship of SuStaIn stage with lung function in the control population

We tested whether non-zero SuStaIn stage could be used as a marker of early disease in the control population by testing for associations with lung function. SuStaIn stage was associated with baseline lung function and longitudinal decline in lung function in the control population (see Figure 3 for FEV_1/FVC ratio and Supplementary Figure 7 for FEV_1 %predicted, and Supplementary Results).

Longitudinal SuStaIn subtype and stage in the control population

We tested the consistency of the SuStaIn subtype assignments in the smoking controls at five-year follow-up (Supplementary Table 10). At five-year follow-up the assignment to Tissue→Airway and Airway→Tissue subtypes remained consistent in 86% individuals. We verified that the SuStaIn stages were broadly similar at follow-up in the control population. The SuStaIn stages at baseline

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

and follow-up showed a strong correlation (Supplementary Figure 8): r=0.48 (p<0.001) in the Tissue→Airway subgroup, and r=0.61 (p<0.001) in the Airway→Tissue subgroup.

Progression to COPD in the control population

Finally, we tested whether those controls assigned to Tissue→Airway and Airway→Tissue subtypes had a greater individual risk of disease progression compared to those who were normal (SuStaIn stage 0), as measured by a classification of GOLD stage 1 or greater at follow-up. 8.7% of the SuStaIn 0 controls progressed to GOLD stage 1 at follow-up, compared to 23.0% of the Tissue→Airway subtype and 20.9% of the Airway→Tissue subtype. This represents a significantly higher rate of progression to COPD amongst those assigned to SuStaIn subtypes compared to those with normal imaging metrics (p<0.001, Chi-squared test) and therefore that SuStaIn provides a biomarker of 'early COPD'.

Discussion

We report the first application of SuStaIn in COPD, replicating the subtypes identified by SuStaIn in a separate cohort at baseline, and over time in the original cohort. SuStaIn identifies two distinct patterns (subtypes) of COPD disease progression, and early stages of both subtypes were detectable in 29% of 'healthy smokers'. 70% of COPD subjects comprise a Tissue→Airway subtype who follow a progression model in which abnormalities in the small airways (functional small airway disease) and emphysema develop before measurable changes in larger airways. A minority of subjects (30%) comprise an Airway→Tissue subtype in whom disease starts in the larger airways before the later development of emphysema and small airway dysfunction. SuStaIn disease stages correlate with cross-sectional and longitudinal markers of spirometric impairment, with greater loss of lung function at earlier SuStaIn stages of disease. The assignment to Tissue→Airway and Airway→Tissue subtypes remained consistent at five-year follow-up, whilst individuals tended to progress in stage. Progression through stages was more rapid in earlier disease. We therefore identify two distinct

patterns of subtype progression in COPD, of potential utility in the clinic and clinical trials, and provide a biomarker of early COPD in smoking controls.

The long natural history of COPD, over decades, has prevented any single study reporting on longitudinal disease progression in individual patients. Disease progression modelling provides a potential solution to this. Our findings are important for a number of reasons. First, we show that different subjects are on different disease trajectories and may therefore represent distinct endotypes requiring different interventions. Second, we provide early identification of people at risk of developing COPD whilst spirometry is still normal. Reducing the future burden of COPD requires both early identification of smokers likely to develop the condition, and targeted therapy. Finally, our modelling suggests that later stages of COPD progress more slowly, and therefore that disease activity may be greatest in early disease, where treatment and prevention should be targeted.

The Tissue→Airway and Airway→Tissue subtypes we have defined mirror, to some extent, recognised descriptions of COPD, whilst providing a novel imaging biomarker for early disease stratification. Historically, typical phenotypes of COPD have been referred to as “pink puffers” and “blue bloaters” (23). The relative presence of chronic bronchitis or emphysema in addition to significant differences in BMI characterised these classic phenotypes. Such features are also seen in our results, with patients in the Tissue→Airway subtype having a significantly lower BMI and lower incidence of chronic bronchitis compared to those in the Airway→Tissue subtype.

Various studies have shown that inflammatory changes in the small airways are fundamental processes driving the progression and severity of COPD (24). Our results also suggest that the small airways, emphysema and bronchitis are the principal drivers of COPD progression, but that these occur in different proportions and at different times in the two different groups. Just as Hogg (24) showed that a cascade of inflammatory processes lead to small-airways disease and lung function impairment, it is possible that the distinct subtypes we have identified are a function of distinct inflammatory mechanisms (25) with consequent differences in progression patterns. The ability of

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

SuStaIn to separate patients into distinct subtypes at early stages could enable the characterisation of different COPD endotypes.

SuStaIn posits that cross-sectional patient measurements arise from different stages along a disease time course, and that there are distinct groups of individuals (disease subtypes) that undergo different patterns of disease progression. The assumption is that variation in both subtype and stage produces heterogeneity in observed disease biomarkers. Previous findings align with this assumption. The study by Vestbo (2) demonstrates highly-variable decline in FEV_1 in 3-year longitudinal data. As lung function impairment arises from the bulk effect of complex pathological abnormalities in lung structure, different proportions and types of structural damage could explain this variability across patients. The fact that we observed different rates of FEV_1 decline within different subtypes supports this explanation. We therefore demonstrate that changes measured solely by imaging may be used to disentangle subtypes of patients who experience different trajectories of lung function impairment, imperceptible with bulk physiological measurements. Early life factors might also affect the trajectory of lung function decline and risk of developing COPD, but information on these are unfortunately not available in the COPDGene and ECLIPSE cohorts.

Previous research has provided a strong case for early detection of COPD yet this remains challenging in practice. Fletcher and Peto (26) described the rate of lung function decline in COPD, suggesting slow decline at onset followed by a more rapid phase in advanced disease. Recent studies have suggested that faster decline in lung function impairment occurs earlier in disease (27), particularly in mild-to-moderate COPD (27, 28). These results are mirrored in studies showing that smokers may develop emphysema on CT before abnormal lung function (29, 30). Undetected structural alterations may be critical in the early, accelerated decline of lung function and the subsequent course of COPD. Our results support this as we show that early, undetected pathological changes are present in a proportion of healthy smokers, whilst lung function decline is accelerated at earlier stages of disease in the Tissue→Airway subtype. Moreover, our work adds a

new dimension to existing models of disease progression in COPD (26) (27) by disentangling how lung function changes with disease across the COPD population, helping to explain heterogeneity in lung function decline (2).

Our findings are clinically and statistically significant despite the limited precision of some CT metrics. The attenuation value of a voxel is dependent on several factors such as radiation dose, scanner modality, the reconstruction kernel and inspiration level (31). CT scans in COPDGene were not spirometrically gated. Variations in inspiration across patients may cause errors in the measurement of emphysema. Moreover, measurements relating to the airway tree are averages of six bronchial paths in the upper and intermediate zones of the lung (18). Nonetheless, we demonstrate that the SuStaIn subtype trajectories derived from these imaging metrics are reproducible in both a separate cohort and in the same cohort over time, and have strong stratification capabilities in separating individuals with distinct clinical characteristics and patterns of lung function decline. SuStaIn does assume that progression is one directional and that disease cannot 'regress' – it is not known if this may occur in early stages of disease, and the explanation for CT abnormalities in a proportion of people with normal spirometry requires further study.

In conclusion, we report the first use of SuStaIn to study disease progression in COPD, as an exemplar chronic respiratory disease. Using this technique, we report the following novel findings. First, there are two distinct subtypes of COPD – the majority of patients develop small airway disease and emphysema before large airway wall changes, but a significant minority (30%) develop large airway wall changes first. Second, the relationship with lung function in these subtypes is different, with a more rapid initial decline in lung function (greater disease activity) observed in the Tissue→Airway group. This may explain the heterogeneity observed in FEV_1 decline across COPD populations. Finally, the technique suggests that a group of healthy subjects with 'early COPD' at risk of disease progression can be identified using CT biomarkers. In heterogeneous long-term

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

conditions such as COPD there is real need to better stratify patients for targeted therapy. SuStaIn provides a novel technique to achieve this, and a mechanism for detection of early disease.

References

1. Mannino DM, Buist AS. Global burden of COPD: risk factors, prevalence, and future trends. *Lancet* 2007;

2. Vestbo J, Edwards LD, Scanlon PD, Yates JC, Agusti A, Bakke P, Calverley PMA, Celli B, Coxson HO, Crim C, Lomas DA, MacNee W, Miller BE, Silverman EK, Tal-Singer R, Wouters E, Rennard SI, ECLIPSE Investigators. Changes in forced expiratory volume in 1 second over time in COPD. *N Engl J Med* 2011;365:1184–92.

3. Bhatt SP, Soler X, Wang X, Murray S, Anzueto AR, Beaty TH, Boriek AM, Casaburi R, Criner GJ, Diaz AA, Dransfield MT, Curran-Everett D, Galban C, Hoffman EA, Hogg JC, Kazerooni EA, Kim V, Kinney GL, Lagstein A, Lynch DA, Make BJ, Martinez FJ, Ramsdell JW, Reddy R, Ross B, Rossiter H, Steiner RM, Strand M, Van Beek EJR, *et al.* Association between functional small airways disease and FEV 1 decline in COPD. *Am J Respir Crit Care Med* 2016;1164:201511–2219O.

4. Castaldi PJ. Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. *Thorax* 2014;1–8.doi:10.1136/thoraxjnl-2013-203601.

5. Burgel P-R, Paillasseur J-L, Caillaud D, Tillie-Leblond I, Chanez P, Escamilla R, Court-Fortune I, Perez T, Carré P, Roche N. Clinical COPD phenotypes: a novel approach using principal component and cluster analyses. *Eur Respir J* 2010;36:531–9.

6. Ross JC, Castaldi PJ, Cho MH, Chen J, Chang Y, Dy JG, Silverman EK, Washko GR, Jose Estepar RS. A Bayesian Nonparametric Model for Disease Subtyping: Application to Emphysema Phenotypes. *IEEE Trans Med Imaging* 2017;36:343–354.

7. Ross JC, Castaldi PJ, Cho MH, Hersh CP, Rahaghi FN, Sanchez-Ferrero G V., Parker MM, Litonjua AA, Sparrow D, Dy JG, Silverman EK, Washko GR, San Jose Estepar R. Longitudinal modeling of lung function trajectories in smokers with and without Chronic Obstructive Pulmonary Disease. *Am J Respir Crit Care Med* 2018;198:1033–1042.

8. Bragman FJS, McClelland JR, Jacob J, Hurst JR, Hawkes DJ. Manifold learning of COPD. *Med Image Comput Comput Aided Interv* Springer, Cham; 2017. p. 586–593.

9. Harmouche R, Ross JC, Diaz AA, Washko GR, Estepar RSJ. A Robust Emphysema Severity Measure Based on Disease Subtypes. *Acad Radiol* 2016;23:421–428.

10. Young AL, Oxtoby NP, Daga P, Cash DM, Fox NC, Ourselin S, Schott JM, Alexander DC. A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain* 2014;137:2564–2577.

11. Fonteijn HM, Modat M, Clarkson MJ, Barnes J, Lehmann M, Hobbs NZ, Scahill RI, Tabrizi SJ, Ourselin S, Fox NC, Alexander DC. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *Neuroimage* 2012;60:1880–9.

12. Oxtoby NP, Alexander DC. Imaging plus X: Multimodal models of neurodegenerative disease. *Curr Opin Neurol* 2017;30:371–379.

13. Donohue MC, Jacqmin-Gadda H, Le Goff M, Thomas RG, Raman R, Gamst AC, Beckett LA, Jack CR, Weiner MW, Dartigues J-F, Aisen PS. Estimating long-term multivariate progression from short-term data. *Alzheimer's Dement* 2014;10:S400–S410.

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

14. Bilgel M, Prince JL, Wong DF, Resnick SM, Jedynak BM. A multivariate nonlinear mixed effects model for longitudinal image analysis: Application to amyloid imaging. *Neuroimage* 2016;134:658–670.

15. Wang X, Sontag D, Wang F. Unsupervised Learning of Disease Progression Models. *20th ACM SIGKDD Conf Knowl Disovery Data Min* 2014. p. 85–94.

16. Young AL, Marinescu R-V V, Oxtoby NP, Bocchetta M, Yong K, Firth N, Cash DM, Thomas DL, Dick KM, Cardoso J, Swieten J van, Borroni B, Galimberti D, Masellis M, Tartaglia MC, Rowe JB, Graff C, Tagliavini F, Frisoni G, Laforce R, Finger E, Medonça A, Sorbi S, Warren JD, Crutch S, Fox NC, Ourselin S, Schott JM, Rohrer JD, *et al.* Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nat Commun* 2018;9:4273.

17. Bragman FJS, Young AL, Hawkes DJ, Alexander DC, Hurst JR. Disease progression patterns in COPD. *Eur Respir Soc* 2018. p. OA2139.

18. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, Curran-Everett D, Silverman EK, Crapo JD. Genetic epidemiology of COPD (COPDGene) study design. *COPD* 2010;7:32–43.

19. Vestbo J, Anderson W, Coxson HO, Crim C, Dawber F, Edwards L, Hagan G, Knobil K, Lomas DA, MacNe W, Silverman EK, Tal-Singer R. Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE). *Eur Respir J* 2008;31:869–873.

20. Galbán CJ, Han MK, Boes JL, Chughtai K a, Meyer CR, Johnson TD, Galbán S, Rehemtulla A, Kazerooni E a, Martinez FJ, Ross BD. Computed tomography-based biomarker provides unique signature for diagnosis of COPD phenotypes and disease progression. *Nat Med* 2012;18:1711–5.

21. Nakano Y, Wong JC, De Jong PA, Buzatu L, Nagao T, Coxson HO, Elliott WM, Hogg JC, Paré PD. The prediction of small airway dimensions using computed tomography. *Am J Respir Crit Care Med* 2005;171:142–146.

22. Coxson HO, Dirksen A, Edwards LD, Yates JC, Agusti A, Bakke P, Calverley PMA, Celli B, Crim C, Duvoix A, Fauerbach PN, Lomas DA, Macnee W, Mayer RJ, Miller BE, Müller NL, Rennard SI, Silverman EK, Tal-Singer R, Wouters EFM, Vestbo J. The presence and progression of emphysema in COPD as determined by CT scanning and biomarker expression: A prospective analysis from the ECLIPSE study. *Lancet Respir Med* 2013;1:129–136.

23. Filley GF, Beckwitt HJ, Reeves JT, Mitchell RS. Chronic obstructive bronchopulmonary disease. *Am J Med* 1968;44:26–38.

24. Hogg JC, Macklem PT, Thurlbeck WM. Site and nature of airway obstruction in chronic obstructive lung disease. *N Engl J Med* 1968;278:1355–60.

25. Barnes PJ. Inflammatory mechanisms in patients with chronic obstructive pulmonary disease. *J Allergy Clin Immunol* 2016;

26. Fletcher C, Peto R. The natural history of chronic airflow obstruction. *BMJ* 1977;1:1645–1648.

27. Tantucci C, Modina D. Lung function decline in COPD. *Int J Chron Obstruct Pulmon Dis* 2012;7:95.

28. Rennard SI, Drummond MB. Early chronic obstructive pulmonary disease: definition, assessment, and prevention. *Lancet* 2015;385:1778–1788.

29. Sashidhar K, Gulati M, Gupta D, Monga S, Suri S. Emphysema in heavy smokers with normal chest radiography. Detection and quantification by HCRT. *Acta radiol* 2002;43:60–5.

30. Srinakarin J, Thammaroj J, Boonsawat W. Comparison of high-resolution computed tomography with pulmonary function testing in symptomatic smokers. *J Med Assoc Thail* 2003;86:522–528.

31. Mets OM, de Jong PA, van Ginneken B, Gietema HA, Lammers JWJ. Quantitative computed tomography in COPD: possibilities and limitations. *Lung* 2012;190:133–145.

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

Acknowledgements

COPDGene® Investigators – Core Units

Administrative Center: James D. Crapo, MD (PI); Edwin K. Silverman, MD, PhD (PI); Barry J. Make, MD; Elizabeth A. Regan, MD, PhD

Genetic Analysis Center: Terri Beaty, PhD; Ferdouse Begum, PhD; Peter J. Castaldi, MD, MSc; Michael Cho, MD; Dawn L. DeMeo, MD, MPH; Adel R. Boueiz, MD; Marilyn G. Foreman, MD, MS; Eitan Halper-Stromberg; Lystra P. Hayden, MD, MMSc; Craig P. Hersh, MD, MPH; Jacqueline Hetmanski, MS, MPH; Brian D. Hobbs, MD; John E. Hokanson, MPH, PhD; Nan Laird, PhD; Christoph Lange, PhD; Sharon M. Lutz, PhD; Merry-Lynn McDonald, PhD; Margaret M. Parker, PhD; Dandi Qiao, PhD; Elizabeth A. Regan, MD, PhD; Edwin K. Silverman, MD, PhD; Emily S. Wan, MD; Sungho Won, Ph.D.; Phuwanat Sakornsakolpat, M.D.; Dmitry Prokopenko, Ph.D.

Imaging Center: Mustafa Al Qaisi, MD; Harvey O. Coxson, PhD; Teresa Gray; MeiLan K. Han, MD, MS; Eric A. Hoffman, PhD; Stephen Humphries, PhD; Francine L. Jacobson, MD, MPH; Philip F. Judy, PhD; Ella A. Kazerooni, MD; Alex Kluiber; David A. Lynch, MB; John D. Newell, Jr., MD; Elizabeth A. Regan, MD, PhD; James C. Ross, PhD; Raul San Jose Estepar, PhD; Joyce Schroeder, MD; Jered Sieren; Douglas Stinson; Berend C. Stoel, PhD; Juerg Tschirren, PhD; Edwin Van Beek, MD, PhD; Bram van Ginneken, PhD; Eva van Rikxoort, PhD; George Washko, MD; Carla G. Wilson, MS;

PFT QA Center, Salt Lake City, UT: Robert Jensen, PhD

Data Coordinating Center and Biostatistics, *National Jewish Health, Denver, CO*: Douglas Everett, PhD; Jim Crooks, PhD; Camille Moore, PhD; Matt Strand, PhD; Carla G. Wilson, MS

Epidemiology Core, *University of Colorado Anschutz Medical Campus, Aurora, CO*: John E. Hokanson, MPH, PhD; John Hughes, PhD; Gregory Kinney, MPH, PhD; Sharon M. Lutz, PhD; Katherine Pratte, MSPH; Kendra A. Young, PhD

Mortality Adjudication Core: Surya Bhatt, MD; Jessica Bon, MD; MeiLan K. Han, MD, MS; Barry Make, MD; Carlos Martinez, MD, MS; Susan Murray, ScD; Elizabeth Regan, MD; Xavier Soler, MD; Carla G. Wilson, MS

Biomarker Core: Russell P. Bowler, MD, PhD; Katerina Kechris, PhD; Farnoush Banaei-Kashani, Ph.D

COPDGene® Investigators – Clinical Centers

Ann Arbor VA: Jeffrey L. Curtis, MD; Carlos H. Martinez, MD, MPH; Perry G. Pernicano, MD

Baylor College of Medicine, Houston, TX: Nicola Hanania, MD, MS; Philip Alapat, MD; Mustafa Atik, MD; Venkata Bandi, MD; Aladin Boriek, PhD; Kalpatha Guntupalli, MD; Elizabeth Guy, MD; Arun Nachiappan, MD; Amit Parulekar, MD;

Brigham and Women's Hospital, Boston, MA: Dawn L. DeMeo, MD, MPH; Craig Hersh, MD, MPH; Francine L. Jacobson, MD, MPH; George Washko, MD

Columbia University, New York, NY: R. Graham Barr, MD, DrPH; John Austin, MD; Belinda D'Souza, MD; Gregory D.N. Pearson, MD; Anna Rozenshtein, MD, MPH, FACR; Byron Thomashow, MD

Duke University Medical Center, Durham, NC: Neil MacIntyre, Jr., MD; H. Page McAdams, MD; Lacey Washington, MD

HealthPartners Research Institute, Minneapolis, MN: Charlene McEvoy, MD, MPH; Joseph Tashjian, MD

Johns Hopkins University, Baltimore, MD: Robert Wise, MD; Robert Brown, MD; Nadia N. Hansel, MD, MPH; Karen Horton, MD; Allison Lambert, MD, MHS; Nirupama Putcha, MD, MHS

Los Angeles Biomedical Research Institute at Harbor UCLA Medical Center, Torrance, CA: Richard Casaburi, PhD, MD; Alessandra Adami, PhD; Matthew Budoff, MD; Hans Fischer, MD; Janos Porszasz, MD, PhD; Harry Rossiter, PhD; William Stringer, MD

Michael E. DeBakey VAMC, Houston, *TX*: Amir Sharafkhaneh, MD, PhD; Charlie Lan, DO

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

Minneapolis VA: Christine Wendt, MD; Brian Bell, MD

Morehouse School of Medicine, Atlanta, GA: Marilyn G. Foreman, MD, MS; Eugene Berkowitz, MD, PhD; Gloria Westney, MD, MS

National Jewish Health, Denver, CO: Russell Bowler, MD, PhD; David A. Lynch, MB

Reliant Medical Group, Worcester, MA: Richard Rosiello, MD; David Pace, MD

Temple University, Philadelphia, PA: Gerard Criner, MD; David Ciccolella, MD; Francis Cordova, MD; Chandra Dass, MD; Gilbert D'Alonzo, DO; Parag Desai, MD; Michael Jacobs, PharmD; Steven Kelsen, MD, PhD; Victor Kim, MD; A. James Mamary, MD; Nathaniel Marchetti, DO; Aditi Satti, MD; Kartik Shenoy, MD; Robert M. Steiner, MD; Alex Swift, MD; Irene Swift, MD; Maria Elena Vega-Sanchez, MD

University of Alabama, Birmingham, AL: Mark Dransfield, MD; William Bailey, MD; Surya Bhatt, MD; Anand Iyer, MD; Hrudaya Nath, MD; J. Michael Wells, MD

University of California, San Diego, CA: Joe Ramsdell, MD; Paul Friedman, MD; Xavier Soler, MD, PhD; Andrew Yen, MD

University of Iowa, Iowa City, IA: Alejandro P. Comellas, MD; Karin F. Hoth, PhD; John Newell, Jr., MD; Brad Thompson, MD

University of Michigan, Ann Arbor, MI: MeiLan K. Han, MD, MS; Ella Kazerooni, MD; Carlos H. Martinez, MD, MPH

University of Minnesota, Minneapolis, MN: Joanne Billings, MD; Abbie Begnaud, MD; Tadashi Allen, MD

University of Pittsburgh, Pittsburgh, PA: Frank Sciurba, MD; Jessica Bon, MD; Divay Chandra, MD, MSc; Carl Fuhrman, MD; Joel Weissfeld, MD, MPH

University of Texas Health Science Center at San Antonio, San Antonio, TX: Antonio Anzueto, MD; Sandra Adams, MD; Diego Maselli-Caceres, MD; Mario E. Ruiz, MD

ECLIPSE

The ECLIPSE study was sponsored by GlaxoSmithKline. The study sponsor did not place any restrictions regarding statements made in this manuscript. A Steering Committee and a Scientific Committee comprising academic and sponsor representatives developed the original ECLIPSE study design, had full access to the study data, and were responsible for decisions regarding publications.

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

Figure Legends

FIGURE 1: Disease progression patterns predicted by SuStaIn.

COPD is characterised by two distinct disease progression models (top row). In the "Tissue→Airway" subtype (70%; top left) the presence of emphysema and functional small airways disease initiates disease progression followed by later emergence of pathology in larger airways (the overall tissue damage measure captures the presence of both functional small airways disease and emphysema). In the "Airway→Tissue" subtype (30%; top right), disease progression is initiated by pathology in the larger airways before the development of functional small airways disease and emphysema. At each SuStaIn stage a new z-score event occurs when a feature transitions to a new severity level, as indexed by a z-score with respect to the control population; z-scores of z=1 (orange) and z=2 (red). Higher opacity represents a higher confidence in the ordering. The bottom row visualises the PRM images and airway wall thickness values for representative patients at different SuStaIn subtypes and stages. The airway wall thickness values are visualised using a purple colour scale on top of an airway tree segmentation, with the minimum value of the colour scale corresponding to the 1st percentile of airway wall thickness values across the population, and the maximum value of the colour scale corresponding to the 99th percentile. In the Tissue→Airway subtype, the first individual (early stage) has early tissue damage visible at the outer edges of the lung but no airway wall changes, the second individual (middle stage) has visible tissue damage but no airway changes, and the third individual (late stage) has severe tissue damage together with airway wall thickening. In the Airway→Tissue subtype, the first individual (early stage) has early signs of airway wall thickening but no visible tissue damage, the second individual (middle stage) has clear signs of airway wall thickening but very little visible tissue damage, and the third individual (late stage) has severe airway wall thickening and tissue damage.

FIGURE 2: Relationship between SuStaIn stage and lung function.

(A) Scatter plot of cross-sectional spirometry versus SuStaIn stage for the Tissue→Airway and Airway→Tissue subtypes. A linear and a quadratic model are fitted to the data via a least-squares estimation to gauge the relationship between SuStaIn stage and markers of lung function. In the Tissue→Airway subtype, there is a visible non-linear relationship between lung function and SuStaIn stage with a more rapid decrease in lung function at earlier SuStaIn stages. The decline in lung function in the Airway→Tissue subgroup is linear and less rapid at earlier SuStaIn stages. (B) Scatter plot of measured decline in spirometry versus baseline SuStaIn stage for the Tissue→Airway and Airway→Tissue subtypes in GOLD 1-2 subjects. In both the Tissue→Airway and Airway→Tissue subtypes, SuStaIn stage at baseline correlated with future decline in lung function measured using FEV_1/FVC.

FIGURE 3: Relationship between lung function and SuStaIn stage in smoking controls.

Baseline SuStaIn Stage is associated with cross-sectional and longitudinal changes in airflow obstruction in smoking controls. A) Scatter plot of baseline values FEV_1/FVC versus SuStaIn stage in the control population. B) Scatter plot of longitudinal change in FEV_1/FVC per year versus SuStaIn stage in the control population.

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

Tables

TABLE 1: Basic demographics for the COPDGene control and COPD populations used in deriving the SuStaIn subtype trajectories.

Parameter	Control subjects	COPD subjects
Subjects, *n*	3479	3698
Age (years)*, mean (SD)*	56.90 (8.45)	63.13 (8.61)
Male, *n (%)*	1816 (52)	2087 (56)
Female, *n (%)*	1663 (48)	1611 (44)
GOLD Stage 1, *n (%)*	NA	643 (17)
GOLD Stage 2, *n (%)*		1616 (44)
GOLD Stage 3, *n (%)*		960 (26)
GOLD Stage 4, *n (%)*		479 (13)
Smoking history (pack-years), *mean (SD.)*	37.33 (20.04)	51.91 (26.99)
Exacerbations (n/year), *mean (SD.)*	0.13 (0.53)	0.64 (1.18)

TABLE 2: Demographics of patients in the Tissue→Airway and Airway→Tissue subtypes. We report two sample t-test for continuous variables, chi-squared test for categorical variables, and Mann-Whitney U-test results for frequency data. Only patients at SuStaIn stages ≥1 were included.

Feature	Tissue→Airway	Airway→Tissue	
Number of patients, *n (%)*	2354 (70.4%)	988 (29.6%)	
Male, *n (%)*	1230 (52.3)	657 (66.5)	$p < 0.001$
Female, *n (%)*	1124 (47.7)	331 (33.5)	
Age (years), *mean (SD.)*	63.18 (8.14)	63.17 (9.49)	$p = 0.92$
BMI (kg/m^2), *mean (SD.)*	26.65 (5.43)	30.54 (6.28)	$p < 0.001$
FEV_1 (% predicted), *mean (SD.)*	53.63 (23.05)	58.64 (17.74)	$p < 0.001$
FEV_1/FVC ratio, *mean (SD.)*	0.49 (0.14)	0.56 (0.11)	$p < 0.001$
GOLD Stage 1, *n (%)*	340 (14.4)	103 (10.4)	
GOLD Stage 2, *n (%)*	908 (38.6)	559 (56.6)	$p < 0.001$
GOLD Stage 3, *n (%)*	680 (28.9)	273 (27.6)	
GOLD Stage 4, *n (%)*	426 (18.1)	53 (5.4)	
Smoking history (pack-years), *mean (SD.)*	53.10 (26.42)	50.35 (26.12)	$p = 0.006$
Exacerbations (n/year), *mean (SD.)*	0.71 (1.23)	0.62 (1.16)	$p = 0.018$
Chronic Bronchitis, *n (%)*	591 (25.1)	314 (31.8)	$p < 0.001$
% Emphysema, *mean (SD.)*	15.17 (13.67)	4.08 (6.46)	$p < 0.001$
% fSAD*, mean (SD.)*	28.89 (11.86)	20.18 (12.83)	$p < 0.001$
% Tissue Damage, *mean (SD.)*	44.06 (20.79)	24.26 (17.57)	$p < 0.001$
Airway Wall Area %, *mean (SD.)*	51.94 (6.81)	61.77 (6.21)	$p < 0.001$
Pi10 SRWA (mm)*, mean (SD.)*	2.52 (0.48)	3.13 (0.56)	$p < 0.001$
Airway Wall Thickness (mm), *mean (SD.)*	1.06 (0.19)	1.34 (0.21)	$p < 0.001$

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

Figures

FIGURE 1:

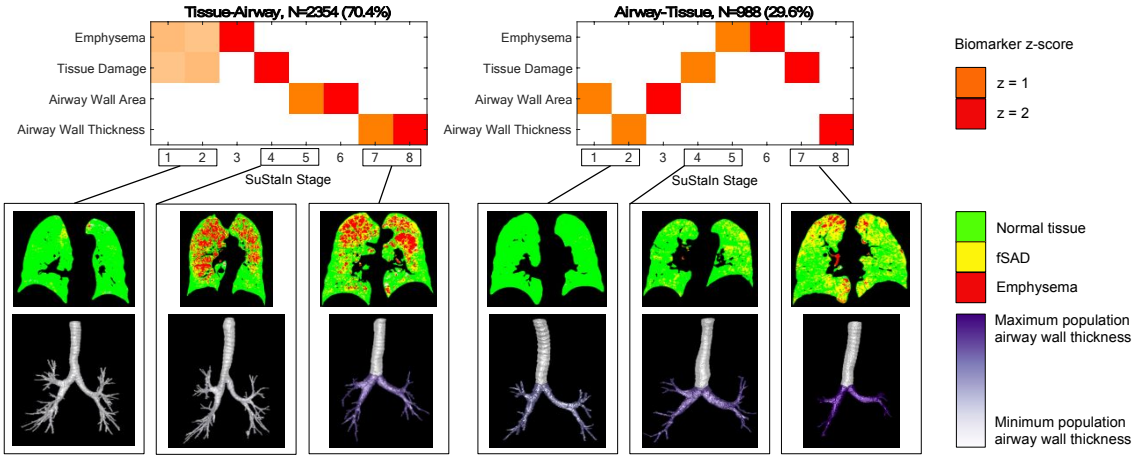
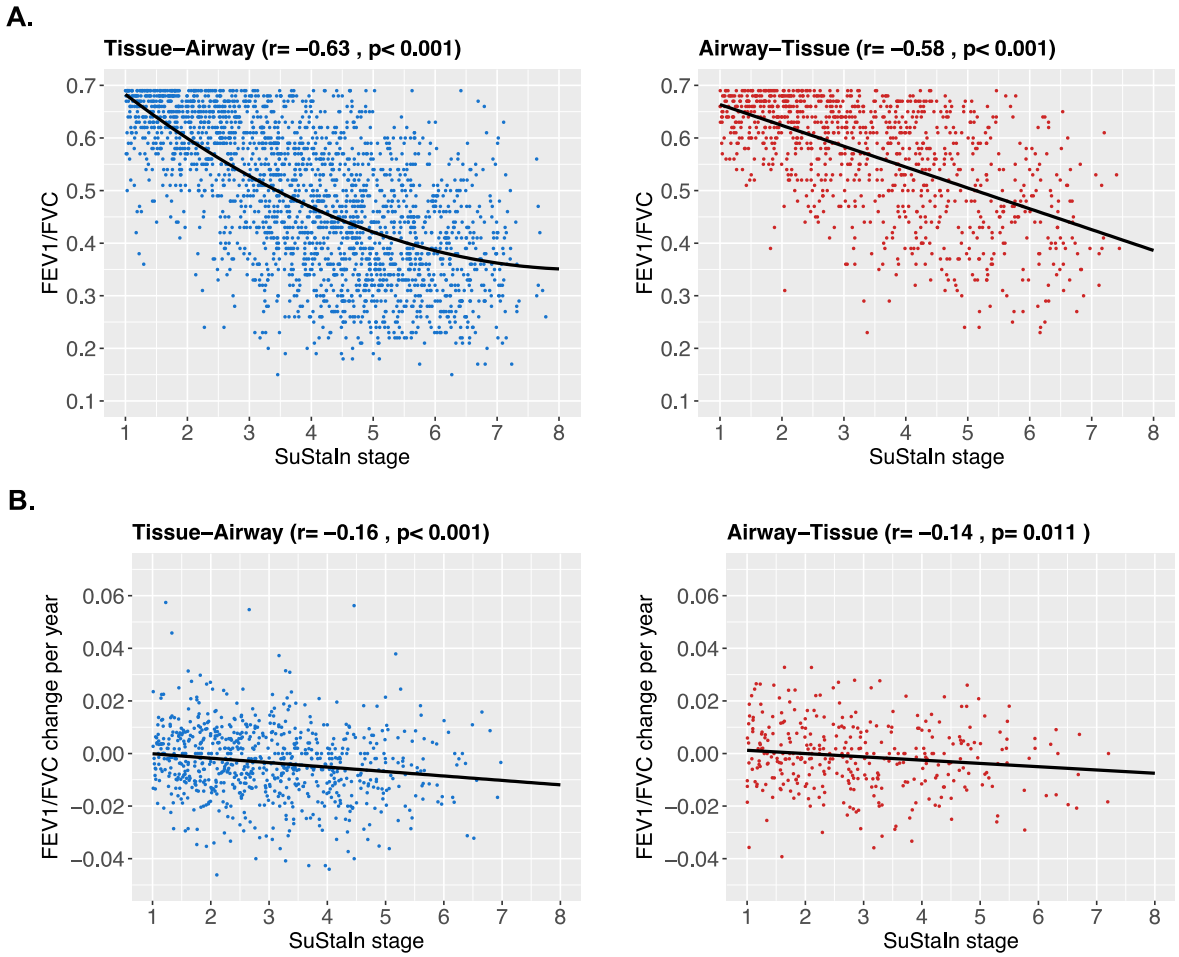
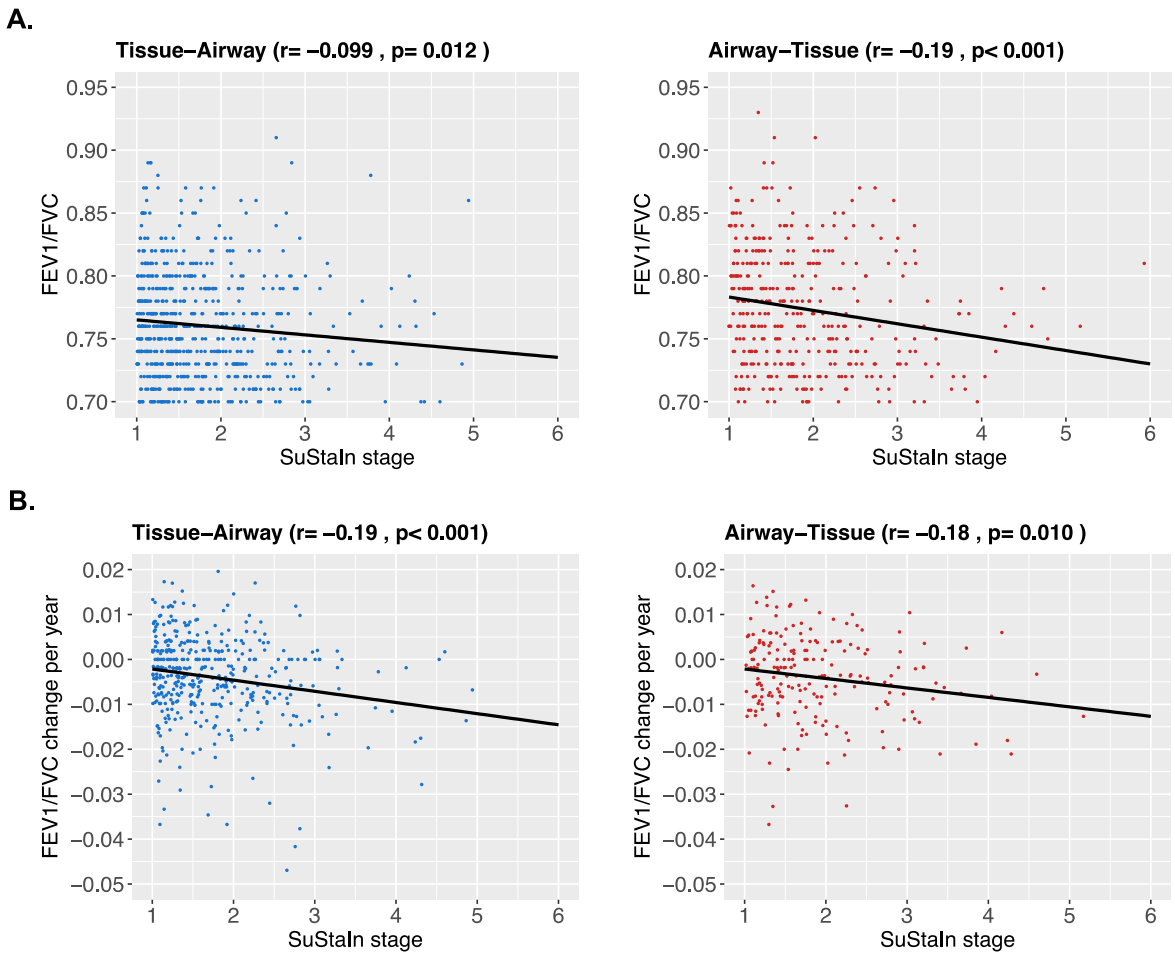


FIGURE 2:



AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

FIGURE 3:



AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

Disease Progression Modelling in

Chronic Obstructive Pulmonary Disease (COPD)

Online Supplementary Material

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

Supplementary Method

The following is an unabridged version of the Method.

Definitions and Overview

Key terminology is defined in the Glossary.

Model development used CT data from COPDGene Phase 1 (18), comprising a cross-sectional dataset of baseline measurements from 3479 smoking controls and 3698 COPD patients. We repeated the SuStaIn algorithm using baseline data from 303 smoking controls and 1809 COPD patients in the ECLIPSE study (19) to verify consistency of SuStaIn output in an independent dataset. We evaluated longitudinal progression using follow-up (Phase 2) COPDGene scans and data to verify the SuStaIn subtype progression patterns (reconstructed from cross-sectional data) against true longitudinal progression of individual subjects. This included 1929 COPD subjects and 2158 controls who had all imaging biomarkers available from the initial scan together with measures of lung function from both time points, and a second dataset of 1675 COPD subjects and 1939 controls who had all imaging biomarkers available at both phases of the COPDGene study.

GLOSSARY:
SUBTYPE – a group of subjects who share a particular trajectory of biomarker evolution.
STAGE – the position on a subtype trajectory of an individual subject at a specific time. In SuStaIn this represents the degree of abnormality in imaging biomarkers and a change in stage occurs when an imaging biomarker becomes more abnormal relative to a control population.
DISEASE PROGRESSION – change in stage with time as the natural history of the condition unfolds. We use the term in two distinct contexts: 1. GROUP (SUBTYPE) LEVEL: referring to the sequence of changes that the typical patient undergoes from start to finish. 2. INDIVIDUAL LEVEL: change in stage or severity of an individual subject as biomarkers become increasingly abnormal.

Cohorts

COPDGene

Data from subjects participating in COPDGene, a large multicentre observational cohort study, were used in this analysis (18). COPDGene enrolled current and former smokers with greater than or equal to 10 pack-years smoking history, with and without airflow obstruction. In this study we used data from COPD patients (GOLD1-4) and smoking controls ($FEV_1/FVC > 0.70$ and FEV_1 predicted > 80%), downloaded on 29th September 2018. In Phase 1, 10371 individuals were recruited, including 4510 COPD patients and 4409 smoking controls. In Phase 2, five years later, 4568 individuals were followed up, including 2025 COPD patients and 2110 smoking controls. We selected subsets of these individuals that had baseline and/or longitudinal measures of various imaging features for use in our analysis as described above.

The CT protocol for COPDGene has previously been described (18). CT analysis to obtain the imaging features included in this study was performed using Thirona lung quantification software (Thirona, Netherlands, http://www.thirona.eu) (18). The derived imaging features were: 1) emphysema

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

obtained using parametric response mapping (PRM) (20), 2) functional small airways disease (fSAD) obtained from PRM, 3) Pi10 square root wall area (SRWA) (21), and 4) segmental airway wall thickness.

ECLIPSE

Data from individuals participating in ECLIPSE (Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints) were used to replicate the subtypes identified in COPDGene. ECLIPSE was an observational, longitudinal, and controlled study where, after the baseline visit, participants were evaluated at 3 months, 6 months, and then every 6 months for 3 years. The study design of ECLIPSE (Clinicaltrials.gov identifier NCT00292552; GSK study code SCO104960) has been published previously (19) ECLIPSE complied with the Declaration of Helsinki and Good Clinical Practice Guidelines and was approved by the ethics committees of the participating centers. All participants provided written informed consent before the performance of all study-related assessments. There were only two imaging measures available in the ECLIPSE study: emphysema and Pi10 SRWA. Both were estimated using VIDA software applied to inspiratory scans. The scan protocol for CT in ECLIPSE has been previously described (22).

Disease progression modelling

Given a cross-sectional data set, SuStaIn simultaneously identifies a set of disease subtypes, each defined by a distinct trajectory of biomarker evolution with a probabilistic assignment of each subject to a subtype and stage along the corresponding trajectory. A conceptual overview is illustrated as Supplementary Figure 1. The trajectory of each subtype is described as a linear z-score model, which consists of a series of stages in which each stage corresponds to a biomarker reaching a particular z-score (relative to a control group). The SuStaIn algorithm simultaneously optimises the progression pattern for each subtype (i.e. linear z-score model) and the assignment of individuals to

subtypes. The optimal number of subtypes is determined using information criterion. The population-level SuStaIn subtype progression patterns can then be used to assign individuals to subtypes and stages probabilistically.

Identification of the subtypes

We applied the SuStaIn algorithm (16) to COPD GOLD1-4 patients from the COPDGene dataset using CT metrics as input. SuStaIn requires monotonic measurements, however fSAD has been shown to convert to emphysema at later stages of COPD (20), potentially violating this assumption by causing fSAD to decrease once replaced by emphysema. To obtain a monotonic representation of fSAD and emphysema, we replaced fSAD with a combined measure we term 'overall tissue damage', computed as the sum of fSAD and emphysema (and thus similar to a measure of air trapping). Using a measure of overall tissue damage in combination with a measure of emphysema provides the same information as using fSAD and emphysema without compromising the monotonicity assumption. As SuStaIn requires input features that are expressed as z-scores relative to a control population, we transformed each dataset into z-scores relative to the smoking controls in COPDGene. Prior to performing the z-score transformation, imaging measures were log transformed to improve normality, and the effects of age, gender, height, race, scanner model and smoking history estimated from the smoking control population were regressed out.

Replication of the subtypes

To replicate the subtype progression patterns in an independent dataset we repeated our analysis in COPD GOLD1-4 patients from ECLIPSE using the subset of CT metrics available from inspiratory scans, and the corresponding smoking controls to perform z-score transformation. As ECLIPSE only has inspiratory scans, to investigate the effect of using a reduced set of imaging features derived

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

from inspiratory scans alone we repeated the SuStaIn algorithm on a COPDGene cross-sectional dataset consisting of baseline measurements from 4102 smoking controls and 4152 COPD patients with inspiratory measurements available for emphysema and Pi10 SRWA. We refer to these data as the 'Inspiratory COPDGene' dataset.

Cross-sectional subtyping and staging

We used the SuStaIn model (i.e. the subtype progression patterns identified using the SuStaIn algorithm) to automatically assign individuals to their most probable subtype and stage. We computed a continuous estimate of SuStaIn stage by evaluating the weighted average of each individual's probability distribution over stages. We used this process to obtain a subtype and stage assignment for all COPD patients and each control subject in the COPDGene dataset at each of the two visits. We repeated the same process of assigning individuals to SuStaIn subtypes and stages in the ECLIPSE and Inspiratory COPDGene datasets, using the corresponding SuStaIn subtype progression patterns estimated in each dataset.

COPD patients at SuStaIn stage 0 in the disease progression trajectories have no detectable changes in imaging features despite abnormal spirometry and cannot be confidently assigned to a SuStaIn subtype. We therefore only included patients at SuStaIn stage greater or equal to one when comparing subtypes in the COPDGene dataset. In the ECLIPSE and Inspiratory COPDGene datasets, we included patients at SuStaIn stage greater or equal to 1/2 for comparing subtypes. This lower threshold was necessary because of the reduced number of stages in the ECLIPSE and Inspiratory COPDGene models consequent on the smaller set of CT features.

Longitudinal subtyping and staging

We assigned individuals from COPDGene Phase 2 to subtypes and stages using the same procedure described above, measuring the subtypes and stages against the subtype progression patterns estimated using the COPDGene Phase 1 dataset.

Statistical analysis

Clinical characteristics of the disease progression subgroups

We compared the clinical characteristics of individuals assigned to each subtype using two sample t-tests for continuous variables, chi-squared tests for categorical variables, and Mann-Whitney U-tests for frequency data.

Relationship between SuStaIn stage and baseline lung function

We verified that SuStaIn stage could be used as a measure of disease severity in COPD by examining whether SuStaIn stage correlated with spirometric impairment as assessed by FEV_1/FVC and FEV_1%predicted. In the COPDGene dataset the relationship between SuStaIn stage and spirometry appeared non-linear. We assessed the statistical significance of this non-linear relationship by fitting a quadratic and a linear model between SuStaIn stage and spirometry and assessing the goodness of fit of each using analysis of variance (ANOVA) for nested models.

Relationship between SuStaIn stage and longitudinal decline in lung function

We further evaluated whether a higher SuStaIn stage could be used as an indicator of future lung function decline (disease progression) by assessing whether baseline SuStaIn stage was correlated

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

with changes in lung function between baseline and follow-up on FEV_1/FVC and FEV_1%predicted in GOLD1-2 and GOLD3-4 patients in each SuStaIn subtype.

Consistency of subtypes

We assessed whether the SuStaIn subtypes remained consistent at five-year follow-up, quantifying consistency as the percentage of individuals in which the subtype assignment remained the same.

Progression of individuals in SuStaIn stage

We assessed whether individuals progressed in SuStaIn stage between baseline and follow-up by comparing the distribution of SuStaIn stages at baseline and follow-up in GOLD1-2 and GOLD3-4 patients belonging to each SuStaIn subtype using two sample t-tests.

Analysis in smoking controls

We repeated the above analyses in the COPDGene smoking control group to test whether SuStaIn subtype and stage might be useful for identification of otherwise healthy individuals at risk of developing COPD.

Supplementary Results

Replication of the subtype progression patterns

The presence of two progression patterns was replicated using data from the ECLIPSE study (Supplementary Figure 5; Supplementary Table 1). Only emphysema and airway wall area measures were available in ECLIPSE. The proportion of individuals assigned to the Tissue→Airway subtype was somewhat larger using the reduced set of imaging measures at 90.7% (n=1461) in the Tissue→Airway group and 9.3% (n=150) in the Airway→Tissue group. This was likely due to differences in the number of imaging features rather than the study populations. Subtypes in COPDGene were consistent with those in ECLIPSE when using the 'Inspiratory COPDGene' dataset comprising the same imaging features (Supplementary Figure 5; Supplementary Table 2): a comparable proportion of individuals were assigned to each subtype (92.8%, n=3300 Tissue→Airway; 7.2%, n=255 Airway→Tissue).

Clinical characteristics of COPD subtypes in ECLIPSE

The clinical characteristics of the disease progression subtypes were broadly replicable in the ECLIPSE dataset (Supplementary Table 3), with the Tissue→Airway group having a smaller proportion of men (61.9% versus 82.0%, p<0.001), lower BMI (26.02 versus 31.30 kgm^{-2}, p<0.001), lower prevalence of chronic bronchitis (32.8% versus 39.3%, p=0.131) and more severe spirometric impairment (FEV_1/FVC ratio 0.43 versus 0.50, p<0.001). The clinical characteristics of the subtypes in the Inspiratory COPDGene dataset are reported as Supplementary Table 4.

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

Non-linear relationship of SuStaIn stage with baseline lung function

We observed that the relationship between lung function decline and SuStaIn stage differed between the Tissue→Airway and Airway→Tissue subtypes. An accelerated early decline in lung function was apparent in the Tissue→Airway subtype with a reduced rate of decline in later SuStaIn stages (Figure 2A and Supplementary Figure 3A): a non-linear model was a better fit (nested models ANOVA; p<0.001 FEV_1/FVC; p<0.001 FEV_1 %predicted). In contrast, we observed no difference between the models in the Airway→Tissue subtype (nested models ANOVA; p=0.20 FEV_1/FVC; p=0.97 FEV_1 %predicted).

Relationship of SuStaIn stage with baseline lung function in ECLIPSE

The correlations between baseline lung function and SuStaIn stage were replicable in the ECLIPSE dataset (Supplementary Figure 4), although weaker using the smaller set of imaging features available in ECLIPSE. This analysis in the Inspiratory COPDGene dataset is reported in Supplementary Figure 5.

Consistency of SuStaIn subtypes at five-year follow-up

SuStaIn assumes that individuals belong to a single disease subtype, progressing only in stage with time. We verified this by testing whether the SuStaIn subtypes remained the same at five-year follow-up using a longitudinal validation dataset consisting of COPDGene individuals who had all imaging biomarkers available at both phases (subject characteristics reported in Supplementary Table 6). After five years follow-up, the assignment to Tissue→Airway and Airway→Tissue subtypes remained consistent in 1283/1472 (87%) individuals (Supplementary Table 8). Amongst those with a high probability of belonging to the Tissue→Airway or Airway→Tissue subgroups at baseline

(probability of ≥ 0.75) the proportion of individuals with a consistent assignment at follow-up was higher still at 95% (Supplementary Table 9).

Progression of individuals in SuStaIn stage

We next assessed whether individuals progressed in SuStaIn stage between baseline and follow-up, hypothesising that their stage would show progression. As expected, the SuStaIn stages at baseline and follow-up showed a strong correlation (Supplementary Figure 6): r=0.81 (p<0.001) in the Tissue→Airway subtype, and r=0.71 (p<0.001) in the Airway→Tissue subtype. On average, those in the Tissue→Airway subtype progressed by 0.387 SuStaIn stages over five years (p<0.001), which corresponds to 4.84% of possible disease progression estimated by SuStaIn (stages 1 to 8). Conversely, those in the Airway→Tissue subtype progressed by 0.158 SuStaIn stages (p=0.008), 1.98% of possible SuStaIn progression. Progression was more rapid in GOLD1-2 patients, with those assigned to the Tissue→Airway subtype progressing by an average of 0.459 SuStaIn stages (p<0.001), and those assigned to the Airway→Tissue subtype by an average of 0.224 SuStaIn stages (p=0.001). By comparison, amongst GOLD3-4 patients, those assigned to the Tissue→Airway subtype progressed by an average of 0.251 SuStaIn stages (p<0.001), whilst GOLD3-4 patients assigned to the Airway→Tissue subtype did not progress (average 0.003 SuStaIn stages, p=0.98).

Relationship of SuStaIn stage with baseline lung function in the control population

We tested whether SuStaIn stage could be used as a measure of early disease severity in the control population by looking for associations between SuStaIn stage and baseline lung function (Figure 3A for FEV_1/FVC ratio and Supplementary Figure 7A for FEV_1 %predicted). We found a weak, but significant relationship between SuStaIn stage and baseline FEV_1/FVC in both the Tissue→Airway (r=-

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

0.099, *p*=0.012) and Airway→Tissue (r=-0.19, *p*<0.001) subtypes, but no significant relationship with baseline FEV_1 %predicted in either subtype (r=-0.039, *p*=0.32 and r = 0.0089, *p*=0.86, respectively).

Relationship of SuStaIn stage with longitudinal decline in lung function in the control population

We next tested for associations between SuStaIn stage and longitudinal decline in lung function in the control group (Figure 3B for FEV_1/FVC ratio and Supplementary Figure 7B for FEV_1 %predicted). We found a correlation between baseline SuStaIn stage and longitudinal change in FEV_1/FVC in both the Tissue→Airway (r=-0.19, *p*<0.001) and Airway→Tissue (r=-0.18, *p*=0.010) subtypes, and significant correlations between baseline SuStaIn stage and longitudinal change in FEV_1 %predicted in both subtypes (r=-0.12, *p*=0.015 and r=-0.20, *p*=0.004 respectively).

Analysis of GOLD1-4 patients and smoking controls assigned to SuStaIn stage 0

For completeness, we further looked at the follow-up scans of individuals assigned to SuStaIn stage 0 at baseline. We found that on average GOLD1-4 patients with a normal appearing scan at baseline progressed by 0.502 stages at follow-up (*p*<0.001), whilst smoking controls with a normal appearing scan at baseline progressed by 0.158 stages at follow-up (*p*<0.001). Of the 203 GOLD1-4 patients that had a normal appearing scan at baseline, 114 (56%) continued to have a normal scan at follow-up, 57 (28%) progressed to the Tissue→Airway subtype and 32 (16%) progressed to the Airway→Tissue subtype. Of the 1371 smoking controls that had a normal appearing scan at baseline, 1151 (84%) continued to have a normal scan at follow-up, 142 (10%) progressed to the Tissue→Airway subtype and 78 (6%) progressed to the Airway→Tissue subtype.

Supplementary Figures and Tables

Supplementary Table 1:

Demographics for the control and COPD populations from ECLIPSE used to replicate the SuStaIn subtype progression patterns.

Parameter	Control subjects	COPD subjects
Subjects, *n*	303	1809
Age (years)*, mean (SD)*	55.25 (8.90)	63.23 (7.11)
Male, *n (%)*	169 (56)	1155 (64)
Female, *n (%)*	134 (44)	654 (36)
GOLD Stage 1, *n (%)*	NA	12 (1)
GOLD Stage 2, *n (%)*		794 (44)
GOLD Stage 3, *n (%)*		761 (42)
GOLD Stage 4, *n (%)*		242 (13)
Smoking history (pack-years), *mean (SD)*	31.43 (22.17)	47.90 (26.41)

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

Supplementary Table 2:

Demographics for the control and COPD populations from the 'Inspiratory COPDGene dataset', i.e. using the reduced set of measures also available in ECLIPSE.

Parameter	Control subjects	COPD subjects
Subjects, *n*	4102	4152
Age (years)*, mean (SD)*	56.69 (8.36)	63.14 (8.61)
Male, *n (%)*	2169 (53)	2340 (56)
Female, *n (%)*	1933 (47)	1812 (44)
GOLD Stage 1, *n (%)*	NA	748 (18)
GOLD Stage 2, *n (%)*		1792 (43)
GOLD Stage 3, *n (%)*		1072 (26)
GOLD Stage 4, *n (%)*		540 (13)
Smoking history (pack-years), *mean (SD)*	37.30 (20.27)	51.55 (27.09)
Exacerbations (n/year), *mean (SD)*	0.13 (0.52)	0.64 (1.19)

Supplementary Table 3:

Relationships between breathlessness and exacerbation frequency with SuStaIn subtype and stage.

Data reported as median (IQR).

	SuStaIn Stage						
Tissue→Airway	**1-2**	**2-3**	**3-4**	**4-5**	**5-6**	**6-7**	**7-8**
Exacerbations (/year)	n= 336 0 (0-0)	n= 369 0 (0-1)	n= 461 0 (0-1)	n= 523 0 (0-1)	n= 384 0 (0-1)	n= 237 1 (0-2)	n= 44 1 (0-2)
MRC Dyspnoea	n= 336 0 (0-2)	n= 368 1 (0-3)	n= 458 2 (1-3)	n= 523 3 (1-3)	n= 384 3 (2-4)	n= 237 3 (2-4)	n= 44 3 (3-4)
Airway→Tissue	**1-2**	**2-3**	**3-4**	**4-5**	**5-6**	**6-7**	**7-8**
Exacerbations (/year)	n= 203 0 (0-0)	n= 219 0 (0-1)	n= 197 0 (0-1)	n= 169 0 (0-1)	n= 120 0 (0-1)	n= 71 0 (0-1)	n= 9 0 (0-0)
MRC Dyspnoea	n= 203 1 (0-2.5)	n= 219 2 (0-3)	n= 197 2 (0-3)	n= 169 2 (1-3)	n= 120 3 (1-3.25)	n= 71 3 (1.5-3)	n= 9 3 (1-4)

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

Supplementary Table 4:

Demographics of patients in the Tissue→Airway and Airway→Tissue subtypes from ECLIPSE. We report two sample t-tests for continuous variables, chi-squared test for categorical variables, and Mann-Whitney U-test for frequency data.

Feature	Tissue→Airway	Airway→Tissue	
Number of patients, *n (%)*	1461 (90.7)	150 (9.3)	
Male, *n (%)*	905 (61.9)	123 (82.0)	*p < 0.001*
Female, *n (%)*	556 (38.1)	27 (18.0)	
Age (years), *mean (SD)*	63.29 (6.90)	62.17 (8.00)	*p = 0.064*
BMI (kg/m^2), *mean (SD)*	26.02 (5.47)	31.30 (5.25)	*p < 0.001*
FEV_1 (% predicted), *mean (SD)*	47.29 (15.95)	49.25 (14.98)	*p = 0.151*
FEV_1/FVC ratio, *mean (SD)*	0.43 (0.11)	0.50 (0.11)	*p < 0.001*
GOLD Stage 1, *n (%)*	10 (0.7)	0 (0.0)	
GOLD Stage 2, *n (%)*	595 (40.7)	65 (43.3)	*p = 0.56*
GOLD Stage 3, *n (%)*	638 (43.7)	67 (44.7)	
GOLD Stage 4, *n (%)*	218 (14.9)	18 (12.0)	
Chronic Bronchitis, *n (%)*	480 (32.8)	59 (39.3)	*p = 0.131*
% Emphysema, *mean (SD)*	16.35 (11.65)	7.72 (6.24)	*p < 0.001*
Pi10 SRWA (mm), *mean (SD)*	4.37 (0.16)	4.76 (0.17)	*p < 0.001*

Supplementary Table 5:

Demographics of patients in the Tissue→Airway and Airway→Tissue subtypes from the Inspiratory COPDGene dataset, i.e. using the reduced set of measures that were available in ECLIPSE. We report two sample t-tests for continuous variables, chi-squared test for categorical variables, and Mann-Whitney U-test for frequency data.

Feature	Tissue→Airway	Airway→Tissue	
Number of patients, *n (%)*	3300 (92.8)	255 (7.2)	
Male, *n (%)*	1835 (55.6)	181 (71.0)	*p < 0.001*
Female, *n (%)*	1465 (44.4)	74 (29.0)	
Age (years), *mean (SD)*	63.65 (8.36)	59.68 (9.16)	*p < 0.001*
BMI (kg/m^2), *mean (SD)*	27.34 (5.92)	31.60 (6.77)	*p < 0.001*
FEV_1 (% predicted), *mean (SD)*	54.67 (22.66)	55.88 (16.63)	*p = 0.405*
FEV_1/FVC ratio, *mean (SD)*	0.50 (0.13)	0.58 (0.09)	*p < 0.001*
GOLD Stage 1, *n (%)*	496 (15.0)	17 (6.7)	
GOLD Stage 2, *n (%)*	1324 (40.1)	139 (54.5)	*p < 0.001*
GOLD Stage 3, *n (%)*	955 (28.9)	86 (33.7)	
GOLD Stage 4, *n (%)*	525 (15.9)	13 (5.1)	
Chronic Bronchitis, *n (%)*	880 (26.7)	88 (34.5)	*p = 0.008*
% Emphysema, *mean (SD)*	14.57 (12.57)	2.48 (3.69)	*p < 0.001*
Pi10 SRWA (mm), *mean (SD)*	3.70 (0.12)	3.97 (0.14)	*p < 0.001*

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

Supplementary Table 6:

Demographics (at baseline) for the control and COPD populations used to examine the relationship between SuStaIn subtype and stage against longitudinal decline in lung function.

Parameter	Control subjects	COPD subjects
Subjects, *n*	2158	1929
Age, *years*	57.86 (8.40)	62.78 (8.24)
Male, *n (%)*	1050 (49)	1065 (55)
Female, *n (%)*	1108 (51)	864 (45)
GOLD classification		
Stage 1, *n (%)*	NA	409 (21)
Stage 2, *n (%)*		928 (48)
Stage 3, *n (%)*		484 (25)
Stage 4, *n (%)*		108 (6)
Smoking history, *pack-years*	37.04 (20.45)	50.23 (25.62)
Exacerbations, *n/year*	0.14 (0.54)	0.51 (1.05)

Supplementary Table 7:

Demographics (at baseline) for the control and COPD populations used in the longitudinal validation of the SuStaIn subtypes and stages.

Parameter	Control subjects	COPD subjects
Subjects, *n*	1939	1675
Age, *years*	58.02 (8.38)	62.92 (8.22)
Male, *n (%)*	951 (49)	919 (55)
Female, *n (%)*	988 (51)	756 (45)
GOLD classification		
Stage 1, *n (%)*	NA	362 (22)
Stage 2, *n (%)*		821 (49)
Stage 3, *n (%)*		410 (24)
Stage 4, *n (%)*		82 (5)
Smoking history, *pack-years*	37.13 (20.52)	50.06 (25.18)
Exacerbations, *n/years*	0.14 (0.52)	0.48 (1.02)

AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

Supplementary Table 8:

SuStaIn subtype assignments at baseline and five-year follow-up. The subtype assignment remained unchanged in 87% of individuals.

		Follow-up	
		Tissue→Airway	Airway→Tissue
Baseline	Tissue→Airway	968	75
	Airway→Tissue	114	315

Supplementary Table 9:

SuStaIn subtype assignments at baseline and five-year follow-up, only classifying individuals who were confidently assigned to a subtype based on the probability estimated using the SuStaIn subtype progression patterns in Figure 1. At a probability threshold of 0.75, 66% of individuals can be confidently assigned to a subtype, increasing the consistency of the subtype assignments at follow-up from 87% (Supplementary Table 7) to 95%.

		Follow-up	
		Tissue→Airway	Airway→Tissue
Baseline	Tissue→Airway	742	19
	Airway→Tissue	30	186

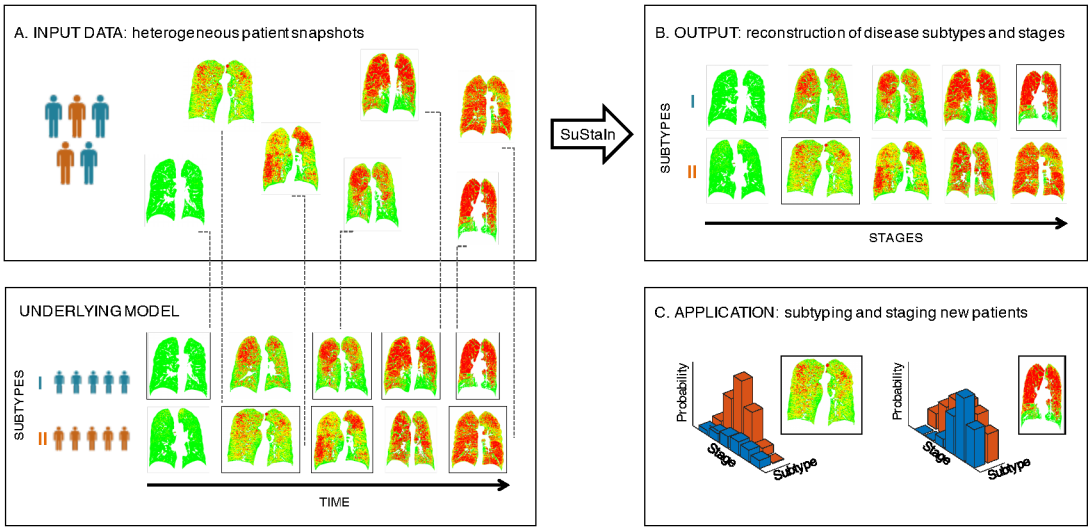
Supplementary Table 10:

Consistency of the subtype assignments in the smoking controls at five-year follow-up. The subtype assignment remains unchanged in 86% of individuals.

		Follow-up	
		Tissue→Airway	Airway→Tissue
Baseline	Tissue→Airway	346	35
	Airway→Tissue	42	145

Supplementary Figure 1: Overview of SuStaIn disease progression modelling.

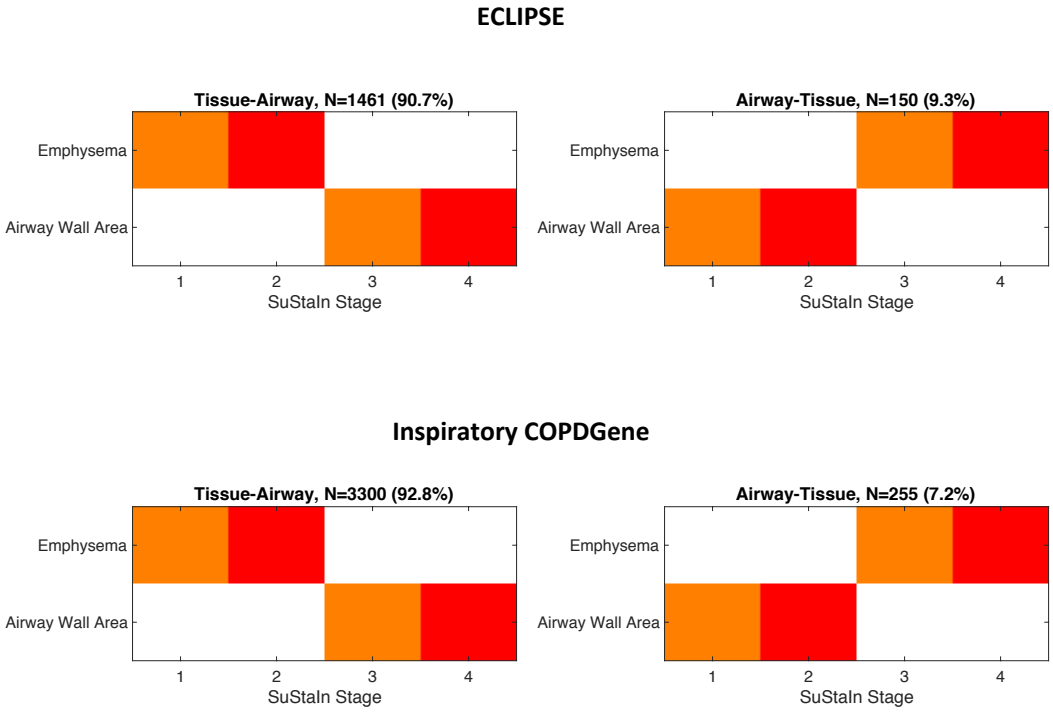
A) COPDGene is a cross-sectional study consisting of COPD patients across disease severity and control subjects. Each patient has a set of measurements from Computed Tomography scans that characterise the extent of emphysema (red), "functional" small airways disease (fSAD; yellow) and airway wall thickening. Normal lung is represented by green. The set of measurements for each patient provides a snapshot of the current disease process yet no indication of past or future disease progression. B) We applied a machine learning algorithm called SuStaIn (**Su**btype and **Sta**ge **In**ference) that identifies clusters of individuals (subtypes) with common disease progression trajectories based on cross-sectional data. Applied to the COPDGene dataset, the output is two subtypes of COPD patients that differ based on the sequence of disease progression (here labelled I and II). C) SuStaIn probabilistically assigns each patient to one of the two subtypes, and to their position (stage) on that respective disease progression trajectory. Images are representative coronal slices. This figure is adapted from the article published by Young in Nature Communications (16) to provide a conceptual illustration of the application of SuStaIn to COPD (Creative Commons Attribution 4.0 International License: http://creativecommons.org/licenses/by/4.0/).



AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

Supplementary Figure 2: Replication of SuStaIn subgroup progression patterns in an external dataset.

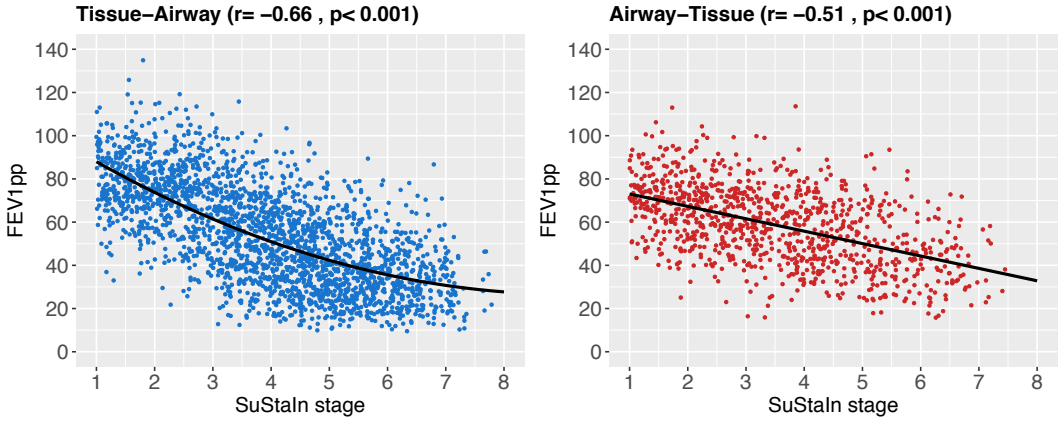
As Figure 1 but for disease progression patterns predicted by SuStaIn in ECLIPSE and COPDGene using the reduced set of inspiratory measures.



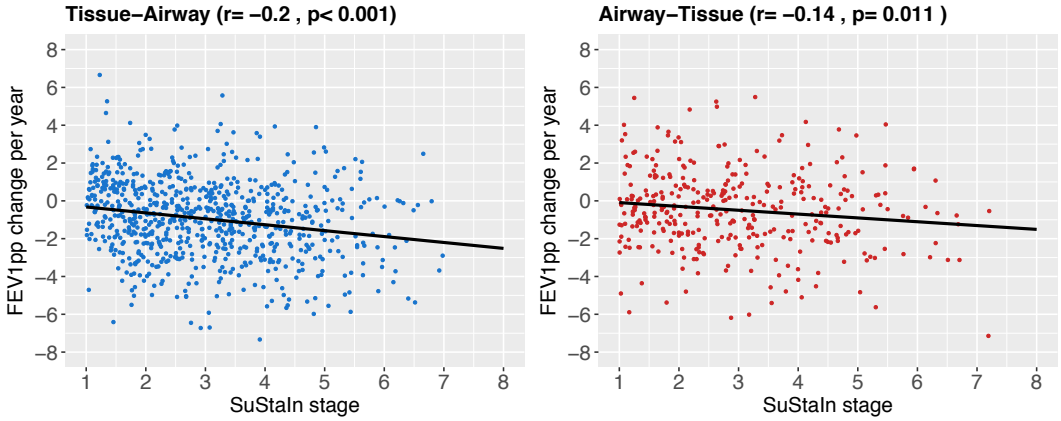
Supplementary Figure 3: Relationship between SuStaIn stage and lung function measured using FEV_1%predicted.

As Figure 2, but for FEV_1%predicted rather than FEV_1/FVC.

A.



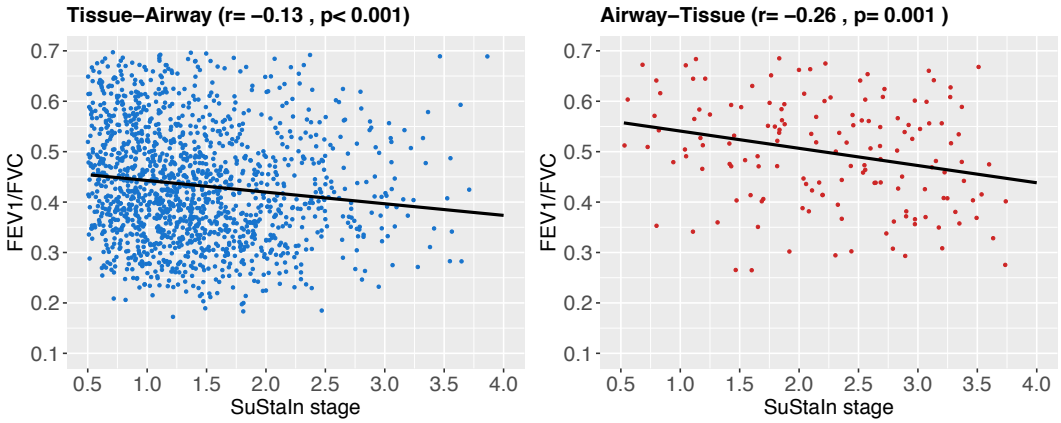
B.



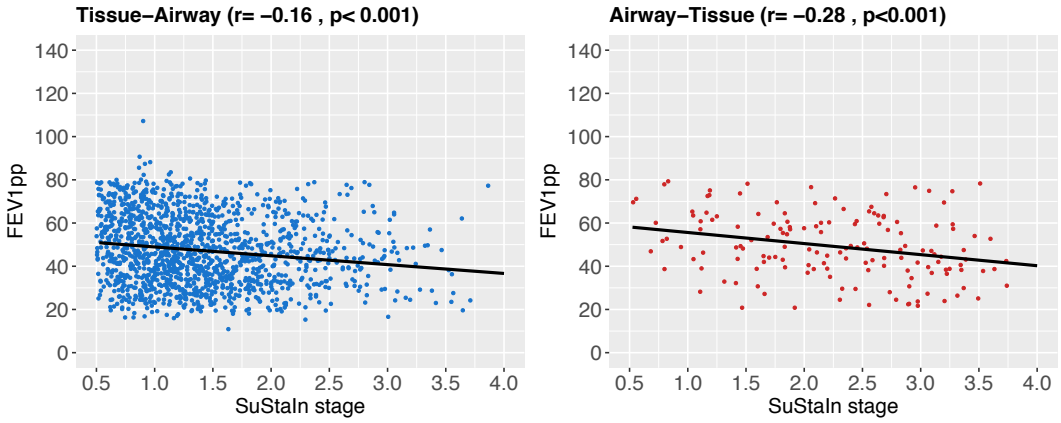
AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

Supplementary Figure 4: Relationship between lung function at baseline and SuStaIn stage in ECLIPSE measured using (A) FEV_1/FVC and (B) FEV_1%predicted.

A.

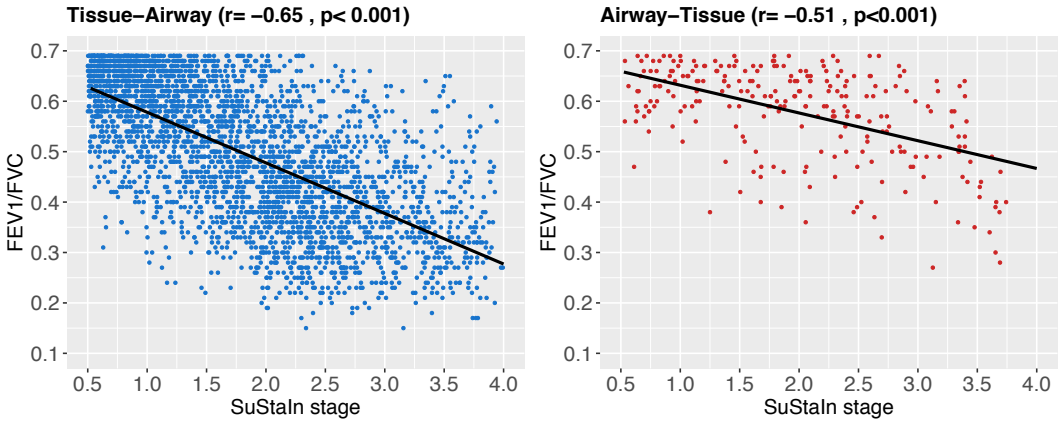


B.

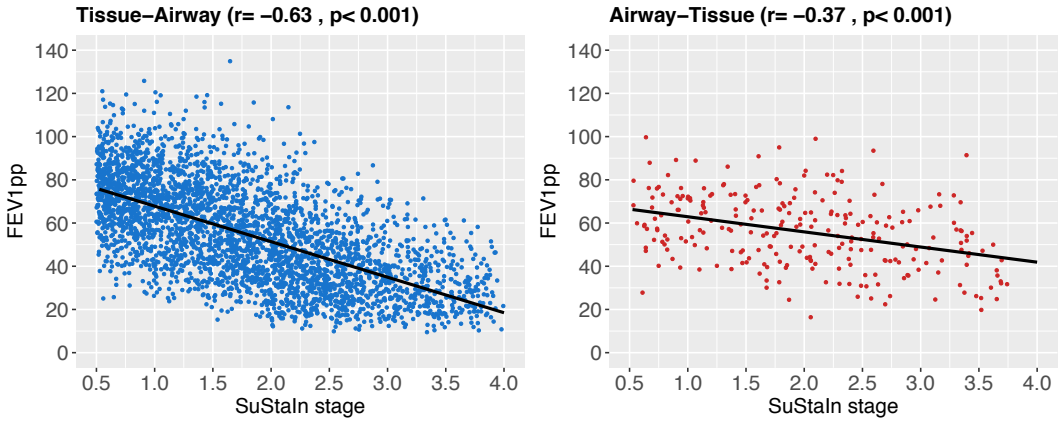


Supplementary Figure 5: Relationship between lung function at baseline and SuStaIn stage in Inspiratory COPDGene, i.e. using the reduced set of measures that had been available in ECLIPSE.

A.



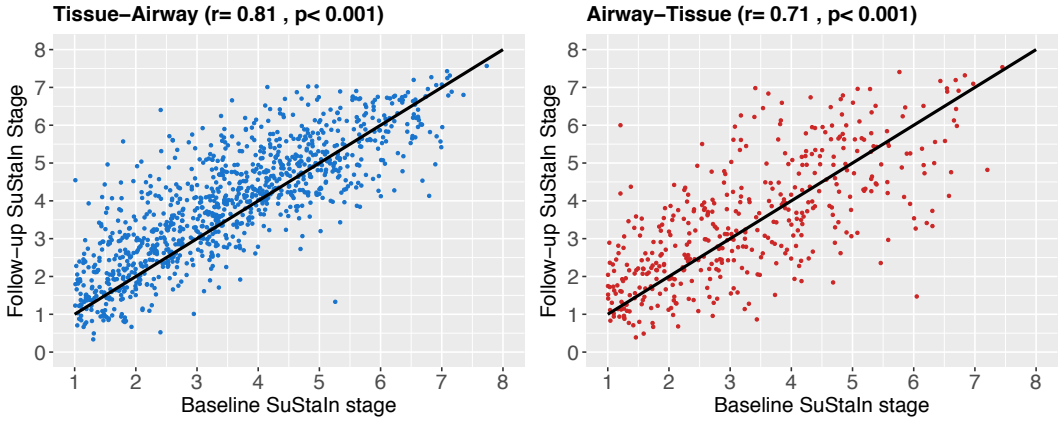
B.



AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

Supplementary Figure 6: Correlation of SuStaIn stages at baseline and follow-up in individuals, stratified by assignment to the Tissue→Airway or Airway→Tissue subtype at baseline.

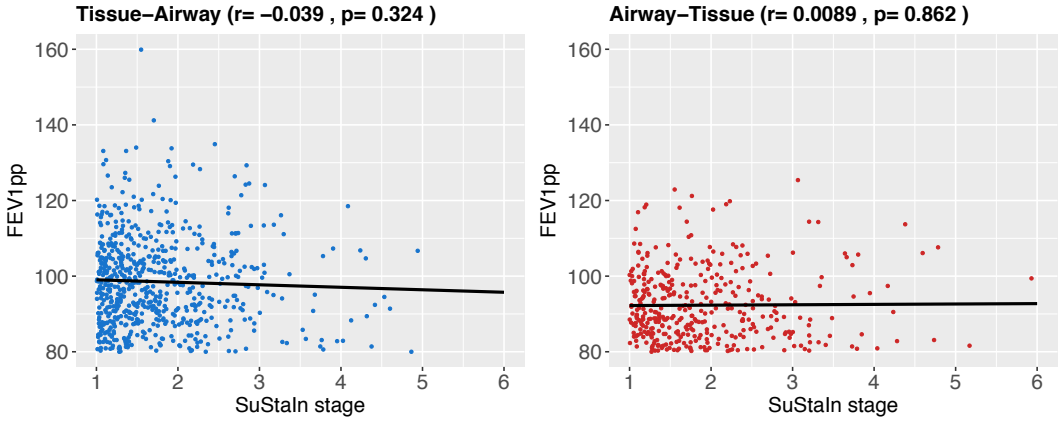
The black line shows y=x, i.e. SuStaIn stage remaining constant at follow-up. The majority of individuals progress slightly in SuStaIn stage (are above the line y=x), indicating worsening of tissue or airway damage. We were expecting to see progression which is why we have not performed Bland-Altman type analysis.



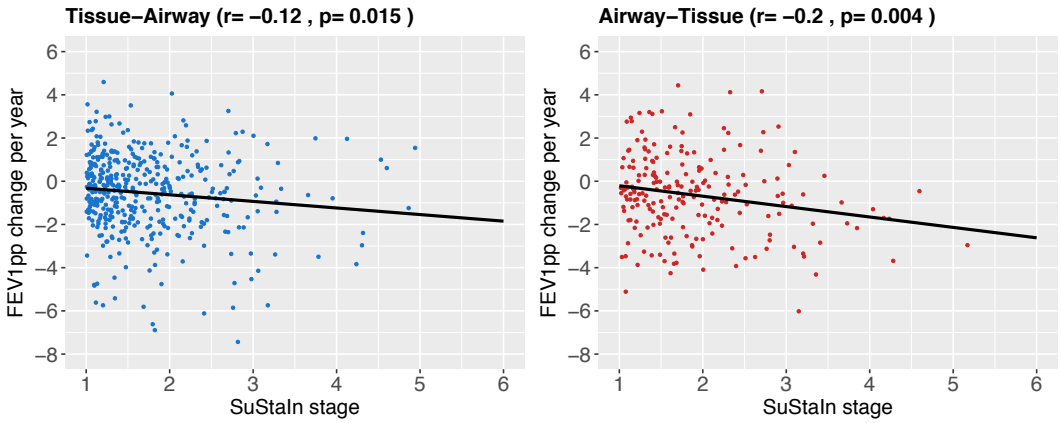
Supplementary Figure 7: Relationship between lung function and SuStaIn stage in smoking controls measured using FEV_1%predicted.

As Figure 3, but for FEV_1%predicted rather than FEV_1/FVC.

A.



B.



AJRCCM Articles in Press. Published on 28-October-2019 as 10.1164/rccm.201908-1600OC

Supplementary Figure 8: Consistency of the stage assignments in the smoking controls at five-year follow-up.

As Supplementary Figure 6, but for smoking controls rather than GOLD1-4 patients.