A prognostic model for predicting recurrence-free survival in hepatocellular carcinoma patients

Wenhua Wang  
Nanchang University  
Lingchen Wang  
Nanchang University  
Xinsheng Xie  
Nanchang University  
Yehong Yan  
First Affiliated Hospital of Nanchang University  
Yue Li  
Nanchang University  
Quqin Lu (✉ quqinlu@ncu.edu.cn)  
Nanchang University  
https://orcid.org/0000-0003-2813-197X

Research article

Keywords: TCGA, hepatocellular carcinoma, recurrence-free survival, risk score, prognostic model

DOI: https://doi.org/10.21203/rs.3.rs-30658/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
Abstract

Hepatocellular carcinoma (HCC) remains the most frequent liver cancer, accounting for approximately 90% of primary liver cancers worldwide. The recurrence-free survival (RFS) of HCC patients is a critical factor in devising a personal treatment plan. Thus, it is necessary to accurately forecast the prognosis of HCC patients in clinical practice. Using the The Cancer Genome Atlas (TCGA) dataset, we identified genes that are associated with RFS. A robust likelihood-based survival modeling approach was used to select the best genes for the prognostic model. Then, the GSE76427 dataset was used to evaluate the prognostic model’s effectiveness. We identified 1331 differentially expressed genes associated with RFS. Seven of these genes were selected to generate the prognostic model. Validation in both the TCGA cohort and the GEO cohort demonstrated that the 7-gene prognostic model has the capability of predicting the RFS of HCC patients. Meanwhile, the result of multivariate Cox regression showed that the 7-gene prognostic model could work as an independent prognostic factor. In addition, according to the time dependent ROC curve, the 7-gene prognostic model performed better in predicting the RFS of the training set and the external validation dataset than the classical TNM staging and BCLC. What’s more, these seven genes were found to be related to the occurrence and development of liver cancer by exploring other three databases. Our study identified a seven-gene signature for HCC RFS prediction that is a novel and convenient prognostic tool. The seven genes might provide potential target genes for metabolic therapy and the treatment of HCC.

Background

In 2018, liver cancer remained among the top six prevalent carcinomas. There were 841,080 new patients, and 781,631 patients died of liver cancer, according to the Global Cancer Statistics (1) (2). Hepatocellular carcinoma (HCC) is the most frequent liver cancer, accounting for approximately 90% of primary liver cancers (3). Despite the continuous development of medical technology, the outcome of many patients who receive treatment remains poor, and the prognosis of liver cancer remains poor, with a 2-year recurrence rate of 76.9% (4). The recurrence-free survival (RFS) of HCC patients is a critical factor in devising a personal treatment plan (5). Thus, it is necessary to accurately forecast HCC prognosis to improve the prognosis of HCC. Most previous studies constructed prognostic models using the TNM (tumor-node-metastasis) staging system to assess the prognosis of HCC patients (6). However, the TNM staging system does not predict the prognosis of HCC. Therefore, it is important to develop a reliable tool for clinicians to predict the prognosis of patients with HCC.

Given the remarkable development of high-throughput technologies, the development of The Cancer Genome Atlas (TCGA) and the intergovernmental Gene Expression Omnibus (GEO) database provides an abundance of high-quality information for HCC (7). Hence, it is urgent to develop methods to identify reliable therapeutic gene targets that could enable earlier prognostic evaluation and better therapeutic strategies (8). Therefore, we considered whether we could build a gene-based risk score model (9). Our goal is to generate simple and effective prognostic tools based on several genes and other factors that may affect RFS (10). Using the TCGA dataset, we selected 7 genes by robust likelihood-based survival
modeling and built a risk score system (11). We used an independent dataset (GSE76427) to validate the effectiveness of the risk score system, demonstrating that its clinical value for predicting RFS in HCC patients is better than that of the TNM staging system.

**Materials And Methods**

2.1 Data collection and survival analyses

First, we downloaded the gene expression profiles and clinical information from The Cancer Genome Atlas-liver hepatocellular carcinoma (TCGA-LIHC)(12) dataset, which included 335 HCC samples. We used GSE76427 as a validation group, which contained 115 HCC samples, including gene expression and clinical information. Samples in TCGA-LIHC and GSE76427 that met the following inclusion criteria were included in this study: all samples had mRNA sequencing data and clinical information on RFS(13).

2.2 Identification of genes associated with RFS

The raw count data were normalized with a log(a+1) transformation. Then, using the “survfit” function in the “survival” package, we plotted the Kaplan-Meier curves for the high and low expression groups of each gene. When the $P$-value of the log rank test is less than 0.05, it was considered statistically significant (14).

2.3 Enrichment analysis of GO functions and KEGG pathways

For the selected genes, we used WebGestalt (http://bioinfo.vanderbilt.edu/webgestalt) based on GO (Gene Ontology) functions and the Kyoto Encyclopedia of Genes and Genomes (KEGG) to understand the biological significance of the identified genes(15).

2.4 Identification of the best genes for modeling

A robust likelihood-based survival approach was used to identify the best genes for modeling after determining the genes associated with RFS (16). We used the “rbsurv” package in R to complete this modeling process. The algorithm for modeling is summarized as follows:

1. We randomly divided the experimental group into two sets: a training set with $N \times (1-p)$ samples and a validation set with $N \times p$ samples ($p = 1/3$). Next, we adjusted a gene in the training set to obtain estimates of the genetic parameters. Log-likelihoods were evaluated using parameter estimates and a set of validation samples. This evaluation was performed in all genes.

2. We repeated the above process 10 times and obtained 10 log-likelihoods for each gene. We selected the best genes with the largest mean log-likelihoods. All of the best genes related to overall survival were selected using a robust likelihood approach.

3. We calculated every two-gene model and selected the one with the largest mean log-likelihood by repeating the two steps above.
4. We continued this process of gene selection. As a result, a series of models was generated using the Akaike Information Criterion (AIC) for all the candidate models and choosing the best model with the smallest AIC.

2.5 Construction and validation of the risk score system

Multivariate Cox regression analysis and “rbsurv” analysis were performed to identify the genes related to recurrence-free survival and to construct the prognostic gene signature. The “survivalROC” package in R was used to investigate the time-dependent prognostic value. Optimal cut-off values based on ROC curves were obtained to classify patients into low-risk groups and high-risk groups. The Calibration curve and concordance index (C-index) were used to evaluate the risk score system.

2.6 External validation of the risk score system

We calculated the risk score in the GSE76427 dataset. Then, the AUCs for 12-month, 15-month, and 18-month RFS and the Kaplan-Meier curves were used to verify the risk score system. The calibration curve was used to validate the risk score system.

In addition, the prognosis-related genes in the risk score system were verified at the protein level by using The Human Protein Atlas database. CBioPortal for cancer genomics was used to study genetic alterations and in the risk score system (17).

2.7 Statistical analysis

Statistical tests were performed using R software and SPSS. Univariate and multivariate Cox regression were performed with a forward stepwise procedure. When the P-value was less than 0.05, it was considered statistically significant (17).

Result

3.1 Acquisition of Gene Expression and Clinical Data

We downloaded the TCGA-LIHC dataset from The Cancer Genome Atlas (http://portal.gdc.cancer.gov/). The TCGA-LIHC dataset included 334 samples, and all samples had data on the recurrence-free survival time and censoring status. The GSE76427 dataset was downloaded from the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/gov/). The GSE76427 dataset included 115 samples from HCC patients, but 7 patients had missing information on recurrence-free survival time and censoring status. Thus, 108 samples were included in this study. The median RFS times of the TCGA and GSE76427 series were 390 and 252 days, respectively, and the two datasets contained clinical information, such as sex, age, and TNM stage.

3.2 Genes associated with recurrence-free survival
We used the “survfit” function from the “survival” package and found 1331 genes associated with RFS. To explore the genetic biological implications, we next analyzed the 1331 genes through Gene Ontology (GO) functional and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses. As shown in Figure 1, through KEGG analysis, we found that these genes are enriched in signaling pathways such as the cell cycle, homologous recombination, DNA replication, the Fanconi anemia pathway, complement and coagulation cascades, and the T cell receptor signaling pathway.

3.3 Construction of the prognostic model in TCGA-LIHC

Then, “rbsurv” identified seven genes to construct the risk score system. The seven genes included in the system were TTK protein kinase (TTK), chromosome 16 open reading frame 54 (C16orf54), phosphoribosyl pyrophosphate amidotransferase (PPAT), CD3e molecule associated protein (CD3EAP), Solute carrier organic anion transporter family member 2A1 (SLCO2A1), acetyl-CoA acetyltransferase 1 (ACAT1), and Growth-arrest specific 2 like 3 (GAS2L3) (Table ).

The risk score was calculated with the following formula: risk score= (-0.038)*expression of TTK + (-0.357)*expression of C16orf54 + 0.634* expression of PPAT +0.221 * expression of CD3EAP +(-0.076) * expression of SLCO2A1 + (-0.184) * expression of ACAT1 + 0.277 * expression of GAS2L3.

A total of 334 patients were divided into two groups (134 high-risk patients and 200 low-risk patients) with a cut-off of 4.97976 for the risk score. Furthermore, the survival curve revealed that the recurrence-free survival was significantly poorer in the high-risk group than in the low-risk group (p<0.0001; Figure 2).

3.4 Validation of the prognostic model in GSE76427

We validated the risk score system in the GSE76427 cohort. A total of 108 patients were divided into two groups (45 high-risk patients and 63 low-risk patients) with a cut-off of 3.4144 for the risk score. Furthermore, the survival curve revealed that the recurrence-free survival was significantly poorer in the high-risk group than in the low-risk group (p=0.011; Figure 3).

In summary, these results indicate that the prognostic model has moderate sensitivity and specificity.

3.5 Association between the prognostic model and the clinical characteristics of patients

In clarifying the correlation between the seven-gene signature and the clinical characteristics of HCC patients, we found that a high risk score was significantly correlated with TNM stage (p<0.001), grade (p=0.001), and AFP (p=0.014) but was not significantly associated with the sex, age, BMI, or Child-Pugh score of patients with HCC (Table ).

In GSE76427, the results showed that the 7-gene signature was not significantly associated with sex, age, BCLC (Barcelona Clinic Liver Cancer) or TNM stage (Table ).

3.6 Independent prognostic role of the prognostic gene signature
Moreover, the results of univariate Cox regression showed that the TNM stage (HR=1.680, \( p<0.001 \)) and our prognostic model (HR=3.607, \( p<0.001 \)) were both independent factors for recurrence-free survival among the 334 TCGA-LIHC patients. However, among the 108 patients in the GSE76427 cohort, TNM stage was not an independent prognostic factor for recurrence-free survival (18). The prognostic model (HR=2.407, \( p=0.014 \)) was also an independent factor for recurrence-free survival. (Figure 4).

3.7 Comparison with the TNM stage model and BCLC model

To compare the accuracy of the prognostic model and the TNM model, we calculated the AUC for 12-month, 15-month, and 18-month recurrence-free survival. In the TCGA-LIHC dataset, the prognostic model's AUCs for 12-month, 15-month, and 18-month RFS were 0.7768, 0.7934, and 0.7529; the TNM model's AUCs for 12-month, 15-month, and 18-month RFS were 0.6884, 0.7026, and 0.6721, respectively (Figure 5). In the GSE76427 dataset, the prognostic model's AUCs for 12-month, 15-month, and 18-month RFS were 0.6159, 0.6118, and 0.6217; the TNM model's AUCs for 12-month, 15-month, and 18-month RFS were 0.6122, 0.6009, and 0.5762, respectively. In addition, the BCLC model's AUCs for 12-month, 15-month, and 18-month RFS were 0.5669, 0.5627, and 0.5684, respectively (Table ii). Overall, our prognostic model showed benefit for predicting recurrence-free survival, which might help doctors with targeted treatment (Figure 6).

3.8 Development of the calibration curve

We calculated the C-index and drew the calibration curves for the 12-, 15- and 18-month survival predictions to evaluate the calibration in the TCGA-LIHC dataset and the GSE76427 dataset. The C-index in the TCGA-LIHC dataset is 0.717 and in the GSE76427 dataset is 0.647, as shown in Figures 7 and 8.

3.9 External validation in an online database

The representative protein expression of SLCO2A1, PPAT, GAS2L3, CD3EAP, ACAT1 were explored in the Human Protein Profiles. Then we explored the TTK, C16orf54, PPAT, CD3EAP, SLCO2A1, ACAT1, and GAS2L3 genes in the CBioPortal for cancer genomics. TTK exhibited the most frequent genetic alterations (3%), of which deep deletion was the most frequent alteration. The second most altered gene was CD3EAP (1.3%), and the most frequent alterations were amplification mutations (Figure 9). The expression of the seven genes in different cancers were shown in Figure 10. In summary, the aberrant expression of these seven genes may explain some of the abnormal expression of these genes.

Discussion

In this study, we developed a risk score of seven genes that had the ability to predict the probability of recurrence-free survival in HCC patients and is more accurate than clinical indicators. Using this model, we can predict which patients with HCC have a higher risk of recurrence, indicating that they need more attention. In the TCGA-LIHC dataset, a total of 1331 genes were found to be associated with the recurrence-free survival of HCC patients. Through KEGG analysis, we found that the 1331 genes were
enriched in signaling pathways such as the cell cycle, homologous recombination, DNA replication, the Fanconi anemia pathway, complement and coagulation cascades, and the T cell receptor signaling pathway. This finding suggests that the 7-gene signature might affect the RFS of HCC through those pathways. Then, we selected the best 7 genes to develop the risk score model: TTK, C16orf105, PPAT, CD3EAP, SLCO2A1, ACAT1, and GAS2L3. Additionally, our study also showed that the TNM staging system was not an accurate indicator for predicting RFS in HCC patients, which was consistent with the results of other studies. According to the prognostic model, we divided the patients into low- and high-risk groups, which had significant differences in recurrence-free survival. This result indicated that the prognostic model could be used as a conventional tool in predicting the RFS of HCC patients.

The prognostic model was validated using another independent dataset, GSE76427. The area under the curve showed the ability of the prognostic model to differentiate the patients’ prognoses; the survival curve represents the survival of the high-risk group, which had a worse prognosis, compared with that of the low-risk group. These findings demonstrate that the prognostic model has the ability to forecast RFS in HCC patients.

Most of the seven genes in our prognostic model have been reported to be involved in cancer. TTK protein levels are different in human liver cancer between liver cancer cells and adjacent noncancerous liver cells (19). The study also tested the utility of TTK-targeted inhibition and demonstrated therapeutic potential in an experimental model of liver cancer in vivo. Furthermore, our study demonstrated its effectiveness and incorporated it into the prognostic model. PPAT, a member of the purine/pyrimidine phosphoribosyl transferase family, regulates pyruvate kinase activity and cell proliferation and invasion and is a biomarker for lung adenocarcinoma. Acetyl-CoA acetyltransferase (ACAT) was recently reported to be elevated in human cancer cell lines (20). ACAT1 has acetyltransferase activity, and it can acetylate pyruvate dehydrogenase (PDH), which affects tumor growth (21).

In other scholars' prognostic analysis of HCC, CD3EAP is also a predictor, suggesting that CD3EAP is an important predictor of HCC prognosis, but the function of CD3EAP is not completely clear (22). The function of GAS2L3 is still unknown; it may be involved in mediating the absorption and clearance of prostaglandins, but it has not reported in liver cancer (13). Moreover, SLCO2A1 and C16orf105 have not been reported in previous HCC studies, indicating that these genes may be potential factors in the treatment of HCC. Understanding the function of these genes may promote the development of HCC treatment.

However, despite the potential substantial clinical significance of our results, this study still has some limitations. One was that, although the calibration curve performance and the AUC value were excellent in the validation group, multicenter clinical application is also needed to evaluate the external utility of the prognostic model (23). Second, only 1331 genes were defined as genes associated with RFS and evaluated for the prognostic model construction. Some important genes could have been excluded before building the prognostic model (24). In addition, signaling pathways are urgently needed to reveal the functions of these genes in HCC.
Conclusions

In conclusion, we developed and validated a prognostic model for predicting the recurrence-free survival probability of HCC patients. The simple prognostic model had the ability to predict RFS, and it could be a useful tool for doctors conducting an evaluation of HCC and selecting treatment plans for HCC patients.

Declarations

Conflict of Interest

The authors declare no competing interests.

Author Contributions Statement

WW, LW and YY conceived and designed the study. WW analyzed the data. XX and YL performed literature searches. WW and LW wrote the paper. QL reviewed and edited the manuscript. All authors read and approved the manuscript.

Funding

This research was supported in part by a grant from the National Natural Science Foundation of China (91180525 to Q.L.).

Acknowledgements

Not applicable.

References

1. Siegel RL, Miller KD and Jemal A: Cancer statistics, 2019. CA Cancer J Clin 69: 7-34, 2019.
2. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J and Jemal A: Global cancer statistics, 2012. CA Cancer J Clin 65: 87-108, 2015.
3. Li G, Xu W, Zhang L, et al.: Development and validation of a CIMP-associated prognostic model for hepatocellular carcinoma. EBioMedicine 47: 128-141, 2019.
4. Chun YH, Kim SU, Park JY, et al.: Prognostic value of the 7th edition of the AJCC staging system as a clinical staging system in patients with hepatocellular carcinoma. Eur J Cancer 47: 2568-2575, 2011.
5. Gu JX, Zhang X, Miao RC, et al.: Six-long non-coding RNA signature predicts recurrence-free survival in hepatocellular carcinoma. World J Gastroenterol 25: 220-232, 2019.
6. Amin MB, Greene FL, Edge SB, et al.: The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA Cancer J Clin 67: 93-99, 2017.
7. Liao X, Yang C, Huang R, et al.: Identification of Potential Prognostic Long Non-Coding RNA Biomarkers for Predicting Survival in Patients with Hepatocellular Carcinoma. Cell Physiol Biochem 48: 1854-1869, 2018.

8. Gao Z, Zhang D, Duan Y, et al.: A five-gene signature predicts overall survival of patients with papillary renal cell carcinoma. PLoS One 14: e0211491, 2019.

9. Chen SH, Wan QS, Zhou D, et al.: A Simple-to-Use Nomogram for Predicting the Survival of Early Hepatocellular Carcinoma Patients. Front Oncol 9: 584, 2019.

10. Yuan SX, Yang F, Yang Y, et al.: Long noncoding RNA associated with microvascular invasion in hepatocellular carcinoma promotes angiogenesis and serves as a predictor for hepatocellular carcinoma patients' poor recurrence-free survival after hepatectomy. Hepatology 56: 2231-2241, 2012.

11. Lee JH, Jung S, Park WS, et al.: Prognostic nomogram of hypoxia-related genes predicting overall survival of colorectal cancer-Analysis of TCGA database. Sci Rep 9: 1803, 2019.

12. Joyce S and Nour AM: Blocking transmembrane219 protein signaling inhibits autophagy and restores normal cell death. PLoS One 14: e0218091, 2019.

13. Wang Y, Sun L, Li Z, et al.: Hepatoid adenocarcinoma of the stomach: a unique subgroup with distinct clinicopathological and molecular features. Gastric Cancer 22: 1183-1192, 2019.

14. Liu GM, Zeng HD, Zhang CY and Xu JW: Identification of a six-gene signature predicting overall survival for hepatocellular carcinoma. Cancer Cell Int 19: 138, 2019.

15. Wang L, Yan Z, He X, Zhang C, Yu H and Lu Q: A 5-gene prognostic nomogram predicting survival probability of glioblastoma patients. Brain Behav 9: e01258, 2019.

16. Luo D, Deng B, Weng M, Luo Z and Nie X: A prognostic 4-IncRNA expression signature for lung squamous cell carcinoma. Artif Cells Nanomed Biotechnol 46: 1207-1214, 2018.

17. Liu GM, Xie WX, Zhang CY and Xu JW: Identification of a four-gene metabolic signature predicting overall survival for hepatocellular carcinoma. J Cell Physiol 2019.

18. Buti S, Karakiewicz PI, Bersanelli M, et al.: Validation of the GRade, Age, Nodes and Tumor (GRANT) score within the Surveillance Epidemiology and End Results (SEER) database: A new tool to predict survival in surgically treated renal cell carcinoma patients. Sci Rep 9: 13218, 2019.

19. Miao R, Wu Y, Zhang H, et al.: Utility of the dual-specificity protein kinase TTK as a therapeutic target for intrahepatic spread of liver cancer. Sci Rep 6: 33121, 2016.

20. Goudarzi A: The recent insights into the function of ACAT1: A possible anti-cancer therapeutic target. Life Sci 232: 116592, 2019.

21. Chen L, Peng T, Luo Y, et al.: ACAT1 and Metabolism-Related Pathways Are Essential for the Progression of Clear Cell Renal Cell Carcinoma (ccRCC), as Determined by Co-expression Network Analysis. Front Oncol 9: 957, 2019.

22. Zhang G, Xue P, Cui S, et al.: Different splicing isoforms of ERCC1 affect the expression of its overlapping genes CD3EAP and PPP1R13L, and indicate a potential application in non-small cell
23. Abdelnabi M, Almaghraby A, Saleh Y and Abd Elsamad S: Hepatocellular carcinoma with a direct right atrial extension in an HCV patient previously treated with direct-acting antiviral therapy: a case report. Egypt Heart J 71: 5, 2019.

24. Abou-Alfa GK, Shi Q, Knox JJ, et al.: Assessment of Treatment With Sorafenib Plus Doxorubicin vs Sorafenib Alone in Patients With Advanced Hepatocellular Carcinoma: Phase 3 CALGB 80802 Randomized Clinical Trial. JAMA Oncol2019.

Tables

Table 1: the best genes predicting recurrence-free survival of hepatocellular carcinoma patients

| Gene symbol | nloglik  | AIC       | Select |
|------------|----------|-----------|--------|
| TTK        | 808.79   | 1619.59   | *      |
| C16orf105  | 797.58   | 1599.16   | *      |
| PPAT       | 791.22   | 1588.43   | *      |
| CD3EAP     | 788.83   | 1585.66   | *      |
| SLC02A1    | 787.91   | 1585.83   | *      |
| ACAT1      | 786.25   | 1584.50   | *      |
| GAS2L3     | 784.91   | 1583.83   | *      |
| SH2D5      | 784.84   | 1585.68   |        |
| ATP8A2     | 784.75   | 1587.50   |        |
| PABPC5     | 784.74   | 1589.49   |        |

Table 2: characteristics of HCC patients in TCGA-LIHC dataset
| Variables | 7-gene signature | The chi-square test | Univariate cox regression |
|-----------|------------------|---------------------|--------------------------|
| Score     | Low-risk (200)   | High-risk (134)     | P value                  | HR          | P value |
| gender    |                  |                     | 0.330                    | 0.975       | 0.879   |
|           | Male             | Male                | 140                      | 87          |
|           | female           | female              | 60                       | 47          |
| Age(years)|                  |                     | 0.785                    | 1.048       | 0.769   |
| <60       | 91               | 63                  |
| ≥60       | 109              | 71                  |
| BMI (kg/m²)|                 |                     | 0.061                    | 0.900       | 0.509   |
| <25       | 91               | 75                  |
| ≥25       | 109              | 59                  |
| TNM       |                  |                     | <0.001                   | 1.680       | <0.001  |
| I         | 123              | 44                  |
| II        | 44               | 39                  |
| III       | 31               | 50                  |
|           | 2                | 1                   |
| grade     |                  |                     | 0.001                    | 1.112       | 0.515   |
| 1+2       | 139              | 68                  |
| 3+4       | 61               | 64                  |
| NA        | 0                | 2                   |
| AFP (ng/ml)|                 |                     | 0.014                    | 0.976       | 0.913   |
| <300      | 134              | 63                  |
| ≥300      | 31               | 30                  |
| NA        | 35               | 41                  |
| Child-Pugh score |              |                     | 0.082                    | 1.202       | 0.581   |
| A         | 136              | 68                  |
| B-C       | 10               | 11                  |
| NA        | 56               | 55                  |
Note: BMI: body mass index; TNM: tumor-node-metastasis; AFP: alpha fetoprotein; grade: tumor grade; BCLC: Barcelona Clinic Liver Cancer NA: not available

**Table 3: characteristics of HCC patients in GEO76427 dataset**

| Variables | 7-gene signature | The chi-square test | Univariate cox regression |
|-----------|------------------|---------------------|--------------------------|
|           | Low-risk (63)    | High-risk (45)      |                           |
| Score     |                  | P value 2.047       | HR 2.047 P value 0.014    |
| gender    |                  | 0.374               | 0.609 0.208              |
| Male      | 11               | 11                  |
| female    | 52               | 34                  |
| Age(years)|                  | 0.161               | 1.048 0.769              |
| <60       | 21               | 21                  |
| ≥60       | 42               | 24                  |
| TNM       |                  | 0.877               | 1.267 0.191              |
| I         | 36               | 16                  |
| II        | 15               | 19                  |
| III       | 10               | 9                   |
|          | 2                | 1                   |
| BCLC      |                  | 0.877               | 1.112 0.515              |
| 0         | 2                | 2                   |
| A         | 41               | 30                  |
| B         | 16               | 9                   |
| C         | 4                | 4                   |

Note: BMI: body mass index; TNM: tumor-node-metastasis; AFP: alpha fetoprotein; grade: tumor grade; BCLC: Barcelona Clinic Liver Cancer NA: not available

**Table 4: comparison of the prognostic model with the TNM and BCLC model**
| model      | TNM model | BCLC model | Prognostic model |
|------------|-----------|------------|------------------|
| TCGA-LIHC  |           |            |                  |
| 12-month AUC | 0.6884 (0.6272-0.7496) |            | 0.7768 (0.7180-0.8356) |
| 15-month AUC | 0.7026 (0.6416-0.7636) |            | 0.7934 (0.7367-0.8501) |
| 18-month AUC | 0.6721 (0.6086-0.7356) |            | 0.7529 (0.6905-0.8153) |
| GSE76427   |           |            |                  |
| 12-month AUC | 0.6122 (0.4733-0.7511) | 0.5669 (0.4408-0.6931) | 0.6159 (0.4596-0.7722) |
| 15-month AUC | 0.6009 (0.4692-0.7326) | 0.5627 (0.4400-0.6853) | 0.6118 (0.4679-0.7575) |
| 18-month AUC | 0.5762 (0.4453-0.7072) | 0.5684 (0.4458-0.6910) | 0.6217 (0.4828-0.7605) |

*Note: AUC: area under the curve; CI, confidence interval.*

**Figures**
Figure 1

(A): The differentially expressed genes and pathway enrichment of GO slim summary. Red, blue, and green bars represent the biological process, cellular component, and molecular function categories, respectively. The height of the bar represents the number of differentially expressed genes observed in that category. (B): The top 10 pathways of genes associated with recurrence-free survival.
Figure 2

Analysis of the seven-gene signature in HCC: (A): risk score of each patient; (B): the RFS time and the RFS status of the HCC patients; (C): the expression levels of TTK, C16orf105, PPAT, CD3EAP, SLCO2A1, ACAT1 and GAS2L3 in the signature; Kaplan-Meier analysis in the TCGA dataset.
Figure 3

Analysis of the seven-gene signature in HCC. (A): risk score of each patient; (B): the RFS time and the RFS status of the HCC patients; (C): the expression levels of TTK, C16orf105, PPAT, CD3EAP, SLCO2A1, ACAT1 and GAS2L3 in the signature; Kaplan-Meier analysis in the GSE76427 dataset.
### A

| Factor    | Sample | Hazard Ratio (95% CI) | P-value |
|-----------|--------|-----------------------|---------|
| TNM       | (N=334)| 1.380 (1.111 - 1.664) | 0.00279 ** |
| Score     | (N=334)| 2.526 (1.914 - 3.334) | <0.001 *** |
| Gender    | (N=334)| 1.318 (0.918 - 1.893) | 0.13434 |
| Age       | (N=334)| 1.001 (0.986 - 1.015) | 0.91679 |
| BMI       | (N=334)| 0.999 (0.973 - 1.025) | 0.93857 |

### B

| Factor    | Sample | Hazard Ratio (95% CI) | P-value |
|-----------|--------|-----------------------|---------|
| TNM       | (N=107)| 0.937 (0.570 - 1.538) | 0.7969  |
| Score     | (N=107)| 2.360 (1.055 - 5.277) | 0.03643 * |
| Age       | (N=107)| 1.010 (0.985 - 1.035) | 0.44331 |
| BCLC      | (N=107)| 1.430 (0.626 - 2.473) | 0.20111 |
| Gender    | (N=107)| 1.258 (0.508 - 3.113) | 0.61936 |
Figure 4

Multivariate Cox regression analysis. (A): Multivariate Cox regression analysis of the TCGA dataset. (B): Multivariate Cox regression analysis of the GSE76427 dataset.
Figure 5

(A): The prognostic model's AUC for 12-, 15-, and 18-month RFS in the TCGA-LIHC dataset. (B): The TNM stage model's AUC for 12-, 15-, and 18-month RFS in the TCGA-LIHC dataset.
Figure 6

(A): The prognostic model's AUC for 12-, 15-, and 18- month RFS in the GSE76427 dataset. (B): The TNM stage model's AUC for 12-, 15-, and 18-month RFS in the GSE76427 dataset. (C): The BCLC model's AUC for 12-, 15-, and 18-month RFS in the GSE76427 dataset.
Figure 7

Calibration curve for the 12-month, 15-month, and 18-month periods in the TCGA-LIHC dataset. (A) The prognostic model was used to generate a calibration curve for the 12-month RFS prediction. (B) The prognostic model was used to generate a calibration curve for the 15-month RFS prediction. (C) The prognostic model was used to generate a calibration curve for the 18-month RFS prediction.
Figure 8

Calibration curve for the 12-month, 15-month, and 18-month periods in the GSE76427 dataset. (A) The prognostic model was used to generate a calibration curve for the 12-month RFS prediction. (B) The prognostic model was used to generate a calibration curve for the 15-month RFS prediction. (C) The prognostic model was used to generate a calibration curve for the 18-month RFS prediction.
Figure 9

External validation in an online database (A): The representative protein expression of the seven genes in HCC and normal liver tissue (B): Genetic alterations of the seven genes.
Figure 10

The expression of the seven genes in different cancers