Proceedings of Anticancer Research, 2026, Volume 10, Issue 2
http://ojs.bbwpublisher.com/index.php/PAR
ISSN Online: 2208-3553
ISSN Print: 2208-3545

Construction of a Prognostic Model for Lung Adenocarcinoma Based on Bioinformatics Analysis of Glycolysis-Related Genes

Yongming Kang*

Dalian Women and Children's Medical Group, Dalian 116011, Liaoning, China

**Author to whom correspondence should be addressed.*

Abstract: *Objective*: This study aims to collect lung adenocarcinoma samples from the Cancer Genome Atlas (TCGA) database and explore the differential expression of glycolysis-related genes between lung adenocarcinoma tissues and adjacent normal tissues. By combining differentially expressed genes with prognostic data, we investigate the correlation between them and establish a prognostic prediction model for the survival rate of lung adenocarcinoma. *Methods*: Raw expression data were downloaded from the TCGA database and organized using the Perl language. Differential analysis was performed using the "limma" package in R software. Univariate Cox regression analysis was employed to screen glycolysis-related genes associated with the survival of lung adenocarcinoma patients. Correlation analysis and consensus clustering analysis were then conducted. Lasso regression analysis and 10-fold cross-validation were used to screen glycolysis-related genes associated with prognosis. Kaplan-Meier survival curves were plotted to confirm significant differences between high- and low-risk groups, and the receiver operating characteristic (ROC) curve was plotted to calculate the area under the curve (AUC). Finally, a risk model was constructed. *Results*: Based on data from the TCGA database, 19 differentially expressed glycolysis-related genes were identified (17 upregulated and 2 downregulated). Univariate Cox regression analysis revealed that 14 genes were significantly associated with prognosis, among which five genes, including PGAM1 and NUP50, were identified as risk factors, while HK3 and PRKACA were protective factors. Following consensus clustering analysis, lung adenocarcinoma patients were classified into three subtypes. Survival analysis demonstrated significant prognostic differences among these subtypes, with subtype 2 exhibiting the worst prognosis. Using LASSO regression, 11 key glycolysis-related genes were selected, and a risk scoring model was constructed based on these genes. According to this model, patients were divided into high- and low-risk groups, revealing significant differences in survival rates between the two groups ($P < 0.001$). The ROC curve demonstrated the model's good predictive ability for 1-, 2-, and 3-year survival rates (AUCs of 0.742, 0.725, and 0.673, respectively). *Conclusion*: This study found a correlation between glycolysis-related genes and the prognosis of lung adenocarcinoma. A risk scoring formula based on 11 key glycolysis-related genes was developed, and a risk model was constructed to predict the survival rate of lung adenocarcinoma patients using their risk scores along with T stage, N stage, and overall stage. This model provides valuable assistance for clinical research and individualized treatment of lung adenocarcinoma.

Keywords: Glycolysis; Lung adenocarcinoma, TCGA database; Prognostic model; Bioinformatics

Online publication: April 14, 2026

1. Introduction

As one of the most common malignant tumors, lung cancer has shown an increasing trend in severity year by year, with lung adenocarcinoma being the most prevalent type. The exploration of new indicators and methods for the early diagnosis and prognostic evaluation of lung adenocarcinoma has consistently been a constant theme of research [1]. The Cancer Genome Atlas (TCGA) database is centered around tumor patient samples, and through data mining of the TCGA database, numerous potential tumor therapeutic targets and clinical biomarkers have been identified. Glycolysis refers to the process by which glucose or glycogen in the body is broken down into lactic acid while producing a small amount of ATP under anaerobic or hypoxic conditions. The abnormal energy metabolism in tumor cells, particularly the high level of glycolysis, indirectly reflects the occurrence and growth of tumors. Therefore, studying new intervention methods and therapeutic drugs from this perspective has become a direction of efforts [2]. Kaplan-Meier survival analysis and Cox regression analysis are common methods for survival analysis. Kaplan-Meier survival analysis can analyze events based on a single influencing factor, with the time range for each independent individual extending from the recording point to the event occurrence point. Cox regression analysis is a multi-parameter regression model that uses survival outcome and survival time as dependent variables, allowing for the simultaneous analysis of the effects of multiple factors on survival.

2. Materials and methods

2.1. Case data

Obtain data from The Cancer Genome Atlas (TCGA) database and download gene expression data for a total of 551 lung adenocarcinoma samples from the official data repository website (https://portal.gdc.cancer.gov/). This includes 497 tumor tissue samples of lung adenocarcinoma and 54 adjacent tissue samples, with the cutoff date for sample acquisition being December 2019.

3. Research methods

3.1. Data organization and expression analysis

A gene set related to glycolysis (EACTOME_GLYCOLYSIS) containing 72 genes was obtained from the GSEA website. Using R software (version 4.0.2), the expression levels of glycolysis genes were extracted from the transcriptomic information of the samples, and the gene expression data were normalized. The “Corrplot” package in R software was used to conduct co-expression analysis of glycolysis genes. The expression of glycolysis genes in lung adenocarcinoma tissues and normal tissues was analyzed, and the “limma” package in R software (http://bioconductor.org/) was used to perform differential analysis of glycolysis genes. A p-value < 0.05 was considered statistically significant, and glycolysis genes with differential expression were screened based on the condition of $|\log FC| \geq 1$. Finally, the “pheatmap” package was used to visualize the results as a heatmap.

3.2. Screening of prognostic markers and establishment of a prognostic risk scoring model

We first merged gene expression data with survival data, and then used the “survival” package in R software to conduct univariate Cox regression analysis on glycolysis-related genes. This allowed us to calculate the hazard ratio (HR), 95% confidence interval, and P-value for each gene in relation to the survival of lung adenocarcinoma patients. Glycolysis genes significantly associated with the prognosis of lung adenocarcinoma patients were

screened out based on a *P*-value threshold of < 0.05, and a forest plot was generated.

Consensus clustering analysis was performed on the expression levels of glycolysis genes to determine the optimal number of clusters (k). Subsequently, survival analysis was conducted on the identified clusters. Using the ConsensusClusterPlus package, all glycolysis genes were divided into k distinct subtypes, and consensus clustering plots, cumulative distribution function curves, Delta plots, and survival curves were generated.

Lasso regression analysis was performed on glycolysis genes associated with the prognosis of lung adenocarcinoma using the "glmnet" package in R to further screen for prognostic variables. Lasso regression employed 10-fold cross-validation to determine the optimal λ value, thereby obtaining regression coefficients for each variable and constructing a risk score equation (risk score) based on glycolysis gene expression. The risk score equation was established as Risk score = β1×mRNA1EXP + β2×mRNA2EXP + ... + βn×mRNAnEXP, incorporating the screened glycolysis differentially expressed genes associated with prognosis. In the formula, β represents the regression coefficient of the corresponding mRNA, mRNAEXP represents the expression level of the corresponding gene, and n represents each glycolysis gene used as a variable.

4. Results

4.1. Abnormal expression of glycolytic genes in lung adenocarcinoma tissues

Among the 72 glycolysis-related genes, 62 exhibited differential expression, with 8 downregulated and 54 upregulated ($P < 0.05$). Further screening using the criterion of $|\log FC| \geq 1$ revealed that 19 genes showed significant differential expression, with 2 downregulated and 17 upregulated. In **Figure 1**, red indicates upregulation of the gene in lung adenocarcinoma tissues compared to normal tissue glycolytic genes, while blue indicates downregulation. The greater the difference in gene expression, the darker the color of the block.

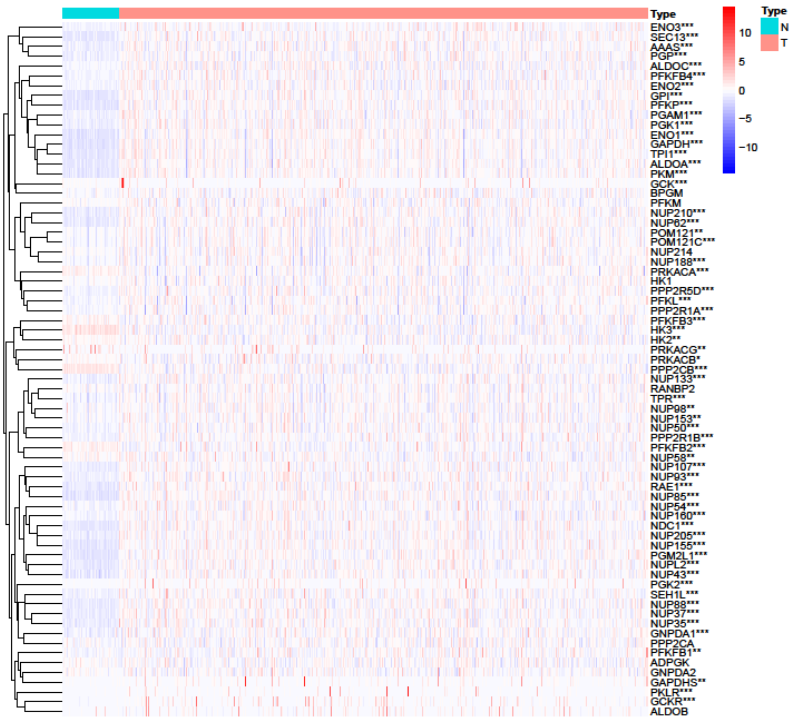

Figure 1. Heatmap of glycolytic gene expression.

4.2. Screening glycolytic genes associated with the prognosis of lung adenocarcinoma

After integrating glycolytic gene expression data with sample survival information, univariate Cox regression analysis was performed using the "survival" package in R software to calculate the hazard ratio (HR) and *P*-value for each glycolytic gene in relation to lung adenocarcinoma patients. Fourteen genes were identified as being associated with patient survival ($P < 0.05$). A forest plot (**Figure 2**) was generated, revealing that the genes identified as risk factors (HR > 1) included PGAM1, NUP50, PPP2R1A, NUP37, and HK2, while the genes identified as protective factors (HR < 1) included HK3 and PRKACA (**Figure 2**).

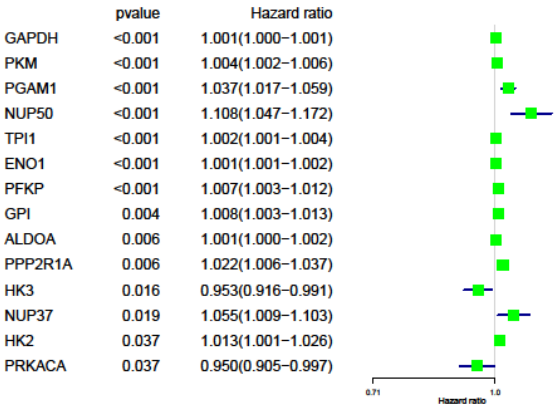

Figure 2. Forest plot of differentially expressed genes.

4.3. Co-expression analysis of glycolytic genes

Correlation analysis between every two glycolytic genes in lung adenocarcinoma tumor tissues and adjacent normal lung tissues was conducted using the Corrplot software package. As shown in **Figure 3**, red indicates a positive correlation, blue indicates a negative correlation, and the deeper the color, the higher the degree of correlation.

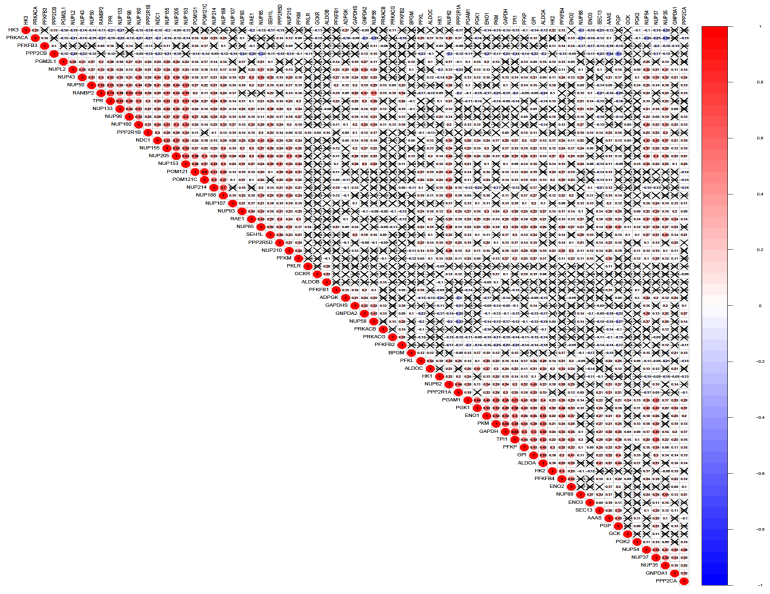

Figure 3. Co-expression analysis of glycolytic genes in lung adenocarcinoma.

4.4. Consensus clustering analysis

Consensus clustering analysis was performed on the expression levels of glycolytic genes to select the optimal number of clusters. Subsequently, survival analysis was conducted on the clustered groups. Using the ConsensusClusterPlus software package, all glycolytic genes were divided into k different subtypes (**Figure 4** shows matrix heatmaps for k = 2 to 9). When k =3, the optimal partition was obtained based on the CDF curve of the consensus score (**Figure 5**). Finally, the optimal number of clusters, assessed through the area under the curve (**Figure 6**), was determined to be 3 (**Figure 7**). Therefore, we divided the samples into three groups and conducted survival analysis, revealing a significant relationship with prognosis, as shown in **Figure 7**.

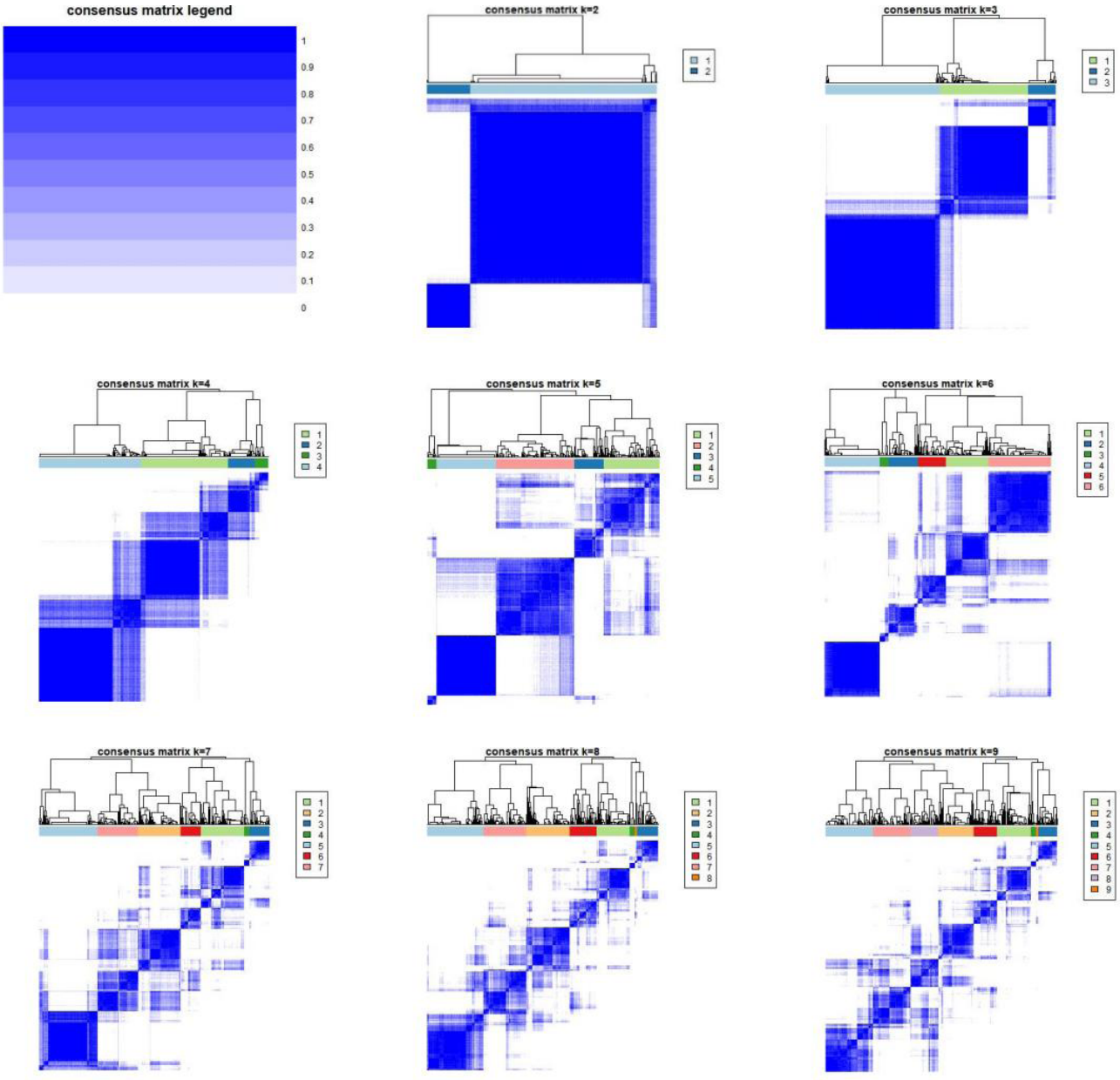

Figure 4. Consensus clustering matrix of glycolytic genes for k = 2 to 9.

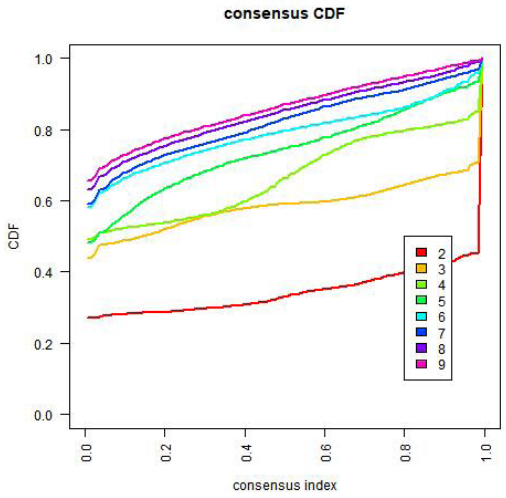

Figure 5. Consensus cumulative distribution function (CDF) curve for determining the optimal number of clusters.

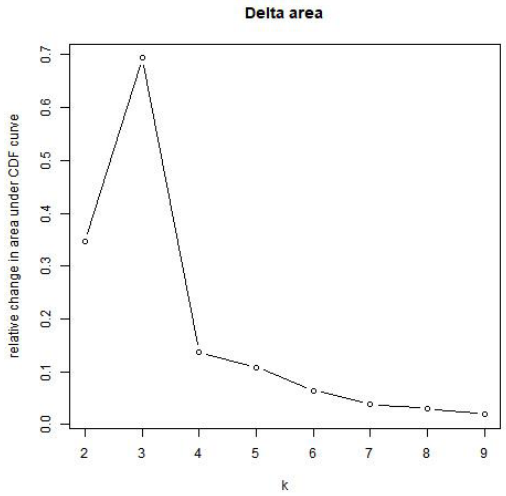

Figure 6. Relative change in area under the CDF curve.

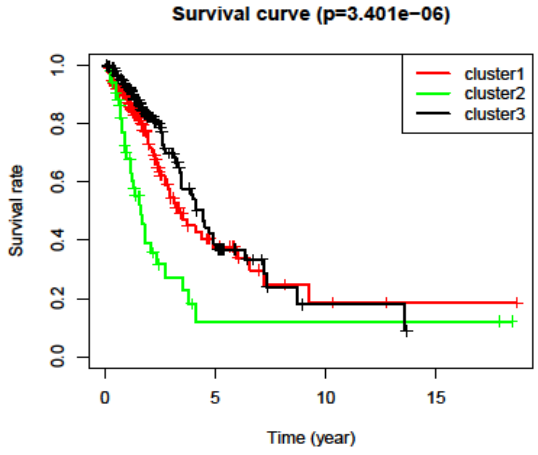

Figure 7. Kaplan-Meier survival analysis for three glycolytic gene clusters.

4.5. Establish a risk prediction model

Lasso regression analysis was conducted on glycolysis-related differentially expressed genes using the "glmnet" package in R software to screen variables. Lasso regression employed 10-fold cross-validation to determine the optimal λ value (**Figure 8**), thereby identifying 11 glycolysis-related genes associated with prognosis (PRKACA, PPP2R1A, PKM, PGAM1, PFKP, NUP50, HK3, GAPDH, ENO3, ENO1, ALDOA) along with their corresponding regression coefficients (**Figures 9 and 10**). Subsequently, a risk score equation based on the expression of glycolysis-related genes was constructed.

The risk score equation was established by incorporating the identified 11 glycolysis-related genes associated with prognosis, following the formula Riskscore = β1×mRNA1EXP + β2×mRNA2EXP + ... + βn×mRNAnEXP. In this formula, β represents the regression coefficient of the corresponding mRNA, and mRNAEXP denotes the expression level of the corresponding gene.

Based on the prognostic model, the risk score for each lung adenocarcinoma patient can be calculated. Using the median risk score value, the model stratifies the included lung adenocarcinoma patients into high-risk and low-risk groups (**Figure 12**). A heatmap reveals significant differences between the two groups in terms of stage, T stage, gender, age, and final outcome. The Kaplan-Meier curve (**Figure 11**) demonstrates significant differences in overall survival rates between the high-risk and low-risk groups ($p < 0.05$), with the high-risk group exhibiting a worse prognosis compared to the low-risk group.

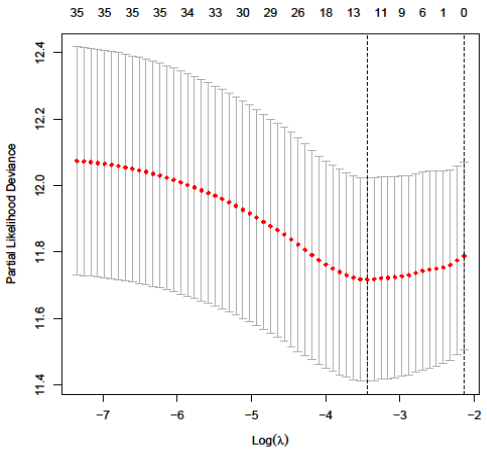

Figure 8. Lasso regression cross-validation plot for selecting the optimal tuning parameter (λ).

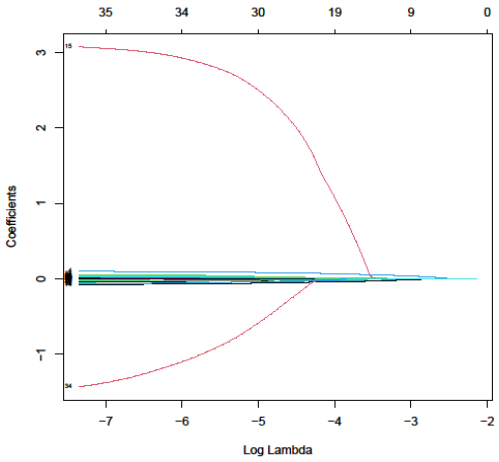

Figure 9. Lasso coefficient profiles of glycolysis-related differentially expressed genes.

Gene	Coef
GAPDH	0.000299433616986084
PKM	0.000442082242810737
PGAM1	0.0059626716663652
NUP50	0.0519997732227822
ENO1	0.000260433037189509
PFKP	0.0020475602804881
ALDOA	0.000104245306775322
PPP2R1A	0.00588082227240346
HK3	−0.0223346157432718
PRKACA	−0.0295740350832226
ENO3	−0.00602431857039597

Figure 10. Partial likelihood deviance plot for variable selection in Lasso regression.

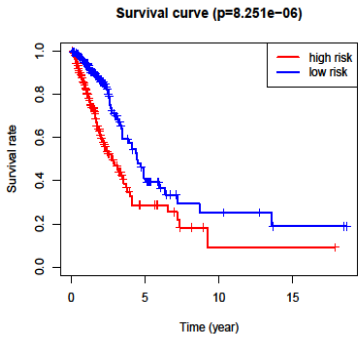

Figure 11. Kaplan-Meier survival analysis of high-risk and low-risk groups stratified by the prognostic model.

Figure 12. Heatmap of high-risk and low-risk groups.

5. Discussion

Lung adenocarcinoma, the most common pathological subtype of lung cancer, accounts for 40% of all lung cancer cases and poses a significant threat to human life and health [3]. Therefore, the search for new indicators and methods for early diagnosis, prognostic evaluation, and personalized treatment of lung adenocarcinoma has always been a consistent theme of research. In recent years, there have been numerous research findings on survival analysis of gene data from the TCGA database. By utilizing data mining techniques, researchers have obtained many gene lineages that can guide tumor diagnosis, treatment, and survival prediction.

High levels of glycolysis represent a significant metabolic characteristic of tumors. Regardless of whether the environment is aerobic or anaerobic, tumor cells rapidly obtain energy through glycolysis, a relatively short reaction pathway, to meet their rapid proliferation and growth needs. Meanwhile, the large amounts of intermediate products generated also serve as raw materials for tumor growth [4].

This study successfully established a prognostic prediction model for lung adenocarcinoma based on the data analysis of glycolysis-related genes in lung adenocarcinoma tissues, involving 11 glycolysis genes.

The study obtained a glycolysis-related gene set (EACTOME_GLYCOLYSIS) containing 72 genes from the GSEA website. The expression levels of glycolysis genes were extracted from the transcriptome information of the samples. The gene expression data were normalized using R software (version 4.0.2). Statistical significance was set at $P < 0.05$. Differential analysis of glycolysis genes was conducted using the “limma” package in R software (http://bioconductor.org/), revealing that 62 out of the 72 glycolysis-related genes exhibited differential expression. The above analysis indicated significant differences in the expression of glycolysis-related genes between tumor cells and normal cells in patients with lung adenocarcinoma. After merging the glycolysis gene expression data with sample survival information, we performed univariate Cox regression analysis on the differentially expressed

genes using R software and identified 14 genes associated with the survival of patients with lung adenocarcinoma ($P < 0.05$). Among them, genes classified as risk factors (HR value > 1) included PGAM1, NUP50, PPP2R1A, NUP37, and HK2, while genes classified as protective factors (HR value < 1) included HK3 and PRKACA [5]. We also conducted consensus clustering analysis on the sample data using ConsensusClusterPlus and found that the samples could be divided into three groups, each with similar glycolysis gene expression characteristics. Survival analysis revealed differences in prognostic survival among the three groups, with the second group having the worst prognosis. This suggests a potential association between specific glycolysis gene expression and the survival prognosis of patients with lung adenocarcinoma [6]. Lasso regression analysis was performed on differentially expressed genes using the R package "glmnet", resulting in the selection of 11 glycolysis-related genes associated with prognosis, along with their corresponding regression coefficients, for the construction of a prognostic model for lung adenocarcinoma based on glycolytic gene expression. From this, the risk score for each sample could be calculated [7]. Based on the median risk score, samples were divided into high- and low-risk groups. Survival analysis revealed a significant difference in overall survival (OS) between the two groups ($p < 0.05$), with the high-risk group exhibiting a worse prognosis than the low-risk group. The predictive model demonstrated good performance in forecasting 1-year(**Figure 13**), 2-year(**Figure 14**), and 3-year survival for lung adenocarcinoma patients, as validated by ROC curve analysis(**Figure 15**).

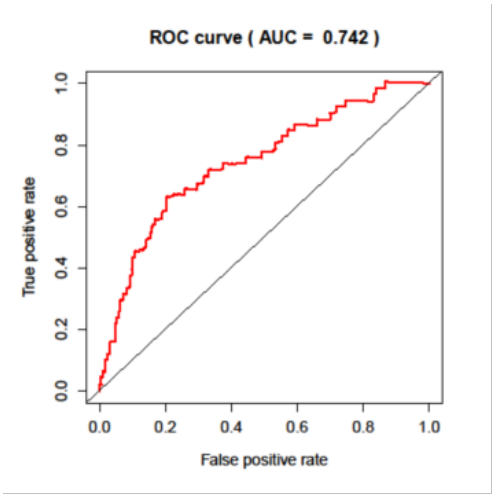

Figure 13. AUCs of 0.742

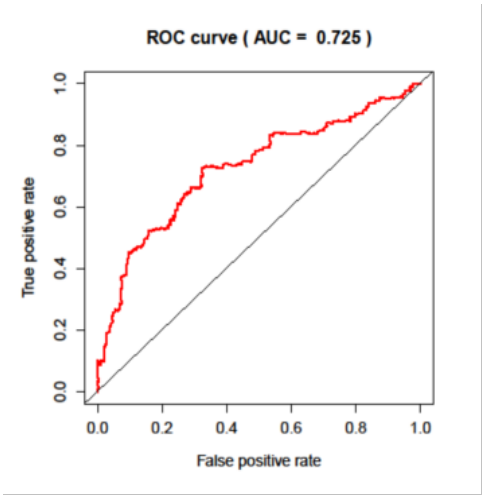

Figure 14. AUCs of 0.725

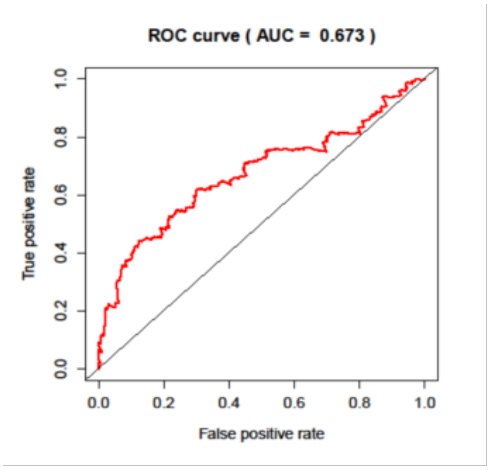

Figure 15. AUCs of 0.673

6. Conclusion

In this study, through data mining and analysis, we identified and demonstrated a link between abnormal expression of glycolysis-related genes and prognosis in lung adenocarcinoma patients. We selected 11 glycolysis-related genes associated with prognosis in these patients and constructed a model to predict survival rates using these genes, validating its sensitivity and effectiveness. By accurately assessing the prognosis of lung adenocarcinoma patients, we can develop more individualized treatment plans, enhancing the precision and foresight of clinical treatment for lung adenocarcinoma while reducing the waste of medical resources and benefiting patients.

Disclosure statement

The author declares no conflict of interest.

References

[1] Liu B, Sun C, Wang X, et al., 2025, Bioinformatics Analysis of Differentially Expressed Genes in Multiple Primary Lung Cancers Based on the GEO Database. Journal of Jilin University (Medical Edition), 51(02): 437–446.

[2] Ding D, Zhao R, Ding Y, et al., 2024, Construction of a Prognostic Model for Lung Adenocarcinoma Patients Using Glycolysis-Related LncRNAs. Medical Information, 37(05): 1–11 + 19.

[3] Wang X, Yang Y, Pan Z, et al., 2023, Exploring the Mechanism of Ginseng in Regulating Ferroptosis in Lung Adenocarcinoma Based on Bioinformatics, Network Pharmacology, and Molecular Docking. Special Wild Economic Animal and Plant Research, 45(05): 57–65.

[4] Lu W, Chen T, Yao Y, 2023, Analysis of the Expression and Prognosis of LDHA and LDHB Genes in Lung Adenocarcinoma Based on Public Databases. Zhejiang Journal of Integrated Traditional Chinese and Western Medicine, 33(05): 412–417.

[5] Zhang Y, Wang J, 2021, Bioinformatics Analysis of the Expression and Clinical Significance of Lactate Dehydrogenase A in Lung Adenocarcinoma. Chongqing Medicine, 50(16): 2804–2812.

[6] Zhang Y, Jia F, Wang Q, et al., 2020, Bioinformatics Analysis of the Expression and Clinical Significance of PKM2 in Lung Adenocarcinoma. Journal of Cancer Prevention and Treatment, 33(09): 760–766.

[7] Yang H, Lai H, Rao Y, et al., 2020, Evaluation of the Expression and Clinical Significance of the UBE2 Family and UBE2T Gene in Lung Adenocarcinoma Based on Bioinformatics Approaches. Journal of Sun Yat-sen University (Medical Sciences), 41(03): 445–451.