International Journal of Chronic Obstructive Pulmonary Disease Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

A Model Using Support Vector Machines Recursive Feature Elimination (SVM-RFE) Algorithm to Classify Whether COPD Patients Have Been Continuously Managed According to GOLD Guidelines

This article was published in the following Dove Press journal:
International Journal of Chronic Obstructive Pulmonary Disease

Jie Xia[1–4,*]
Lina Sun[5,*]
Suqin Xu[1–4]
Qiu Xiang[1–4]
Jianping Zhao[1–4]
Weining Xiong[1–4]
Yongjian Xu[1–4]
Shuyuan Chu[1–4,6]

[1]Department of Respiratory and Critical Care Medicine, Tongji Hospital, Tongji Medical College, Huazhong University of Sciences & Technology, Wuhan, People's Republic of China; [2]Wuhan Clinical Medical Research Center for Chronic Airway Diseases, Tongji Hospital, Tongji Medical College, Huazhong University of Sciences & Technology, Wuhan, People's Republic of China; [3]Key Laboratory of Pulmonary Diseases of Health Ministry, Tongji Hospital, Tongji Medical College, Huazhong University of Sciences & Technology, Wuhan, People's Republic of China; [4]Key Site of National Clinical Research Center for Respiratory Disease, Tongji Hospital, Tongji Medical College, Huazhong University of Sciences & Technology, Wuhan, People's Republic of China; [5]Department of Respiratory and Critical Care Medicine, Peking University Third Hospital, Beijing, People's Republic of China; [6]Laboratory of Respiratory Disease, Affiliated Hospital of Guilin Medical University, Guilin, People's Republic of China

*These authors contributed equally to this work

Correspondence: Shuyuan Chu; Jie Xia
Department of Respiratory and Critical Care Medicine, Tongji Hospital, Tongji Medical College, Huazhong University of Sciences & Technology, Wuhan, People's Republic of China
Tel +8613978345180; +8627-83665520
Email emilyyuanchu@163.com; xiajhzu1215@126.com

Purpose: Patients with chronic obstructive pulmonary disease (COPD) would have a poor prognosis if they were not continuously managed according to the Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines. We aim to develop a model to classify whether COPD patients have been continuously managed according to GOLD in the previous year.

Methods: The Managed group were COPD patients from a prospective cohort from November 2017 to November 2019, who have been continuously managed according to GOLD for 1 year. The Control group were COPD patients who were not continuously managed according to GOLD. They were from a retrospective cohort from October 2016 to October 2017 in the same hospitals as the Managed group. A synthetic minority oversampling technique (SMOTE) algorithm was used to up-sample the Managed group in a training dataset. Features for classification were selected using a support vector machine recursive feature elimination (SVM-RFE) algorithm. The classification model was developed using LibSVM, and its performance was assessed on the testing dataset.

Results: The final analysis included 15 subjects in the Managed group and 191 in the Control group. SVM-RFE selects nine features including smoking history, post-bronchodilator (post-)FVC before management, and those after 1-year follow-up (BMI, moderate and severe AECOPD frequency in previous 12 months, mMRC score, post-FEV1, post-FEV1%pred, post-FVC, and post-FEV1/FVC). For our model, positive predictive value is 66.7%, F1 score is 0.978, and AUC is 0.987.

Conclusion: SVM classifier combined with SVM-REF feature selection algorithm could achieve good classification between COPD patients who are or are not continuously managed. This model could be applied in clinical practice to help doctors make decisions and enhance COPD patients' compliance with standard treatment.

Keywords: COPD, GOLD, continuous management, support vector machine

Introduction

Chronic obstructive pulmonary disease (COPD) has been a major public health problem in the world for its high prevalence, morbidity, and mortality in recent decades. The World Health Organization estimates it will become the third leading

cause of death globally in 2030.[1] The situation is similar in China, where the overall prevalence of COPD is as high as 8.6% based on a national cross-sectional study in recent years.[2] Even more, the prevalence of COPD was 11.9% in men and 13.7% in people older than 40 years.[2] However, even for COPD patients living in a metropolis, more than 90% of them are not continuously managed according to the Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines.[3] Not only in China, but also in the US, Italy and South Korea, about 40–70% of COPD patients are not continuously managed according to GOLD guidelines.[4–6] For COPD patients, inappropriate treatment usually increases the rate of exacerbations, symptoms, and medical cost, worsens the quality-of-life and physical activity, and decreases survival.[7] Thus, it's crucial to classify COPD patients who are not continuously managed according to GOLD guidelines. Those patients could get more help from physicians to get appropriate treatment and improve their compliance with continuous and standard treatment, resulting in an improvement of their prognosis. However, recall bias and subjectivity of the information from patients could interfere with the doctor's decision, particularly when COPD patients are older in the majority. Therefore, it's also helpful for the doctor to get objective information about the treating history of COPD patients from a classification model.

In our study, we develop a model to classify whether COPD patients were continuously managed according to the GOLD guidelines using support vector machines (SVMs). SVMs are supervised machine learning techniques, which have been widely used in classification. SVMs work as a classifier by constructing a multidimensional hyper plane, which optimally discriminates two classes by maximizing the margin between the data clusters. SVMs are an effective approach for classification by using linear functions or special nonlinear functions, namely kernels, to transform the input space into a multidimensional space.[8] SVMs are trained with a training dataset, in which each case is marked as one category of the two. The trained algorithm of SVMs develops a model which could predict the category of a new case. Nowadays SVMs have been used in medical study.[9–11]

Based on the model from SVMs, we could make an objective judgment for COPD patients as to whether they'd been continuously managed according to GOLD guidelines in the previous 1 year, which could help to improve compliance with standard treatment, resulting in an improvement of therapeutic effect and prognosis.

Methods

Study Design and Participants

A prospective cohort was conducted in Tongji Hospital, Tongji Medical College, Huazhong University of Sciences & Technology in Wuhan and Peking University Third Hospital in Beijing, China, from November 2017 to November 2019. Adult patients with acute exacerbation of COPD (AECOPD) diagnosed by pulmonologists according to the GOLD guideline in 2017[12] were enrolled in the study. Briefly, COPD was diagnosed based on a forced expiratory volume in 1 s (FEV1)/forced vital capacity (FVC) ratio <70%, with a reversibility of less than 15% after inhaling 200 mg of salbutamol. AECOPD is the stage that respiratory symptoms acutely worsen, leading to additional therapy. When AECOPD patients had become stable and retained that level for at least 1 month, they were recruited into the study. All the pulmonologists were trained to diagnose COPD and manage patients according to GOLD guidelines in 2017 before the beginning of the trial.

Exclusion criteria for patients in the study were: 1) had other chronic pulmonary disease besides of COPD; 2) had a history of intubation within 3 years of enrollment; 3) were pregnant or prepared for pregnancy; or 4) had psychiatric disorders. Those subjects received COPD management according to GOLD guidelines in 2017[12] for 1 year. During this year, the subjects were received a face-to-face interview every 3 months. In each interview, subjects received a COPD assessment and adjusted therapeutic strategy according to GOLD guideline in 2017 based on the assessment. They also got education on smoking cessation and COPD self-care by a pulmonologist. That was the Managed group.

The Control group was from a retrospective cohort, who were hospitalized patients for AECOPD from October 2016 to October 2017 in the same hospitals as the Managed group. They received COPD treatment according to GOLD guidelines in 2016[13] in the hospital. They were recruited into the study 1 month after leaving hospital when they were in a stable condition of COPD. They had a face-to-face interview 1 year after recruiting into the study. In the interview, they received COPD assessment by a pulmonologist and completd a survey about the period of being COPD managed after leaving hospital. The inclusion and exclusion criteria

were the same as subjects in the Managed group. In addition, subjects were excluded if they were continuously managed for more than 1 month after leaving hospital.

The study protocol was approved by the Institutional Review Board (Research Ethics Committee of Tongji Medical College, Huazhong University of Sciences & Technology). All methods were performed in accordance with the declaration of Helsinki, and informed consent was obtained from each subject. The ClinicalTrials.gov ID is NCT03314077.

In the Managed group, one of 16 subjects was excluded for not continuously being managed according to GOLD guidelines for personal reasons within the 3 months. In the Control group, 217 subjects were retrospectively interviewed. Twenty-three of them were excluded for being continuously managed according to GOLD guidelines for more than 1 month after leaving hospital. Three subjects were lost to follow-up. Thus, 15 subjects in the Managed group and 191 in the Control group were included in the final sample (Figure 1).

Candidate Features to Classify Patients

The candidate predictors included demography characteristics of patients, and indicators of COPD assessment based on GOLD guidelines 2017. They were sex, age, education level, smoking history, BMI before and after 1-year follow-up, moderate and severe AECOPD frequency in previous 12 months before and after 1-year follow-up, modified medical British Research Council (mMRC) score after 1-year follow-up, COPD assessment test (CAT)[14,15] score after 1-year follow-up, lung function testing results after inhaling bronchodilator including FEV1, FEV1% predicted value (FEV1% pred), FVC, and FEV1/FVC before and after 1-year follow-up. Moderate and severe AECOPD were defined as GOLD guidelines.[12] Briefly, the moderate AECOPD is defined as requiring systemic corticosteroids or antibiotics or both, and the severe one is defined as receiving treatment at the emergency department or hospital.[12]

Statistical Analysis

Data were analyzed using SAS 9.4 (SAS Institute Inc., Cary, NC, USA). Group data were expressed as the mean ±standard deviation (SD). Significant differences in patient characteristics were evaluated using independent-samples *t*-test or chi-square test. *P*-values<0.05 were considered to be statistically significant.

Support Vector Machines

The classification model was developed using the LibSVM algorithm.[16] The model was built using Python 3.5.5 programming language, scikit-learn 20.0 library,[17,18] which is a powerful tool for scientific research.[19,20] In each group of subjects, 80% were randomly selected (training sample), who were used to develop the model. The remaining 20% (testing sample) served to test the model. Twelve patients in the Managed group and 151 in the Control group were randomly selected as the training sample, leaving three patients in the Managed group and 40 in the Control group as the testing sample. Since there was an imbalance in sample size between the Managed group and Control group (15 vs 191), we applied a synthetic minority over-sampling technique (SMOTE)[21] procedure to up-sample the Managed group. In order to keep the testing efficiency of the testing dataset in a real background, only the training dataset was up-sampled, but not the testing dataset. In our SMOTE procedure, four nearest neighbors of each

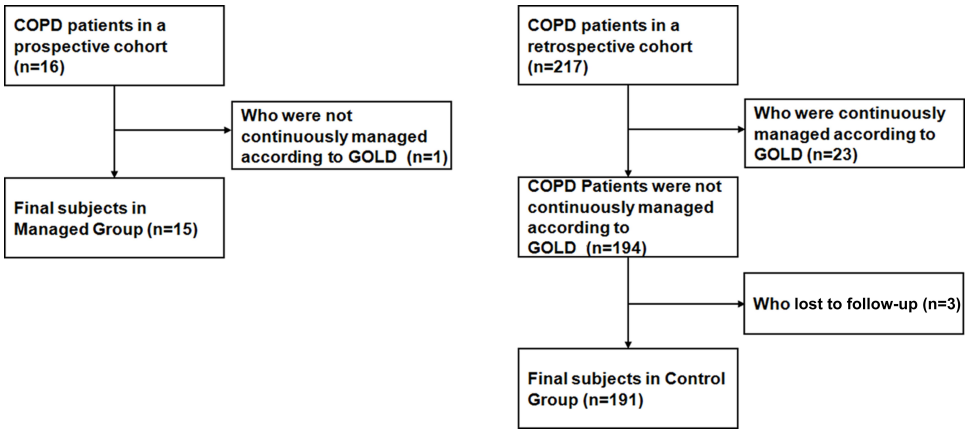

Figure 1 Population flow chart.

sample in the minority class were used in the training model,[22] which finally included 60 samples in the Managed group training dataset. The training data were standardized using z-score transformation, and the testing data were also transformed using the same parameters as those from the training data.

The support vector machine recursive feature elimination (SVM-RFE) algorithm[23] was used to find the features that could optimize the performance of the classifier. We used the grid-search and 10-fold cross-validation to train and estimate SVM hyperparameters. The grid-search was performed on the ranges of C = 0.01–10, kernel = linear or radial basis function (RBF). The best hyper parameters were RBF kernel, cost parameter as 1, gamma as scale. These hyperparameters and 10-fold cross-validation were used to train the classifier using the training set. The performance of the classifier was assessed on the testing dataset, which was not used during the training step. Since being lost to follow-up lead to missing data in our study, we excluded data from those who were lost to follow-up.

Results

Patient Characteristics

Table 1 illustrates participants' characteristics. The Managed group showed a higher proportion of smoking than the Control group. After 1-year of continuous management according to GOLD guidelines, the Managed Group showed less AECOPD frequency in the previous 12 months and a lower mMRC score than the Control group, whereas the AECOPD frequency was not significantly different before the follow-up. For patients in the COPD group according to the standard of GOLD, there were more patients in Group A and Group D in the Managed group than the Control group after 1-year follow-up, whereas there were less patients in Group B and Group C in the Managed group.

For lung function (Table 2), when inhaling a bronchodilator, FEV1 was significantly higher in the Managed group than the Control group before 1-year following, while FEV1 was higher in the Managed group after 1-year follow-up even though not significant. In contrast, FVC after inhaling a bronchodilator was significantly higher in the Managed group after 1-year follow-up, and was also higher before the follow-up, although not significant.

Model

When selected by SVM-RFE, the features in the model were smoking history, post-bronchodilator FVC before management, and those after 1-year follow-up including BMI, moderate and severe AECOPD frequency in the previous 12 months, mMRC score, post-bronchodilator FEV1, post-bronchodilator FEV1%pred, post-bronchodilator FVC, and post-bronchodilator FEV1/FVC. Those features could classify whether COPD patients were continuously managed according to GOLD guidelines using SVM. When the model was tested in the testing dataset, the positive predictive value (PPV) is 66.7% and F1 score is 0.978. The under the ROC (AUC) is 0.987. The kappa is 0.788. Even though the SVM hyperparameters were estimated by grid-search, a sensitivity analysis was conducted by reducing hyperparameter C in order to observe the robustness of the model. When we decreased the C-value to 0.1, the testing results were the same as those of the primary model, suggesting the robustness of our model.

Discussion

This study shows that we were able to develop a multivariable classifier as a promising tool to identify whether COPD patients were continuously managed according to GOLD guidelines in the previous 1 year. This model includes variables from spirometry, BMI, and moderate and severe AECOPD frequency, which could be used by both pulmonologists and general practitioners in everyday care.

In our study, the classification model is developed using LibSVM,[16] in which the hyperparameters were trained and estimated using the grid-search. The hyperparameters for a model influence the efficiency of the classifier and the result of classification. The optimal values of hyperparameters could eliminate the possibility of overfitting and underfitting. However, it's difficult to get optimal values of hyperparameters (C and kernel) in the training dataset on a given problem. Thus, we used a grid-search and 10-fold cross-validation technique to get the optimal values of hyperparameters for a kernel-based SVM model with the training dataset. The grid-search performs a comprehensive search over the specified parameter values for an estimator. Since SVMs using grid-search have shown good performance in medical research,[24–26] the selected hyperparameters in our study should be the best for the model and contribute to the best

Table 1 Participants' Characteristics in the Managed Group and the Control Group

Variables	Managed Group (N=15)	Control Group (N=191)	*P*-values
Sex (male)	13 (86.7%)	158 (82.7%)	0.972
Age (years)	62.2±10.6	66.5±10.3	0.145
Education (years)			0.055
≤6	3 (20.0%)	82 (42.9%)	
7–9	4 (26.7%)	51 (26.7%)	
10–12	2 (13.3%)	32 (16.8%)	
≥13	6 (40.0%)	26 (13.6%)	
Smoking (yes)	12 (80.0%)	95 (49.7%)	0.024
BMI before 1-year follow-up (kg/m^2)	22.2±3.2	22.8±3.8	0.571
BMI after 1-year follow-up (kg/m^2)	22.6±3.1	23.7±2.0	0.184
AECOPD_12 before 1-year follow-up	2.4±2.4	1.5±1.0	0.150
AECOPD_12 within 1-year follow-up	0.6±1.0	2.0±1.4	<0.001
mMRC score after 1-year follow-up	0.9±0.5	1.6±0.9	<0.001
CAT score after 1-year follow-up	8.7±3.9	11.1±6.2	0.132
GOLD group before 1-year follow-up			0.059
A	5 (33.3%)	20 (10.5%)	
B	3 (20.0%)	81 (42.4%)	
C	5 (33.3%)	61 (31.9%)	
D	2 (13.3%)	29 (15.2%)	
GOLD group after 1-year follow-up			<0.001
A	5 (33.3%)	1 (0.5%)	
B	5 (33.3%)	149 (78.0%)	
C	2 (13.3%)	39 (20.4%)	
D	3 (20.3%)	2 (1.0%)	

Abbreviations: BMI, body mass index; AECOPD, acute exacerbation of chronic obstructive pulmonary disease; AECOPD_12, AECOPD frequency in previous 12 months.

Table 2 Participants' Lung Function Post Inhaling Bronchodilator in the Managed Group and the Control Group

Variables	Managed Group (N=15)	Control Group (N=191)	*P*-values
Lung function before 1-year follow-up			
FEV1 (L)	1.84±0.89	1.43±0.69	0.032
FEV1%pred (%)	62.2±26.4	52.4±21.1	0.093
FVC (L)	3.19±1.21	2.84±0.88	0.144
FEV1/FVC (%)	47.0±15.8	53.8±39.3	0.507
Lung function after 1-year follow-up			
FEV1 (L)	1.74±0.83	1.38±0.29	0.112
FEV1%pred (%)	62.2±26.7	50.7±6.1	0.119
FVC (L)	3.20±0.94	2.62±0.36	0.033
FEV1/FVC (%)	53.6±15.6	51.2±5.7	0.572

Abbreviations: FEV1, forced expiratory volume in 1 second; % pred, % predicted; FVC, forced vital capacity.

results. In addition, our sensitivity analysis shows the robustness of the results of our model, confirming the optimal values of hyperparameters.

The imbalance problem in classification is quite common in medical data. The imbalanced datasets result in the classifier, which has a bias towards the majority class and tends to produce a majority class classifier.[27] In most cases, the class of interest is the minority class, which is the cause of lower sensitivity. In our study, the sample size of the Managed group and the Control group is significantly imbalanced (15 vs 191). To alleviate the effect of the imbalance in the training dataset, we adopted the powerful and effective SMOTE algorithm in the Managed group of the training dataset. In contrast, the SMOTE algorithm was not performed in the testing dataset, so the model could be tested in a real background. SMOTE performs better than simple oversampling when it's with SVM as a base classifier.[28] Moreover, since the overall effect of the model is significantly influenced by correctly classifying the majority class, we assessed the performance of the model, in our study, using f1 score, PPV, and AUC, which take into account the performance regarding the minority class as well. Our results showed a good performance of the model to classify patients between the Managed group and Control group in the testing sample, suggesting a good generalization of the model.

In our study, nine features to classify those subjects were selected using SVM-RFE, which are smoking history, post-bronchodilator FVC before management and those after 1-year follow-up including BMI, moderate, and severe AECOPD frequency in previous 12 months, mMRC score, post-bronchodilator FEV1, post-bronchodilator FEV1%pred, post-bronchodilator FVC, and post-bronchodilator FEV1/FVC. According to GOLD guidelines, spirometry is the basis for COPD diagnosis and

airflow limitation assessment.[12] And mMRC score and moderate or severe exacerbation history are the basis for assessing symptoms and the risk of exacerbation.[12] First, the mMRC score is useful to assess dyspnea symptoms, the risk of exacerbation, and hospitalization.[29] Second, moderate and severe AECOPD accelerates the process of COPD, which relates with high mortality and a repaid decline in health status.[12,30,31] Thus, moderate and severe AECOPD is considered in COPD assessment. Third, BMI is associated with FEV1 decline in COPD patients,[32] indicating the therapeutic efficiency. Therefore, mMRC score, moderate and severe exacerbation history, and BMI are closely related with the efficiency of COPD treatment and the prognosis of patients. On the other hand, in GOLD guidelines, the overall goals of COPD management are to optimize pulmonary function, to prevent progression, to improve quality-of-life, and to prevent and reduce the frequency and severity of exacerbations.[12] Thus, those features in our model were consistent with the overall goal of COPD management according to GOLD guidelines. Moreover, those predicative indicators are objective and accessible in clinical practice. Therefore, our model has implementation feasibility in clinical practice.

Since the overall goals of COPD management in GOLD guidelines are related with pulmonary function, prevent progression, exacerbations, and symptoms,[12] we selected candidate features for the model as follows: demographic characteristics, smoking history, BMI, moderate and severe AECOPD frequency in previous 12 months, mMRC score, CAT score, post-bronchodilator FEV1, FEV1%pred, FVC, and FEV1/FVC before and/or after 1-year follow-up. Those candidate predictors are involved in the overall goals of COPD standard management. The GOLD COPD group of patients was not selected as a candidate in our model, because the factors to assess patient's COPD group have been included in the model, such as post-bronchodilator FEV1 and FEV1% pred, mMRC score, CAT score, and moderate and severe AECOPD frequency in previous 12 months. Those factors that were respectively selected into the model could keep more information than using the COPD group of patients. That could avoid multicollinearity between features in the model as well.

We acknowledge limitations in this study. First, the controls were not parallel to the Managed group. The controls were recruited from a retrospective cohort, comprising hospitalized patients with COPD 1 year before the Managed group recruitment. Although the controls were diagnosed according to GOLD 2016 guidelines, the diagnosis standard and the majority of management in GOLD 2016 guidelines were the same as GOLD 2017 which was used for the Managed Group.[12,13] Moreover, controls were recruited from the same hospitals as the Managed group, as well as using the same inclusion and exclusion criteria. Thus, it may not bias the results of this study. Second, the sample size in the Managed group was significantly smaller than the Control group (15 vs 191). We adopted the SMOTE[21] up-sampling algorithm to increase the sample size of the Managed group in the training dataset. The SMOTE up-sampling algorithm adds synthetic data between the minority sample and its nearest neighbors based on a distance which is calculated by standard Euclidean distance between minority samples. That SMOTE algorithm avoids problems caused by simple oversampling with replacement and undersampling, such as low generalization of the final model or not taking full advantage of the original dataset.[33–35] Furthermore, when we tested the model in the testing dataset without up-sampling, our model showed a good performance in f1 score, PPV, and AUC. Thus, SMOTE could solve the problem of imbalanced sample size between two groups in our SVM model.

Conclusions

In conclusion, a limited number of quantitative indicators could classify whether COPD patients are continuously managed according to GOLD guidelines in the previous 1 year. This classification model could be useful and readily applicable in clinical practice to help doctors make decisions and enhance COPD patients' compliance with continuous and standard treatment according to GOLD guidelines in the long-term, which is crucial to improve the prognosis of COPD patients.

Data Sharing Statement

The ClinicalTrials.gov ID is NCT03314077. The individual deidentified participant data, specific data, and other study documents are available from the corresponding author upon reasonable request. The data has been made available from July 2020, which will last 3 years.

Funding

This work was supported by the National Key Technologies R&D Program (No. 2016YFC1304700, 1303900) and the National Natural Science Foundation of China (No. 81700042).

Disclosure of Interest

The authors report no conflicts of interest in this work. Jie Xia and Lina Sun contributed equally to this study and should be considered co-first authors.

References

1. The WHO website. Available from: https://www.who.int/respiratory/copd/burden/en/. Accessed April 17, 2020.
2. Wang C, Xu J, Yang L, et al. Prevalence and risk factors of chronic obstructive pulmonary disease in China (the China pulmonary health [CPH] study): a national cross-sectional study. *Lancet*. 2018;391(10131):1706–1717. doi:10.1016/S0140-6736(18)30841-9
3. Xiao T, Chen X, Wang N, Zhao Q, Fu C, Xu B. Study on the situation of drug use in patients with chronic obstructive pulmonary diseases in the Chinese communities of large cities. *Chin J Epidemiol*. 2017;38(2):142–146.
4. Surani S, Aiyer A, Eikermann S, et al. Adoption and adherence to chronic obstructive pulmonary disease GOLD guidelines in a primary care setting. *SAGE Open Med*. 2019;7:2050312119842221. doi:10.1177/2050312119842221
5. Palmiotti GA, Lacedonia D, Liotino V, et al. Adherence to GOLD guidelines in real-life COPD management in the Puglia region of Italy. *Int J Chron Obstruct Pulmon Dis*. 2018;13:2455–2462. doi:10.2147/COPD.S157779
6. Kim TO, Shin HJ, Kim YI, et al. Adherence to the GOLD guideline in COPD management of South Korea: findings from KOCOSS study 2011–2018. *Chonnam Med J*. 2019;55(1):47–53. doi:10.4068/cmj.2019.55.1.47
7. Vogelmeier CF, Criner GJ, Martinez FJ, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease 2017 report. GOLD executive summary. *Am J Respir Crit Care Med*. 2017;195(5):557–582. doi:10.1164/rccm.201701-0218PP
8. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Med Inform DecisMak*. 2010;10:16. doi:10.1186/1472-6947-10-16
9. Cornforth DM, Dees JL, Ibberson CB, et al. Pseudomonas aeruginosa transcriptome during human infection. *Proc Natl Acad Sci U S A*. 2018;115(22):E5125-E5134.
10. Ekins S, Puhl AC, Zorn KM, et al. Exploiting machine learning for end-to-end drug discovery and development. *Nat Mater*. 2019;18(5):435-441.
11. Irimia A, Lei X, Torgerson CM, Jacokes ZJ, Abe S, Van Horn JD. Support vector machines, multidimensional scaling and magnetic resonance imaging reveal structural brain abnormalities associated with the interaction between autism spectrum disorder and sex. *Front Comput Neurosci*. 2018;12:93. doi:10.3389/fncom.2018.00093
12. The GOLD website. Available from: https://goldcopd.org/. Accessed October 1, 2017.
13. The GOLD website. Available from: https://goldcopd.org/. Accessed October 1, 2016.
14. Zhou QT, Mei JJ, He B, et al. Chronic obstructive pulmonary disease assessment test score correlated with dyspnea score in a large sample of Chinese patients. *Chin Med J (Engl)*. 2013;126(1):11–15.
15. Tu YH, Zhang Y, Fei GH. Utility of the CAT in the therapy assessment of COPD exacerbations in China. *BMC Pulm Med*. 2014;14:42. doi:10.1186/1471-2466-14-42
16. Chang CC, Lin C. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol*. 2013;2:1–39. doi:10.1145/1961189.1961199
17. Python software website. http://www.python.org. Available from: Accessed October 5, 2019.
18. Scikit-learn website. https://scikit-learn.org/stable/. Availabel from: Accessed October 5, 2019.
19. Lima I. Python for scientific computing python overview. *Mar Chem*. 2006;9:10–20.
20. Millman KJ, Aivazis M. Python for scientists and engineers. *Comput Sci Eng*. 2011;13(2):9–12. doi:10.1109/MCSE.2011.36
21. Bowyer KW, Chawla NV, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321–357.
22. Wolff P, Graña M, Ríos SA, Yarza MB. Machine learning readmission risk modeling: a pediatric case study. *Biomed Res Int*. 2019;2019:8532892. doi:10.1155/2019/8532892
23. Li F, Zhao C, Xia Z, Wang Y, Zhou X, Li GZ. Computer-assisted lip diagnosis on traditional Chinese medicine using multi-class support vector machines. *BMC Complement Altern Med*. 2012;12:127. doi:10.1186/1472-6882-12-127
24. Arslan AK, Colak C, Sarihan ME. Different medical data mining approaches based prediction of ischemic stroke. *Comput Methods Programs Biomed*. 2016;130:87-92. doi:10.1016/j.cmpb.2016.03.022
25. Gupta Y, Lama RK, Kwon GR. Prediction and classification of Alzheimer's disease based on combined features from apolipoprotein-E genotype, cerebrospinal fluid, MR, and FDG-PET imaging biomarkers. *Front Comput Neurosci*. 2019;13:72.
26. Zhao D, Liu H, Zheng Y, He Y, Lu D, Lyu C. A reliable method for colorectal cancer prediction based on feature selection and support vector machine. *Med Biol Eng Comput*. 2019;57(4):901-912. doi:10.1007/s11517-018-1930-0
27. Wei Q, Dunbrack RL. The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PLoS One*. 2013;8(7):e67863. doi:10.1371/journal.pone.0067863
28. Wallace B, Small K, Brodley C, Trikalinos T. Data mining (ICDM), 2011 IEEE 11th International Conference on; 2011; Vancouver, Canada: Class imbalance, Redux. 754–763.
29. Cheng SL, Lin CH, Wang CC, et al. Comparison between COPD assessment test (CAT) and modified medical research council (mMRC) dyspnea scores for evaluation of clinical symptoms, comorbidities and medical resources utilization in COPD patients. *J Formos Med Assoc*. 2019;118(1 Pt 3):429–435. doi:10.1016/j.jfma.2018.06.018
30. Suissa S, Dell'Aniello S, Ernst P. Long-term natural history of chronic obstructive pulmonary disease: severe exacerbations and mortality. *Thorax*. 2012;67(11):957–963. doi:10.1136/thoraxjnl-2011-201518
31. Müllerova H, Maselli DJ, Locantore N, et al. Hospitalized exacerbations of COPD: risk factors and outcomes in the ECLIPSE cohort. *Chest*. 2015;147(4):999–1007. doi:10.1378/chest.14-0655
32. Sun Y, Milne S, Jaw JE, et al. BMI is associated with FEV1 decline in chronic obstructive pulmonary disease: a meta-analysis of clinical trials. *Respir Res*. 2019;20(1):236.
33. Nakamura M, Kajiwara Y, Otsuka A, Kimura H. LVQ-SMOTE-learning vector quantization based synthetic minority over-sampling technique for biomedical data. *BioData Min*. 2013;6:1–10. doi:10.1186/1756-0381-6-16
34. López V, Fernandez A, Garcia S, Palade V, Herrera F. An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. *Inf Sci*. 2013;250:113–141. doi:10.1016/j.ins.2013.07.007
35. Luengo J, Fernandez A, Garcia S, Herrera F. Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary undersampling. *Soft Comput*. 2011;15:1909–1936. doi:10.1007/s00500-010-0625-8