BMC Pulmonary Medicine

RESEARCH

Open Access

Quantitative CT and COPD: cluster analysis reveals five distinct subtypes with varying exacerbation risks

Chusheng Peng[1], Zizheng Chen[1,2], Haobin Zhou[1], Chaoyue Dai[1], Haolei Yuan[1], Yuan Gao[1], Fengyan Wang[2] and Zhenyu Liang[2*]

Abstract

Background The heterogeneity of chronic obstructive pulmonary disease (COPD) is increasingly recognized. To characterize the heterogeneity of COPD, we aimed to identify subtypes related to quantitative CT by using principal component analysis (PCA) and cluster analysis.

Methods The data of 1879 participants in the SPIROMICS study were obtained from the NHLBI Biologic Specimen and Data Repository Information Coordinating Center. A combination of PCA and k-means clustering was used to analyze the data from these participants in the SPIROMICS study. We randomly split the samples into training and validation sets. Clusters were evaluated for their relationship with acute exacerbation risk throughout the entire follow-up period. The results of the training set were confirmed in the validation set. To avoid sampling errors, we conducted 10 random sampling cycles. Normalized mutual information (NMI) was applied in every cycle to evaluate the stability of clustering.

Results We identified five clusters related to quantitative CT characterized as follows: (1) male-dominated low disease impact cluster, (2) obesity with relatively high symptom burden cluster, (3) airway wall lesion cluster, (4) lung upper region zone-predominant emphysema cluster, (5) severe emphysema cluster. There are significant differences in acute exacerbation risk among these five clusters.

Conclusions Cluster analysis identified 5 clusters related to quantitative CT of all participants in the SPIROMICS cohort with significant differences in baseline characteristics and acute exacerbation risk. The stability of clustering results was validated through NMI in 10 sampling cycles. In addition, dimensionality reduction results showed high reproducibility in different studies.

Keywords Chronic obstructive pulmonary disease, Cluster analysis, Disease axes, Longitudinal outcomes

*Correspondence:
Zhenyu Liang
490458234@qq.com
[1]Department of Clinical Medicine, Guangzhou Medical University, Guangzhou 511436, China
[2]Guangzhou Institute of Respiratory Health, State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory Disease, National Center for Respiratory Medicine, The First Affiliated Hospital of Guangzhou Medical University, 151Yanjiang Road, Yuexiu District, Guangzhou 510120, Guangdong, China

Background

Chronic obstructive pulmonary disease (COPD) is now one of the top three causes of death worldwide and most of these deaths occur in low-income and middle-income countries [1, 2]. At present, the diagnosis of COPD, which is most widely used, is confirmed based on FEV1/FVC, and the severity of the disease is classified based on the FEV1% predicted [3]. However, GOLD 2024 describes COPD as a heterogeneous lung condition characterized by chronic respiratory symptoms resulting from abnormalities of the airways and/or alveoli [4]. This definition emphasizes the heterogeneity of COPD and suggests that abnormal lung changes in early COPD are a continuous process that may also exist in subjects who have not yet experienced standard airflow obstruction, such as the Pre COPD or PRISM population [5–8]. In fact, this method of diagnosis and classification based on the degree of airflow obstruction has led to improved diagnosis and treatment of the disease [9, 10], but it can also lead to significant overlap between the different disease features and the proposed subtypes, which may not reflect phenotypic heterogeneity. With the increasing availability of CT imaging, CT-based measurements are becoming a reliable and objective method for assessing the risk of acute exacerbations in COPD [11]. Therefore, while maintaining the diagnostic approach for COPD, researchers have continued to verify multiple subtypes by integrating CT images with clinical variables [12–14], which have important consequences for clinical management, such as asthma-COPD overlap [15] and upper lobe-predominant emphysema [16].

With the increasing number of measurements related to COPD, researchers have applied unsupervised algorithms to discover COPD-related subtypes [17]. The advantage of this approach is that it can uncover complex relationships among diverse variables, thereby identifying potential subtypes. The two main types of unsupervised machine learning algorithms are the clustering and dimensionality reduction algorithms [12, 18]. Disease axes, which are generated by specific algorithms such as dimensionality reduction, represent continuous trajectories of disease phenotypic features [17, 19]. Unlike discrete clusters that group individuals into distinct subtypes, disease axes provide continuous measures that are composed of many contributing variables, making them more suitable for situations where the variable set follows a continuous distribution [18, 20]. Principal component analysis (PCA) is a widely used technique for reducing dimensionality by generating linear combinations of features that maximally explain the observed population variance and thus can be applied to define disease axes. However, because unsupervised algorithms are purely data-driven and do not rely on predefined labels, the reproducibility and stability of the obtained clusters or disease axes in different studies have not been extensively recognized.

Therefore, to identify subtypes related to quantitative CT and explore the reproducibility of disease axes and clusters, we employed principal component analysis (PCA) and clustering on clinical data obtained from the SPIROMICS cohort. Additionally, we assessed the reproducibility of the obtained disease axes and clusters.

Methods

Research data

The data analysis in this study utilized the open-access SPIROMICS (Subpopulations and Intermediate Markers in COPD Study) (Clinical Trial Registry: ClinicalTrials.gov, Identifier: NCT01969344, registered on October 25, 2013) dataset from the NHLBI Biologic Specimen and Data Repository Information Coordinating Center. The SPIROMICS study has previously been described in detail [21]. Briefly, between November 2011 and January 2015, 2982 participants, aged 40–80 years, were enrolled in a multicenter prospective observational study funded by the NIH [21] and distributed across four enrollment strata: never-smokers, smokers without COPD, smokers with mild or moderate COPD, and smokers with severe COPD. We excluded individuals with missing data (Supplementary Data: Fig. S1). The median follow-up period for the 1879 subjects in this study was 1611 days.

Sample splitting

To evaluate the effectiveness of clustering solutions, the SPIROMICS data were randomly split into equally sized training and validation sets. All subsequent model building was performed on the training data. The validation set was used to validate clustering characteristics.

Dimensionality reduction

To understand the phenotypic spectrum of COPD, we applied PCA to standardized individual feature, spirometric data and CT quantitative data from the SPIROMICS training set. (Supplementary Data: Table S1). We aimed to include enough principal components in the clustering analysis to account for a sufficient proportion of the variance. However, since the explained variance ratios for principal component 9 and those thereafter were similar and relatively low, we chose to retain the first eight principal components. (Supplementary Data: Fig. S2)

To better interpret the principal components, we applied varimax rotation to the first 8 principal components using the method of maximizing variance. Varimax rotation does not alter the relative positions of data points, and therefore does not affect the results of k-means clustering. The principal component scores were calculated by multiplying the rotated principal

component weights with the standardized original variables and summing.

Identifying clusters

We chose the k-means clustering method for this research, because it allows to verify the clustering results in the validation set by using the centers of the identified clusters. Then, we selected k = 5 as the optimal number of clusters, based on the greatest differentiation in clinical characteristics and acute exacerbation risk among the clusters. We used logistic regression to calculate the odds ratios (OR) of exacerbation for each subtype. To validate the significance of the subtyping, we adjusted the GOLD stage and calculated the OR of exacerbation for each cluster. Cox regression analysis was performed to estimate risk of exacerbation and adjusted for race.

Clusters validation

We computed the principal component scores for the validation set using the principal component score formulas derived from the training set. Based on the centers learned by the k-means algorithm in the training sample, we assigned the validation set to the nearest cluster center and allocated clusters in the validation sample accordingly. We utilized Wilcoxon rank sum tests to examine the differences in clustering features between the training and validation samples. K-means clustering was performed using the k-means function. The PCA was conducted using the psych package in R language. All statistical methods were executed in R language 4.3.2. Overview of methods is shown in Fig. 1.

Cross-validation

To avoid sampling errors, we conducted 10 random sampling cycles. In each cycle, cross validation was applied to each cycle (Details in Supplementary Data). We used normalized mutual information (NMI) to compare the two clustering solutions for each independent dataset.

Results

The characteristics of the training and validation samples are shown in Supplementary Data: Table S2, demonstrating their comparability.

The PCA with varimax rotation identified 8 principal components, explaining 73% of the variance in all variables. Detailed PCA loadings are shown in Fig. 2. PCA loadings represent the correlation between the original variables and the principal components, indicating how much each variable contributes to each component. Higher loadings reflect stronger contributions. We defined disease axes using these principal component scores.

Rotated principal component 1 (RC1), high representing by quantitative CT measurements of emphysema and air trapping, was interpreted as representing a multidimensional air trapping disease axis. Rotated principal component 2 (RC2), high representing by lung function measurement representation, was interpreted as a negative lung ventilation function disease axis. Rotated principal component 3 (RC3), mostly correlated with airway wall area percentages (loading scores 0.44 to 0.70), was interpreted as an airway wall lesion disease axis. Rotated principal component 4 (RC4), mostly correlated with high Pi10 index (loading scores 0.62 to 0.96), was interpreted as an airway wall thickness disease axis. Pi10 was calculated by regressing the square-root wall area on internal perimeter of included airways to predict the square-root wall area of a single hypothetical airway with internal perimeter of 10 mm. Notably, airway wall area percentages and Pi10 appear as independent components in the principal component analysis, representing principal component 3 and principal component 4, respectively, with low correlation between the two variables. This indicates that changes in Pi10 and airway wall area percentages are not linearly related in the progression of COPD. Rotated principal component 5 (RC5) to rotated principal component 8 (RC8) were not explained due to their low variance contribution.

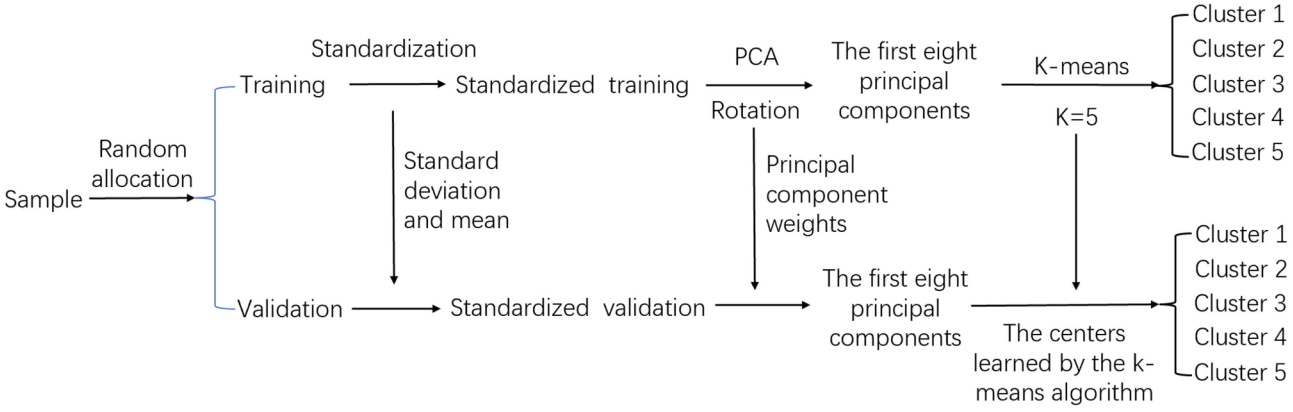

Fig. 1 Overview of clustering, principal component analysis, and validation

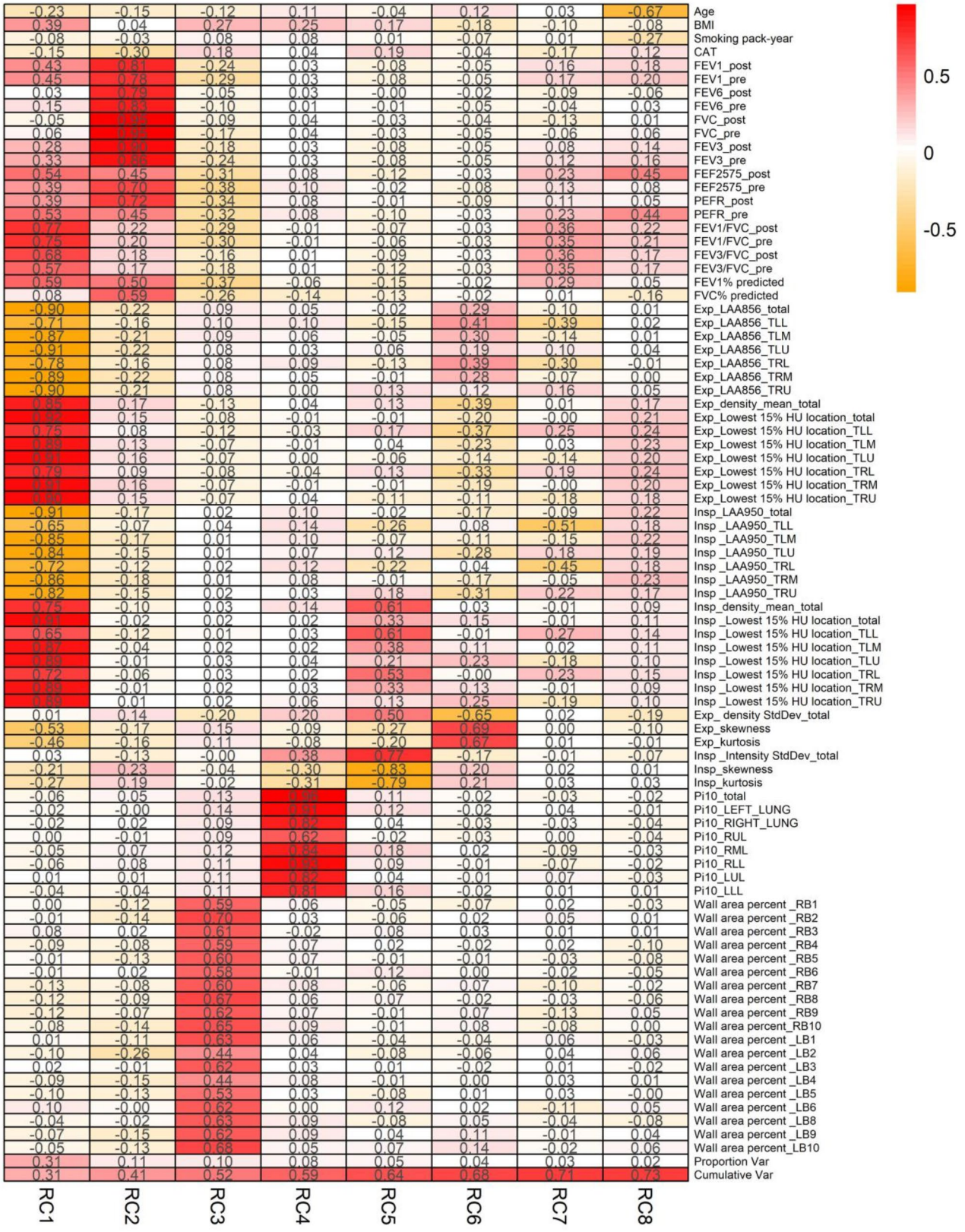

Fig. 2 Principal component analysis description

Through PCA and cluster analysis, we identified five clusters and ranked them according to the acute exacerbation rate, with the lowest exacerbation rate cluster labeled as cluster 1.

Cluster 1: male-dominated low disease impact

Cluster 1 represents 26% of the SPIROMICS training sample and is characterised by the highest predicted FVC% and FEV1%, the lowest CAT score and mainly composed of males (Male individuals account for 80% of this cluster) (Table 1). The majority of individuals in cluster 1 is primarily composed of control subjects and individuals classified as GOLD stage 0 and GOLD stage 1. (Table 2)

In terms of quantitative CT data, this cluster exhibits the highest segmental airway wall area, the highest lumen area and total bronchial area compared to other clusters, while simultaneously having the lowest airway wall area percentage.

Cluster 2: obesity with relatively high symptom burden

Cluster 2 represents 14% of the SPIROMICS training sample and characterised by the highest BMI among the 5 clusters and a relatively high burden of COPD symptoms (CAT score greater than 10) (Table 1). This cluster mainly includes members of GOLD stage 0, GOLD stage 2 and the control group. (Table 2)

Compared to cluster 1, this cluster has a decrease in FEV1% predicted by 0.09. However, due to a decrease in predicted FVC% predicted of 0.14, the FEV1/FVC ratio of cluster 2 increases by 0.05. Additionally, this cluster has the lowest level of emphysema among the 5 clusters, with the skewness and kurtosis of the lung density histogram being the lowest among the 5 clusters, which may suggest a possible association with pulmonary fibrosis [22, 23]. Female individuals account for 67% of this cluster.

Cluster 3: airway wall lesion

Cluster 3 represents 37% of the SPIROMICS training sample and characterised by the lowest airway wall area among the five clusters. However, the percentage of airway wall area in this cluster is relatively high, suggesting that the increase rate of airway wall area towards the airway lumen exceeds the decrease rate of airway wall area, resulting in a net increase in wall area percent (Table 1). This cluster mainly includes members of GOLD stage 0, GOLD stage 1, GOLD stage 2 and GOLD stage 3 (Table 2). Females represent 59% of this cluster.

Table 1 Cluster characteristics in training set, k = 5

Characteristic	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
N	242	130	343	113	111
Age	64.00 (56.00, 68.00)	54.00 (49.00, 60.00)	67.00 (62.00, 72.00)	65.00 (58.00, 71.00)	64.00 (59.00, 70.00)
BMI	27.00 (25.00, 30.00)	31.00 (27.00, 35.00)	29.00 (25.00, 33.00)	25.00 (22.00, 29.00)	25.00 (22.00, 27.50)
GENDER % female	48 (20%)	87 (67%)	203 (59%)	57 (50%)	35 (32%)
CAT	7.00 (3.00, 13.00)	15.00 (7.00, 22.00)	12.00 (7.00, 18.00)	16.00 (10.00, 22.00)	17.00 (12.00, 24.00)
BODE index	0.00 (0.00, 0.00)	0.00 (0.00, 1.00)	1.00 (0.00, 2.00)	2.00 (1.00, 4.00)	3.00 (2.00, 5.00)
FEV1/FVC	0.76 (0.69, 0.81)	0.81 (0.74, 0.85)	0.64 (0.56, 0.72)	0.50 (0.41, 0.60)	0.39 (0.34, 0.45)
FEV1% predicted	0.98 (0.89, 1.08)	0.89 (0.76, 0.96) *	0.74 (0.60, 0.86)	0.56 (0.42, 0.72)	0.40 (0.33, 0.51)
FVC% predicted	1.02 (0.93, 1.12)	0.88 (0.80, 0.95)	0.90 (0.80, 0.99)	0.87 (0.74, 0.99)	0.83 (0.72, 0.96)
Insp_LAA950_total	2.19 (1.17, 4.65)	0.36 (0.26, 0.60)	2.13 (0.98, 4.75)	19.16 (13.21, 28.45)	18.81 (11.61, 26.52)
Exp_density StdDev	173.74 (166.82, 181.98)	173.48 (166.00, 184.30)	169.18 (162.88, 178.78)	184.23 (172.87, 196.72)	163.24 (153.40, 172.22)
Exp_skewness	1.90 (1.66, 2.16)	1.61 (1.37, 1.86)	2.10 (1.85, 2.40)	2.11 (1.62, 2.38)	2.62 (2.37, 2.92)
Exp_kurtosis	3.53 (2.02, 5.39)	2.43 (1.12, 3.88)	4.52 (2.95, 6.11)	3.95 (2.03, 6.26)	7.46 (5.22, 10.04)
Exp_LAA856_TLU	7.10 (2.44, 15.53)	1.42 (0.42, 5.40)	11.92 (5.05, 26.22)	57.35 (36.29, 70.04)	50.92 (38.98, 63.43)
Exp_LAA856_TLL	6.26 (2.62, 14.95)	1.49 (0.41, 3.60)	9.55 (4.22, 21.58)	18.19 (6.39, 29.06)	55.29 (43.78, 65.93)
Exp_LAA856_TRU	6.03 (1.83, 13.96)	1.19 (0.33, 4.69)	10.48 (3.80, 27.27)	64.90 (47.74, 76.56)	48.31 (33.34, 60.02) *
Exp_LAA856_TRL	10.42 (4.50, 19.29)	2.81 (1.10, 5.30)	16.30 (6.80, 27.33)	28.96 (15.32, 41.83)	58.67 (49.20, 68.65)
Wall area†	39.36 (36.15, 43.29) *	32.99 (29.58, 37.14)	31.52 (28.45, 35.45)	32.71 (29.55, 37.46)	33.02 (29.81, 36.69)
Wall area percent†	0.58 (0.56, 0.59)	0.60 (0.58, 0.62)	0.61 (0.60, 0.63)	0.60 (0.58, 0.62)	0.61 (0.59, 0.63)
Lumen area†	30.01 (26.36, 33.67) *	22.77 (19.21, 26.71)	21.01 (18.37, 23.76)	23.82 (19.85, 28.11)	22.06 (18.71, 26.03)
Total bronchial area†	69.35 (62.78, 76.43) *	56.22 (48.83, 63.69)	52.71 (47.95, 58.82)	56.89 (50.55, 65.48)	55.29 (48.20, 61.69)

Notes: values are median (IQR) unless otherwise noted

*: *p* Value comparing mean in training to validation < 0.05 for Wilcoxon rank sum tests

†: Each subject's value is replaced by the mean value for 19 bronchi of lung segments

Abbreviations: exp, expiratory; insp, inspiratory; wall area = total bronchial area - lumen area; wall area percent = wall area / total bronchial area; TLU, upper left third of the lung; TLL, lower left third of the lung; TRU, upper right third of the lung; TRL, lower right third of the lung; kurtosis, kurtosis of lung density histogram; skewness, skewness of lung density histogram; StdDev, standard deviation of mean lung density; BMI, body mass index; HU, Hounsfield units

Table 2 The composition of COPD stages and exacerbation risk for each cluster

	Training					Validation				
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
GOLD_STAGE										
0	131 (54%)	83 (64%)	83 (24%)	6 (5.3%)	1 (0.9%)	123 (50%)	82 (66%)	84 (23%)	4 (3.8%)	0 (0%)
1	63 (26%)	2 (1.5%)	57 (17%)	11 (9.7%)	1 (0.9%)	59 (24%)	5 (4.0%)	60 (17%)	13 (12%)	2 (1.9%)
2	18 (7.4%)	20 (15%)	142 (41%)	47 (42%)	29 (26%)	35 (14%)	18 (14%)	158 (44%)	37 (35%)	22 (21%)
3	0 (0%)	6 (4.6%)	46 (13%)	41 (36%)	60 (54%)	0 (0%)	0 (0%)	33 (9.2%)	35 (33%)	50 (49%)
4	0 (0%)	0 (0%)	4 (1.2%)	8 (7.1%)	20 (18%)	0 (0%)	0 (0%)	5 (1.4%)	17 (16%)	29 (28%)
Unknown†	30 (12%)	19 (15%)	11 (3.2%)	0 (0%)	0 (0%)	31 (13%)	20 (16%)	18 (5.0%)	0 (0%)	0 (0%)
Exacerbation (%)	52 (21%)	47 (36%)	155 (45%)	76 (67%)	88 (79%)	56 (23%)	43 (34%)	182 (51%)	72 (68%)	79 (77%)
Exacerbation (OR)	1	2.069**	3.012***	7.505***	13.980***	1	1.798*	3.545***	7.261***	11.286***
Exacerbation (Adjusted OR)	1	2.140**	1.951**	3.353***	4.621***	1	2.016**	2.252***	2.711**	3.091**
Hazards ratio	1	1.942**	2.493***	4.906***	6.936***	1	1.620*	2.662***	4.375***	5.680***

Notes: GOLD unknown†: control group with less than 1 pack-year

***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$

OR (odds ratio): calculated from logistic regression

Adjusted OR (adjusted odds ratio): calculated from logistic regression, adjusting for GOLD stage

hazards ratio: calculated from cox proportional hazards model, cox regression analysis was adjusted for race

Cluster 4: lung upper region zone-predominant emphysema

Cluster 4 represents 12% of the SPIROMICS training sample and characterised by emphysema with marked upper zone-predominance. Air trapping in the upper regions are more severe compared to the lower regions, with the area of air trapping in the upper regions being more than twice that of the lower regions. Furthermore, this cluster exhibits the highest standard deviation in CT values, which may be related to the predominance of air trapping in the upper lung area of the cluster (Table 1). This cluster mainly includes members of GOLD stage 2 and GOLD stage 3. Females represent 50% of this cluster.

Cluster 5: severe emphysema

This cluster represents 12% of the training sample and is characterized by severe emphysema, airflow obstruction and the highest CAT scores. This cluster has the lowest BMI among the clusters and predominantly consists of males (Table 1). This cluster mainly includes members of GOLD stage 2, GOLD stage 3 and GOLD stage 4. Female individuals account for 32% of this cluster.

Validation of the clusters

All displayed baseline characteristics of the clusters have undergone Wilcoxon rank sum tests between the training and validation sets. Values marked with * indicate Wilcoxon rank sum tests p-values less than 0.05. The similarity of clustering features between the training and validation samples suggests that clustering can be reliably reproduced in separate data samples. The clustering characteristics of the validation set can be seen in Supplementary Data: Table S3. In cross validation, the results showed that the median NMI value was 0.66 for both the training and validation sets, which confirms the reproducibility of clustering results. (Supplementary Data: Fig. S3)

Acute exacerbation risk of each cluster

The exacerbation rate across the follow-up period in each cluster is shown in Table 2. Unless stated otherwise, the exacerbations mentioned in this study include mild, moderate, and severe exacerbation. Using cluster 1 as a reference, the hazards ratio and OR of acute exacerbation for other clusters are shown in Table 2.

To determine whether the association observed with these clusters and acute exacerbation risk was driven by severity of airflow obstruction, we repeated the cluster association tests adjusting for GOLD stage. The associations with exacerbation remained significant ($p < 0.01$). This suggests that the discovered clusters provide information on the risk of acute exacerbation independent from COPD severity as defined by GOLD.

Discussion

By employing PCA and cluster analysis on clinical data obtained from the SPIROMICS cohort, we distinguished 5 clusters as follows: (1) male-dominated low disease impact cluster, (2) obesity with relative high symptom burden cluster, (3) airway wall lesion cluster, (4) lung upper region zone-predominant emphysema, (5) severe emphysema cluster. The clinical characteristics and acute exacerbation risk of these clusters were validated through validation set.

This analysis reveals novel insights into COPD subtypes, particularly describing two clusters: the male-dominated low disease impact cluster and the airway wall lesion cluster, which have not been extensively described in prior research. Several factors contributed to the identification of these subtypes: (1) this study includes the

most comprehensive set of clinical variables compared to previous COPD clustering studies, encompassing lung function tests, emphysema, airway wall thickness, and individual characteristic variables. (2) our study population also includes subjects with no airflow obstruction. Our work also addresses the issue of reproducibility in clustering analysis across independent samples derived from the same cohort. Furthermore, we evaluated whether the correlation of clinical characteristics within the clusters matched those observed in previous studies, thereby enhancing the credibility of the clustering.

This study offers a novel insight into airway wall progression in COPD, suggesting that airway wall thickness alone may not accurately reflect airway pathology. In the male-dominated low disease impact cluster, which has the highest airway wall area, the exacerbation rate during follow-up was the lowest. Conversely, in the airway wall lesion cluster, despite having the lowest airway wall area, it also has the smallest total bronchial wall and lumen areas and a higher exacerbation rate than clusters 1 and 2. This indicates that COPD progression may involve not only thickening of the airway wall towards the lumen but also destruction of the airway wall, manifested as a decrease in total bronchial wall area.

The marked gender differences between clusters are of particular interest because gender was not included in the dimension reduction process in machine learning as an original variable. However, the variability caused by gender differences was captured by other selected variables, such as FEV1, indicators related to airway wall thickness and so on [24, 25].

This study further validates previous discoveries concerning subtypes. Firstly, Peter J. Castaldi et al. identified a subtype characterized by emphysema in the upper lung region using k-means clustering in the COPDGene cohort, which closely aligns with our cluster 4. However, our cluster 4 in the training set exhibits higher airflow obstruction levels compared to theirs [13]. Secondly, we identified a cluster characterized by obesity with a high symptom burden, similar to the PRISM cohort. This cluster shows a decrease in FEV1% predicted compared to cluster 1, but due to a synchronous decrease in FVC% predicted, it has the highest FEV1/FVC ratio among the 5 clusters. Our results partially confirm PRISM's features, such as higher BMI and higher proportion of females [8, 26]. Moreover, this cluster is primarily composed of individuals without COPD, with 64% of the patients being classified as GOLD stage 0. Thirdly, subtypes related to severe emphysema have been proposed in previous studies, and our study proposes a severe emphysema subtype with high airflow obstruction and high airway wall thickness [13, 14].

Additionally, this study confirms that the disease axis exhibits relatively high reproducibility in different studies. Compared with the factor loading done by KINNEY et al., where their factor 1 predominantly represented quantitative CT measures of air trapping and emphysema, and factor 2 predominantly represented lung function test measures, our study exhibits the similar results [19]. In addition, rotated principal component 3 (RC3) and rotated principal component 4 (RC4) in our study differ from factors 3 and 4 identified by Kinney et al. This discrepancy is attributed to the inclusion of a larger number of Pi10 and wall area percent indicators in our study.

This work has several limitations. Firstly, the study was based on a single cohort, and the external validity of the identified COPD subtypes needs to be confirmed in other independent datasets. Validation in different populations, regions, and clinical settings is crucial to assess whether the identified subtypes and their associated exacerbation risks are generalizable. Secondly, due to missing data, we excluded subjects with incomplete information. Although our sample size remains large and potential bias is minimal, this exclusion may still introduce some bias compared to the overall cohort.

Conclusion

In summary, through the utilization of PCA and k-means cluster analysis, we identified five clusters and their stability was validated in both the validation set and cross-validation. This will contribute to further describing and studying subtypes related to quantitative CT. In addition, this method shows high repeatability in independent data samples derived from the same cohort, providing a feasible approach for future research to address the issue of subtype repeatability obtained through unsupervised learning.

Abbreviations

COPD	Chronic obstructive pulmonary disease
SPIROMICS	Subpopulations and intermediate markers in COPD study
PCA	Principal component analysis
OR	Odds ratios
NMI	Normalized mutual information
RC1-8	Rotated principal component 1–8

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12890-025-03562-8.

Supplementary Material 1

Acknowledgements

This manuscript was prepared using SPIROMICS Research Materials obtained from NHLBI Biologic Specimen and Data Repository Information Center and does not necessarily reflect the opinions or views of the SPIROMICS of the NHLBI.

Author contributions

C.P is the guarantor of the content of the manuscript, has full access to all the data in the study and takes responsibility for the integrity of the data and the

accuracy of the data analysis. All authors participated in drafting and revising the submitted manuscript. Y.G, Z.C, C.D, and H.Y contributed to the analysis and interpretation of data. C.P, F.W, H.Z, and Z.L contributed to the conception and design of the study. All authors read and approved the final manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding
This work was supported by the National Natural Science Foundation of China (No.82270044, 82200044), the Guangzhou Science and Technology Planning Project (No.202201020451), the China International Medical Foundation (No. Z-2017-24-2301) and the State Key Laboratory of Respiratory Disease, Guangzhou Medical University (No. SKLRD-Z-202317).

Data availability
The dataset supporting the conclusions of this article is available in the NHLBI Biologic Specimen and Data Repository Information Center repository, [HLB01461719a, https://biolincc.nhlbi.nih.gov/studies/spiromics/].

Declarations

Ethics approval and consent to participate
This study was approved by the Ethics Review Committee of The First Affiliated Hospital of Guangzhou Medical University (ES-2023-046-01). The procedures used in this study adhere to the tenets of the Declaration of Helsinki. All SPIROMICS sites that enrolled patients obtained informed consent from patients.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 12 October 2024 / Accepted: 19 February 2025
Published online: 26 February 2025

References

1. Meghji J, Mortimer K, Agusti A, et al. Improving lung health in low-income and middle-income countries: from challenges to solutions. Lancet. 2021;397(10277):928–40. https://doi.org/10.1016/s0140-6736(21)00458-x
2. Halpin DMG, Celli BR, Criner GJ, et al. The GOLD summit on chronic obstructive pulmonary disease in low- and middle-income countries. Int J Tuberc Lung Dis. 2019;23(11):1131–41. https://doi.org/10.5588/ijtld.19.0397
3. Agusti A, Celli BR, Criner GJ, et al. Global initiative for chronic obstructive lung disease 2023 report: GOLD executive summary. Eur Respir J. 2023;61(4). https://doi.org/10.1183/13993003.00239-2023. [published Online First: 2023/03/02].
4. Celli B, Fabbri L, Criner G, et al. Definition and nomenclature of chronic obstructive pulmonary disease: time for its revision. Am J Respir Crit Care Med. 2022;206(11):1317–25. https://doi.org/10.1164/rccm.202204-0671PP. [published Online First: 2022/08/02].
5. Han MK, Agusti A, Celli BR, et al. From GOLD 0 to Pre-COPD. Am J Respir Crit Care Med. 2021;203(4):414–23. https://doi.org/10.1164/rccm.202008-3328PP
6. Wan ES. The clinical spectrum of PRISm. Am J Respir Crit Care Med. 2022;206(5):524–25. https://doi.org/10.1164/rccm.202205-0965ED. [published Online First: 2022/05/26].
7. Martinez FJ, Agusti A, Celli BR, et al. Treatment trials in young patients with chronic obstructive pulmonary disease and Pre-Chronic obstructive pulmonary disease patients: time to move forward. Am J Respir Crit Care Med. 2022;205(3):275–87. https://doi.org/10.1164/rccm.202107-1663SO. [published Online First: 2021/10/22].
8. Wan ES, Castaldi PJ, Cho MH, et al. Epidemiology, genetics, and subtyping of preserved ratio impaired spirometry (PRISm) in COPDGene. Respir Res. 2014;15(1):89. https://doi.org/10.1186/s12931-014-0089-y. [published Online First: 2014/08/07].
9. Buist AS, McBurnie MA, Vollmer WM, et al. International variation in the prevalence of COPD (the BOLD Study): a population-based prevalence study. Lancet. 2007;370(9589):741–50. https://doi.org/10.1016/S0140-6736(07)61377-4. [published Online First: 2007/09/04].
10. Calverley PM. The GOLD classification has advanced Understanding of COPD. Am J Respir Crit Care Med. 2004;170(3):211–2. https://doi.org/10.1164/rccm.2405008. discussion 14.
11. Smith BM, Traboulsi H, Austin JHM, et al. Human airway branch variation and chronic obstructive pulmonary disease. Proc Natl Acad Sci U S A. 2018;115(5):E974–81. https://doi.org/10.1073/pnas.1715564115. [published Online First: 2018/01/18].
12. Bell AJ, Ram S, Labaki WW, et al. Temporal exploration of COPD phenotypes: insights from the COPDGene and SPIROMICS cohorts. Am J Respir Crit Care Med. 2024. https://doi.org/10.1164/rccm.202401-0127OC. [published Online First: 2024/09/13].
13. Castaldi PJ, Dy J, Ross J, et al. Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. Thorax. 2014;69(5):416–23. https://doi.org/10.1136/thoraxjnl-2013-203601
14. Rennard SI, Locantore N, Delafont B, et al. Identification of five chronic obstructive pulmonary disease subgroups with different prognoses in the ECLIPSE cohort using cluster analysis. Annals Am Thorac Soc. 2015;12(3):303–12. https://doi.org/10.1513/AnnalsATS.201403-125OC
15. Gibson PG, Simpson JL. The overlap syndrome of asthma and COPD: what are its features and how important is it? Thorax 2009;64(8):728– 35. https://doi.org/10.1136/thx.2008.108027
16. Fishman A, Martinez F, Naunheim K, et al. A randomized trial comparing lung-volume-reduction surgery with medical therapy for severe emphysema. N Engl J Med. 2003;348(21):2059–73. https://doi.org/10.1056/NEJMoa030287. [published Online First: 2003/05/22].
17. Yuan NF, Hasenstab K, Retson T, et al. Unsupervised learning identifies computed tomographic measurements as primary drivers of progression, exacerbation, and mortality in chronic obstructive pulmonary disease. Ann Am Thorac Soc. 2022;19(12):1993–2002. https://doi.org/10.1513/AnnalsATS.202110-1127OC. [published Online First: 2022/07/14].
18. Castaldi PJ, Boueiz A, Yun J, et al. Machine learning characterization of COPD subtypes. Chest. 2020;157(5):1147–57. https://doi.org/10.1016/j.chest.2019.11.039
19. Kinney GL, Santorico SA, Young KA, et al. Identification of chronic obstructive pulmonary disease axes that predict All-Cause mortality: the COPDGene study. Am J Epidemiol. 2018;187(10):2109–16. https://doi.org/10.1093/aje/kwy087. [published Online First: 2018/05/18].
20. Castaldi PJ, Benet M, Petersen H, et al. Do COPD subtypes really exist? COPD heterogeneity and clustering in 10 independent cohorts. Thorax. 2017;72(11):998–1006. https://doi.org/10.1136/thoraxjnl-2016-209846
21. Couper D, LaVange LM, Han M, et al. Design of the subpopulations and intermediate outcomes in COPD study (SPIROMICS). Thorax. 2014;69(5):491–4. https://doi.org/10.1136/thoraxjnl-2013-203897. [published Online First: 2013/09/14].
22. Mascalchi M, Camiciottoli G, Diciotti S. Lung densitometry: why, how and when. J Thorac Disease. 2017;9(9):3319–45. https://doi.org/10.21037/jtd.2017.08.17
23. Hansell DM, Goldin JG, King TE, et al. CT staging and monitoring of fibrotic interstitial lung diseases in clinical practice and treatment trials: a position paper from the Fleischner society. Lancet Respiratory Med. 2015;3(6):483–96. https://doi.org/10.1016/s2213-2600(15)00096-x
24. Bhatt SP, Bodduluri S, Nakhmani A et al. Sex Differences in Airways at Chest CT: Results from the COPDGene Cohort. Radiology 2022;305(3):699–708. https://doi.org/10.1148/radiol.212985
25. Celli B, Vestbo J, Jenkins CR, et al. Sex differences in mortality and clinical expressions of patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2011;183(3):317–22. https://doi.org/10.1164/rccm.201004-0665OC
26. Higbee DH, Granell R, Smith GD, et al. Prevalence, risk factors, and clinical implications of preserved ratio impaired spirometry: a UK biobank cohort analysis. Lancet Respiratory Med. 2022;10(2):149–57. https://doi.org/10.1016/s2213-2600(21)00369-9

Publisher's note