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 Sciruba,21 Dawn L DeMeo,1,5 Craig P Hersh,1,5 Edwin K Silverman,1,5 Michael H Cho1,5

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.

ABSTRACT

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.

ORIGINAL ARTICLE

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.

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 FEV1/FVC and classifies disease severity based on FEV1.
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 FEV1 and FEV1/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 clustering method, would identify novel, clinically relevant 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.

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 clustering method, would identify novel, clinically relevant 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.

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 clustering method, would identify novel, clinically relevant 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 clustering method, would identify novel, clinically relevant 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 clustering method, would identify novel, clinically relevant 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 clustering method, would identify novel, clinically relevant 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.
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, FEV1% predicted, CT-quantified emphysema at −950 Hounsfield units (HU) and bronchodilator responsiveness as a % of FEV1. For the core feature set, correlation filtering yielded a set of four features—FEV1% 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.

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.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Baseline characteristics of the training and validation data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>Validation</td>
</tr>
<tr>
<td>N</td>
<td>4187</td>
</tr>
<tr>
<td>Age</td>
<td>59.5 (9.0)</td>
</tr>
<tr>
<td>Gender, % female</td>
<td>46.7</td>
</tr>
<tr>
<td>Race, % African-American</td>
<td>32.0</td>
</tr>
<tr>
<td>FEV1, % of predicted</td>
<td>76.9 (25.2)</td>
</tr>
<tr>
<td>FEV1/FVC</td>
<td>0.67 (0.16)</td>
</tr>
<tr>
<td>Pack-years, median (IQR)</td>
<td>39.3 (28.0)</td>
</tr>
<tr>
<td>BMI</td>
<td>28.9 (6.3)</td>
</tr>
<tr>
<td>Emphysema at −950HU, median (IQR)</td>
<td>1.8 (5.8)</td>
</tr>
<tr>
<td>Upper/lower emphysema ratio (IQR)</td>
<td>0.8 (1.1)</td>
</tr>
<tr>
<td>Segmental airway wall thickness</td>
<td>61.4 (3.2)</td>
</tr>
<tr>
<td>Upper/lower lobe emphysema difference (IQR)</td>
<td>−0.17 (2.0)</td>
</tr>
<tr>
<td>Gas trapping (IQR)</td>
<td>14.5 (24.8)</td>
</tr>
<tr>
<td>GOLD unclassifiable*, %</td>
<td>12.0</td>
</tr>
<tr>
<td>Smoking controls, %</td>
<td>43.8</td>
</tr>
<tr>
<td>GOLD 1, %</td>
<td>8.3</td>
</tr>
<tr>
<td>GOLD 2, %</td>
<td>19.2</td>
</tr>
<tr>
<td>GOLD 3, %</td>
<td>11.3</td>
</tr>
<tr>
<td>GOLD 4, %</td>
<td>5.4</td>
</tr>
</tbody>
</table>

*GOLD unclassifiable refers to subjects with a FEV1% predicted <80 but FEV1/FVC >0.7.
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).

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=4.4×10^−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.

Cluster 4: severe emphysema
Cluster 4 represents 24% of the training sample and is characterised by severe airflow obstruction and severe emphysema (p values compared with other clusters <0.001). The average amount of emphysema in this group is high (mean emphysema = 15.26%), with nearly half of this cluster having greater than 20% emphysema. As shown in figure 3, most of the individuals in the severe emphysema cluster are GOLD stages 3–4, with a higher proportion of women compared with the smoking-resistant cluster (p<0.001) and a lower proportion of AAs (p<0.001).

Cluster 1 (relatively smoking resistant individuals) consists largely of control smokers and GOLD 1–2 individuals. Cluster 4 (severe emphysema) consists largely of GOLD 3–4 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.

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

<table>
<thead>
<tr>
<th></th>
<th>Training sample</th>
<th>Validation sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C1: mean</td>
<td>C2: mean</td>
</tr>
<tr>
<td></td>
<td>C1: mean</td>
<td>C2: mean</td>
</tr>
<tr>
<td>N</td>
<td>1598</td>
<td>623</td>
</tr>
<tr>
<td>Age</td>
<td>58.9</td>
<td>58.0*</td>
</tr>
<tr>
<td>Gender, % female</td>
<td>0.44</td>
<td>0.53</td>
</tr>
<tr>
<td>Race, % African-American</td>
<td>0.30</td>
<td>0.46</td>
</tr>
<tr>
<td>FEV1, per cent of predicted</td>
<td>95.3</td>
<td>81.9</td>
</tr>
<tr>
<td>FEV1/FVC</td>
<td>0.76</td>
<td>0.70</td>
</tr>
<tr>
<td>BMI</td>
<td>28.7</td>
<td>27.9</td>
</tr>
<tr>
<td>Pack years</td>
<td>38.0</td>
<td>45.8</td>
</tr>
<tr>
<td>Emphysema at −950HU</td>
<td>2.6</td>
<td>3.3</td>
</tr>
<tr>
<td>Segmental airway wall thickness</td>
<td>58.8</td>
<td>61.5</td>
</tr>
<tr>
<td>Upper/lower emphysema ratio</td>
<td>0.7</td>
<td>6.7</td>
</tr>
<tr>
<td>Upper/lower emphysema difference</td>
<td>−0.3</td>
<td>1.4</td>
</tr>
<tr>
<td>Gas trapping†</td>
<td>12.9</td>
<td>16.5</td>
</tr>
</tbody>
</table>

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.

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 3–4 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 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=5×10⁻⁸) 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 ≤ 0.05 in the validation sample. Furthermore, of the 11 genetic associations observed with p ≤ 0.05 in the training sample, 7 were replicated at p ≤ 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 ≤ 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

<table>
<thead>
<tr>
<th>Training Validation</th>
<th>C2: OR (95% CI)</th>
<th>C3: OR (95% CI)</th>
<th>C4: OR (95% CI)</th>
<th>C2: OR (95% CI)</th>
<th>C3: OR (95% CI)</th>
<th>C4: OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exacerbations</td>
<td>2.27 (2.19 to 2.35)***</td>
<td>3.10 (2.92 to 3.27)***</td>
<td>3.61 (3.48 to 3.78)***</td>
<td>2.81 (2.60 to 3.03)***</td>
<td>3.24 (3.02 to 3.48)***</td>
<td>3.78 (3.48 to 4.10)***</td>
</tr>
<tr>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>MMRC</td>
<td>2.81 (2.57 to 3.07)***</td>
<td>3.31 (3.02 to 3.62)***</td>
<td>4.67 (4.24 to 5.15)***</td>
<td>2.63 (2.32 to 2.99)***</td>
<td>3.65 (3.27 to 4.07)***</td>
<td>5.26 (4.78 to 5.78)***</td>
</tr>
<tr>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>BODE</td>
<td>3.37 (3.06 to 3.78)***</td>
<td>4.63 (4.24 to 5.15)***</td>
<td>66.52 (60.06 to 73.67)***</td>
<td>2.62 (2.32 to 2.99)***</td>
<td>4.23 (3.87 to 4.61)***</td>
<td>52.64 (47.62 to 58.19)***</td>
</tr>
<tr>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>Hospitalisations/ER visits</td>
<td>4.07 (3.75 to 4.43)***</td>
<td>5.05 (4.42 to 5.78)***</td>
<td>11.82 (9.98 to 14.00)***</td>
<td>3.05 (2.53 to 3.68)***</td>
<td>4.13 (3.52 to 4.86)***</td>
<td>8.03 (6.86 to 9.39)***</td>
</tr>
<tr>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>rs7671167 (FAM13A)</td>
<td>0.95 (0.87 to 1.04) NS</td>
<td>0.95 (0.87 to 1.04) NS</td>
<td>0.95 (0.87 to 1.04) NS</td>
<td>0.95 (0.87 to 1.04) NS</td>
<td>0.95 (0.87 to 1.04) NS</td>
<td>0.95 (0.87 to 1.04) NS</td>
</tr>
<tr>
<td>rs1980057 (HHIP)</td>
<td>0.64 (0.58 to 0.70)***</td>
<td>0.92 (0.85 to 0.98) NS</td>
<td>0.79 (0.72 to 0.86)***</td>
<td>0.79 (0.72 to 0.86)***</td>
<td>0.92 (0.85 to 0.98) NS</td>
<td>1.09 (1.02 to 1.17) NS</td>
</tr>
<tr>
<td>rs13180 (Chr15q25)</td>
<td>0.82 (0.75 to 0.89)***</td>
<td>1.04 (0.96 to 1.11) NS</td>
<td>0.82 (0.72 to 0.93)***</td>
<td>0.82 (0.72 to 0.93)***</td>
<td>1.04 (0.96 to 1.11) NS</td>
<td>0.82 (0.72 to 0.93)***</td>
</tr>
<tr>
<td>rs8034191 (Chr15q25)</td>
<td>1.33 (1.21 to 1.48)***</td>
<td>1.25 (1.15 to 1.36)***</td>
<td>1.50 (1.39 to 1.62)***</td>
<td>1.30 (1.18 to 1.42)***</td>
<td>1.30 (1.18 to 1.42)***</td>
<td>1.50 (1.39 to 1.62)***</td>
</tr>
<tr>
<td>rs7937 (Chr19q13)</td>
<td>1.30 (1.18 to 1.42)***</td>
<td>1.16 (1.08 to 1.24)***</td>
<td>1.46 (1.36 to 1.57)***</td>
<td>1.30 (1.18 to 1.42)***</td>
<td>1.16 (1.08 to 1.24)***</td>
<td>1.46 (1.36 to 1.57)***</td>
</tr>
</tbody>
</table>
| 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.

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 literature by (1) providing additional insight regarding the role of emphysema distribution in COPD heterogeneity, (2) substantially increasing the size of our sample is large and consists of a broad spectrum of smoking-related disease, and (3) individuals with severe obstruction and emphysema are more likely to be described as (1) relatively smoking-resistant individuals, (2) individuals with mild upper zone-predominant emphysema, (3) individuals with airway-predominant disease, and (4) individuals with severe obstruction and emphysema.

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 26 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.28 29 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
1Channing Division of Network Medicine, Brigham and Women’s Hospital, Boston, Massachusetts, USA
2Division of General Medicine, Brigham and Women’s Hospital, Boston, Massachusetts, USA
3Department of Computer Science, Northeastern University, Boston, Massachusetts, USA
4Surgical Planning Laboratory and Laboratory of Mathematics in Imaging, Brigham and Women’s Hospital, Boston, Massachusetts, USA
5Pulmonary and Critical Care Division, Brigham and Women’s Hospital and Harvard Medical School, Boston, Massachusetts, USA
6Division of Biostatistics and Bioinformatics, National Jewish Health, Denver, Colorado, USA
7Department of Radiology, National Jewish Health, Denver, Colorado, USA
8Department of Medicine, National Jewish Health, Denver, Colorado, USA
9Department of Epidemiology, Colorado School of Public Health, University of Colorado Denver, Denver, Colorado, USA
10Department of Internal Medicine, Division of Pulmonary and Critical Care, University of Michigan Health System, Ann Arbor, Michigan, USA
11Department of Medicine, Division of Pulmonary and Critical Care Medicine, University of California, San Diego, USA
12Departments of Medicine and Epidemiology, Columbia University Medical Center, New York, USA
13Morehouse School of Medicine, Atlanta, USA
14Clinical Research Imaging Centre, Queen’s Medical Research Institute, University of Edinburgh, Edinburgh, UK
15Rehabilitation Clinical Trials Center, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, California, USA
16Temple University, Philadelphia, Pennsylvania, USA
17Department of Biostatistics, University of Colorado, Denver, USA
18Anschutz Medical Campus, Aurora, USA
19University of Nebraska Medical Center, Omaha, USA
20Department of Mathematics and Statistics, University of Colorado, Denver, USA
21Division of Pulmonary, Allergy, and Critical Care Medicine, University of Pittsburgh, Pittsburgh, USA
22Department of Medicine, National Jewish Health, Denver, Colorado, USA
23Departments of Medicine and Biostatistics, University of Colorado, Denver, USA
24Department of Biostatistics, University of Colorado, Denver, USA
25Division of Pulmonary and Critical Care Medicine, University of Colorado, Denver, USA
26Division of Pulmonary, Allergy, and Critical Care Medicine, University of Pittsburg, Pittsburgh, USA
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, Axel Kruger, MD, Carlos Restrepo, MD, Michael Lane, MD

Contributors PJC, JD, JR, JEH, SS, EK, and MHC were responsible for conception and design. PJC, JD, JR, YC, GRW, DC-E, DAL, BJM, IDC, RPB, EAR, JEH, MKH, XS, JWR, RGB, MF, EvB, RC, GC, 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 R01HL102265 (Castaldi), K08HL097029 (Cho), P01HL105339 (Silverman) and by Award Numbers R01HL089897 (Crapo) and R01HL089856 (Silverman) (Silverman) for Boehringer Ingelheim, Pfizer, GSX, 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, Cogen, Breathe, Merck, Sunovion, Boehringer Ingelheim, MedImmune, Ikaria and Novartis, served as a consultant for Astellas, offers 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, BP, Pfizer, Forest and AstraZeneca. 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, Nymco, 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
7 Calverley PM. The GOLD classification has advanced understanding of COPD. Am J Respir Crit Care Med 2004;170:211–12.


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

Peter J Castaldi, Jennifer Dy, James Ross, Yale Chang, George R Washko, Douglas Curran-Everett, Andre Williams, David A Lynch, Barry J Make, James D Crapo, Russ P Bowler, Elizabeth A Regan, John E Hokanson, Greg L Kinney, Meilan K Han, Xavier Soler, Joseph W Ramsdell, R Graham Barr, Marilyn Foreman, Edwin van Beek, Richard Casaburi, Gerald J Criner, Sharon M Lutz, Steven I Rennard, Stephanie Santorico, Frank C Sciurba, Dawn L DeMeo, Craig P Hersh, Edwin K Silverman and Michael H Cho

Thorax published online February 21, 2014

Updated information and services can be found at: http://thorax.bmj.com/content/early/2014/02/21/thoraxjnl-2013-203601

These include:

Supplementary Material
Supplementary material can be found at: http://thorax.bmj.com/content/suppl/2014/02/21/thoraxjnl-2013-203601.DC1

References
This article cites 26 articles, 4 of which you can access for free at: http://thorax.bmj.com/content/early/2014/02/21/thoraxjnl-2013-203601#BIBL

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Topic Collections
Articles on similar topics can be found in the following collections

Health education (1223)
Smoking (1037)
Tobacco use (1039)

Notes

To request permissions go to: http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to: http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to: http://group.bmj.com/subscribe/