Article 25fa pilot End User Agreement

This publication is distributed under the terms of Article 25fa of the Dutch Copyright Act (Auteurswet) with explicit consent by the author. Dutch law entitles the maker of a short scientific work funded either wholly or partially by Dutch public funds to make that work publicly available for no consideration following a reasonable period of time after the work was first published, provided that clear reference is made to the source of the first publication of the work.

This publication is distributed under The Association of Universities in the Netherlands (VSNU) ‘Article 25fa implementation’ pilot project. In this pilot research outputs of researchers employed by Dutch Universities that comply with the legal requirements of Article 25fa of the Dutch Copyright Act are distributed online and free of cost or other barriers in institutional repositories. Research outputs are distributed six months after their first online publication in the original published version and with proper attribution to the source of the original publication.

You are permitted to download and use the publication for personal purposes. Please note that you are not allowed to share this article on other platforms, but can link to it. All rights remain with the author(s) and/or copyrights owner(s) of this work. Any use of the publication or parts of it other than authorised under this licence or copyright law is prohibited. Neither Radboud University nor the authors of this publication are liable for any damage resulting from your (re)use of this publication.

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please contact the Library through email: copyright@ubn.ru.nl, or send a letter to:

University Library
Radboud University
Copyright Information Point
PO Box 9100
6500 HA Nijmegen

You will be contacted as soon as possible.
Subphenotypes of Mild-to-Moderate COPD by Factor and Cluster Analysis of Pulmonary Function, CT Imaging and Breathomics in a Population-Based Survey

Niki Fens, Annelot G.J. van Rossum, Pieter Zanen, Bram van Ginneken, Rob J. van Klaveren, Aeilko H. Zwinderman & Peter J. Sterk

To cite this article: Niki Fens, Annelot G.J. van Rossum, Pieter Zanen, Bram van Ginneken, Rob J. van Klaveren, Aeilko H. Zwinderman & Peter J. Sterk (2013) Subphenotypes of Mild-to-Moderate COPD by Factor and Cluster Analysis of Pulmonary Function, CT Imaging and Breathomics in a Population-Based Survey, COPD: Journal of Chronic Obstructive Pulmonary Disease, 10:3, 277-285, DOI: 10.3109/15412555.2012.744388

To link to this article: https://doi.org/10.3109/15412555.2012.744388

Published online: 28 Mar 2013.

Submit your article to this journal

Article views: 566

View related articles

Citing articles: 5 View citing articles
**Original Research**

Subphenotypes of Mild-to-Moderate COPD by Factor and Cluster Analysis of Pulmonary Function, CT Imaging and Breathomics in a Population-Based Survey

Niki Fens,¹ Annelot G.J. van Rossum,¹ Pieter Zanen,² Bram van Ginneken,³ Rob J. van Klaveren,⁴ Aeilko H. Zwinderman,⁵ and Peter J. Sterk¹

**Abstract**

**Introduction:** Classification of COPD is currently based on the presence and severity of airways obstruction. However, this may not fully reflect the phenotypic heterogeneity of COPD in the (ex-) smoking community. We hypothesized that factor analysis followed by cluster analysis of functional, clinical, radiological and exhaled breath metabolomic features identifies subphenotypes of COPD in a community-based population of heavy (ex-) smokers. **Methods:** Adults between 50–75 years with a smoking history of at least 15 pack-years derived from a random population-based survey as part of the NELSON study underwent detailed assessment of pulmonary function, chest CT scanning, questionnaires and exhaled breath molecular profiling using an electronic nose. Factor and cluster analyses were performed on the subgroup of subjects fulfilling the GOLD criteria for COPD (post-BD FEV₁/FVC < 0.70). **Results:** Three hundred subjects were recruited, of which 157 fulfilled the criteria for COPD and were included in the factor and cluster analysis. Four clusters were identified: cluster 1 (n = 35; 22%): mild COPD, limited symptoms and good quality of life. Cluster 2 (n = 48; 31%): low lung function, combined emphysema and chronic bronchitis and a distinct breath molecular profile. Cluster 3 (n = 60; 38%): emphysema predominant COPD with preserved lung function. Cluster 4 (n = 14; 9%): highly symptomatic COPD with mildly impaired lung function. In a leave-one-out validation analysis an accuracy of 97.4% was reached. **Conclusions:** This unbiased taxonomy for mild to moderate COPD reinforces clusters found in previous studies and thereby allows better phenotyping of COPD in the general (ex-) smoking population.

**Introduction**

Chronic obstructive pulmonary disease (COPD) is a major cause of mortality and disability, and is defined by the Global Initiative for Chronic Obstructive Lung Disease (GOLD) as the presence of not fully reversible airflow obstruction, confirmed by postbronchodilator spirometry (ratio of forced expiratory volume in 1 second to forced vital capacity: FEV₁/FVC < 0.70) (1). However, it is increasingly recognized that COPD is a complex and heterogeneous disease that cannot be described adequately by the severity of airflow limitation alone (2). Patients vary widely in clinical presentation, response to therapy, decline in lung function, CT imaging and exacerbation frequency. By identification and prospective evaluation of subphenotypes of COPD, the response to therapies and prognosis may be better predicted. Besides this, differential
pathophysiological mechanisms may thus be identified thereby allowing better targeted COPD studies (2).

It has been proposed that a new taxonomy for defining COPD is required to recognize such subphenotypes (3–7). Using unbiased statistical methods such as unsupervised cluster analysis, individuals are grouped based on their similarities and differences in multi-scale data (5,8). This is a way to describe patterns of the disease based on clinical, functional, and pathogenetic features.

Several studies have adopted this approach in order to identify meaningful subphenotypes of obstructive lung diseases (3,6,7,9-13). However, validation and generalisability of these clusters has been limited by the lack of postbronchodilator spirometry and limited availability of CT scans to assess the presence of emphysema. A comprehensive phenotypic characterization of patients with COPD should be based on patient-related outcomes such as health-related quality of life as well as functional, clinical, radiological and inflammatory features.

In a recent study we showed that metabolomic profiles of volatile organic compounds (VOCs) in exhaled air reflect different patterns of inflammatory metabolism and oxidative stress, being associated with eosinophilic and neutrophilic airway inflammation in subjects with mild and moderate COPD (14). Combining exhaled air metabolomics with functional, clinical and radiological measurements may therefore identify relevant subphenotypes of COPD.

The identification and clinical usefulness of clusters is highly dependent on the population under investigation (8). In the community or primary care centers, many individuals with COPD are underdiagnosed (15) and severe cases are scarce, in contrast to the homogeneous patient populations in referral centers. A population-based survey may overcome the selection of a certain subset of patients and may well reflect the primary care population of patients with COPD (8).

In this study we hypothesized that unsupervised cluster analysis in a community-based population of patients with COPD reveals subphenotypes of COPD using lung function parameters, CT scanning, symptoms and other clinical parameters, combined with exhaled breath metabolomics. To that end a survey on a population of heavy (ex-) smokers was undertaken. Cluster analysis was performed on parameters from those subjects in this survey fulfilling the current diagnostic criteria for COPD (1). The validity of the clusters was then tested in a cross-validation analysis.

Methods

Subjects

Subjects in this study were a random selection at a single location (Utrecht, The Netherlands) of the NELSON project. The NELSON project is a population-based Dutch-Belgian multicenter lung cancer screening trial of heavy smokers and ex-smokers (16). Participants of the NELSON study were recruited by sending a questionnaire about smoking history and health-related items to all citizens between the ages of 50 and 75 years living in the area of the participating centers. Of the respondents, subjects meeting the inclusion criterion of a smoking history of at least 15 pack-years were invited for participation in the study (Figure 1). COPD was defined according to the GOLD criteria (postbronchodilator FEV1/FVC ratio < 0.70) (1). Chronic bronchitis was defined as cough and sputum production on most days for at least 3 months per year in at least 2 consecutive years (17).

The trial was approved by the Dutch Ministry of Health and by the local investigational review boards of participating centers. All subjects gave their written informed consent. This COPD substudy of the NELSON study was registered at the Netherlands Trial Register, under NTR 1285.

Design

The study had a cross-sectional design and participants paid a single visit to the hospital. Chest CT scanning, exhaled breath sampling, pre- and postbronchodilator spirometry and diffusion capacity were performed and questionnaires were taken.

Measurements

Questionnaires

Subjects completed a validated questionnaire of the European Community Respiratory Health Survey (ECRHS) concerning sociodemographic data, respiratory symptoms, co-morbidity, treatments and previous diagnoses, use of medications and smoking history (18). Subjects also completed the clinical COPD questionnaire (CCQ) which measures the disease-related health status (19).

Lung function testing

Spirometry was performed before and 10 minutes after inhalation of 400 μg of salbutamol via a spacer according to the latest ERS recommendations using daily calibrated equipment (20). Diffusion capacity for carbon monoxide corrected for alveolar volume (DL,CO/VA) was measured according to the recommendations using the single breath method and an inhalation mixture of 0.3% CO and 10% He with air (21).

Chest CT scanning

CT scanning was performed using a multidetector-row CT scanner (Mx8000 IDT or Brilliance 16P, Philips Medical Systems, Cleveland, OH, USA). Scans were performed in end-inspiration without contrast. Scanning time was within 12 s, in spiral mode with 16 × 0.75 mm collimation, 1.0 mm reconstruction thickness, 0.7 mm increment, voltage of 120 kVp for subjects ≤ 80 kg or 140 kVp for subjects > 80 kg (30 mAs). Scans were analyzed using ImageXplorer (iX) (Image Sciences Institute, Utrecht, NL) (22). The extent of emphysema was estimated by quantifying the percentage of the total
lung voxels below 950 and 910 Hounsfield units (HU) (22). In addition, the 15th percentile point (perc15) was measured (23). The perc15 is defined as the threshold value in Hounsfield units (HU) for which 15% of all lung voxels has a lower value.

**Exhaled breath collection**
Collection of exhaled breath was done as previously described (14,24,25). In short, patients breathed normally for 5 minutes with the nose clipped through an inspiratory VOC-filter (A2, North Safety, NL), connected to a three-way non re-breathing valve and exhaled through a silica reservoir. Then, patients took a single deep maximal inspiration and exhaled one vital capacity volume into a 10 L Tedlar bag (SKC Inc., Eighty Four, PA, USA) connected to the expiratory port and the silica reservoir. In parallel, a Tedlar bag was filled with VOC-filtered room air for comparison.

**Breath analysis by electronic nose**
Within 5 minutes after breath was collected the Cyranose 320 electronic nose (Smiths Detection, Pasadena, Ca, USA) was connected to the Tedlar bag, followed by 1 minute (100 ml) sampling of the exhaled air. This was done in parallel with sampling a Tedlar bag filled with VOC-filtered room air for comparison. Raw eNose data consists of changes in electrical resistance of each of the 32 polymer sensors (26,27) and was used for further analysis with offline pattern-recognition software.

**Statistical analysis**
SPSS (version 18.0) and R were used for data analysis. The total number of 100 variables was reduced to 32 variables representing measures of airflow limitation, response to bronchodilation, gas transfer, metabolomics breath profile, health-related quality of life, sputum production, smoking habits and radiological features of emphysema. Redundant variables covering the same domain were excluded from analysis (e.g., postbronchodilator FEV1 %pred was chosen for the lung function domain). eNose raw data (change in resistance of sensors) were restructured by principal component analysis from the original 32 sensors to 4 principal components (PC) capturing 95.2% of the variance within the data set.

Missing values made up a small proportion (3.9%) and were considered to be at random (28,29). Multiple imputation by the method of Rubin (30) was used for the generation of missing data replacements based on values drawn from the distribution posited by the prediction model, taking into account the relationships between the incomplete variables and all other variables (28,29).

**Factor analysis**
Data was further reduced by factor analysis with orthogonal varimax rotation (31). Based on the loading pattern 12 factors were identified with an eigenvalue >1 representing the different domains of COPD.

One variable representative for each factor was selected, based on the highest loading: pbFEV1, reversibility in FEV1, the presence of chronic bronchitis,
diabetes, coronary vascular disease, BMI, dyspnea at rest, pack-years, use of long-acting bronchodilators, 2 electronic nose principal components and chest CT-scan emphysema score (perc15).

Cluster analysis
Cluster analysis is an unbiased way to assign subjects to groups with similar features without a priori assumptions for classification. Clusters are constructed in such a way that individuals within in a cluster are highly associated and weakly to individuals in other clusters (8).

Subjects were selected for inclusion in the cluster analysis in case they met the spirometric criteria for COPD: post-bronchodilator FEV1/FVC ratio < 0.70 (1).

Cluster analysis was performed following a multi-step approach. Hierarchical cluster analysis by Ward’s method resulted in a dendrogram for estimation of the number of clusters (Figure 2). The number of clusters was determined by the stage associated with the largest distance coefficient delta in the agglomeration schedule by Ward’s method and by the Dunn index (cluster.stats, library fpc in R) and gap statistic (clusGAP in R) (32). The Dunn index identifies clusters that are compact and well separated. The number of clusters where the Dunn index was maximal was taken as the optimal number of clusters (33).

The cluster quality was then checked by two-step cluster analysis, which was ‘good’ for 4 clusters. Finally, a non-hierarchical k-means cluster analysis was performed with the number of clusters set to 4 (34). The internal validity of this model was tested by repeating the k-means cluster analysis 157 times in a leave-one-out validation model (every subject was left out once during cluster formation and used for validation, while the remaining patients served as the training set). The accuracy number was obtained by assessing the percentage of correctly classified cases in each of the 157 runs as the number was obtained by assessing the percent-accuracy number was obtained by assessing the per-

Figure 2. Dendrogram. Hierarchical cluster analysis by Ward’s method resulted in a dendrogram identifying 4 clusters.

Results
Baseline characteristics
Three hundred subjects completed the tests, of which 157 (52%) fulfilled the criteria for COPD and entered the factor and cluster analysis. For the flow chart of the NELSON study in The Netherlands and the COPD sub-study, see Figure 1. The characteristics of the subjects that participated in the cluster analysis are shown in Table 1.

Cluster analysis
Variables that were included in the factor analysis and their values per cluster are listed in Tables 1 to 3. The cluster analysis identified 4 distinct clusters based on the 12 variables included in the model after factor analysis (Figure 2). Clusters showed contrasts with respect to airflow limitation, exhaled molecular profile, health-related quality of life, sputum production, dyspnea, smoking history, co-morbidity, radiologic lung density and gender. Other demographic variables, response to bronchodilation and current smoking status showed no differences between clusters. One out of 4 electronic nose principal components (PC4) showed differences between clusters, whereas the other 3 PC did not (Table 2). The gap statistic was 0.63 (SE 0.02) and the Dunn index 0.074 for 4 clusters (range 0.049–0.074 for 2 to 6 clusters).

Cluster 1 grouped 35 subjects (22% of total) with relatively mild disease. Subjects exhibited mild airflow obstruction, good quality of life and limited symptoms of cough or chronic sputum production. Subjects reported symptoms of dyspnea mainly during exercise and not in rest. Electronic nose PC4 showed low values representing a characteristic molecular profile. Patients showed a relatively high lung density with a mean perc15 of -918 (SD 19) HU indicating little emphysema. This cluster contained the highest percentage females (31%).

Cluster 2 grouped 48 subjects (31% of total) with characteristics of both chronic bronchitis and emphysema. Airflow limitation was moderately severe as this cluster contained all patients with GOLD stage 3 disease that were present in the study. This group showed a distinct breath profile with high PC4 values. Diffusion capacity was impaired and chest CT showed emphysema.

Cluster 3 was the largest cluster, consisting of 60 sub-
jects (38% of total). GOLD stage 1 subjects made up 95% of all subjects in this cluster and as such, mean pBFEV1 was best in this cluster as compared to other clusters (mean 106 %pred, SD 10). Patients had little symptoms, but chest CT did show a lower lung density (mean -950
HU, SD 15) indicating the presence of (asymptomatic) emphysema with preserved lung function. Hypercholesterolaemia was common in this cluster, although cardiovascular co-morbidity was not increased. eNose PC4 was in the same range as in cluster 1, suggesting a similar exhaled molecular profile in these clusters.

Cluster 4 grouped 14 subjects (9% of total) that were highly symptomatic with respect to dyspnea, both at

Table 1. Demographics and clinical characteristics of the total cohort and COPD clusters identified by factor and cluster analysis

<table>
<thead>
<tr>
<th></th>
<th>Total COPD cohort</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
<th>p-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of subjects</td>
<td>157</td>
<td>35</td>
<td>48</td>
<td>60</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Sex (% female)</td>
<td>24 (5.7)</td>
<td>31 (5.7)</td>
<td>25 (5.7)</td>
<td>23 (5.7)</td>
<td>0 (4.2)</td>
<td>0.046*</td>
</tr>
<tr>
<td>Age (yr)</td>
<td>63.2 (5.7)</td>
<td>63.1 (5.7)</td>
<td>63.7 (6.2)</td>
<td>63.3 (5.7)</td>
<td>62.3 (4.2)</td>
<td>0.877</td>
</tr>
<tr>
<td>BMI (kg/m2)</td>
<td>26.8 (3.9)</td>
<td>26.5 (3.8)</td>
<td>26.7 (4.5)</td>
<td>26.6 (3.4)</td>
<td>27.9 (3.5)</td>
<td>0.705</td>
</tr>
<tr>
<td>Post-BD FEV1/FVC</td>
<td>0.61 (0.08)</td>
<td>0.62 (0.07)</td>
<td>0.56 (0.09)</td>
<td>0.64 (0.05)</td>
<td>0.60 (0.06)</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>Pre-BD FEV1 (%pred)</td>
<td>88 (18)</td>
<td>85 (13)</td>
<td>75 (15)</td>
<td>102 (13)</td>
<td>83 (13)</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>Post-BD FEV1 (%pred)</td>
<td>92 (17)</td>
<td>90 (13)</td>
<td>77 (14)</td>
<td>106 (10)</td>
<td>86 (12)</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>GOLD stage 1/2/3 (%)</td>
<td>70/27/3</td>
<td>63/37/0</td>
<td>40/44/10</td>
<td>95/5/0</td>
<td>64/36/0</td>
<td></td>
</tr>
<tr>
<td>Reversibility in FEV1 (%)</td>
<td>3.4 (6.4)</td>
<td>4.4 (4.5)</td>
<td>2.0 (4.5)</td>
<td>4.3 (7.2)</td>
<td>2.6 (5.6)</td>
<td>0.219</td>
</tr>
<tr>
<td>KCO (%)</td>
<td>79 (26)</td>
<td>80 (21)</td>
<td>75 (19)</td>
<td>76 (19)</td>
<td>75 (14)</td>
<td>0.642</td>
</tr>
<tr>
<td>Pack-years</td>
<td>45 (22)</td>
<td>38 (11)</td>
<td>45 (12)</td>
<td>36 (12)</td>
<td>96 (27)</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>Current smokers, %</td>
<td>55 (66)</td>
<td>66 (50)</td>
<td>50 (53)</td>
<td>50 (50)</td>
<td></td>
<td>0.156</td>
</tr>
</tbody>
</table>

*p-value from analysis of variance or Chi-square test between the clusters.
Numeric data is expressed as mean (SD). Post-BD: post-bronchodilator values after 400 µg salbutamol. Variables listed in table 3 were used for cluster formation.

Table 2. Description of the total cohort and COPD clusters identified by factor and cluster analysis

<table>
<thead>
<tr>
<th></th>
<th>Total COPD cohort</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
<th>p-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of subjects</td>
<td>157</td>
<td>35</td>
<td>48</td>
<td>60</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Sputum production %</td>
<td>28</td>
<td>29</td>
<td>38</td>
<td>23</td>
<td>14</td>
<td>0.111</td>
</tr>
<tr>
<td>Chronic bronchitis %</td>
<td>22</td>
<td>14</td>
<td>31</td>
<td>20</td>
<td>14</td>
<td>0.062</td>
</tr>
<tr>
<td>Dyspnea at exercise %</td>
<td>53</td>
<td>54</td>
<td>58</td>
<td>43</td>
<td>71</td>
<td>0.060</td>
</tr>
<tr>
<td>Dyspnea at rest %</td>
<td>5</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>21</td>
<td>0.003*</td>
</tr>
<tr>
<td>Chronic unproductive cough %</td>
<td>11</td>
<td>14</td>
<td>13</td>
<td>7</td>
<td>14</td>
<td>0.301</td>
</tr>
<tr>
<td>2nd hand smoke %</td>
<td>24</td>
<td>20</td>
<td>29</td>
<td>17</td>
<td>14</td>
<td>0.632</td>
</tr>
<tr>
<td>Cancer, any form, %</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>7</td>
<td>0.258</td>
</tr>
<tr>
<td>Inhaled corticosteroid use %</td>
<td>7</td>
<td>6</td>
<td>10</td>
<td>7</td>
<td>0</td>
<td>0.324</td>
</tr>
<tr>
<td>Long-acting bronchodilator use %</td>
<td>9</td>
<td>14</td>
<td>13</td>
<td>5</td>
<td>0</td>
<td>0.163</td>
</tr>
<tr>
<td>Hypercholesterolaemia %</td>
<td>32</td>
<td>31</td>
<td>23</td>
<td>42</td>
<td>29</td>
<td>0.041*</td>
</tr>
<tr>
<td>Asthma %</td>
<td>10</td>
<td>14</td>
<td>8</td>
<td>5</td>
<td>14</td>
<td>0.216</td>
</tr>
<tr>
<td>Diabetes %</td>
<td>13</td>
<td>3</td>
<td>13</td>
<td>15</td>
<td>21</td>
<td>0.121</td>
</tr>
<tr>
<td>Cardiovascular disease %</td>
<td>20</td>
<td>17</td>
<td>21</td>
<td>22</td>
<td>21</td>
<td>0.676</td>
</tr>
<tr>
<td>Hypertension %</td>
<td>24</td>
<td>17</td>
<td>33</td>
<td>22</td>
<td>14</td>
<td>0.101</td>
</tr>
<tr>
<td>CCQ, median (IQR)</td>
<td>6 (4–11)</td>
<td>6 (3–10)</td>
<td>8 (5–12)</td>
<td>6 (3–11)</td>
<td>12 (5–16)</td>
<td>0.025*</td>
</tr>
<tr>
<td>Perc15 (HU), mean (SD)</td>
<td>−943 (22)</td>
<td>−918 (19)</td>
<td>−953 (13)</td>
<td>−950 (15)</td>
<td>−934 (18)</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>VI-HU910 (%), mean (SD)</td>
<td>42 (15)</td>
<td>23 (8)</td>
<td>49 (8)</td>
<td>46 (11)</td>
<td>35 (13)</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>VI-HU950 (%), mean (SD)</td>
<td>14 (9)</td>
<td>5 (2)</td>
<td>18 (7)</td>
<td>16 (7)</td>
<td>10 (6)</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>eNose PC1, mean (SD)</td>
<td>0.00 (1.00)</td>
<td>−0.05 (0.98)</td>
<td>−0.05 (0.98)</td>
<td>0.06 (1.07)</td>
<td>0.03 (0.90)</td>
<td>0.931</td>
</tr>
<tr>
<td>eNose PC2, mean (SD)</td>
<td>0.00 (1.00)</td>
<td>−0.07 (0.95)</td>
<td>−0.04 (1.09)</td>
<td>0.07 (0.99)</td>
<td>−0.01 (0.92)</td>
<td>0.903</td>
</tr>
<tr>
<td>eNose PC3, mean (SD)</td>
<td>0.00 (1.00)</td>
<td>−0.18 (0.83)</td>
<td>0.18 (1.08)</td>
<td>−0.05 (1.02)</td>
<td>0.08 (1.03)</td>
<td>0.410</td>
</tr>
<tr>
<td>eNose PC4, mean (SD)</td>
<td>0.00 (1.00)</td>
<td>−0.14 (0.96)</td>
<td>0.33 (0.93)</td>
<td>−0.18 (1.09)</td>
<td>−0.04 (0.73)</td>
<td>0.047*</td>
</tr>
</tbody>
</table>

CCQ: clinical COPD questionnaire. Perc15: 15th percentile point. VI-HU910: Voxel index at a threshold of -910 Hounsfield Units. VI-HU950: Voxel index at a threshold of -950 Hounsfield Units. PCx: principal component x of exhaled breath electronic nose. Chronic bronchitis is defined as the presence of cough and sputum production for at least 3 months in each of 2 consecutive years. Variables listed in table 3 were used for cluster formation.
exercise and at rest. Patients had a reduced quality of life, but lung function was relatively preserved and inhalation medication use was absent. This cluster showed the highest amount of pack-years (mean 96, SD 27). Radiologic emphysema score and breath profile (electronic nose PC4) were in the intermediate range (Tables 1-2).

**Internal validation of clusters**

In order to validate the clusters that were identified, the k-means cluster analysis in a leave-one-out validation model was performed 157 times, resulting in an accuracy of 97.4%. The average silhouette width for all 4 clusters was 0.22 (cluster 1 0.24; cluster 2 0.27; cluster 3 0.25; cluster 4 0.17).

**Discussion**

The present cluster analysis derived from a community-based population of heavy (ex-) smokers identified 4 distinct clusters of patients with COPD. It appeared that symptoms, spirometry, CT lung density and exhaled molecular profiling all contributed significantly ($p < 0.05$) to distinguish these COPD subphenotypes. Cluster 1 consisted mainly of mild COPD. Patients in cluster 2 showed the largest impairment in lung function and signs of emphysema and many subjects reported symptoms of dyspnea and chronic bronchitis. Cluster 3 was characterized by emphysema with preserved lung function. Cluster 4 consisted of highly symptomatic subjects with preserved lung function. This study validates and extends phenotypes of COPD of other studies, and could therefore provide a new taxonomy for COPD (5).

Multidimensional assessment of COPD is increasingly recognized as an effective approach to discover clinically relevant phenotypes that may differ in natural course of disease and response to therapies (4,35). A previous factor analysis provided good evidence of heterogeneity of COPD when using clinical, physiological and cellular biomarkers (36). Cluster analysis comprises a next step aimed to identify COPD phenotypes by classification of subjects into homogeneous groups based on a heterogeneous set of variables (8). This method is considered to be hypothesis-generating rather than potentially biased by a priori assumptions.

Previously, several groups adopted similar approaches to identify subphenotypes in patients with COPD (6,7,9-13). Subphenotypes that were identified were severe COPD (9-11,13), mild COPD (10,11,13), chronic bronchitis (9-12), emphysematous COPD (12,13) and a phenotype with features of both chronic bronchitis and emphysema (9). These studies were mainly focused on clinical COPD subphenotypes, and often did not include biomarkers or CT parameters.

In contrast, subjects in the present study were randomly recruited from a community-based survey among heavy (ex-) smokers. This makes our results better generalisable in the primary care population. In addition to the regular diagnostic workup for COPD, we also included spiral CT scanning of the chest in order to assess the presence and extent of emphysema, and exhaled breath metabolomics using an electronic nose. The latter is a non-invasive integrative analysis that appears to be associated with the inflammatory profile in both asthma and COPD (14,37).

The clusters show distinct breath molecular profiles as assessed by electronic nose. Cluster 1 has high PC4 values, indicating a distinct profile as compared to the other clusters. This cluster also shows chronic bronchitis and emphysema. Although sputum measurements were not performed in this study, it is likely that the high PC4 value in this cluster reflects the presence of airway inflammation associated with chronic bronchitis.

The clusters identified in this study confirm and extend the COPD phenotypes found by others using cluster analysis, including the mild COPD group (cluster 1) (9,10), the combined chronic bronchitis and emphysema group (cluster 2) (9-11) and the asymptomatic emphysematous group (cluster 3) (9,11,13). This not only serves as an promising result warranting further external validation (38), but also indicates that clinical subphenotypes of COPD can also be found in a general population of heavy smokers. Most importantly, this study confirms the COPD-clusters found in the community-based study by Weatherall et al. (9), thereby serving as validation. Interestingly, the smallest cluster in this study, the highly symptomatic COPD group with no use of inhalation medication with preserved lung function (cluster 4) was not reported previously. This group might represent patients that are not optimally treated (inhaled medication use is 0%) despite their relatively normal lung function.

The newest GOLD guidelines recommend the inclusion of exacerbation risk (high risk defined as ≥2 exacerbations/year) as an extra domain in characterizing COPD (1). In this population-based study, only 5 exacerbations were reported in the past year. This was not unexpected.
as most patients (97%) were classified as GOLD A or B based on their spirometry results. However, 3 out of 5 exacerbations were reported in Cluster 2, the combined emphysema-chronic bronchitis cluster.

We made large efforts to ensure the validity of our findings by careful methodological choices. First, COPD was established based on the GOLD criteria (1), being the current gold-standard. This makes the results easily comparable to other studies. Second, the recruitment of subjects from the general population, rather than selected clinical groups of patients, ensured that the data can be generalized to COPD in the community, predominantly featuring mild to moderate airways obstruction. Third, the variables in the cluster analysis were chosen to reflect different aspects of COPD, including functional, clinical, inflammatory, radiological and patient-centered outcomes. Fourth, most studies using cluster analysis excluded subjects with missing data. This markedly decreases the available sample size, as it often occurs that a variable is missing due to logistic reasons or technical issues. Rather than excluding subjects, we used multiple imputation to generate replacements of the small proportion of randomly missing data (28-30). This procedure has been validated and is considered a reliable method for unbiased analysis of missing data when taking into account all observed data (29).

The choice for a community-derived sample of subjects may also be considered as a limitation in this study. The range of FEV1 in the study population is narrow with a mean FEV1 of 88%pred, and only 3 patients are GOLD stage III. We recognize that the different clusters that were found are directly influenced by the population in which the study was carried out. However, the fact that all COPD clusters in this study are consistent with those in other studies as well strengthens the validity of our findings. Notably, these phenotypes appear to be already apparent at mild or early stages of the disease (4).

A second potential limitation is that the use of disease-modifying treatments such as inhaled corticosteroids and the presence of co-morbidities may have affected the results. Co-morbidities were assessed by questionnaires and by medication usage rather than by systematic diagnostic check-up. Therefore, we cannot completely exclude missing a proportion of yet undiagnosed co-morbidities. On the other hand, medication use and co-morbidity were not used as selection criteria, thereby reflecting a real life community-derived COPD population and daily practice in primary care.

Each of the clusters showed a distinct profile: a combination of features covering the different disease domains of COPD. This cannot be translated directly into (partly) distinct, driving pathogenetic mechanisms. However, one may speculate that inflammatory subtype and -activity plays an important role. This study is likely to have implications for future COPD studies. First, a better understanding of the COPD phenotypes that are associated with differential treatment responses or prognosis may facilitate the development of more specific diagnostic procedures and targeted therapies. COPD subphenotyping will benefit the design of randomized controlled trials, especially in mild or early stages of the disease in the primary care population (5,15).

It has been estimated that only 1 in 20 patients identified as having COPD in a community survey would meet the current criteria for inclusion in an RCT for COPD (39). Representative samples of the most prevalent phenotypes of COPD will strengthen the implications of clinical trials. Second, exhaled metabolomic profiling contributed to the subphenotyping of COPD. This indicates not only that inflammatory status (14) and thereby the potential response to anti-inflammatory and other therapies differs between clusters, but also that such molecular subphenotyping can be obtained by adding a rapid, non-invasive breath test to the diagnostic work-up of COPD. Finally, in order to examine the stability and clinical course of the COPD subphenotypes, longitudinal follow-up studies and treatment response studies should be carried out (4).

Conclusions

In a community-derived population of patients with predominantly mild to moderate COPD, 4 distinct subphenotypes could be identified using factor and cluster analysis of clinical, functional, CT lung density and metabolomics data: mild COPD, combined emphysema and chronic bronchitis, asymptomatic emphysema and symptomatic COPD with preserved lung function. Using such new taxonomy could result in better phenotyping of COPD and thereby to potentially better management in daily practice and more focussed clinical trials.

Acknowledgments

The authors would like to thank Marc P.C. van der Schee and Eva Roos for their help with exhaled breath measurements. The study was funded by an unrestricted grant from the Netherlands Asthma Foundation, grant no 3.2.06.17. The Netherlands Asthma Foundation did not have a role in the design, collection, analysis and interpretation of data, writing of the manuscript and in the decision to submit the manuscript for publication.

NF conceived of the study, carried out the breath measurements, performed the statistical analysis and wrote the manuscript. AvR carried out the breath measurements, constructed the database and performed the statistical analysis. PZ facilitated the lung function measurements and participated in the design and coordination of the study. BvG carried out the CT-scans and participated in the design and coordination of the study. RvK participated in the design and coordination of the study. AZ performed the statistical analysis and helped to draft the manuscript. PS conceived of the study, participated in the design of the study and helped to draft the manuscript. All authors read and approved the final manuscript.
Declaration of Interest Statement

The authors declare that there are no financial, consulting, and personal interests that could influence the authors’ work. The study was funded by an unrestricted grant from the Netherlands Asthma Foundation, grant no 3.2.06.17.

References