13 April, 10:30 - 10:50, Room A3

Clinical characteristics of patients with chronic obstructive pulmonary disease (COPD): are they different?

Vera Enes[1], **Ana Helena Tavares**[2], **Vera Afreixo**[2], **Filipa Machado**[3], **Alda Marques**[1,3]

[1] Institute of Biomedicine (iBiMED), University of Aveiro, vera.enes@ua.pt

[2] Center for Research & Development in Mathematics and Applications (CIDMA), University of Aveiro

[3] Respiratory Research and Rehabilitation Laboratory (Lab3R), School of Health Sciences (ESSUA), University of Aveiro

---

Chronic Obstructive Pulmonary Disease is a major public health problem known to affect 800.000 people in Portugal. Symptoms include breathing difficulty, cough, fatigue and sputum. Although known that the disease progresses differently in patients with the same level of airway obstruction, the clinical characteristics of patients that may be associated with different disease phenotypes are not fully understood. This study aims to enhance our knowledge on the clinical characteristics of patients with COPD. A clustering procedure was performed, based on lung function, oxygen saturation, muscle strength and impact of the disease on patients' daily life and well-being.

Keywords: COPD, Clustering, Principal Component Analysis

Chronic Obstructive Pulmonary Disease (COPD) is a condition characterized by progressive and persistent airflow limitation resulting from a chronic inflammatory response of the airways and lungs in response to inhaled harmful gases and particles. Clinical diagnosis is based on airflow obstruction (assessed with lung function test-spirometry) and symptoms. Its prognosis depends on several factors including acute exacerbations (define as worsening of symptoms that result in additional therapy), environmental exposures, comorbidities and genetic predisposition [2]. COPD is burdensome not only for economic and social systems but most importantly to patients since it significantly affect their quality of life. It is known that the disease does not progress in the same way in all patients and that lung function, symptoms and reduction of quality of life may not be correlated. In fact, the interplay between patients' clinical characteristics and different disease phenotypes is not fully understood.

This study aims to enhance our knowledge on the clinical characteristics of patients with COPD. We retrospectively reviewed 394 patients with COPD. A clustering procedure is designed to stratify patients with COPD. From the 70 registered variables, we focus

on the most commonly assessed clinical variables: body mass index (BMI), age (AGE), the modified British Medical Research Council questionnaire (mMRC), number of acute exacerbations (AECOPD), number of hospitalization by respiratory cause (nHosp), the Charlson comorbidity index (CCI), Peripheral oxygen saturation (SpO2), forced expiratory volume in one second (FEV1pp), quadriceps muscle strength (QMSpp), 1-minute sit-to-stand test (1STS), COPD assessment test total score (CAT), St. George's Respiratory Questionnaire (SGRQ), dyspnoea and fatigue Borg scores (dysp.Borg and fat.Borg), and Hospital Anxiety and Depression Scale (anx.HADS and dep.HADS). Other variables were excluded due to missing values.

Clustering aims to find groups in a dataset. Since k-means looks for spherical clusters, it works best when the input variables are uncorrelated and have similar scales. In our dataset several variables were strongly correlated, e.g. the Pearson correlation between CAT and SGRQ was 0.79. We apply Principal Component (PC) Analysis on these vectors, obtaining a set of values of linearly uncorrelated variables. The number of components to retain is selected such that at least a given percentage of the variance is explained. The scores associated to those first PCs yield a data matrix, on which the k-means clustering algorithm is applied. The result of k-means depends on the number of clusters k, which is often hard to choose a priori. Therefore it is common practice to run the method for several values of k, and then select the 'best' value of k as the one which optimizes a certain criterion called a validity index. Many such indices have been proposed in the literature. Here we consider the Calinski-Harabasz index [1] and the GAP statistic [3].

Our procedure retained 6 principal components that explained 70% of the total variance of the dataset. Carrying out k-means clustering for different numbers of clusters yields and evaluating the obtained validation indices it appears that 3 or 5 clusters are appropriate. By looking at the composition of each cluster we define a patient prototype of each cluster.

Acknowledgements This work was funded by Programa Operacional de Competitividade e Internacionalização – POCI, through Fundo Europeu de Desenvolvimento Regional - FEDER (POCI-01-0145-FEDER-007628 and POCI-01-0145-FEDER-028806), Fundação para a Ciência e Tecnologia (PTDC/DTP-PIC/2284/2014 and PTDC/SAU-SER/28806/2017). Moreover, the costs resulting from the FCT hirings is funded by national funds (OE), through FCT, I.P., in the scope of the framework contract foreseen in the numbers 4, 5 and 6 of the article 23, of the Decree-Law 57/2016, of August 29, changed by Law 57/2017, of July 19. The work of VA and AT is partially funded by FCT under project UID/MAT/04106/2019.

References

[1] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3(1):1–27, 1974.

[2] Global Initiative for Chronic Obstructive Lung Disease. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease (2019 report), 2019.

[3] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.