

Introduction

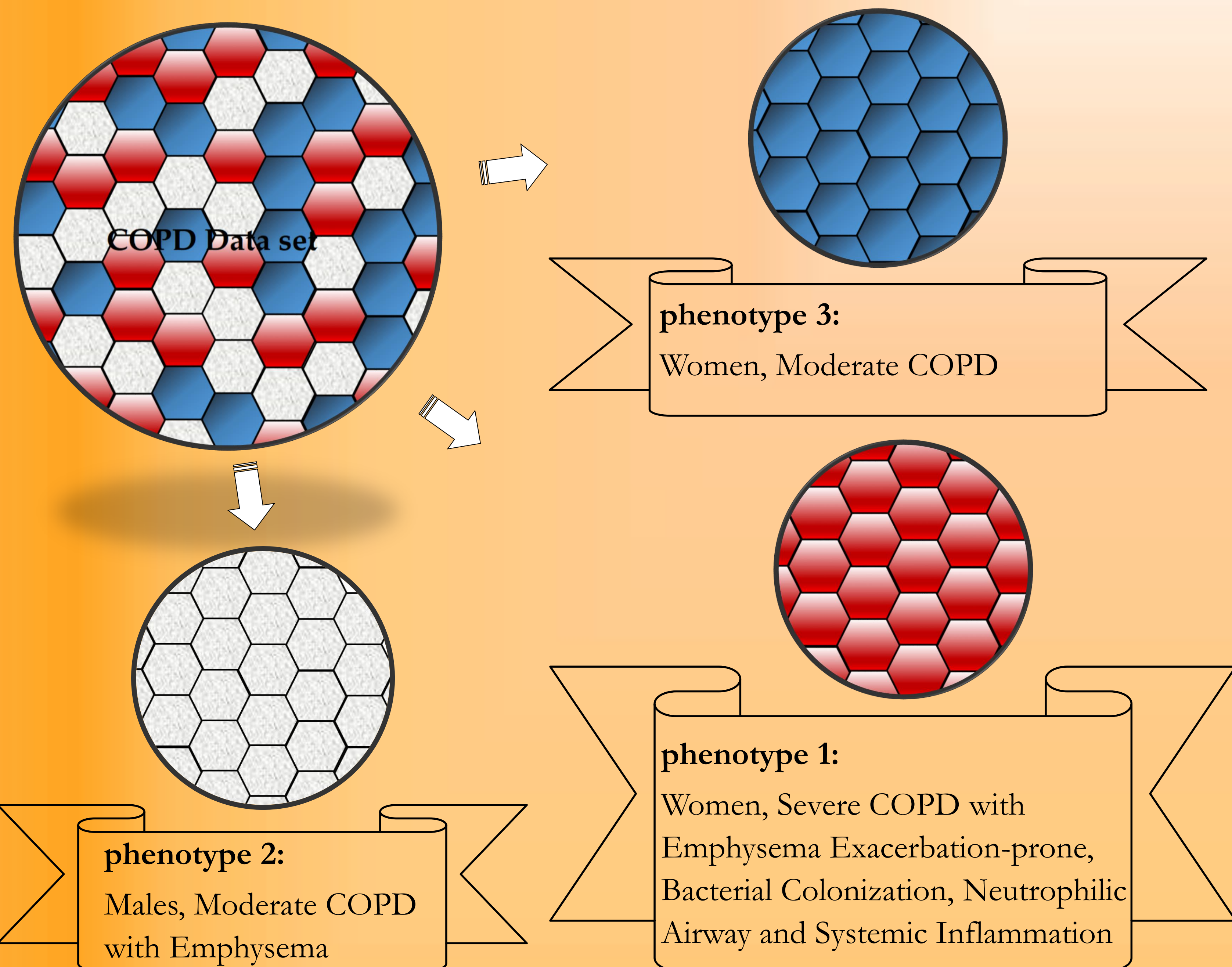
Chronic obstructive pulmonary disease (COPD) is a complex, multidimensional and heterogeneous disease with a large number of subtypes and multifactorial background. It may be caused by different pathophysiologic mechanisms (sometimes referred to as endotypes) but may share similar observed characteristics (phenotypes). These phenotypes divide all patients into several groups with common features which helps patients to receive effective care and achieve better clinical results.

Objective

The main purpose of this study is, through the development of statistical methods, to identify clinical phenotypes among adults suffering from COPD. This problem of grouping objects may be solved by cluster analysis. Clustering was applied to understand, manage and better predict future risks and optimize treatment selection based on the new groupings of patients. Furthermore, in this study, missing data and dimension-reduction, which are present in any large dataset of observational data, were handled.

Cluster Analysis

In this application, 178 patients were described by 85 multiple and huge sets of variables that structured into seven groups. This study was conducted by the Pneumology Department of the University Hospital of Liege. At the first step, the missing values were imputed by multiple factor analysis (MFA). After single imputation, MFA was applied for reducing the complexity of high-dimensional data. After this step, hierarchical clustering was performed using Ward's criterion on the selected principal components. In this step, the optimal number of clusters was selected based on the hierarchical tree, total within-cluster sum of square and silhouette method. In the final step, K-means was performed to improve the initial partition obtained from hierarchical clustering. All statistical analyses were performed using R software.



Conclusion

In the past years, classification methods in COPD have been applied based on ignoring missing values with limited or selected number of variables which have missed more complex phenotypes. In this study, these two issues are solved. Then, with advanced statistical methods, patients are divided into three distinct clusters. These clinically meaningful clusters of patients with common characteristics can be used to predict outcomes of patients with COPD, to aid in the development of personalized therapy.

Statistical Analysis

Three different phenotypes were defined in COPD. All variables were compared between three clusters by using Kruskal-Wallis and Chi-squared tests for quantitative and qualitative variables, respectively.

Table. Characteristics of patients with COPD according to variable groups after imputation, Median(IQR)/Percentage(frequency)

Variable	Cluster 1 (n=39)	Cluster 2 (n=72)	Cluster 3 (n=67)	P - value
	Women, Severe COPD with Emphysema Exacerbation-prone, Bacterial Colonization, Neutrophilic Airway and Systemic Inflammation	Males, Moderate COPD with Emphysema	Women, Moderate COPD	
Demographic				
Age (year)	69 (62 - 75)	63.5 (57 - 71.25)	64 (56 - 71)	0.059
Sex (Female)	79.49% (31)	1.39% (1)	97.01% (65)	<0.0001
Height (cm)	169 (163.5 - 175)	160.55 (155 - 165)	174 (169 - 178)	<0.0001
Weight (kg)	67 (56 - 77)	60 (52.75 - 66.12)	78 (67 - 90)	<0.0001
BMI (kg/m ²)	23.18 (20.49 - 25.08)	23.16 (20.06 - 24.91)	25.88 (22.13 - 30.08)	0.0002
Cigarette Packs (year)	37.7 (20.87 - 50)	38 (21.45 - 46)	37.5 (27.75 - 51)	0.385
Cigarettes (day)	19.14 (10 - 20)	20 (10 - 20)	20 (11.5 - 25)	0.072
Smoking Duration(year)	45 (30 - 50)	44 (36 - 49)	41 (32.5 - 50)	0.432
OCS Course	0 20.51% (8)	61.11% (44)	82.09% (55)	
1 46.15% (18)	26.39% (19)	11.94% (8)		
≥2 33.33% (13)	12.5% (9)	5.97% (4)		
0 20.51% (8)	31.94% (23)	50.75% (34)		
Antibiotic Course	1 58.97% (23)	58.33% (42)	47.79% (28)	0.003
≥2 20.51% (8)	7.42% (7)	7.46% (5)		
0 20.51% (8)	31.94% (23)	50.75% (34)		
Emergency Room Admission for asthma or COPD	1 30.77% (12)	12.5% (9)	7.46% (5)	0.001
≥2 2.56% (1)	0% (0)	0% (0)		
0 66.67% (26)	87.5% (63)	92.54% (62)		
Number of hospitalizations for asthma or COPD	1 30.77% (12)	11.11% (8)	2.98% (2)	<0.0001
≥2 2.56% (1)	1.39% (1)	1.49% (1)		
Pulmonary				
FeNO (ppb)	20 (11.15 - 26)	14 (9.75 - 19.89)	17 (11 - 29.5)	0.051
FEV1 predicted (mL)	1050 (780 - 1315)	1295 (987.5 - 1505)	1840 (1475 - 2175)	<0.0001
FEV1 predicted (%)	36 (30 - 48)	57 (47.75 - 68)	57 (48 - 68)	<0.0001
FEV1 PD (mL)	1110 (825 - 1380)	1330 (1082.5 - 1595)	1970 (1660 - 2350)	<0.0001
FEV1 PD (%)	38 (31 - 51.5)	61 (51.75 - 71.25)	63 (53 - 74)	<0.0001
Reversibility (%)	7 (1 - 13)	6.5 (2 - 13)	7 (0.5 - 11)	0.982
FVC predicted (mL)	2250 (1760 - 2750)	2150 (1710 - 2442.5)	3200 (2650 - 3825)	<0.0001
FVC predicted (%)	62 (52.5 - 73.5)	78.5 (66.5 - 90.25)	82 (70.5 - 90.5)	<0.0001
FVC post (mL)	2340 (1795 - 2835)	2190 (1850 - 2555)	3390 (2890 - 3900)	<0.0001
FVC post (%)	69 (56 - 77)	80.5 (69.75 - 94)	88 (73.5 - 97)	<0.0001
FEV1/ FVC pre (%)	47.2 (43.25 - 50.45)	61.85 (54.37 - 66.85)	57.5 (50.05 - 64.25)	<0.0001
FEV1/ FVC post (%)	47.5 (42.7 - 53.2)	63 (55.42 - 69.4)	58.6 (52.75 - 66.2)	<0.0001
DEM 25/75 (mL)	470 (355 - 643.06)	726.52 (530.64 - 831.40)	881.38 (769.3 - 1020.3)	<0.0001
DEM 25/75 (%)	17 (13.45 - 22.58)	26.01 (20.27 - 31)	29 (24.71 - 33.25)	<0.0001
TLC (mL)	6899.66 (6127.1 - 7762.4)	5625.33 (5090 - 5960)	6911.76 (6630 - 7192)	<0.0001
TLC predicted (%)	114.17 (108.08 - 120.5)	114.54 (109 - 124.03)	105.85 (97 - 108.82)	<0.0001
RV (mL)	4624.24 (4028.1 - 5237.1)	3334.51 (2940 - 3690.69)	3796.9 (3320 - 4191.8)	<0.0001
RV (%)	193 (183.98 - 218)	172.37 (156.15 - 188.5)	161.76 (138.5 - 179.6)	<0.0001
RV/TLC (%)	66 (61.35 - 68.41)	58.27 (54.95 - 63.64)	54.06 (49.5 - 58.85)	<0.0001
DLCO (mmol/kPa.min)	3.4 (2.75 - 4.27)	3.46 (2.98 - 3.94)	4.91 (4.44 - 5.57)	<0.0001
DLCO predicted (%)	43.48 (35 - 50.33)	47 (43 - 51.15)	56.14 (52.13 - 61)	<0.0001
DLCO/AV	0.95 (0.77 - 0.99)	0.95 (0.85 - 1.02)	1.01 (0.96 - 1.1)	0.0004
DLCO/AV predicted (%)	70.32 (59.96 - 77.42)	65.01 (55.62 - 68.33)	75.70 (71.64 - 83)	<0.0001
sGaw (1/kPa*sec)	0.39 (0.26 - 0.46)	0.55 (0.38 - 0.64)	0.52 (0.44 - 0.65)	<0.0001
FRC PL (L)	5.49 (4.99 - 6.19)	4.22 (3.90 - 4.64)	5.12 (4.86 - 5.47)	<0.0001
FRC PL predicted (%)	172.44 (162.16 - 181.70)	158.06 (149.47 - 169.8)	149 (140.5 - 157.92)	<0.0001
CAT score	26.2 (19.4 - 31.2)	24.4 (17.5 - 29.4)	22.5 (14.4 - 30)	0.245
Treatment				
Treatment (Yes)	87.18% (34)	65.28% (47)	44.78% (30)	<0.0001
ICS (Yes)	74.35% (29)	58.33% (42)	40.28% (27)	0.002
OCS (Yes)	17.95% (7)	0% (0)	2.98% (2)	0.0001
LABA - LAMA (Yes)	100% (39)	80.56% (58)	59.70% (40)	<0.0001
SABA (Yes)	53.85% (21)	43.05% (31)	29.85% (20)	0.045
LTRA (Yes)	0% (0)	6.94% (5)	1.49% (1)	0.087
Theophylline (Yes)	7.69% (3)	1.39% (1)	1.49% (1)	0.114
Blood				
Leucocytes (μL)	9.17 (7.35 - 11.77)	7.96 (6.86 - 9.37)	7.42 (6.48 - 8.84)	0.001
Neutrophils (%)	70.1 (65.15 - 76.05)	59.05 (53.55 - 64.57)	58.2 (53.35 - 62.95)	<0.0001
Lymphocytes (%)	20.7 (14.45 - 23.33)	30 (24.8 - 34.70)	29 (25.85 - 35)	<0.0001
Monocytes (%)	7.8 (6.8 - 9.3)	7.15 (6.4 - 8.6)	8.8 (7.75 - 10.05)	0.0002
Eosinophils (%)	1.8 (0.9 - 2.7)	1.9 (0.9 - 2.65)	2.38 (1.5 - 3.45)	0.033
Basophils (%)	0.3 (0.2 - 0.45)	0.4 (0.3 - 0.55)	0.42 (0.3 - 0.6)	0.006
Neutrophils (μL)	6255.06 (4844.1 - 8336.9)	4945.6 (3825.1 - 5821.5)	4503.6 (3723.7 - 5303)	<0.0001
Lymphocytes (μL)	1814.92 (1282.2 - 2265.5)	2295.7 (1914.5 - 2735.9)	2297.8 (1819.8 - 2781)	0.0003
Monocytes (μL)	797.79 (580.2 - 904.23)	624.48 (481.51 - 703.15)	674.25 (534.6 - 790.02)	0.012
Eosinophils (μL)	169.97 (74.38 - 230.37)	146.09 (75.82 - 221.60)	186.78 (111.2 - 288.64)	0.215
Basophils (μL)	25.8 (18.49 - 44.52)	33.15 (23.44 - 47.32)	32.04 (23.74 - 48.96)	<0.0001
Fibrinogen (g/L)	4.02 (3.47 - 4.63)	3.48 (2.99 - 3.82)	3.36 (2.86 - 3.76)	0.0005
CRP (mg/L)	3.4 (2.25 - 13.43)	2.15 (1 - 4.3)	2.1 (1.25 - 6.25)	0.027
Alpha 1 antitrypsin (g/l)	1.66 (1.51 - 1.77)	1.51 (1.37 - 1.61)	1.46 (1.35 - 1.55)	0.0002
Calcium (mmol/L)	2.4 (2.35 - 2.47)	2.43 (2.38 - 2.48)	2.41 (2.37 - 2.45)	0.173
25(OH) Vitamine D (ng/ml)	18.19 (10 - 29.5)	22.56 (16 - 30.72)	19 (12.5 - 27)	0.105
Phosphate (mmol/L)	0.9 (0.69 - 1.11)	1.01 (0.93 - 1.1)	0.85 (0.75 - 0.92)	<0.0001
IgE (KU/L)	162 (46.5 - 440.73)	70.5 (22 - 145.25)	96 (30 - 336.18)	0.065
RAST DPT (dI) %>0.35 (KU/L)	17.95% (7)	2.78% (2)	29.85% (20)	<0.0001
RAST Cat (eI), %>0.35 (KU/L)	2.56% (1)	1.39% (1)	13.43% (9)	<0.0001
RAST Dog (eS), %>0.35 (KU/L)	0% (0)	0% (0)	7.46% (5)	<0.0001
RAST Grass (GX3), %>0.35 (KU/L)	12.82% (5)	2.78% (2)	11.94% (8)	<0.05
RAST microog (MIX1), %>0.35 (KU/L)	17.95% (7)	6.94% (5)	7.46% (5)	<0.05
RAST Birch (t3), %>0.35 (KU/L)	2.56% (1)	0% (0)	1.49% (1)	0.151
Positive Aerobic Sputum Culture	17.95% (7)	6.94% (5)	1.49% (1)	0.007
Weight of sputum (g)	1.89 (0.99 - 2.38)	2.18 (1.34 - 2.81)	2.38 (1.31 - 2.98)	0.143
Total Cell Counts (10 ⁶ /g)	11.3 (4.90 - 18.48)	2.62 (1.08 - 5.10)	1.89 (0.89 - 4.60)	<0.0001
Squamous (%)	7 (0.5 - 15.13)	13.62 (5.75 - 21.63)	22.96 (11.5 - 34.5)	<0.0001
Viability (%)	72 (60 - 87)	63.74 (59.75 - 71)	68.80 (57 - 79)	0.034
Macrophages (%)	6 (2.47 - 11.46)	18.86 (13.95 - 24.3)	12 (5.3 - 18.6)	<0.0001
Lymphocytes (%)	0.4 (0 - 1)	2.4 (0.47 - 3.19)	1.37 (0.2 - 2)	<0.0001
Neutrophils (%)	86.6 (76 - 95.4)	63.39 (54.05 - 74.4)	74.6 (63.25 - 88.5)	<0.0001
Eosinophils (%)	3.2 (0.45 - 7.42)	3.5 (0.75 - 5.48)	2.6 (0.3 - 6.02)	0.888
Epithelial cells (%)	2.4 (0.2 - 5.1)	7.49 (1.8 - 9.61)	3.39 (0.7 - 6.7)	0.003
Macrophages (10 ³ /g)	538.51 (150 - 718.29)	540.19 (151.02 - 622.21)	269.64 (75.39 - 477.92)	0.018
Lymphocytes (10 ³ /g)	26 (2.31 - 47.47)	50.75 (5.98 - 65.67)	21.08 (4.48 - 39.49)	0.017
Neutrophils (10 ³ /g)	9608.94 (4630.9 - 16824)	1354.41 (472.6 - 3771.9)	1178.1 (494.2 - 3473.6)	<0.0001
Eosinophils (10 ³ /g)	372.96 (44.27 - 663.51)	34.76 (5.24 - 223.68)	41.14 (6.25 - 293.95)	0.001
Epithelial cells (10 ³ /g)	131.21 (38.45 - 260.3)	141.03 (17.91 - 293.95)	50.52 (11.49 - 120.2)	0.009

Reference

Haldar P., Pavord I.D., Shaw D.E., et al (2008). Cluster analysis and clinical asthma phenotypes. *American Journal of Respiratory and Critical Care Medicine*, 178: 218–224.

Husson F., Josse J., Pagès J. (2010). Principal Component Methods-Hierarchical Clustering-Partitional Clustering: Why Would we Need to Choose for Visualizing Data?, *Technical Report-Agrocampus, Applied Mathematics Department*, 1-17.