



# Clinical COPD phenotypes: a novel approach using principal component and cluster analyses

P-R. Burgel<sup>\*</sup>, J-L. Paillasseur<sup>#</sup>, D. Caillaud<sup>†</sup>, I. Tillie-Leblond<sup>+</sup>, P. Chanez<sup>§</sup>, R. Escamilla<sup>‡</sup>, I. Court-Fortune<sup>\*\*</sup>, T. Perez<sup>+</sup>, P. Carré<sup>##</sup>, and N. Roche<sup>††</sup> on behalf of the Initiatives BPCO Scientific Committee<sup>++</sup>

**ABSTRACT:** Classification of chronic obstructive pulmonary disease (COPD) is usually based on the severity of airflow limitation, which may not reflect phenotypic heterogeneity. Here, we sought to identify COPD phenotypes using multiple clinical variables.

COPD subjects recruited in a French multicentre cohort were characterised using a standardised process. Principal component analysis (PCA) was performed using eight variables selected for their relevance to COPD: age, cumulative smoking, forced expiratory volume in 1 s (FEV<sub>1</sub>) (% predicted), body mass index, exacerbations, dyspnoea (modified Medical Research Council scale), health status (St George's Respiratory Questionnaire) and depressive symptoms (hospital anxiety and depression scale). Patient classification was performed using cluster analysis based on PCA-transformed data.

322 COPD subjects were analysed: 77% were male; median (interquartile range) age was 65.0 (58.0–73.0) yrs; FEV<sub>1</sub> was 48.9 (34.1–66.3)% pred; and 21, 135, 107 and 59 subjects were classified in Global Initiative for Chronic Obstructive Lung Disease (GOLD) stages 1, 2, 3 and 4, respectively. PCA showed that three independent components accounted for 61% of variance. PCA-based cluster analysis resulted in the classification of subjects into four clinical phenotypes that could not be identified using GOLD classification. Importantly, subjects with comparable airflow limitation (FEV<sub>1</sub>) belonged to different phenotypes and had marked differences in age, symptoms, comorbidities and predicted mortality.

These analyses underscore the need for novel multidimensional COPD classification for improving patient care and quality of clinical trials.

**KEYWORDS:** Chronic obstructive pulmonary disease, cluster analysis, comorbidities, dyspnoea, phenotypes, principal component analysis

Chronic obstructive pulmonary disease (COPD) is a major cause of mortality and disability worldwide [1]. The disease is characterised by airflow limitation that is not fully reversible. Classification of COPD is usually based on the severity of airflow obstruction, as assessed using the forced expiratory volume in 1 s (FEV<sub>1</sub>) [1]. In recent years, it has emerged that COPD is a complex disease with multiple clinical manifestations and that COPD subjects cannot be described by only using the severity of airflow limitation.

Thus, many other independent predictors of outcomes have been identified, including worsening dyspnoea, frequency and severity of exacerbations, malnutrition, depression and health-related quality of life (HRQoL) impairment [2]. Furthermore, comorbidities (e.g. cardiovascular diseases and cancer) are major causes of death and hospitalisations in COPD subjects [3, 4].

Large clinical trials performed in COPD subjects have shown that current treatments improved

## AFFILIATIONS

<sup>\*</sup>Service de Pneumologie, Hôpital Cochin, AP-HP and Paris-Descartes University,

<sup>††</sup>Service de Pneumologie, Hôpital de l'Hôtel Dieu, AP-HP and Paris-Descartes University, Paris,  
<sup>#</sup>Clindatafirst, St Quentin en Yvelines,

<sup>†</sup>Service de Pneumologie, Hôpital Gabriel Montpied, CHU Clermont-Ferrand,

<sup>+</sup>Service de Pneumologie, Hôpital Calmette, Lille,

<sup>§</sup>Département des Maladies Respiratoires, AP-HM, Université de la Méditerranée, Marseille,

<sup>‡</sup>Clinique des voies respiratoires Hôpital Larrey, Toulouse,

<sup>\*\*</sup>Service de Pneumologie, CHU Saint Etienne, Saint Etienne,

<sup>##</sup>Service de Pneumologie, Hôpital Antoine Gayrard, Carcassonne, France.

<sup>†††</sup>For a complete list of study group members, please refer to the Acknowledgements section.

## CORRESPONDENCE

P-R. Burgel  
Service de Pneumologie  
Hôpital Cochin  
Assistance Publique Hôpitaux de Paris  
27 rue du Faubourg St Jacques  
75679 Paris Cedex 14  
France  
E-mail: pierre-regis.burgel@cch.aphp.fr

Received:  
Nov 05 2009  
Accepted after revision:  
Dec 24 2009  
First published online:  
Jan 14 2010

European Respiratory Journal  
Print ISSN 0903-1936  
Online ISSN 1399-3003

For editorial comments see page 472.

This article has online supplementary material available from [www.erj.ersjournals.com](http://www.erj.ersjournals.com)

Earn CME accreditation by answering questions about this article. You will find these at the back of the printed copy of this issue or online at [www.erj.ersjournals.com/misc/cmeinfo.dtl](http://www.erj.ersjournals.com/misc/cmeinfo.dtl)

several outcomes (e.g. exacerbations, dyspnoea and HRQoL), but the authors reported disappointing data on mortality and rates of decline in FEV<sub>1</sub> [5, 6]. One explanation may be that COPD subjects are heterogeneous and that not all subjects benefit from the same therapy. This point has been best exemplified by the National Emphysema Treatment Trial, in which some phenotypic characteristics were associated with increased mortality after lung volume reduction surgery, whereas this therapy reduced mortality in other COPD subjects [7]. Thus, dismantlement of phenotypes appears as one of the current major challenges in subjects with COPD.

Phenotypic characterisation of COPD subjects may rely on clinical manifestations, assessment of patient-related outcomes (e.g. depression and HRQoL) using validated questionnaires, imaging and biological measurements [8]. Many studies are currently trying to identify biomarkers related to severity or prognosis of COPD subjects [9]. Adequate clinical categorisation of subjects would be of utmost importance in these studies. Furthermore, identification of phenotypes using clinical variables would be useful in primary care, where imaging and biological measurements are not widely used.

Identification of clinical COPD phenotypes has been described as early as the 1950s, when DORNHORST [10] proposed the distinction between pink puffers and blue bloaters. These descriptions were based on rather subjective clinical assessment of subjects. In recent years, it has been proposed that statistical methods can be applied to clinical medicine for examining phenotypic heterogeneity. Cluster analysis, which seeks to organise information so that heterogeneous groups of variables can be classified into relatively homogeneous groups, has been proposed to examine phenotypic heterogeneity in airway diseases [11]. In the present study, we used this method to analyse clinical data obtained in a well-characterised group of COPD subjects recruited throughout France [12]. Because information obtained using clinical data and validated questionnaires contained redundancy, cluster analysis was performed using principal component analysis (PCA)-transformed data. This original methodology allowed for testing the hypothesis that COPD subjects could be grouped into clinical phenotypes.

## METHODS

### Subjects

The present study is based on a cross-sectional analysis of a cohort of COPD subjects (Initiatives BPCO study group) recruited between January 2005 and August 2008 in 17 pulmonary units in university hospitals located throughout France [12]. Respiratory physicians prospectively recruited subjects in stable condition (no history of exacerbation requiring medical treatment for the previous 4 weeks) with a diagnosis of COPD based on the presence of a post-bronchodilator FEV<sub>1</sub>/forced vital capacity (FVC) ratio <70% [1]. Subjects with a main diagnosis of bronchiectasis, asthma or any significant respiratory diseases were excluded. The study was approved by the Ethics Committee of Versailles, France and all subjects provided informed written consent.

### Data collection

We used a standardised characterisation process that covered demographic data, cumulative tobacco smoking and COPD characteristics (including symptoms, spirometry and therapy)

in stable condition. Pulmonary function tests were performed according to international standards [13]. Severity of airflow obstruction was evaluated according to Global Initiative for Chronic Obstructive Lung Disease (GOLD) classification [1]. Numbers of acute exacerbations of COPD during the previous year were determined according to patient's self-reported exacerbations. Comorbidities (including congestive heart failure, coronary artery disease, systemic hypertension and diabetes mellitus) were identified from the patient files. We calculated the multidimensional BOD index (body mass index (BMI), obstruction (FEV<sub>1</sub> % pred) and dyspnoea evaluated on the modified Medical Research Council (MMRC) scale), which was reported to be a better predictor of mortality than FEV<sub>1</sub> [14, 15].

The hospital anxiety and depression (HAD) scale was used to examine mood disorders. This 14-item self-questionnaire has two seven-item subscales for anxiety (HAD-A) and depression (HAD-D). Scores range from 0 to 21 for each subscale, and a score of 8 or higher on either subscale is conventionally used to define anxiety and depression [16]. A score of 11 or higher on either subscale is even more closely associated with the presence of the mood disorder. HRQoL was evaluated using the St George's Respiratory Questionnaire (SGRQ) [17].

### Statistical analysis plan

Statistics followed a step by step process detailed below, beginning with selection of relevant clinical variables by the Scientific Committee. Subjects with complete information for these variables were analysed. Correlations between GOLD classes [1] and other variables were assessed using Kendall  $\tau_b$  rank correlation or logistic regression, as appropriate. Next, correlations within the group of selected variables were studied using cluster analysis. These analyses were useful in determining whether information provided by each clinical variable was independent from the others. Because redundancy was found, PCA was performed on these variables as a method for reducing interaction between variables. Then, cluster analysis based on the main components of the PCA was performed to search for COPD phenotypes. Data are presented as median (interquartile range) or % unless otherwise specified. A  $p < 0.05$  was considered statistically significant. Analyses were performed using the SAS package version 9.01 (SAS Institute, Cary, NC, USA).

### Selection of variables for analyses

The scientific committee selected eight variables for their relevance to pulmonary and/or extrapulmonary manifestations of COPD. The variables were: age (years), tobacco smoking (pack-years), severity of airflow obstruction (assessed by FEV<sub>1</sub> % pred), exacerbations (number per patient per year), nutritional status (assessed by BMI kg·m<sup>-2</sup>), dyspnoea (assessed by the MMRC scale), HRQoL (assessed by the SGRQ total score), and anxiety and depression (assessed by the HAD total score).

The cohort contained 584 individual subjects at the time of analysis. Complete data for these eight variables, which were necessary for principal component and cluster analyses, were available for 322 subjects. Most of the remaining 262 subjects were excluded from analyses due to the lack of data on SGRQ or HAD questionnaires. Both populations did not differ in terms of

age, cumulative tobacco smoking, FEV<sub>1</sub>, MMRC scale, BMI and exacerbations (online supplementary material). Male subjects represented 76.7 versus 84.0% of subjects included and excluded in the analysis, respectively ( $p=0.03$ , Chi-squared test).

### Correlations between clinical variables

Relationships between the eight selected variables were studied by cluster analysis, using the VARCLUS procedure. This procedure, which organises a set of numeric variables into hierarchical clusters, can be used to examine redundancy between variables. Results were presented in a dendrogram showing variables in each grouping, and the distance between groupings.

### Identification of COPD phenotypes

Because we found that information obtained using these clinical variables was not independent from one another, we transformed clinical data using PCA [18]. Linear combinations of the eight selected variables were used to form eight new independent variables (eigenvectors) called components [19]. The eigenvalue of each component is a measure of its variability. A component with an eigenvalue  $<1$  contributes little to explain the relationships between original variables and thus is not subjected to further analysis. Next, we performed a cluster analysis based on significant components identified in the PCA (*i.e.* with an eigenvalue  $>1$ ). Cluster analysis was performed using Ward's method. In this method, grouping was based on quantitative measures of similarity procedure (minimum within cluster sum of square), such that subjects in the same cluster were more similar to each other than to subjects in another cluster. We used pseudo-F and pseudo- $t^2$  statistics to determine the optimal number of clusters in the data. Relatively large pseudo-F values were considered to indicate a stopping point. For the pseudo- $t^2$  statistic, we moved down the column until we found the first value markedly larger than the previous value and moved back up the column by one cluster.

## RESULTS

### Classification of COPD subjects according to GOLD classification

Clinical characteristics of the 322 COPD subjects according to GOLD classification are presented in table 1. Variables correlated with increasing GOLD stages included FEV<sub>1</sub>, FVC and BMI (inverse correlations), and MMRC, BOD score, SGRQ total score and numbers of exacerbations per patient per year (positive correlations). Systemic hypertension was less prevalent in subjects with GOLD stage 3 and 4 COPD, and a similar trend existed for coronary disease. There was no significant correlation between GOLD stage and age, smoking history, HAD total score, and other comorbidities (*i.e.* chronic heart failure and diabetes mellitus).

Long-acting  $\beta$ -agonists and inhaled corticosteroids were prescribed in ~50% subjects in GOLD stages 1 and 2, and in up to 84% of subjects in GOLD stage 4. No significant difference was observed among GOLD classes for tiotropium prescription.

### Relationships between clinical variables

Cluster analysis of the eight selected variables resulted in a dendrogram (fig. 1), which illustrates how these variables

relate to each other. This analysis showed that clinical information obtained using these eight variables could be grouped into three clusters (fig. 1), indicating that these variables were not completely independent.

### PCA of clinical variables

We performed PCA to transform data contained in the eight selected variables into eight independent components. The first three components that contributed significantly to explaining the relationships among the eight selected variables (eigenvalues  $>1$ ) accounted for 61% of the information. Correlations of the selected variables with these three independent components are shown in table 2. Component 1 correlated with SGRQ total score and MMRC score, and inversely correlated with FEV<sub>1</sub> (% pred), but was independent from age. Component 2 highly correlated with age and cumulative tobacco smoking, but was independent from FEV<sub>1</sub> (% pred) and SGRQ score. Component 3 mostly correlated with BMI and FEV<sub>1</sub> (% pred). Components 4–8 explained little variability of the original data (eigenvalues  $<1$ ; online supplementary material) and therefore were not subjected to further analysis.

### Classification of COPD subjects using cluster analysis

Classification of the 322 COPD subjects using cluster analysis based on the first three components identified in the PCA resulted in a dendrogram that showed the progressive joining of the clustering process (fig. 2). Pseudo-F and pseudo- $t^2$  statistics determined that the data could be optimally grouped into four clusters (phenotypes). Clinical characteristics of the 322 COPD subjects according to these four phenotypes are presented in table 3.

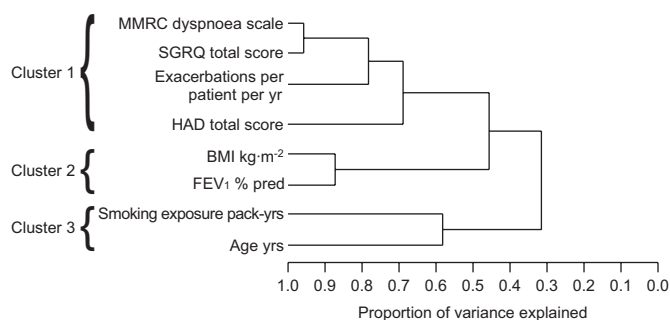
Major differences were found among groups. First, two extreme phenotypes were identified. The first phenotype contained young subjects ( $n=44$ , median age 58 yrs) with severe airflow limitation (GOLD stage 3 and 4), low BMI, severe dyspnoea, frequent exacerbations, anxiety, depression and severely impaired HRQoL. Cardiovascular comorbidities were infrequent in this group of subjects. The second phenotype was composed of older subjects ( $n=89$ , median age 68 yrs) with mild airflow limitation (GOLD stage 1 or 2 in 85.4% subjects), low dyspnoea, low levels of anxiety and depression, almost no exacerbations, mild impairment in HRQoL and who were mildly overweight. These older subjects had higher prevalence of comorbidities, including hypertension (median 57.5%), coronary artery disease (19.5%), diabetes mellitus (17.5%) and chronic heart failure (12.8%).

Phenotypes 3 and 4, which were composed of subjects with moderate to severe airflow limitation (GOLD stages 2 and 3 in about three-quarters of subjects), could not be distinguished based on FEV<sub>1</sub>, but differed in terms of age, symptoms and comorbidities: significant differences (all  $p<0.05$ ) were found for all variables presented in table 3 except for sex ratio, FEV<sub>1</sub>, FVC, HAD anxiety subscale and percentage of subjects treated with tiotropium. Compared with subjects in phenotype 3, subjects in phenotype 4 were older, and had higher prevalence of depressive symptoms and other comorbidities, including cardiovascular comorbidities (especially chronic heart failure). Furthermore, subjects in phenotype 4 had higher BMI and

**TABLE 1** Characteristics of the 322 chronic obstructive pulmonary disease subjects according to Global Initiative for Chronic Obstructive Lung Disease (GOLD) stages

	GOLD 1	GOLD 2	GOLD 3	GOLD 4	p-value
<b>Subjects</b>	21 (6.5)	135 (41.9)	107 (33.2)	59 (18.3)	
<b>Male/female %</b>	71.4/28.6	78.5/21.5	71.0/29.0	84.8/15.2	0.20
<b>Age yrs</b>	66.0 (58.0–75.0)	66.0 (58.0–72.0)	64.0 (57.0–73.0)	63.0 (58.0–72.0)	0.48
<b>Smoking exposure pack-yrs</b>	41.2 (28.0–56.0)	42.0 (26.0–55.0)	38.0 (25.0–50.0)	43.8 (30.0–72.0)	0.74
<b>FEV<sub>1</sub> % pred</b>	84.7 (81.9–86.4)	65.4 (59.1–71.4)	40.3 (34.4–44.8)	25.1 (21.2–28.8)	<0.0001
<b>FVC % pred</b>	109.8 (99.9–115.2)	87.4 (78.8–96.4)	74.6 (59.4–86.8)	57.0 (48.6–70.3)	<0.0001
<b>BMI kg·m<sup>-2</sup></b>	26.5 (24.1–28.1)	25.3 (23.0–29.1)	24.1 (20.4–27.4)	22.3 (18.3–25.6)	<0.0001
<b>MMRC dyspnoea score</b>	1.0 (0.0–1.0)	1.0 (1.0–2.0)	2.0 (1.0–3.0)	3.0 (2.0–4.0)	<0.0001
<b>BOD score</b>	1.0 (1.0–2.0)	2.0 (1.0–2.0)	4.0 (3.0–5.0)	5.0 (4.0–6.0)	<0.0001
<b>SGRQ total score</b>	27.2 (16.3–56.6)	36.3 (26.8–52.3)	50.9 (37.7–61.0)	63.9 (46.4–72.0)	<0.0001
<b>HAD scale</b>					
Total score	15.0 (8.0–19.0)	13.0 (9.0–17.0)	13.0 (8.0–18.0)	14.0 (8.0–21.0)	0.46
Anxiety score	8.0 (5.0–11.0)	7.0 (5.0–10.0)	7.0 (5.0–10.0)	8.0 (4.0–11.0)	0.66
Depression score	6.0 (3.0–10.0)	6.0 (3.0–8.0)	5.0 (3.0–9.0)	7.0 (3.0–11.0)	0.33
<b>Exacerbations per patient per yr</b>	2.0 (0.0–3.0)	1.0 (0.0–2.0)	2.0 (1.0–3.0)	2.0 (1.0–5.0)	<0.0001
<b>Comorbidities %</b>					
Coronary artery disease	19.1	20.3	18.3	5.4	0.06
Chronic heart failure	9.5	17.7	18.3	19.3	0.49
Diabetes mellitus	9.5	5.1	11.5	10.9	0.95
Hypertension	47.6	44.4	35.6	23.2	0.005
<b>Inhaled therapy %</b>					
LABA	52.4	56.3	73.8	84.8	<0.0001
ICS	52.4	54.8	72.9	78.0	<0.0001
Tiotropium	19.1	23.0	17.8	18.6	0.44

Data are presented as n (%) or median (interquartile range), unless otherwise stated. FEV<sub>1</sub>: forced expiratory volume in 1 s; % pred: % predicted; FVC: forced vital capacity; BMI: body mass index; MMRC: modified Medical Research Council; BOD score: BMI, obstruction (FEV<sub>1</sub> % pred) and dyspnoea (MMRC scale); SGRQ: St George's Respiratory Questionnaire; HAD: hospital anxiety and depression; LABA: long-acting  $\beta$ -agonist; ICS: inhaled corticosteroids.



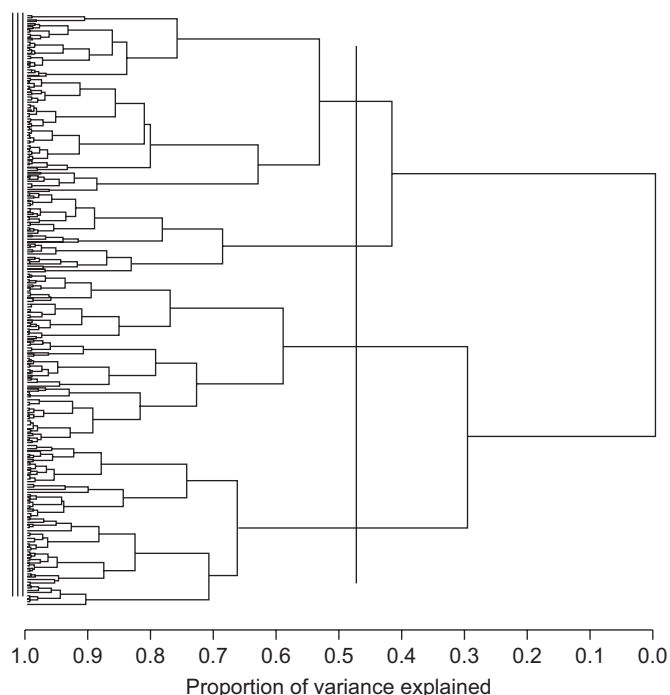
**FIGURE 1.** Dendrogram illustrating the results of the cluster analysis of clinical variables. Clinical variables obtained in 322 chronic obstructive pulmonary disease subjects were classified using VARCLUS cluster analysis. If the patterns of response to two variables were similar for most individuals, these variables were grouped, whereas different response patterns suggested variables were rated more independently. Each horizontal line represents an individual variable and the length of horizontal lines represents the degree of similarity between variables. The original variables could be grouped into three major clusters. MMRC: modified Medical Research Council; SGRQ: St George's Respiratory Questionnaire; HAD: hospital anxiety and depression; BMI: body mass index; FEV<sub>1</sub>: forced expiratory volume in 1 s; % pred: % predicted.

**TABLE 2** Correlations of the eight original variables with the three main components derived from the principal component analysis in the 322 chronic obstructive pulmonary disease subjects

	Components		
	1	2	3
<b>BMI</b>	-0.229487	0.370158	0.622298
<b>MMRC dyspnoea score</b>	0.487056	0.292108	0.134858
<b>Smoking exposure pack-yrs</b>	0.069677	0.454797	-0.415740
<b>HAD total score</b>	0.329744	-0.220896	0.412890
<b>Exacerbations per patient per yr</b>	0.368287	-0.127880	0.096095
<b>Age yrs</b>	-0.026188	0.707358	0.041551
<b>FEV<sub>1</sub> % pred</b>	-0.398280	-0.042196	0.445006
<b>SGRQ total score</b>	0.549161	0.059434	0.205549

The variance of components 1, 2 and 3 were 31.4, 15.7 and 13.7%, respectively. Complementary data and Screeplot depicting eigenvalue of each component are presented in the online supplementary material. BMI: body mass index; MMRC: modified Medical Research Council; HAD: hospital anxiety and depression; FEV<sub>1</sub>: forced expiratory volume in 1 s; % pred: % predicted; SGRQ: St George's Respiratory Questionnaire.





**FIGURE 2.** Dendrogram illustrating the results of the cluster analysis in 322 chronic obstructive pulmonary disease subjects. Subjects were classified using agglomerative hierarchical cluster analysis based on principal component analysis-transformed clinical variables (see Methods section). Each horizontal line represents an individual subject and the length of horizontal lines represents the degree of similarity between subjects. The vertical line identifies the optimal number of clusters in the data, as determined by pseudo-F and pseudo- $t^2$  statistics (see Methods section). Data can be optimally grouped into four clusters (phenotypes). Characteristics of subjects in each phenotype are presented in tables 3 and 4.

more severe dyspnoea, which were responsible for increased BOD scores.

A summary of these four COPD phenotypes is presented in table 4.

#### **Relationship between BOD index and phenotypes**

We examined predicted mortality using the BOD index, which was not used to elaborate our analysis. BOD scores were markedly different among phenotypes, but BOD index was not sufficient for the discrimination of these phenotypes (fig. 3).

#### **DISCUSSION**

We used an original statistical approach to analyse clinical data obtained in a large group of COPD subjects. In this heterogeneous COPD population defined with the current FEV<sub>1</sub>-based GOLD classification, this methodology resulted in the identification of four COPD phenotypes, as follows. Phenotype 1: young subjects with predominant severe to very severe respiratory disease; phenotype 2: older subjects with mild airflow limitation, mild symptoms and mild age-related comorbidities; phenotype 3: young subjects with moderate to severe airflow limitation, but few comorbidities and mild symptoms; and phenotype 4: older subjects with moderate to severe airflow limitation and severe symptoms ascribed, at least in part, to major comorbidities (e.g. chronic heart failure).

Importantly, our results indicated that age, dyspnoea, HRQoL, exacerbations and comorbidities (e.g. chronic heart failure and depression) were markedly different among subjects with the same GOLD classification, underscoring the need for multi-dimensional assessment of COPD subjects.

We searched for COPD phenotypes using cluster analysis, a method for classifying heterogeneous groups of variables into relatively homogeneous groups [11]. Previously, few studies used cluster analysis for assessing phenotypes in patients with airway diseases. These studies were performed in a mixed population of 27 asthmatics and 22 COPD subjects [11], in 175 subjects from a community-based study [20], and in three different populations of asthmatic subjects [21]. Our study is original because we applied this method to a large cohort of well-characterised COPD subjects. This exploratory statistical approach allowed the identification of several COPD phenotypes. Among these, some have been suggested using conventional methods. Indeed, phenotype 1 (subjects with severe respiratory disease and nutritional depletion) and phenotype 4 (mildly overweight subjects with moderate to severe airflow limitation) would correspond to classical descriptions of severe respiratory disease in pink puffers and blue bloaters, respectively [10]. We also identified groups of subjects with milder phenotypes. Thus, phenotype 2 was composed of older subjects in whom mild airflow limitation was accompanied by mild symptoms, few exacerbations and relatively preserved HRQoL. These older subjects were mostly classified as GOLD stage 2, suggesting that severity of airflow limitation in this population is independent from age. Finally, subjects in phenotype 3 were young subjects with moderate to severe airflow limitation and few comorbidities. Longitudinal follow-up will be necessary to improve our knowledge of the natural history of subjects within these phenotypes.

We used PCA as a mean for transforming variables included in the cluster analysis. This methodology, which was applied for the first time in airway diseases, is especially useful for the elimination of noisy variables that may corrupt the cluster structure [18, 22]. To confirm the yield of PCA as a preliminary step before cluster analysis, we performed another cluster analysis based on the total number of initial variables, *i.e.* without previously using PCA for variable reduction. This analysis identified three phenotypes which overlapped markedly for several variables, including age, BMI, dyspnoea and SGRQ (complete data are provided in the online supplementary material). Such overlaps suggest that phenotypes identified by clustering without initial PCA may be less clinically useful, confirming that PCA is a useful way for transforming variables before cluster analysis.

This study has important strengths. First, clinical data were collected prospectively by respiratory physicians. Secondly, the diagnosis of COPD was based on GOLD criteria and validated questionnaires were used to measure patient-related outcomes. Thirdly, the studied population contained subjects in all GOLD classes. Finally, the statistical methods used allow unbiased analyses that are not based on any *a priori* assumptions. Some limitations also have to be taken into account when interpreting the results. Subjects with incomplete datasets were excluded from the analyses, which necessitated complete data. Importantly, no clinically significant difference was found

**TABLE 3** Characteristics of the 322 chronic obstructive pulmonary disease subjects according to the four phenotypes identified using principal component analysis-based cluster analysis

	Phenotype 1	Phenotype 2	Phenotype 3	Phenotype 4
<b>Subjects</b>	44 (13.7)	89 (27.6)	93 (28.9)	96 (29.8)
<b>Male/female %</b>	70.4/29.6	84.3/15.7	74.2/25.8	75.0/25.0
<b>Age yrs</b>	58.0 (55.0–63.0)	68.0 (60.0–74.0)	59.0 (50.0–65.0)	72.5 (67.0–77.0)*
<b>Smoking exposure pack-yrs</b>	39.5 (25.3–50.5)	40.5 (26.3–54.0)	37.5 (27.0–50.0)	45.1 (28.3–72.0)*
<b>FEV<sub>1</sub> % pred</b>	31.2 (21.3–37.5)	68.2 (57.4–75.9)	46.3 (35.3–60.3)	42.9 (32.5–63.5)
<b>FVC % pred</b>	63.3 (55.2–83.2)	88.1 (78.2–99.9)	81.2 (67.9–91.1)	77.8 (57.9–91.8)
<b>GOLD stage %</b>				
1	2.2	14.6	1.1	6.2
2	0	70.8	41.9	34.4
3	47.8	13.5	41.9	36.5
4	50.0	1.1	15.1	22.9
<b>BMI kg·m<sup>-2</sup></b>	19.4 (17.7–23.5)	28.1 (25.2–31.9)	21.6 (19.0–23.7)	26.4 (23.7–30.1)*
<b>MMRC dyspnoea score</b>	3.0 (2.0–4.0)	1.0 (0.0–1.0)	1.0 (1.0–2.0)	3.0 (2.0–3.0)*
<b>BOD score</b>	5.0 (4.0–6.0)	1.0 (1.0–2.0)	3.0 (2.0–3.0)	4.0 (3.0–6.0)*
<b>SGRQ total score</b>	69.5 (59.8–75.4)	27.2 (18.6–34.6)	39.1 (29.2–52.6)	58.5 (46.8–67.0)*
<b>HAD scale</b>				
Total score	20.0 (16.5–24.0)	11.0 (6.0–14.0)	12.0 (7.0–17.0)	14.0 (11.0–20.0)*
Anxiety score	12.0 (9.0–13.5)	6.0 (3.0–8.0)	7.0 (5.0–10.0)	7.5 (5.0–10.5)
Depression score	10.0 (5.5–11.5)	5.0 (2.0–7.0)	4.0 (2.0–7.0)	7.0 (4.0–10.0)*
<b>Exacerbations per patient per yr</b>	4.0 (3.0–6.0)	0.0 (0.0–1.0)	1.0 (0.0–2.0)	2.0 (1.0–3.0)*
<b>Comorbidities %</b>				
Coronary artery disease	14.2	19.5	9.7	22.8*
Chronic heart failure	4.7	12.8	10.8	35.6*
Diabetes mellitus	0.0	17.2	3.2	19.8*
Hypertension	19.1	57.5	20.4	45.7*
<b>Inhaled therapy %</b>				
LABA	84.1	50.6	60.2	81.2*
ICS	88.6	47.1	59.1	76.0*
Tiotropium	18.1	11.2	25.8	24.0

Data are presented as n (%) or median (interquartile range), unless otherwise stated. FEV<sub>1</sub>: forced expiratory volume in 1 s; % pred: % predicted; FVC: forced vital capacity; GOLD: Global Initiative for Chronic Obstructive Lung Disease; BMI: body mass index; MMRC: modified Medical Research Council; BOD score: BMI, obstruction (FEV<sub>1</sub> % pred) and dyspnoea (MMRC scale); SGRQ: St George's Respiratory Questionnaire; HAD: hospital anxiety and depression; LABA: long-acting  $\beta$ -agonist; ICS: inhaled corticosteroids. \*: p<0.05, compared with phenotype 3.

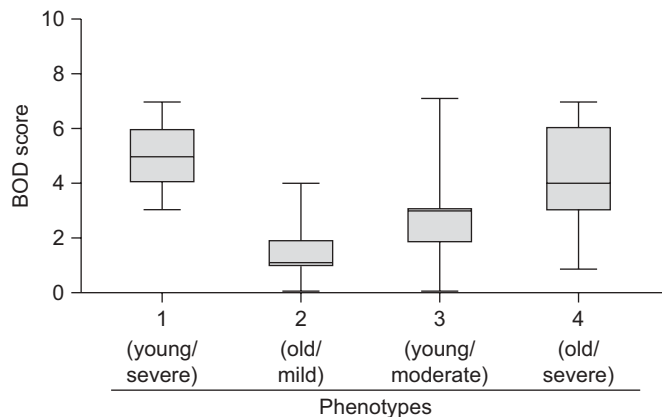
between included and excluded subjects (see Methods section), except for sex, with female subjects representing 23.2% versus 16.0% of subjects included and excluded in the analysis, respectively, suggesting that females are more likely than

males to answer questionnaires. Subjects were recruited in university hospitals and may represent a specific population of COPD subjects. However, the entire range of GOLD severity stages of COPD was represented. Assessment of comorbidities

**TABLE 4** Summary of chronic obstructive pulmonary disease phenotypes identified using principal component analysis-based cluster analysis

	Phenotype 1: young/severe	Phenotype 2: old/mild	Phenotype 3: young/moderate	Phenotype 4: old/severe
<b>Age</b>	Young	Old	Young	Old
<b>Respiratory disease<sup>#</sup></b>	Very severe	Mild	Moderate	Moderate
<b>Nutritional status</b>	Underweight	Overweight	Normal	Overweight
<b>Chronic heart failure</b>	None	None	None	Frequent
<b>Depression</b>	Very frequent	None	None	Frequent
<b>HRQoL impairment</b>	Very severe	Mild	Moderate	Severe

HRQoL: health-related quality of life. #: taking into account airflow obstruction, dyspnoea and exacerbation rate.



**FIGURE 3.** BOD scores (body mass index, obstruction (forced expiratory volume in 1 s % predicted) and dyspnoea evaluated on the modified Medical Research Council scale) in 322 chronic obstructive pulmonary disease (COPD) subjects grouped by phenotypes. COPD subjects ( $n=322$ ) were grouped in four phenotypes, according to the results of the principal component analysis-based cluster analysis. BOD scores were calculated as described previously [14, 15]. Increased BOD score predicts increased mortality. Each box plot is composed of five horizontal lines that display minimum and maximum values, and 25th, 50th and 75th percentiles of the variable.

was based on diagnosed comorbidities and not on systematic diagnostic work-up, preventing us from taking clinically occult diseases into account. Exacerbations were analysed as self-reported exacerbations, which may result in underestimation of exacerbation numbers [23]. However, this approach corresponds to what happens in real life when a physician characterises one of his subjects. Phenotyping was exclusively based on clinical variables, spirometry and questionnaires, but no imaging or biomarkers data were analysed. Our approach was appropriate for identification of clinically based phenotypes that can be used in daily practice. It is possible that inclusion of other variables relevant to the pathogenesis of COPD (e.g. bronchodilator reversibility, peak flow variability, atopy,  $\alpha_1$ -antitrypsin status, emphysema or sputum production, eosinophilic airway inflammation, exhaled nitric oxide fraction or other biomarkers) may have increased our ability to identify phenotypes. It is also possible that current treatments have disease-modifying effects that affected our phenotypes and/or that some treatment responses vary depending on the clinical phenotype. Further studies will be required to explore these hypotheses.

In the present study, we defined COPD using a fixed FEV<sub>1</sub>/FVC <0.7 ratio, which is sex- and age-dependent. The purposes of this choice were to stay in line with the current GOLD guidelines [1], to use the criteria that is mostly referred to in routine practice, and to allow comparison of the data with previous literature. It may have resulted in the exclusion from our cohort of young subjects with mild airflow obstruction (FEV<sub>1</sub>/FVC less than the lower limit normal (LLN) but >0.7). Furthermore, it has resulted in the inclusion of subjects with an FEV<sub>1</sub>/FVC <0.7 but FEV<sub>1</sub>/FVC greater than the LLN ( $n=56$ , 17.3% of subjects). In order to examine whether the definition of airflow limitation using the LLN instead of fixed ratio had an impact on our conclusion, we performed another cluster

analysis based on the 266 patients in this cohort with a FEV<sub>1</sub>/FVC less than the LLN (results are provided in the online supplementary material). An important and reassuring finding was that this analysis identified four phenotypes that were very comparable to those identified by our previous set of analysis in the 322 patients with FEV<sub>1</sub>/FVC <0.7. Thus, although our "mild phenotype in older patient" largely disappeared (due to the removal of older patients with mild airflow limitation using LLN criteria), the other phenotypes were very similar and our main conclusion remains: patients with similar airflow limitation (FEV<sub>1</sub>) had different symptoms (dyspnoea) and outcomes (exacerbation numbers and predicted mortality), and differed in terms of age and comorbidities, further indicating the robustness of our analyses.

There was a strong inverse correlation between GOLD classification and BMI. These data confirm and extend findings by VESTBO *et al.* [24], who reported that BMI was reduced in GOLD stage 4 subjects compared with subjects with milder airflow obstruction. However, the results of our PCA-based cluster analysis suggested that the relationship between BMI and FEV<sub>1</sub> is affected by age, with higher BMI found in older subjects.

We found major differences in the levels of dyspnoea in subjects classified at similar GOLD stages. Cluster analysis of the relationships between clinical variables indicated that dyspnoea showed only moderate correlation with FEV<sub>1</sub>, confirming previous studies [25]. Comparing subjects who could not be differentiated based on FEV<sub>1</sub> (phenotypes 3 and 4), we found that subjects with increased dyspnoea (phenotype 4) were older, had increased prevalence of chronic heart failure, and were mildly overweight. We speculate that chronic heart failure and reduced physical activity may explain, at least in part, increased dyspnoea in these subjects. Indeed, other studies found that daily activity (which was not assessed in our cohort) is decreased even in subjects with mild to moderate airflow limitation, and that this decrease is associated with increased dyspnoea, chronic heart failure and increased mortality [26, 27].

Although BOD scores (and, therefore, predicted mortality) differed significantly between the four identified phenotypes, a marked overlap was observed. Indeed, these phenotypes take into account important patient characteristics that are not captured in the BOD score (e.g. age, exacerbation history, quality of life, comorbidities and depression).

These data have important implications for patient care. International guidelines recommend adaptation of therapies based on severity of COPD, as assessed by post-bronchodilator FEV<sub>1</sub> and symptoms [1]. We suggest that this strategy is appropriate for subjects with predominant respiratory disease (phenotype 1), but not in subjects with important extrapulmonary disease, *i.e.* comorbidities (phenotype 4). The development of patient-oriented rather than single-disease guidelines may prove very useful for management of subjects with multiple chronic diseases.

These data also have major implication for clinical trials. We showed that subjects with similar GOLD stages had very different clinical characteristics, including symptoms, comorbidities and predicted mortality (as determined using BOD

index). We speculate that the inclusion of subjects with similar FEV<sub>1</sub>, but different risks or causes of mortality, may have resulted in negative results reported in large therapeutic trials of inhaled therapies in COPD subjects [5, 6]. Indeed, relative mortality risk reduction depends not only on beneficial effect of treatment but also on the distribution of baseline mortality risk in the population and on the causes of mortality in each subgroup of subjects [28]. This has been best underscored in subjects with high blood pressure for whom mortality risk depends not only on blood pressure levels but also on various coexisting conditions (e.g. age, diabetes, high cholesterol level and smoking) [28]. We suggest that future clinical studies should analyse results based on risk assessment (e.g. mortality risk) rather than on a single parameter (e.g. FEV<sub>1</sub>).

In summary, current FEV<sub>1</sub>-based GOLD classification appears inappropriate for guiding therapy and for stratification of subjects in clinical trials, because it does not discriminate subjects with markedly different phenotypes. Our study described an original statistical methodology that allowed the identification of clinical COPD phenotypes. This methodology could be applied to other COPD cohorts to examine whether similar, or different, phenotypes are present in different populations. The prognostic value of these phenotypes should also be evaluated in longitudinal studies. Such studies will provide data on the relevance of these new phenotypes and will allow comparisons of outcome prediction between phenotypes and the GOLD classification or composite indices. We propose that dissemination of this original approach could result in better phenotypic characterisation, which may prove useful in daily practice and clinical trials. We further propose that data from large clinical trials should be re-analysed using this methodology for classification of patients according to their clinical characteristics at study entry.

#### SUPPORT STATEMENT

This work was funded by unrestricted grants from Boehringer Ingelheim France (Paris, France) and Pfizer France (Paris, France).

#### STATEMENT OF INTEREST

A statement of interest for this study can be found at [www.erj.ersjournals.com/misc/statements.dtl](http://www.erj.ersjournals.com/misc/statements.dtl)

#### ACKNOWLEDGEMENTS

The Initiatives BPCO study group: G. Brinchault-Rabin (Rennes), P-R. Burgel (Cochin, Paris), D. Caillaud (Clermont-Ferrand), P. Carré (Carcassonne), P. Chanez (Marseille), A. Chaouat (Vandoeuvre les Nancy), I. Court-Fortune (Saint-Etienne), A. Cuvelier (Rouen), R. Escamilla (Toulouse), C. Gut-Gobert (Brest), G. Jebrak (Paris), F. Lemoigne (Nice), P. Nesme-Meyer (Lyon), T. Perez and I. Tillie-Leblond (Lille), C. Perrin (Cannes), C. Pinet (Toulon), C. Raheison (Bordeaux) and N. Roche (Hôtel Dieu, Paris, France).

#### REFERENCES

- 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: 532–555.
- Celli BR, Cote CG, Lareau SC, *et al.* Predictors of survival in COPD: more than just the FEV<sub>1</sub>. *Respir Med* 2008; 102: Suppl. 1, S27–S35.
- Mannino DM, Thorn D, Swensen A, *et al.* Prevalence and outcomes of diabetes, hypertension and cardiovascular disease in COPD. *Eur Respir J* 2008; 32: 962–969.
- McGarvey LP, John M, Anderson JA, *et al.* Ascertainment of cause-specific mortality in COPD: operations of the TORCH Clinical Endpoint Committee. *Thorax* 2007; 62: 411–415.
- Calverley PM, Anderson JA, Celli B, *et al.* Salmeterol and fluticasone propionate and survival in chronic obstructive pulmonary disease. *N Engl J Med* 2007; 356: 775–789.
- Tashkin DP, Celli B, Senn S, *et al.* A 4-year trial of tiotropium in chronic obstructive pulmonary disease. *N Engl J Med* 2008; 359: 1543–1554.
- Criner GJ, Sternberg AL. National Emphysema Treatment Trial: the major outcomes of lung volume reduction surgery in severe emphysema. *Proc Am Thorac Soc* 2008; 5: 393–405.
- Celli BR. Roger S. Mitchell lecture. Chronic obstructive pulmonary disease phenotypes and their clinical relevance. *Proc Am Thorac Soc* 2006; 3: 461–465.
- Turino GM. COPD and biomarkers: the search goes on. *Thorax* 2008; 63: 1032–1034.
- Dornhorst AC. Respiratory insufficiency (Frederick Price Memorial Lecture). *Lancet* 1955; 1: 1185–1187.
- Wardlaw AJ, Silverman M, Siva R, *et al.* Multi-dimensional phenotyping: towards a new taxonomy for airway disease. *Clin Exp Allergy* 2005; 35: 1254–1262.
- Burgel PR, Nesme-Meyer P, Chanez P, *et al.* Cough and sputum production are associated with frequent exacerbations and hospitalizations in COPD subjects. *Chest* 2009; 135: 975–982.
- Quanjer PH, Tammeling GJ, Cotes JE, *et al.* Lung volumes and forced ventilatory flows. *Eur Respir J* 1993; 6: Suppl. 16, 5–40.
- Celli B, Jones P, Vestbo J, *et al.* The multidimensional BOD: association with mortality in the TORCH trial. *Eur Respir J* 2008; 32: Suppl. 52, 42s.
- Celli BR, Cote CG, Marin JM, *et al.* The body-mass index, airflow obstruction, dyspnea, and exercise capacity index in chronic obstructive pulmonary disease. *N Engl J Med* 2004; 350: 1005–1012.
- Ng TP, Niti M, Tan WC, *et al.* Depressive symptoms and chronic obstructive pulmonary disease: effect on mortality, hospital readmission, symptom burden, functional status, and quality of life. *Arch Intern Med* 2007; 167: 60–67.
- 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; 145: 1321–1327.
- Vogt W, Nagel D. Cluster analysis in diagnosis. *Clin Chem* 1992; 38: 182–198.
- Casanova C, Cote C, de Torres JP, *et al.* Inspiratory-to-total lung capacity ratio predicts mortality in patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2005; 171: 591–597.
- Weatherall M, Travers J, Shirtcliffe PM, *et al.* Distinct clinical phenotypes of airways disease defined by cluster analysis. *Eur Respir J* 2009; 34: 812–818.
- Haldar P, Pavord ID, Shaw DE, *et al.* Cluster analysis and clinical asthma phenotypes. *Am J Respir Crit Care Med* 2008; 178: 218–224.
- Ben-Hur A, Guyon I. Detecting stable clusters using principal component analysis. In: Brownstein MJ, Kohodursky A, eds. *Functional Genomics: Methods and Protocols*. Totowa, Humana Press, 2003; pp. 159–182.
- Seemungal TA, Donaldson GC, Paul EA, *et al.* Effect of exacerbation on quality of life in patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 1998; 157: 1418–1422.
- Vestbo J, Prescott E, Almdal T, *et al.* Body mass, fat-free body mass, and prognosis in patients with chronic obstructive pulmonary disease from a random population sample: findings from the Copenhagen City Heart Study. *Am J Respir Crit Care Med* 2006; 173: 79–83.



- 25** Curtis JR, Deyo RA, Hudson LD. Pulmonary rehabilitation in chronic respiratory insufficiency. 7. Health-related quality of life among patients with chronic obstructive pulmonary disease. *Thorax* 1994; 49: 162–170.
- 26** Garcia-Aymerich J, Serra I, Gomez FP, *et al.* Physical activity and clinical and functional status in COPD. *Chest* 2009; 136: 62–70.
- 27** Watz H, Waschki B, Boehme C, *et al.* Extrapulmonary effects of chronic obstructive pulmonary disease on physical activity: a cross-sectional study. *Am J Respir Crit Care Med* 2008; 177: 743–751.
- 28** Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *JAMA* 2007; 298: 1209–1212.