Identification of a Seven-Gene Signature and Establishment of a Nomogram Predicting Overall Survival in Head and Neck Squamous Cell Carcinoma

Haige Zheng
Jinan University First Affiliated Hospital
Xiangkun Wu
Guangzhou Medical College First Affiliated Hospital
Huixian Liu
Jinan University First Affiliated Hospital
Yumin Lu
Guangxi Medical University First Affiliated Hospital
Hengguo Li (lhgjnu@263.net)
Jinan University First Affiliated Hospital https://orcid.org/0000-0002-6770-0162

Primary research

Keywords: overall survival, immune infiltration, risk score model, nomogram, head and neck squamous cell carcinoma

Posted Date: September 8th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-70821/v1

Abstract

Background: Head and neck squamous cell carcinoma (HNSCC) is a highly heterogeneous tumor with high incidence and poor prognosis. Therefore, effective predictive models are needed to evaluate patient outcomes and optimize treatment.

Methods: Ten gene microarray datasets were obtained from the gene expression omnibus (GEO) database. Level 3 mRNA expression and clinical data were obtained in The Cancer Genome Atlas (TCGA) database. We identified highly robust differentially-expressed genes (DEGs) between HNSCC and normal tissue in nine GEO and TCGA datasets using Robust Rank Aggregation (RRA) method. Univariate Cox regression analysis and lasso Cox regression analysis were performed to identify DEGs related to the Overall-survival (OS) and to construct a prognostic gene signature. External validation was performed using GSE65858. Moreover, gene set enrichment analyses (GSEA) analysis was used to analyze significantly rich pathways in high-risk and low-risk groups, and tumor immunoassays were used to clarify immune correlation of the prognostic gene. Finally, integrate multiple forecast indicators were used to build a nomogram using the TCGA-HNSCC dataset. Kaplan–Meier analysis, receiver operating characteristic (ROC), a calibration plot, Harrell's concordance index (C-index), and decision curve analysis (DCA) were used to test the predictive capability of the seven genetic signals and the nomogram.

Results: A novel seven-gene signature (including SLURP1, SCARA5, CLDN10, MYH11, CXCL13, HLF, and ITGA3) was established to predict overall survival in HNSCC patients. ROC curve performed well in the training and validation data sets. Kaplan–Meier analysis demonstrated that low-risk groups had a longer survival time. The nomogram containing seven genetic markers and clinical prognostic factors was a good predictor of HNSCC survival and showed a certain net clinical benefit through the DCA curve. Further research demonstrated that the infiltration degree of CD8 + T cells, B cells, neutrophils, and NK cells were significantly lower in the high-risk group.

Conclusion: Our analysis established a seven-gene model and nomogram to accurately predict the prognosis status of HNSCC patients, immune relevance was also described, which may provide a new possibility for individual treatment and medical decision-making.

Background

Head and neck squamous cell carcinoma (HNSCC) ranks sixth among cancer-related deaths worldwide, and includes tumors originating from the mouth, oropharynx, nasopharynx, hypopharynx, larynx, and neck [1]. Head and neck cancer is one of the main tumors that threatens human health, with over 600,000 new patients and over 350,000 deaths worldwide every year. Despite the substantial improvement in surgery, radiation therapy and chemotherapy in the past 30 decades, early diagnosed patients may achieve good results with surgery or radiation therapy, but survival rates for patients with advanced cancer is only 34.9%, and the median survival time of individuals has been reported to be 6 to 9 months [2, 3]. Thus, there is an urgent need to identify effective HNSCC predictive biomarkers for an accurate and

effective evaluation of a patient's disease status to improve prognosis and reduce mortality. As we know, the immune system is an important factor in tumorigenesis. As an emerging effective anti-tumor therapy approach, immunotherapy represented mainly by the programmed cell death 1 (PD-1)/programmed cell death ligand 1 (PD-L1) pathway has shown great therapeutic potential in many tumor types [4, 5]. Thus, it is important to study the tumor-immunity correlation and provide additional treatment options for HNSCC patients.

With the development of genome-sequencing technology, there is growing evidence that prognostic gene markers can predict head and neck cancer overall survival (OS). For example, Shen et al. (2017) have identified a 7-gene signature that associated with the survival of patients with head and neck carcinoma [6]. Liu et al (2018) selected 5-lncRNA and constructed a prognostic score model for prediction of OS [7]. She et al (2020) predicted the prognosis of HNSCC using immune-related genes[8].

In our study, we use the two large public databases—The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) to collect data from different microarray platforms. For the data of these two databases, we have corrected the batch effect using the combat function of the 'sva' package in R software [9]. We identified highly robust differentially-expressed genes (DEGs) between HNSCC and normal tissue in nine GEO and TCGA datasets using Robust Rank Aggregation (RRA) method. Uni-variate, lasso and multi-variate Cox analysis were conducted to screen the DEGs associated with OS and to establish a seven-gene prognostic model using the TCGA-FPKM RNA-seq dataset. The prognostic model was validated using GSE65858 dataset. In addition, we also conducted a subgroup analysis of head and neck cancer patients in the TCGA cohort, and showed significant statistical significance in different grading and staging groups, further verifying the credibility of the model. We established a nomogram combining prognostic genetic characteristics and several reliable clinical parameters to predict patient survival. The tumor-immune correlation and its molecular mechanism as a prognostic gene signal and its potential to guide immunotherapy were also studied. Overall, our seven-gene model and nomogram have strong diagnostic performance in terms of predicting the prognostic status of patients and is beneficial to guide treatment and management of HNSCC patients.

Materials And Methods

Source of Data

The 10 gene chip datasets of HNSCC were derived from the GEO database. Nine gene microarray data sets (GSE6631, GSE13601, GSE30784, GSE31056, GSE33205, GSE37991, GSE51985, GSE59102, and GSE138206) were used for DEGs analysis. GSE65858 with 270 tumor tissues and corresponding clinical information were used to verify the prognostic model. We directly downloaded the standardized matrix data from GEO platform and match the probe to the genetic symbol using the manufacturer's annotation file. When a single gene symbol corresponded to several probes, the method of median sequence value was adopted.

HNSCC patient's normalized HTSeq-counts and HTSeq-fpkm data along with corresponding clinical data were obtained from TCGA databases, which contained data from 502 cancer samples and 44 normal samples. The ensemble ID was converted to a gene symbol through the annotation file. HTSeq-counts data were used to screen for DEGs. Finally, the HTSeq-fpkm data were applied to establish a prognostic model, after removing 10 samples with survival time less than one month, ultimately, 491 samples were incorporated into the modeling analysis.

Comprehensive analysis of data sets and DEGs screening

Background correction, standardization and acquisition of DEGs were performed using the nine datasets from GEO using Impute and Limma packages from R software [10]. The edgeR package in R was used for DEG analysis of the TCGA queue. The RRA method was used to analyze the DEGs identified from the nine GEO datasets using R package 'RobustRankAggreg'. All genes with | LogFC | >1.0, FDR < 0.05 and p < 0.05 were selected as DEGs. More reliable HNSCC-specific DEGs were obtained by the intersection of nine GEO and TCGA dataset results. The 'sva' package in R was run to eliminate or reduce batch effects between different data sets.

Functional analysis of genes

GO and KEGG enrichment analyses were applied to identify potential biological processes, cellular components, molecular functions, as well as KEGG pathway terms. Important signal pathways related to DEGs were identified using the cluster profiler R package.

Identification of the Survival-Related Prognostic Model

Uni-variate, the least absolute shrinkage and selection operator (Lasso) and Multi-variate Cox analyses were conducted to obtain prognostic genes significantly correlated with OS in 491 HNSCC patients with survival time > 30 days [11]. First, Uni-variate analysis was used to identify genes associated with prognosis. Next, Lasso analysis was conducted to decrease the amount of mRNA to obtain a more meaningful prognostic gene. Ultimately, a seven-gene signature of HNSCC was established based on a linear combination of the regression coefficient derived from the Lasso Cox regression model (β) multiplied with its expression level. the risk model = (β* SLURP1) + (β * SCARA5) + (β * CLDN10) + (β* MYH11) + (β * CXCL13) + (β * HLF) + (β * ITGA3). Using the median value as the cut-off, 491 HNSCC patients were separated into high- and low-risk groups. Kaplan–Meier analysis and receiver operating characteristic curve (ROC) curve analyses were performed to evaluate the accuracy and sensitivity of the model using the R package 'survival' and 'survivalROC', respectively [12]. The C-index was also used to measure discrimination between the predicted value and the real value of the Cox model.

GSE65858 cohort for external validation

The GSE65858 dataset from the GEO database was used for external validation. Risk scores for each included patient were calculated using the same prognostic gene signature. Next, the predictive capability of the gene model was also tested based on the Kaplan–Meier curve, the ROC curve, and the C-index.

Independence of prognostic genes

Only 378 patients with complete clinicopathological information including age, sex, survival time, survival status, grade, American Joint Committee on Cancer (AJCC) staging, T staging, and N staging were included in our subsequent analysis. Since most patients lacked clinical information on M staging, M staging was not included in this study. Uni-variate and multi-variate analyses were performed to test whether the prognostic risk score could be independent of other clinical variables (such as sex, age, tumor grade, and AJCC staging).

Establish and verify a prognostic nomogram

A nomogram can be used in combination with multiple indicators to diagnose or predict disease onset or progression. We constructed the nomogram using age, sex, tumor grade, staging, and risk score to predict the 1-, 3-, and 5-year OS rate of head and neck cancer. A nomogram is generally verified by two indicators, namely, discrimination and calibration. Thus, the nomogram's calibration curves were drawn to compare actual survival rate and the predicted survival rate, with the ordinate indicating the actual survival rate, and the abscissa indicating the predicted survival rate. While, the C-index was used to determine the discrimination of the nomogram using a bootstrap approach, repeated 1000 times. Kaplan–Meier analysis, AUC of the ROC curve, and DCA curve were also applied to assess the predictability of the prognostic nomogram.

DCA was used to predict clinical outcome variables and was performed to quantify the clinical utility of the nomogram and to determine its clinical usefulness [13]. The nomogram was used to calculate the total score and patients were separated into two groups. The Kaplan-Meier method was used to plot the survival curves for different risk assessment groups.

Enrichment and Tumor Immunity Analyses

In order to clarify the potential pathobiological mechanism underlying the gene signature, GSEA analysis was conducted to identify rich GO terms and KEGG pathways between high-risk and low-risk groups, including the KEGG pathway in C2, GO term in C5, and oncogenic signatures of gene sets in C6.

The stromal, immunity, and estimate scores of each head and neck cancer sample were calculated using the ESTIMATE algorithm. Single-sample Gene set enrichment analysis (ssGSEA) approaches can be used to quantify tumor-infiltrating immune cells using the R package ‘GSVA’ [14].The infiltration abundance of 29 immune cells (including B cells, CD4 + T cells, CD8 + T cells, macrophages, neutrophils, dendritic cells, etc.) in each head and neck cancer sample were calculated. Finally, the correlation between stroma, immunity, estimated score and risk score, and the difference in immune cell infiltration between the high-risk and low-risk groups was analyzed.

Statistical analysis

R software (version 3.6.2; http://www.Rproject.org) and GraphPad Prism (v. 8.0) were used for statistical analysis and to prepare figures. The unpaired t-test was used to estimate the statistical significance of two groups of normally distributed variables. Kaplan-Meier curve analysis was used for survival analysis, the logarithmic rank test was conducted to assess the survival risk in high-risk and low-risk groups.

Results

Screening for differentially expressed genes

Our study follows the flowchart shown in (Fig. 1**)**. The details of the 10 GEO data sets are shown in Table 1. From the GSE6631, GSE13601, GSE30784, GSE33205, GSE37991, GSE51985, GSE59102 and GSE138206 data sets, 145, 1140, 1582, 2082, 420, 1848, 620, 2758, and 901 differential genes were selected, respectively. After a comprehensive study of nine GSE datasets based on the RRA means, 193 DEGs were identified, including 69 upregulated and 124 downregulated genes. The top 20 upregulated and downregulated DEGs found in the comprehensive analysis were shown in the volcano diagram (Fig. 2a). For the TCGA-HNSCC dataset, 3033 DEGs (1359 up-regulated and 1673 down-regulated) were selected, the representative heat map of DEG showed that DEG could effectively distinguish between tumor and normal tissue (Fig. 2c). Finally, after the intersection of GEO and TCGA results, 255 reliable DEGs were identified (Fig. 2b).

Functional analysis of genes

GO enriched sets ($p < 0.0005$) were analyzed relative to genes involved in serine hydrolase activity, cytokine activity, muscle structure component, serine type peptidase activity, serine-type endopeptidase activity, collagen binding, CXCR chemokine receptor binding, receptor ligand activity, actin binding, heparin binding, extracellular matrix structural constituent (Fig. 3a). The KEGG pathway analysis of the differential genes mainly involved 3 pathways ($P < 0.05$): tyrosine, retinol metabolism, and drug metabolism cytochromes (Fig. 3b).

Establishment of a prognostic model

A total of 491 TCGA patients with a survival greater than one month were incorporated into this model analysis. The basic clinical parameters of the included patients are shown in Table 2. Univariate analysis identified 166 genes that significantly associated with OS. Through layer upon layer screening, seven genes finally obtained through multi-variate regression analysis were selected to build a predictive model. The formula for the risk score is as follows: risk model = (-0.00173 * SLURP1) + (-0.21582 * SCARA5) + (-0.0458* CLDN10) + (0.030114* MYH11) + (-0.00725 * CXCL13) + (-0.0839 * HLF) + (0.005721* ITGA3). Next, Kaplan–Meier analysis disclosed a significantly better prognosis in the low-risk group ($p < 0.0001$) (Fig. 4a). In addition, the ROC and C-indexes also evaluated the good predictive power of the seven-gene model. The time-dependent AUCs of risk scores were 0.666 at 3-year and 0.739 at 5-year OS (Fig. 4c). The C-index of the risk score was 0.623 (95% CI, 0.538–0.708).

External validation set and performance

The GSE65858 dataset with 267 HNSCC patients were selected to validate the seven-genes signature. We calculated risk score with the same risk formula and divided the patients in the GSE65858 dataset into high-risk and low-risk groups. The model established in this study was also meaningful in the validation set, the results show that significant differences in OS between high- and low-risk groups ($p < 0.0001$)

(Fig. 4b). The predicted AUC for the 3- and 5-year OS were 0.706 and 0.618, respectively, indicating good diagnostic performance (Fig. 4d). The C-index of the gene signature was 0.641(95% CI, 0.528–0.754).

Subgroup analysis on the seven-gene signature

We performed subgroup analysis on the TCGA data set to further verify the validity of the seven gene model and the results revealed that the model also had certain diagnostic performance in different subgroups (Fig. 5a-d). They were stage T1-2 patients ($p < 0.001$), T3-4 patients ($p < 0.001$), N0-1 patients ($p < 0.001$), N2-3 patients ($p = 0.004$), G1-2 patients ($p < 0.001$), G3-4 patients ($p < 0.001$), and stage III-IV ($p < 0.001$).

Independent predictability of prognostic models

Uni-variate and multi-variate Cox regression were performed to assess the independent predictive capacity of the seven gene prognostic model based on 378 HNSCC patients with available clinicopathological parameters. The p-value for grade, T staging and N staging was < 0.05 in univariate Cox regression analysis. (Fig. 6a). Next, multi-variate Cox regression analysis demonstrated that N staging and the risk score can independently predict OS for head and neck cancer (Fig. 6b). The same approach was used for the validation set, and results also indicated that the risk score was an independent prognostic factor (Fig. 6c-d).

Establishment and verification of the nomogram

We constructed a nomogram to predict 1-, 3-, and 5-year OS of HNSCC patients based on the TCGA-HNSCC cohort using independent prognostic factors including age, sex, grade, AJCC staging, T staging, N staging and risk score (Fig. 7a). The C-index of nomogram was 0.678 (95% CI, 0.582–0.774). The AUCs for the 3- and 5-year OS were 0.723 and 0.684, respectively (Fig. 7c). The calibration curve intuitively demonstrated that predicted value of the nomogram was close to the true value and had reliable prediction performance (Fig. 7b). The group with a lower risk score had a significantly better prognosis (Fig. 7d). The DCA results further suggested that the nomogram was more clinically useful than the gene-based risk model or staging system alone (Fig. 7e).

Gene Set Enrichment Analysis (GSEA)

To investigate the possible underlying pathobiological mechanisms of the prognostic genes, GSEA was used to analyze important enrichment pathways in different risk groups of TCGA.A total of 491 TCGA-HNSCC patients were enrolled in the GSEA comparison high-risk and low-risk cohort. The results indicated that 1 KEGG pathway, 2 GO terms, and 3 oncological signatures were enriched in the high-risk group. The identified enriched KEGG pathway and GO terminology mainly involved the proteasome pathway, proteasome accessory and endopeptidase signal transduction pathway, and three oncological signatures included early serum response (CSR), glioma-associated oncogene (GLI1) and B cell-specific moloney murine leukemia virus integration site (BMI1) (Fig. 8a-c).

Tumor immune mechanism of prognostic genes

Expression data obtained from the TCGA HNSCC dataset were used to calculate the matrix, immunity, and estimated scores using an estimation algorithm. The results revealed that the immunity scores and estimated scores of the high-risk group were significantly lower, indicating that there was less infiltration of immune cells in the cancerous tissue ($p < 0.05$; Fig. 8d). The difference in the matrix score between the high-risk group and the low-risk group was not statistically significant ($p > 0.05$).

To further study the relationship between the seven prognostic genes and immune infiltration in head and neck carcinoma, we used ssGSEA method to assess the correlation between these genes and the level of immune cell infiltration, including CD4 + T, CD8 + T, B cell, dendritic cells, neutrophils, macrophages and other immune cells. The results suggested that there was a relatively higher level of CD8 + T cells, B cells, neutrophils, and NK cell infiltration in the low risk group (Fig. 8e).

Discussion

HNSCC is highly heterogeneous tumor with very poor prognosis and its incidence is increasing every year. The five-year survival rate is only 40–50% [2]. With the expanding research in the field of gene sequencing technology, gene expression profiling has attracted increasing attention, and has been used to identify prognostic markers associated with the heterogeneity of tumors. Therefore, screening for prognostic molecular markers that fully reflect the tumor biological characteristics may provide clinicians with novel tools to treat HNSCC patients. In this study, we screened DEGs in head and neck cancer tissues and adjacent non-tumor tissues, and performed Uni-variate, Lasso, and Multi-variate Cox analysis to establish a prognostic risk model for HNSCC. We identified seven DEGs: SLURP1, SCARA5, CLDN10, MYH11, CXCL13, HLF, and ITGA3, and among them MYH11 and ITGA3 were defined risk factors and the remainder as protective factors.

Five of the seven gene signatures have previously been linked to head and neck cancer. SCARA5, a member of scavenger receptor Type A family, is able to bind lipopolysaccharide, bacteria, and nucleotides to charge residues in the cysteine domain. The reduction of SCARA5 expression can promote invasiveness and proliferation of oral tumor cells, and the expression of SCARA5 was found to be down-regulated in oral tumors [15]. In addition, the SCARA5 receptor can recognize and engulf pathogens, and then transmit intracellular signals to generate an immune response. When the number of receptors is reduced, the cell is unable to mount an immune-induced defense. SCARA5 expression is usually downregulated in hepatocellular carcinoma due to high promoter methylation and allele imbalance [16].

MYH11, a smooth muscle myosin, may be involved in cell migration and adhesion, intracellular transport, and signal transduction, and as a contractile protein, it converts the chemical energy hydrolyzed by adenosine triphosphate (ATP) into mechanical energy [17]. MYH11 is closely related to the survival of HNSCC, acute myeloid leukemia, colorectal cancer, bladder cancer, and other tumors[18–20].

CXCL13 is an independent and cloned B lymphocyte chemokine named Angie, which is an antimicrobial peptide. CXC chemokines are highly expressed in spleen follicles, lymph nodes, and Peyer's patch. It promotes B-cell migration by stimulating chemotaxis into cells expressing Burkitt lymphoma receptor

1(Blr-1) and calcium influx [21, 22]. Previous studies have shown that CXCL13 is associated with the prognosis of various cancers. For example, oral squamous cell carcinoma, breast cancer, and prostate carcinoma [23–25].

Hlf-encoded proteins are a member of the proline and acid-rich (PAR) protein family, which activate transcription by forming homotypic or heterotypic dimers with other PAR family members and binding specific promoter elements. It has been reported that, HLF is closely related to the prognosis in liver cancer, gastric cancer, and lung cancer among others [26–28].

The ITGA3-encoded protein belongs to the integrin family. Integrin is an isomeric membrane protein composed of chains and chains, which acts as a cell surface adhesion molecule. the down-regulation of ITGA3 reduces the phosphorylation of AKT, ERK1/2, and FAK in SAS cells and significantly inhibited migration of cancer cells and invasion of HNSCC cells. High expression of ITGA3 predicted poor survival in patients with HNSCC [29]. Moreover, several studies have confirmed that ITGA3 is a marker of glioblastoma, pancreatic cancer, and thyroid cancer [30–32].

The role of SLURP1 and CLDN10 in head and neck cancer has not been described. The protein encoded by SLURP1, a member of the Ly6/uPAR family, has anti-tumor activity [33]. SLURP1 is related to the occurrence and development of pancreatic cancer. First, it can not only reduced invasion of cancer cells by controlling AKT, ERK, and NF-kB signaling, but also attenuates nicotine-mediated migration and invasion possibly through competing binding sites [34]. Furthermore, mutations in SLURP1 can increase the incidence of melanoma and mucosal skin cancer [35].

Studies have shown that CLDN10 is highly expressed in thyroid papillary carcinoma, and can affect cell proliferation, migration, and invasion in vitro; further, it plays the role of tumor promoter, and the up-regulation of CLDN10 is related to lymph node metastasis [36]. In contrast, low expression of CLDN10 indicated poor prognosis in lung cancer patients. The expression of CLDN10 was negatively correlated with the expression of c-fos. c-fos is considered a recognized oncogene, CLDN10 may control the invasion and metastasis of lung cancer cells by inhibiting the c-fos pathway [37]. In our study, the down-regulation of CLDN10 was associated with poor survival rates for HNSCC, which was consistent with the latter view. The mechanism of SLURP1 and CLDN10 in head and neck tumors deserves further study.

Many studies have shown that the immune system acts to control tumor growth and progression, and the prognosis of the tumor is related to lymphocyte infiltration. As an emerging anti-tumor force, immunotherapy has shown great therapeutic potential in many cancers, HNSCC is no exception. In addition, HNSCC patients with highly infiltrated CD8 T cells have a better prognosis, especially in patients who are HPV positive [38].

Our research has many advantages. First, the number of samples was much larger than in previously published studies, we integrated 9 GEO data sets and TCGA data sets, which provided full verification of the signature and made the prognostic gene model more reliable. Second, we identified several previously under-evaluated genes with unknown functions. Furthermore, we also analyzed the relevance between

prognostic gene expression and the immune microenvironment. Nonetheless, the present study also includes certain limitations. First, we originally wanted to include more datasets to better validate our biomarkers. However, due to different platforms, there is a possibility of sampling deviation in gene expression values, even though we have attempted to corrected for this potential error. Second, the results obtained by bioinformatics analysis alone are not sufficient and need to be confirmed by experimental verification. Therefore, further identification of prognostic biomarkers requires experimental studies with larger samples and experimental validation.

Conclusion

In conclusion, our results suggested that the seven genetic markers were closely related to the prognosis and progression of HNSCC. The predicted nomogram is reliable in predicting the OS of HNSCC and may be conducive to individualized treatment, management, and medical decision making.

Abbreviations

HNSCC: head and neck squamous cell carcinoma;

GEO: gene expression omnibus;

TCGA: The Cancer Genome Atlas;

DEGs: differentially-expressed genes;

RRA: Robust Rank Aggregation;

OS: overall survival;

GSEA: gene set enrichment analyses;

ROC: receiver operating characteristic;

C-index: Harrell’s concordance index;

DCA: decision curve analysis;

PD-1: programmed cell death 1;

PD-L1: programmed cell death ligand 1;

FC: fold change;

FDR: false discovery rate;

GO: gene ontology;

KEGG: Kyoto Encyclopedia of Genes and Genomes;

ssGSEA: single-sample Gene set enrichment analysis;

LASSO: least absolute shrinkage and selection operator;

ATP: adenosine triphosphate;

Blr-1: Burkitt lymphoma receptor 1;

PAR: proline and acid-rich.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The TCGA-HNSCC dataset analyzed in this study can be obtained from the TCGA database (https://cancergenome.nih.gov/). Ten GEO datasets (GSE6631, GSE13601, GSE30784, GSE31056, GSE33205, GSE37991, GSE51985, GSE59102, GSE138206 and GSE65858) analyzed in this study can be obtained from the GEO database (https://www.ncbi.nlm.nih.gov/geo/)

Competing interests

Authors have no conflict of interest to declare.

Funding

Not applicable

Authors' contributions

HZ and XW: study design and the drafting of the manuscript, HZ, XW and HXL: data acquisition, analysis and manuscript modification, HZ, YL and HGL: the research oversight and manuscript reviews.

Acknowledgments

The authors would like to express gratitude to the editors and reviewers for reviewing this research and the hard work behind it.

References

1. Siegel RL, Miller KD, Jemal A: **Cancer statistics, 2016**. *CA Cancer J Clin* 2016, **66**(1):7-30.
2. Leemans CR, Braakhuis BJ, Brakenhoff RH: **The molecular biology of head and neck cancer**. *Nat Rev Cancer* 2011, **11**(1):9-22.
3. Gildener-Leapman N, Ferris RL, Bauman JE: **Promising systemic immunotherapies in head and neck squamous cell carcinoma**. *Oral Oncol* 2013, **49**(12):1089-1096.
4. Lin W, Chen M, Hong L, Zhao H, Chen Q: **Crosstalk Between PD-1/PD-L1 Blockade and Its Combinatorial Therapies in Tumor Immune Microenvironment: A Focus on HNSCC**. *Front Oncol* 2018, **8**.
5. Ohaegbulam KC, Assal A, Lazar-Molnar E, Yao Y, Zang X: **Human cancer immunotherapy with antibodies to the PD-1 and PD-L1 pathway**. *Trends Mol Med* 2015, **21**(1):24-33.
6. Shen S, Bai J, Wei Y, Wang G, Li Q, Zhang R, Duan W, Yang S, Du M, Zhao Y *et al*: **A seven-gene prognostic signature for rapid determination of head and neck squamous cell carcinoma survival**. *Oncol Rep* 2017, **38**(6):3403-3411.
7. Liu G, Zheng J, Zhuang L, Lv Y, Zhu G, Pi L, Wang J, Chen C, Li Z, Liu J *et al*: **A Prognostic 5-lncRNA Expression Signature for Head and Neck Squamous Cell Carcinoma**. *Sci Rep* 2018, **8**(1):15250.
8. She Y, Kong X, Ge Y, Yin P, Liu Z, Chen J, Gao F, Fang S: **Immune-related gene signature for predicting the prognosis of head and neck squamous cell carcinoma**. *Cancer Cell Int* 2020, **20**.
9. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD: **The sva package for removing batch effects and other unwanted variation in high-throughput experiments**. *Bioinformatics* 2012, **28**(6):882-883.
10. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies**. *Nucleic Acids Res* 2015, **43**(7):e47.
11. Tibshirani R: **The lasso method for variable selection in the Cox model**. *Stat Med* 1997, **16**(4):385-395.
12. Heagerty PJ, Lumley T, Pepe MS: **Time-dependent ROC curves for censored survival data and a diagnostic marker**. *Biometrics* 2000, **56**(2):337-344.
13. Vickers AJ, Elkin EB: **Decision curve analysis: a novel method for evaluating prediction models**. *Med Decis Making* 2006, **26**(6):565-574.
14. Li Y, Lu Z, Che Y, Wang J, Sun S, Huang J, Mao S, Lei Y, Chen Z, He J: **Immune signature profiling identified predictive and prognostic factors for esophageal squamous cell carcinoma**. *Oncoimmunology* 2017, **6**(11).
15. Liu H, Hu J, Wei R, Zhou L, Pan H, Zhu H, Huang M, Luo J, Xu W: **SPAG5 promotes hepatocellular carcinoma progression by downregulating SCARA5 through modifying beta-catenin degradation**. *J Exp Clin Cancer Res* 2018, **37**(1):229.
16. Huang J, Zheng DL, Qin FS, Cheng N, Chen H, Wan BB, Wang YP, Xiao HS, Han ZG: **Genetic and epigenetic silencing of SCARA5 may contribute to human hepatocellular carcinoma by activating FAK signaling**. *J Clin Invest* 2010, **120**(1):223-241.

17. Islam T, Rahman R, Gov E, Turanli B, Gulfidan G, Haque A, Arga KY, Haque Mollah N: **Drug Targeting and Biomarkers in Head and Neck Cancers: Insights from Systems Biology Analyses**. *OMICS* 2018, **22**(6):422-436.
18. Zhao B, Baloch Z, Ma Y, Wan Z, Huo Y, Li F, Zhao Y: **Identification of Potential Key Genes and Pathways in Early-Onset Colorectal Cancer Through Bioinformatics Analysis**. *Cancer Control* 2019, **26**(1).
19. Faber ZJ, Chen X, Gedman AL, Boggs K, Cheng J, Ma J, Radtke I, Chao JR, Walsh MP, Song G *et al*: **The Genomic Landscape of Core-Binding Factor Acute Myeloid Leukemias**. *Nat Genet* 2016, **48**(12):1551-1556.
20. Hu J, Zhou L, Song Z, Xiong M, Zhang Y, Yang Y, Chen K, Chen Z: **The identification of new biomarkers for bladder cancer: A study based on TCGA and GEO datasets**. *J Cell Physiol* 2019.
21. Liu X, Asokan SB, Bear JE, Haugh JM: **Quantitative analysis of B-lymphocyte migration directed by CXCL13**. *Integr Biol (Camb)* 2016, **8**(8):894-903.
22. Barstad B, Tveitnes D, Dalen I, Noraas S, Ask IS, Bosse FJ, Oymar K: **The B-lymphocyte chemokine CXCL13 in the cerebrospinal fluid of children with Lyme neuroborreliosis: associations with clinical and laboratory variables**. *Infect Dis (Lond)* 2019, **51**(11-12):856-863.
23. Sambandam Y, Sundaram K, Liu A, Kirkwood KL, Ries WL, Reddy SV: **CXCL13 Activation of c-Myc Induce RANK Ligand Expression in Stromal/Preosteoblast Cells in the Oral Squamous Cell Carcinoma Tumor-Bone Microenvironment**. *Oncogene* 2013, **32**(1):97-105.
24. Chen L, Huang Z, Yao G, Lyu X, Li J, Hu X, Cai Y, Li W, Li X, Ye C: **The expression of CXCL13 and its relation to unfavorable clinical characteristics in young breast cancer**. *J Transl Med* 2015, **13**.
25. El-Haibi CP, Singh R, Sharma PK, Singh S, Lillard JW: **CXCL13 mediates prostate cancer cell proliferation through JNK signalling and invasion through ERK activation**. *Cell Prolif* 2011, **44**(4):311-319.
26. Hou JY, Wang YG, Ma SJ, Yang BY, Li QP: **Identification of a prognostic 5-Gene expression signature for gastric cancer**. *J Cancer Res Clin Oncol* 2017, **143**(4):619-629.
27. Xiang DM, Sun W, Zhou T, Zhang C, Cheng Z, Li SC, Jiang W, Wang R, Fu G, Cui X *et al*: **Oncofetal HLF transactivates c-Jun to promote hepatocellular carcinoma development and sorafenib resistance**. *Gut* 2019, **68**(10):1858-1871.
28. He R, Zuo S: **A Robust 8-Gene Prognostic Signature for Early-Stage Non-small Cell Lung Cancer**. *Front Oncol* 2019, **9**.
29. Koshizuka K, Hanazawa T, Kikkawa N, Arai T, Okato A, Kurozumi A, Kato M, Katada K, Okamoto Y, Seki N: **Regulation of ITGA3 by the anti-tumor miR-199 family inhibits cancer cell migration and invasion in head and neck cancer**. *Cancer Sci* 2017, **108**(8):1681-1692.
30. Wang Z, Gao L, Guo X, Feng C, Lian W, Deng K, Xing B: **Development and validation of a nomogram with an autophagy-related gene signature for predicting survival in patients with glioblastoma**. *Aging (Albany NY)* 2019, **11**(24):12246-12269.

31. Jiao Y, Li Y, Liu S, Chen Q, Liu Y: **ITGA3 serves as a diagnostic and prognostic biomarker for pancreatic cancer**. *Onco Targets Ther* 2019, **12**:4141-4152.
32. Zhai T, Muhanhali D, Jia X, Wu Z, Cai Z, Ling Y: **Identification of gene co-expression modules and hub genes associated with lymph node metastasis of papillary thyroid cancer**. *Endocrine* 2019, **66**(3):573-584.
33. Campbell G, Swamynathan S, Tiwari A, Swamynathan SK: **The secreted Ly-6/uPAR related protein-1 (SLURP1) stabilizes epithelial cell junctions and suppresses TNF-alpha-induced cytokine production**. *Biochem Biophys Res Commun* 2019, **517**(4):729-734.
34. Throm VM, Männle D, Giese T, Bauer AS, Gaida MM, Kopitz J, Bruckner T, Plaschke K, Grekova SP, Felix K *et al*: **Endogenous CHRNA7-ligand SLURP1 as a potential tumor suppressor and anti-nicotinic factor in pancreatic cancer**. *Oncotarget* 2018, **9**(14):11734-11751.
35. Radiono S, Pramono ZAD, Oh GGK, Surana U, Widiyani S, Danarti R: **Identification of novel homozygous SLURP1 mutation in a Javanese family with Mal de Meleda**. *Int J Dermatol* 2017, **56**(11):1161-1168.
36. Zhou Y, Xiang J, Bhandari A, Guan Y, Xia E, Zhou X, Wang Y, Wang O: **CLDN10 is Associated with Papillary Thyroid Cancer Progression**. *J Cancer* 2018, **9**(24):4712-4717.
37. Zhang Z, Wang A, Sun B, Zhan Z, Chen K, Wang C: **Expression of CLDN1 and CLDN10 in lung adenocarcinoma in situ and invasive lepidic predominant adenocarcinoma**. *J Cardiothorac Surg* 2013, **8**:95.
38. de Ruiter EJ, Ooft ML, Devriese LA, Willems SM: **The prognostic role of tumor infiltrating T-lymphocytes in squamous cell carcinoma of the head and neck: A systematic review and meta-analysis**. *Oncoimmunology* 2017, **6**(11):e1356148.

Tables

Table 1. Details of the GEO datasets included in this study.

Datasets	Platform	Sample size (tumor / control)	Application
GSE6631	[HG_U95Av2] Affymetrix Human Genome U95 Version 2 Array	44 (22/22)	Identification of DEGs
GSE13601	[HG_U95Av2] Affymetrix Human Genome U95 Version 2 Array	58 (31/27)	Identification of DEGs
GSE30784	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	212 (167/45)	Identification of DEGs
GSE31056	[HG-U133_Plus_2]Affymetrix Gene Chip HumanGenomeHG-U133Plus2 Array	47 (23/24)	Identification of DEGs
GSE33205	[HuEx-1_0-st]Affymetrix Human Exon 1.0 ST Array [transcript (gene) version]	69 (44/25)	Identification of DEGs
GSE37991	Illumina HumanRef-8 v3.0 expression bead chip	80 (40/40)	Identification of DEGs
GSE51985	Illumina HumanHT-12 V4.0 expression bead chip	20 (10/10)	Identification of DEGs
GSE59102	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F	42 (29/13)	Identification of DEGs
GSE138206	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	18 (12/6)	Identification of DEGs
GSE65858	Illumina HumanHT-12 V4.0 expression bead chip	270	Validation of prognostic gene

Table 2. Patients' information of this study from the TCGA and GEO databases.

Characteristic		TCGA training datasets (n=491)	GSE65858(n=270)
Age (years)	<=50	87	47
	>50	404	223
Survival Status	Living	302	176
	Dead	189	94
Gender	Female	130	47
	male	361	223
Grade	G1	60	\
	G2	293	\
	G3	117	\
	G4	2	\
	GX/unknow	19	\
Pathologic T	T1	44	35
	T2	129	80
	T3	96	58
	T4	166	97
	Unknow	56	\
Pathologic N	N0	167	94
	N1	65	32
	N2	159	132
	N3	7	12
	NX/unknow	93	\
Pathologic M	M0	181	263
	M1	1	7
	MX/unknow	309	\
Tumor stage	Stage I	25	18
	Stage II	69	37
	Stage III	78	37
	Stage IV	251	178
	Unknow	68	\

Figures

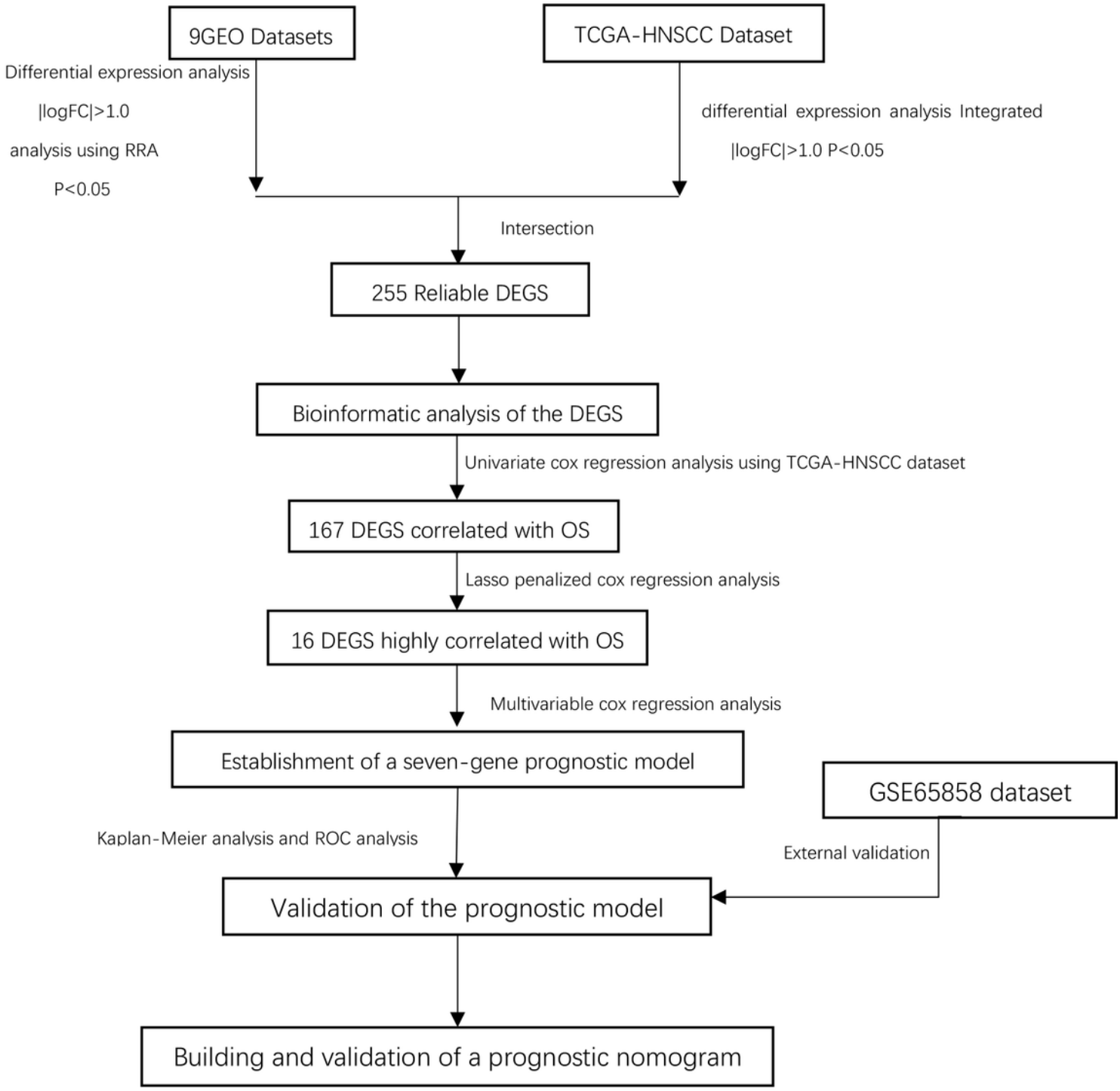

Figure 1

Overall flowchart of this study.

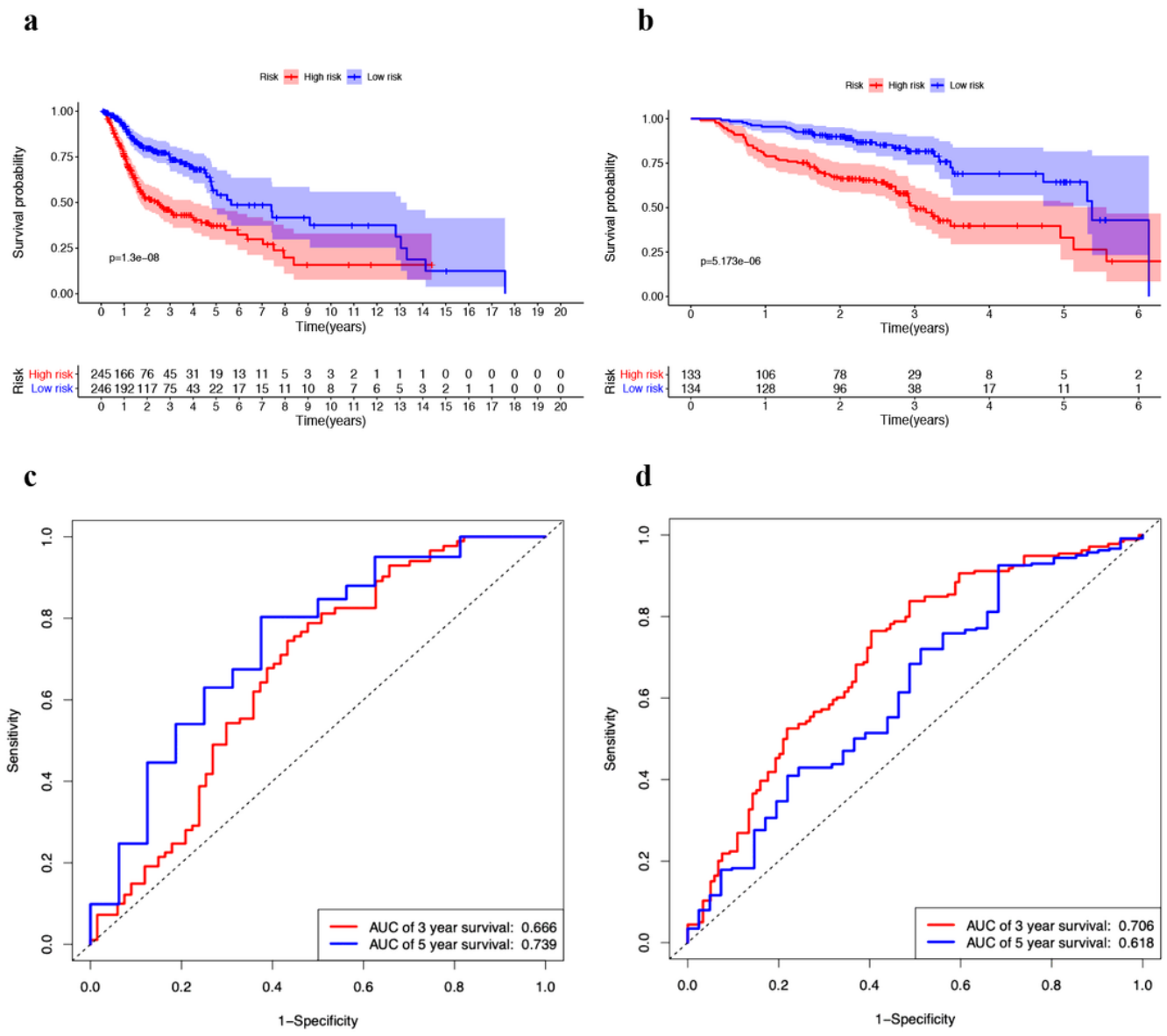

Figure 2

Verification of the seven gene signature in training and validation cohort. Kaplan–Meier analysis shows that patients with high risk scores have poorer OS, whether in the TCGA (a) or in the GSE65858 data set (b). The ROC curve shows the accuracy of predicting the 3-year and 5-year OS of patients in the TCGA (c) and GSE65858 data sets (d).

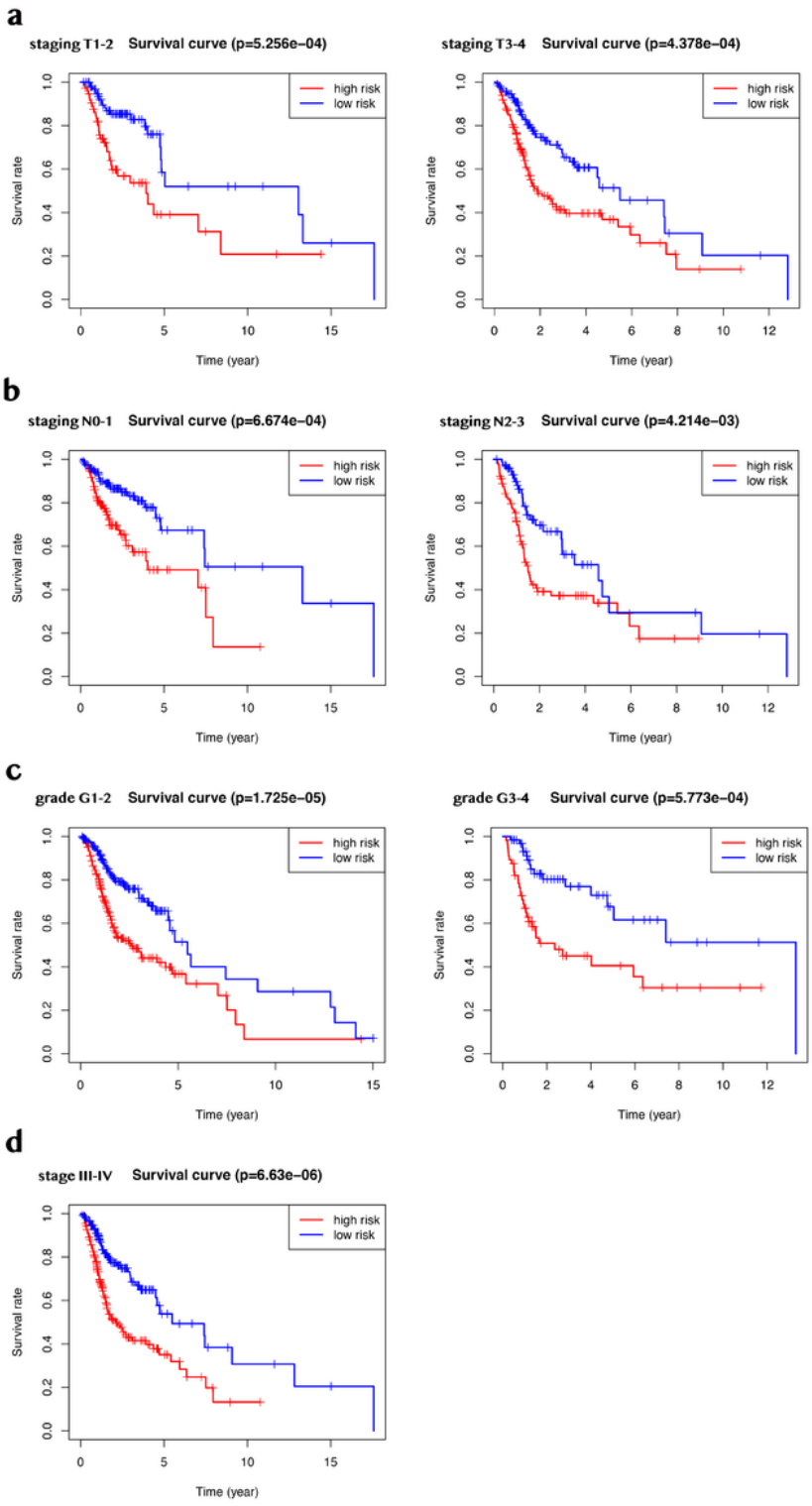

Figure 3

Kaplan–Meier analysis of different subgroups in the TCGA test set.

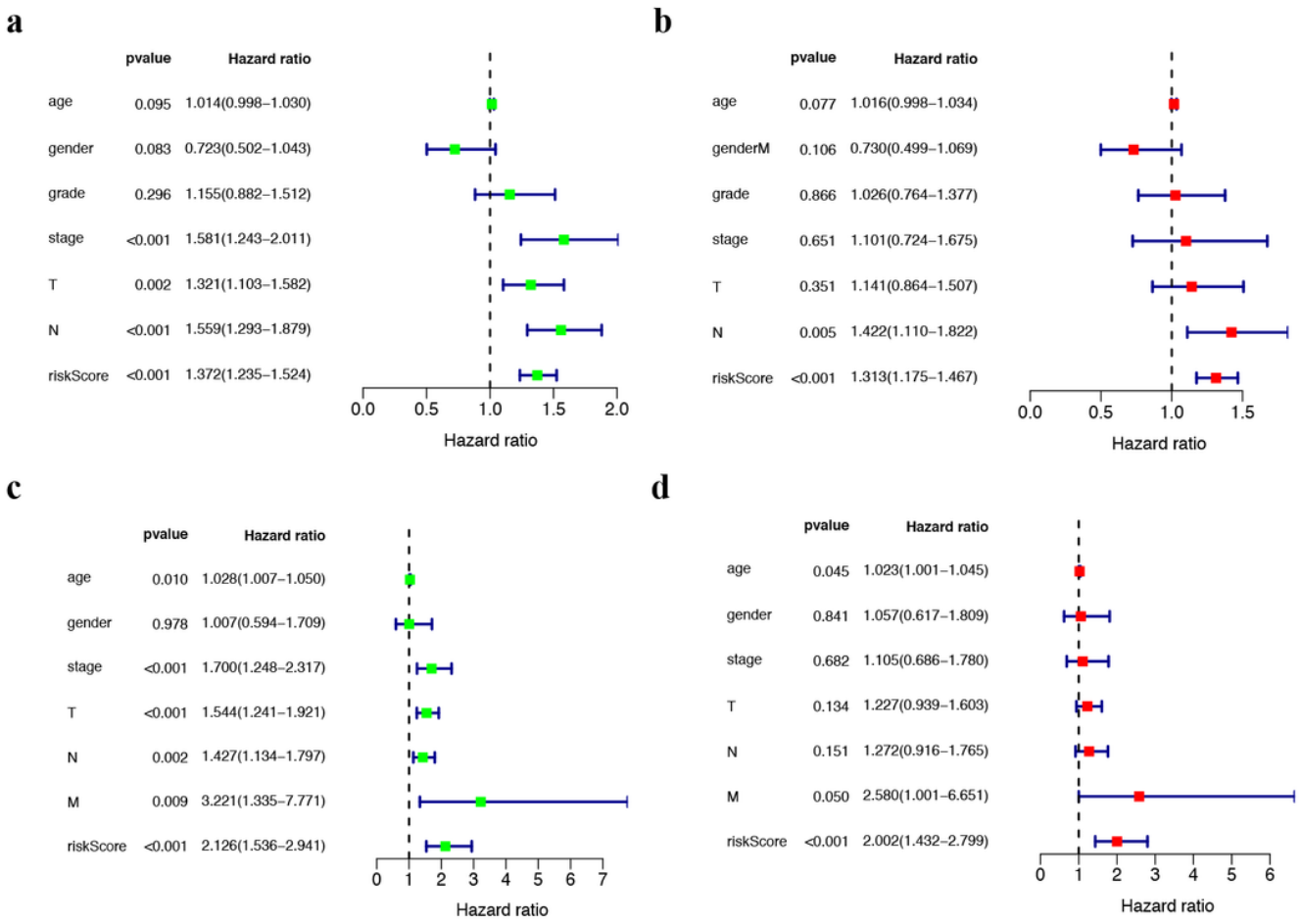

Figure 4

Cox regression analysis to detect clinical independence of risk scores. Uni-variate (a) and Multi-variate analysis (b) indicated that the risk score is an independent predictor variable in the TCGA dataset. (c-d) The Cox regression analysis also show that the risk score can be independent of other clinical variables in the validation set GSE65858.

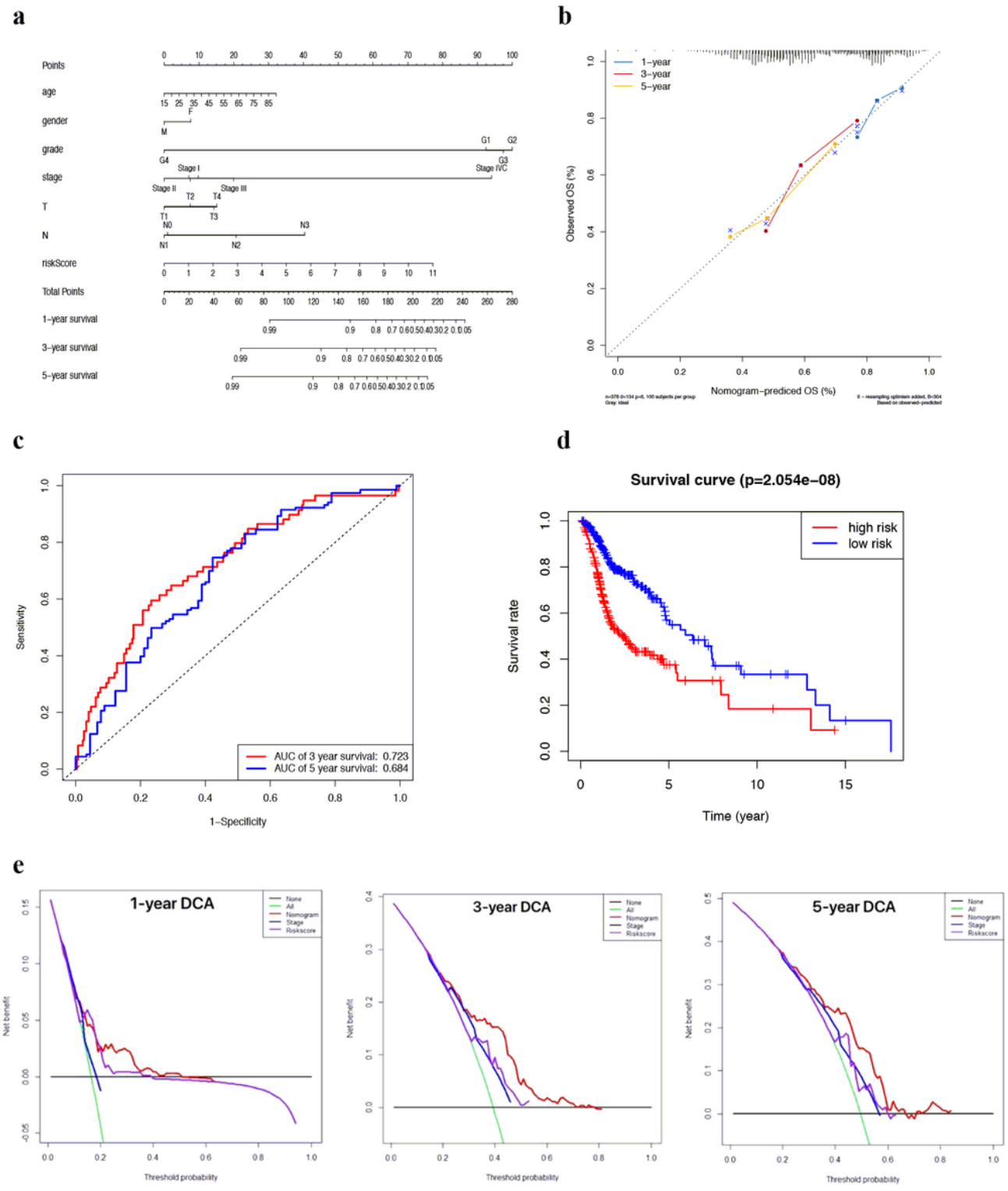

Figure 5

Verification of nomogram prediction performance. (a) A nomogram based on 7 genes and related clinical parameters. (b) The calibration curve reflects the accuracy of the nomogram to estimate the risk. (c) The ROC curve of nomogram to predict 3, 5 years OS in HNSCC patients. (d) The Kaplan–Meier analysis of the nomogram. (e) The DCA curves evaluate the clinical benefit and the application range of the nomograms.

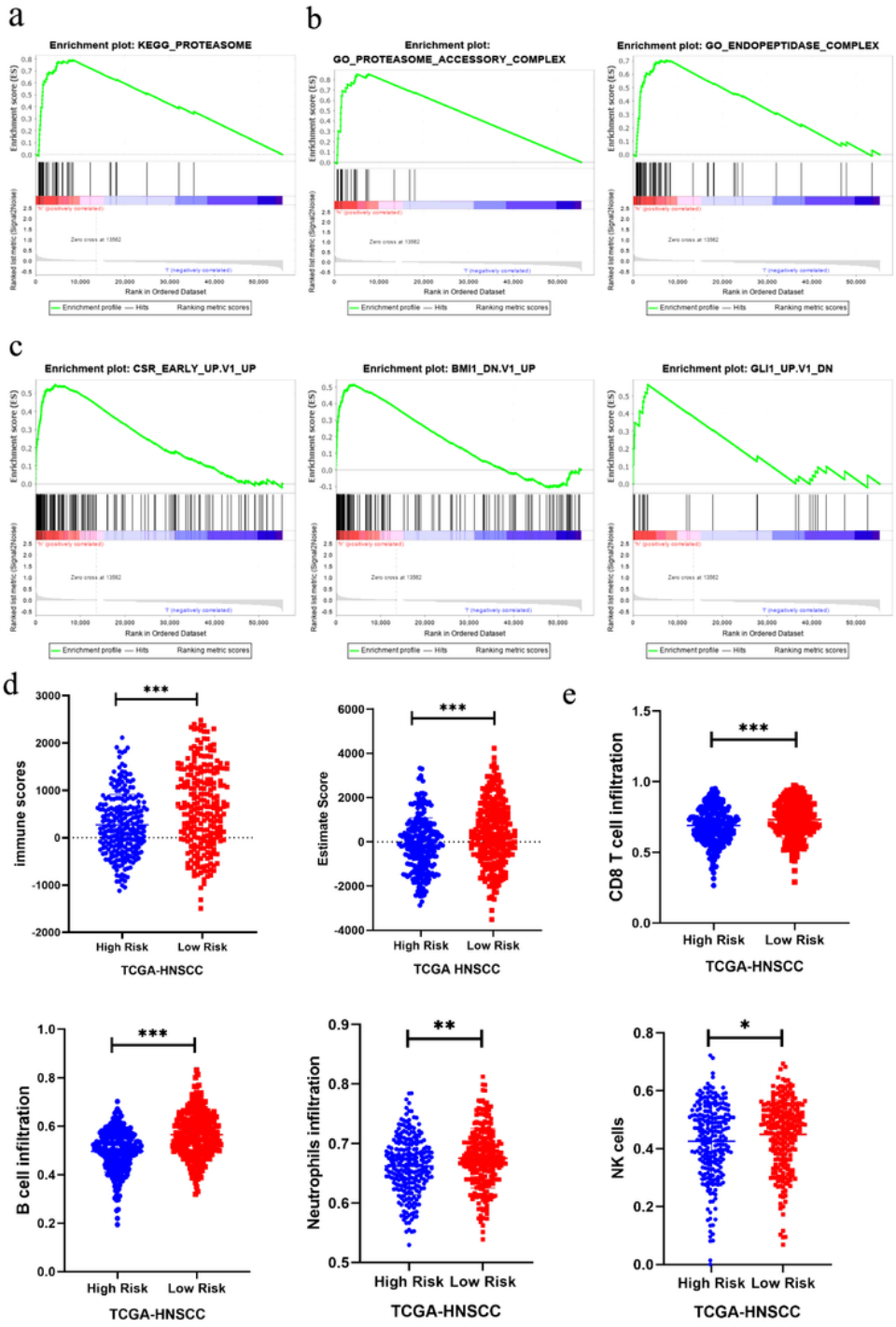

Figure 6

GSEA and immunity correlation analysis of seven prognostic gene. (a-c) The results showed that 1 KEGG pathway, 2 GO terms, and 3 oncological signatures were enriched in the high-risk group. (d) The immune score and estimated score of the high-risk group were significantly lower in the TCGA cohort. (e) The infiltration level of immune cells (including CD8+ T cells, B cells, neutrophils, and NK cells) in the low-risk group was significantly higher than that in the high-risk group in the TCGA cohort.