Comprehensive Cluster Analysis for COPD Including Systemic and Airway Inflammatory Markers

Halehsadat Nekoee Zahraei[a,b]* Françoise Guissard[b] Virginie Paulus[b] Monique Henket[b] Anne-Françoise Donneau[a] and Renaud Louis[b]

[a]*Biostatistics Unit, Department of Public Health, University of Liège, Belgium;*
[b]*Department of Pneumology, GIGA, University of Liège, Belgium*

*corresponding author
Bat.B23 Biostatistics, Quartier Hopital, Avenue de l'Hopital, 3, 4000 Liege, Belgium
Email H.Nekoee@uliege.be

Multiple Imputation

One of the most common problems in any large observational dataset which researchers have been experienced in clinical data analysis is the missing values and methods for handling it. Before starting to impute missing value is important to investigate the reason for the missing value. In COPD dataset, we can consider that missing value is related to the patients (Not Missing at Random, NMAR). The patients do not have enough ability to do the experiments and lead to missing values, and alternative argument is that the data is missing randomly (Missing at Random, MAR) and missingness related to the latent variable. In the present study, missing values were imputed using multiple imputation technique with exceptional properties, aiming to obtain complete datasets. Multiple imputation process contains three steps. At the first step, imputation step, missing value is imputed m (>1) times and creates m different complete datasets with the same number of patients and variables, also the method preserves type and range of variables. In this study, missing values were imputed by draw from the posterior predictive distribution of Bayesian model and Predictive Mean Matching (PMM) was used as a robust method to model misspecification in imputing values. In this study, we set m equals to 100. In the second step, analysis step, three visualized statistical analyses were applied, FAMD and HCPC, which has many analytical decisions. In this process, each imputed dataset has its distinctive and individual outcomes of the analysis. In the final step, combination step, the overall result is derived from m distinct outcomes by consensus clustering method. Details are given in the following sections. Multiple imputation method and PMM are implemented in the R-package MICE[1].

Factor Analysis Methodology

Another important issue in cluster analysis is number of variables contained in clustering. Since we consider large number of variables, the poor discrimination of distance, correlated variables, and redundant variables are caused to degrade final classification especially when the variables can be combined and considered on lower dimension. Variable reduction is a crucial step for accelerating model building without losing the potential predictive power of the data. After multiple imputation, for reducing the complexity of huge multi-dimensional variables, FAMD was performed to create the new uncorrelated continuous components from the linear combination of the existing variables. The main idea of factor analysis for mixed data is to reduce the number of variables to the smaller number of new components while the new components containing as much information as possible[2]. The procedure for choosing the number of components to be retained is another important question. Several methods have been proposed for determining the number of components that should be kept for further analysis. Kaiser proposed dropping components whose eigenvalues are less than one or keep components with cumulative variance larger than 90[3]. Based on these two criterions, the number of components was directly chosen for each imputed dataset. Therefore, thirty-four new components were considered for the six imputed datasets, ninety-three of imputed datasets have thirty-five components and only one of imputed datasets have thirty-six components for the next step, cluster analysis[4].

Cluster Analysis Methodology

Cluster analysis is a collection of methods for partitioning objects into several groups such objects in one cluster are similar to each other and have high differences into other clusters. The ultimate goal is the members of created clusters are closer than the members of different clusters. Several methods are utilized in the world of clustering analysis. Depend on the nature of research, the sort of data and ease execution of clustering methods, researchers select one of the clustering methods.

In this study, patients are usually collected by huge multi-dimensional variables which in the previous step was reduced to several continuous components. Distance between these components represents the similarity and dissimilarity between two patients. Two patients are close to each other when the distance between two data is small. Among all of the methods in this field, here, we focused on two common approaches of the clustering method, hierarchical clustering, and partitioning method.

In cluster analysis, it is necessary to define the number of clusters. One of the methods to find the number of clusters is the hierarchical method based on agglomerative technique using Ward criterion with squared Euclidean distance[5]. In this method, each patient creates a single cluster and then pairs of close clusters (minimum between variance) iteratively merge to new cluster and this process continues until one cluster is formed. In the hierarchical clustering method, the selection of numbers is determined by dendrogram. When the increase of between variance between *k* and *k-1* cluster is much greater than the one between *k* and *k+1* clusters, then the *k* clusters are chosen for the best number of clusters[6]. However, determination of the number of clusters is still a big problem in cluster analysis. There is no unique and acceptable answer to find the best number of clusters. In this study, we afford to exclude every arbitrary decision from the user. For each imputed dataset, the Nbclust package was applied[6]. This R package provides 30 different methods for distinguish high classification. In this step, regarding to majority vote of those 30 available methods, number of cluster was determined to the best discrimination[6]. So, user isn't able to select and change the number of clusters, like similarity profile analysis and dendrogram plot. Therefore, among 100 imputed datasets, ninety-four imputed datasets proposed three clusters and only six clusters are chosen for four imputed datasets. After determining the number of clusters for each imputed dataset, *k* initial starting points are randomly selected, and observations were assigned to *k* groups based on the closeness of each subject to *k* initial point. Then, the centroid of each cluster is updated and re-assigned group to subjects based on the nearest point to cluster centroid. This procedure was repeated until no improvement was observed. We performed all of these processes on each imputed dataset with different numbers of components and different numbers of clusters. FAMD and cluster analysis were computed and visualized using FactoMineR and factoextra R packages[4].

Consensus Clustering

In the final step of clustering based on Rubin's rule for multiple imputation, m individual clustering solutions are obtained from each imputed dataset. In view of this fact that each imputed dataset involves bias and not considering errors of the imputations, we avoid selecting one particular result as the final clustering result. In literature, the simplest method, high frequency in m imputed datasets, is used to find the final clustering solution and assigning the patient to the cluster[7]. However, the main disadvantage of this method is that all clustering has to be grouped in the same number of clusters. Therefore, in this study, consensus clustering was applied which choosing the best solution of combining ensemble clustering to the final cluster. This method minimizes the sum of squared distance of existing clustering results. Consensus clustering is implemented in the R-package CLUE[8].

Clustering Validation

Cluster analysis is a powerful but unsupervised method to find a structure in the dataset such that patients in one group are closer than other patients from the next groups. The big issue in this unsupervised method is, to evaluate the quality of the clustering framework for classification. There exist three statistical methods that show the validation's clustering, internal, external, and clustering stability validation. In the following Table two indices for internal clustering validation and two indices for clustering stability validation are reported.

Table. Clustering validation

	Indices	**Value**
Internal measures	Silhouette Width	0.6101
	Dunn Index	0.5362
Stability measures	Average Proportion of Non-overlap(APN)	0.0213
	Average Distance between Means (ADM)	0.0081

Silhouette width measures the average degree of confidence in the clustering. The Silhouette width is in the interval [-1,1] and values near 1 present well clustered. The Dunn Index divides the smallest distance between two observations in the different clusters to the largest intra-cluster distance. The Dunn Index is in the interval [0, ∞] and the maximum values are preferable. Average Proportion of Non-overlap (APN) divides the observations in different clusters by clustering in full data. The APN is in the interval [0, 1] and the minimum values are preferred. Average Distance between Means (ADM) calculates the distance of observations in the same clusters and clustering in full data. The ADM is in the interval [0, ∞] and smaller values show well clustering[9].

References

1. Buuren S van. *Flexible Imputation of Missing Data*. CRC Press; 2012.

2. Escofier B, Pagès J. Multiple factor analysis (AFMULT package). *Comput Stat Data Anal*. 1994;18(1):121-140. doi:10.1016/0167-9473(94)90135-X

3. Kaiser HF. The Application of Electronic Computers to Factor Analysis. *Educ Psychol Meas*. Published online April 1, 1960. doi:10.1177/001316446002000116

4. Kassambara A. *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*.

5. Ward JH. Hierarchical Grouping to Optimize an Objective Function. *J Am Stat Assoc*. 1963;58(301):236-244. doi:10.1080/01621459.1963.10500845

6. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *J Stat Softw*. 2014;61(1):1-36. doi:10.18637/jss.v061.i06

7. Basagaña X, Barrera-Gómez J, Benet M, Antó JM, Garcia-Aymerich J. A framework for multiple imputation in cluster analysis. *Am J Epidemiol*. 2013;177(7):718-725. doi:10.1093/aje/kws289

8. Hornik K. A CLUE for CLUster Ensembles. *J Stat Softw*. 2005;14(1):1-25. doi:10.18637/jss.v014.i12

9. Brock G, Pihur V, Datta S, Datta S. **clValid** : An *R* Package for Cluster Validation. *J Stat Softw*. 2008;25(4). doi:10.18637/jss.v025.i04