ERJ Express. Published on January 14, 2010 as doi: 10.1183/09031936.00175109

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

Pierre-Régis Burgel MD PhD[1], Jean-Louis Paillasseur PhD[2], Denis Caillaud MD[3], Isabelle Tillie-Leblond MD PhD[4], Pascal Chanez MD PhD[5], Roger Escamilla MD[6], Isabelle Court-Fortune MD[7], Thierry Perez MD[4], Philippe Carré MD[8], Nicolas Roche MD PhD[9], on behalf of the Initiatives BPCO Scientific Committee

[1] Service de Pneumologie, Hôpital Cochin, AP-HP and Paris-Descartes University, Paris, France
[2] Clindatafirst, St Quentin en Yvelines, France.
[3] Service de Pneumologie, Hôpital Gabriel Montpied, CHU Clermont-Ferrand, France
[4] Service de Pneumologie, Hôpital Calmette, Lille, France
[5] Département des Maladies Respiratoires, AP-HM, Université de la Méditerranée, Marseille, France
[6] Clinique des voies respiratoires Hopital Larrey, Toulouse, France
[7] Service de Pneumologie, CHU Saint Etienne, France
[8] Service de Pneumologie, Hôpital Antoine Gayrard, Carcassonne, France
[9] Service de Pneumologie, Hôpital de l'Hôtel Dieu, AP-HP and Paris-Descartes University, Paris, France

Word count : 3400.

Short Title: Clinical COPD phenotypes

Keywords (max 6): principal component analysis, cluster analysis, COPD comorbidities, COPD phenotypes, dyspnea.

Address for correspondence:
Dr Pierre-Régis Burgel
Service de Pneumologie, Hôpital Cochin,
Assistance Publique Hôpitaux de Paris,
27 rue du Faubourg St Jacques
75679 Paris Cedex 14, France
Ph : 33 1 58 41 23 71
Fax : 33 1 46 33 82 53
Email : pierre-regis.burgel@cch.aphp.fr

Abstract (196 words)
Rationale: Classification of 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.

Methods: COPD subjects recruited in a French multicenter cohort were characterized using a standardized process. Principal component analysis (PCA) was performed using eight variables selected for their relevance to COPD: age, cumulative smoking, FEV_1 (% predicted), body mass index, exacerbations, dyspnea (MMRC scale), health status (St Georges Respiratory Questionnaire), and depressive symptoms (Hospital Anxiety-Depression scale). Patient classification was performed using cluster analysis based on PCA-transformed data.

Results: Data are median [IQR]. 322 COPD subjects were analyzed: male 77%, age 65.0 yr [58.0; 73.0], FEV_1 48.9 % [34.1; 66.3], GOLD 1/2/3/4: 21/135/107/59 subjects. PCA showed that three independent components accounted for 61% of variance. PCA-based cluster analysis resulted in the classification of subjects into 4 clinical phenotypes that could not be identified using GOLD classification. Importantly, subjects with comparable airflow limitation (FEV_1) belonged to different phenotypes and had marked differences in age, symptoms, comorbidities, and predicted mortality.

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

Introduction

Chronic obstructive pulmonary disease (COPD) is a major cause of mortality and disability worldwide [1]. The disease is characterized 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 one second (FEV_1) [1]. In recent years, it has emerged that COPD is a complex disease with multiple clinical manifestations and that COPD subjects cannot be described only using the severity of airflow limitation. Thus, many other independent predictors of outcomes have been identified, including worsening dyspnea, frequency and severity of exacerbations, malnutrition, depression, and health-related quality of life (HRQoL) impairment [2]. Further, comorbidities (e.g., cardiovascular diseases and cancer) are major causes of death and hospitalizations in COPD subjects [3, 4].

Large clinical trials performed in COPD subjects have shown that current treatments improved several outcomes (e.g., exacerbations, dyspnea, HRQoL), but the authors reported disappointing data on mortality and rates of decline in FEV_1 [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 Therapy 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 characterization 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 categorization of subjects would be of utmost importance in these studies. Further, 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 1950's, when Dornhorst proposed the distinction between pink puffers and blue bloaters [10]. 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 organize 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 analyze clinical data obtained in a well-characterized 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-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 wk) with a diagnosis of COPD based on the presence of a post-bronchodilator FEV_1/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 standardized characterization 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 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 (B: Body mass index, O: obstruction (FEV_1%), D: dyspnea evaluated on the modified Medical Research Council – MMRC- scale), which was reported to be a better predictor of mortality than FEV_1 [14, 15].

The hospital anxiety and depression (HAD) scale was used to examine mood disorders. This 14-item self-questionnaire has two 7-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. Health related quality of life 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 analyzed. Correlations between GOLD classes [1] and other variables were assessed using Kendall τ_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, principal component analysis (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; IQR] or % unless otherwise specified. A $P<0.05$ was considered statistically significant. Analyses were performed using the SAS® 9.1.3 statistical software.

Selection of variables for analyses

The Scientific Committee selected 8 variables for their relevance to pulmonary and/or extrapulmonary manifestations of COPD. The variables were: age (yr), tobacco smoking (pack-

yr), severity of airflow obstruction (assessed by FEV_1, % predicted), exacerbations (number/patient/yr), nutritional status (assessed by body mass index –BMI-, kg/m^2), dyspnea (assessed by the –MMRC- scale), heath-related quality of life (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 8 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_1, MMRC scale, BMI and exacerbations/patient/yr (see **online supplement**). Male subjects represented 76.7 vs. 84.0 % subjects included and excluded in the analysis, respectively (P=0.03, chi square test).

Correlations between clinical variables

Relationships between the 8 selected variables were studied by cluster analysis, using the VARCLUS procedure. This procedure, which organizes 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 principal component analysis (PCA) [18]. Linear combinations of the 8 selected variables were used to form 8 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 the 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 stages

Clinical characteristics of the 322 COPD subjects according to GOLD stages are presented in **Table 1**. Variables correlated with increasing GOLD stages included FEV_1, FVC and BMI (inverse correlations), and MMRC, BOD score, SGRQ total score and numbers of exacerbations/patient/yr (positive correlations). Systemic hypertension was less prevalent in subjects with GOLD stage 3 and 4, 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, diabetes mellitus).

Long-acting beta agonists and inhaled corticosteroids (ICS) were prescribed in about 50% subjects in GOLD 1 and 2, and in up to 84% of subjects in GOLD 4. No significant difference was observed among GOLD classes for tiotropium prescription.

Table 1. Characteristics of the 322 COPD subjects according to GOLD stages.

	GOLD 1 n=21 (6.5%)	**GOLD 2** n=135 (41.9%)	**GOLD 3** n=107 (33.2%)	**GOLD 4** n=59 (18.3%)	*P* values
Male/Female, %	71.4/28.6	78.5/21.5	71.0/29.0	84.8/15.2	0.20
Age, yr	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
Smoking, pack-yr	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
FEV_1, % pred	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
FVC, % pred	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
BMI, kg/m^2	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
MMRC	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
BOD score*	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
SGRQ, total score	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
HAD scale					
Total scale	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 (HAD-A)	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 (HAD-D)	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
Exacerbation/patient/yr	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
Comorbidities					
CAD %	19.1	20.3	18.3	5.4	0.06
CHF %	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
Inhaled therapy					
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

All data are median [IQR] unless otherwise specified. MMRC: Modified Medical Research Council scale.* BOD score: Body mass index (BMI), Obstruction (FEV_1 %), Dyspnea (MMRC). SGRQ: St Georges Respiratory Questionnaire. HAD: Hospital Anxiety Depression Scale. CAD: coronary artery disease ; CHF: chronic heart failure. LABA: long acting beta agonist; ICS: inhaled corticosteroids.

Relationships between clinical variables

Cluster analysis of the 8 selected variables resulted in a dendrogram (**Figure 1),** which illustrates how these variables relate to each other. This analysis showed that clinical information obtained using these 8 variables could be grouped into 3 clusters (**Figure 1**), indicating that these variables were not completely independent.

Principal component analysis of clinical variables

We performed principal component analysis (PCA) to transform data contained in the 8 selected variables into 8 independent components. The first 3 components that contributed significantly to explain the relationships among the 8 selected variables (eigenvalues>1) accounted for 61% of the information. Correlations of the selected variables with these 3 independent components are shown in **Table 2**. Component 1 correlated with SGRQ total score, and MMRC score, and inversely correlated with FEV_1 (% pred), but was independent from age. Component 2 highly correlated with age and cumulative tobacco-smoking, but was independent from FEV_1 (% pred) and SGRQ score. Component 3 mostly correlated with BMI and FEV_1 (% pred). Components 4 to 8 explained little variability of the original data (eigenvalues<1, **see Online supplement**) and therefore were not subjected to further analysis.

Table 2. Correlations of the 8 original variables with the 3 main components derived from the principal component analysis in COPD subjects (n=322).

Components (% variance) VARIABLES	Comp 1 (31.4%)	Comp 2 (15.7%)	Comp 3 (13.7%)
BMI	-.229487	0.370158	0.622298
MMRC	0.487056	0.292108	0.134858
Pack-yr	0.069677	0.454797	-.415740
HAD total score	0.329744	-.220896	0.412890
Exacerbations/patient/yr	0.368287	-.127880	0.096095
Age, yr	-.026188	0.707358	0.041551
FEV_1 % pred	-.398280	-.042196	0.445006
SGRQ total score	0.549161	0.059434	0.205549

Complementary data and Screeplot depicting eigenvalue of each component are presented in an **Online supplement**.

Classification of COPD subjects using cluster analysis

Classification of the 322 COPD subjects using cluster analysis based on the first 3 components identified in the PCA resulted in a dendrogram that showed the progressive joining of the clustering process (**Figure 2**). Pseudo F and pseudo t^2 statistics determined that the data could be optimally grouped into 4 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 yr) with severe airflow limitation (GOLD 3 and 4), low BMI, severe dyspnea, 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 yr) with mild airflow limitation (GOLD 1 or 2 in 85.4% subjects), mild overweight, low dyspnea, low levels of anxiety and depression, almost no exacerbations, and mild impairment in HRQoL. 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 2 and 3 in about ¾ of subjects), could not be distinguished based on FEV_1, but differed in terms of age, symptoms and comorbidities: significant differences (each, $P<0.05$) were found for all variables presented in Table 3 except for sex ratio, FEV_1, FVC, HAD anxiety subscale and % 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). Further,

subjects in phenotype 4 had higher BMI and more severe dyspnea, which were responsible for increased BOD scores.

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

Table 3. Characteristics of the 322 COPD subjects according to the four phenotypes identified using principal component analysis-based cluster analysis

	Phenotype 1 n=44 (13.7%)	**Phenotype 2** n=89 (27.6%)	**Phenotype 3** n=93 (28.9%)	**Phenotype 4** n=96 (29.8%)
Male/Female, %	70.4 / 29.6	84.3 / 15.7	74.2 / 25.8	75.0 / 25.0
Age, yr	58.0 [55.0 ; 63.0]	68.0 [60.0 ; 74.0]	59.0 [50.0 ; 65.0]	72.5 [67.0 ; 77.0]**
Smoking, pack-yr	39.5 [25.3 ; 50.5]	40.5 [26.3 ; 54.0]	37.5 [27.0 ; 50.0]	45.1 [28.3 ; 72.0]**
FEV_1, % pred	31.2 [21.3 ; 37.5]	68.2 [57.4 ; 75.9]	46.3 [35.3 ; 60.3]	42.9 [32.5 ; 63.5]
FVC, % pred	63.3 [55.2 ; 83.2]	88.1 [78.2 ; 99.9]	81.2 [67.9 ; 91.1]	77.8 [57.9 ; 91.8]
GOLD 1, %	2.2	14.6	1.1	6.2
GOLD 2, %	0	70.8	41.9	34.4
GOLD 3, %	47.8	13.5	41.9	36.5
GOLD 4, %	50.0	1.1	15.1	22.9
BMI, kg/m^2	19.4 [17.7 ; 23.5]	28.1 [25.2 ; 31.9]	21.6 [19.0 ; 23.7]	26.4 [23.7 ; 30.1]**
MMRC	3.0 [2.0 ; 4.0]	1.0 [0.0 ; 1.0]	1.0 [1.0 ; 2.0]	3.0 [2.0 ; 3.0]**
BOD score*	5.0 [4.0 ; 6.0]	1.0 [1.0 ; 2.0]	3.0 [2.0 ; 3.0]	4.0 [3.0 ; 6.0]**
SGRQ, total score	69.5 [59.8 ; 75.4]	27.2 [18.6 ; 34.6]	39.1 [29.2 ; 52.6]	58.5 [46.8; 67.0]**
HAD scale				
Total scale	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 (HAD-A)	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 (HAD-D)	10.0 [5.5 ; 11.5]	5.0 [2.0 ; 7.0]	4.0 [2.0 ; 7.0]	7.0 [4.0 ; 10.0]**
Exacerbation/patient/yr	4.0 [3.0 ; 6.0]	0.0 [0.0 ; 1.0]	1.0 [0.0 ; 2.0]	2.0 [1.0 ; 3.0]**
Comorbidities				
CAD %	14.2	19.5	9.7	22.8**
CHF %	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**
Inhaled therapy				
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

All data are median [IQR] unless otherwise specified. * BOD score: Body mass index, Obstruction (FEV_1 %), Dyspnea (MMRC). CAD: coronary artery disease; CHF: chronic heart failure. LABA: long acting beta agonist; ICS: inhaled corticosteroids.
** $P<0.05$ compared with phenotype 3

Table 4. Summary of COPD phenotypes identified using PCA-based cluster analysis

	Phenotype 1: young/severe	**Phenotype 2: old/mild**	**Phenotype 3: young/moderate**	**Phenotype 4: old/severe**
Age	young	old	young	old
Respiratory Disease*	very severe	mild	moderate	moderate
Nutritional status	underweight	overweight	normal	overweight
Chronic heart failure	none	none	none	frequent
Depression	very frequent	none	none	frequent
HRQoL impairment	very severe	mild	moderate	severe

* Taking into account airflow obstruction, dyspnea and exacerbation rate.

Relationship between BOD index and phenotypes

We examined predicted mortality using BOD index, which was not used to elaborate our analysis. BOD scores were markedly different among phenotypes, but BOD index was not sufficient to discriminate these phenotypes (see **Figure 3**).

Discussion

We used an original statistical approach to analyze clinical data obtained in a large group of COPD subjects. In this heterogeneous COPD population defined with the current FEV_1-based GOLD classification, this methodology resulted in the identification of four COPD phenotypes. These phenotypes included: (A) young subjects with predominant severe to very severe respiratory disease (phenotype 1) (B) older subjects with mild airflow limitation, mild symptoms and mild age-related comorbidities (phenotype 2) (C) young subjects with moderate to severe airflow limitation, but few comorbidities and mild symptoms (phenotype 3) and (D) older subjects with moderate to severe airflow limitation and severe symptoms ascribed, at least in part, to major comorbidities (e.g., chronic heart failure) (phenotype 4). Importantly, our results indicated that age, dyspnea, HRQoL, exacerbations and comorbidities (e.g., chronic heart failure, depression) were markedly different among subjects in the same GOLD class, underscoring the need for multidimensional 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 3 different populations of asthmatic subjects [21]. Our study is original because we applied this method to a large cohort of well-characterized COPD subjects. This exploratory statistical approach allowed identifying 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 (subjects with mild overweight and 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 in 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 principal component analysis 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 to eliminate noisy variable 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 3 phenotypes which overlapped markedly for several variables including age, BMI, dyspnea, SGRQ (*complete data are provided in the online supplement*). 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. Firstly, 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. Last, statistics used allow unbiased analyses that are not based on any a priori assumption. 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

between included and excluded subjects (see Methods) except for gender, female subjects representing 23.2% vs. 16.0% of subjects included and excluded in the analysis, respectively, suggesting that women are more prone than men 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 was based on diagnosed comorbidities and not on systematic diagnostic work-up, preventing us from taking clinically occult diseases into account. Exacerbations were analyzed as self-reported exacerbations, which may result in underestimation of exacerbation numbers [23]. On the other hand, this approach corresponds to what happens in the real life when a physician characterizes one of his subjects. Phenotyping was exclusively based on clinical variables, spirometry and questionnaires, but no imaging or biomarkers data were analyzed. 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, FeNO, or other biomarkers) may have increased our ability for phenotype identification. It is also possible that current treatment 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 $FEV_1/FVC<0.7$ 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 ($FEV_1/FVC<$lower limit normal-LLN but >0.7). Further, it has resulted in the inclusion of subjects with an FEV1/FVC<0.7 but

FEV_1/FVC>LLN (n=56, 17.3% of subjects). To examine whether the definition of airflow limitation using 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_1/FVC<LLN (*results are provided in the online supplement*). An important and reassuring finding was that this analysis identified 4 phenotypes that were very comparable to those identified by our previous set of analysis in the 322 patients with FEV_1/FVC<0.7. Thus, although our "mild phenotype in older patient" largely disappeared (due to the removal of mainly mild older patients using <LLN), the other phenotypes were very similar and our main conclusion remain: Patients with similar airflow limitation (FEV_1) had different symptoms (dyspnea), outcomes (exacerbation numbers, 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. who reported that BMI was reduced in GOLD stage 4 subjects compared with subjects with milder airflow obstruction [24]. However, results of our PCA-based cluster analysis suggested that the relationship between BMI and FEV_1 is affected by age, with higher BMI found in older subjects.

We found major differences in the levels of dyspnea in subjects with similar GOLD stages. Cluster analysis of the relationships between clinical variables indicated that dyspnea showed only moderate correlation with FEV_1, confirming previous studies [25]. Comparing subjects who could not be differentiated based on FEV_1 (phenotypes 3 and 4), we found that subjects with increased dyspnea (phenotype 4) were older, had increased prevalence of chronic heart failure, and mild overweight. We speculate that chronic heart failure and reduced physical activity may explain, at least in part, increased dyspnea 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 dyspnea, 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_1 and symptoms [1]. We suggest that this strategy is appropriate for subjects with predominant respiratory disease (e.g., phenotype 1) but not in subjects with important extrapulmonary disease –i.e., comorbidities- (e.g., 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_1, 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, smoking) [28].

We suggest that future clinical studies should analyze results based on risk assessment (e.g., mortality risk) rather than on single parameter (e.g., FEV_1).

In summary, current FEV_1-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 identifying clinical COPD phenotypes. This methodology could be applied to other COPD cohorts to examine whether similar or different phenotypes are present in different populations. 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 comparison 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 characterization, which may prove useful in daily practice and clinical trials. We further propose that data from large clinical trials should be re-analyzed using this methodology for classification of patients according to their clinical characteristics at study entry.

ACKNOWLEDGEMENTS:

The Initiatives BPCO study group:

Graziella Brinchault-Rabin (Rennes), Pierre-Régis Burgel (Paris, Cochin), Denis Caillaud (Clermont-Ferrand), Philippe Carré (Carcassonne), Pascal Chanez (Marseille), Ari Chaouat (Vandoeuvre les Nancy), Isabelle Court-Fortune (Saint-Etienne), Antoine Cuvelier (Rouen), Roger Escamilla (Toulouse), Christophe Gut-Gobert (Brest), Gilles Jebrak (Paris), Franck Lemoigne (Nice), Pascale Nesme-Meyer (Lyon), Thierry Perez and Isabelle Tillie-Leblond (Lille), Christophe Perrin (Cannes), Christophe Pinet (Toulon), Chantal Raherison (Bordeaux), Nicolas Roche (Paris, Hôtel Dieu).

FUNDING: This work was funded by unrestricted grants from Boehringer Ingelheim France and Pfizer.

References

1. Rabe KF, Hurd S, Anzueto A, Barnes PJ, Buist SA, Calverley P, Fukuchi Y, Jenkins C, Rodriguez-Roisin R, van Weel C, Zielinski J. 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.
2. Celli BR, Cote CG, Lareau SC, Meek PM. Predictors of Survival in COPD: more than just the FEV1. *Respir Med* 2008; 102 Suppl 1: S27-35.
3. Mannino DM, Thorn D, Swensen A, Holguin F. Prevalence and outcomes of diabetes, hypertension and cardiovascular disease in COPD. *Eur Respir J* 2008; 32: 962-969.
4. McGarvey LP, John M, Anderson JA, Zvarich M, Wise RA. Ascertainment of cause-specific mortality in COPD: operations of the TORCH Clinical Endpoint Committee. *Thorax* 2007; 62: 411-415.
5. Calverley PM, Anderson JA, Celli B, Ferguson GT, Jenkins C, Jones PW, Yates JC, Vestbo J. Salmeterol and fluticasone propionate and survival in chronic obstructive pulmonary disease. *N Engl J Med* 2007; 356: 775-789.
6. Tashkin DP, Celli B, Senn S, Burkhart D, Kesten S, Menjoge S, Decramer M. A 4-year trial of tiotropium in chronic obstructive pulmonary disease. *N Engl J Med* 2008; 359: 1543-1554.
7. 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.
8. Celli BR. Roger s. Mitchell lecture. Chronic obstructive pulmonary disease phenotypes and their clinical relevance. *Proc Am Thorac Soc* 2006; 3: 461-465.
9. Turino GM. COPD and biomarkers: the search goes on. *Thorax* 2008; 63: 1032-1034.
10. Dornhorst AC. Respiratory insufficiency (Frederick Price Memorial Lecture). *Lancet* 1955; 1: 1185-1187.
11. Wardlaw AJ, Silverman M, Siva R, Pavord ID, Green R. Multi-dimensional phenotyping: towards a new taxonomy for airway disease. *Clin Exp Allergy* 2005; 35: 1254-1262.
12. Burgel PR, Nesme-Meyer P, Chanez P, Caillaud D, Carre P, Perez T, Roche N. Cough and sputum production are associated with frequent exacerbations and hospitalizations in COPD subjects. *Chest* 2009; 135: 975-982.
13. Quanjer PH, Tammeling GJ, Cotes JE, Pedersen OF, Peslin R, Yernault JC. Lung volumes and forced ventilatory flows. Report Working Party Standardization of Lung Function Tests, European Community for Steel and Coal. Official Statement of the European Respiratory Society. *Eur Respir J Suppl* 1993; 16: 5-40.
14. Celli B, Jones P, Vestbo J, Anderson J, Ferguson G, Yates J, Jenkins C, Calverley P. The multidimensional BOD: association with mortality in the TORCH trial. *Eur Resp J Suppl* 2008: A384.
15. Celli BR, Cote CG, Marin JM, Casanova C, Montes de Oca M, Mendez RA, Pinto Plata V, Cabral HJ. The body-mass index, airflow obstruction, dyspnea, and exercise capacity index in chronic obstructive pulmonary disease. *N Engl J Med* 2004; 350: 1005-1012.
16. Ng TP, Niti M, Tan WC, Cao Z, Ong KC, Eng P. 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.
17. Jones PW, Quirk FH, Baveystock CM, Littlejohns P. 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.

18. Vogt W, Nagel D. Cluster analysis in diagnosis. *Clin Chem* 1992; 38: 182-198.
19. Casanova C, Cote C, de Torres JP, Aguirre-Jaime A, Marin JM, Pinto-Plata V, Celli BR. 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.
20. Weatherall M, Travers J, Shirtcliffe PM, Marsh SE, Williams MV, Nowitz MR, Aldington S, Beasley R. Distinct clinical phenotypes of airways disease defined by cluster analysis. *Eur Respir J* 2009; 34: 812-818.
21. Haldar P, Pavord ID, Shaw DE, Berry MA, Thomas M, Brightling CE, Wardlaw AJ, Green RH. Cluster analysis and clinical asthma phenotypes. *Am J Respir Crit Care Med* 2008; 178: 218-224.
22. Ben-Hur A, Guyon I. Detecting stable clusters using principal component analysis. *In:* Brownstein MJ, Kohodursky A, eds. Functional genomics: methods and protocols. Humana press, 2003; pp. 159-182.
23. Seemungal TA, Donaldson GC, Paul EA, Bestall JC, Jeffries DJ, Wedzicha JA. Effect of exacerbation on quality of life in patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 1998; 157: 1418-1422.
24. Vestbo J, Prescott E, Almdal T, Dahl M, Nordestgaard BG, Andersen T, Sorensen TI, Lange P. 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, Farrero E, Balcells E, Rodriguez DA, de Batlle J, Gimeno E, Donaire-Gonzalez D, Orozco-Levi M, Sauleda J, Gea J, Rodriguez-Roisin R, Roca J, Agusti AG, Anto JM. Physical activity and clinical and functional status in COPD. *Chest* 2009; 136: 62-70.
27. Watz H, Waschki B, Boehme C, Claussen M, Meyer T, Magnussen H. 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.

Figure Legends

Figure 1. Dendrogram illustrating the results of the cluster analysis of clinical variables.

Clinical variables obtained in 322 COPD 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.

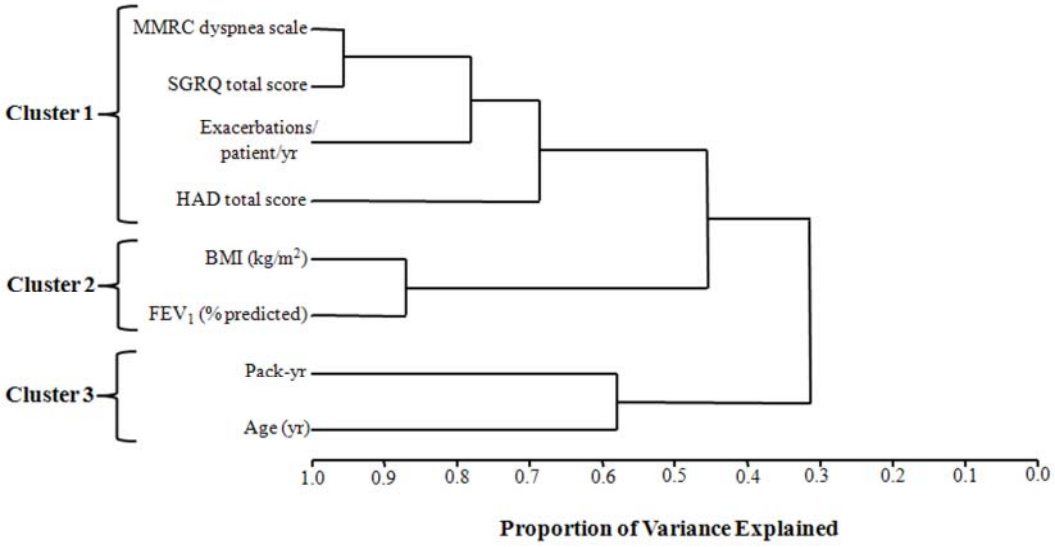


Figure 2: Dendrogram illustrating the results of the cluster analysis in 322 COPD subjects.

Subjects were classified using agglomerative hierarchical cluster analysis based on principal component analysis-transformed clinical variables (see Methods). 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 cluster in the data, as determined by pseudo *F* and pseudo t^2 statistics (See Methods). Data can be optimally grouped

into 4 clusters (phenotypes). Characteristics of subjects in each phenotype are presented in **Table 3 and 4**.

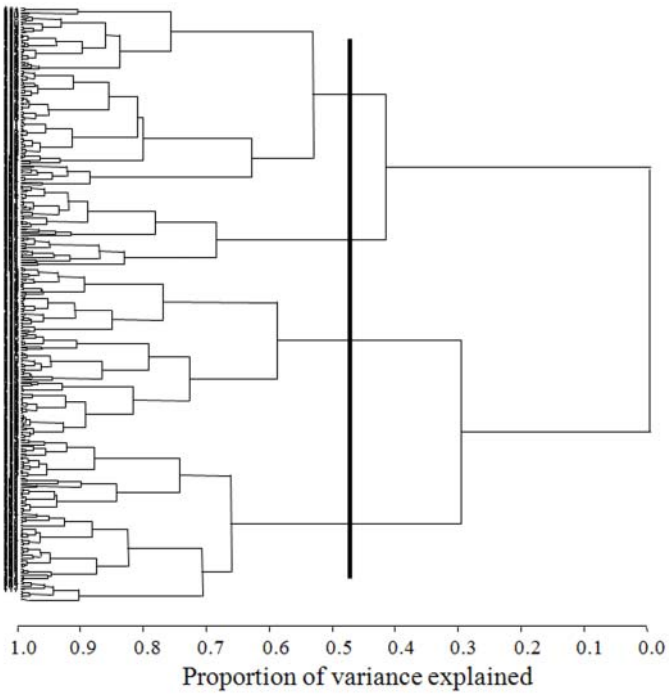
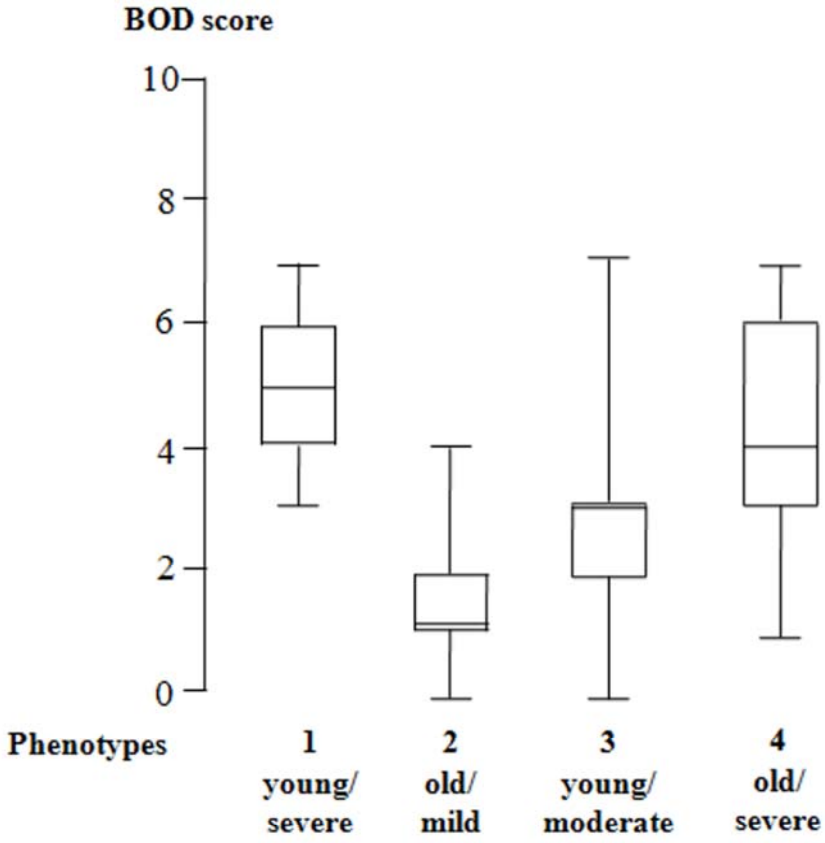


Figure 3: BOD scores in 322 COPD subjects grouped by phenotypes.

COPD subjects (n=322) were grouped in 4 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, 25th, 50th, 75th percentiles and maximum of the variable.


BOD score
10
8
6
4
2
0
Phenotypes
1
young/
severe
2
old/
mild
3
young/
moderate
4
old/
severe