Cell J
Cell Journal(Yakhteh)
Homepage: https://www.celljournal.org/
Original Article
Vol 25, No 3, March 2023, Pages: 203-211

Chronic Obstructive Pulmonary Disease: Novel Genes Detection with Penalized Logistic Regression

Kimiya Gohari, Ph.D.[1], Anoshirvan Kazemnejad, Ph.D.[1]*, Shayan Mostafaei, Ph.D.[2], Samaneh Saberi, Ph.D.[3], Ali Sheidaei, Ph.D.[4]

1. Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran
2. Division of Clinical Geriatrics, Department of Neurobiology, Care Sciences and Society, Karolinska Institute, Stockholm, Sweden
3. HPGC Research Group, Medical Biotechnology Department, Biotechnology Research Center, Pasteur Institute of Iran, Tehran, Iran
4. Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran

Abstract

Objective: This study aimed to introduce novel techniques for identifying the genes associated with developing chronic obstructive pulmonary disease (COPD) and to prioritize COPD candidate genes using regression methods.

Materials and Methods: This is a secondary analysis of the data from an experimental study. We used penalized logistic regressions with three different types of penalties included least absolute shrinkage and selection operator (LASSO), minimax concave penalty (MCP), and smoothly clipped absolute deviation (SCAD). The models were trained using genome-wide expression profiling to define gene networks relevant to the COPD stages. A 10-fold cross-validation scheme was used to evaluate the performance of the methods. In addition, we validate our results by the external validity approach. We reported the sensitivity, specificity, and area under curve (AUC) of the models.

Results: There were 21, 22, and 18 significantly associated genes for LASSO, SCAD, and MCP models, respectively. The most statistically conservative method (detecting less significant features) was MCP detected 18 genes that were all detected by the other two approaches. The most appropriate approach was a SCAD penalized logistic regression (AUC= 96.26, sensitivity= 94.2, specificity= 86.96). In this study, we have a common panel of 18 genes in all three models that show a significant positive and negative correlation with COPD, in which RNF130, STX6, PLCB1, CACNA1G, LARP4B, LOC100507634, SLC38A2, and STIM2 showed the odds ratio (OR) more than 1. However, there was a slight difference between penalized methods.

Conclusion: Regularization solves the serious dimensionality problem in using this kind of regression. More exploration of how these genes affect the outcome and mechanism is possible more quickly in this manner. The regression-based approaches we present could apply to overcoming this issue.

Keywords: COPD, Gene Expression, LASSO, MCP, Panelized Logistic Regression

Citation: Gohari K, Kazemnejad A, Mostafaei Sh, Saberi S, Sheidaei A. Chronic obstructive pulmonary disease: novel genes detection with penalized logistic regression. Cell J. 2023; 25(3): 203-211. doi: 10.22074/CELLJ.2022.557389.1048.

Introduction

Chronic obstructive pulmonary disease (COPD), a progressive inflammatory disorder, that is characterized mainly by airway obstruction is predicted to be one of the three first death causes worldwide by 2030 (1). The disease has exhibited by emphysema, small airway obstructions, and chronic bronchitis. Smoking is the first-line risk factor in a healthy airway and alveolar development background. Some other general risks are a history of maternal/paternal asthma, maternal smoking, and childhood asthma or respiratory infections. Polluted air, second-hand smoke, and malnutrition could also lead to COPD in the susceptible population (2).

This complex disease may be influenced by different gene interactions (2). The incidence of this disease is increasing worldwide. Its mortality rate has shown a 12.3 % increase from 1990 to 2017, while this rate increased to 60 % in 2020 in comparison with 1999 (3). Thus, the disease is a global public health challenge with high prevalence, mortality, and disability rates, while diagnosis is usually based on spirometry results and clinical signs with an inability to diagnose recurrences (4). It seems that identifying more effective initial diagnostic methods based on reliable biomarkers is essential. DNA microarrays now permit scientists to screen thousands of genes simultaneously and determine whether the expression pattern of these genes, changes in tissues, particularly pulmonary tissue. So, new analytical methods

Received: 08/July/2022, Revised: 10/December/2022, Accepted: 24/December/2022

*Corresponding address: P.O.Box: 14115-111, Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran
Email: Kazem_an@modares.ac.ir

Royan Institute
Cell Journal (Yakhteh)

must be developed to select COPD-related genes (2). Since genome-wide association studies (GWAS) have shown the different phenotypes or severity of COPD are associated with numerous genetic variants such as *CHRNA3/CHRNA5* (5), *HHIP* (6), *FAM13A*, and *CYP2A6* (7). On the other side, many studies have performed expression profiles of COPD-related genes, which have screened thousands of different expression genes (DEGs) involved in the development and progression of the disease. Among these studies, due to sample heterogeneity and diagnostic platform differences, they did not have very successful results and could not help in the correct diagnosis. Therefore, the associations between genomics and disease incidence and progression could be studied precisely through machine-learning techniques (8). Also, network medicine has been introduced to facilitate the investigations of genomics, transcriptomics, proteomics, and other "omics" to cast a more elucidating light on the pathogenesis complexity of diseases likewise COPD. One of the properties of microarray data is that the number of genes (parameter number in statistics) exceeds the number of samples (number of observations in statistics). They are dealing with the situation commonly known as the high dimensional dataset. However, logistic regression as a highly appropriate classification tool for such high-dimensional datasets from the microarray technique has drawbacks, such as the emergence of irrelevant data (9).

Moreover, regression analysis has been established to overlook the multicollinearity problem such as the strong correlation between two or more genes in the regression model. So, overfitting and multicollinearity are the most common problems in high-dimensional data when applying statistical classification and prediction methods (10). Nowadays, researchers update, improve, and optimize the techniques such as the least absolute shrinkage and selection operator (LASSO), minimax concave penalty (MCP), and smoothly clipped absolute deviation (SCAD) to introduce statistical learning models to overcome this issue. Penalized Logistic Regression models represent spares and interpretable models in high-dimensional datasets and control the multicollinearity (9). There has been no study on the omics data integration on COPD to compare these approaches.

Although, LASSO has many excellent properties, it is a biased estimator that does not always tend to zero as the sample size is increased. The bias of the LASSO estimation for a truly non-zero variable is constant even for significant regression coefficients. One approach to reducing this bias is using the weighted penalty approach. If we choose the weights that give lower weight to the variables with significant coefficients, we can reduce the estimation bias of the LASSO. It is the motivation of adaptive the LASSO approaches. The SCAD penalty retains the penalization rate (and bias) of the LASSO for small coefficients but continuing relaxes the rate of penalization as the absolute value of the coefficient increases. The idea behind the MCP is very similar. In comparison with the SCAD, the MCP immediately relaxes the penalization rate, although, its rate remains flat before decreasing (11). As, these approaches may lead to different results, selecting the appropriate one needs to consider various clinical and statistical aspects.

Based on the biological perspective, a smaller subset of genes may cause a definite disease (12). Therefore, the present study was designed to apply statistical-learning methods for a better understanding of the genetic etiology of COPD affected by the previous smoking habits.

Materials and Methods

Ethics declarations

Our project was founded under the Ethical Committee of the National Institute for Medical Research Development (NIMAD), Tehran, Iran (I.R.NIMAD.REC.1398.115).

Study population and dataset

This is a secondary analysis on an experimental study data that used genome-wide expression profiling to define gene networks relevant to the COPD stages. The raw data of gene expression architecture in the small airway epithelium (SAE) of COPD affected was retrieved from the Gene Expression Omnibus (GEO) site in the National Center of Biotechnology Information (NCBI) database (13), with the accession number "GSE22148", with 54,675 probes from 143 patients with GOLD stage of COPD, from 2 to 4 stages. Genome-wide gene expression analysis was performed using Affymetrix Human Genome U133 Plus 2.0 array (GPL570) (14).

The subjects were selected of the evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE). We chose a three-year multicenter longitudinal study with four specific aims: i. Definition of clinically relevant COPD subtypes; ii. Identification of parameters that predict disease progression in these subtypes; iii. Examination of biomarkers that correlate with COPD subtypes and may predict disease progression; and iv. Identification of novel genetic factors and/or biomarkers that both correlate with clinically relevant COPD subtypes and predict disease progression (15). Totally, the gene arrays on 140 subjects passed the quality control criteria (69 GOLD stage II and 71 GOLD stages III/IV) which we entered in our study. More details about this database and data gathering were published elsewhere (14).

Normalization and filtering of primary probes

At first, the data was transformed according to the logarithmic scale 2 for better distribution. The "sva" and "affy" packages were used respectively for removing batch effects and other unwanted variations in data and for statistical comparisons. Also, the standardization and normalization in the "limma" package were performed. In addition, differential analysis of gene expression data was conducted using the adjusted P value based on the Benjamini-Hochberg-FDR correction at α=0.05. The penalized regression approaches shrink the coefficients to zero and eliminate the unrelated features. Therefore, the large dimensionality is not mattering further, and the P value of the adjusted model effectively reports the statistically significant. All statistical analyses were performed using R version 3.5.2.

LASSO, MCP, and SCAD logistic regressions

The LASSO regression uses the absolute value of the magnitude of the coefficient as a penalty term and hence provides an automatic gene selection. The penalty-based models tend to shrink the coefficients of correlated variables toward each other, which is suitable for multicollinearity and grouped selection. However, the LASSO penalty is indifferent to choose a set of solid but correlated variables. Therefore, LASSO is good for simultaneous estimation and eliminating trivial genes, but not for grouped selection. However, it is known that LASSO requires rather stringent conditions on the design matrix to be variable selection consistent (9).

Non-convex penalized high-dimensional regression has recently received considerable attention, especially for identifying the unknown sparsity pattern. Fan and Li recommended the SCAD penalty, which enjoys the oracle property for variable selection. The zero coefficients can be estimated as an exact zero with probability approaching one and estimate the non-zero coefficients as efficiently as if the actual sparsity pattern is known in advance (16).

Zhang (17) proposed the MCP penalized regression and devised a novel algorithm that, when used together, can achieve the oracle property under certain regularity conditions. The mentioned logistic classifiers were done by "ncvreg" R packages.

Cross-validation, stability, and accuracy

The K-fold Cross-Validation scheme (K-CV), a common technique, evaluates the classifier performance. The K-CV estimation of the error is the average value of the errors committed in each fold. Thus, the K-CV error estimator depends on the training set and the partition into folds (18). In the present study, the algorithms split the data set by using 100 times repeated random sub-sampling in 10-fold cross-validation, permuting the sample labels every time. The cross-validated performance is summarized by observing sensitivity and specificity and the Youden index. Furthermore, the receiver operator characteristic (ROC) curve and its area under curve (AUC) was used to calculate the accuracy of classifier performance. We used the "cv.ncvreg" and "roc" function in the "ncvreg" and "pROC" R packages for K-CV and ROC analysis, respectively.

Interactive cluster heatmap

A heatmap is a popular graphical method for visualizing high-dimensional data. Rows and columns are sorted using a hierarchical clustering technique. The interactive cluster heatmap was applied by the "heatmap" R package.

External validity

We used completely independent data to explore the external validity of the findings. This data is available online with accession number "GSE20257" on the NCBI database. We fitted to the regression models to this data and a test data set. The sensitivity, specificity and AUC of each model were calculated to predict the COPD stages.

Results

Differential analysis of genes expression data

Differential analysis was performed on the array expression profiling of 54675 probes. The expression profiling of 140 patients at the GOLD stages 2-4 COPD was used in this study. The differential expression analysis results showed significant expressions for the top 250 genes after adjusting P values by the Benjamini-Hochberg-false discovery rate (FDR) correction at α =0.05 and the logarithm of fold change (Table S1, See Supplementary Online Information at www.celljournal.org). In addition, the visualize quality control test results include the volcano plot (P value logarithm vs. fold change logarithm) and MD plot (fold change logarithm vs. mean logarithm of expression) are presented in Figure S1 (See Supplementary Online Information at www.celljournal.org).

Gene selection and model validation

All the genes normalized expression and 20 surrogate variables entered the logistic regression models. The tuning parameter of Lambda was set according to the minimization of cross-validation errors in each approach. This study identified 23 significant genes associated with COPD progression based on all three logistic regression models, while 18 of them were observed in all three models, and 9 of them, RNF130, PLCB1, CACNA1G, LARP4B, CALD1, TMEM182, PARD3B, PELP1, and RPIA, were positively correlated. These genes were the most significantly regulated novel genes selected based on the Z scores and may represent novel biomarkers in COPD prognosis. In this way, the LASSO model minimized the cross-validation error when we entered 21 features. The minimum values were achieved at 22, and 18 selected genes for SCAD and MCP approach (Fig.1). The Venn diagram shows all the 18 genes detected by the MCP approach overlap in all three models, and 20 overlap only between SCAD and LASSO models (Fig.2).

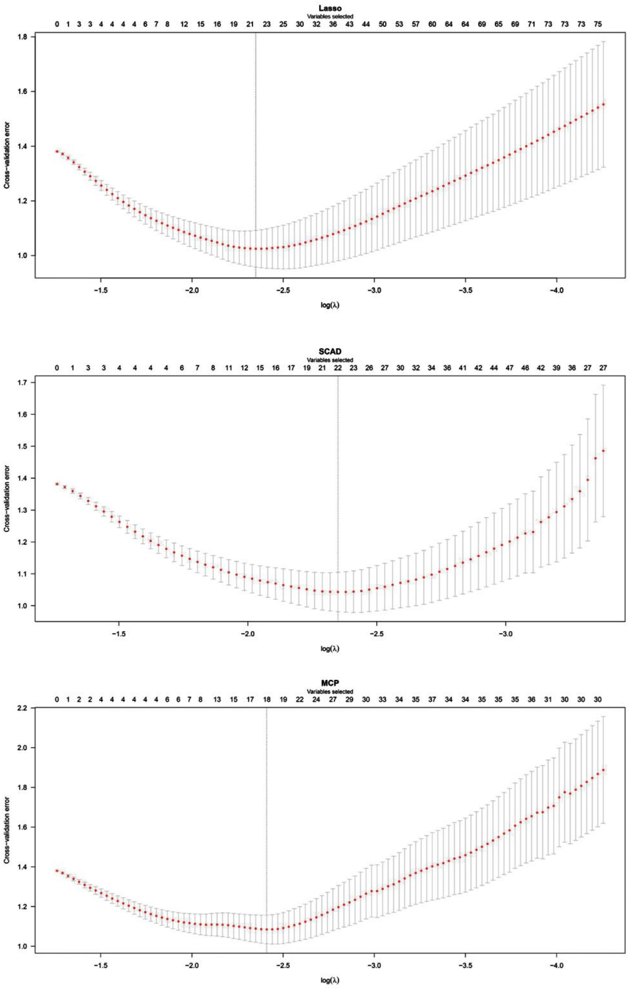

Fig.1: Cross-validation error and the selected number of features in LASSO, SCAD, and MCP for different values of the tuning parameter. LASSO; Least absolute shrinkage and selection operator, SCAD; Smoothly clipped absolute deviation, and MCP; Minimax concave penalty.

The list of the selected genes for each model is presented in Table 1. Among 21 genes in the LASSO approach, 12 of them negatively correlate with the outcome. Therefore, higher expressions of these genes were more likely for individuals in moderate stages of COPD patients. All the models confirm these negative associations. The *STX6* presents the highest significant positive association in the LASSO model, with the odds ratio (OR) equal to 1.63. This gene is placed in the second importance rank with the OR of 1.65 in comparison with *RNF130* with OR 1.69 in the SCAD approach. The SCAD pattern is repeated in the MCP approach, and the ORs for *RNF130* and *STX6* are 1.76 and 1.72, respectively.

The ROC curves of the models are presented in Figure 3. The AUC values are 95.61, 96.26, and 96.37 for LASSO, SCAD, and MCP models. In addition, the MCP, both sensitivity and specificity, are 89.86. Also, the specificity of LASSO is 85.51, and the corresponding value for SCAD is 86.96. It means that the extra genes were detected by the LASSO and SCAD approach are more relevant to detecting severe patients than moderate stage ones. All the second-order interactions between genes were explored in the final models, and none were significant. Therefore, we only report the main effects in Table 1.

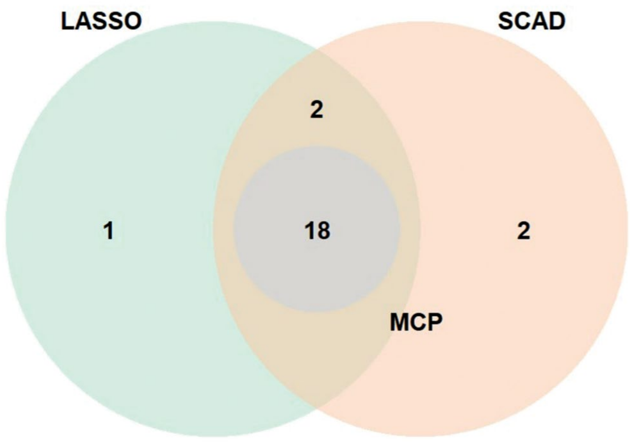

Fig.2: The Venn diagram of overlapping between 3 different regularization methods. LASSO; Least absolute shrinkage and selection operator, SCAD; Smoothly clipped absolute deviation, and MCP; Minimax concave penalty.

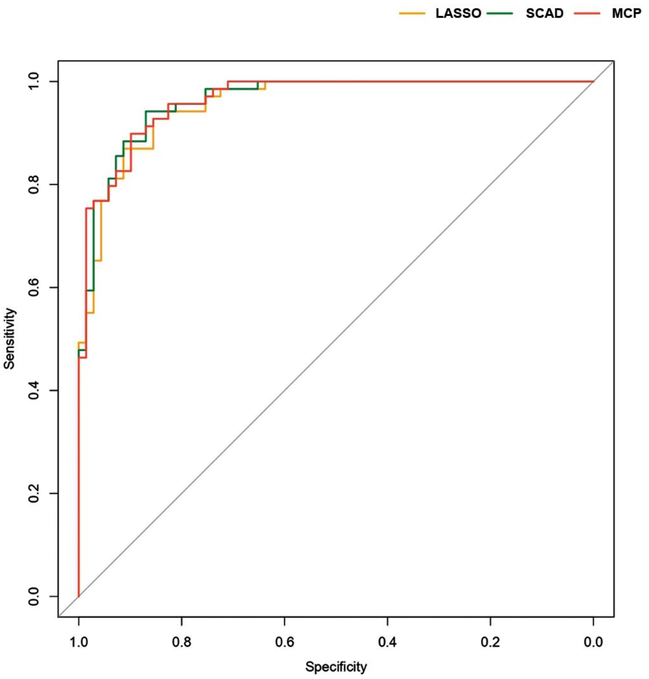

Fig.3: The ROC curves for logistic regressions using LASSO, SCAD, and MCP regularization. ROC; Receiver operator characteristic, LASSO; Least absolute shrinkage and selection operator, SCAD; Smoothly clipped absolute deviation, and MCP; Minimax concave penalty.

Table 1: Results of gene selection by LASSO, SCAD, and MCP logistic regression

Gene symbol	Gene title	LASSO			SCAD			MCP		
		Estimate	OR (95% CI)	Adj P	Estimate	OR (95% CI)	Adj P	Estimate	OR (95% CI)	Adj P
LOC100507634	Uncharacterized LOC100507634	0.16	1.17 (1.04-1.32)	5.20E-05	0.19	1.21 (1.05-1.4)	5.20E-05	0.23	1.26 (1.06-1.49)	5.20E-05
LINC00693	Long intergenic non-protein coding RNA 693	-2.41	0.09 (0.02-0.38)	0.0105	-2.68	0.07 (0.01-0.35)	0.0105	-3.13	0.04 (0.01-0.29)	0.0105
TAF15	TATA-box binding protein associated factor 15	-0.79	0.45 (0.28-0.75)	7.78E-05	-0.91	0.4 (0.22-0.72)	7.78E-05	-0.9	0.41 (0.23-0.73)	7.78E-05
CACNA1G	Calcium voltage-gated channel subunit alpha1 G	0.19	1.21 (1.05-1.39)	0.000203	0.27	1.31 (1.07-1.6)	0.0002	0.3	1.34 (1.08-1.69)	0.000203
UCP2	Uncoupling protein 2	-0.06	0.94 (0.9-0.98)	7.21E-06	-0.05	0.95 (0.92-0.99)	7.21E-06	-		-
NR2F1	Nuclear receptor subfamily 2 group F member 1	-		-	-0.02	0.98 (0.96-1)	0.00499	-		-
CALD1	Caldesmon 1	-0.05	0.95 (0.92-0.99)	2.04E-05	-0.08	0.92 (0.87-0.98)	2.04E-05	-0.01	0.99 (0.98-1)	2.04E-05
RPIA	Ribose 5-phosphate isomerase A	-1.61	0.2 (0.09-0.46)	6.05E-06	-1.73	0.18 (0.07-0.44)	6.05E-06	-2.08	0.12 (0.04-0.36)	6.05E-06
PLCB1	Phospholipase C beta 1	0.28	1.32 (1.1-1.59)	1.07E-05	0.31	1.37 (1.1-1.68)	1.07E-05	0.32	1.38 (1.11-1.71)	1.07E-05
PELP1	Proline, glutamate and leucine rich protein 1	-1.02	0.36 (0.2-0.64)	5.43E-06	-0.95	0.39 (0.22-0.67)	5.43E-06	-0.81	0.44 (0.27-0.73)	5.43E-06
SLC38A2	Solute carrier family 38-member 2	0.08	1.08 (1.02-1.15)	6.40E-06	0.11	1.12 (1.03-1.21)	6.40E-06	0.04	1.04 (1.01-1.07)	6.40E-06
ESYT2	Extended synaptotagmin 2	0.04	1.04 (1.01-1.07)	2.99E-06	-		-	-		-
STIM2	Stromal interaction molecule 2	-0.01	0.99 (0.98-1)	0.000307	-0.1	0.9 (0.84-0.98)	0.00031	-0.1	0.9 (0.84-0.98)	0.000307
EPC1	Enhancer of polycomb homolog 1	0.1	1.1 (1.03-1.18)	1.22E-05	0.14	1.15 (1.04-1.27)	1.22E-05	-		-
LAMA1	Laminin subunit alpha 1	-0.4	0.67 (0.53-0.85)	3.17E-06	-0.44	0.64 (0.49-0.84)	3.17E-06	-0.43	0.65 (0.5-0.85)	3.17E-06
RNF130	Ring finger protein 130	0.56	1.74 (1.27-2.42)	4.12E-07	0.52	1.69 (1.22-2.31)	4.12E-07	0.57	1.76 (1.25-2.5)	4.12E-07
AMOTL1	Angiomotin like 1	-1.76	0.17 (0.07-0.42)	1.96E-05	-1.67	0.19 (0.08-0.45)	1.96E-05	-1.89	0.15 (0.06-0.4)	1.96E-05
TMEM182	Transmembrane protein 182	-0.41	0.67 (0.51-0.87)	0.0226	-0.4	0.67 (0.51-0.88)	0.0226	-0.3	0.74 (0.6-0.91)	0.0226
PARD3B	Par-3 family cell polarity regulator beta	-0.69	0.5 (0.31-0.8)	0.0902	-0.99	0.37 (0.19-0.73)	0.0902	-1.1	0.33 (0.16-0.7)	0.0902
LARP4B	La ribonucleoprotein domain family member 4B	0.19	1.2 (1.06-1.39)	0.0014	0.22	1.25 (1.06-1.47)	0.0014	0.16	1.18 (1.04-1.32)	0.0014
CARD8-AS1	CARD8 antisense RNA 1	-		-	0.02	1.02 (1-1.04)	0.00046	-		-
STX6	Syntaxin 6	0.49	1.63 (1.2-2.22)	6.47E-05	0.5	1.65 (1.19-2.29)	6.47E-05	0.55	1.72 (1.21-2.48)	6.47E-05
PRDX2	Peroxiredoxin 2	-0.5	0.61 (0.45-0.81)	0	-0.53	0.59 (0.43-0.81)	0	-0.6	0.55 (0.38-0.78)	0
Sensitivities		92.75			94.2			89.86		
Specificities		85.51			86.96			89.86		
Youden index		78.26			81.16			79.71		
AUC (95% CI)		95.61 (92.68-95.61)			96.26 (93.59-96.26)			96.37 (93.79-96.37)		

LASSO; Least absolute shrinkage and selection operator, SCAD; Smoothly clipped absolute deviation, MCP; Minimax concave penalty, OR; Odds ration, CI; Confidence interval, Adj P; Adjusted P values, and AUC; Area under the curve.

Finally, The Spearman's rank correlation, co-expression matrix between the selected genes and heatmap for hierarchical clustering of the twenty-three candidate genes based on their gene expression pattern was presented in Figure 4.

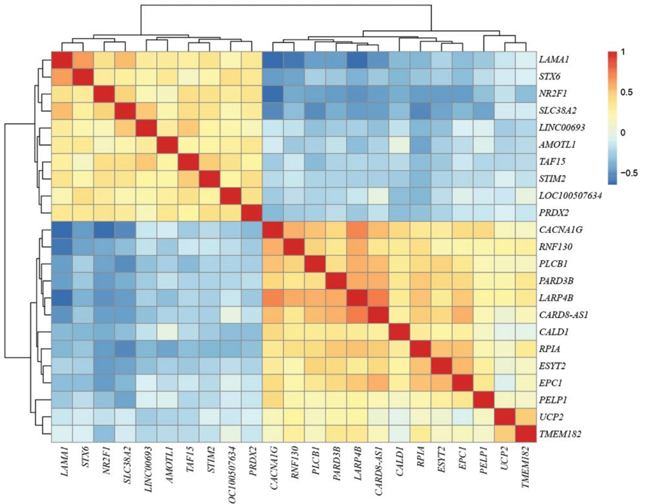

Fig.4: Spearman's rank correlation, co-expression matrix between the selected genes: heatmap of hierarchical clustering, the twenty-three candidate genes based on their gene expression pattern.

The predictive power of the model was evaluated in the test dataset (Table 2). According to our findings, the MCP model predicts 75% of cases (higher stage) and 67% of non-cases (lower stage) correctly on the test data.

Table 2: Predictive power of genes in test data

Criteria	LASSO	SCAD	MCP
Sensitivities	83.33	83.33	75.00
Specificities	44.44	44.44	66.67

LASSO; Least absolute shrinkage and selection operator, SCAD; Smoothly clipped absolute deviation, and MCP; Minimax concave penalty.

Discussion

In this study, the highest positive correlation with COPD in all three models was related to *RNF130*, an E3 ubiquitin ligase RING finger (RNF) protein with high levels of similarity to human GRAIL protein. This protein degrades *CD3ζ* in response to T-cell receptor (TCR) activation and is reported as a negative regulator of TCR signalling (19). On the other hand, it has been reported that tumor-adjacent and COPD lungs show a higher T cell density than the lungs of healthy donors, reflecting the inflammation in these patients (20). Therefore, reducing the expression level of *RNF130* can lead to more activity of T cells and more inflammation and tissue damage in the lungs. Hence, examining the *RNF130* expression level can help in COPD process identification in early step.

Moreover, the phospholipase C beta (PLCB) class of phospholipases comprises four isozymes (B1-B4) that showed an association with several inflammatory diseases and cancers. It has been shown that *PLCB1*, one of the responsible genes for *PLCB* phospholipases, is highly expressed by neuronal tissue and plays a substantial role in the stimulation of neuroendocrine growth factors that promote the progression of small cell lung carcinoma by increasing the proliferation of tumour cells (21) and significantly associated with poor overall survival (O.S.) of lung adenocarcinoma (22). It has also been shown that *PLCB1* plays a role in endothelial inflammation, inhibiting the effect of lipopolysaccharide-induced endothelial cell inflammation with varying degrees of proinflammatory cytokine expression (23). We also observed that the expression of PLCB1 positively correlated with COPD. It seems that the expression of *PLCB1* can be used as an informative early gene in the early detection of inflammatory disease and eventually lung cancer. In addition, Calcium voltage-gated channel subunit alpha1 G (CACNA1G) expressed significantly higher in the non-small cell lung cancer (NSCLC) tissues or cell lines than that in para-carcinoma normal tissues or cells and was relative to more lymph node metastasis and distant metastasis and epithelial-mesenchymal transition (EMT) (24). In addition, the RNA-binding protein la ribonucleoprotein 4B (LARP4B) has a la motif (lam) that allows it to participate in posttranscriptional control of RNA and play an important role in translation. It has been shown that LARP4B mRNA is highly expressed in liver cancer tissue and was correlated with survival status, where genes involved in the G2M checkpoint, E2F targets, and mitotic spindle were differentially enriched (25). Also, Caldesmon 1 gene (*CALD1*), as the unique gene of the MCP regression model, is a novel gene associated with both overall and disease-free survival in bladder cancer patients (26), diabetic nephropathy (27), and glioma neovascularization. Furthermore, it has been shown that the oncogenic activity of the TNF-α/miR-450a/*TMEM182* axis is primarily through activating the extracellular signal-regulated kinase 1/2 (ERK1/2) signaling pathway, which was first discovered in cancer cells, and drugs that reverse the signaling are being investigated as cancer treatments. It has been revealed that *ERK1/2* inhibitor prevented the TNF-α-induced miR-450a expression and enhanced adhesion ability which seems that TNF-α-induced ERK1/2-dependent miR-450a against *TMEM182* expression exerts a great influence on increasing oral squamous cell motility (28) which may result in metastasis of cancer cells. The par-3 family cell polarity regulator beta (*PARD3B*) is highly expressed in the kidney, lung, and skeletal muscle, and is localized at tight junctions with tight junction protein ZO-1. Recently, it has been revealed that PARD3B binds to the tumor suppressor protein Lkb1 and suppresses its kinase activity, whereas ablation of *PARD3B* causes rapid and profound stem cell loss that is vital for mammary gland stem cell maintenance (29). Therefore, it seems that increasing the expression of partitioning defective 3 homolog B (*PARD3B*) may lead to increased proliferation of lung stem cells and gradually the incidence of lung cancer. *PELP1* is an active member of this pathway, closely related to cancer metastasis reported in the lung cancer. It has been shown that *PELP1* transcript and protein levels in tumor tissues compared to adjacent pathologically unchanged tissues are significantly increased in all patients, which correlated with lung cancer stages (I/ II stages), tumor size, and lymph node metastasis (30). It has also has been reported that *PELP1* dysregulates in the non-small cell lung carcinoma, especially in lung adenocarcinoma, which significantly positively correlated with more minor differentiated features of carcinoma cells, positive lymph node metastasis, higher clinical stage as well as the status of ERα, ERβ, and *PCNA* (22). In other words, it can be said that since the *PELP1* gene represents the adverse clinical outcome of lung cancer patients, examining PELP1 expression in COPD can give a reasonable prognosis of the disease process. Ribose-5-phosphate isomerase A (RPIA) is an important integral member of the PPP and regulates cancer cell growth and tumorigenesis in pancreatic ductal adenocarcinoma (PDAC) and hepatocellular carcinoma (HCC) (31). It has been reported that in the *RPIA* is significantly up-regulated in the CRC tissue, and it is expressed at multiple stages of tumorigenesis, including early stages. It has also been shown that RPIA increases the expression of β-catenin and its target genes, and induces tumorigenesis in gut-specific promoter-carrying RPIA transgenic zebrafish in which RPIA enters the nucleus and stabilizes β-catenin activity (32). Since this process could be one of the first events in the cancer progression, hence it seems that RPIA can be used as an important marker in early detection of cancer. It is worth noting that above mention genes are not previously detected in COPD studies and because they all play a role in the inflammatory cycle and cancer, they may represent novel biomarkers in the diagnosis or prognosis of COPD.

On the other hand, in this study, we have a panel of genes that show a significant negative correlation with COPD, which includes STX6, LOC100507634, SLC38A2, STIM2, LAMA1, PRDX2, TAF15, AMOTL1, and LINC00693. In this regard, studies showed that STX6 is involved in diverse cellular functions in various cell types and has been shown to regulate many intracellular membrane trafficking events such as endocytosis, recycling, anterograde and retrograde trafficking. The oncogenic roles of STX6 in the progression of esophageal squamous cell carcinoma (ESCC) are established, and it might be a valuable target for ESCC therapy (33). Moreover, an epigenome-wide association study (EWAS) found that STX6 may be one of the genes associated with atopic asthma. They also reported that STX6 might have a role in the methylation process seen in this disease (34). The STX6 may have a role in regulating neutrophil secondary granule exocytosis and stimulating cells by Ca^{2+}. This role may influence the inflammatory process that obstructs airways in the COPD. Also, the only report for the solute carrier family 38 members 2, SLC38A2, has been reported that the mRNA expression levels of the solute carrier were higher in the tumour lung than in the healthy lung (35), the mechanism of which is unknown.

Moreover, Stromal interaction molecules, *STIM2*, as the unique gene of the LASSO regression model, regulates store-operated calcium (Ca^{2+}) entry and basal cytoplasmic Ca^{2+} levels in human cells. As is known, E2 exposure inhibits *STIM1* translocation in airway epithelia, and prevents SOCE. The E2 can signal non gnomic by inhibiting basal phosphorylation of *STIM1*, and *STIM2*, leading to a reduction in SOCE (36). Another study showed that *STIM1* and *STIM2* were significant as up-regulated genes versus healthy controls and healthy smokers (37). On the other hand, haplotype-based computational genetic analysis and gene expression profiling of lung tissue obtained from fibrosis-susceptible and -resistant mouse models identified *LAMA1* as a genetic modifier of susceptibility to pulmonary fibrosis. The *LAMA1* gene is a genetic modifier of TGF-β1 effector responses, such as macrophage activation, fibroblast proliferation, myofibroblast transformation, and the extracellular matrix production that significantly affects the development of pulmonary fibrosis (38). Therefore, it seems that the study of the *LAMA1* gene can give us a reasonable prognosis of the COPD process. Also, *AMOTL1* via the activation of LKB1/AMPK signalling and IFN-γ-induced hyperpermeability of cultured human lung microvascular endothelial cells by maintaining the levels of *AMOTL1* is related to lung function. Angiomotin Like *AMOTL1* and caldesmon, one has a role in involvement in Adhesion and Cell Motility in lung airway and alveolar and may have a role in the obstruction of the airway by a problem in the expulsion of produced mucosa and destruction of alveolar walls or spasm in small airways. The *STIM2* and *AMOTL1* were selected as the essential genes in this study to reveal these genes as a novel target in treating COPD (36).

COPD is a progressive health problem accompanied by dyspnea, cough, and sputum production. Two mechanisms cause dyspnea: i. Alveolar cell destruction and the inability of the alveolar wall to maintain its structure and decrease available respiratory gases exchange surface area, and ii. Inflammation of the airways that causes narrowing of small airways, and this can result in a problem with the passing of air in the small airways. Several molecular pathophysiology pathways induce similar clinical symptoms and signs, such as limitations in pulmonary function and caught. The studies showed that chronic inflammation and increased oxidative stress by smoking might have a role in the COPD progression. The inflammatory cells could release mediators, such as proteases and cytokines; these mediators may contribute to tissue remodeling. Chemoattractant factors and chemokines attract other inflammatory cells to pulmonary tissue, including epithelial and proinflammatory cytokines, chemokines, and other mediators (2).

It seems that two main mechanisms participate in the COPD development, and several genes may be involved in these processes. The first mechanism is oxidative stress and the response of immune cells such as neutrophils, CD4, and CD8 lymphocytes and macrophages, which have an essential role in this inflammatory process. It is reported that macrophage 5-10 times increased in the airways, lung parenchyma, BAL fluid, and sputum in patients with COPD (2). Gens such as *AMOTL1, Syntaxin 6*, and *PRX2* may have a role in inflammation or oxidative stress response. Some other genes, such as Cacna1g and smooth muscles, exist in small airways. Some genes, such as *CALD1*, may have a role in cell members maintenance and may destroy alveolar walls. Furthermore, these genes may induce another mechanism such as production of glycoproteins and amyloids that help obstruct the small airways. On the other hand, genes such as *PELP1, LAMA1,* and *RNF130* are associated with increased inflammation, metastasis, TCR down-regulation, and lung cancer, giving us a reasonable prognosis for the COPD process.

Conclusion

Differential analysis of gene expression data can reduce the number of possible genes for further exploration. Regularization solves the serious dimensionality problem in using this kind of regression. More exploration of how these genes affect the outcome and mechanism is possible more quickly in this manner. The regression-based approaches we present could apply to overcoming this issue. However, it should be considered that most of the mentioned genes do not have supporting data in experimental studies of lung inflammation and cancer, and they need to be validated in such investigations.

Acknowledgments

Research reported in this publication was supported by the Elite Researcher Grant Committee under award number [No. 982759] from the National Institutes for Medical Research Development (NIMAD), Tehran, Iran. The authors declare no conflict of interest.

Authors' Contributions

A.K., K.G.; Designed the study and drafting the manuscript. A.K., K.G., A.Sh., Sh.M.; Collected the data and performed the data analysis. S.S.; Describe genetic discussion. A.K., K.G., A.S., S.S.; Wrote the manuscript. All authors read and approved the final manuscript.

References

1. Quan Z, Yan G, Wang Z, Li Y, Zhang J, Yang T, et al. Current status and preventive strategies of chronic obstructive pulmonary disease in China: a literature review. J Thorac Dis. 2021; 13(6): 3865-3877.
2. Szalontai K, Gémes N, Furák J, Varga T, Neuperger P, Balog JÁ, et al. Chronic obstructive pulmonary disease: epidemiology, biomarkers, and paving the way to lung cancer. J Clin Med. 2021; 10(13): 2889.
3. GBD 2017 Causes of Death Collaborators. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018; 392(10159): 1736-1788.
4. Halpin DMG, Criner GJ, Papi A, Singh D, Anzueto A, Martinez FJ, et al. Global initiative for the diagnosis, management, and prevention of chronic obstructive lung disease. The 2020 GOLD science committee report on COVID-19 and chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2021; 203(1): 24-36.
5. Pillai SG, Ge D, Zhu G, Kong X, Shianna KV, Need AC, et al. A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. PLoS Genet. 2009; 5(3): e1000421.
6. Wilk JB, Chen TH, Gottlieb DJ, Walter RE, Nagle MW, Brandler BJ, et al. A genome-wide association study of pulmonary function measures in the Framingham Heart Study. PLoS Genet. 2009; 5(3): e1000429.
7. Cho MH, Castaldi PJ, Wan ES, Siedlinski M, Hersh CP, Demeo DL, et al. A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. Hum Mol Genet. 2012; 21(4): 947-957.
8. Zhao J, Cheng W, He X, Liu Y, Li J, Sun J, et al. Chronic obstructive pulmonary disease molecular subtyping and pathway deviation-based candidate gene identification. Cell J. 2018; 20(3): 326-332.
9. Huang HH, Liu XY, Liang Y. Feature selection and cancer classification via sparse logistic regression with the hybrid l1/2 +2 regularization. PLoS One. 2016; 11(5): e0149675.
10. Cui Y, Zheng CH, Yang J, Sha W. Sparse maximum margin discriminant analysis for feature extraction and gene selection on gene expression data. Comput Biol Med. 2013; 43(7): 933-941.
11. Xie H, Huang J. SCAD-penalized regression in high-dimensional partially linear models. Ann Stat. 2009; 37(2): 673-696.
12. Hardin M, Silverman EK. Chronic obstructive pulmonary disease genetics: a review of the past and a look into the future. Chronic Obstr Pulm Dis. 2014; 1(1): 33-46.
13. Bahr TM, Hughes GJ, Armstrong M, Reisdorph R, Coldren CD, Edwards MG, et al. Peripheral blood mononuclear cell gene expression in chronic obstructive pulmonary disease. Am J Respir Cell Mol Biol. 2013; 49(2): 316-323.
14. Singh D, Fox SM, Tal-Singer R, Plumb J, Bates S, Broad P, et al. Induced sputum genes associated with spirometric and radiological disease severity in COPD ex-smokers. Thorax. 2011; 66(6): 489-495.
15. Vestbo J, Anderson W, Coxson HO, Crim C, Dawber F, Edwards L, et al. Evaluation of COPD longitudinally to identify predictive surrogate end-points (ECLIPSE). Eur Respir J. 2008; 31(4): 869-873.
16. Fan J, Peng H. Nonconcave penalized likelihood with a diverging number of parameters. Ann Stat. 2004; 32(3): 928-961.
17. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. Ann Statist. 2010; 38(2): 894-942.
18. Rodríguez JD, Pérez A, Lozano JA. Sensitivity analysis of kappa-fold cross validation in prediction error estimation. IEEE Trans Pattern Anal Mach Intell. 2010; 32(3): 569-575.
19. Watanabe T, Yamashita S, Ureshino H, Kamachi K, Kurahashi Y, Fukuda-Kurahashi Y, et al. Targeting aberrant DNA hypermethylation as a driver of ATL leukemogenesis by using the new oral demethylating agent OR-2100. Blood. 2020; 136(7): 871-884.
20. Pentcheva-Hoang T, Corse E, Allison JP. Negative regulators of T-cell activation: potential targets for therapeutic intervention in cancer, autoimmune disease, and persistent infections. Immunol Rev. 2009; 229(1): 67-87.
21. Li J, Zhao X, Wang D, He W, Zhang S, Cao W, et al. Up-regulated expression of phospholipase C, β1 is associated with tumor cell proliferation and poor prognosis in hepatocellular carcinoma. Onco Targets Ther. 2016; 9: 1697-1706.
22. Zhang D, Dai J, Pan Y, Wang X, Qiao J, Sasano H, et al. Overexpression of PELP1 in lung adenocarcinoma promoted e2 induced proliferation, migration and invasion of the tumor cells and predicted a worse outcome of the patients. Pathol Oncol Res. 2021; 27: 582443.
23. Lin YJ, Chang JS, Liu X, Tsang H, Chien WK, Chen JH, et al. Genetic variants in PLCB4/PLCB1 as susceptibility loci for coronary artery aneurysm formation in Kawasaki disease in Han Chinese in Taiwan. Sci Rep. 2015; 5: 14762.
24. Yu PF, Kang AR, Jing LJ, Wang YM. Long non-coding RNA CACNA1G-AS1 promotes cell migration, invasion and epithelial-mesenchymal transition by HNRNPA2B1 in non-small cell lung cancer. Eur Rev Med Pharmacol Sci. 2018; 22(4): 993-1002.
25. Li Y, Jiao Y, Li Y, Liu Y. Expression of la ribonucleoprotein domain

family member 4B (LARP4B) in liver cancer and their clinical and prognostic significance. Dis Markers. 2019; 2019: 1569049.

26. Liu Y, Wu X, Wang G, Hu S, Zhang Y, Zhao S. CALD1, CNN1, and TAGLN identified as potential prognostic molecular markers of bladder cancer by bioinformatics analysis. Medicine (Baltimore). 2019; 98(2): e13847.
27. Wang Z, Wang Z, Zhou Z, Ren Y. Crucial genes associated with diabetic nephropathy explored by microarray analysis. BMC Nephrol. 2016; 17(1): 128.
28. Hsing EW, Shiah SG, Peng HY, Chen YW, Chuu CP, Hsiao JR, et al. TNF-α-induced miR-450a mediates TMEM182 expression to promote oral squamous cell carcinoma motility. PLoS One. 2019; 14(3): e0213463.
29. Huo Y, Macara IG. The Par3-like polarity protein Par3L is essential for mammary stem cell maintenance. Nat Cell Biol. 2014; 16(6): 529-537.
30. Słowikowski BK, Gałęcki B, Dyszkiewicz W, Jagodziński PP. Increased expression of proline-, glutamic acid- and leucine-rich protein PELP1 in non-small cell lung cancer. Biomed Pharmacother. 2015; 73: 97-101.
31. Ciou SC, Chou YT, Liu YL, Nieh YC, Lu JW, Huang SF, et al. Ribose-5-phosphate isomerase A regulates hepatocarcinogenesis via PP2A and ERK signaling. Int J Cancer. 2015; 137(1): 104-115.
32. Chou YT, Jiang JK, Yang MH, Lu JW, Lin HK, Wang HD, et al. Identification of a noncanonical function for ribose-5-phosphate isomerase A promotes colorectal cancer formation by stabilizing and activating β-catenin via a novel C-terminal domain. PLoS Biol. 2018; 16(1): e2003714.
33. Du J, Liu X, Wu Y, Zhu J, Tang Y. Essential role of STX6 in esophageal squamous cell carcinoma growth and migration. Biochem Biophys Res Commun. 2016; 472(1): 60-67.
34. Hoang TT, Sikdar S, Xu CJ, Lee MK, Cardwell J, Forno E, et al. Epigenome-wide association study of DNA methylation and adult asthma in the Agricultural Lung Health Study. Eur Respir J. 2020; 56(3): 2000217.
35. Sudo H, Tsuji AB, Sugyo A, Okada M, Kato K, Zhang MR, et al. Direct comparison of 2amino[311C]isobutyric acid and 2amino[11C] methylisobutyric acid uptake in eight lung cancer xenograft models. Int J Oncol. 2018; 53(6): 2737-2744.
36. Sheridan JT, Gilmore RC, Watson MJ, Archer CB, Tarran R. 17β-Estradiol inhibits phosphorylation of stromal interaction molecule 1 (STIM1) protein: implication for store-operated calcium entry and chronic lung diseases. J Biol Chem. 2013; 288(47): 33509-33518.
37. Deng F, Dong H, Zou M, Zhao H, Cai C, Cai S. Polarization of neutrophils from patients with asthma, chronic obstructive pulmonary disease and asthma-chronic obstructive pulmonary disease overlap syndrome. Zhonghua Yi Xue Za Zhi. 2014; 94(48): 3796-3800.
38. Lee CM, Cho SJ, Cho WK, Park JW, Lee JH, Choi AM, et al. Laminin α1 is a genetic modifier of TGF-β1-stimulated pulmonary fibrosis. JCI Insight. 2018; 3(18): e99574.