Stratification of Patients with Chronic Obstructive Pulmonary Disease Using Volatile Organic Compounds

K. P. Exarchos, C. Chronis, L. Lipirou, V. Sakkas, and K. Kostikas

Abstract

Chronic obstructive pulmonary disease (COPD) is a common disease that causes long-term disability and death. Its natural history is punctuated by acute worsening of symptoms, called exacerbations, which are associated with increased mortality and hospitalization. In this work, we aim to stratify patients with COPD based on their risk for exacerbation; for this purpose, we employ non-invasive biomarkers, that is, volatile organic compounds (VOCs), acquired from the patients' exhaled breath coupled with their spirometry and age. We utilize a series of classification schemes with the best performing one achieving overall Accuracy = 93.5%. The yielded results are, therefore, encouraging and prompt for further investigation toward the utilization of VOCs in the management of COPD.

Keywords

Chronic obstructive pulmonary disease · COPD · Emphysema · Chronic bronchitis · Volatile organic compounds · VOCs · Exacerbation · Feature selection · Classification

1 Introduction

Chronic obstructive pulmonary disease (COPD) is chronic inflammatory disease-causing breathing difficulties; emphysema (damage to the air sacs of the lungs) and chronic bronchitis (long-term inflammation of the airways) comprise COPD pathophysiology causing a wide array of symptoms. The most common symptoms are cough, primarily productive, shortness of breath that is worse on exertion, therefore, causing limitation of physical activities and frequent respiratory tract infections. The predominant cause of COPD is tobacco smoking; however, air pollution and occupational exposure to certain chemicals and fumes increase the risk of developing COPD in smokers and non-smokers [14].

COPD poses a major health challenge from several perspectives; the symptoms of the disease that *are progressive* affect the day-to-day activities of the patients suffering from COPD leading to significant limitations and low quality of life. Subsequently, this also poses a significant load to their families especially in the last stages of the disease where the patients need more help with daily activities and self-care, coupled with the psychological burden. Moreover, COPD patients,

K. P. Exarchos (✉) · C. Chronis · K. Kostikas
Respiratory Medicine Department, School of Medicine, University of Ioannina, Ioannina, Greece

L. Lipirou · V. Sakkas
Department of Chemistry, University of Ioannina, Ioannina, Greece

P. Vlamos (ed.), *GeNeDis 2020*, Advances in Experimental Medicine and Biology 1337,
https://doi.org/10.1007/978-3-030-78771-4_1

especially in the late stages of the disease often suffer a spiral of infections and hospitalizations that impose considerable burden to the healthcare system overall. It should be highlighted that COPD is currently the fourth leading cause of death worldwide and in terms of cost is one of the most expensive conditions accounting for approximately 6% of the total annual healthcare budget in the European Union.

The course of the disease over time is marked by acute *worsenings* called exacerbations that are associated with increased hospitalizations, mortality and account for the greatest proportion of the total COPD burden on the healthcare system. Therefore, it is of utmost importance to identify the patients *that* are at high risk of having exacerbations and if possible identify early such events, in order to treat them early and/or adjust treatment to prevent future exacerbations. To this end, several risk stratification tools have been proposed in the literature featuring several biomarkers aiming to identify those patient subgroups that have higher risk for exacerbations, yet their utilization in clinical practice remains minimal.

Potential candidate biomarkers are volatile organic compounds (VOCs) that have been around for several years, but their utilization has been hampered by lack of standardization and validation. The exhaled breath consists of inorganic compounds (O_2, CO_2, and NO), nonvolatile organic compounds (isoprostane, leukotrienes, cytokines, and H_2O_2), and volatile organic compounds [4]. VOCs are the products of human metabolism and constitute a diverse group of carbon-based chemicals that are volatile at room temperature. VOCs are captured from the exhaled breath of patients and represent certain pathophysiological processes in the body. Moreover, the use and refinement of sensitive chemical methodologies such as gas chromatography and mass spectrometry have led to the capture and quantification of VOCs with considerable accuracy.

VOCs have been employed in a large number of studies and applications in healthcare and elsewhere. In the healthcare setting, VOCs have been used for the diagnosis of several conditions, for example, type II diabetes [7], Alzheimer's disease [13]; they have been associated with certain cancer types, such as breast cancer [1, 5], oral cancer [9], and lung cancer [2, 6, 10]. Due to the affinity of VOCs with the respiratory system, several applications have also been presented pertaining to pulmonary diseases. Specifically, [17] is a systematic review regarding the clinical use of VOCs especially in terms of diagnosis and monitoring in several respiratory-related diseases: asthma, COPD, cystic fibrosis, lung cancer, tuberculosis, mesothelioma, etc. The majority of studies focuses on asthma and lung cancer, and only sporadic applications in COPD exist; some exemplar applications are the following: discrimination between COPD patients and healthy nonsmokers [16], as well as between COPD patients undergoing exacerbation and stable COPD patients, differentiation between bacterial and viral infections in COPD patients [11].

In this study, we evaluated a decision support system utilizing VOCs as input aiming to stratify patients with COPD into two categories based on their risk of exacerbations. Based on their risk for exacerbation, patients can be managed more effectively, either by avoiding unneeded visits in low-risk patients or by monitoring more closely high-risk patients.

2 Materials and Methods

2.1 Study Design

In this study, we have enrolled 27 patients, all diagnosed with COPD of variable severity. From these patients, we have acquired breath samples using the RTubeVOC tubes. All samples have been collected during the steady state of the patients, none presenting any symptoms pertaining to an exacerbation. Moreover, during the same visit, spirometry was also performed. Each patient has been subsequently assigned as high or low risk, based on a composite index of their exacerbation history and higher blood eosinophil counts, both within the past year. Specifically, patients with two or more moderate/severe exacerbations and blood eosinophil count $\geq$300 cells/μL were assigned in the high-risk group, whereas

Table 1 Characteristics of the two patient subgroups

Variable	High risk	Low risk
Age	69 (±2)	67.9 (±7.3)
Sex (male/female)	4/0	23/1
Eosinophils (cells/μL)	447.5 (±61.8)	185.2 (±101)
FEV1%	33 (±23.6)	52.5 (±15.5)
FVC %	48 (±20.3)	66.4 (±13.8)
FEV1/FVC	49.3 (±14.2)	59.7 (±9)
FEF25-75%	32.3 (±41.5)	29.4 (±16.9)
Exacerbations	3.3 (±1.4)	0.7 (±1)

the rest were assigned to the low-risk group. The specific characteristics of the two groups are shown in Table 1.

2.2 Data Extraction

The samples were collected with special RTubeVOC tubes. These tubes are strictly single-exhalation devices and use two single-direction valves to maintain the one-way flow throughout the breathing cycle as the person exhales through the mouthpiece. It has a capacity of 65 ml with the aim of expelling the first fractions of exhaled air and trapping the last fraction that is representative of the internal lung. It consists of the tube which is made of polypropylene, the spout made of polyethylene, the stoppers that serve to trap the air and prevent losses and are made of medical vinyl, the adjustment valve made of silicone rubber (FDA approved components) and from two interventions in the form of a ring with circular cross-section used to seal the connection to the tube. All patients exhaled through the tube for 6 s in order to collect the last part of the exhalation. Then the tube remains sealed in order to prevent alterations in the composition of the air sample.

Samples must be strictly processed within 2 h otherwise the volatile compounds are deposited on the walls of the tube resulting in significant losses. The adsorption of the VOCs to be analyzed is done by the solid phase microextraction method (SPME). For the adsorption, a fiber made from a combination of divinylbenzene/carboxene/polydimethylsiloxane (CAR/DVB/PDMS) and diameter 30/50 μm is used. Before extracting VOCs, it is necessary to preprocess the fiber in order to clean it for avoiding impurities that may cause noise in the chromatogram. The pre-treatment of the fiber is done in the gas chromatography machine combined with mass spectrometer (Trace GC Ultra, Thermo Scientific-ISQ-Single Quadrupole-Thermo Scientific) with capillary column He (99.999%), which was selected as the carrier of gas and its flow rate was 2 ml/min that is going to be used for the analysis of samples as well. The pre-treatment stage includes heating the fiber to 200 °C for 2 h. After pre-treatment, the fiber is introduced into the sample through the adjustment valve. The system is sealed at the interconnection point with parafilm to minimize sample losses. The VOCs in the sample are adsorbed by the fiber for 37 min. Then the fiber is exposed to the gas chromatography machine injection system where the fiber-absorbed compounds are absorbed. In order for the VOCs to be absorbed, the temperature of the column is initially at 40 °C for 2 min. Then there is an increase in temperature from 7 °C/min to 200 °C and from that point the temperature rises 20 °C/min to 230 °C where it is kept for 3 min and maintaining the same rate, it reaches 270 °C and is kept there for 5 min. The total duration of the chromatographic analysis was 37 min. The temperature of the injection point was 200 °C and of the interface point was 285 °C.

For the formation of the ions of the analyzers, gas phase source of electron impact was used, and positive ionization took place. The voltage applied to accelerate electrons was 70 eV. The source temperature was 250 °C. The mass spectrometer consists of a simple tetrapolar mass analyzer (ISQ-Single Quadrupole-Thermo Scientific). The analysis was carried out with the function of the full scan and 0.5 s scan time. The range of mass area was 35–200 amu.

After the non-targeted GC/MS analysis, a chromatogram is received with peaks, which has different areas, retention times (rts), and heights corresponding to compounds. The raw data (area and rt) from the chromatograms were used for the subsequent analysis.

2.3 Data Preprocessing

The aforementioned procedure results in eight features that are used along with spirometry and age in the next steps of our methodology. Besides the small number of patients in our dataset, the most significant issue is the class imbalance. For this purpose, we applied Synthetic Minority Oversampling Technique (SMOTE) [3]. It should be highlighted that SMOTE is not applied a priori but as an intermediate step of the classification in order to avoid any bias. Next, we either employ the entire feature vector as is, or use two feature selection techniques, namely, Correlation-based Feature Selection [8] and the Wrapper algorithm [12]. Same as with the SMOTE algorithm, for bias purposes feature selection is applied as part of the classification process.

2.4 Classification

The resulting feature vectors are fed as input to a series of classification algorithms [15], specifically, Bayes Network (BN), Naive Bayes (NB), Artificial Neural Networks (ANN), Support Vector Machines (SVM), AdaBoostM1, Decision Tree (DT), and Random Forests (RF). All algorithms, both for preprocessing as well as for classification have been run using the Weka workbench (2017).

3 Methodology

Figure 1 depicts the steps followed in the methodology described in this work. Specifically, we employ a set of COPD patients that have been labeled as high and low risk based on their exacerbation propensity. Exhaled breath is captured from these patients and is subject to certain steps involving gas chromatography coupled with mass spectrometry in order to extract a set of meaningful features representing the VOCs in the patients' breath. The resulting VOCs along with each patient's spirometry and age comprise the feature vector that is used in the next steps of the methodology. Specifically, we deal with the class imbalance in the dataset by applying the SMOTE algorithm and then feature selection is performed. Next, we perform supervised classification, aiming to discriminate between the patients of the two classes.

4 Results and Discussion

As mentioned previously, we have utilized the feature vector either unchanged, that is, without performing any sort of feature selection or we have applied feature selection using two popular algorithms, namely: CFS and Wrapper. Next, we utilized seven classification algorithms in order to discern the patients of the two classes. For evaluation purposes, we have calculated the following performance metrics: Sensitivity, Specificity, Accuracy, and AUC (area under ROC curve). The results gained with each of these classification schemes are shown in three consecutive tables; specifically, Table 2 shows the results obtained without performing feature selection, Table 3 contains the results obtained after applying the CFS algorithm for feature selection, and Table 4 shows the respective results yielded after applying the Wrapper algorithm.

The features maintained by the CFS algorithm are the following: Age, FVC %, Area 7_94, Area 14_63.

Since the Wrapper algorithm is tailored to the classification algorithm invoked, different feature sets are retained. Specifically, with Bayes Network the following features are maintained: FEV1 (L), FEV1%, FVC %, Area 7_94, Area 14_63; with Naive Bayes: Age, Area 7_94, Area 8_29, Area 14_63; with ANN: FEV1%, FVC %, FEV1/FVC, FEF25-75 (L), Area 9_36, Area 14_63; with SVM: Age, FEV1 (L), FVC %, FEV1/FVC, FEF25-75 (L), FEF25-75%, Area 10_77; with AdaBoostM1: Area 7_94, Area 10_77, Area 11_26; with Decision Tree: Area 7_94; and with Random Forests: FEV1%, FEF25-75 (L), Area 7_94.

We observe that the best performance is achieved when the entire feature vector is fed as input to the Naive Bayes algorithm, yielding overall Accuracy = 93.5% and AUC = 0.983.

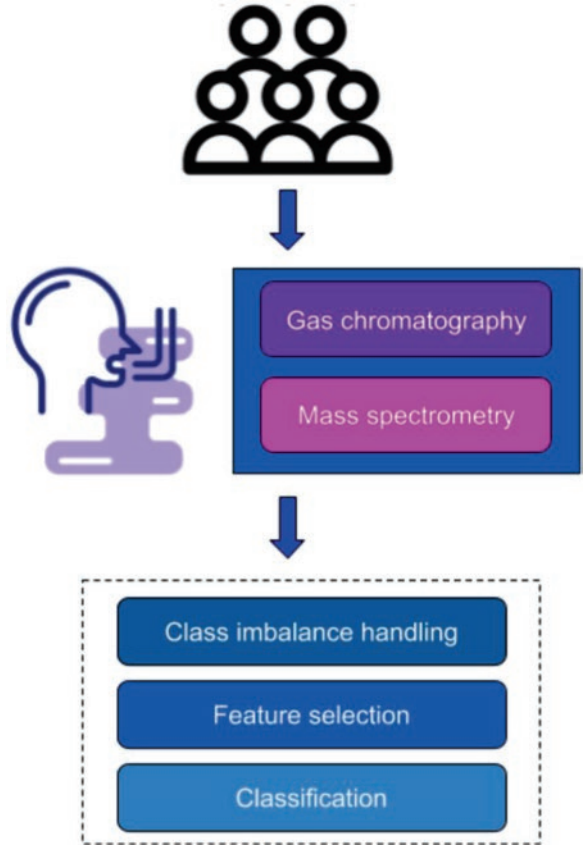

Fig. 1 Flowchart of the proposed methodology

Table 2 Results obtained without performing feature selection

Classification algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
Bayes Network	87	95.7	78.3	0.898
Naive Bayes	93.5	100	87	0.983
ANN	78.3	91.3	65.2	0.851
SVM	84.8	91.3	78.3	0.848
AdaBoostM1	87	95.7	78.3	0.934
Decision Tree	87	91.3	82.6	0.826
Random Forest	84.8	91.3	78.3	0.915

Table 3 Results obtained after applying the CFS algorithm for feature selection

Classification algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
Bayes Network	84.8	87	82.6	0.886
Naive Bayes	89.1	100	78.3	0.904
ANN	60.9	65.2	56.5	0.628
SVM	54.3	39.1	69.6	0.543
AdaBoostM1	76.1	91.3	60.9	0.879
Decision Tree	82.6	91.3	73.9	0.821
Random Forest	76.1	87	65.2	0.914

Table 4 Results obtained after applying the Wrapper feature selection algorithm

Classification algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
Bayes Network	80.4	87	73.9	0.881
Naive Bayes	87	95.7	78.3	0.960
ANN	76.1	69.6	82.6	0.864
SVM	76.1	78.3	73.9	0.761
AdaBoostM1	73.9	87	60.9	0.886
Decision Tree	82.6	87	78.3	0.880
Random Forest	82.6	87	78.3	0.938

Overall, the best results are obtained when no feature selection algorithm is applied. This is to be expected as the feature vector is relatively small and does not add significant complexity to the task under consideration. However, if the features pinpointed by the two algorithms are observed indifferent to the final outcome, we can see that the following features are more frequently maintained: FVC (%), Area 7_94, Area 14_63, FEV1 (L), FEV1%, and Area 10_77. Even though the results are quite encouraging, this can be partly attributed to the limited number of patients enrolled, leading to overtraining. Therefore, these preliminary results are yet to be validated with richer and more diverse patient sets.

It is important that all features used throughout the aforementioned methodology constitute non-invasive biomarkers that can be easily attained in an outpatient clinic. The standardization remains currently under fine-tuning but based on the obtained results, it could be an interesting and promising prospect.

5 Conclusion

In this chapter, we present a methodology utilizing non-invasive biomarkers for the identification of COPD patients that are at higher risk of having exacerbations over the course of the disease. Using VOCs coupled with spirometry, we developed a classification scheme that is able to pinpoint high-risk patients with significant accuracy. Nevertheless, further validation is needed in order to port this methodology in the clinical practice.

Acknowledgments This research is co-financed by Greece and the European Union (European Social Fund – ESF) through the Operational Program "Human Resources Development, Education and Lifelong Learning 2014–2020" in the context of the project "VOCs for the identification of high-risk COPD patients" (5047637).

References

1. Barash O, Haick H (2014) Exhaled volatile organic compounds as noninvasive markers in breast cancer. In: Omics approaches in breast cancer. Springer, New Delhi, pp 461–481
2. Barash O, Tisch U, Haick H (2013) Volatile organic compounds and the potential for a lung cancer breath test. Lung Cancer Manage 2:471–482
3. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
4. Dent AG, Sutedja TG, Zimmerman PV (2013) Exhaled breath analysis for lung cancer. J Thorac Dis 5(Suppl 5):S540–S550
5. Díaz de León-Martínez L, Rodríguez-Aguilar M, Gorocica-Rosete P et al (2020) Identification of profiles of volatile organic compounds in exhaled breath by means of an electronic nose as a proposal for a screening method for breast cancer: a case-control study. J Breath Res 14:046009
6. Gasparri R, Sedda G, Spaggiari L (2018) Volatile organic compounds and lung cancer: a tight link useful for diagnosis. Shanghai Chest 2:95
7. Greiter MB, Keck L, Siegmund T et al (2010) Differences in exhaled gas profiles between patients with type 2 diabetes and healthy controls. Diabetes Technol Ther 12:455–463
8. Hall MA (1999) Correlation-based feature selection for machine learning
9. Hartwig S, Raguse JD, Pfitzner D et al (2017) Volatile organic compounds in the breath of oral squamous cell carcinoma patients: a pilot study. Otolaryngol Head Neck Surg 157:981–987
10. Janssens E, van Meerbeeck JP, Lamote K (2020) Volatile organic compounds in human matrices as lung cancer biomarkers: a systematic review. Crit Rev Oncol Hematol 153:103037

11. Kamal F, Kumar S, Singanayagam A et al (2018) Volatile organic compound (VOC) analysis to differentiate between bacterial and viral respiratory infections in COPD. Respiratory infections. Aliment Pharmacol Ther 51(3):334–346
12. Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97:273–324
13. Mazzatenta A, Pokorski M, Sartucci F et al (2015) Volatile organic compounds (VOCs) fingerprint of Alzheimer's disease. Respir Physiol Neurobiol 209:81–84
14. Rabe KF, Hurd S, Anzueto A et al (2007) Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. Am J Respir Crit Care Med 176:532–555
15. Tan P-N, Steinbach M, Karpatne A, Kumar V (2019) Introduction to data mining. Addison-Wesley
16. van Berkel J, Dallinga J, Moller G et al (2009) A profile of volatile organic compounds in breath discriminates COPD patients from controls. Respir Med 104(4):557–563
17. van de Kant KDG, van der Sande LJTM, Jöbsis Q et al (2012) Clinical use of exhaled volatile organic compounds in pulmonary diseases: a systematic review. Respir Res 13:117. (2017) The WEKA workbench. Data mining 553–571