Identification of Chronic Obstructive Pulmonary Disease Phenotype using PCA and Clustering Methodologies

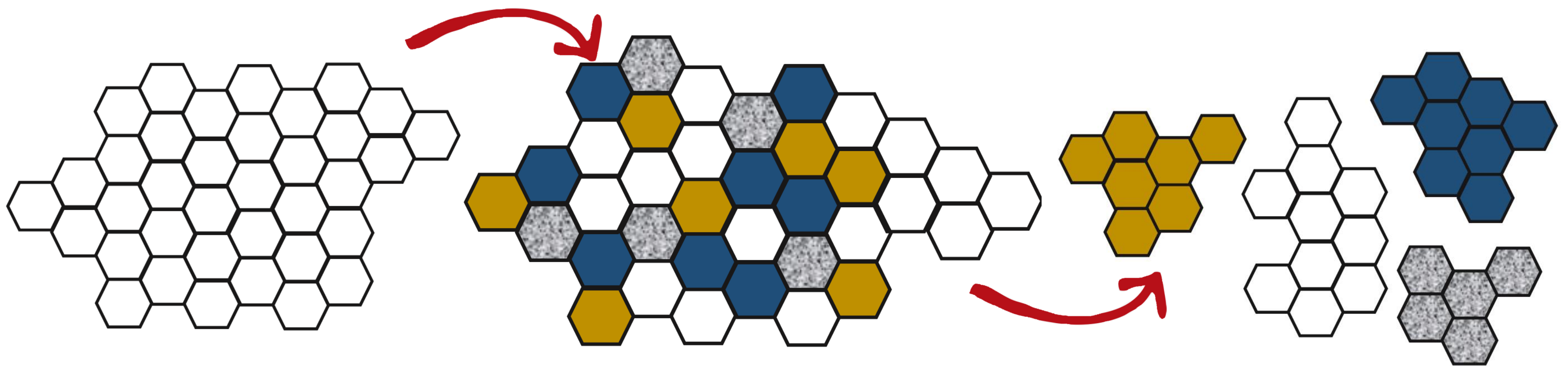
H. Nekoee Zahraei[1,2], V. Paulus[2], M. Henket[2], R. Louis[2], A.F. Donneau[1]

[1] Department of Public Health, Bstat, University of Liège, Belgium

[2] Department of Pneumology, GIGA, University of Liège, Belgium

Introduction

One of the most well-known methods, within data mining framework, for discovering knowledge in multivariate data set is clustering. The goal of clustering methods is to identify groups of patients (i.e. clusters) such that patients in the same group are more similar to each other than to patients in other groups.



Objective

In this study, we aim to identity distinct phenotypes of adults suffering from chronic obstructive pulmonary disease (COPD). Clustering leads to better understanding, management, future risks, prediction and treatment selection optimization of different patient groups.

Cluster Analysis

Hierarchical and partition clustering method around medoids based on square root of Jensen–Shannon distance was applied to the multiple factor analysis. This method is more robust to noise and outliers.

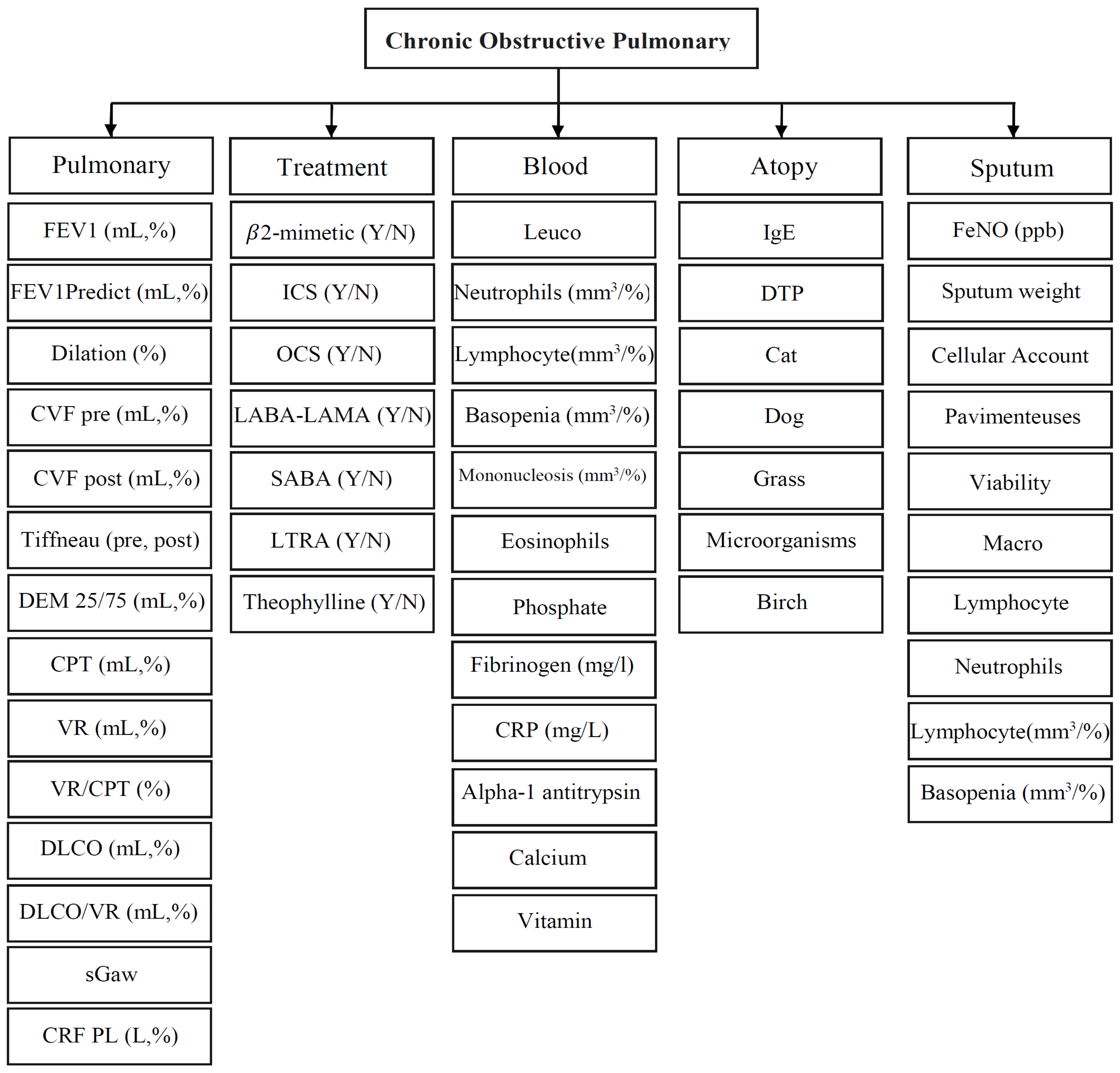
$$D(a,b) = \sqrt{0.5\,KLD\,(a,m) + 0.5\,KLD(b,m)}$$

$$KLD(x,y) = \sum_i x_i \log\left(\frac{x_i}{y_i}\right)$$

The square root of Jensen-Shannon distance is a positive definite measure, symmetric and triangular inequality. We added a small value to avoid zero in the numerator and/or denominator of equation. Finally, the partitioning around medoids (PAM) clustering algorithm was applied to cluster the COPD data.

Variables Reduction

Classification of COPD is complicated by the heterogeneous and multidimensional nature of the disease. In the past, application of classification methods in this context had been developed by limited and selected variables to define phenotypes with comparable results. However, in this study we used multiple factor analysis (MFA), which is an extension of principal component analysis (PCA), for reducing the complexity of high-dimensional data.



Chronic Obstructive Pulmonary

Pulmonary: FEV1 (mL,%); FEV1Predict (mL,%); Dilation (%); CVF pre (mL,%); CVF post (mL,%); Tiffneau (pre, post); DEM 25/75 (mL,%); CPT (mL,%); VR (mL,%); VR/CPT (%); DLCO (mL,%); DLCO/VR (mL,%); sGaw; CRF PL (L,%)

Treatment: β2-mimetic (Y/N); ICS (Y/N); OCS (Y/N); LABA-LAMA (Y/N); SABA (Y/N); LTRA (Y/N); Theophylline (Y/N)

Blood: Leuco; Neutrophils (mm^3/%); Lymphocyte(mm^3/%); Basopenia (mm^3/%); Mononucleosis (mm^3/%); Eosinophils; Phosphate; Fibrinogen (mg/l); CRP (mg/L); Alpha-1 antitrypsin; Calcium; Vitamin

Atopy: IgE; DTP; Cat; Dog; Grass; Microorganisms; Birch

Sputum: FeNO (ppb); Sputum weight; Cellular Account; Pavimenteuses; Viability; Macro; Lymphocyte; Neutrophils; Lymphocyte(mm^3/%); Basopenia (mm^3/%)

Result

The best cluster number was found to be equal to 5. Those clusters were identified as: mild COPD, Eosinophilic COPD, Atopic COPD, COPD + Emphysema and neutrophilic COPD + Emphysema + Systemic inflammation.

Table1. Part of Descriptive of Cluster Analysis (mean (std.error))

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
	mild COPD	Eosinophilic + COPD	Atopic + COPD	Emphysema + COPD	Emphysema Systemic inflammation + neutrophilic + COPD
Variable	n=68	n=18	n=4	n=61	n=23
Demographic					
Age(year)	63.18(9.38)	62.28(9.14)	65(9.20)	66.06(9.14)	64.83(10.18)
Sex (Female)	40% (27)	61%(11)	25%(1)	59%(36)	22%(5)
Cigarette Packs (year)	40.48(29.52)	37.34(18.53)	60.25(46.08)	40.18(28.88)	37.18(18.01)
Smoking Duration (years)	39.30(13.51)	40.94(13.76)	43.5(6.56)	41.67(12.55)	41.54(14.02)
Pulmonary Function					
FEV1Predict (%)	66.97(14.24)	47.39(10.15)	47.5(17.41)	48.69(13.79)	39.69(14.79)
CVF post (%)	89.94(16.06)	72.44(15.83)	65.75(21.93)	78.06(16.56)	65.95(15.44)
DEM 25/75 (%)	32.84(12.95)	26.87(12.33)	18.33(4.04)	21.65(9.69)	16.07(7.49)
VR/CPT (%)	52.83(9.25)	58.79(10.18)	69.67(2.08)	61.80(6.82)	63.36(8.89)
DLCO/VR (%)	74.98(21.49)	76.22(11.28)	92(19.67)	63.27(20.68)	69.43(26.80)
sGaw	0.65(0.33)	0.52(0.26)	0.35(0.08)	0.45(0.20)	0.35(0.19)
CRF PL (%)	145.78(35.29)	165.0(36.14)	150.67(30.24)	167.37(30.47)	165.78(39.35)
Treatment					
β2-mimetic (Y/N)	0.28%(19)	78%(0.43)	0(0)	85%(52)	73%(17)
LABA-LAMA (Y/N)	47%(32)	94%(17)	50%(2)	97%(59)	100%(23)
SABA (Y/N)	18%(12)	56%(10)	50%(2)	59%(36)	39%(9)
Blood					
Leuco	8.24(2.26)	8.59(1.72)	9.67(2.07)	7.21(1.81)	11.60(3.25)
Neutrophils (%)	58.25(7.94)	56.89(8.87)	59.45(9.06)	60.97(8.56)	75.66(11.0)
Lymphocyte(%)	30.33(7.93)	29.56(7.38)	32.72(6.95)	28.54(8.41)	15.83(7.69)
Mononucleosis (%)	8.66(2.30)	8.59(3.22)	5.87(1.80)	8.01(2.01)	6.56(3.11)
Eosinophils	2.26(1.67)	4.42(2.43)	1.95(1.56)	2.05(1.24)	1.66(1.46)
Basopenia (%)	0.49(0.28)	0.58(0.26)	0.27(0.09)	0.45(0.26)	0.29(0.17)
Fibrinogen (mg/l)	3.41(0.73)	3.63(0.66)	3.25(0.43)	3.51(0.79)	4.36(1.14)
CRP (mg/L)	4.57(6.19)	3.58(2.92)	4.57(3.37)	6.29(10.31)	24.99(36.83)
Atopy					
IgE	108.18(213.52)	869.12(1402.68)	417.25(284.6)	328.31(882.54)	234.19(284.57)
DTP	0.08(0.48)	0.23(0.64)	7.07(9.04)	0.58(2.07)	0.02(0.11)
Sputum					
FeNO (ppb)	20.22(18.65)	32.44(24.99)	26(16.75)	16.60(11.44)	22.23(13.10)
Cellular Account	2.75(4.82)	3.31(2.63)	6.59(2.58)	4.16(5.04)	29.94(41.17)
Macro	18.29(14.15)	22.37(13.55)	41.2	12.86(11.39)	5.78(6.39)
Lymphocyte	1.76(2.34)	5.7(16.65)	6.6	1.28(1.40)	0.96(2.08)
Neutrophils	72.14(18.28)	26.64(16.18)	48.8	78.95(13.84)	84.54(16.43)
Eosinophils	2.57(4.24)	26.72(28.72)	2.8	2.9(4.13)	6.33(11.54)
Epithelial	5.08(60.)	18.56(15.14)	0.6	4.01(3.94)	2.38(4.54)

Conclusion

Application of this new cluster approach allowed to identify five groups within stable COPD patients based on huge clinical variables while accounting for noise and outlier. Derived results can be used to predict outcomes of patients with COPD and to aid in development of personalized therapy.

Reference

Rousseeuw, P. J. Silhouettes (1987). a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.

Everitt, B.S., Landau, S. and Leese, M. (2001), Cluster Analysis, Fourth edition, Arnold.

Arumugam, M., Raes, J., et al. (2011). Enterotypes of the human gut microbiome, Nature, 473, 174–180

H.Nekoee@uliege.be

LIÈGE université