ORIGINAL ARTICLE

Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema

Peter J Castaldi,[1,2] Jennifer Dy,[3] James Ross,[4] Yale Chang,[3] George R Washko,[5] Douglas Curran-Everett,[6] Andre Williams,[6] David A Lynch,[7] Barry J Make,[8] James D Crapo,[8] Russ P Bowler,[8] Elizabeth A Regan,[8] John E Hokanson,[9] Greg L Kinney,[9] Meilan K Han,[10] Xavier Soler,[11] Joseph W Ramsdell,[11] R Graham Barr,[12] Marilyn Foreman,[13] Edwin van Beek,[14] Richard Casaburi,[15] Gerald J Criner,[16] Sharon M Lutz,[17] Steven I Rennard,[18,19] Stephanie Santorico,[20] Frank C Sciurba,[21] Dawn L DeMeo,[1,5] Craig P Hersh,[1,5] Edwin K Silverman,[1,5] Michael H Cho[1,5]

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/thoraxjnl-2013-203601).

For numbered affiliations see end of article.

Correspondence to
Dr Peter Castaldi, Channing Division of Network Medicine, Brigham and Women's Hospital, 181 Longwood Ave., Boston, MA 02115, USA; peter.castaldi@channing.harvard.edu

Received 19 March 2013
Revised 17 January 2014
Accepted 22 January 2014
Published Online First
21 February 2014

ABSTRACT

Background There is notable heterogeneity in the clinical presentation of patients with COPD. To characterise this heterogeneity, we sought to identify subgroups of smokers by applying cluster analysis to data from the COPDGene study.

Methods We applied a clustering method, k-means, to data from 10 192 smokers in the COPDGene study. After splitting the sample into a training and validation set, we evaluated three sets of input features across a range of *k* (user-specified number of clusters). Stable solutions were tested for association with four COPD-related measures and five genetic variants previously associated with COPD at genome-wide significance. The results were confirmed in the validation set.

Findings We identified four clusters that can be characterised as (1) relatively resistant smokers (ie, no/mild obstruction and minimal emphysema despite heavy smoking), (2) mild upper zone emphysema-predominant, (3) airway disease-predominant and (4) severe emphysema. All clusters are strongly associated with COPD-related clinical characteristics, including exacerbations and dyspnoea ($p<0.001$). We found strong genetic associations between the mild upper zone emphysema group and rs1980057 near *HHIP*, and between the severe emphysema group and rs8034191 in the chromosome 15q region ($p<0.001$). All significant associations were replicated at $p<0.05$ in the validation sample (12/12 associations with clinical measures and 2/2 genetic associations).

Interpretation Cluster analysis identifies four subgroups of smokers that show robust associations with clinical characteristics of COPD and known COPD-associated genetic variants.

Key messages

What is the key question?
- Can distinct subtypes of pulmonary damage be identified in smokers?

What is the bottom line?
- Cluster analysis in the COPDGene study identifies four clusters of smokers with distinct patterns of airway wall thickness, emphysema and emphysema distribution, and these subtypes show strong association with relevant clinical measures and known COPD-associated genetic variants.

Why read on?
- This paper demonstrates robust clustering results that identify clinically important subgroups of smokers in the largest COPD subtyping study to date.

BACKGROUND

The clinical presentation of COPD is heterogeneous. Smoking-related damage manifests as airway wall thickening, loss of small airways, emphysematous lung destruction and a range of extrapulmonary manifestations. However, these specific manifestations may vary in individual smokers. COPD heterogeneity has been broadly characterised as emphysema-predominant and airway-predominant disease,[1 2] and the varying amounts of airway obstruction and emphysema present in an individual can be described with quantitative CT measures. In addition to the emphysema-airway characterisation, additional subtypes have been proposed in an effort to further refine our understanding of smoking-related lung damage. Some of these, such as upper lobe-predominant emphysema and the 'frequent-exacerbator' subtype, have important consequences for clinical management.[3–5]

The most widely accepted current definition of COPD is that of the Global Initiative for Chronic Obstructive Lung Disease (GOLD 2007).[6] Based primarily on spirometry, GOLD 2007 confirms the diagnosis of COPD based on FEV_1/FVC and classifies disease severity based on FEV_1. This simplicity

To cite: Castaldi PJ, Dy J, Ross J, *et al*. *Thorax* 2014;**69**:415–422.

has arguably led to improved recognition, diagnosis and treatment of the disease.[6 7] However, the GOLD 2007 criteria do not fully describe the heterogeneity of COPD,[8 9] and the most recent GOLD 2011 criteria add clinical characteristics to define new classes.[10] GOLD provides clear cut-offs to define presence/absence of COPD based on FEV_1 and FEV_1/FVC; however, spirometric measures, as well as associated CT scan characteristics such as emphysema, have a continuous distribution in the population, indicating that the smoking-related damage characteristic of COPD is likely a continuous process that can also be present in subjects who have not yet developed airflow obstruction meeting standard criteria.

One rationale for the simplicity of the GOLD 2007 criteria is that there is substantial overlap between different disease characteristics and among proposed subtypes. It is a challenge to synthesise the various smoking-related subtypes proposed in the literature because subtypes may overlap or be defined in ways that are not complementary. In an effort to derive data-driven COPD classifications, investigators have recently employed unsupervised machine learning approaches.[11–13] The benefit of such approaches is that they employ quantitative methods to define subtypes, but the challenge in applying these approaches for clinical subtype identification is that they are designed primarily for data exploration rather than specific hypothesis testing. As a result, the generalisability and reproducibility of machine-learned COPD subtype classifications in independent data samples has been largely unexplored.

We hypothesised that k-means, a widely used unsupervised clustering method, would identify novel, clinically relevant subtypes when applied to quantitative chest CT, spirometric and clinical measures from the COPDGene study. The COPDGene study is a large epidemiological and genetic study of over 10 000 current and former smokers with and without COPD that includes demographic and clinical information, spirometry, genome-wide single-nucleotide polymorphisms (SNP) genotyping data, and inspiratory and expiratory CT scans. We specified a priori a set of clinically relevant clinical and genetic variables that would be used only to evaluate and interpret (but not to generate) clusters, and we split our data into a training and validation set to provide rigorous assessment of the reproducibility of our results.

METHODS

Study population

The COPDGene study has previously been described in detail.[14] Briefly, between 2007 and 2011, 10 192 non-Hispanic Whites (NHW, n=6784) and African-American (AA, n=3408) smokers were enrolled in a multicentre study designed to investigate the genetic and epidemiological associations of COPD. Subjects with respiratory disease other than asthma, COPD or emphysema were excluded. All subjects had blood collected for genetic analysis, and they completed questionnaires, spirometry and chest CT scans. The institutional review boards of all participating centres approved the COPDGene study, and written informed consent was obtained from all subjects.

Sample splitting, feature selection and clustering

In order to assess the validity of cluster solutions, the COPDGene data were randomly split into equally sized training and validation sets. All subsequent model building was conducted in the training data, with the validation set used only for the validation of cluster characteristics and associations.

In this paper, we use the term *feature* to refer to a variable that is used as an input for clustering. A set of continuous input features for k-means clustering, hereafter referred to as the *comprehensive feature set*, was selected to represent key aspects of COPD-related physiology, particularly spirometry and quantitative chest CT data. Detailed feature descriptions are included in online supplemental table 1.

Since the quality of clustering results can be improved by eliminating uninformative features,[15] we used two approaches to generate filtered subsets of the comprehensive feature set using a *top factor* and a *core feature* approach. In the top factor approach, we identified factors that individually accounted for 5% or greater of the overall variance and then selected the top loading feature for each factor to constitute the *top factor set*. In the *core feature* approach, we considered spirometric and quantitative CT variables from the comprehensive feature set and filtered these variables based on Pearson correlation such that no variables in the core variable set were correlated at 0.7 or greater.

k-means clustering was performed using the k-means function in version 2.13,[16] and the stability of cluster solutions was assessed by average normalised mutual information (NMI) as assessed by fivefold cross-validation in the training data. Data were scaled and centred prior to clustering.

Evaluation of cluster significance in the training sample

To prioritise and evaluate the clinical relevance of clustering solutions, we specified a priori a set of COPD-related measures and genetic variables to test for association with cluster membership. These variables were not used as inputs to the clustering process. The COPD-related measures were BODE index, MMRC dyspnoea score, number of COPD exacerbations over the previous year and self-report of a lung-related emergency room visit or hospitalisation over the previous year (lung-related healthcare use). COPD-related measures were related to cluster membership using logistic regression or ordinal logistic regression as appropriate.

Genetic variables consisted of five SNPs previously associated with COPD at genome-wide significance (COPD SNPs—rs7671167,[17] rs1980057,[18 19] rs13180,[19] rs8034191,[19] rs7937[20]). Genetic associations with cluster membership were tested by logistic regression using the 'healthiest' cluster (ie, with the highest average FEV_1) as the reference, and comparisons with other cluster as reference were also performed. As a sensitivity analysis, all cluster associations were evaluated with and without adjustment for study centre, GOLD 2007 stage and GOLD 2011 classifications by including these as covariates in separate regression models.

Validation of cluster characteristics, clinical and genetic associations

After prioritising cluster solutions in the training sample by cluster stability and strength of clinical and genetic associations, a single clustering result was selected for independent validation. Clusters were assigned in the validation sample by assigning each subject to the closest cluster centre using the centres learned by the k-means algorithm in the training sample. T-tests were used to test for differences in average cluster characteristics between the training and validation samples, and cluster associations with the prespecified clinical and genetic measures were examined as described above. Additional details are included in the online supplement.

RESULTS

The characteristics of the training and validation samples are shown in table 1, and the samples are comparable. The

Table 1 Baseline characteristics of the training and validation data

	Training	Validation
N	4187	4101
Age	59.5 (9.0)	59.7 (9.0)
Gender, % female	46.7	45.9
Race, % African-American	32.0	31.4
FEV_1, % of predicted	76.9 (25.2)	77.1 (25.2)
FEV_1/FVC	0.67 (0.16)	0.67 (0.16)
Pack-years, median (IQR)	39.3 (28.0)	39.7 (27.0)
BMI	28.9 (6.3)	28.9 (6.1)
Emphysema at −950HU, median (IQR)	1.8 (5.8)	2.0 (6.1)
Upper/lower emphysema ratio (IQR)	0.8 (1.1)	0.8(1.2)
Segmental airway wall thickness	61.4 (3.2)	61.4 (3.3)
Upper/lower lobe emphysema difference (IQR)	−0.17 (2.0)	−0.14 (2.2)
Gas trapping (IQR)	14.5 (24.8)	14.7 (25.3)
GOLD unclassifiable*, %	12.0	12.6
Smoking controls, %	43.8	43.8
GOLD 1, %	8.3	7.7
GOLD 2, %	19.2	19.4
GOLD 3, %	11.3	11.3
GOLD 4, %	5.4	5.3

Values are mean (SD) unless otherwise noted.
*GOLD unclassifiable refers to subjects with a FEV_1% predicted <80 but FEV_1/FVC >0.7.

difference in sample size between the training and validation samples is due to differences in missing data (see online supplement).

Defining feature subsets

Factor analysis on the comprehensive feature set identified four factors that individually accounted for at least 5% of the variance in the data. Features with the top loadings for these factors were functional residual capacity (FRC) % predicted, FEV_1% predicted, CT-quantified emphysema at −950 Hounsfield units (HU) and bronchodilator responsiveness as a % of FEV_1. For the core feature set, correlation filtering yielded a set of four features—FEV_1% predicted, CT-quantified emphysema, segmental wall area% and emphysema distribution (log ratio of upper third/lower third emphysema).

Prioritising clustering solutions by cluster stability

Cluster stability for the three feature sets is shown in figure 1. Seven stable clustering solutions with NMI > 0.9 were prioritised for further evaluation. We examined the clinical and genetic associations of these seven solutions in the training sample. For the comprehensive and top factor feature sets, the highest stability results were for k=2. These solutions largely replicated the traditional COPD case–control distinction and were likely driven by the case–control design and recruitment strategy of COPDGene.

For the core feature set, highly stable clustering was observed for a range of k from 2 to 5. Figure 2 shows the characteristics of the clustering features for the k=3 to k=5 solutions and the pattern in which clusters emerge as k increases. Based on the strong pattern of cluster-specific clinical and genetic associations, the k=4 core feature (CF4) solution was selected for further validation.

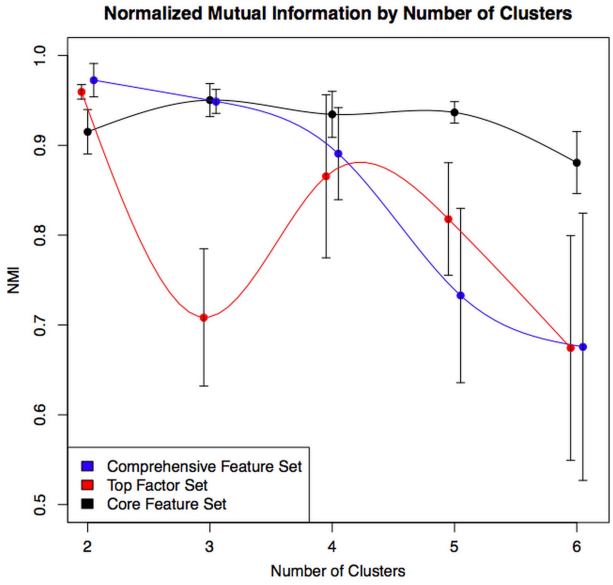

Figure 1 Cluster stability as measured by average normalised mutual information (NMI) by number of clusters across the three input feature sets. High NMI values indicate high cluster stability. For the comprehensive and top factor feature sets, stability is greatest for the k=2 solution. For the core feature set, very high stability is observed up to k=5. Dots and SEs bars represent average NMI and SEs over fivefold cross-validation, respectively. Dots are slightly offset to improve visualisation.

Cluster characteristics

Cluster characteristics for the CF4 solution are shown in table 2. The four clusters can be characterised as low susceptibility smokers, mild upper zone emphysema-predominant, airway-predominant and severe emphysema.

k=3

	C1	C2	C3
N	1833	1347	1007
FEV_1	94.7	76.0	46.0
% Emphysema	2.6	1.3	18.3
Emph Upper/Lower	1.4	0.9	3.9
Segmental WA %	58.9	63.9	62.7

1516 317 1114 843

k=4

	C1	C2	C3	C4
N	1598	623	1122	844
FEV_1	95.3	81.9	74.9	41.2
% Emphysema	2.6	3.3	1.3	20.5
Emph Upper/Lower	0.7	6.7	1.6	2.2
Segmental WA %	58.8	61.5	64.1	62.7

838 760 520 736 792

k=5

	C1	C2	C3	C4	C5
N	934	1147	531	778	797
FEV_1	96.1	91.0	80.9	67.6	40.7
% Emphysema	4.1	1.0	2.9	1.8	21.3
Emph Upper/Lower	1.0	0.5	7.5	0.7	2.3
Segmental WA %	58.0	60.7	62.0	65.1	62.6

Figure 2 Average values of clustering features from core feature set solutions k=3 through 5. Arrows indicate relationships between these k-means derived clusters that share large numbers of individuals.

Table 2 Cluster characteristics in training and validation data for core feature set cluster solution, k=4

	Training sample				Validation sample			
	C1: mean	C2: mean	C3: mean	C4: mean	C1: mean	C2: mean	C3: mean	C4: mean
N	1598	623	1122	844	1595	620	1060	826
Age	58.9	58.0*	56.8	65.4	58.7	58.9*	57.3	65.4
Gender, % female	0.44	0.53	0.52	0.40	0.43	0.51	0.53	0.40
Race, % African-American	0.30	0.46	0.37	0.19	0.29	0.45	0.37	0.17
FEV_1, per cent of predicted	95.3	81.9	74.9	41.2	95.7	81.6	73.8	42.0
FEV_1/FVC	0.76	0.70	0.71	0.42	0.76	0.69	0.71	0.42
BMI	28.7	27.9	31.4*	26.7	28.3	27.6	32.0*	26.8
Pack years	38.0	45.8	42.8	56.8	38.3	46.9	43.1	55.9
Emphysema at −950HU	2.6	3.3	1.3	20.5	2.7	3.6	1.4	20.7
Segmental airway wall thickness	58.8	61.5	64.1	62.7	58.8	61.4	64.2	63.0
Upper/lower emphysema ratio	0.7	6.7	0.6	2.2	0.7	8.3	0.6	2.3
Upper/lower emphysema difference	−0.3	1.4	−0.3	2.6	−0.3	1.7	−0.3	2.9
Gas trapping†	12.9	16.5	13.4	52.1	13.1	17.3	13.3	52.7

Values represent the mean of each variable for each cluster unless otherwise specified.
Only the variables shown in bold were used as input variables for the primary clustering solution (CF4).
*p Value comparing mean in training to validation <0.05 for t test.
†%LAA using −856 Hounsfield unit threshold on expiratory CT scan.
C1, relatively resistant smokers; C2, mild upper zone-predominant emphysema; C3, airway-predominant; C4, severe emphysema.

Cluster 1: relatively resistant smokers

Cluster 1 represents 38% of the COPDGene training sample and is characterised by heavy smoking exposure with no or minimal airflow obstruction, as well as lower emphysema ($p<0.001$ for comparison with clusters 2 and 4) and airway wall thickness ($p<0.001$ for all cluster comparisons) compared with the more severely affected clusters. The majority of individuals in the relatively resistant cluster are control smokers or GOLD stage 1 (figure 3).

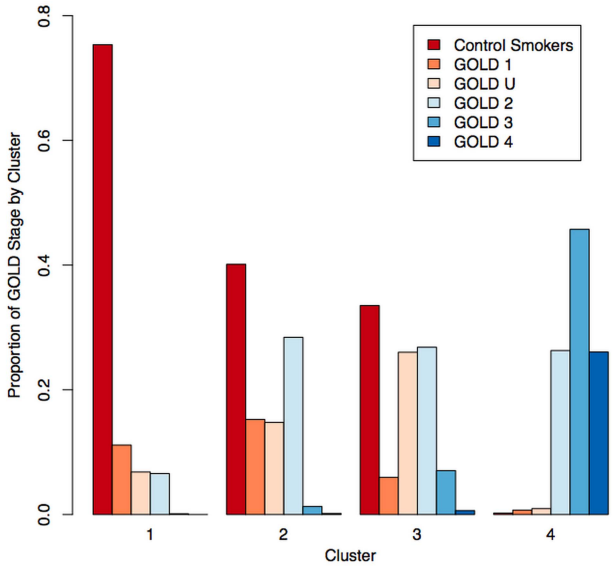

Figure 3 Proportion of individuals in each Global Initiative for Chronic Obstructive Lung Disease (GOLD 2007) stage by core feature set clustering solution (k=4). Cluster 1 (relatively smoking resistant individuals) consists largely of control smokers and GOLD 1–2 individuals. Cluster 4 (severe emphysema) consists largely of GOLD 2–4 individuals. Clusters 2 and 3 (upper zone emphysema and airway-predominant) consist largely of control smokers, GOLD 1–2 and GOLD unclassifiable (GOLD U) individuals.

Cluster 2: mild upper zone-predominant emphysema

Cluster 2 represents 15% of the training sample and is characterised by mild airflow obstruction and mild emphysema with marked upper zone-predominance (p values compared with other clusters <0.001). The average amount of emphysema in this group is modest (mean emphysema = 3.31%), though the range is broad and nearly a quarter of this cluster has greater than 5% emphysema. As shown in figure 3, most of the individuals in the mild upper zone emphysema cluster are control smokers or GOLD stages 1–2, with 15% unclassifiable by GOLD criteria.

When compared with the relatively resistant cluster, this cluster was more likely to experience an exacerbation, have a higher MMRC dyspnoea score and BODE index, and more likely to have used the emergency room or been admitted to the hospital for a respiratory issue (table 3). The NHW subjects in this group show a strong genetic association with rs1980057 near the *HHIP* gene ($p=4.4\times10^{-6}$). This cluster has a higher proportion of AAs than the airway-predominant and severe emphysema clusters ($p<0.001$) and a higher proportion of women compared with the relatively smoking-resistant and severe emphysema clusters ($p<0.001$).

Cluster 3: airway-predominant disease

Cluster 3 represents 27% of the training sample and is characterised by thicker airway walls, the lowest average emphysema of all clusters, and high BMI ($p<0.001$ for all measures). The overall distribution of GOLD 2007 stages in this group is similar to the mild upper zone emphysema cluster, with the exception of a higher proportion of GOLD stage 3 and unclassifiable individuals (figure 3).

This cluster is more likely than the relatively smoking-resistant cluster to report COPD exacerbations and lung-related healthcare use, and they have higher MMRC score and BODE index (table 3). It has a significantly higher proportion of women than the smoking-resistant and severe emphysema clusters ($p<0.001$), and the overall strength of genetic associations between this cluster and COPD SNPs is weak.

Table 3 Cluster associations with COPD-related measures and COPD SNPs in training and validation data for core feature set cluster solution, k = 4

	Training			Validation		
	C2: OR (95% CI)	C3: OR (95% CI)	C4: OR (95% CI)	C2: OR (95% CI)	C3: OR (95% CI)	C4: OR (95% CI)
Exacerbations	2.27 (1.97 to 2.61)***	3.16 (2.82 to 3.55)***	8.93 (7.97 to 10.01)***	2.17 (1.90 to 2.49)***	2.66 (2.38 to 2.98)***	7.80 (7.00 to 8.70)***
MMRC	2.81 (2.57 to 3.07)***	3.39 (3.14 to 3.66)***	10.88 (10.00 to 11.83)***	2.02 (2.01 to 2.40)***	3.00 (2.78 to 3.23)***	10.07 (9.26 to 10.94)***
BODE	3.37 (3.06 to 3.70)***	4.63 (4.27 to 5.02)***	66.52 (60.06 to 73.67)***	2.62 (2.38 to 2.88)***	4.23 (3.90 to 4.58)***	52.64 (47.62 to 58.19)***
Hospitalisations/ER visits	4.07 (3.34 to 4.95)***	5.05 (4.24 to 6.01)***	11.82 (9.98 to 14.00)***	3.05 (2.53 to 3.68)***	4.13 (3.52 to 4.86)***	8.03 (6.86 to 9.39)***
rs7671167 (FAM13A)	0.95 (0.87 to 1.04)^{NS}	0.87 (0.81 to 0.93)*	0.84 (0.78 to 0.91)*	1.01 (0.92 to 1.10)^{NS}	0.89 (0.83 to 0.95)^{NS}	0.91 (0.85 to 0.98)^{NS}
rs1980057 (HHIP)	0.64 (0.58 to 0.70)***	0.92 (0.85 to 0.98)^{NS}	0.79 (0.73 to 0.85)***	0.80 (0.73 to 0.87)*	1.09 (1.01 to 1.17)^{NS}	0.74 (0.69 to 0.80)***
rs13180 (Chr15q25)	0.82 (0.75 to 0.90)*	1.04 (0.96 to 1.11)^{NS}	0.82 (0.76 to 0.88)**	0.72 (0.66 to 0.79)***	0.99 (0.92 to 1.07)^{NS}	0.82 (0.76 to 0.88)**
rs8034191 (Chr15q25)	1.33 (1.21 to 1.46)**	1.03 (0.96 to 1.11)^{NS}	1.50 (1.39 to 1.61)***	1.30 (1.19 to 1.43)**	0.89 (0.83 to 0.96)^{NS}	1.17 (1.09 to 1.26)*
rs7937 (Chr 19q13)	1.30 (1.18 to 1.42)**	1.16 (1.08 to 1.24)*	1.20 (1.12 to 1.29)*	1.08 (0.99 to 1.18)^{NS}	1.06 (0.99 to 1.14)^{NS}	1.46 (1.36 to 1.57)***

Effect sizes represent OR from logistic regression or proportional odds logistic regression in the case of exacerbations, MMRC score and BODE index.
In all instances, cluster 1 (ie, the cluster with the highest mean FEV1% of predicted) serves as the reference.
Effect allele for rs7671167 = C, rs1980057 = T, rs13180 = C, rs8034191=C, rs7937=T.
*0.01<p≤0.05; **0.001<p≤0.01; ***p≤0.001, NS p>0.05.

Cluster 4: severe emphysema

Cluster 4 represents 20% of the sample and is characterised by high emphysema, gas trapping and severe airflow obstruction ($p<0.001$ for all measures). This group consists primarily of GOLD 2–4 individuals. It has the lowest BMI, highest lifetime pack-years exposure, oldest average age ($p<0.001$ for all measures) and it is the most severely affected cluster in terms of COPD-related measures. The effect sizes of the associations between the severe emphysema cluster and the four COPD-related clinical variables are roughly twice as large as those observed for the upper zone emphysema and airway-predominant clusters.

This cluster is strongly associated with rs1980057 ($p=0.001$) near *HHIP* and rs8034191 ($p=5\times10^{-8}$) in the chromosome 15q locus that includes the nicotinic receptor genes *CHRNA3* and *CHRNA5* as well as *IREB2* (table 3). It has a significantly higher proportion of NHWs than all other clusters and a higher proportion of male subjects than the mild upper zone emphysema and airway-predominant clusters ($p<0.001$).

Validation of the CF4 clustering solution

To validate the CF4 clustering solution, we examined the characteristics and associations of CF4 clusters in the validation data sample. The characteristics of the CF4 clusters in the training and validation samples were similar (table 2), demonstrating that the clusters can reliably be reproduced in a separate data sample.

The associations in the training and validation sample between CF4 clusters, COPD-related clinical measures and COPD SNPs are shown in table 3. For the clinical variables, all 12 of the associations are highly significant in training and validation. For the genetic risk factors, the two associations in the training sample with p values below the Bonferroni-determined threshold of $p=0.0007$ were both replicated at $p \leq 0.05$ in the validation sample. Furthermore, of the 11 genetic associations observed with $p \leq 0.05$ in the training sample, 7 were replicated at $p \leq 0.05$ in validation.

Robustness of CF4 clusters after adjustment for GOLD stage

To determine whether the associations observed with these clusters and COPD-related clinical and genetic variables were driven by severity of airflow obstruction, we repeated the cluster association tests adjusting for GOLD 2007 stage and GOLD 2011 classes A–D (see online supplemental tables 2 and 3). All of the associations with clinical measures remained significant ($p \leq 0.001$). This suggests that the discovered clusters provide information independent from COPD severity as defined by GOLD.

In regard to genetic associations, the cluster associations showed divergent behaviour in response to adjustment for GOLD 2007 stage and GOLD A–D classes. The genetic associations with cluster 4 were attenuated, whereas the strong association observed between cluster 2 (upper zone emphysema) and rs1980057 near *HHIP* was unaffected, suggesting that this association is due to properties of this cluster that are distinct from disease severity as assessed by the severity of airflow obstruction.

DISCUSSION

Using a large sample of smokers with a wide range of airflow obstruction and well characterised with respect to COPD features, cluster analysis identified solutions demonstrating strong association with clinically relevant COPD-related measures and high repeatability in cross-validation. A filtered subset of input

features yielded a four-cluster result that is informative beyond the traditional COPD case–control distinction. These clusters can be described as (1) relatively smoking-resistant individuals, (2) individuals with mild upper zone-predominant emphysema and airflow obstruction, (3) individuals with airway-predominant disease and (4) individuals with severe obstruction and emphysema. In addition to being relevant clinically, some of these clusters are strongly associated with known COPD-associated variants. These clusters and associations were validated in a second data sample from the same study population.

This analysis presents novel findings about smoking-related pulmonary subtypes. We describe a mild upper zone emphysema-predominant cluster that has not been extensively described in previous studies and demonstrate that membership in this cluster is associated with a genetic variant in the *HHIP* gene. This cluster was identified in our study population for at least three reasons: first, our study population included CT scans from a range of smokers, including those with mild or no obstruction; second, we included emphysema distribution as an input feature for clustering; and third, our sample size is substantially larger than previously reported COPD cluster analysis studies. Our work also adds to the field by explicitly addressing the reproducibility of cluster analyses and by using intrinsic (ie, cluster stability) and extrinsic (ie, clinical and genetic associations) criteria for assessing multiple potential clustering solutions.

These results confirm some of the findings from previous subtyping efforts in COPD. First, most studies have identified a severely affected group, though the severity of emphysema and airway wall thickness in this group has been variable.[12 21–23] Second, these findings affirm the concept of emphysema-predominant and airway-predominant COPD while providing additional insight regarding the role of emphysema distribution in COPD heterogeneity.[2 5 13 21 22 24 25] The identification of emphysema-predominant and airway-predominant groups, however, has not been universal. Garcia-Aymerich *et al* did not identify an airway-predominant group, and instead identified a group with elevated BMI and increased comorbidities but with less prominent airway wall thickness on CT scan.[12] In our study, the high average BMI and over-representation of women in the airway-predominant group is of clinical and epidemiological interest, and the female airway predominance recapitulates observations by Martinez *et al* in NETT.[26]

We examined the association of clusters with known COPD GWAS SNPs. While the directionality of associations varied between clusters for some SNPs, the analysed SNPs did show a consistent direction of effect compared with the previous COPD susceptibility association literature in the comparison of the relatively smoking-resistant cluster to the severe obstruction/emphysema cluster. The weak associations in our airway-predominant group are consistent with the findings in the ECLIPSE cohort, where no associations were identified with Pi10.[27] In contrast, consistent associations with the *HHIP* and 15q loci were found for the severe and mild upper lobe-predominant emphysema groups. This association in the latter group is particularly notable since the airway-predominant group, with similar average lung function to the upper lobe-predominant group, shows no strong genetic associations. These results are congruent with ECLIPSE where the associations of these loci with radiologist-scored emphysema were stronger than that for *FAM13A*.[27] Together, these findings suggest that genetic associations in COPD may be subtype dependent.

This work has some limitations. It focuses primarily on continuous spirometric and quantitative CT measures; however, other aspects of COPD such as biomarker measurements and comorbidities were not included either due to their absence from our data or due to limitations of the k-means clustering method, which can yield spurious results when applied to a mixture of continuous and categorical variables. In the future, approaches that evaluate a range of clustering methods and a wider set of variables will be of interest. However, as this work demonstrates, the inclusion of more input features does not necessarily yield better clustering results. The optimal selection of features for clustering (ie, feature selection) is a critical area for the application of unsupervised learning to disease subtyping that requires further exploration. This analysis is cross-sectional, and it is possible that these results may be confounded by differences in disease severity. This is an important limitation for all clustering efforts using cross-sectional data that could be addressed through analyses of longitudinal data or through the development of novel clustering methods. A number of subjects from the overall study were excluded from the clustering analysis due to missing data, primarily from CT scan-related variables, and there is some bias in the clustering subset compared with the excluded subjects. This limits the generalisability of the sample on which clustering was performed, though the included sample is large and consists of a broad spectrum of smoking-related disease.

In summary, k-means clustering in the COPDGene study identifies four groups of smokers that are associated with important COPD-related measures even after adjustment for GOLD stage. Genetic association analysis with known COPD-associated variants shows strong, cluster-specific associations with these known genetic risk factors. This clustering approach is reproducible in independent data sets, facilitating the further study and characterisation of these groups of smokers.

Author affiliations
[1]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA
[2]Divison of General Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA
[3]Department of Computer Science, Northeastern University, Boston, Massachusetts, USA
[4]Surgical Planning Laboratory and Laboratory of Mathematics in Imaging, Brigham and Women's Hospital, Boston, Massachusetts, USA
[5]Pulmonary and Critical Care Division, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA
[6]Division of Biostatistics and Bioinformatics, National Jewish Health, Denver, Massachusetts, USA
[7]Department of Radiology, National Jewish Health, Denver, Massachusetts, USA
[8]Department of Medicine, National Jewish Health, Denver, Massachusetts, USA
[9]Department of Epidemiology, Colorado School of Public Health, University of Colorado Denver, Denver, Colorado, USA
[10]Department of Internal Medicine, Division of Pulmonary and Critical Care, University of Michigan Health System, Ann Arbor, Michigan, USA
[11]Department of Medicine, Division of Pulmonary and Critical Care Medicine, University of California, San Diego, USA
[12]Departments of Medicine and Epidemiology, Columbia University Medical Center, New York, USA
[13]Morehouse School of Medicine, Atlanta, USA
[14]Clinical Research Imaging Centre, Queen's Medical Research Institute, University of Edinburgh, Edinburgh, UK
[15]Rehabilitation Clinical Trials Center, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, California, USA
[16]Temple University, Philadelphia, Pennsylvania, USA
[17]Department of Biostatistics, University of Colorado, Denver, USA
[18]Anschutz Medical Campus, Aurora, USA
[19]University of Nebraska Medical Center, Omaha, USA
[20]Department of Mathematics and Statistical Sciences, University of Colorado, Denver, USA
[21]Division of Pulmonary, Allergy, and Critical Care Medicine, University of Pittsburgh, Pittsburgh, USA

Acknowledgements

COPDGene Investigators—Core Units

Administrative Core: James Crapo, MD (PI), Edwin Silverman, MD, PhD (PI), Barry Make, MD, Elizabeth Regan, MD, PhD, Sarah Moyle, MS, Amy Willis, MA, Rochelle Lantz, Lori Stepp, Sandra Melanson, Douglas Stinson

Genetic Analysis Core: Terri Beaty, PhD, Barbara Klanderman, PhD, Nan Laird, PhD, Christoph Lange, PhD, Michael Cho, MD, Stephanie Santorico, PhD, John Hokanson, MPH, PhD, Dawn DeMeo, MD, MPH, Nadia Hansel, MD, MPH, Craig Hersh, MD, MPH, Peter Castaldi, MD, MSc, Jacqueline Hetmanski, MS, Margaret Parker, MS, Tanda Murray, MS

Imaging Core: David Lynch, MB, Joyce Schroeder, MD, John Newell, Jr., MD, John Reilly, MD, Harvey Coxson, PhD, Philip Judy, PhD, Eric Hoffman, PhD, George Washko, MD, Raul San Jose Estepar, PhD, James Ross, MSc, Ho Yun Lee, MD, Joon Beom Seo, MD, PhD, Atsushi Nambu, MD, PhD, Gongyoung Jin, MD, PhD, Song Soo Kim, MD, Mustafa Al Qaisi, MD, Rebecca Leek, Jordan Zach, Alex Kluiber, Jered Sieren, Heather Baumhauer, Verity McArthur, Demitry Kazlouski, Andrew Allen, Tanya Mann, Anastasia Rodionova, Deanna Richert, Joshua Jaramillo, Alexander McKenzie, Thomas Gethin-Jones, Jaleh Akhavan, Douglas Stinson

PFT QA Core, LDS Hospital, Salt Lake City, UT: Robert Jensen, PhD

Biological Repository, Johns Hopkins University, Baltimore, MD: Homayoon Farzadegan, PhD, Stacey Meyerer, Shivam Chandan, Samantha Bragan

Data Coordinating Center and Biostatistics, National Jewish Health, Denver, USA: Douglas Everett, PhD, Andre Williams, PhD, Carla Wilson, MS, Anna Forssen, MS, Amber Powell, Joe Piccoli

Epidemiology Core, University of Colorado School of Public Health, Denver, USA: John Hokanson, MPH, PhD, Marci Sontag, PhD, Jennifer Black-Shinn, MPH, Gregory Kinney, MPH, PhDc, Sharon Lutz, MPH, PhD

COPDGene Investigators—Clinical Centers

Ann Arbor VA: Jeffrey Curtis, MD, Ella Kazerooni, MD

Baylor College of Medicine, Houston, TX: Nicola Hanania, MD, MS, Philip Alapat, MD, Venkata Bandi, MD, Kalpalatha Guntupalli, MD, Elizabeth Guy, MD, Antara Mallampalli, MD, Charles Trinh, MD, Mustafa Atik, MD, Hasan Al-Azzawi, MD, Marc Willis, DO, Susan Pinero, MD, Linda Fahr, MD, Arun Nachiappan, MD, Collin Bray, MD, L. Alexander Frigini, MD, Carlos Farinas, MD, David Katz, MD, Jose Freytes, MD, Anne Marie Marciel, MD

Brigham and Women's Hospital, Boston, MA: Dawn DeMeo, MD, MPH, Craig Hersh, MD, MPH, George Washko, MD, Francine Jacobson, MD, MPH, Hiroto Hatabu, MD, PhD, Peter Clarke, MD, Ritu Gill, MD, Andetta Hunsaker, MD, Beatrice Trotman-Dickenson, MBBS, Rachna Madan, MD

Columbia University, New York, NY: R. Graham Barr, MD, DrPH, Byron Thomashow, MD, John Austin, MD, Belinda D'Souza, MD

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

Fallon Clinic, Worcester, MA: Richard Rosiello, MD, Timothy Bresnahan, MD, Joseph Bradley, MD, Sharon Kuong, MD, Steven Meller, MD, Suzanne Roland, MD

Health Partners Research Foundation, Minneapolis, MN: Charlene McEvoy, MD, MPH, Joseph Tashjian, MD

Johns Hopkins University, Baltimore, MD: Robert Wise, MD, Nadia Hansel, MD, MPH, Robert Brown, MD, Gregory Diette, MD, Karen Horton, MD

Los Angeles Biomedical Research Institute at Harbor UCLA Medical Center, Los Angeles, CA: Richard Casaburi, MD, PhD, Janos Porszasz, MD, PhD, Hans Fischer, MD, PhD, Matt Budoff, MD, Mehdi Rambod, MD

Michael E. DeBakey VAMC, Houston, TX: Amir Sharafkhaneh, MD, Charles Trinh, MD, Hirani Kamal, MD, Roham Darvishi, MD, Marc Willis, DO, Susan Pinero, MD, Linda Fahr, MD, Arun Nachiappan, MD, Collin Bray, MD, L. Alexander Frigini, MD, Carlos Farinas, MD, David Katz, MD, Jose Freytes, MD, Anne Marie Marciel, MD

Minneapolis VA: Dennis Niewoehner, MD, Quentin Anderson, MD, Kathryn Rice, MD, Audrey Caine, MD

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

National Jewish Health, Denver, USA: Russell Bowler, MD, PhD, Adam Friedlander, MD, David Lynch, MB, Joyce Schroeder, MD, John Newell, Jr., MD, Valerie Hale, MD, John Armstrong, II, MD, Debra Dyer, MD, Jonathan Chung, MD, Christian Cox, MD, Hakan Sahin, MD

Temple University, Philadelphia, PA: Gerard Criner, MD, Victor Kim, MD, Nathaniel Marchetti, DO, Aditi Satti, MD, A. James Mamary, MD, Robert Steiner, MD, Chandra Dass, MD, Libby Cone, MD

University of Alabama, Birmingham, AL: William Bailey, MD, Mark Dransfield, MD, Michael Wells, MD, Surya Bhatt, MD, Hrudaya Nath, MD, Satinder Singh, MD

University of California, San Diego, CA: Joe Ramsdell, MD, Paul Friedman, MD

University of Iowa, Iowa City, IA: Geoffrey McLennan, MD, PhD, Edwin JR van Beek, MD, PhD, Brad Thompson, MD, Dwight Look, MD, Alejandro Cornellas, MD

University of Michigan, Ann Arbor, MI: Fernando Martinez, MD, MeiLan Han, MD, Ella Kazerooni, MD

University of Minnesota, Minneapolis, MN: Christine Wendt, MD, Tadashi Allen, MD

University of Pittsburgh, Pittsburgh, PA: Frank Sciurba, MD, Joel Weissfeld, MD, MPH, Carl Fuhrman, MD, Jessica Bon, MD, Danielle Hooper, MD

University of Texas Health Science Center at San Antonio, San Antonio, TX: Antonio Anzueto, MD, Sandra Adams, MD, Carlos Orozco, MD, Mario Ruiz, MD, Amy Mumbower, MD, Ariel Kruger, MD, Carlos Restrepo, MD, Michael Lane, MD

Contributors PJC, JD, JR, JEH, SS, EKS and MHC were responsible for conception and design. PJC, JD, JR, YC, GRW, DC-E, DAL, BJM, JDC, RPB, EAR, JEH, MKH, XS, JWR, RGB, MF, EvB, RC, GJC, SML, SIR, SS, FCS, DLD, CPH, EKS and MHC were responsible for acquisition, analysis and/or interpretation. All authors were responsible for drafting the manuscript.

Funding This work was supported by U.S. National Institutes of Health (NIH) grants K08HL102265 (Castaldi), K08HL097029 (Cho), P01HL105339 (Silverman) and by Award Numbers R01HL089897 (Crapo) and R01HL089856 (Silverman) 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 comprising AstraZeneca, Boehringer Ingelheim, Novartis, Pfizer, Siemens and Sunovion.

Competing interests CPH reports grants from National Heart Lung and Blood Institute, grants from National Institute of Nursing Research, grants from Alpha-1 Foundation, personal fees from Novartis and personal fees from CSL Behring. In the past three years, EKS received honoraria and consulting fees from Merck and grant support from GlaxoSmithKline. RT-S is a current employee of GlaxoSmithKline. DAL is a consultant for, and receives honoraria and grant support, from GSK. He is the Chair of the GSK Respiratory Area Therapy Board. MKH has served as a consultant for Boehringer Ingelheim, Pfizer, GSK, Medimmune, Novartis, Grifols Therapeutics and United Biosource Corporation. She has received royalties from UpToDate and has developed educational presentations for National Association for Continuing Education and WebMD. DAL has received grant support from Siemens and Centocor and served as a consultant for Perceptive Imaging, Intermune and Gilead. BJM has served on advisory boards for Forest, AstraZeneca, Novartis, Coviden, Breathe, Merck, Sunovion, Boehringer Ingelheim, MedImmune, Ikaria and Novartis, served as a consultant for Astellas, reports grant support from AstraZeneca, GlaxoSmithKline, NABI, Boehringer Ingelheim, Sunovion and Forest, received lecture fees from GlaxoSmithKline, Boehringer Ingelheim, Pfizer and Forest and has received royalties from UpToDate. FCS has participated in consulting for GSK, AstraZeneca and Pfizer and has received research grant funding from the NIH, GSK, BI, Pfizer, Forest and Actelion. SIR received fees for serving on advisory boards, consulting or honoraria from Almirall, APT Pharma, Aradigm, Argenta, AstraZeneca, Boehringer Ingelheim, Chiesi, Dey, Forest, GlaxoSmithKline, Hoffmann-La Roche, MedImmune, Mpex, Novartis, Nycomed, Oriel, Otsuka, Pearl, Pfizer, Pharmaxis, Merck and Talecris.

Ethics approval Brigham and Women's Hospital IRB and participating study centre IRBs.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Clinical and genetic data from the COPDGene study are available through dbGaP.

REFERENCES

1 Burrows B, Niden AH, Fletcher CM, *et al*. Clinical types of chronic obstructive lung disease in london and in Chicago. A study of one hundred patients. *Am Rev Respir Dis* 1964;90:14–27.
2 Burrows B, Fletcher CM, Heard BE, *et al*. The emphysematous and bronchial types of chronic airways obstruction. A clinicopathological study of patients in London and Chicago. *Lancet* 1966;1:830–5.
3 Hurst JR, Vestbo J, Anzueto A, *et al*. Susceptibility to exacerbation in chronic obstructive pulmonary disease. *N Engl J Med* 2010;363:1128–38.
4 Fishman A, Martinez F, Naunheim K, *et al*. A randomized trial comparing lung-volume-reduction surgery with medical therapy for severe emphysema. *N Engl J Med* 2003;348:2059–73.
5 Ziegler-Heitbrock L, Frankenberger M, Heimbeck I, *et al*. The EvA study: aims and strategy. *Eur Respir J* 2012;40:823–9.
6 Rabe KF, Hurd S, Anzueto A, *et al*. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am J Respir Crit Care Med* 2007;176:532–55.
7 Calverley PM. The GOLD classification has advanced understanding of COPD. *Am J Respir Crit Care Med* 2004;170:211–12.
8 Agusti A, Calverley PM, Celli B, *et al*. Characterisation of COPD heterogeneity in the ECLIPSE cohort. *Respir Res* 2010;11:122.
9 Rennard SI, Vestbo J. The many "small COPDs": COPD should be an orphan disease. *Chest* 2008;134:623–7.
10 Vestbo J, Hurd SS, Agusti AG, *et al*. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am J Respir Crit Care Med* 2013;187:347–65.

11 Cho MH, Washko GR, Hoffmann TJ, *et al*. Cluster analysis in severe emphysema subjects using phenotype and genotype data: an exploratory investigation. *Respir Res* 2010;11:30.
12 Garcia-Aymerich J, Gomez FP, Benet M, *et al*. Identification and prospective validation of clinically relevant chronic obstructive pulmonary disease (COPD) subtypes. *Thorax* 2011;66:430–7.
13 Paoletti M, Camiciottoli G, Meoni E, *et al*. Explorative data analysis techniques and unsupervised clustering methods to support clinical assessment of Chronic Obstructive Pulmonary Disease (COPD) phenotypes. *J Biomed Inform* 2009;42:1013–21.
14 Regan EA, Hokanson JE, Murphy JR, *et al*. Genetic epidemiology of COPD (COPDGene) study design. *COPD* 2010;7:32–43.
15 Fraiman R, Justel A, Svarc M. Selection of variables for cluster analysis and classification rules. *J Am Stat Assoc* 2008;103:1294–303.
16 R Development Core Team. R: A Language and Environment for Statistical Computing. 2011. http://www.R-project.org
17 Cho MH, Boutaoui N, Klanderman BJ, *et al*. Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nat Genet* 2010;42:200–2.
18 Wilk JB, Chen TH, Gottlieb DJ, *et al*. A genome-wide association study of pulmonary function measures in the Framingham Heart Study. *PLoS Genet* 2009;5:e1000429.
19 Pillai SG, Ge D, Zhu G, *et al*. A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet* 2009;5:e1000421.
20 Cho MH, Castaldi PJ, Wan ES, *et al*. A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. *Hum Mol Genet* 2012;21:947–57.
21 Fujimoto K, Kitaguchi Y, Kubo K, *et al*. Clinical analysis of chronic obstructive pulmonary disease phenotypes classified using high-resolution computed tomography. *Respirology* 2006;11:731–40.
22 Pistolesi M, Camiciottoli G, Paoletti M, *et al*. Identification of a predominant COPD phenotype in clinical practice. *Respir Med* 2008;102:367–76.
23 Burgel PR, Paillasseur JL, Caillaud D, *et al*. Clinical COPD phenotypes: a novel approach using principal component and cluster analyses. *Eur Respir J* 2010;36:531–9.
24 Patel BD, Coxson HO, Pillai SG, *et al*. Airway wall thickening and emphysema show independent familial aggregation in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2008;178:500–5.
25 Hogg JC. A pathologist's view of airway obstruction in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2012;186:v–vii.
26 Martinez FJ, Curtis JL, Sciurba F, *et al*. Sex differences in severe pulmonary emphysema. *Am J Respir Crit Care Med* 2007;176:243–52.
27 Pillai SG, Kong X, Edwards LD, *et al*. Loci identified by genome-wide association studies influence different disease-related phenotypes in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2010;182:1498–505.