Explorative data analysis techniques and unsupervised clustering methods to support clinical assessment of Chronic Obstructive Pulmonary Disease (COPD) phenotypes

Matteo Paoletti a,*, Gianna Camiciotto b, Eleonora Meoni b, Francesca Bigazzi b, Lucia Cestelli b, Massimo Pistolesi b, Carlo Marchesi a

aDepartment of Systems and Computer Science, University of Florence, Florence, Italy
bDepartment of Internal Medicine, Section of Respiratory Medicine, University of Florence, Florence, Italy

A R T I C L E   I N F O
Article history:
Received 5 September 2008
Available online 6 June 2009

Keywords:
Explorative biomedical data analysis
Homogeneity analysis
Biomedical data clustering
Biomedical clustering
Multiple Correspondence Analysis
Principal Component Analysis
K-Harmonic Means
COPD
Chronic Obstructive Pulmonary Disease
Biomedical data mining

A B S T R A C T
Chronic Obstructive Pulmonary Disease (COPD) is the fourth leading cause of death worldwide and represents one of the major causes of chronic morbidity. Cigarette smoking is the most important risk factor for COPD. In these patients, the airflow limitation is caused by a mixture of small airways disease and parenchyma destruction, the relative contribution of which varies from person to person. The twofold nature of the pathology has been studied in the past and according to some authors each patient should be classified as presenting a predominantly bronchial or emphysematous phenotype. In this study we applied various explorative analysis techniques (PCA, MCA, MDS) and recent unsupervised clustering methods (KHM) to study a large dataset, acquired from 415 COPD patients, to assess the presence of hidden structures in data corresponding to the different COPD phenotypes observed in clinical practice. In order to validate our methods, we compared the results obtained from a training set of 415 patients with lung density data acquired in a test set of 93 patients who underwent HRCT (High Resolution Computed Tomography).

1. Introduction

Chronic Obstructive Pulmonary Disease (COPD) is the fourth leading cause of death worldwide and represents one of the major causes of chronic morbidity. Cigarette smoking is the most important risk factor for COPD [1]. In the past this disease has been relatively neglected and unfortunately there are no current therapies that reduce the progression of the disease. Due to the enormous burden of the disease and rising healthcare cost, there is now renewed interest in the underlying cellular and molecular mechanisms and a search for new therapies, resulting in a general re-evaluation of the disease. However, despite its highly relevant health impact, there has been relatively little research into COPD, and at present, it is the most under-funded disease in relation to its global burden [2].

The Global Initiative on Obstructive Lung Disease (GOLD) has recently adopted a new definition of COPD: “a disease state character-
Since we had relatively little prior knowledge about the hidden nature of those data and since a general reference statistical model was missing, we proceeded with data driven analysis to understand the structure of this multidimensional dataset and to assess the existence of particular subgroups of patients related to the two different phenotypes that may be observed in clinical practice. This approach may also be useful in studying other complex chronic pathologies, in which different variables and clinical parameters may influence the outcome of the disease.

We avoided hypotheses about raw data (unlabeled samples) and assumptions of particular parametric forms that in this case may lead to poor or meaningless results.

We were mainly interested in extracting knowledge from our dataset, not in imposing a structure on it.

The methods we used for the numerical and graphical explorative analysis are mostly derived from pattern recognition, cluster analysis, and classic multivariate statistics theories.

After numerical and visual inspection of the data, we applied a recently introduced unsupervised clustering algorithm, called KHM (K-Harmonic Means), to study the homogeneity of the dataset at hand and to investigate the presence of hidden structures in data. Finally, in order to test the correspondence between the clusters observed in our analysis and the groups of patients characterized by different COPD phenotypes, we compared the calculated memberships with standardized lung density measures performed on a test set composed of 93 patients that underwent HRCT (High Resolution Computed Tomography).

2. Materials

Four hundred and fifteen consecutive COPD patients with mild to severe airflow limitation underwent clinical tests and standardized questionnaires to derive clinical, functional and radiological characteristics that could represent the broad spectrum of COPD manifestations. Study enrollment was based on stability of clinical conditions, availability of chest radiography within one week of the functional evaluation, and a patient’s willingness to participate. For each patient we considered 19 categorical variables and 15 numerical-continuous variables as shown in Table 1.

Categorical variables were derived from the clinical history, physical examination and chest radiography, while the numerical continuous parameters were collected using a mass-flow sensor and a blood gas analyzer according to standard methodologies.

Table 1

The dataset is composed of 19 categorical variables (parameters) and 15 numerical-continuous variables. Categorical variables were derived from clinical history, physical examination and chest radiography, while the numerical continuous parameters were collected using a mass-flow sensor and a blood gas analyzer according to standard methodologies.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Values</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Categorical variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chest radiography</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) Increased vascular markings</td>
<td>0/1 (Absent/present)</td>
<td>IVM</td>
</tr>
<tr>
<td>(2) Reduced vascular markings</td>
<td>0/1 (Absent/present)</td>
<td>RVM</td>
</tr>
<tr>
<td>(3) Interstitial changes</td>
<td>0/1 (Absent/present)</td>
<td>IC</td>
</tr>
<tr>
<td>(4) Reduced lung density</td>
<td>0/1 (Absent/present)</td>
<td>RLD</td>
</tr>
<tr>
<td>(5) Increased lung volume</td>
<td>0/1 (Absent/present)</td>
<td>ILV</td>
</tr>
<tr>
<td>(6) Bronchial wall thickening</td>
<td>0/1 (Absent/present)</td>
<td>BWT</td>
</tr>
<tr>
<td>Clinical history</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(7) Cough</td>
<td>0/1/2 (Absent/occasional/chronic)</td>
<td>Cough</td>
</tr>
<tr>
<td>(8) Sputum</td>
<td>0/1/2 (Occasional/chronic/purulent)</td>
<td>Sputum</td>
</tr>
<tr>
<td>(9) MRC dyspnoea scale</td>
<td>0/1/2/3/4 (None/slight/moderate/severe/very severe)</td>
<td>MRC</td>
</tr>
<tr>
<td>Physical examination</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(10) Cyanosis</td>
<td>0/1 (Absent/present)</td>
<td>CYA</td>
</tr>
<tr>
<td>(11) Pursued lips breathing</td>
<td>0/1 (Absent/present)</td>
<td>PLB</td>
</tr>
<tr>
<td>(12) Neck veins dilatation</td>
<td>0/1 (Absent/present)</td>
<td>NVD</td>
</tr>
<tr>
<td>(13) Hypertrophy of accessory muscles of respiration</td>
<td>0/1 (Absent/present)</td>
<td>HAM</td>
</tr>
<tr>
<td>(14) Barrel chest</td>
<td>0/1 (Absent/present)</td>
<td>BC</td>
</tr>
<tr>
<td>(15) Hoover’s sign</td>
<td>0/1 (Absent/present)</td>
<td>HS</td>
</tr>
<tr>
<td>(16) Peripheral edema</td>
<td>0/1 (Absent/present)</td>
<td>PE</td>
</tr>
<tr>
<td>(17) Chest hyper ressonance</td>
<td>0/1 (Absent/present)</td>
<td>CH</td>
</tr>
<tr>
<td>(18) Reduced breath sounds</td>
<td>0/1 (Absent/present)</td>
<td>RBS</td>
</tr>
<tr>
<td>(19) Adventitious breath sounds</td>
<td>0/1 (Absent/present)</td>
<td>ABS</td>
</tr>
<tr>
<td><strong>Numeric/continuous variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respiratory function measurements (% of predicted)</td>
<td>Continuous/numeric</td>
<td>FVC (%)</td>
</tr>
<tr>
<td>(1) Forced vital capacity</td>
<td>Continuous/numeric</td>
<td>FEV₁ (%)</td>
</tr>
<tr>
<td>(2) Forced expiratory volume (in one second)</td>
<td>Continuous/numeric</td>
<td>VC (%)</td>
</tr>
<tr>
<td>(3) Slow vital capacity</td>
<td>Continuous/numeric</td>
<td>IC (%)</td>
</tr>
<tr>
<td>(4) Inspiratory Capacity</td>
<td>Continuous/numeric</td>
<td>FEV₁/VC</td>
</tr>
<tr>
<td>(5) Tiffeneau index</td>
<td>Continuous/numeric</td>
<td>FRC (%)</td>
</tr>
<tr>
<td>(6) Functional residual capacity</td>
<td>Continuous/numeric</td>
<td>RV (%)</td>
</tr>
<tr>
<td>(7) Residual volume</td>
<td>Continuous/numeric</td>
<td>TLC (%)</td>
</tr>
<tr>
<td>(8) Total lung capacity</td>
<td>Continuous/numeric</td>
<td>DLCO (%)</td>
</tr>
<tr>
<td>(9) CO diffusing capacity</td>
<td>Continuous/numeric</td>
<td>RV/TLC</td>
</tr>
<tr>
<td>(10) Motlhey index</td>
<td>Continuous/numeric</td>
<td>pH</td>
</tr>
<tr>
<td>Blood gas measurements</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) pH</td>
<td>Continuous/numeric</td>
<td>pH</td>
</tr>
<tr>
<td>(2) O₂ partial arterial tension (mmHg)</td>
<td>Continuous/numeric</td>
<td>PaO₂</td>
</tr>
<tr>
<td>(3) CO₂ partial arterial tension (mmHg)</td>
<td>Continuous/numeric</td>
<td>PaCO₂</td>
</tr>
<tr>
<td>HRCT measurements</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) Pixel Index (percentage of lung area with attenuation &lt; −950 Hounsfield Units (HU))</td>
<td>Continuous/numeric</td>
<td>PI</td>
</tr>
<tr>
<td>(2) Mean Lung Attenuation (HU)</td>
<td>Continuous/numeric</td>
<td>MLA</td>
</tr>
</tbody>
</table>
flow sensor and a blood gas analyzer according to standard methodology [6,7]. Ninety three of the 415 patients also underwent spirometrically-gated High Resolution Computerized Tomography (HRCT) and served as the validation set for this study. HRCT was obtained at predefined inspiratory lung volume to avoid the influence on density measurements of different lung inflation levels during scanning [8]. Patients breathed through a spirometer connected to the scanner and performed reproducible vital capacity (VC) maneuvers. Subsequently, the airflow through the spirometer was interrupted at 90% of VC by a shutter triggering the scanner to acquire images at three levels (carina, 5 cm above, 5 cm below). Frequency histograms of lung attenuation values of each section were averaged [9] to derive X-ray mean attenuation (MLA variable in Table 1), as overall assessment of lung density, and percentage area with X-ray attenuation values below –950 (PI variable in Table 1) Hounsfield Units (HU), as index of emphysema extent [10].

3. Methods

When approaching a medical data analysis problem, we may already have some hypotheses (derived by scientific or suggested by clinical practice) that we would like to test. In this case, we would use hypothesis-testing procedures to confirm our a priori hypotheses regarding the relationships between the variables in our dataset [11]. However, we do not always have a priori knowledge of the expected relationships among the variables in the dataset or we do not want contradictory opinions or clinical debates, often regarding application fields in which a reference mathematical model is missing, to bias the analysis process.

To delve into our large unknown data set we used exploratory data analysis (EDA) instead of classic statistical methods. The interrelationships among the various attributes were evaluated with a “data driven” approach. Such an approach could be also easily exported to other clinical contexts and the described methodologies could be generalized across other biomedical domains and applications.

We started analyzing the dataset at hand to identify interesting subsets of the observations and to assess the presence of separated clusters in our sample without a priori hypotheses. We used the following methods to explore the dataset:

- **PCA** (Principal Component Analysis) to study numerical variables (measured values of functional parameters) [12].
- **MCA** (Multiple Correspondence Analysis) to analyze categorical variables (clinical and radiographic parameters) [13].

In particular these methods have been used to allow clinical specialists to visualize and understand the dataset at hand before applying clustering algorithms.

Although data were collected from patients affected by the same disease, the exploratory analysis revealed the presence of partially overlapping subgroups or clusters in the underlying data structure, which suggests that further investigation was necessary to find possible correspondence with clinical phenotypes observed by physicians.

We proceeded, in subsequent steps, to process the whole data matrix using MDS [14] (Multi Dimensional Scaling or Principal Coordinates Analysis) to reduce the space dimensions and applying an unsupervised clustering [15] method called KHM [16], to characterize the observed subgroups. This recently introduced algorithm is essentially insensitive to the initialization of the class centers and it shows rapid convergence.

Finally we compared the cluster memberships of the 93 patients of the validation set with the MLA (Mean Lung Attenuation) and PI (Pixel Index) variables acquired using HRCT to verify possible correspondences between the obtained results and clinical evidences. MLA and PI are derived from high-resolution images obtained using computerized tomography and represent a reliable measure of pulmonary density, an important index of emphysema extent [9,10]. Table 2 presents a summary of the methods applied during this study. In the next sections we provide some details about the techniques we used during this research. The “data quality” theme is not considered in this paper, in facts during the data acquisition phase of the study the incoming measurements and variables were controlled step by step by the clinical specialists (outliers, missing data) using a data entry system utility.

### 3.1. Principal Component Analysis

Since our dataset is composed of a large number of observed variables, we simplify the explorative analysis process by considering a smaller number of linear combinations of the original variables. In Principal Component Analysis (PCA), we seek to represent the d-dimensional data in a lower-dimensional space [17]. This reduces the degrees of freedom, reduce the space and time complexities and allow a visual exploration of the whole dataset at hand. The goal of this well-known method is to concentrate the information about the differences between samples into a small number of dimensions. In particular a set of n-dimensional vector samples $\mathbf{x} = \{x_1, x_2, x_3, \ldots, x_m\}$ should be transformed into another set $\mathbf{y} = \{y_1, y_2, \ldots, y_m\}$ of the same dimensionality, but $y_i$ have the property that most of their information (in terms of variance) content is stored in the first few dimensions. This will allow us

---

**Table 2**

<table>
<thead>
<tr>
<th>Step #</th>
<th>Variables set</th>
<th>Data type</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Visual exploratory analysis</td>
<td>Respiratory function measurements + blood gas parameters, Chest radiography, Clinical history, Physical examination</td>
<td>Numeric</td>
<td>PCA</td>
</tr>
<tr>
<td>2. Parameters-space dimension reduction</td>
<td>Entire dataset at hand</td>
<td>Numeric and categorical</td>
<td>MDS</td>
</tr>
<tr>
<td>3. Clustering</td>
<td>Reduced dataset</td>
<td>Numeric</td>
<td>KHM</td>
</tr>
<tr>
<td>4. Validation</td>
<td>HRCT test set + clustering results</td>
<td>Numeric</td>
<td>Box and whisker plots, percentiles analysis</td>
</tr>
</tbody>
</table>
to reduce the data set into a smaller number of dimensions with minimal loss of information [18].

After we have decided how many components to consider, we can visualize the data through a special graph called biplot [19], in which information on both samples and original variables are displayed allowing an easy visual interpretation of relationships between different patients.

3.2. Multiple Correspondence Analysis

In the previous section we described how to use PCA to project a multivariate data set onto the plane in which the data has maximum dispersion. This allows us to examine the data visually, while sacrificing the minimum amount of information. While PCA represents an effective method to study continuous numerical variables, it is not suitable for the analysis of categorical variables. Correspondence Analysis (CA) is an exploratory data analytic technique designed to analyze two-way and multi-way tables containing some measure of correspondence between the rows and columns. These methods were originally developed primarily in France by Jean-Paul Benzécri in the early 1960’s and 1970’s, but have only recently gained popularity in English-speaking countries [20].

An important feature of CA is the multivariate treatment of the data through simultaneous consideration of multiple categorical variables. This method allows to graphically display row (patients, cases) and column (variables, parameters) points in biplots, which help to detect structural relationships among the variable categories and the various observations. We can consider CA a special case of PCA of the rows and columns of a dataset, especially applicable to a cross-tabulation. However CA and PCA are used under different circumstances. PCA is used for dataset consisting of continuous measurements, whereas CA is applied to contingency tables (for example cross-tabulations). The main goal is to transform a table of numerical information into a graphical display, in which each row and each column is depicted as a point [21]. The usual procedure for analyzing a cross-tabulation is to determine the probability of global association between rows and columns. The significance of association is tested by the Chi-square test, but this test provides no information as to which are the significant individual associations between row–column pairs of the data matrix. CA shows however that the variables are related, not just that a relationship exists. Multiple Correspondence Analysis (MCA) may be considered an extension of simple correspondence analysis to more than two variables. As in PCA, the results of MCA are presented on graphs that represent the configurations of points in projection planes, defined by the first axes considered two or three at a time [22]. With this kind of graphics we can summarize the row and column coordinates in a single plot. However, it is important to remember that in such plots, one can only interpret the distances between row points, and the distances between column points, but not the distances between row points and column points. The joint display of row and column points shows the relation between a point from one set and all points of another set, not between individual points from each set.

3.3. Multi Dimensional Scaling

Many difficulties may arise during biomedical multivariate data analysis. Most of them derive from the inability to plot and visualize the structure of multidimensional data. Multi Dimensional Scaling (MDS) method, also known as Principal Coordinates Analysis, seeks to represent the original data points in a lower dimensional Euclidean space while preserving, as far as possible, the dissimilarities between data points [22]. This method is used to represent dissimilarities through well defined distances in a geometric space (where a metric is defined). MDS is often used to project the original data points into a 2-D or 3-D reduced space allowing a visual inspection of the whole dataset at hand. MDS operates on a dissimilarity matrix calculated from the original dataset and is also useful when the dataset is composed of continuous measurements, categorical variables and binary variables (heterogeneous datasets). The computation of the dissimilarity between two objects depends on the type of the original variables. In particular, since in our dataset each patient is represented through several parameters of different types, we calculate the dissimilarities \( d(i,j) \) between the various patients as follows:

\[
d(i,j) = \frac{\sum_{t=1}^{p} \phi_t^{(i)} d_t^{(j)}}{\sum_{t=1}^{p} \phi_t^{(i)}}
\]

The term \( d_t^{(j)} \) represents the contribution of the variable \( t \) and depends on its specific type:

- If \( t \) is a variable derived by continuous measurements we have

\[
d_t^{(j)} = \frac{|x_j - x_t|}{\text{max}_x x_j - \text{min}_x x_t}
\]

- If \( t \) is a categorical or binary variable we have \( d_t^{(j)} = 0 \) if \( x_j = x_t \) and \( d_t^{(j)} = 1 \) otherwise.

- If \( t \) is a discrete ordinal variable (has \( M_t \) possible values or scores which are ordered) first we replace the \( x_j \) by their rank \( r_j \in \{1, \ldots, M_t\} \), then we calculate the transformed variables \( u_t = \frac{r_j - 1}{M_t - 1} \). We proceeded as in the first step (Eq. (2)) to calculate dissimilarities using \( u_t \) instead of the original values.

In Eq. (1), the term \( \phi_t^{(i)} = 0 \) if \( x_j \) or \( x_t \) are missing, or if \( x_j = x_t = 0 \) and \( t \) is an asymmetric binary variable. Otherwise, \( \phi_t^{(i)} = 1 \).

Let us consider \( n \) samples \( x_1, \ldots, x_n \) and their lower-dimensional projections \( z_1, \ldots, z_n \). The algorithm tries to find a configuration of image points \( z_j \) for which the \( n(n-1)/2 \) Euclidean distances \( h_{ij} \) are as close as possible to the corresponding original dissimilarities \( d_{ij} \). An optimal configuration \( z_1, \ldots, z_n \) is defined as one that minimizes the following criterion function:

\[
J = \sum_{i=1}^{n} (h_{ij} - d_{ij})^2
\]

Since \( J \) considers only distances between points, the function values are invariant to rigid-body motions of the configurations. An optimal data points configuration in the reduced space can be calculated by a simple gradient-descent procedure, starting with an initial configuration and changing the \( z_i \) in the direction of greatest decrease in \( J \). The gradient of \( h_{ij} \) with respect to \( z_j \) is a unit vector in the \( z_i - z_j \) direction, in fact we have \( h_{ij} = \|z_i - z_j\| \), thus the gradient to be considered is:

\[
\nabla_{z_j} J = \frac{2}{\sum_{i} h_{ij}^2} \sum_{k \neq j} (h_{ij} - d_{ij}) \frac{z_k - z_j}{h_{ij}}
\]

Fig. 5 shows the 2-D optimal configuration obtained with this method considering all the original variables in our COPD dataset.

3.4. Cluster analysis

Cluster analysis is a set of methodologies allowing automatic grouping of samples, or cases, into classes of similar objects. Unlike classification, clustering is unsupervised and there is no target variables [24]. With clustering methods the goal is not to classify, estimate, or predict the value of a target variable but to segment the entire data set into relatively homogeneous subgroups or clusters, where the similarity of the records within the cluster is maximized, and the similarity to records outside the cluster is min-
imized [25]. In other words, those methods seek to construct groups of records such that the between-cluster variation is large compared to the within-cluster variation. Examples of clustering tasks in business and research include: target marketing, segmentation of financial behavior into benign and suspicious categories, dimension-reduction tools, gene expression clustering. Clustering has been often used also in some areas of medicine. Our aim was to identify whether there are different subtypes of disease lumped together under a single diagnosis.

In this study we used a recently introduced partitioning method called KHM (K-Harmonic Means) [16] instead of the well known K-Means algorithm. In facts a major problem with K-Means is that it is sensitive to the selection of the initial partition and may converge to a local minimum of the objective function if the initial centers are not properly chosen [26]. The KHM algorithm, recently introduced by Zhang et al. [27] solves this problem by replacing the minimum distance from a data point to the centers, used in K-Means, by the harmonic averages of the distances from the data point to all centers. The harmonic average of \( N \) numbers is defined as:

\[
H(a_1, ..., a_N) = \frac{N}{\sum_{k=1}^{N} \frac{1}{a_k}}
\]

This quantity is small also if at least one of the numbers is small.

The association of data points with the centers is distributed and the transition becomes continuous during convergence of KHM. According to the authors, the algorithm has a “built-in” dynamic weighting function, which boosts the data that are not close to any center by giving them a higher weight in the next iteration. The recursive expression used to calculate center coordinates at each iteration is:

\[
c_k = \frac{\sum_{i=1}^{N} \frac{1}{d_{i,k}} \cdot x_i}{\sum_{i=1}^{N} \frac{1}{d_{i,k}}}
\]

where \( c_k \) with \( k = 0, ..., K \) are the center positions, \( x_i \) are the coordinates of the \( i \)th data point, the terms \( d_{i,k} = ||x_i - c_k|| \) represent the distances of data points from centers. The KHM algorithm starts with a set of initial positions of the centers, the terms \( d_{i,k} \) are calculated and then the new positions of the centers are obtained using (6).

The results are that KHM performs consistently better than KM and EM [28,29]. The superiority of the K-Harmonic Means algorithm to identify clustering of high quality in experimental dataset has been demonstrated by comparison of the three clustering methods [30]. For these reasons we preferred KHM to classic KM/EM methods in our applications.

4. Results and discussion

Although data were collected from patients suffering with the same disease (COPD), the exploratory analysis revealed the presence of partially overlapping subgroups or clusters in the underlying data structure. We started data exploration calculating the Principal Components relatively to the continuous numerical variables of the dataset. In order to perform a better discriminating visual exploratory analysis, we introduce, in this paper, a particular graph type (that we named “DB-Plot”, Density Bi-Plot) derived from the standard Principal Component Analysis biplots [31]. Instead of overlapping the original variables vectors and a scatter plot of the transformed data we used a density plot as background. This simple detail allowed us to enhance sub-clusters and groups differences appreciable using the standard biplots. Fig. 1a and b show DB-Plots relative to the KL transform (first three components planes) performed considering the functional respiratory variables dataset at hand (see Table 1 for details and variable labels).

Two high density partially overlapping structures are well distinguishable in the dataset. Observing the directions of the original variable vectors it is possible to derive a subset of parameters that better discriminate individuals belonging to two subgroups. In particular in Fig. 1a it is possible to observe a parameter set composed of FRC, TLC, RV, RV/TLC in contrast to DLCO, FEV1, FEV1/VC along
the two-cluster axis direction. The PaO₂, pH, PaCO₂ parameters seems not to be significative in this projection because of their low modules. IC, FVC, VC variables directions are orthogonal to the two-cluster axis, and do not give useful information for discriminating the two groups. Fig. 1b shows the projection obtained considering the Principal Components I and III. This projection includes a major contribution of the three variables PaO₂, pH, PaCO₂ along the Y axis (high modules along the III principal component). Those variables directions seem to be orthogonal to the twofold structure axis, demonstrating low discriminating power. In summary the visual exploratory analysis of the functional respiratory variables, shows the presence of one patient’s subgroup characterized by higher values for DLCO, FEV1, FEV1/VC and simultaneous lower values for RV/TLC, FRC, TLC, RV. Another useful conclusion arising from this analysis is the redundancy of the acquired dataset. As we can see from the figures, the following variable sets seem to be linearly related or acting like indicators of the same physical phenomena: RV-FRC and IC-FVC-VC. In facts, calculating the correlations for the variables above, we obtained the following values: 0.91 for RV-FRC, 0.89 for FVC-VC, 0.72 for FVC-IC and 0.71 for VC-IC. These considerations may be also useful to plan data acquisition campaign or to design questionnaires.

We obtained interesting results and useful information about our COPD patients also applying Multiple Correspondence Analysis to the categorical data described in Table 1. In order to study the specific contribution of the available variables groups, we split the analysis process in three parts related to: anamnestic variables, clinical variables, radiographic variables. Figs. 2–4 show the MCA results through various output data plots.

Due to the high number of variables we have plotted, for each parameters set, a couple of graphs, showing the projection of the original variables in the factorial space and the density plot of the individuals’ projections. We considered only the most interesting and informative projections, avoiding planes with no interesting structures.

Fig. 2b shows the distribution of the original variables relative to chest radiography in the I–II factorial space after applying MCA technique. The first panel shows a density projection of the individuals in the factorial space (I–II factors plane). The plot shows a clear twofold structure for the radiological parameters in the dataset. The discriminating direction is along the F1 axis, which is mainly supported (in module) by the variables RVM, BWT, RLD, IVM, ILV.

Fig. 3. The second panel shows the original anamnestic variables projection in the plane defined by the first two MCA factors; the first panel shows a zoomed vision of the points projections (individuals’ density) in the same plane.
PC analysis, we can gain useful information about the radiographic parameters and their contribution to the individuals space partition. To interpret these plots we can consider the quadrants (I–IV) and the directions of the imaginary line joining the points to the origin (consider each point as a vector as in PCA examples). For example, the graph plotted in Fig. 2b suggests that patients with reduced lung density (RLD = 1), absence of bronchial wall thickening (BWT = 0), and increased lung volume (ILV = 1) and of increased vascular markings (IVM = 0). This is the typical chest radiographic pattern of patients with predominant emphysema.

The Fig. 2b is informative about the variables distribution while the plot in the upper panel shows a density projection of the individuals in the factorial space (I–II factors plane). Similar considerations can be derived for the association of other variables displayed in the MCA plots.

The plot shows a clear twofold structure for these radiological parameters. In particular the discriminating direction seems to be along the F1 axis, which is mainly supported (in module) by the variables RVM, BWT, RLD, IVM, ILV. It seems that two typical configurations are present in this space, one characterized by an increased pulmonary volume, reduced lung density, no bronchial thickening, the other characterized by an opposite parameters combination. Only a few points are projected in the central area near the origin, indicating that only a low number of patients show a profile different from the two configurations described above. To better display the twofold nature of the point density along the F1 factor, in the upper panel we provide a density projection that enhances the bimodal distribution of the subjects.

Similar projections, obtained applying MCA to variables derived from anamnesis questionnaire and from clinical evaluation, are shown in Figs. 3 and 4, respectively.

The graphs displayed in Fig. 3b show the projection of the original variables in the plane defined by the first two factors while the upper panel shows a zoomed vision of the points projections (individuals’ density) in the same plane. It can be observed that individuals with Cough = 0 (absent cough) are often characterized by Sputum = 0 (occasional sputum). Similarly the various cough properties (0 = absent, 1 = occasional, 2 = chronic) correspond to the sputum characteristics (0 = occasional, 1 = chronic, 2 = purulent) while the MRC scale seems not to follow this path. In fact, MRC = 0 is in the same direction of MRC = 4. As far as patients density distribution, the upper panel shows a fragmented situation where several subgroups and clusters are visible, indicating that the sample is characterized by the presence of subgroups of patients with same parameters combination. Furthermore, overlapping the two panels, it is possible to identify a high density group characterized by Sputum = 2 and Cough = 2 (chronic cough and purulent sputum).

Fig. 4. MCA results obtained from data acquired during COPD patients physical examination. In this projection it is possible to distinguish distinct data clouds (second panel) along the F2 factorial axis. Two main different patients group are visible in this space while a minor points cloud is positioned between the main groups. The first panel displays the original variables distribution in this reduced space.
Finally, we can observe in Fig. 4 the MCA results obtained analyzing data acquired during patients physical examination.

The plane defined by the first two factors shows distinct data clouds (Fig. 4b) along the F2 factorial axis. As shown in the figures, few points are located between the two visible main groups. The individuals projected in the lowest part of the plot seems to be characterized by RBS = 1, CH = 1, ABS = 0, while patients belonging to the other group (cluster with positive F2 values) show higher probability to be characterized by RBS = 0, CH = 0, ABS = 1. Other variables-individuals correspondence can be assessed with the same logical criteria and considerations.

In summary, we can assume that this dataset is fragmented in different subgroups suggesting that the patient profile for this COPD population does not show a continuous spectrum. Different clusters and groups are clearly visible in several projections and using different parameters combination. Thus, exploratory considerations suggest to further investigate to numerically characterize the various individuals and to verify possible correspondences between the observed clusters and different phenotypes observed in clinical practice.

In the following step we applied MDS (Multi Dimensional Scaling) and KHM clustering algorithm to calculate a space partition capable to summarize the data characteristics observed above in a quantitative manner.

The analysis (Fig. 5) was performed considering at the same time both categorical and continuous/numeric data.

Visual inspection reveals again the presence of two partially overlapping subsets of aggregation. One group has negative values along principal coordinate I and higher dispersion of the data points with respect to a second group with positive values of principal coordinate I. Fig. 5 also shows the results of KHM cluster analysis. The highest value of between-cluster variance ratio was obtained when considering two as the optimum number of clusters within the series of patients data points. The two clusters show only slight overlaps (see Fig. 5 for details).

The following steps of the analysis was aimed to evaluate whether the two clusters represent the bronchial and the emphysematous phenotypes observed in COPD.

To test the correspondence between the subgroups observed in this study and the different clinical COPD phenotypes, we analyzed the 93 individuals (validation set) in whom high-resolution computerized tomography (HRCT) density parameters were measured. In particular we compared the calculated cluster memberships with the following measures: MLA10 (Mean Lung Attenuation for scans at 10% of vital capacity), MLA90 (Mean Lung Attenuation for scans at 90% of vital capacity), PI10 (Pixel Index, % of lung area with attenuation values below −950 HU at 10% of vital capacity), PI90 (Pixel Index, % of lung area with attenuation values below −950 HU at 90% of vital capacity).

Fig. 6 shows the box and whisker plots of the MLA10, MLA90, PI10, PI90 as a function of the cluster memberships derived from the application of MDS and KHM methods. The plots also show the confidence intervals. In this type of graphs, if the confidence intervals on two boxes do not overlap, this indicates a difference in location at a rough 5% significance level.

The two identified subgroups, C0 and C1, seems to refer to two different ranges of MLA10, MLA90, PI10, PI90 values, showing that the two subgroups are significantly different. In particular, patients belonging to C0 are characterized by a significantly lower pulmonary density typical for the emphysematous COPD phenotype. Patients belonging to C1 seem not to be affected by parenchyma destruction and are characterized by a higher lung density. The fact that the two subgroups had a minimal overlap may also suggest that, in our sample, individuals were affected by a predominantly emphysematous COPD phenotype or by a predominantly bronchial COPD phenotype.
dominant emphysematous COPD or by a predominant bronchial COPD.

5. Conclusions

In conclusion, analyzing a large dataset acquired from a population of COPD patients with exploratory and unsupervised clustering methods, it appears that each patient, although being the individual clinical expression of a wide and continuous spectrum of pathologic changes (different lesions may coexist) causing expiratory airflow limitation, could be classified as being affected by predominant airway disease or by predominant parenchymal destructive changes. The evaluation of the cluster membership calculated for each COPD patient may be considered as a useful profile indicator and may well impact on understanding the results of pharmacologic trials, on clinician's approach to patient treatment, and on deeper knowledge of COPD natural history.

Furthermore, this study may be considered as a methodological example showing possible applications of intelligent data analysis and visual exploratory techniques to investigate clinical aspects of chronic pathologies where a mathematical referring model is generally missing. In this field it is advisable to proceed with a data driven approach to avoid possible bias introduced by a priori hypotheses. Such an approach may be useful to define patient profiles related to different pathologic mechanisms of disease as in patients with COPD.

The same methodological approach may be of value to study the temporal evolution of chronic diseases by evaluating data acquired during long term longitudinal monitoring campaigns.

References