Supplemental Materials

Methods

Data Collection

Quantitative measures of emphysema and airway wall thickness were generated with SLICER (http://www.slicer.org) and VIDA software (VIDA Diagnostics, Iowa City, IA; http://www.vidadiagnostics.com), respectively.(1) Dyspnea and lung disease-specific quality of life measures were obtained through the use of previously validated questionnaire items.(2;3)

Cross-Validation Estimates of Cluster Stability

To assess the stability of various cluster solutions, we used five-fold cross validation to derive estimates of cluster stability as quantified by the average normalized mutual information (NMI). Normalized mutual information quantifies the dependency between variables, and it ranges from 0 (no dependency) to 1 (high dependency). Unlike Pearson correlation, NMI captures nonlinear in addition to linear dependency between variables. This procedure was carried out entirely in the training portion of the data. Four-fifths of the training sample served as the cross-validation training set (CV Train) and the remaining one-fifth of the data served as cross-validation test set (CV Test). Using the learned centroids from the CV Train set, clusters were predicted in the CV Test set and then compared to the cluster results for that fold obtained by running k-means on the entire (original) training sample. NMI quantified the degree of agreement, and the average NMI results obtained from each of the five rounds of cross-validation were used to prioritize cluster solutions by stability.

Genetic Association Testing

Genetic associations were performed in non-Hispanic white (NHW) subjects only using additive genetic coding and adjusted for principal components of genetic ancestry. A Bonferonni-adjusted statistical significance of p=0.0007 for genetic associations in the training set was defined based on 70 genetic association tests performed. The threshold for validation in the independent sample was p=0.05.

Missing Data

We employed a complete cases approach and excluded individuals from analysis who were missing data in any of the variables used for clustering, cluster association testing or interpretation. There was no difference in age of pack-years between included and excluded subjects (Supplemental Table 8). There was statistically significant but relatively minor differences in FEV_1 and FEV_1/FVC, and there were significant differences in gender and racial composition. Subjects with missing data were more likely to be female and African-American. Of the 10,300 individuals enrolled in COPDGene, 108 non-smokers were excluded from analysis, as well as 63 individuals with inadequate spirometry data. Of the remaining 10,129 individuals, 511 did not receive an inspiratory or expiratory scan. An additional 953 subjects failed quality control for either the inspiratory or expiratory scan, and 64 subjects were excluded for an FRC/TLC ratio >1. Of the remaining 8,601 subjects, 143 had incomplete data for emphysema distribution. An additional 170 individuals were excluded due to missing data for the following variables: airway wall thickness (n=4), gas

trapping (n=44), resting oxygen saturation (n=2), MMRC dyspnea score (n=11), and BODE (n=109).

Supplemental Table 1. Feature Descriptions for Comprehensive Feature Set

	Variables	Descriptions
Spirometry-Defined Variables	Post-bronchodilator FEV_1 % of predicted	observed FEV_1 (liters)/predicted FEV_1 (liters) from Hankinson equations
	FVC	observed forced vital capacity (liters)
	FEV_1/FVC	observed FEV_1 (liters)/observed FVC (liters)
	BDR as % of FEV_1	% Change in FEV_1 volume: (post-bronchodilation FEV_1 - pre-bronchodilation FEV_1, liters) /post-bronchodilation FEV_1 (liters)
	BDR as % of FVC	% Change in FVC volume: (post-bronchodilation FVC - pre-bronchodilation FVC, liters) /post-bronchodilation FVC (liters)
CT-Defined Variables	log-transformed emphysema (%LAA -950HU*)	log(%LAA -950HU + 1)
	Log ratio of Upper Third to Lower Third Emphysema	log(%LAA -950 in upper third of lung/%LAA -950 in lower third of lung)
	Segmental Wall Area %	Area of airway wall/area of entire airway for 6 selected segmental level airways (RB1, RB4, RB10, LB1, LB4, LB10)
	TLC % of Predicted	TLC measured from inspiratory CT (liters)/predicted TLC (liters)
	FRC % of Predicted	FRC measured from expiratory CT (liters)/predicted FRC (liters)
	Gas Trapping	%LAA -856HU on expiratory scan
Other Physiologic Measures	BMI	weight (kg)/height (m^2)
	Oxygen Saturation	peripheral oxygen saturation, %
* HU = Hounsfield units		

Supplemental Table 2. Cluster Associations in Training Sample for CF4 Solution Adjusting for GOLD 2007 Stage

	Training					
	C2:OR	C2:pval	C3:OR	C3:pval	C4:OR	C4:pval
Exacerbations	1.62	<0.001	2.03	<0.001	2.98	<0.001
MMRC	2.24	<0.001	2.27	<0.001	2.46	<0.001
BODE	2.45	<0.001	2.54	<0.001	4.27	<0.001
Hospitalizations/ER Visits	3.15	<0.001	3.45	<0.001	3.44	<0.001
rs7671167 (*FAM13A*)	0.98	0.81	0.96	0.59	0.89	0.53
rs1980057 (*HHIP*)	0.66	7.81E-05	0.92	0.35	0.78	0.16
rs13180 (Chr15q)	0.89	0.26	1.16	0.08	1.06	0.75
rs8034191 (Chr15q)	1.19	0.09	0.90	0.22	1.13	0.49
rs7937 (Chr 19q)	1.31	0.01	1.11	0.21	1.21	0.31
OR = odds ratio. Effect sizes represent odds ratio from logistic regression or proportional odds logistic regression in the case of Exacerbations, MMRC Score, and BODE index.						

Supplemental Table 3. Cluster Associations in Training Sample for CF4 Solution Adjusting for GOLD 2011 A-D Classes

	Training					
	C2:OR	C2:pval	C3:OR	C3:pval	C4:OR	C4:pval
Exacerbations	1.75	<0.001	1.96	<0.001	2.11	<0.001
MMRC	2.19	<0.001	2.30	<0.001	4.15	<0.001
BODE	2.55	<0.001	2.93	<0.001	22.87	<0.001
Hospitalizations/ER Visits	3.28	<0.001	3.42	<0.001	4.54	<0.001
rs7671167 (*FAM13A*)	0.96	0.64	0.93	0.35	0.87	0.24
rs1980057 (*HHIP*)	0.64	1.36E-05	0.91	0.24	0.83	0.11
rs13180 (Chr15q)	0.89	0.24	1.11	0.19	0.95	0.67
rs8034191 (Chr15q)	1.21	0.06	0.92	0.31	1.28	0.04
rs7937 (Chr 19q)	1.33	0.004	1.13	0.11	1.29	0.03
OR = odds ratio. Effect sizes represent odds ratio from logistic regression or proportional odds logistic regression in the case of Exacerbations, MMRC Score, and BODE index.						

Supplemental Table 4. Cluster Associations in Training Sample Using Cluster 2 (ULP) as Reference

Response	Group	OR (CI)	P
Exacerbations	C1	0.44 (0.38-0.51)	<0.001
	C3	1.39 (1.22-1.58)	0.01
	C4	3.93 (3.47-4.46)	<0.001
BODE	C1	0.3 (0.27-0.33)	<0.001
	C3	1.37 (1.25-1.51)	<0.001
	C4	19.75 (17.72-22.03)	<0.001
MMRC	C1	0.36 (0.33-0.39)	<0.001
	C3	1.21 (1.1-1.32)	0.04
	C4	3.87 (3.52-4.26)	<0.001
Hospitalizations/ ER Visits	C1	0.25 (0.2-0.3)	<0.001
	C3	1.24 (1.06-1.45)	0.17
	C4	2.91 (2.5-3.38)	<0.001
rs7671167	C1	0.95 (0.87-1.04)	0.59
	C3	0.91 (0.83-1)	0.33
	C4	0.9 (0.82-0.99)	0.29
rs1980057	C1	0.64 (0.58-0.7)	<0.001
	C3	1.42 (1.28-1.56)	<0.001
	C4	1.21 (1.1-1.34)	0.05
rs13180	C1	0.82 (0.75-0.9)	0.03
	C3	1.29 (1.17-1.42)	0.01
	C4	1.01 (0.91-1.11)	0.93
rs8034191	C1	1.33 (1.21-1.46)	0.002
	C3	0.76 (0.69-0.84)	0.01
	C4	1.1 (1-1.22)	0.31
rs7937	C1	1.3 (1.18-1.42)	0.004
	C3	0.9 (0.81-0.99)	0.26
	C4	0.92 (0.83-1.02)	0.41
Reference cluster -= C2 (Upper Lobe Predominant)			

Supplemental Table 5. Cluster Associations in Training Sample Using Cluster 3 (AP) as Reference

Response	Group	OR (CI)	P
Exacerbations	C1	0.32 (0.28-0.36)	<0.001
	C2	0.72 (0.63-0.82)	0.01
	C4	2.82 (2.56-3.12)	<0.001
BODE	C1	0.22 (0.2-0.23)	<0.001
	C2	0.73 (0.66-0.8)	<0.001
	C4	14.37 (13.06-15.81)	<0.001
MMRC	C1	0.29 (0.27-0.32)	<0.001
	C2	0.83 (0.76-0.91)	<0.001
	C4	3.21 (2.95-3.48)	<0.001
Hospitalizations/ ER Visits	C1	0.2 (0.17-0.24)	<0.001
	C2	0.81 (0.69-0.94)	<0.001
	C4	2.34 (2.08-2.64)	<0.001
rs7671167	C1	0.32 (0.28-0.36)	0.05
	C2	0.72 (0.63-0.82)	0.33
	C4	2.82 (2.56-3.12)	0.86
rs1980057	C1	0.22 (0.2-0.23)	0.22
	C2	0.73 (0.66-0.8)	<0.001
	C4	14.37 (13.06-15.81)	0.05
rs13180	C1	0.29 (0.27-0.32)	0.62
	C2	0.83 (0.76-0.91)	0.01
	C4	3.21 (2.95-3.48)	0.004
rs8034191	C1	0.2 (0.17-0.24)	0.66
	C2	0.81 (0.69-0.94)	0.01
	C4	2.34 (2.08-2.64)	<0.001
rs7937	C1	0.32 (0.28-0.36)	0.03
	C2	0.72 (0.63-0.82)	0.26
	C4	2.82 (2.56-3.12)	0.75
Reference cluster -= C3 (Airway Predominant)			

Supplemental Table 6. Cluster Associations in Training Sample Using Cluster 4 (SE) as Reference

Response	Group	OR (CI)	P
Exacerbations	C1	0.11 (0.1-0.13)	<0.001
	C2	0.25 (0.22-0.29)	<0.001
	C3	0.35 (0.32-0.39)	<0.001
BODE	C1	0.02 (0.01-0.02)	<0.001
	C2	0.05 (0.05-0.06)	<0.001
	C3	0.07 (0.06-0.08)	<0.001
MMRC	C1	0.09 (0.08-0.1)	<0.001
	C2	0.26 (0.23-0.28)	<0.001
	C3	0.31 (0.29-0.34)	<0.001
Hospitalizations/ ER Visits	C1	0.08 (0.07-0.1)	<0.001
	C2	0.34 (0.3-0.4)	<0.001
	C3	0.43 (0.38-0.48)	<0.001
rs7671167	C1	0.84 (0.78-0.91)	0.95
	C2	0.9 (0.82-0.99)	0.10
	C3	0.99 (0.91-1.07)	0.19
rs1980057	C1	0.79 (0.73-0.85)	0.01
	C2	1.21 (1.1-1.34)	0.24
	C3	0.86 (0.8-0.93)	<0.001
rs13180	C1	0.82 (0.76-0.88)	<0.001
	C2	1.01 (0.91-1.11)	0.90
	C3	0.79 (0.73-0.86)	0.01
rs8034191	C1	1.5 (1.39-1.61)	0.004
	C2	1.1 (1-1.22)	0.13
	C3	1.42 (1.31-1.54)	0.03
rs7937	C1	1.2 (1.12-1.29)	0.38
	C2	0.92 (0.83-1.02)	0.38
	C3	1.03 (0.95-1.11)	<0.001
Reference cluster -= C3 (Airway Predominant)			

Supplemental Table 7. Comparison of Clustering Assignment in GOLD 2-4 Subjects from Clustering Performed in All Subjects and Clustering Performed in GOLD 2-4 Only

		Classification of GOLD 2-4 Subjects in All Subjects Analysis			
		C1	C2	C3	C4
Classification of GOLD 2-4 Subjects in Case Only Clustering	C1	107	179	0	0
	C2	0	0	379	1
	C3	0	7	8	827

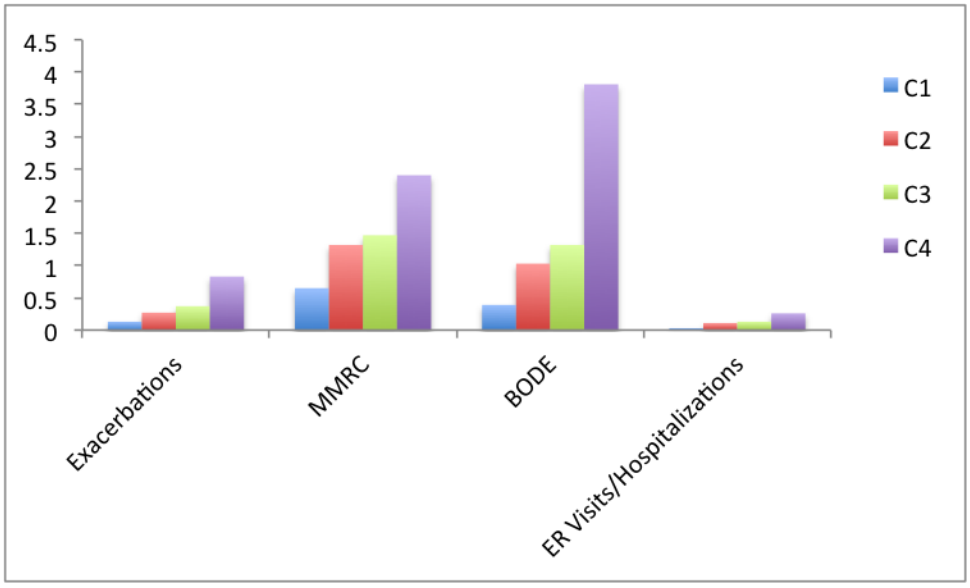
Supplemental Table 8. Characteristics of Subjects Excluded for Missing Data Compared to Analyzed Subjects

Characteristic	Subjects with Complete Data	Subjects Excluded for Missing Data	P-value
N	8288	1904	---
Gender, % Female	0.46	0.50	0.01
Race, % African-American	0.32	0.42	<0.001
Age	58.9 (14.4)	58.4 (14.8)	0.15
Pack Years	39.5 (27.8)	38.3 (26.9)	0.15
FEV_1, % of predicted	81.0 (34.3)	78.8 (39.6)	<0.001
FEV_1/FVC	0.72 (0.20)	0.71 (0.23)	0.006
63 excluded subjects were missing data for the analyzed characteristics above. P-value obtained by Pearson's chi-square test (proportions) or Wilcoxon rank sum test.			

Supplemental Table 9. Detailed Smoking Information for Training and Validation Data

Characteristic	Training	Validation
N	4187	4101
Pack-Years, median (IQR)	39.3 (28.0)	39.7 (27.0)
Smoking Intensity, median (IQR)	20 (10)	20 (10)
Smoking Duration in years	36.4 (10.1)	36.4 (10.2)
Age Started Smoking	16.9 (4.5)	16.8 (4.7)

Supplemental Figure 1. Average Number of Exacerbations of Past Year, MMRC Score, BODE Index, and Number of ER Visits and Hospitalizations by Cluster



Reference List

(1) Hu S, Hoffman EA, Reinhardt JM. Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images. IEEE Trans Med Imaging 2001 June;20(6):490-8.

(2) Jones PW, Quirk FH, Baveystock CM, et al. A self-complete measure of health status for chronic airflow limitation. The St. George's Respiratory Questionnaire. Am Rev Respir Dis 1992 June;145(6):1321-7.

(3) Bestall JC, Paul EA, Garrod R, et al. Usefulness of the Medical Research Council (MRC) dyspnoea scale as a measure of disability in patients with chronic obstructive pulmonary disease. Thorax 1999 July;54(7):581-6.