Technology in Cancer Research & Treatment
Volume 21: 1-9
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/15330338221136724
journals.sagepub.com/home/tct
SAGE

Construction and Validation of Early Warning Model of Lung Cancer Based on Machine Learning: A Retrospective Study

Siyu Ye, BA[1,*], Jiongwei Pan, BA[2,*], Zaiting Ye, BA[3], Zhuo Cao, MD[2], Xiaoping Cai, MM[2], Hao Zheng, BA[2], and Hong Ye, BA[4]

Abstract

Background: This study is a retrospective study. The purpose of this study is to construct and validate an early warning model of lung cancer through machine learning. **Methods:** The CDKN2A gene expression profile and clinical information were downloaded from The Cancer Genome Atlas (TCGA) database and divided into a tumor group and a normal group (n = 57). The top 5 somatic mutation-related genes were extracted from 567 somatic mutation data downloaded from TCGA database using random forest algorithm. Cox proportional hazard model and nomogram were constructed combining CDKN2A, 5 somatic mutation-related genes, gender, and smoking index. Patients were divided into high-risk and low-risk groups according to risk score. The predictability of the model in the prognosis of lung cancer was estimated by Kaplan–Meier survival analysis and receiver operating characteristics curve. **Results:** We constructed a prognostic model consisting of 5 somatic mutation-related genes (sphingosine 1-phosphate receptor 1 [S1PR1], dedicator of cytokinesis 7 [DOCK7], DEAD-box helicase 4 [DDX4], laminin subunit beta 3 [LAMB3], and importin 5 [IPO5]), cyclin-dependent kinase inhibitor 2A (CDKN2A), gender, and smoking indicators. The high-risk group had a lower overall survival rate compared to the low-risk group (hazard ratio = 2.14, P = 0 .0323). The area under the curve predicted for 3-year, 5-year, and 10-year survival rates are 0.609, 0.673, and 0.698, respectively. The accuracy, sensitivity, and specificity of the model for predicting the 10-year survival rate of lung cancer are 76.19%, 56.71%, and 86.23%. **Conclusion:** The lung cancer early warning model and nomogram may provide an essential reference for patients with lung cancer management in the clinic.

Keywords

lung cancer, random forest, machine learning, somatic mutation, cyclin-dependent kinase inhibitor 2A

Received: July 11, 2022; Revised: September 26, 2022; Accepted: September 30, 2022.

Introduction

Lung cancer is a malignant tumor with the highest morbidity and mortality worldwide.[1] Early detection and diagnosis will reduce the mortality of patients with lung cancer.[2] However, in fact, about 60% of patients with non-small cell lung cancer are in the advanced stage at the time of diagnosis. The diagnosis of lung cancer needs comprehensive judgment of many disciplines, including histological diagnosis, complete staging examination, and comprehensive evaluation.[3] However, the traditional diagnosis is subjective, which is prone to divergence and misdiagnosis. With the continuous development of medical big data, more and more researches apply machine learning technology to the fields of early tumor screening, risk factor analysis, and classification.[4,5] At

[1] School of Public Administration, Wenzhou Medical University, Wenzhou, Zhejiang, China
[2] Respiratory Department, The Sixth Affiliated Hospital of Wenzhou Medical University, Lishui, China
[3] Radiology Department, The Sixth Affiliated Hospital of Wenzhou Medical University, Lishui, China
[4] PE Center, The Sixth Affiliated Hospital of Wenzhou Medical University, Lishui, China

*These authors contributed equally to this work.

Corresponding Author:
Hong Ye, PE Center, The Sixth Affiliated Hospital of Wenzhou Medical University, 15# Dazhong street, Liandu District, Lishui City, Zhejiang Province, 323000, China.
Email: yehong_yhh@163.com

present, machine algorithms are effectively used in the diagnosis of lung cancer.[6,7] However, there is still a lack of early warning models for lung cancer in clinical practice, and the existing early warning models for lung cancer cannot meet the clinical needs of prognosis evaluation. Therefore, it has potential clinical value to construct an early warning model of lung cancer through machine learning.

The occurrence of cancer is due to the long-term accumulation of a large number of gene mutations in somatic cells, which provide advantages for the transformation of cancer.[8] Somatic mutation not only causes tumor occurrence, but also affects tumor development, such as tumor subtype, metastasis, drug resistance, and immune microenvironment.[9] Lung cancer is characterized by extensive genomic instability. Studies have shown that low genomic instability is related to better survival rate of patients with lung adenocarcinoma (LUAD), suggesting that it may be more practical to construct survival prediction or risk assessment model of lung cancer based on somatic mutation.[10,11]

Smoking is the main risk factor for the development of lung cancer.[12] In China, the use of tobacco has accelerated the prevalence of lung cancer, and about three-quarters of male lung cancer deaths can be attributed to smoking.[13] In addition, the epidemiology, histological types, and prognosis of lung cancer show strong gender differences.[14] There is convincing evidence that the risk, morbidity, and mortality of lung cancer in women who never smoke are higher than those in men who never smoke.[15] This suggests that smoking and gender may be important indicators affecting the diagnosis and prognosis of lung cancer.

Cyclin-dependent kinase inhibitor 2A (CDKN2A) is a tumor suppressor gene that is easily inactivated in cancer. It has been found that CDKN2A is a prognostic marker or a transcriptome marker for treatment decisions of hepatocellular carcinoma, colorectal cancer, bladder cancer, and other cancers.[16–18] In addition, studies have shown that the absence of CDKN2A indicates a poor prognosis of lung cancer and promotes the development of lung cancer.[19] Therefore, it is necessary to use CDKN2A as an indicator for the diagnosis and prognosis of lung cancer.

Therefore, this study aims to construct and validate an early warning model for lung cancer by combining somatic mutation, CDKN2A, smoking, and gender indicators through machine learning.

Materials and Methods

Data

In this study, LUAD somatic mutation data and corresponding clinical information (a total of 567 cases) were downloaded from The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/). The 567 cases were the only available data in TCGA database. LUAD transcriptome expression data also downloaded from TCGA database were subjected to extraction of data about 57 tumor samples and 57 corresponding normal tissues from 57 patients who were all cases for which the control sample was available in the 567 cases mentioned above. Since gender and smoking history were involved in the model construction, the samples with unknown gender and smoking history were excluded.

Somatic Mutation Analysis

Random forests are known for their high performance and generalizability.[20] Somatic mutation indices were screened using a random forest algorithm (R package “randomForestSRC”, with variable relative importance > 0.4). According to the outcomes

Table 1. Clinical Characteristics of Patients Involved in the Study.

Characteristics	Number
Gender	**567**
Male	239 (42.15%)
Female	276 (48.68%)
Unknown	52 (9.17%)
Age (years)	
≤65	220 (38.80%)
>65	276 (48.68%)
Unknown	71 (12.52%)
Stage	
I	277 (48.85%)
II	121 (21.34%)
III	84 (14.81%)
IV	26 (4.59%)
Unknown	59 (10.41%)
T (tumour)	
T1	171 (30.16%)
T2	276 (48.68%)
T3	46 (8.11%)
T4	19 (3.35%)
TX	3 (0.53%)
Unknown	52 (9.17%)
N (lymph node)	
N0	330 (58.2%)
N1	97 (17.11%)
N2	74 (13.05%)
N3	2 (0.35%)
NX	11 (1.94%)
Unknown	53 (9.35%)
M (metastasis)	
M0	347 (61.20%)
M1	25 (4.41%)
MX	139 (24.51%)
Unknown	56 (9.88%)
Smoking history	
1	74 (13.05%)
2	120 (21.16%)
3	134 (23.63%)
4	169 (29.81%)
5	4 (0.71%)
Unknown	14 (2.47%)
Survival status	
Death	186 (32.80%)
Alive	329 (58.02%)
Unknown	52 (9.17%)

Note: Lifelong nonsmoker (<100 cigarettes smoked in lifetime) = 1.
Current smoker (includes daily smokers and nondaily smokers or occasional smokers) = 2.
Current reformed smoker for >15 years (>15 years) = 3.
Current reformed smoker for ≤15 years (≤15 years) = 4.
Current reformed smoker, duration not specified = 5’’.
Abbreviations: Mx, metastasis cannot be measured; Nx,Cancer in nearby lymph nodes cannot be measured; Tx, Main tumor cannot be measured.

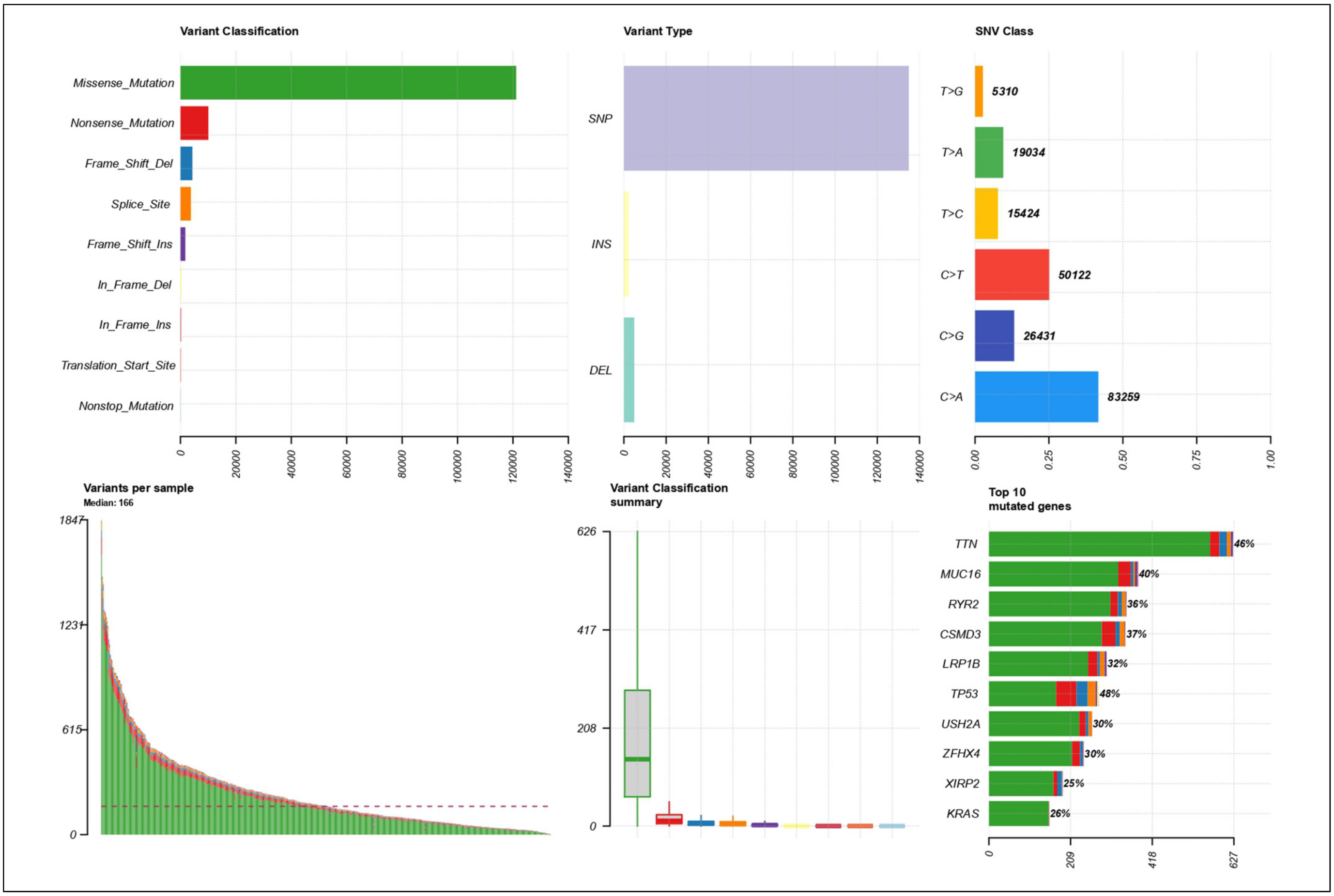

Figure 1. Characterization of somatic mutations in all samples. Abbreviations: CSMD3, CUB and Sushi multiple domains 3; DEL, deletion; Ins, insertion; LRP1B, LDL (low-density lipoprotein) receptor-related protein 1B; MUC16, mucin 16, RYR2, ryanodine receptor 2; SNV, single nucleotide variation; SNP, single nucleotide Polymorphism; TP53, tumor protein p53; TTN, titin; USH2A, usherin; XIRP2, xin actin binding repeat containing 2; ZFHX4, zinc finger homeobox 4.

(discharge or death) in the clinical information, 30 genes that were the most important for the outcomes were obtained, and the top 5 genes extracted according to the Gini index were regarded as somatic mutation indicators.

Differential Expression Analysis

R package edgeR was used to analyze the differential expression of CDKN2A in 114 samples. Log2FC > 1 was regarded as CDKN2A upregulation, log2FC < −1 was regarded as CDKN2A downregulation, and other conditions were regarded as normal. CDKN2A expression was regarded as a transcriptome expression indicator.

Construction and Validation of a Prognostic Model

The cox risk proportion model was constructed based on gender and smoking index in clinical data, together with somatic mutation index and transcriptome expression indicator. The risk score for each sample was calculated using the predict function, with the median of all sample risk scores as the cutoff value. The samples were divided into high-risk and low-risk groups for survival analysis. We obtained the risk score as follows: risk score = (0.225946*S1PR1) + (−0.136905*DOCK7) + (0.622192*DDX4) + (0.008847*LAMB3) + (0.117005*IPO5) + (0.390656*CDKN2A). Risk prediction models were presented in nomogram and their predictive performance was evaluated using a receiver operator characteristic (ROC) curve. The area under the curve (AUC) value was calculated to verify the reliability and the accuracy, sensitivity, and specificity were calculated as previously described.[21]

Statistics

This study is a retrospective study. The reporting of this study conforms to TRIPOD guidelines.[22] All statistical data were carried out in R program (version 4.0.0). Kaplan–Meier analysis was used to assess survival difference, and log-rank test was used for statistical significance. Cox proportional hazards regression was used to analyze the factors affecting the survival of patients with lung cancer. $P < 0.05$ was considered to be statistically significant.

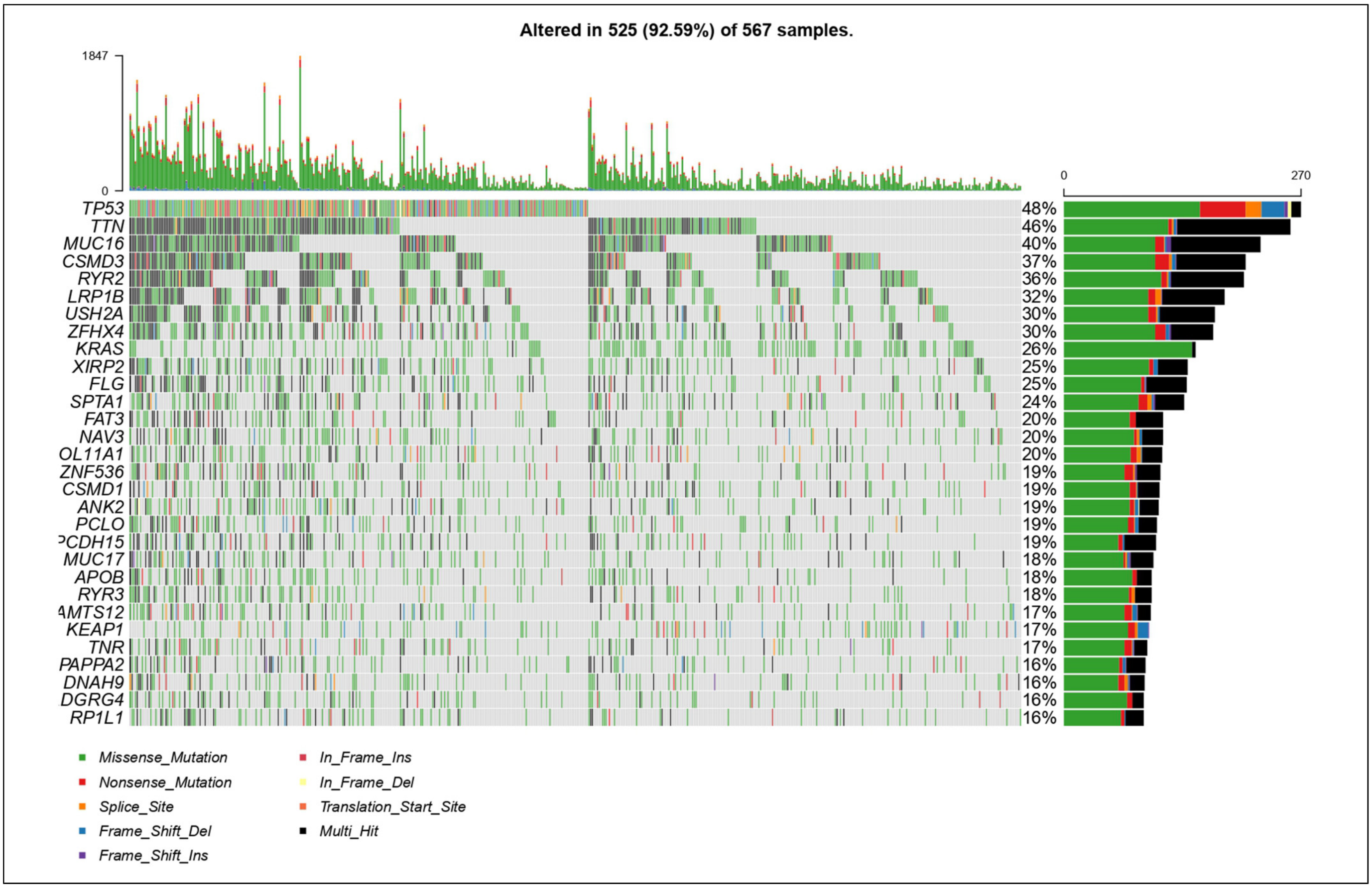

Figure 2. The top 30 genes ranked by the number of mutated samples. Abbreviations: CSMD3, CUB and Sushi multiple domains 3; LRP1B, LDL (low-density lipoprotein) receptor-related protein 1B; MUC16, mucin 16, RYR2, ryanodine receptor 2; TP53, tumor protein p53; TTN, titin; USH2A, usherin; XIRP2, xin actin binding repeat containing 2; ZFHX4, zinc finger homeobox 4; FLG, filaggrin; SPTA1, spectrin-alpha erythrocytic 1; FAT3, FAT atypical cadherin 3; NAV3, neuron navigator 3; COL11A1, collagen type XI alpha 1 chain; ZNF536, zinc finger protein 536; CSMD1, CUB and Sushi multiple domains 1; ANK2, ankyrin 2; PCLO, piccolo presynaptic cytomatrix protein; PCDH15, protocadherin related 15; ADAMTS12, ADAM metallopeptidase with thrombospondin type 1 motif 12; KEAP1, kelch like ECH associated protein 1; TNR, tenascin R; PAPPA2, pappalysin 2; DNAH9, dynein axonemal heavy chain 9; ADGRG4, adhesion G protein-coupled receptor G4; RP1L1, RP1 like 1.

Results

Characterization of Somatic Mutations in All Samples

First, we analyzed the somatic mutation in 567 samples. The clinical characteristics of 567 samples were shown in Table 1. As shown in Figure 1, missense mutation accounted for the highest proportion of all variant classification and single nucleotide polymorphism (SNP) accounted for the highest proportion of variant types. As for single nucleotide variation (SNV), the mutation from C to A accounted for the most. The average number of mutations in each sample was 166. According to the number of variants, the top 10 genes were titin (TTN), mucin 16 (MUC16), ryanodine receptor 2 (RYR2), complement C1r/C1s, Uegf, Bmp1 (CUB) and Sushi multiple domains 3 (CSMD3), LDL (low-density lipoprotein) receptor-related protein 1B (LRP1B), tumor protein p53 (TP53), usherin (USH2A), zinc finger homeobox 4 (ZFHX4), xin actin binding repeat containing 2 (XIRP2), and KRAS proto-oncogene. Then we sorted according to the number of mutated samples and showed the top 30 genes (Figure 2).

Screening of Somatic Mutation Indicators and Transcriptome Expression Indicators

Taking the survival or death status of the follow-up information in clinical data as a factor, we use random forest algorithm to screen the indicators of somatic mutation according to Gini indicators. The first 30 predictors were shown in Figure 3. We put the top 5 genes into the risk prediction model as somatic mutation indicators (sphingosine 1-phosphate receptor 1 [SIPR1], dedicator of cytokinesis 7 [DOCK7], DEAD-box helicase 4 [DDX4], laminin subunit beta 3 [LAMB3], and importin 5 [IPO5]).

CDKN2A has been widely reported as a clinical prognostic factor for lung cancer.[19,23,24] Therefore, we downloaded the transcriptome expression data of TCGA–LUAD, extracted

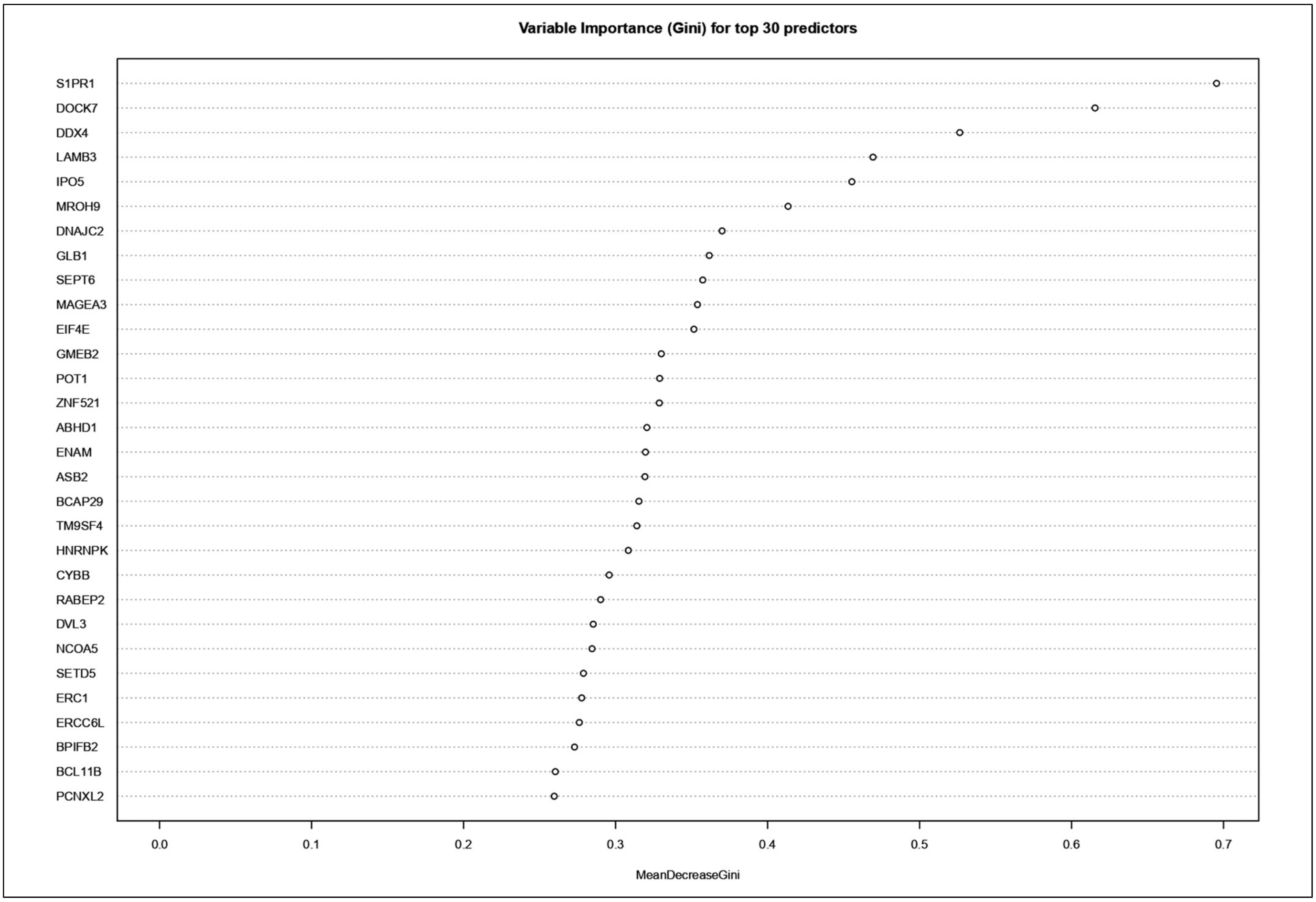

Figure 3. Random forest algorithm was used to screen the indicators of somatic mutation according to Gini indicators. Abbreviations: CSMD3, CUB and Sushi multiple domains 3; LRP1B, LDL (low-density lipoprotein) receptor-related protein 1B; MUC16, mucin 16, RYR2, ryanodine receptor 2, TP53, tumor protein p53; TTN, titin; USH2A, usherin; XIRP2, xin actin binding repeat containing 2; ZFHX4, zinc finger homeobox 4.

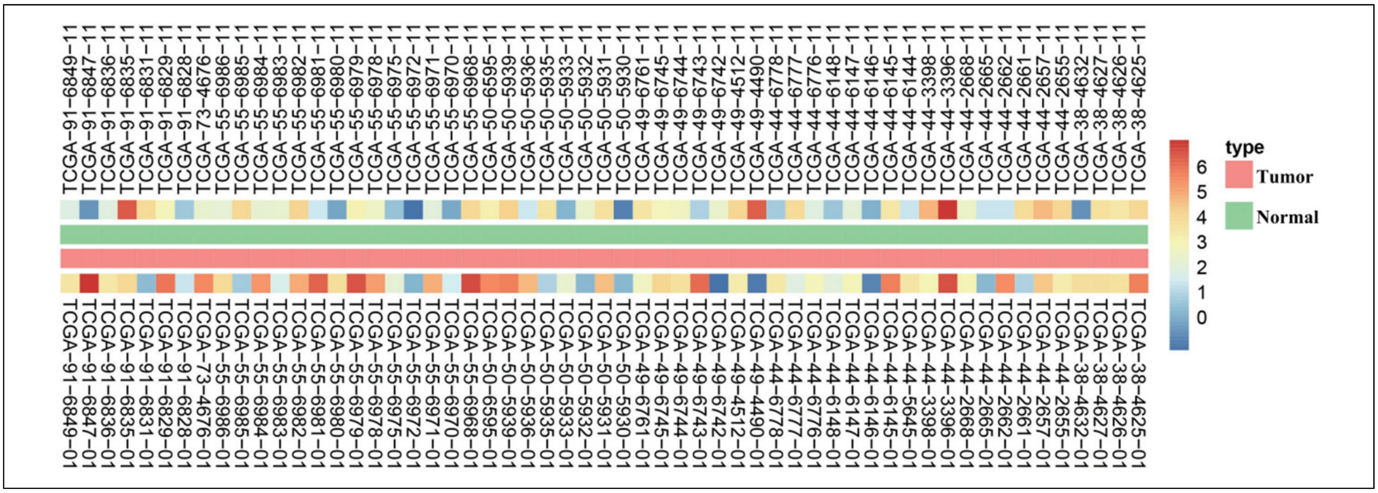

Figure 4. CDKN2A gene expression between the samples (n = 57) and the control (n = 57). Abbreviations: CDKN2A, cyclin-dependent kinase inhibitor 2A; TCGA, The Cancer Genome Atlas.

57 pairs of samples, and viewed the difference in CDKN2A gene expression between the samples and the control (Figure 4). It could be seen that CDKN2A was differentially expressed in these 57 pairs of samples, so we included CDKN2A as a transcriptome expression indicator in the risk prediction model.

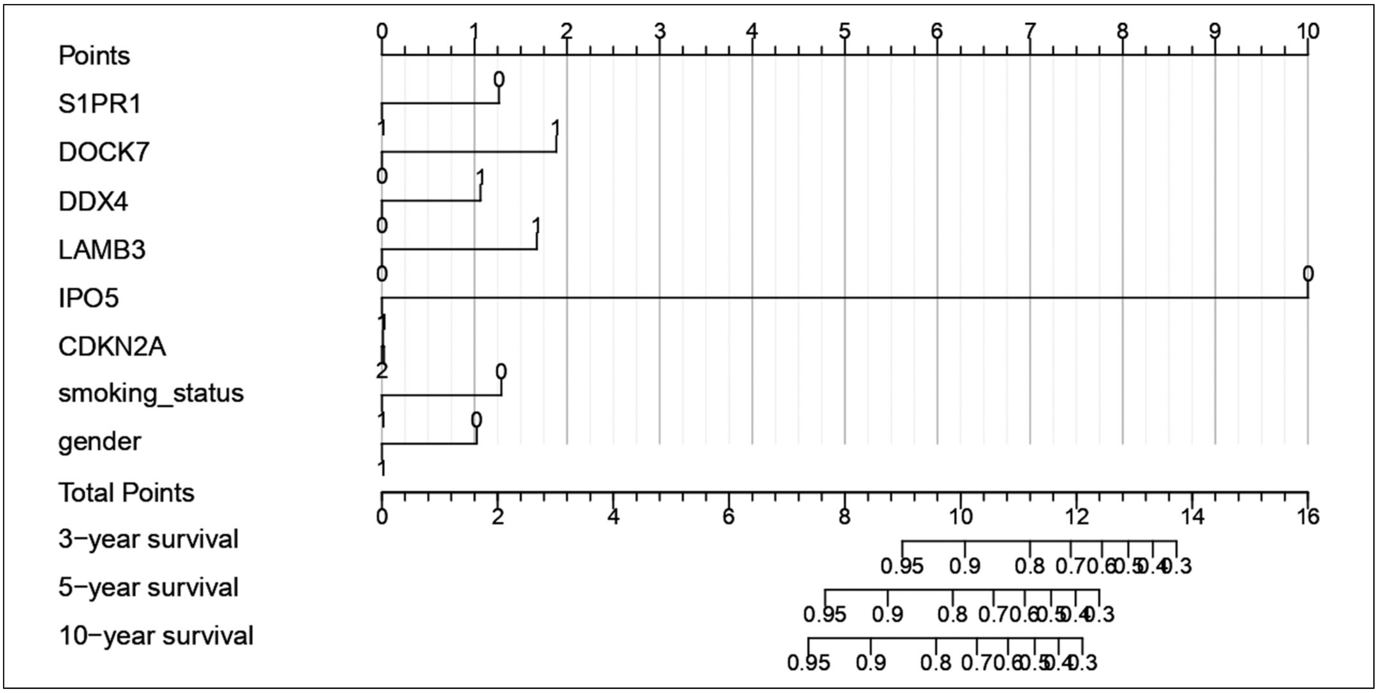

Figure 5. Nomogram for predicting 3-year, 5-year, and 10-year survival for patients with lung cancer in TCGA data set based on somatic mutation indicators, CDKN2A and clinicopathological parameters (smoking and gender). Abbreviations: CDKN2A, cyclin-dependent kinase inhibitor 2A; DDX4, DEAD-box helicase 4; DOCK7, dedicator of cytokinesis 7; IPO5, importin 5; LAMB3, laminin subunit beta 3; SIPR1, sphingosine 1-phosphate receptor 1; TCGA, The Cancer Genome Atlas.

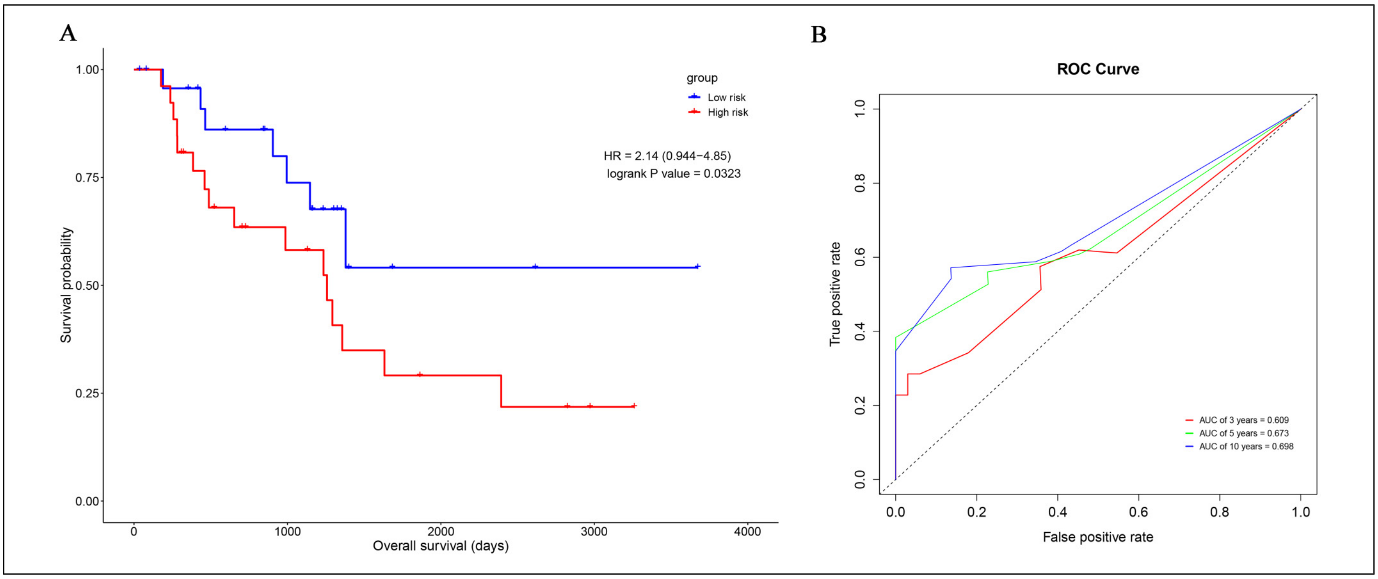

Figure 6. Kaplan–Meier survival analysis of the low-risk group and high-risk group (A). ROC curves were used to confirm the discriminative ability of nomogram (B). Abbreviations: AUC, area under the curve; ROC, receiver operator characteristic.

Construction of a Prognostic Nomogram

We developed a prognostic nomogram by combining somatic mutation index, transcriptome expression indicator, gender, and smoking (Figure 5). Kaplan–Meier survival curve proved the difference of overall survival rate between high-risk and low-risk score groups (Figure 6A, $P = 0.0323$), indicating that the model has a certain latent capacity in forecasting the prognosis of lung cancer sufferers. ROC curve confirmed the reliability of the risk model in predicting the survival rates of

3-year (AUC = 0.609), 5-year (AUC = 0.673), and 10-year (AUC = 0.698) (Figure 6B). Each AUC value was >0.6, clearly indicating that this model has good potential in calculating the prognosis of patients with lung cancer. Furthermore, the accuracy, sensitivity, and specificity of our model are 60.48%, 56.83%, and 63.47% for predicting the 3-year survival rate of lung cancer; 69.85%, 55.35%, and 74.62% for predicting the 5-year survival rate of lung cancer; and 76.19%, 56.71%, and 86.23% for predicting the 10-year survival rate of lung cancer.

Discussion

In this study, 5 somatic mutation indicators (SIPR1, DOCK7, DDX4, LAMB3, and IPO5) were included in the risk model. SIPR1 is the receptor of sphingolipid product S1P, which is highly expressed in breast cancer, gastric cancer, and hepatocellular carcinoma, and indicates the poor prognosis of patients, which is related to the regulation of drug resistance and metastasis of cancer cells by S1P/SIPR1 signal.[25–27] DOCK7 is a replication stress regulator, which mediates the replication stress response by activating Rac to promote replication protein A stability. DOCK7 is highly expressed in ovarian cancer and glioblastoma and negatively correlated with the overall survival rate of patients.[28,29] DDX4 is an IgG autoantigen, which is located in mitotic apparatus and widely expressed in somatic cell-derived cancer cell lines, and mainly plays a role in promoting tumors metastasis.[30] LAMB3-encoded LM-332 protein (extracellular matrix protein) participates in important biological behaviors such as cell differentiation, adhesion, and survival, and is related to the metastatic ability of many types of cancers (colorectal, pancreatic, thyroid, and lung cancers).[31–33] IPO5 is a nuclear transporter, which can lead to abnormal localization of oncogenes and tumor suppressor genes, thus causing drug resistance and abnormal proliferation of cancer cells.[34] Pauline J van der Watt et al[35] incorporated IPO5 into the diagnostic markers of cervical and esophageal cancers, and obtained high sensitivity and specificity. In conclusion, all 5 indicators are related to drug resistance, metastasis or proliferation of cancer, and some indicators have been considered as markers for early diagnosis or prognosis of cancer. However, the prediction efficiency of single gene is obviously not as good as that of multigene and multi-index models. Jianbo Pan et al[36] included DDX4 in the biomarker group of early diagnosis of lung cancer, with a sensitivity of 73.5% and a specificity of more than 85%. Similarly, DDX4 was also used as one feature gene in our model. However, as the model of Jianbo Pan et al is a diagnostic model while our model is a prognostic model, we could not make a direct comparison. Han-Jun Cho et al[37] identified 6 mutant genes related to the prognosis of LUAD through machine learning, but the ROC curve and AUC were unknown. The study of Han-Jun Cho et al[37] also pointed out the relationship between gene mutation and the prognosis of LUAD, which indicates that the possibility of applying somatic mutation genes to clinically guide the prognosis of LUAD.

Lung cancer is the result of individual or combined action of a variety of risk factors. To comprehensively consider a variety of risk factors can more effectively screen out the high-risk population of lung cancer. In this study, we included smoking and gender as 2 major risk factors, which are also important risk factors of lung cancer that are often considered clinically. Lung cancer risk prediction models have been developed that incorporate both gender and smoking indicators with convincing discrimination.[38,39] In addition, the combined use of multi-angle, multifactor tumor molecular markers for the detection of lung cancer is more accurate than a single test. Therefore, we also included the indicator CDKN2A. CDKN2A was silenced in more than 70% of lung squamous cell carcinoma samples.[40] Chunkang et al[21] constructed a lung cancer diagnosis model with 6 genes including CDKN2A (p.16), which can effectively diagnose early lung cancer and indicate cancer risk. In addition, Wei Liu et al[19] reported that CDKN2A indicates a poor prognosis of lung cancer, which was consistent with our findings. In this study, 5 somatic mutation indicators and CDKN2A were combined with smoking and gender to construct a lung cancer early warning model, and the model may provide certain help for clinical lung cancer early warning by predicting the survival rate of patients with lung cancer.

However, this study had some limitations. Due to the limitation of objective conditions, this study only collected 2 groups of samples of lung cancer and healthy controls for the construction of early warning model. In view of the complexity of clinical tumor diagnosis, it is necessary to collect benign lung diseases and other tumor cases in the future to improve the specificity and accuracy of discrimination. In addition, this study only performed internal validation, and future external validation through large-sample, multicenter, prospective studies is required.

Conclusion

The novelty of this study is established as an early warning model for lung cancer by machine learning based on clinical characteristics of smoking and gender in the samples, the somatic mutation gene and CDKN2A gene. The predictive effect of this early warning model for patients with lung cancer may be suitable for clinical practice, which may provide targeted guidance for the early prediction of patients with lung cancer.

Availability of Data and Materials

The analyzed data sets generated during the study are available from the corresponding author on reasonable request.

Data Availability Statement

The data sets presented in this study can be found in online repositories. https://portal.gdc.cancer.gov/, TCGA-LUAD.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Zhejiang Medical and Health Research Project (grant number: 2021KY1235), Lishui Science and Technology Bureau Project (grant numbers: 2020077571 and 2022ZDYF11).

ORCID iD

Hong Ye https://orcid.org/0000-0002-9983-8342

References

1. Thai AA, Solomon BJ, Sequist LV, et al. Lung cancer. *Lancet*. 2021;398(10299):535-554. doi: 10.1016/s0140-6736(21)00312-3
2. Freitas C, Sousa C, Machado F, et al. The role of liquid biopsy in early diagnosis of lung cancer. *Front Oncol.* 2021;11:634316. doi: 10.3389/fonc.2021.634316
3. Nooreldeen R, Bach H. Current and future development in lung cancer diagnosis. *Int J Mol Sci.* 2021;22(16):8661. doi: 10.3390/ijms22168661
4. Lynch CM, Abdollahi B, Fuqua JD, et al. Prediction of lung cancer patient survival via supervised machine learning classification techniques. *Int J Med Inf.* 2017;108:1-8. doi: 10.1016/j.ijmedinf.2017.09.013
5. Cammarota G, Ianiro G, Ahern A, et al. Gut microbiome, big data and machine learning to promote precision medicine for cancer. *Nat Rev Gastroenterol Hepatol*. 2020;17(10):635-648. doi: 10.1038/s41575-020-0327-3
6. Hsu C-H, Chen X, Lin W, et al. Effective multiple cancer disease diagnosis frameworks for improved healthcare using machine learning. *Measurement*. 2021;175:109145. DOI: 10.1016/j.measurement.2021.109145
7. Binson VA, Subramoniam M. Artificial intelligence based breath analysis system for the diagnosis of lung cancer. *J Phys Conf Ser.* 2021;1950:012065. doi: 10.1088/1742-6596/1950/1/012065
8. Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. *Science*. 2015;349(6255):1483-1489. doi: 10.1126/science.aab4082
9. Negrini S, Gorgoulis VG, Halazonetis TD. Genomic instability—an evolving hallmark of cancer. *Nat Rev Mol Cell Biol.* 2010;11(3):220-228. doi: 10.1038/nrm2858
10. Chen Y, Tang WF, Lin H, et al. Wait-and-see treatment strategy could be considered for lung adenocarcinoma with special pleural dissemination lesions, and low genomic instability correlates with better survival. *Ann Surg Oncol.* 2020;27(10):3808-3818. doi: 10.1245/s10434-020-08400-1
11. Geng W, Lv Z, Fan J, et al. Identification of the prognostic significance of somatic mutation-derived LncRNA signatures of genomic instability in lung adenocarcinoma. *Front Cell Dev Biol.* 2021;9:657667. doi: 10.3389/fcell.2021.657667
12. Bade BC, Dela Cruz CS. Lung cancer 2020: epidemiology, etiology, and prevention. *Clin Chest Med.* 2020;41(1):1-24. doi: 10.1016/j.ccm.2019.10.001
13. Cao M, Chen W. Epidemiology of lung cancer in China. *Thorac Cancer.* 2019;10(1):3-7. doi: 10.1111/1759-7714.12916
14. Mederos N, Friedlaender A, Peters S, et al. Gender-specific aspects of epidemiology, molecular genetics and outcome: lung cancer. *ESMO Open*. 2020;5(Suppl 4):e000796. doi: 10.1136/esmoopen-2020-000796
15. Thun MJ, Hannan LM, Adams-Campbell LL, et al. Lung cancer occurrence in never-smokers: an analysis of 13 cohorts and 22 cancer registry studies. *PLoS Med.* 2008;5(9):e185. doi: 10.1371/journal.pmed.0050185
16. Luo JP, Wang J, Huang JH. CDKN2A is a prognostic biomarker and correlated with immune infiltrates in hepatocellular carcinoma. *Biosci Rep.* 2021;41(10):BSR20211103. doi: 10.1042/bsr20211103
17. Rasmussen SL, Krarup HB, Sunesen KG, et al. Hypermethylated DNA as a biomarker for colorectal cancer: a systematic review. *Colorectal Dis*. 2016;18(6):549-561. doi: 10.1111/codi.13336
18. Worst TS, Weis CA, Stöhr R, et al. CDKN2A as transcriptomic marker for muscle-invasive bladder cancer risk stratification and therapy decision-making. *Sci Rep.* 2018;8(1):14383. doi: 10.1038/s41598-018-32569-x
19. Liu W, Zhuang C, Huang T, et al. Loss of CDKN2A at chromosome 9 has a poor clinical prognosis and promotes lung cancer progression. *Mol Genet Genomic Med.* 2020;8(12):e1521. doi: 10.1002/mgg3.1521
20. Bashir UA-O, Kawa B, Siddique M, et al. Non-invasive classification of non-small cell lung cancer: a comparison between random forest models utilising radiomic and semantic features. *Br J Radiol*. 2019;92(1099):20190159. Doi: 10.1259/bjr.20190159
21. Kang C, Wang D, Zhang X, et al. Construction and validation of a lung cancer diagnostic model based on 6-gene methylation frequency in blood, clinical features, and serum tumor markers. *Comput Math Methods Med.* 2021;2021:9987067. doi: 10.1155/2021/9987067
22. Moons KG, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* 2015;162(1):W1-W73. doi: 10.7326/m14-0698
23. Gutiontov SI, Turchan WT, Spurr LF, et al. CDKN2A loss-of-function predicts immunotherapy resistance in non-small cell lung cancer. *Sci Rep.* 2021;11(1):20059. doi: 10.1038/s41598-021-99524-1
24. Győrffy B, Surowiak P, Budczies J, et al. Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. *PLoS ONE*. 2013;8(12):e82241. doi: 10.1371/journal.pone.0082241
25. Pyne NJ, McNaughton M, Boomkamp S, et al. Role of sphingosine 1-phosphate receptors, sphingosine kinases and sphingosine in cancer and inflammation. *Adv Biol Regul.* 2016;60:151-159. doi: 10.1016/j.jbior.2015.09.001
26. Yeon M, Kim Y, Pathak D, et al. The CAGE-MiR-181b-5p-S1PR1 axis regulates anticancer drug resistance and autophagy in gastric cancer cells. *Front Cell Dev Biol.* 2021;9:666387. doi: 10.3389/fcell.2021.666387

27. Zhang SL, Liu L. MicroRNA-148a inhibits hepatocellular carcinoma cell invasion by targeting sphingosine-1-phosphate receptor 1. *Exp Ther Med.* 2015;9(2):579-584. doi: 10.3892/etm.2014.2137
28. Gao M, Guo G, Huang J, et al. DOCK7 protects against replication stress by promoting RPA stability on chromatin. *Nucleic Acids Res.* 2021;49(6):3322-3337. doi: 10.1093/nar/gkab134
29. Murray DW, Didier S, Chan A, et al. Guanine nucleotide exchange factor Dock7 mediates HGF-induced glioblastoma cell invasion via Rac activation. *Br J Cancer.* 2014;110(5):1307-1315. doi: 10.1038/bjc.2014.39
30. Schudrowitz N, Takagi S, Wessel GM, et al. Germline factor DDX4 functions in blood-derived cancer cell phenotypes. *Cancer Sci.* 2017;108(8):1612-1619. doi: 10.1111/cas.13299
31. Zhang H, Pan YZ, Cheung M, et al. LAMB3 mediates apoptotic, proliferative, invasive, and metastatic behaviors in pancreatic cancer by regulating the PI3K/Akt signaling pathway. *Cell Death Dis.* 2019;10(3):230. doi: 10.1038/s41419-019-1320-z
32. Zhu Z, Song J, Guo Y, et al. LAMB3 promotes tumour progression through the AKT-FOXO3/4 axis and is transcriptionally regulated by the BRD2/acetylated ELK4 complex in colorectal cancer. *Oncogene*. 2020;39(24):4666-4680. doi: 10.1038/s41388-020-1321-5
33. Jung SN, Lim HS, Liu L, et al. LAMB3 mediates metastatic tumor behavior in papillary thyroid cancer by regulating c-MET/Akt signals. *Sci Rep.* 2018;8(1):2718. doi: 10.1038/s41598-018-21216-0
34. Zhang W, Lu Y, Li X, et al. IPO5 promotes the proliferation and tumourigenicity of colorectal cancer cells by mediating RASAL2 nuclear transportation. *J Exp Clin Cancer Res*. 2019;38(1):296. doi: 10.1186/s13046-019-1290-0
35. van der Watt PJ, Okpara MO, Wishart A, et al. Nuclear transport proteins are secreted by cancer cells and identified as potential novel cancer biomarkers. *Int J Cancer.* 2022;150(2):347-361. doi: 10.1002/ijc.33832
36. Pan J, Yu L, Wu Q, et al. Integration of IgA and IgG autoantigens improves performance of biomarker panels for early diagnosis of lung cancer. *Mol Cell Proteomics*. 2020;19(3):490-500. doi: 10.1074/mcp.RA119.001905
37. Cho HJ, Lee S, Ji YG, et al. Association of specific gene mutations derived from machine learning with survival in lung adenocarcinoma. *PLoS ONE. 2018;13(11):e0207204.*
38. Muller DC, Johansson M, Brennan P. Lung cancer risk prediction model incorporating lung function: development and validation in the UK Biobank Prospective Cohort Study. *J Clin Oncol*. 2017;35(8):861-869. doi: 10.1200/jco.2016.69.2467
39. Lyu Z, Li N, Chen S, et al. Risk prediction model for lung cancer incorporating metabolic markers: development and internal validation in a Chinese population. *Cancer Med.* 2020;9(11):3983-3994. doi: 10.1002/cam4.3025
40. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 2012;489(7417):519-525. doi: 10.1038/nature11404