Medicine®

OPEN

Exploring the survival prognosis of lung adenocarcinoma based on the cancer genome atlas database using artificial neural network

Na Jiang, MD, Xianrong Xu, MD*

Abstract

The aim of this study was to investigate the clinical factors affecting the survival prognosis of lung adenocarcinoma, and to establish a predictive model of survival prognosis of lung adenocarcinoma by artificial neural network.

Download the cancer genome atlas (TCGA) database for lung adenocarcinoma research data, perform cox regression analysis and descriptive statistics on the obtained clinical data, draw the survival curve by Kaplan–Meier method, select the independent variables that are statistically significant for constructing the artificial neural networks (ANN) model, and establish artificial neural network model.

The number of valid cases included in the study was 524, including 280 men and 244 women, with an age range of 33 to 88 years, mean age 66.87 years, and median progression-free survival (PFS) was 37.7 months. The median overall survival time (OS) was 41.1 months. Cox multivariate analysis showed that smoking history, tumor stage, and surgical margin resection status were independently associated with PFS, and tumor stage and surgical margin resection status were independently associated with OS. The accuracy of the established ANN model itself was predicted to be 65.8%. The accuracy of correctly predicting the prognosis of the predicted samples was 75.0%, and the area under the receiver operating characteristic curve was 0.712.

The clinical prognostic factors of lung adenocarcinoma include: smoking history, tumor stage, and surgical margin resection status. The established ANN model can be used to predict the prognosis of lung adenocarcinoma.

Abbreviations: ANN = artificial neural networks, GDC = Genomic Data Commons, LUAD = lung adenocarcinoma, NSCLC = non-small cell lung cancer, OS = overall survival, PFS = progression-free survival, TCGA = the cancer genome atlas.

Keywords: adenocarcinoma, artificial neural network, lung neoplasms, prognostic factors, the cancer genome atlas

1. Introduction

Lung cancer is the most common cause of cancer-related mortality worldwide, causing >1 million deaths each year.[1] Lung adenocarcinoma (LUAD) is the most common histological type of lung cancer, and its effect on recurrence and metastasis is still unsatisfactory.[2] Therefore, it is important to improve the rate of early diagnosis and individualized treatment options, which forces clinicians and researchers to continuously study the prognosis of their survival in order to guide the clinical.

The cancer genome atlas (TCGA) database is a joint project of the National Cancer Institute and the National Human Genome Research Institute, allowing researchers to search and download tumor-related data for analyze. The Genomic Data Commons Data Portal (GDC) is the TCGA database data-driven platform. As of November 2018, the GDC platform has released 43 projects, including 33,096 samples.

Artificial neural networks (ANN) is an artificial intelligence model with a highly nonlinear super-large-scale continuous-time dynamic system, a network formed by a large number of processing units (neurons) interconnected. In the field of oncology, ANN is increasingly used in cancers such as gastric cancer, prostate cancer, colon cancer, and breast cancer for diagnosis, differential diagnosis, prognosis, etc.,[3–6] and has been continuously recognized by relevant medical professionals.

In this study, the data of lung adenocarcinoma are deeply explored through the TCGA database, and the clinical data are statistically analyzed to explore the prognostic factors, and an artificial neural network prediction model is established.

2. Materials and methods

2.1. Source of information

Data were collected from the TCGA database (https://portal.gdc.cancer.gov/), and accessed its external link GDC platform (http://www.cbioportal.org/) to obtain data on lung adenocarcinoma (TCGA, provisional) published on the GDC platform. A total of 586 lung adenocarcinoma samples were included in this study. In the total of all samples, 62 cases of clinical information deficiency were excluded and 524 valid data were included.

In this study, a total of 524 patients with pathological diagnosis of lung adenocarcinoma were included, including 280 men and

Editor: Jianxun Ding.

This study was supported by the Medical Health Science and Technology Project of Zhejiang Provincial Health Commission (grant no: 2016KYB053).

The authors report no conflicts of interest.

Department of Respiration, Tongde Hospital of Zhejiang Province, Hangzhou, Zhejiang Province, China.

* Correspondence: Xianrong Xu, Department of Respiration, Tongde Hospital of Zhejiang Province, Hangzhou, 310000, Zhejiang Province, China (e-mail: xxr_tongde@163.com).

Medicine (2019) 98:20(e15642)

Received: 27 December 2018 / Received in final form: 28 March 2019 / Accepted: 19 April 2019

http://dx.doi.org/10.1097/MD.0000000000015642

244 women. Among these patients, the age fluctuated between 33 and 88 years, with an average age of 66.87 years. Three hundred forty six cases of smoking history can be traced, and each tumor stage (IA-IV) can be seen. The lungs of each lung can be seen in the tumor site. After 88.80% of patients had surgery, the pathological grade of the margin was R0, and 43.89% of the cases had different types of gene mutations, including Kras and ALK mutations (Table 1).

The TCGA database is a joint project in which the receipt collection phase of the database comes from the National Cancer Institute and the National Human Genome Research Institute, and it is ethical. All the data involved in our research are from the public platform TCGA database, and does not involve ethical issues and patient consent. It is not necessary.

2.2. Statistical methods

The clinical data obtained by SPSS22.0 software (Chicago, Illinois) were processed and statistically analyzed, including log-rank test and cox regression analysis and descriptive statistics. It was considered that $P < .05$ had significant statistical difference.

Table 1

Clinical data statistics for case samples.

Clinical information	Statistical results
Total number of patients	524
Gender	
Male	244 (46.56%)
Female	280 (53.44%)
Age	
Average age (STD), y	66.87 (9.587)
Age range, y	33–88
History of smoking	
Yes	346 (66.03%)
No	178 (33.97%)
Tumor staging	
Stage IA	139 (26.94%)
Stage IB	142 (27.52%)
Stage IIA	51 (9.88%)
Stage IIB	73 (14.15%)
Stage IIIA	74 (14.34%)
Stage IIIB	11 (2.13%)
Stage IV	26 (5.03%)
Primary tumor site	
L-Upper	125 (24.56%)
L-Lower	79 (15.52%)
R-Upper	185 (36.35%)
R-Middle	21 (4.13%)
R-Lower	99 (19.45%)
Surgical margin resection status	
R0	349 (88.80%)
R1	13 (3.31%)
R2	5 (1.27%)
RX	26 (6.62%)
Genic mutation	
Yes	230 (43.89%)
No	294 (56.11%)
Patient's vital status	
Alive	336 (64.12%)
Dead	188 (35.88%)
Median progression-free survival, month (±SE)	37.71 (±0.52799)
Median overall survival, month (±SE)	41.10 (±0.41363)

Note: Because there are values that are not available, the numbers in the table are not equal to the total. Data display form: frequency (frequency) (n [%]), SE=standard error, STD=standard deviation.

Through single factor and multi-factor analysis, the independent variables which are statistically significant for constructing the ANN model are selected, and the artificial neural network model is established. The prediction performance of the ANN model is evaluated by the area under the receiver operating characteristic (ROC) curve (AUC). When AUC >0.5, the more tending to 1, the better the prediction performance.

3. Results

Univariate and multivariate analyses were performed using COX regression to determine the relationship between the selected clinical variables and progression-free survival (PFS) or overall survival (OS), with bold emphasis on statistically significant differences (Tables 2 and 3).

In the univariate analysis, we can conclude that tumor stage is associated with PFS and OS in patients with lung adenocarcinoma ($P < .001$).

In the multivariate analysis, $P = .002 < .05$, rejecting the null hypothesis, and considering that the partial regression coefficient is not 0, it is worth further analysis. At the significant level of 0.05, smoking history, tumor stage, and surgical margin level have statistics. Learning differences, tips: smoking history, tumor stage, and surgical margin resection status were independently associated with PFS ($P = .026$; $P = .009$; .001). Tumor staging and surgical margin resection status were independently associated with OS ($P < .001$; $P = .003$).

Through the cox regression univariate and multivariate analysis, the variables with statistical significance for the survival prognosis of lung adenocarcinoma were screened, and the log-rank test was performed by Kaplan–Meier method (Figs. 1 and 2). Smoking history is related to PFS, and different tumor stages have significant differences in survival prognosis, the higher the stage, the shorter the PFS and OS. Different surgical margins also have significant differences in PFS and OS. Therefore, it can be concluded that the smoking history is related to PFS, tumor stage, and different surgical margin resection status are related to PFS and OS. Through the above analysis, the independent variables which are statistically significant for constructing the ANN model are selected and an artificial neural network model is established. The topology includes input layer, hidden layer, and output layer. The data analysis selects "neural network" → "multilayer perceptron." The input node has 13 neural nodes and 8 hidden layer neural nodes. Transfer function, 2 output neural nodes, corresponding to the survival state: survival=1, death=0, transmitted by softmax function, the established ANN model is as follows (Fig. 3).

The ANN model was used to predict the training set samples, and the accuracy of correctly predicting the prognosis was 65.8%. The ANN model was used to predict the prediction set samples. The accuracy of correct prognosis was 75.0%, and the area under the ROC curve was 0.712 (Fig. 4).

To evaluate the predictive power of the ANN model, we used SPSS software to reconstruct the traditional binary logistic regression model and compare the 2 models. In this model, "smoking history, tumor staging, and surgical margin resection status" were used as independent variables, and "patient's vital status" was used as a dependent variable to construct a binary logistic regression model. The model was used to predict the 524 cases included. And using the joint probability to draw the ROC curve, the results show that the accuracy of correctly predicting the prognosis is 60.6%, and the area under the ROC curve is 0.662 (Fig. 5).

Table 2

Univariate analysis of overall survival and progression-free survival of case samples.

	Progression free survival		Overall survival	
	HR (95% CI)	***P***	**HR (95% CI)**	***P***
Gender	0.935 (0.638–1.371)	.732	1.055 (0.790–1.410)	.715
Age	1.193 (0.812–1.752)	.386	1.264 (0.943–1.694)	.117
History of smoking	0.992 (0.668–1.473)	.968	0.962 (0.713–1.297)	.799
Tumor staging	1.312 (1.186–1.452)	**1.465E–7**	1.313 (1.218–1.417)	**1.6022E–12**
Primary tumor site	1.032 (0.897–1.186)	.663	1.037 (0.932–1.153)	.505
Surgical margin resection status	1.124 (0.870–1.453)	.372	1.194 (0.971–1.468)	.093
Genic mutation	1.291 (0.883–1.886)	.187	1.044 (0.781–1.396)	.772

CI = confidence interval, HR = hazard ratio.

Bold fonts emphasize statistically significant independent associations (*P* < .05).

Table 3

Multivariate analysis of overall survival and progression-free survival of case samples.

	Progression free survival		Overall survival	
	HR (95% CI)	***P***	**HR (95% CI)**	***P***
Gender	1.021 (0.643–1.622)	.930	1.131 (0.796–1.606)	.493
Age	0.947 (0.592–1.514)	.819	1.121 (0.786–1.597)	.528
History of smoking	0.566 (0.342–0.936)	.026	0.713 (0.491–1.035)	.075
Tumor staging	0.386 (0.117–1.276)	**.009**	0.282 (0.129–0.616)	**<.001**[*]
Primary tumor site	0.822 (0.415–1.629)	.521	0.768 (0.464–1.271)	.385
Surgical margin resection status	1.188 (0.455–3.104)	**.001**	1.174 (0.530–2.598)	**.003**
Genic mutation	0.845 (0.531–1.345)	.478	0.925 (0.649–1.320)	.668

[*] *P* = .000006 < .001.

CI = confidence interval, HR = hazard ratio.

Bold fonts emphasize statistically significant independent associations (*P* < .05).

4. Discussion

LUAD is the most common subtype of non-small cell lung cancer (NSCLC), accounting for about 40% of all histological types, and is prone to recurrence. For recurrence and metastatic LUAD, the current treatment effect is poor.[7] Due to the existence of tumor heterogeneity, the choice of individualized programs and the evaluation of prognosis have important clinical significance. Through the study of the clinical features of LUAD, the

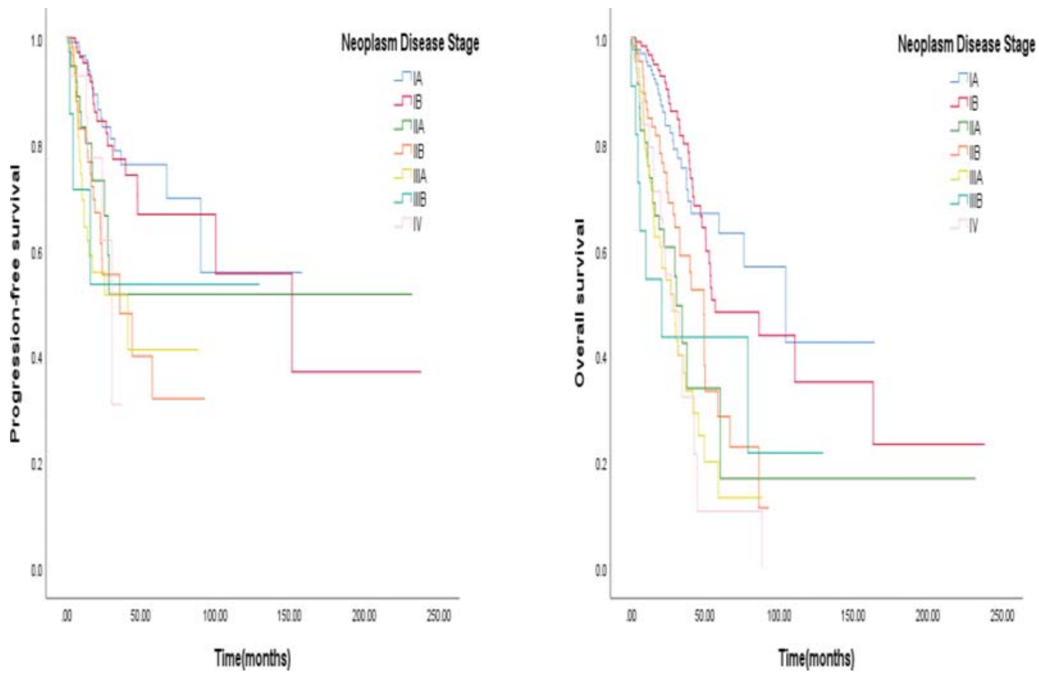

Figure 1. Relationship between tumor stage and PFS or OS. OS = overall survival, PFS = progression-free survival.

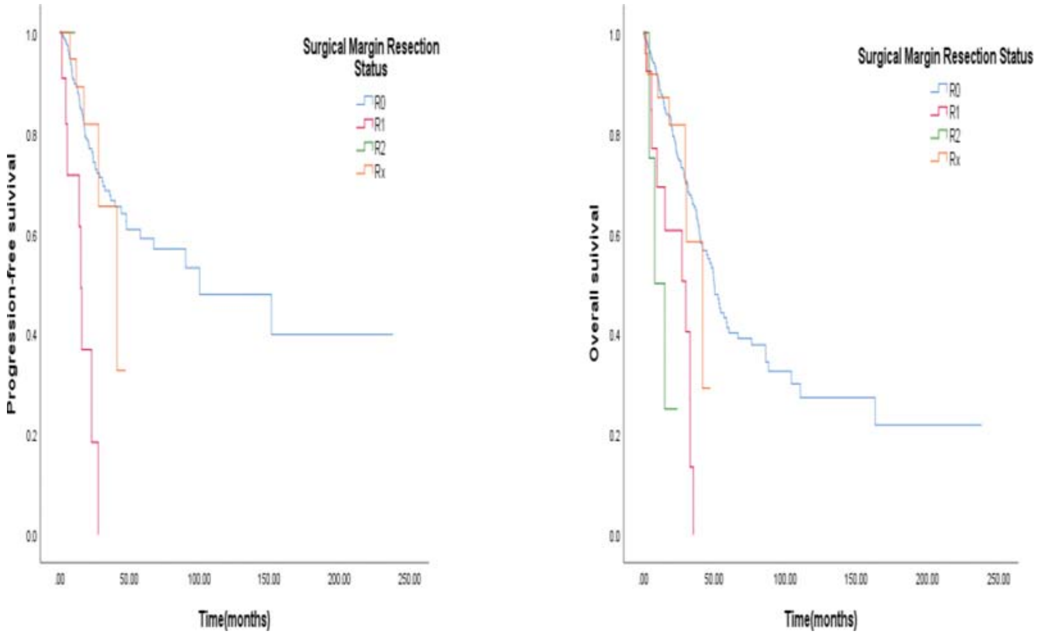

Figure 2. Relationship between surgical margin level and PFS or OS. OS = overall survival, PFS = progression-free survival.

assessment of survival prognosis can provide a basis for clinical individualized treatment.

In this study, by cox regression univariate and multivariate analysis, we found that smoking history, tumor stage, and surgical margin resection status were independently associated with PFS. Tumor staging and surgical margin resection status are independently associated with OS, which is consistent with previous reports.[8,9] The differences in other factors (including

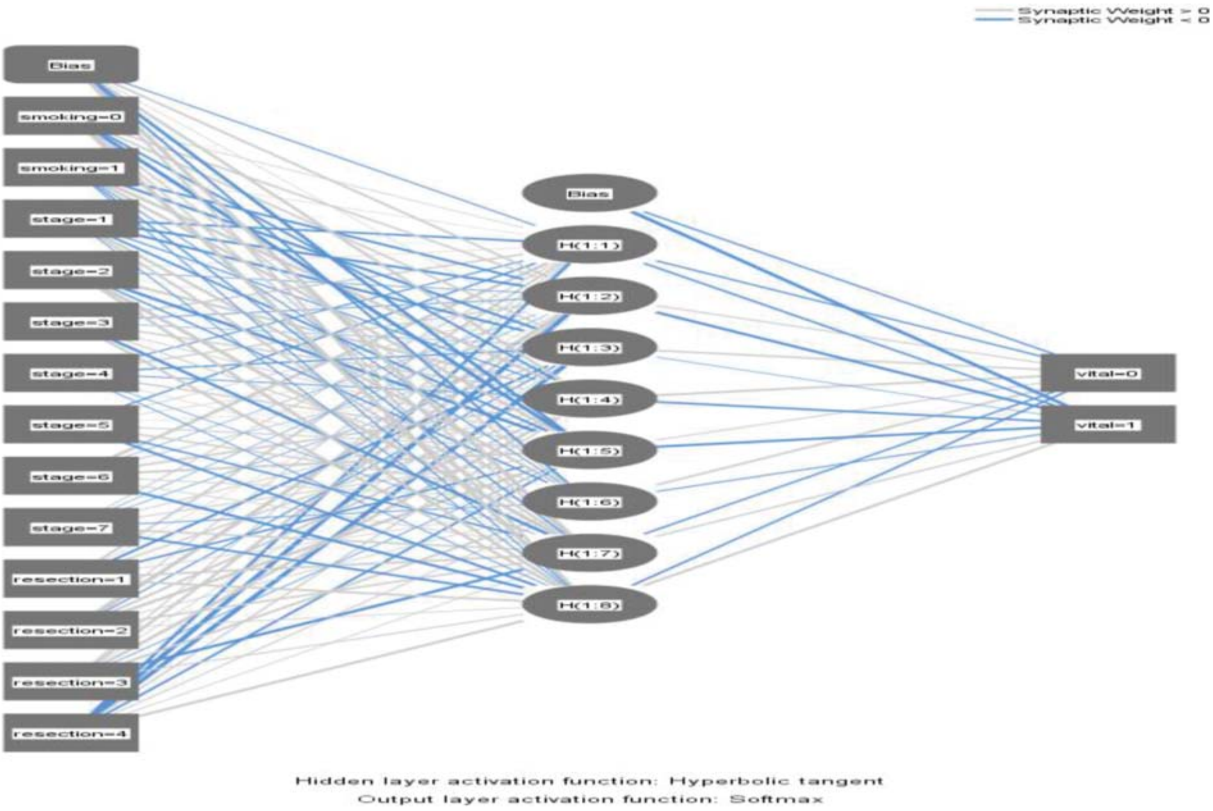

Figure 3. Artificial neural network pattern.

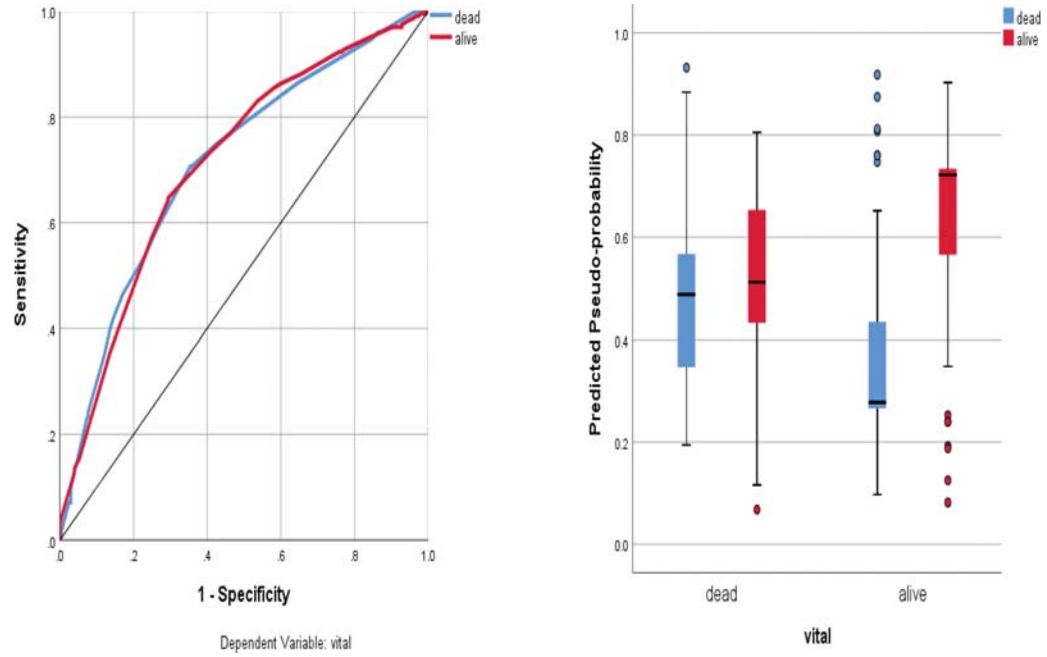

Figure 4. Artificial neural network ROC curve, prediction-measured map. ROC = receiver operating characteristic.

sex, age, smoking history, location of tumors, and whether there were genetic mutations) were not significant.

In previous studies, there were different reports on whether sex and age were the prognostic factors of lung cancer. Jubelirer et al[10] retrospectively analyzed 2207 patients with lung adenocarcinoma, and concluded that sex does not affect the prognosis of lung cancer, and the significant prognosis of death within 5 years. The indicators are staging, grading, age, and histopathology. In the study by Pitz et al,[11] the survival rate and prognosis of women with lung adenocarcinoma were significantly better than those of men. In this study, statistical analysis showed that age was not an independent factor affecting prognosis, which was consistent with previous studies.[8,12]

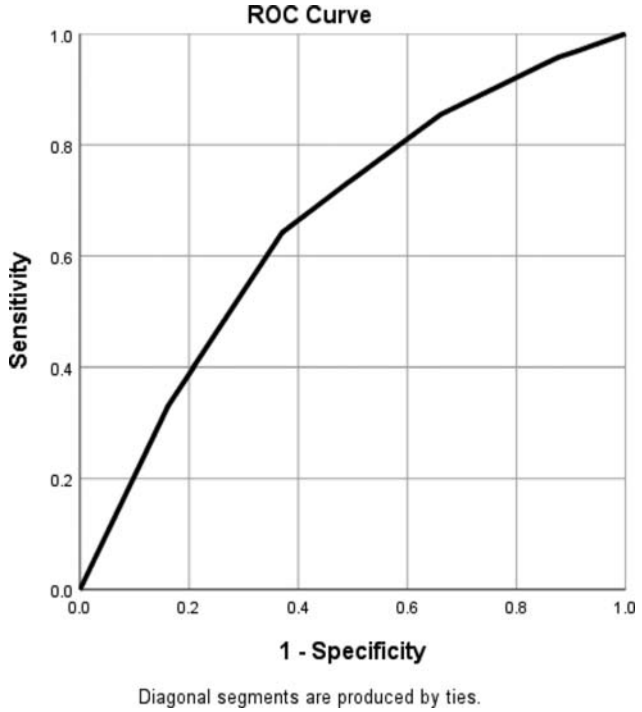

Figure 5. Binary logistic regression model ROC curve. ROC = receiver operating characteristic.

The PS score was considered to be an independent prognostic factor for LUAD, which was not covered in this study. In the literature, Kawaguchi et al[8] in a study of survival prognosis factors in 26,957 patients with NSCLC, by single factor and multivariate analysis found that the median OS of PS = 0 and PS = 1, respectively for 51.5 and 15.4 months, the difference was significant and the PS score was an independent prognostic factor for NSCLC.

The degree of tumor differentiation and the expression of Ki67 are also considered to be important factors influencing survival prognosis. Zhang et al[13] found in a retrospective study of 616 patients with lung adenocarcinoma, the prognosis of patients with poorly differentiated adenocarcinoma is not low. The prognosis of patients with differentiated adenocarcinoma is poor; the prognosis of patients with high expression of Ki67 is significantly worse than that of patients with low or no expression of Ki67.

In addition to the above factors, in recent years, the relationship between miRNA and survival and prognosis of lung adenocarcinoma was found. Lin et al[14] used the R language to analyze the differential miRNAs in LUAD and paracancerous tissues by mining the TCGA database. The results showed that miR-101-3p, miR-148a-3p, miR-192-5p, miR-193b-3p, miR-505-3p, miR-584-5p, and miR-99a-5p are associated with prognosis in LUAD patients.

Wang et al[15] discovered that the methylation level of methylation site cg12013757 of KRI1 gene has an effect on the prognosis of lung adenocarcinoma by mining the whole genome methylation data of TCGA database. The mRNA expression of corresponding gene is also related to lung. It is related to the prognosis of adenocarcinoma and can be further studied as a biomarker for the prognosis of lung cancer.

The above clinical factors are simple, easy to obtain, and have many clinical applications. In this study, we establish the clinical prognostic factors and construct the binary logistic regression model and the ANN model, respectively. The accuracy of correctly predicting prognosis using a binary logistic regression model was 60.6%, and the area under the ROC curve was 0.662. Compared with the traditional binary logistic regression model, the accuracy of the ANN model's own prediction is 65.8%. The accuracy of correct prognosis for predictive samples is 75.0%, and the area under ROC curve is 0.712. The prediction accuracy of the 2 models is roughly the same. The ANN model is slightly more accurate, and it can provide a new idea for the inference of clinical disease prognosis.

In summary, this study used the existing TCGA database to initially analyze the factors affecting the survival prognosis of LUAD, and established the ANN prediction model, which has higher prediction accuracy for the disease. The TCGA database has a large sample size and a wide variety of cancers, which requires us to further digging deeper to provide more guidance for the clinic.

Author contributions

Conceptualization: Xianrong Xu.
Data curation: Na Jiang.
Formal analysis: Na Jiang.
Funding acquisition: Xianrong Xu.
Methodology: Na Jiang.
Project administration: Xianrong Xu.
Resources: Na Jiang.
Software: Na Jiang.
Supervision: Na Jiang, Xianrong Xu.
Validation: Xianrong Xu.
Visualization: Xianrong Xu.
Writing – original draft: Na Jiang.
Writing – review & editing: Xianrong Xu.

References

[1] Cancer Genome Atlas Research NComprehensive molecular profiling of lung adenocarcinoma. Nature 2014;511:543–50.

[2] Little AG, Gay EG, Gaspar LE, et al. National survey of non-small cell lung cancer in the United States: epidemiology, pathology and patterns of care. Lung Cancer 2007;57:253–60.

[3] Yazdani Charati J, Janbabaei G, Alipour N, et al. Survival prediction of gastric cancer patients by Artificial Neural Network model. Gastroenterol Hepatol Bed Bench 2018;11:110–7.

[4] Kim SY, Moon SK, Jung DC, et al. Pre-operative prediction of advanced prostatic cancer using clinical decision support systems: accuracy comparison between support vector machine and artificial neural network. Korean J Radiol 2011;12:588–94.

[5] Afshar S, Afshar S, Warden E, et al. Application of artificial neural network in miRNA biomarker selection and precise diagnosis of colorectal cancer. Iran Biomed J 2019;23:175–83.

[6] Mehdy MM, Ng PY, Shair EF. Artificial neural networks in image processing for early detection of breast cancer. Comput Math Methods Med 2017;2017:2610628.

[7] Goodgame B, Viswanathan A, Miller CR, et al. A clinical model to estimate recurrence risk in resected stage I non-small cell lung cancer. Am J Clin Oncol 2008;31:22–8.

[8] Kawaguchi T, Takada M, Kubo A, et al. Performance status and smoking status are independent favorable prognostic factors for survival in non-small cell lung cancer: a comprehensive analysis of 26,957 patients with NSCLC. J Thorac Oncol 2010;5:620–30.

[9] Babacan NA, Yucel B, Kilickap S, et al. Lung cancer in women: a single institution experience with 50 patients. Asian Pac J Cancer Prev 2014; 15:151–4.

[10] Jubelirer SJ, Varela NL, Welch CA, et al. Does sex make a difference in survival of patients undergoing resection for early stage non-small cell lung cancer (NSCLC)? W V Med J 2009;105:18–22.

[11] Pitz MW, Musto G, Navaratnam S. Sex as an independent prognostic factor in a population-based non-small cell lung cancer cohort. Can Respir J 2013;20:30–4.

[12] Wakelee HA, Dahlberg SE, Brahmer JR, et al. Differential effect of age on survival in advanced NSCLC in women versus men: analysis of recent Eastern Cooperative Oncology Group (ECOG) studies, with and without bevacizumab. Lung Cancer 2012;76:410–5.

[13] Zhang XX, Deng CW, Qu YL, et al. Analysis of clinical prognostic factors in 616 patients with lung adenocarcinoma. J Med Res 2015;44:53–6.

[14] Lin K, Pan B, Xu X, et al. Establishment of a prognostic-related microRNAs risk model for lung adenocarcinoma based on TCGA database. Chin J Clin Lab Manag 2018;6:89–98.

[15] Wang K, Zhao RX, Yang SL, et al. Mining prognostic methylation sites and genes of lung adenocarcinoma based on TCGA database. J Nanjing Med Univ 2016;36:665–9.