Construction of a prognostic model of lung adenocarcinoma based on machine learning

Fan Liu, *Haonan* Jin, *Shuaibing* Jia, *Leifeng* Zhang, *Yingyue* Li, and *Jianhua* Zhang*

School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou, China

Abstract. In order to more accurately predict the prognosis and survival of lung adenocarcinoma patients, this paper used the gene expression and clinical information data of lung adenocarcinoma patients in the open database of TCGA to jointly construct a prognosis model of lung adenocarcinoma. Three difference analysis methods and univariate cox regression analysis were used as the preliminary screening method. By comparing the variable selection ability of lasso regression and random survival forest, comparing the performance of cox proportional risk regression model and random survival forest model, and integrating clinical data, a model that can more accurately predict the prognosis of lung adenocarcinoma patients was constructed. After comparison and selection, lasso regression was used to select variables and cox proportional risk model was used as the prediction model. The consistency index of the model reached 0.712. The AUC for 1-year, 3-year and 5-year survival of lung adenocarcinoma patients in the validation set were 0.808, 0.816 and 0.754, respectively. After the fusion of clinical data, the 1-year, 3-year and 5-year survival prediction AUC in the validation set were 0.840, 0.836 and 0.865, respectively, indicating that the model had good predictive performance.

Keywords: Lung adenocarcinoma, Prognosis model, Lasso regression, Cox proportional risk model.

1 Introduction

Lung cancer is a malignant tumor with the highest morbidity and mortality in the world, and the incidence of lung cancer in China is significantly higher than the world average level, while lung adenocarcinoma is a subtype of non-small cell carcinoma of lung cancer, accounting for 40% of primary lung tumors [1-2]. With the development of medical big data, it has become possible to conduct comprehensive and accurate analysis of tumors from the genetic level [3-6].

In this study, a prognostic model of lung adenocarcinoma was constructed based on machine learning method and combined with patient gene expression and clinical information to evaluate the survival status of lung adenocarcinoma patients, so as to facilitate clinicians to conveniently diagnose lung adenocarcinoma patients.

* Corresponding author email: petermails@zzu.edu.cn

2 Data processing

The data in this study came from the Cancer and Tumor Gene Atlas Project, the obtained data included 600 samples, each with 60,600 gene expression information, which were combined to convert patient ids into TCGA-named ids, which was conducive to distinguishing whether the samples had cancer or not. These data were cleaned and sorted out: First, the expression values of repetitive gene names were averaged, and 59,427 genes remained; Then, low-expression genes were removed, meaning that the average expression of genes in all samples was less than 1, leaving 34,819 genes. In the end, duplicate samples and formalin-soaked samples were removed, and 582 samples were retained, including 58 normal samples and 524 cancer samples.

The retained samples and genetic information were analyzed for genetic differences. DESeq2, edgeR and limma were respectively used for difference analysis. The screening criteria for differential gene selection from the three analyses is that the absolute value of logFC is less than 1 and the corrected P-value is less than 0.01, and 11226, 9653 and 7406 differential genes are obtained respectively. The intersection of these differential genes was used to obtain 6074 genes. The subsequent treatment process should be combined with the survival state and survival time of the patient. For the survival sample, the survival time is the last follow-up time, and for the dead sample, the survival time is the time from birth to death. After integrating the survival information of the samples, the remaining 469 samples were divided into the training set and the test set according to 7:3, and the number of alive and dead samples was also divided according to the proportion. In the training set, 328 samples were included, 208 of which were alive and 120 of which were dead. The training set consisted of 141 samples, of which 89 were alive and 52 were dead.

3 Feature selection

3.1. Univariate cox regression analysis

Single-factor cox regression analysis is a survival analysis method used to evaluate the impact of a single factor on survival time or survival probability. This method is often used in research to explore the relationship between a single variable and survival time, without considering other factors. Univariate cox regression analysis was performed on the data obtained from the difference analysis, and the risk ratio, 95% confidence interval and P-value of each gene could be obtained. According to the P-value of the results was less than 0.05, that is, there was statistical significance as a screening condition, 1308 significant genes were obtained.

3.2. Lasso-cox multivariate regression analysis

Univariate cox regression analysis screened 1308 significant genes, and lasso regression was used for further dimension reduction. Lasso regression uses the R package glmnet, set the parameter alpha to 1. The lasso regression model was cross-verified by 10 folds, and the results were visualized to obtain the deviation change of the regression model with the change of λ, as shown in figure 1:

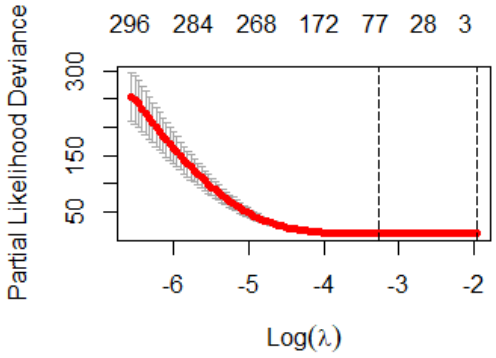

Fig. 1. Lasso regression coefficient selection.

In the figure, the dashed line on the left is lambda.min, and the dashed line on the right is lambda.1se. Here, the corresponding gene variables under the lambda.min parameter are taken, and 73 gene variables are obtained.

After lasso regression and random survival forest selection variables, the number of gene variables was still large. Here, multi-factor cox regression was selected for further reduction and screening. The difference between multi-factor cox regression analysis and single-factor cox regression analysis is that the influence of multiple covariables on survival time is considered. In this way, the gene variables that play a role together are screened out, which makes the model constructed by the final feature subset more simplified.

Multivariate cox regression analysis was performed on the expression data of 73 gene variables obtained by lasso regression screening. The standard was set as P value less than 0.05, and 12 gene variables were obtained after screening.

3.3. Random survival forest - multivariate cox regression analysis

The data obtained from single-factor cox regression analysis was used to construct a random survival forest model, and the relationship between the number of trees and the error rate was observed, the error rate reaches the lowest point when the number of trees is less than 100. Therefore, the number of trees is set to 100, the model is reconstructed, and the importance ranking of variables is obtained.The top 73 gene variables in the importance ranking are selected, which is consistent with the number of variables screened by lasso regression.

Multivariate cox regression analysis was performed on the expression data of 73 gene variables obtained from random survival forest screening. The setting standard was P value less than 0.05, and 12 gene variables were obtained after screening.

4 Model construction

In this section, the performance of the constructed model will be compared. The Index used is Concordance Index (C-index). Concordance index is an index used to evaluate the performance of the survival analysis model, which measures the accuracy of the model's ranking of the sample's survival time, and is especially suitable for evaluating the quality of the survival time prediction model.

After the model parameters are adjusted to the optimum,the consistency index obtained by the four models was compared, and the lasso-cox model was 0.712, the rsf-cox model was 0.702, the lasso-rsf model was 0.666, and the rsf-rsf model was 0.706, among which the lasso-cox model had the highest consistency index.

After model comparison, the cox proportional risk regression model was constructed and the effect of the model was verified by validation set. After the model was constructed, the survival functions of one year, three years and five years were set and the time-dependent ROC curve was obtained by validation set, as shown in figure 2:

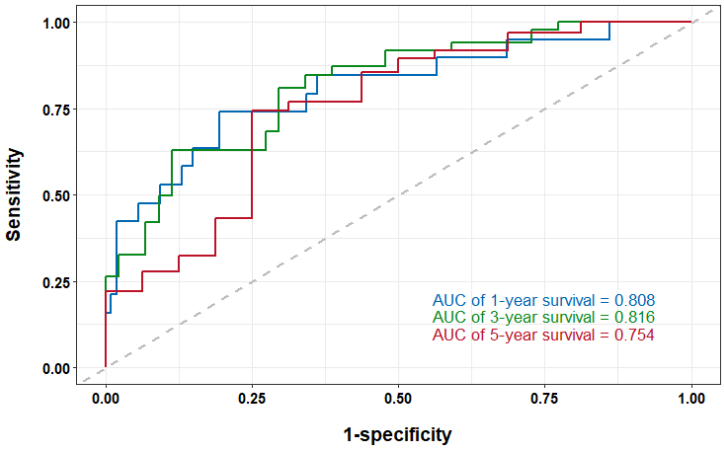

Fig. 2. Cox proportional risk model validation set timeROC curve.

As can be seen from the figure, the forecast AUC of the model for 1 year, 3 years and 5 years is 0.808, 0.816 and 0.754 respectively, achieving a good result.

5 Analyze clinical data

Clinical data were obtained from the TCGA database, including gender, race, age, cancer stage, smoking status, etc. The data of age, gender, cancer stage and annual smoking number of patients were extracted from the data, integrated with the survival status and time of patients, and univariate cox regression analysis was performed to obtain the results as shown in table 1:

Table 1. Results of univariate cox regression analysis of clinical data of lung adenocarcinoma.

Gene	Hazard.Ratio	X95.CI	P.value
age	1.01	0.99-1.03	0.373
gender	1.5	1.03-2.18	0.034
stage	1.47	1.23-1.76	0
pack_smoked_year	1	1-1.01	0.243

It can be seen from the table that gender and cancer stage have a significant impact on the survival of patients with lung adenocarcinoma in clinical data.

Gender and cancer stage data were included in the cox proportional risk regression model, samples with missing data were removed, and the data set was still divided into the training set and validation set according to the ratio of 7:3. The model was constructed using the training set, and the consistency index of the model was 0.740, which was significantly improved. The ROC curve obtained by using the validation set was shown in figure 3:

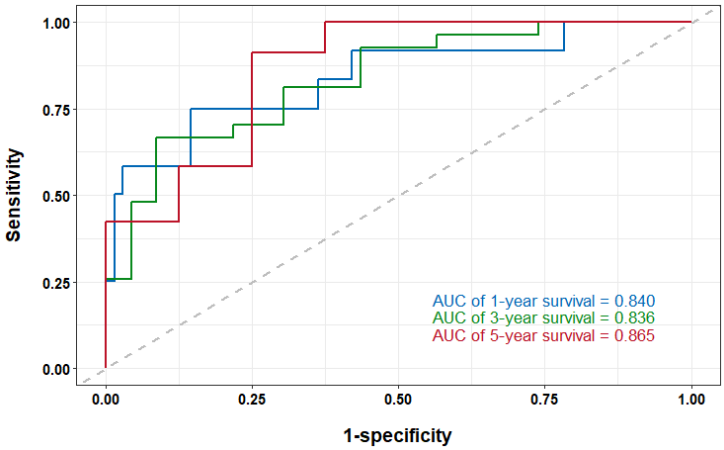

Fig. 3. The set timeROC curve was verified by cox proportional risk model combined with clinical data.

As can be seen from the figure, the forecast AUC of the model for 1 year, 3 years and 5 years is 0.802, 0.828 and 0.789 respectively, which also has a partial improvement compared with before.

The column diagram of the model after adding clinical data is shown in figure 4:

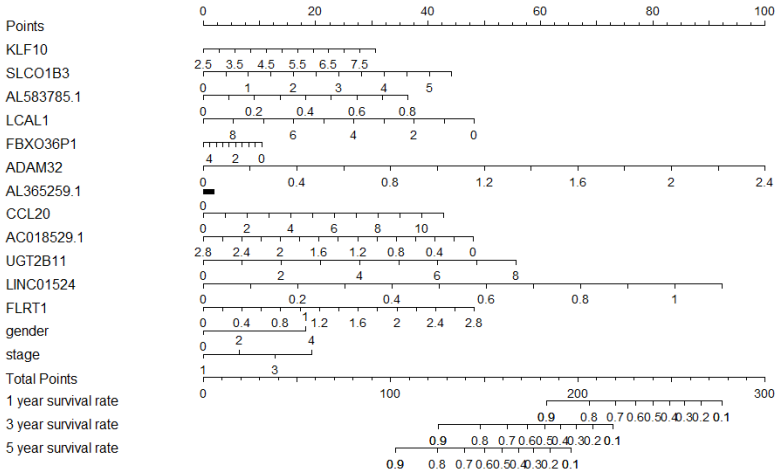

Fig. 4. Combined with clinical data cox proportional risk model diagram.

The chart shows the scores of each variable, with ADAM32 and LINC01524 performing well. Combining scores for all variables can predict a patient's probability of survival.

6 Conclusion

As a major subtype of lung cancer, lung adenocarcinoma poses a great threat to human life and health. Even after surgical treatment, about 30% of patients are at risk of recurrence[7-9]. After cleaning and sorting out the public data obtained from the TCGA database, differential genes were obtained by differential analysis. Gene difference analysis and univariate cox regression analysis were used as the initial screening conditions, and variables were screened by lasso regression for dimension reduction. The lasso model can be used to select variables

by controlling λ parameter, which is widely used in the medical field[9-11]. Finally, multivariate cox regression analysis reduces variables and builds cox proportional risk model, which can achieve better prediction effect. The prediction of patient survival based on gene expression alone can achieve a better prediction effect, but the model with the integration of clinical data will be more accurate. This study shows that gender and cancer stage have a greater impact on patient survival, which is similar to the results of previous studies[12-13].

The process of this method is closely related, and compared with the prediction model constructed by solely using gene expression data or clinical data, the effect is better. However, the clinical information in the acquired data is not comprehensive, and better results will be obtained if the data such as CT images of lung cancer, serum markers and sufficient clinical data can be combined[14-16]. The data used in this study is from a public database, which has a long time, and there are many clinical data missing. To obtain more satisfactory results, recent clinical data can be used to expand the sample size and increase clinical data.

In summary, this method selects 12 key gene variables and 2 clinical data variables according to the lung adenocarcinoma sample data in the TCGA database to jointly construct a prediction model. The model has good prediction accuracy and provides certain help for the follow-up treatment of lung adenocarcinoma patients.

References

1. Yanting Z, Salvatore V, Eileen M, et al. Global variations in lung cancer incidence by histological subtype in 2020: a population-based study. [J]. The Lancet. Oncology, 2023, 24 (11): 1206-1218.
2. Li Xiang, Gao Shen. Trends in incidence, morbidity and mortality of lung cancer in Chinese residents from 1990 to 2019 [J]. Chronic disease prevention and control in China, 2021, 29 (11): 821-826. DOI: 10.16386/j.cjpccd.issn. 1004-6194.2021.11.005
3. Hepp R. Gene Profiling in Recurrent Small Cell Lung Cancer [J]. Oncology Times, 2019, 41 (11): 28-29.
4. Yuhong J, Jun H, Xiaobo W, et al. Identification and validation of core genes in tumor-educated platelets for human gastrointestinal tumor diagnosis using network-based transcriptomic analysis. [J]. Platelets, 2023, 34 (1): 2212071-2212071.
5. Ferreira G L C, Nunes C S, Barbosa L L D, et al. 72. Comprehensive genomic profiling in the diagnosis of Central Nervous System tumors [J]. Cancer Genetics, 2023, 278-279 (S1):
6. Doudou G, Weihua X. Fuzzy-based concept-cognitive learning: An investigation of novel approach to tumor diagnosis analysis [J]. Information Sciences, 2023, 639
7. Liu Dongqi, Dou Fulin, Yang Xiaodong. Bioinformatics study of key genes in lung adenocarcinoma [J]. Chinese experimental diagnostics,2020,24(04):580-586.
8. Shaima B, Gwénaël T L, Riccardo B D, et al. Favoring the hierarchical constraint in penalized survival models for randomized trials in precision medicine. [J]. BMC bioinformatics, 2023, 24 (1): 96-96.
9. Wang Jinsong, Wei Jiayan, Peng Min. Interpretation and implications of the US cancer Statistics Report in 2023 and the latest global cancer statistics [J]. Journal of Practical Oncology, 2023, 38 (06): 523-527. DOI:10.13267/j.cnki.syzlzz.2023.083
10. H C C, A T M, D K R, et al. Design considerations and analytical framework for reliably identifying a beneficial individualized treatment rule. [J]. Contemporary clinical trials, 2022, 123 106951-106951.

11. Rémy J, Dzenis K, Florent C, et al. Prognosis of lasso-like penalized Cox models with tumor profiling improves prediction over clinical data alone and benefits from bi-dimensional pre-screening [J]. BMC Cancer, 2022, 22 (1): 1045-1045.
12. Carsten N, Gyda S A, Astrid D, et al. Sex Differences in Presentation, Treatment, and Survival in Patients Receiving Palliative (Chemo)Radiotherapy for Non-Small Cell Lung Cancer. [J]. Anticancer research, 2024, 44 (1): 301-305.
13. A A A R, Sun Y, Jui K, et al. Sex disparities in lung cancer survival rates based on screening status. [J]. Lung cancer (Amsterdam, Netherlands), 2022, 171 115-120.
14. Wen C, Xuewen H, Ying H, et al. A deep learning-and CT image-based prognostic model for the prediction of survival in non-small cell lung cancer. [J]. Medical physics, 2021, 48 (12): 7946-7958.
15. Shulin C, Hanqing H, Yijun L, et al. A multi-parametric prognostic model based on clinical features and serological markers predicts overall survival in non-small cell lung cancer patients with chronic hepatitis B viral infection [J]. Cancer Cell International, 2020, 20 (1): 555-555.
16. Seungwon O, SaeRyung K, InJae O, et al. Correction: Deep learning model integrating positron emission tomography and clinical data for prognosis prediction in non-small cell lung cancer patients. [J]. BMC bioinformatics, 2023, 23 (S9): 573-573.