Identification and Validation of Prognostic Signature Incorporating Long Non-Coding RNAs and Clinical Factors in White Clear Cell Renal cell Carcinoma Patients

Yang Li
Zhengzhou University
Xue-Zhong Shi
Zhengzhou University
Xiao-Can Jia
Zhengzhou University Library
Xue-Ning Zhang
Zhengzhou University
Jun-Zhe Bao
Zhengzhou University
Jie Lu
Zhengzhou University
Yong-Li Yang (ylyang377@zzu.edu.cn)
Zhengzhou University https://orcid.org/0000-0002-8220-8133

Research

Keywords: ccRCC, long non-coding RNA, prognostic signature, overall survival

Posted Date: March 19th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-326364/v1

Abstract

Background:

In previous studies, the prognostic model of clear cell renal cell carcinoma (ccRCC) was constructed through all racial subjects, however, the intrinsic genomic differences between races may lead to disparity in survival outcomes, hence the accurate prognostic model of white patients in ccRCC was needed to be explored. This research aimed to identify and validate the potential long non-coding RNAs (lncRNAs) and clinical factors model to predict overall survival of ccRCC in white patients.

Methods:

LncRNA expression data and clinical factors of 459 white racial ccRCC subjects were downloaded from The Cancer Genome Atlas database. According to the exclusion criteria, 76 patients were excluded from the analytical dataset, hence 383 ccRCC participants were covered in present study. Then, 383 subjects were randomized into training series (n = 255) and test series (n = 128). The model was constructed using the training series, and validated in the test series. In the training series, the prognostic lncRNA model was constructed through univariate and multivariate Cox regression analysis, and least absolute shrinkage and selection operator regression analysis. Subsequently, the clinical variables were combined with lncRNA model to better predict the prognosis. The prognostic power of the models was verified by concordance index (95%*CI*) and area under time-dependent receiver operator characteristic curve (95%*CI*). Finally, the constructed models were validated in the test and entire series.

Results:

In the training series, 12 lncRNAs were confirmed as prognostic related biomarkers of ccRCC patients, among which AC012404.1, AC092296.1, AC099684.2, AC108752.1, AC131097.1, AL606519.1, and LINC02475 were seven novel candidate prognostic biomarkers. The performance evaluation of combined model incorporating 12-lncRNAs, age, and the tumor node metastasis stage showed that concordance index (95%*CI*) in the training, test and entire series were 0.863(0.830-0.896), 0.863(0.814-0.912) and 0.841(0.812-0.870), respectively, the 5-year area under time-dependent receiver operator characteristic curve (95%*CI*) in the training, test and entire series were 0.923(0.879-0.967), 0.861(0.777-0.945) and 0.879(0.834-0.924), respectively.

Conclusion:

This novel signature incorporating 12-lncRNAs, age, and the tumor node metastasis stage can be applied as an accurate tool for ccRCC prognostic evaluation in white patients.

Introduction

Based on Globocan cancer statistics in 2018, the new diagnostic cases and deaths number of renal cell [illegible] in all cancer cases respectively [1]. The incidence rate of

RCC was increasing rapidly and the incidence rate in North America and Europe is significantly higher than that in Africa and Asia [2].

There are many subtypes of RCC, the most common type is clear cell renal cell carcinoma (ccRCC), which accounts for almost 70%-80% of RCC cases [3]. ccRCC is a cancer with high recurrence and metastasis rate [4]. In recent years, treatment capacity of ccRCC has been improved, the 5-year survival rate for the patients diagnosed at early stage was > 90%, but the 5-year survival rate for the advanced tumor node metastasis (TNM) stage patients was as low as 12% [5]. Currently, the TNM stage of American Joint Committee on Cancer (AJCC) is still the prognostic prediction method commonly used in ccRCC, however, the patients may still have different overall survival (OS) rate even if they have the same TNM stage, which may due to different molecular characteristics of the tumors [6]. Thus, it is necessary to improve the prognostic evaluation method currently in use through exploration of new prognostic factors in ccRCC.

With the rapid development and wide application of molecular biological techniques, high-throughput sequencing techniques, and bioinformatics, a large number of gene transcripts have been detected, most of which were non-coding RNAs that have not been annotated. Hence, exploring novel biomarkers may find potential prognostic factors for ccRCC. The largest amount of the mammalian non-coding transcriptomes were long non-coding RNAs (lncRNAs) [7]. LncRNAs are a class of RNA transcripts longer than 200 nucleotides in length, lacking protein-coding capacity [8]. Numerous studies have indicated that lncRNAs can regulate gene expression through the process of transcription, post transcriptional, and chromatin modification [9, 10]. Moreover, several prognostic lncRNAs for ccRCC have been reported in experimental studies. However, the functions of great amount of lncRNAs in ccRCC were unknown. Thus, more potential and valuable lncRNAs need being identified.

In previous studies, subjects of all races were included in the analytical dataset [11, 12]. However, because of the great majority of data were white racial patients, it is not suitable to conduct subgroup analysis on different races. Meanwhile, multiple studies have demonstrated that regardless of stage, surgical treatment, histologic subtype, sex, or age, white patients with RCC have better OS compared with African American patients [13–15]. The intrinsic genomic differences between races may bring about this difference in survival outcomes [16]. Thus, in this study, only white patients were covered to decrease the bias of results.

The purpose of current research was to construct a prognostic risk score model in white ccRCC patients through OS-related lncRNAs. Then, the prognostic accuracy of this model was assessed and validated in multiple datasets.

Materials And Methods

Data acquisition and preprocessing

The Cancer Genome Atlas (TCGA) is a database that contains a total of 33 types of cancers. The clinical information and transcriptome profiling (mainly containing messenger RNAs and lncRNAs) were downloaded from TCGA kidney renal clear cell carcinoma (KIRC) Data Portal (https://portal.gdc.cancer.gov/),which was fully open and this study was in full compliance with TCGA publication guidelines. The version of this dataset was: February 23, 2020. The white-ethnicity patients make up of 86.6% (459/530) in the entire population, so in this study only 459 white patients were selected. The raw data covered 459 tumor samples and 75 adjacent samples.

There were three steps to preprocess the raw data. Firstly, Perl script was used to transform the raw expression data to a matrix. Subsequently, gene id transformed to Ensembl id using Ensembl database. The Ensembl id contained annotation part to identify the gene type. Finally, the lncRNA expression quantification data was separated from the entire matrix for subsequent prognostic analysis. 14,093 lncRNAs from transcriptome profiling were extracted in this study.

In order to normalize the data, the log2-transformation of each lncRNA expression data was performed. Next, the survival information was combined with the lncRNA matrix using the id number. From the day of diagnosis to the day of final follow-up or death was calculated as the survival time of patients. The censored data were defined as patients still alive until the final follow-up day or with no events.

There were three exclusion criteria because of the concerning of data integrality: (1) incomplete AJCC-TNM characteristics; (2) the follow-up time is missing or less than 30 days; (3) have other tumors. In this study, 76 patients were excluded from the subsequent analysis. Taken together, 383 ccRCC samples were covered in present study. The workflow of this study was shown in (Fig. 1).

The 383 KIRC participants were randomized into training set and test set using block-randomization script, and the ratio of the two series was 2:1. The prognostic model was constructed using training set. The test set was used for validation. Differences in age between the two groups were compared using an independent *t*-test. Chi-Square test was used to compare the two sets in other binary variables. [Figure 1near here]

Differential expression analysis in KIRC and control

In the training set, three KIRC patients all have three cancer tissues, mean counts were calculated to contain all patients only have one cancer tissue and one cancer-adjacent normal tissue for analysis. The R package “edgeR” was utilized for screening differentially expressed lncRNAs in ccRCC samples compared with normal samples [17]. The false discovery rate (FDR) was performed to adjust the *P* statistical value [18]. The threshold criteria for the differential expression were FDR < 0.05, $|\log_2 \text{FoldChange}| > 1$, and row mean expression counts > 1. Additionally, the volcano plot and heatmap were used to visualize the differentially expressed lncRNAs. The R package “gplots” was utilized for constructing heatmap to cluster analysis.

Construction of prognostic lncRNA signature

The normal samples were deleted and only cancer tissues were left after differential expression analysis. In survival analysis, the expression data of the differentially expressed lncRNAs were merged with the survival time and vital status. Univariate Cox regression analysis was utilized to unravel the relevance between survival outcome and the expression of differentially expressed lncRNAs. Hazard ratio (*HR*) and *P* statistical values were calculated to show the results. A statistically significant association was defined as *P*-value < 0.05.

The least absolute shrinkage and selection operator (LASSO) regression was conducted to deal data with high dimensional indicators and strongly relevant variables, and to avoid overfitting in the model. The "glmnet" and "survival" packages were used in this model. LASSO model can conduct a penalty ratio according to their size to shrink the regression coefficients. As a result, several lncRNAs which weak associated with the prognosis of ccRCC were removed from the model. The optimal values of the penalty parameter λ was determined by the 10-times cross-validations method [19]. The log(λ) versus partial likelihood deviance was plotted and the λ was chosen where the partial likelihood deviance is the smallest. Seed 2020 was set in order to make the result repeatable. According to the minimum λ, corresponding lncRNAs were chosen in the model.

Multivariate Cox regression analysis was performed to find the OS-related lncRNAs. Both direction stepwise regression analysis according to Akaike information criterion was conducted in this study. A statistically significant association was defined as *P*-value < 0.05. In addition, *HR* (95% *CI*) were calculated. *HR* > 1 indicates the risk factor while *HR* < 1 indicates the protective biomarker. Furthermore, the model fittness was expressed as concordance index with 95% confidence intervals. The formula risk score= $\Sigma(\beta_n \times e_n)$ was used to compute the risk score of single patient.

In this formula, the *e* value is the expression data of the lncRNA, the regression coefficient of each lncRNAs is the corresponding β value in the results of multivariate Cox regression analysis [20].

In the light of median cutoff point of the risk score, all the subjects were categorized into low-risk and high-risk groups. The Kaplan-Meier survival curve was utilized to two risk groups. The comparation of the survival differences between two groups was using the log-rank test. The median survival time of two groups were illustrated through ggsurvplot function and risk table.

Model evaluation in the training set

In different stages of ccRCC, the pathological characteristics are different, so the Kaplan-Meier survival curve was performed for TNM Ⅰ-Ⅱ set and TNM Ⅲ-Ⅳ set to bear out model's reliability. The median survival time of two groups were also calculated. The subgroup analysis were also performed in age (more than 65 years/ less than 65 years), gender (male/female), and treatment type (radiation therapy/pharmaceutical therapy).

The 1-, 3-, 5-, and 10-year time-dependence receiver operator characteristic (time-dependence ROC) curves were generated to assess the prognostic power of model. The area under time-dependent ROC curve

timeROC" and "survival" packages [21].

The nomogram and calibration curve were constructed using the “rms” and “survival” packages. Nomogram is a good tool through multiple indicators to predict or diagnose the incidence or progress of cancer [22]. A 12-lncRNAs based nomogram was built for predicting the 1-, 3-, 5-, and 10-year OS of each ccRCC patient. The performance of the prognostic nomogram was assessed through the corresponding calibration curves.

Risk score signature combined with clinical factors in the training set

Excluding the influence of molecular levels and genes, the prognosis of ccRCC often rely on many clinical variables. For instance, the preliminary tumor prognostic prediction in clinic usually using the AJCC-TNM stage as a reference factor. Additionally, other clinical factors such as gender and age may enhance the prediction ability of the model. Hence, a model containing prognostic lncRNA and clinical prognostic factors was constructed to optimize the prognostic ability.

To determine the relationship between the clinical variables and OS rate, univariate Cox regression analysis was performed and statistically significant association was defined as *P*-value < 0.05. According to the results, the lncRNAs within the risk sore model and statistically significant clinical variables were combined in multivariate Cox regression analysis to enhance model’s reliability. The fittness of combined model were expressed as concordance index with 95% confidence intervals and AUC(t) with 95% confidence intervals.

Demonstration prognostic models in the test and entire sets

The model constructed in the training set was applied to the test and entire sets, in order to validate the effectiveness of the risk score model. The process of validation sets was similar as training set, in the light of their corresponding median value of risk score, the subjects were portioned into low-risk and high-risk groups. Concordance index (95%*CI*) and (AUC(t)) (95%*CI*) were calculated to evaluate the performance of models. The Kaplan-Meier curves were performed for low-risk and high-risk patients.

Clinical variables combined risk score model was also applied to the test set and entire set. Concordance index (95%*CI*) and (AUC(t)) (95%*CI*) were calculated to evaluate the combined model.

Study checklist and statistical software

This article was presented in accordance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) checklist [23]. All samples were run using R Project for Statistical Computing,version 4.0.2 (https://www.R-project.org/). The Perl (version 5.30.2) was used for data preprocessing. The entire subjects were randomized into training series and test series using Stata16/SE. The description and comparation of baseline characteristics in the training and test series were performed with the IBM SPSS Statistics version 27.0 (SPSS Inc., Chicago).

Results

Baseline characteristics of participants

In this study, 383 KIRC patients were stratified into two series: training series (255 patients) and test series (128 patients). The detailed baseline characteristics of two sets were shown in Table 1. Except for age, no statistical differences were found in the distribution of these variables between the two independent sets ($P > 0.05$).

Table 1
Characteristics of the study patients in training-test cohorts at baseline.

Characteristic	Training cohort ($N = 255$) no. (%)	Test cohort ($N = 128$) no. (%)	χ^2/t	P
Gender			0.260	0.610
Male	172(67.5)	83(64.8)		
Female	83(32.5)	45(35.2)		
AJCC-TNM stage			0.833	0.361
□-□	146(57.3)	67(52.3)		
□-□	109(42.7)	61(47.7)		
AJCC-T stage			0.604	0.437
T1-T2	154(60.4)	72(56.3)		
T3-T4	101(39.6)	56(43.8)		
AJCC-M stage			0.533	0.465
M0	211(82.7)	102(79.7)		
M1	44(17.3)	26(20.3)		
Age in years (mean ± SD)	59.2 ± 11.8	61.8 ± 12.2	2.052	0.041
Treatment type			0.012	0.913
Radiation Therapy	122(47.8)	62(48.4)		
Pharmaceutical Therapy	133(52.2)	66(51.6)		
The P value was calculated by χ^2 for categorical variables and t test for age.				
AJCC, American Joint Committee on Cancer; SD, standard deviance				

The median survival time was 3.76 years (interquartile range [IQR]: 1.90–5.66) in the training cohort, [redacted] IQR: 1.65–5.35). For the vital status, there were 87 (34.1%)

patients dead and 168 (65.9%) patients alive in the training series, and 46 (35.9%) patients dead and 82 (64.1%) patients alive in the test series. [Table 1 near here]

Screening of differentially expressed lncRNAs in the training set

Total 616 differentially expressed lncRNAs were identified using 14,093 lncRNAs expression profiles between 255 KIRC samples and 38 matched adjacent normal tissues. These differentially expressed lncRNAs were clearly showed in the volcano plot, among which 282 were up-regulated ($\log_2$FoldChange > 1) and 334 were down-regulated ($\log_2$FoldChange < -1) (Fig. 2A). The top 30 up-regulated and down-regulated lncRNAs were selected to cluster analysis, and the results were shown in the heatmap (Fig. 2B). The top 20 down-regulated lncRNAs sorted by |$\log_2$FoldChange| were listed in Additional Table 1 and the top 20 up-regulated lncRNAs sorted by |$\log_2$FoldChange| were listed in Additional Table 2. [Figure 2 near here]

Table 2
12 prognostic-related lncRNAs in ccRCC.

LncRNA	Univariate analysis		Multivariable analysis			
	HR	*P*-value	*β*	*SE*	*HR* (95% *CI*)	*P*-value
Protective biomarker						
AC008556.1	0.766	0.015	-0.285	0.114	0.752(0.602–0.940)	0.012
AC012404.1	0.792	0.026	-0.317	0.112	0.728(0.585–0.908)	0.005
AC092296.1	0.668	< 0.001	-0.268	0.102	0.765(0.626–0.935)	0.009
AC099684.2	0.834	0.007	-0.169	0.081	0.844(0.721–0.990)	0.037
SPINT1-AS1	0.589	< 0.001	-0.444	0.111	0.641(0.515–0.798)	< 0.001
Risk biomarker						
AC108752.1	1.088	0.023	0.144	0.046	1.155(1.056–1.264)	0.002
AC131097.1	1.370	0.002	0.325	0.113	1.385(1.109–1.729)	0.004
AL606519.1	1.428	< 0.001	0.246	0.098	1.279(1.055–1.551)	0.012
FOXP4-AS1	1.273	< 0.001	0.337	0.106	1.400(1.138–1.723)	0.001
LINC00261	1.186	< 0.001	0.154	0.071	1.166(1.016–1.339)	0.029
LINC02446	1.332	< 0.001	0.143	0.070	1.153(1.005–1.324)	0.042
LINC02475	1.365	< 0.001	0.335	0.078	1.399(1.200–1.630)	< 0.001

HR,hazard ratio; *β*, regression coefficient in the model; *SE*□standard error; *CI*, confidence interval; lncRNAs, long non-coding RNAs;

ccRCC, clear cell renal cell carcinoma.

Screening of prognostic-related lncRNAs and constructing the prognostic 12-lncRNA signature in ccRCC

To pick out the OS-related lncRNAs, 616 differentially expressed lncRNAs were initially subjected to univariate Cox proportional hazards regression analysis in the training series. Subsequently, 217 lncRNAs associated with OS of ccRCC patients ($P < 0.05$) were selected into LASSO-Cox regression model. The 38 lncRNAs were identified with the minimum lambda 0.046 (Fig. 3). Finally, these 38 lncRNAs were subjected to multivariate Cox proportional hazards regression analysis. [Figure 3 near here]

Based on the above process, in the training cohort, 12 lncRNAs (AC008556.1, AC012404.1, AC092296.1, AC099684.2, AC108752.1, AC131097.1, AL606519.1, FOXP4-AS1, LINC00261, LINC02446, LINC02475, and SPINT1-AS1) were ultimately screened as prognostic lncRNAs of ccRCC patients (all $P <$

0.05),among those five were protective biomarkers and seven were risk factors (Table 2). The concordance index (95%*CI*) of the model was 0.832(0.785–0.879).

According to the above results, the risk score model was constructed as follows:

Risk score= -0.285×$e_{AC008556.1}$ -0.317×$e_{AC012404.1}$-0.268×$e_{AC092296.1}$-0.169×$e_{AC099684.2}$-0.444×$e_{SPINT1-AS1}$+0.144×$e_{AC108752.1}$+0.325×$e_{AC131097.1}$+0.246×$e_{AL606519.1}$+0.337×$e_{FOXP4-AS1}$ +0.154×$e_{LINC00261}$+0.143×$e_{LINC02446}$+0.335×$e_{LINC02475}$

In light of median value of risk score 0.824, all the subjects enrolled into the study were portioned into low-risk group (n = 128) and high-risk group (n = 127). The results of Kaplan-Meier survival curves suggesting that subjects with low risk score suffered better OS compared with subjects who had high risk score (*P* < 0.001) (Fig. 4A). The median survival time in high-risk group was 3.92 years. [Table 2 near here] [Figure 4 near here]

Model evaluation results in the training set

The results of stratified survival analysis showed that TNM Ⅰ-Ⅱ and TNM Ⅲ-Ⅳ subjects with high risk score both suffered worse OS compared with patients who had low risk score (*P* < 0.001) (Fig. 5A). The median survival time of two high risk groups were 6.74 and 2.42 years respectively. Age less than 65 years and more than 65 years subjects with high risk score both suffered worse OS compared with patients who had low risk score, the corresponding *P* value were 0.034 and less than 0.001 respectively (Fig. 5B). The median survival time of two high risk groups were 4.49 and 3.76 years respectively. Male and female subjects with high risk score both suffered worse OS compared with patients who had low risk score (*P* < 0.001) (Fig. 5C). The median survival time of two high risk groups were 3.92 and 3.48 years respectively. Radiation therapy and pharmaceutical therapy subjects with high risk score both suffered worse OS compared with patients who had low risk score (*P* < 0.001) (Fig. 5D). The median survival time of two high risk groups were 3.39 and 4.29 years respectively.

The 1-, 3-, 5-, and 10-year time-dependence ROC curves were visualized in Fig. 6A. The 3- and 5-year (AUC(t)) (95%*CI*) were 0.854(0.791–0.918) and 0.882(0.829–0.936) respectively.

To bring convenience to clinical practice, the nomogram was conducted with 12-lncRNA signature (Fig. 7A). The corresponding 1-, 3-, 5-, and 10-year calibration curves were shown in Fig. 7B-7E. [Figure 5 near here] [Figure 6 near here] [Figure 7 near here]

Signature including lncRNAs and clinical factors in the training set

The univariate Cox regression analysis was performed to identify significant prognostic clinical features in the training group, among which the AJCC-TNM stage, risk score, AJCC-T stage, AJCC-M stage, and age were included (*P* < 0.05). Among those variables, age was grouped into four categories (“younger than 50 year old”, “aged 50 to 60 years”, “range in age from 60 to 70 years”, and “70 years of age or

older”) and AJCC-TNM stage had four categories (stage Ⅰ to stage Ⅳ). Due to the partial overlap of information among AJCC-M stage, AJCC-T stage, and AJCC-TNM stage, for decreasing collinearity of the model, only age and AJCC-TNM stage were selected to combine with 12-lncRNAs signature in multivariate Cox regression analysis. The results shown that age (> 70 years), AJCC-TNM stage Ⅲ and Ⅳ, and risk score were significantly correlated with inferior survival time ($P < 0.001$) (Table 3).

Table 3
Univariate and multivariate Cox regression analysis for the prognostic value of clinical features in ccRCC patients.

Variables	Univariate analysis		Multivariable analysis	
	HR	*P*-value	*HR (95% CI)*	*P*-value
gender (female/male)	0.750	0.188		
AJCC-M stage (M0/M1)	4.462	< 0.001*		
AJCC-T stage (T1-T2/T3-T4)	3.206	< 0.001*		
treatment type (Radiation/Pharmaceutical)	0.814	0.340		
risk score	1.0002	< 0.001*	1.0002	< 0.001*
age				
<50 years (n = 87)	1(Reference)		1(Reference)	
50–60 years (n = 111)	1.392	0.348	1.641(0.773–3.481)	0.197
60–70 years (n = 99)	1.880	0.059	1.659(0.827–3.324)	0.154
>70 years (n = 86)	3.559	< 0.001*	5.931(2.802–12.555)	< 0.001*
AJCC-TNM stage				
(n = 218)	1(Reference)		1(Reference)	
(n = 49)	1.500	0.321	1.286(0.559–2.958)	0.555
(n = 111)	2.608	0.002*	2.366(1.208–4.636)	0.012*
(n = 77)	6.975	< 0.001*	4.029(1.910–8.500)	< 0.001*

* The *P* value is less than 0.05. ccRCC, clear cell renal cell carcinoma; AJCC, American Joint Committee on Cancer;

HR,hazard ratio; *CI*, confidence interval; TNM, tumor node metastasis.

The 1-, 3-, 5-, and 10-year time-dependence ROC curves for signature including lncRNAs and clinical factors were visualized in Fig. 6D. The 3- and 5-year (AUC(t)) (95%*CI*) were 0.890(0.843–0.937) and 0.923(0.879–0.967) respectively. The concordance index (95%*CI*) of this model was 0.863(0.830–0.896). [Table 3 near here]

Demonstration model results in the test and entire sets

The 12-lncRNAs signature constructed in the training set was verified in the 128 ccRCC patients' test set and entire set of 383 ccRCC subjects. The results of the survival analysis demonstrated that the subjects in the low risk group both had better OS compared with that in the high risk group ($P < 0.001$) (Fig. 4B, C). The time-dependence ROC curves for predicting the 1-, 3-, 5-, and 10-year OS in the test and entire sets were visualized in Fig. 6B, C. In the test set, the 3- and 5-year (AUC(t)) (95%*CI*) were 0.720(0.614–0.825) and 0.733(0.622–0.845) respectively. The concordance index (95%*CI*) of the model was 0.701(0.623–0.779). In the entire set, the 3- and 5-year (AUC(t)) (95%*CI*) were 0.757(0.692–0.822) and 0.778(0.719–0.836) respectively. The concordance index (95%*CI*) of the model was 0.744(0.695–0.793).

For validating the signature including lncRNAs and clinical factors in the training set, the 1-, 3-, 5-, and 10-year time-dependence ROC curves in test and entire sets were visualized in Fig. 6E, F. In the test set, the 3- and 5-year (AUC(t)) (95%*CI*) were 0.908(0.849–0.967) and 0.861(0.777–0.945) respectively. The concordance index (95%*CI*) of the model was 0.863(0.814–0.912). In the entire set, the 3- and 5-year (AUC(t)) (95%*CI*) were 0.876(0.837–0.916) and 0.879(0.834–0.924) respectively. The concordance index (95%*CI*) of this model was 0.841(0.812–0.870).

Discussion

In present study, only the white racial was selected, so the prognostic model we constructed was more accurate to white racial populations. Considering guarantee the effectiveness of the model, the entire datasets were partitioned into training and test sets using Stata16/SE block-randomization script, the model was constructed using the training series, and validated in the test series and entire series. As shown in Table 1, the training series and the test series were randomly allocated. To avoid overfitting, the LASSO-Cox regression method was performed for further select lncRNAs. Considering clinical factors are also important in clinical practice, so the significant clinical variables, especially AJCC-TNM stage and age, were combined with risk score model to better predict the prognosis.

We screened 616 differentially expressed lncRNAs using TCGA-KIRC dataset in the training cohort, of which 334 were down-regulated and 282 were up-regulated. Among them, 12 lncRNAs (AC008556.1, AC012404.1, AC092296.1, AC099684.2, AC108752.1, AC131097.1, AL606519.1, FOXP4-AS1, LINC00261, LINC02446, LINC02475, and SPINT1-AS1) were associated with the prognosis of ccRCC (all $P < 0.05$). Then we constructed a prognostic 12-lncRNAs signature, the results proved that a higher risk score was correlated with inferior OS, suggesting that risk score was related to the prognosis of ccRCC. In addition, an outstanding consistency between the

real survival situation and prediction results through the prognostic nomogram were presented in the calibration plots for 1-, 3-, 5-, and 10-year OS. The 12-lncRNAs signature all remains an effective prognostic model, when stratified by age, gender, treatment type, and the AJCC-TNM stage, indicating that this lncRNAs signature was effective in patients with different clinical features. Then, 12-lncRNAs signature was combined with age and AJCC-TNM stage to better predict the prognosis. Finally, the prognostic effects of this combined model were validated in the test and entire cohorts, the results in two data sets were both consistent with the result in training data set, indicating broad applicability of this signature in ccRCC patients. As shown in the Kaplan-Meier plots for all cohorts, OS rate of subjects with low-risk scores was significantly higher than subjects with high-risk scores ($P < 0.001$). All the results suggesting that this prognostic model is suitable for estimating the OS of ccRCC patients.

The performance evaluation of combined model showed that concordance index (95%*CI*) in the training, test, and entire cohorts were 0.863(0.830–0.896), 0.863(0.814–0.912), and 0.841(0.812–0.870), respectively, the 5-year area under time-dependent receiver operator characteristic curve (95%*CI*) in the training, test, and entire cohorts were 0.923(0.879–0.967), 0.861(0.777–0.945), and 0.879(0.834–0.924), respectively. The results of three data sets were all greater than 0.800, indicating that this combined model has higher prediction ability.

For the prognostic model of RCC reported before, three important signatures were used for different types of patients. The University of California Los Angeles integrated staging system model integrated Fuhrman grade, Eastern Cooperative Oncology Group performance status, and TNM stage to predict survival for RCC patients, this model is superior to stage alone in differentiating the survival of patients and is simple to use, however, because of the heterogeneity of patients and treatments, it may be less accurate in the metastatic RCC patients [24, 25]. The Memorial Sloan-Kettering Cancer Center model used five pretreatment factors to predict the survival of metastatic patients [26, 27]. The International Metastatic Renal Cell Carcinoma Database Consortium model use six prognostic factors to predict the survival of metastatic patients [28]. All three prognostic signatures were stratified the subjects into three different risk groups including high-, intermediate-, and low-risk groups. These models had been validated before, the prognostic factors contained in these models are mostly biochemical indicators, whereas no biomarkers such as genes were included.

In the risk score model, 12 lncRNAs could be seen as potential prognostic factors. Among these, AC008556.1, AC012404.1, AC092296.1, AC099684.2, and SPINT1-AS1 are the protective biomarkers. However, the biological functions of AC012404.1, AC092296.1, and AC099684.2 have not been reported in previous research, only AC008556.1 and SPINT1-AS1 were studied before in other cancers. AC008556.1 (ENSG00000277013), as one of the novel S-phase-upregulated lncRNAs, was used in mechanistic studies to determine the role in cell-cycle progression [29]. Serine peptidase inhibitor, Kunitz Type 1 antisense RNA1 (SPINT1-AS1) is a member of serine protease inhibitors of the Kunitz family, has anti-cancer properties through inhibiting cell proliferation, invasion, migration, and decreased expression in many cancers [30]. Zhou *et al.* [31] confirmed that SPINT1-AS1 is up-regulated in the breast cancer cell [illegible] proliferation of breast cancer cells through regulating miR-let-

7a/b/i-5p. However, SPINT1-AS1 is down-regulated in this research. Huang *et al.* [32]reported that the prognostic role of some lncRNAs are consistent in many tumors, while other lncRNAs may have different functions in different tumors. Thus, the biological function of SPINT1-AS1 in ccRCC need more experiments to explore.

High expression of AC108752.1, AC131097.1, AL606519.1, FOXP4-AS1, LINC00261, LINC02446, and LINC02475 were significantly correlated with an inferior prognosis of ccRCC. However, to date, the biological functions of AC108752.1, AC131097.1, AL606519.1, and LINC02475 were little known, no experimental studies have been performed in cancer before, only LINC02446, FOXP4-AS1, and LINC00261 were studied before in other cancers. LINC02446 (ENSG00000256039) was identified as Epithelial-Mesenchymal Transition-Related lncRNAs correlated with the progression and prognosis in bladder cancer patients [33]. Forkhead box P4 antisense RNA 1 (FOXP4-AS1) is a 24.727 kb lncRNA. Li *et al.* [34] have confirmed that FOXP4-AS1 expression is associated to the development of colorectal cancer (CRC), *in vivo* experiments demonstrated that silencing of FOXP4-AS1 in CRC cell lines can inhibit the ability of tumor cells to form tumors in nude mice. LINC00261 (ENSG00000236384) was found to behave as a tumor suppressor through activating DNA damage signaling pathway to arrest cellular division in lung adenocarcinoma [35]. Yan *et al.* [36] demonstrated that LINC00261 repressed colon cancer progression through regulating the Wnt and miR-324-3p pathway. However, the function and expression of LINC02446, FOXP4-AS1, and LINC00261 in ccRCC are still unknown.

The present study has several important strengths. Firstly, we found seven novel candidate prognostic biomarkers including AC012404.1, AC092296.1, AC099684.2, AC108752.1, AC131097.1, AL606519.1, and LINC02475 in ccRCC, the related research should be conducted to explore their biological functions. Furthermore, the significant clinical factors were combined with the 12-lncRNAs signature to make it more functional in clinical practice. Finally, the inclusion criteria were stricter in this study to decrease the confounding factors.

In the present research, though the prediction model elucidated good accuracy in white patients of ccRCC, some limitations should be acknowledged. Firstly, the research data were downloaded from public available TCGA datasets and no validation set was from a prospective cohort, so experimental studies should be prepared in our later research to further validate prognostic value of the model we constructed in ccRCC patients. Furthermore, the mechanism between the 12-lncRNAs and ccRCC prognosis is unclear, so more mechanism research should be conducted to explore the role of these 12 lncRNAs in ccRCC. We acknowledge that before this model applied in the clinic, it is necessary to conduct more experimental researches to further confirm this model.

Conclusions

In summary, our study find that AC012404.1, AC092296.1, AC099684.2, AC108752.1, AC131097.1, AL606519.1, and LINC02475 are novel candidate prognostic biomarkers in ccRCC. Furthermore, we have successfully demonstrated the effectiveness of the prognostic signature including 12-lncRNAs, age, and

AJCC-TNM stage applied for investigating prognosis of ccRCC in white patients. This new signature can be applied to screen for high-risk white patients in ccRCC and assess prognosis more accurate in white patients.

Supplementary Information

Additional file 1: Table 1. The top 20 down-regulated lncRNAs sorted by $|\log_2 FoldChange|$ in ccRCC.

Additional file 2: Table 2. The top 20 up-regulated lncRNAs sorted by $|\log_2 FoldChange|$ in ccRCC.

Abbreviations

AJCC, American Joint Committee on Cancer; AUC(t), area under time-dependent receiver operator characteristic curve; ccRCC, clear cell renal cell carcinoma; *CI*, confidence interval; CRC, colorectal cancer; FDR, false discovery rate; *HR*,hazard ratio; IQR, interquartile range; KIRC, kidney renal clear cell carcinoma; LASSO, least absolute shrinkage and selection operator; lncRNAs, long non-coding RNAs; OS, overall survival; RCC, renal cell carcinoma; ROC, receiver operator characteristic; SD, standard deviance; *SE*□standard error;TCGA, The Cancer Genome Atlas; TNM, tumor node metastasis; TRIPOD, transparent reporting of a multivariable prediction model for individual prognosis or diagnosis.

Declarations

Acknowledgements

We acknowledge TCGA database for providing their platforms and contributors for uploading their meaningful datasets.

Authors' contributions

Yong-Li Yang and Xue-Zhong Shi conceived and designed the project; Yang Li, Xue-Ning Zhang and Xiao-Can Jia acquired the data; Jun-Zhe Bao and Jie Lu analysed and interpreted the data; Yang Li and Yong-Li Yang wrote the paper; all authors read and approved the final manuscript.

Funding

Not applicable

Availability of data and materials

The datasets downloaded and analyzed in this study are available in the TCGA repository (https://portal.gdc.cancer.gov/).

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA-cancer J Clin. 2018;68(6):394–424.
2. Leslie I, Boos LA, Larkin J, Pickering L. Avelumab and axitinib in the treatment of renal cell carcinoma: safety and efficacy. Expert Rev Anticancer Ther. 2020;20(5):343–54.
3. Hsieh JJ, Purdue MP, Signoretti S, Swanton C, Albiges L, Schmidinger M, et al. Renal cell carcinoma. Nat Rev Dis Primers. 2017;3:17009.
4. Del Vecchio SJ, Ellis RJ. Cabozantinib for the Management of Metastatic Clear Cell Renal Cell Carcinoma. Journal of Kidney Cancer Vhl. 2018;5(4):1–5.
5. Atkins MB, Tannir NM. Current and emerging therapies for first-line treatment of metastatic clear cell renal cell carcinoma. Cancer Treat Rev. 2018;70:127–37.
6. Martinez-Salamanca JI, Huang WC, Millan I, Bertini R, Bianco FJ, Carballido JA, et al. Prognostic Impact of the 2009 UICC/AJCC TNM Staging System for Renal Cell Carcinoma with Venous Extension. Eur Urol. 2011;59(1):120–7.
7. Esteller M. Non-coding RNAs in human disease. Nat Rev Genet. 2011;12(12):861–74.
8. Cui N, Liu J, Xia H, Xu D. LncRNA SNHG20 contributes to cell proliferation and invasion by upregulating ZFX expression sponging miR-495-3p in gastric cancer. J Cell Biochem. 2019;120(3):3114–23.
9. Fatica A, Bozzoni I. Long non-coding RNAs: new players in cell differentiation and development. Nat Rev Genet. 2014;15(1):7–21.

10. Ørom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, et al. Long noncoding RNAs with enhancer-like function in human cells. Cell. 2010;143(1):46–58.
11. Liu T, Sui J, Zhang Y, Zhang XM, Wu WJ, Yang S, et al. Comprehensive analysis of a novel lncRNA profile reveals potential prognostic biomarkers in clear cell renal cell carcinoma. Oncol Rep. 2018;40(3):1503–14.
12. Liu H, Ye T, Yang X, Lv P, Wu X, Zhou H, et al. A Panel of Four-lncRNA Signature as a Potential Biomarker for Predicting Survival in Clear Cell Renal Cell Carcinoma. J Cancer. 2020;11(14):4274–83.
13. Chow WH, Shuch B, Linehan WM, Devesa SS. Racial disparity in renal cell carcinoma patient survival according to demographic and clinical characteristics. Cancer. 2013;119(2):388–94.
14. Berndt SI, Carter HB, Schoenberg MP, Newschaffer CJ. Disparities in treatment and outcome for renal cell cancer among older black and white patients. J Clin Oncol. 2007;25(24):3589–95.
15. Tripathi RT, Heilbrun LK, Jain V, Vaishampayan UN. Racial disparity in outcomes of a clinical trial population with metastatic renal cell carcinoma. Urology. 2006;68(2):296–301.
16. Krishnan B, Rose TL, Kardos J, Milowsky MI, Kim WY. Intrinsic Genomic Differences Between African American and White Patients With Clear Cell Renal Cell Carcinoma. JAMA Oncol. 2016;2(5):664–7.
17. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
18. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics. 2003;19(3):368–75.
19. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med. 1997;16(4):385–95.
20. Zeng JH, Lu W, Liang L, Chen G, Lan HH, Liang XY, et al. Prognosis of clear cell renal cell carcinoma (ccRCC) based on a six-lncRNA-based risk score: an investigation based on RNA-sequencing data. J Transl Med. 2019;17(1):281.
21. Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. JAMA. 2017;318(14):1377–84.
22. Williams SB, Huo J, Chu Y, Baillargeon JG, Daskivich T, Kuo YF, et al. Cancer and All-cause Mortality in Bladder Cancer Patients Undergoing Radical Cystectomy: Development and Validation of a Nomogram for Treatment Decision-making. Urology. 2017;110:76–83.
23. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. J Clin Epidemiol. 2015;68(2):134–43.
24. Zisman A, Pantuck AJ, Dorey F, Said JW, Shvarts O, Quintana D, et al. Improved prognostication of renal cell carcinoma using an integrated staging system. J Clin Oncol. 2001;19(6):1649–57.
25. Patard JJ, Kim HL, Lam JS, Dorey FJ, Pantuck AJ, Zisman A, et al. Use of the University of California Los Angeles integrated staging system to predict survival in renal cell carcinoma: an international multicenter study. J Clin Oncol. 2004;22(16):3316–22.

26. Motzer RJ, Mazumdar M, Bacik J, Berg W, Amsterdam A, Ferrara J. Survival and prognostic stratification of 670 patients with advanced renal cell carcinoma. J Clin Oncol. 1999;17(8):2530–40.
27. Heng DY, Xie W, Regan MM, Warren MA, Golshayan AR, Sahi C, et al. Prognostic factors for overall survival in patients with metastatic renal cell carcinoma treated with vascular endothelial growth factor-targeted agents: results from a large, multicenter study. J Clin Oncol. 2009;27(34):5794–9.
28. Ko JJ, Xie W, Kroeger N, Lee JL, Rini BI, Knox JJ, et al. The International Metastatic Renal Cell Carcinoma Database Consortium model as a prognostic tool in patients with metastatic renal cell carcinoma previously treated with first-line targeted therapy: a population-based study. Lancet Oncol. 2015;16(3):293–300.
29. Hao Q, Zong X, Sun Q, Lin YC, Song YJ, Hashemikhabir S, et al. The S-phase-induced lncRNA SUNO1 promotes cell proliferation by controlling YAP1/Hippo signaling pathway. Elife. 2020;9:e55102.
30. Li C, Li W, Zhang Y, Zhang X, Liu T, Zhang Y, et al. Increased expression of antisense lncRNA SPINT1-AS1 predicts a poor prognosis in colorectal cancer and is negatively correlated with its sense transcript. Onco Targets Ther. 2018;11:3969–78.
31. Zhou T, Lin K, Nie J, Pan B, He B, Pan Y, et al. LncRNA SPINT1-AS1 promotes breast cancer proliferation and metastasis by sponging let-7 a/b/i-5p. Pathol Res Pract. 2020;217:153268.
32. Huang Y, Ling A, Pareek S, Huang RS. Oncogene or tumor suppressor? Long noncoding RNAs role in patient's prognosis varies depending on disease type. Transl Res. 2021;230:98–110.
33. Tong H, Li T, Gao S, Yin H, Cao H, He W. An Epithelial-Mesenchymal Transition-Related Long Noncoding RNA Signature Correlates With The Prognosis And Progression In Bladder Cancer Patients. Biosci Rep. 2020;41(1).
34. Li J, Lian Y, Yan C, Cai Z, Ding J, Ma Z, et al. Long non-coding RNA FOXP4-AS1 is an unfavourable prognostic factor and regulates proliferation and apoptosis in colorectal cancer. Cell Prolif. 2017;50(1).
35. Shahabi S, Kumaran V, Castillo J, Cong Z, Nandagopal G, Mullen DJ, et al. LINC00261 Is an Epigenetically Regulated Tumor Suppressor Essential for Activation of the DNA Damage Response. Cancer Res. 2019;79(12):3050–62.
36. Yan D, Liu W, Liu Y, Luo M. LINC00261 suppresses human colon cancer progression via sponging miR-324-3p and inactivating the Wnt/β-catenin pathway. J Cell Physiol. 2019;234(12):22648–56.

Figures

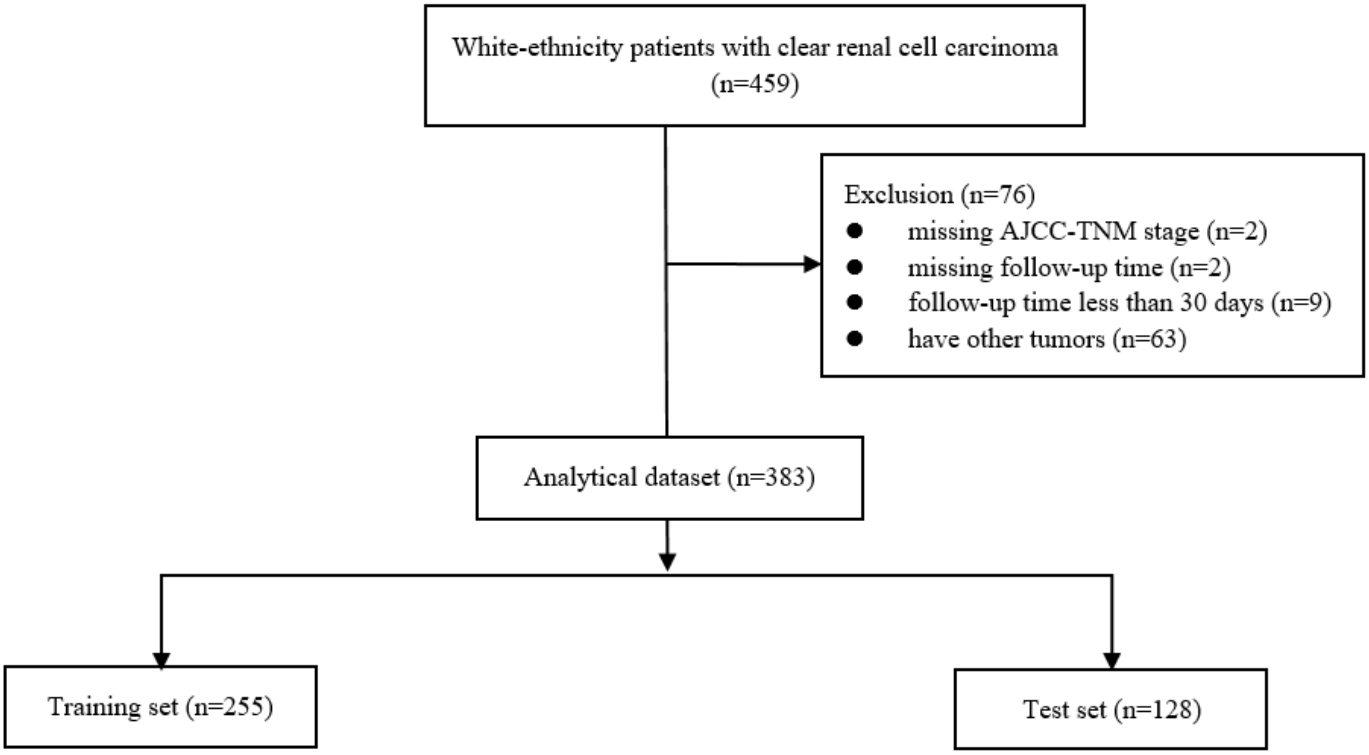

Figure 1

Workflow chart of this study. Abbreviations: AJCC, American Joint Committee on Cancer; TNM, tumor node metastasis.

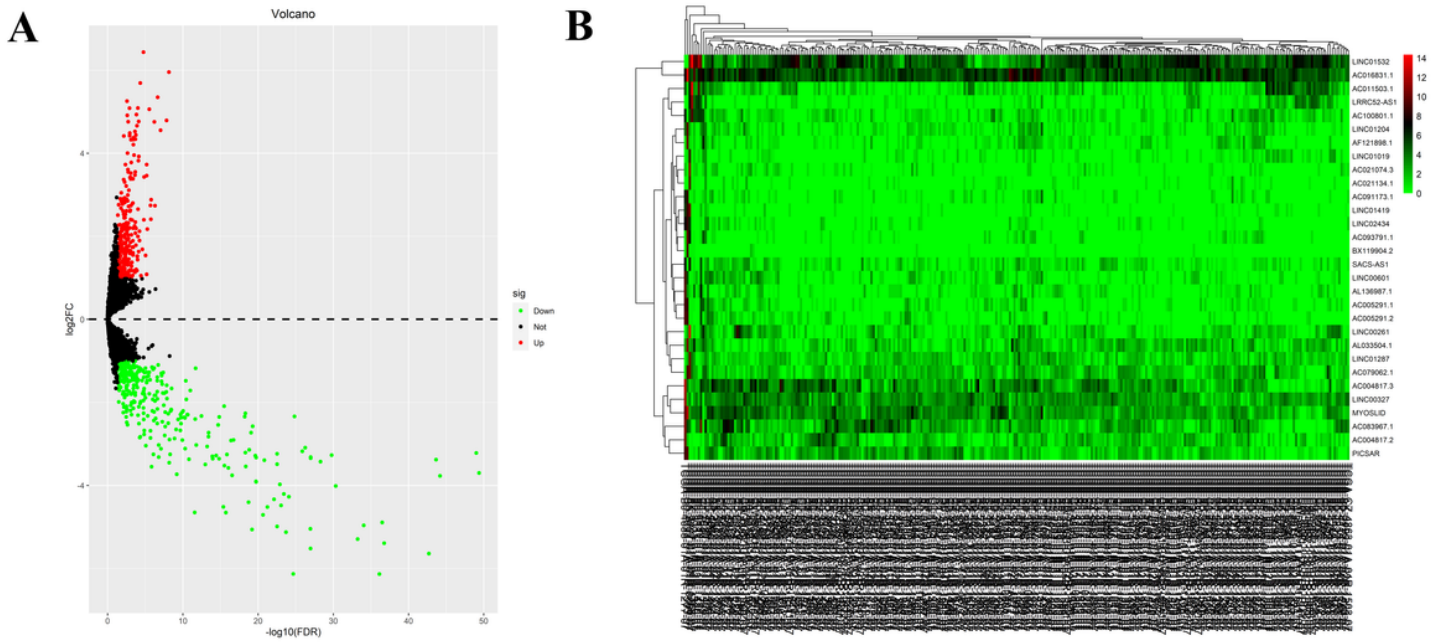

Figure 2

Differentially expressed lncRNAs analysis. Notes: (A). Volcano plot of differentially expressed lncRNAs. dentified, including 282 up-regulated (red dots in the figure)

and 334 down-regulated (green dots in the figure) lncRNAs (FDR <0.05 and |log2FoldChange|>1). (B). Heatmap of the top 30 differentially expressed lncRNAs. Each column represents a sample and each row represents a lncRNA. Short red and green vertical bars represent up-regulated and down-regulated lncRNAs, respectively. Abbreviations: FC, fold change; FDR, false discovery rate; lncRNAs, long non-coding RNAs.

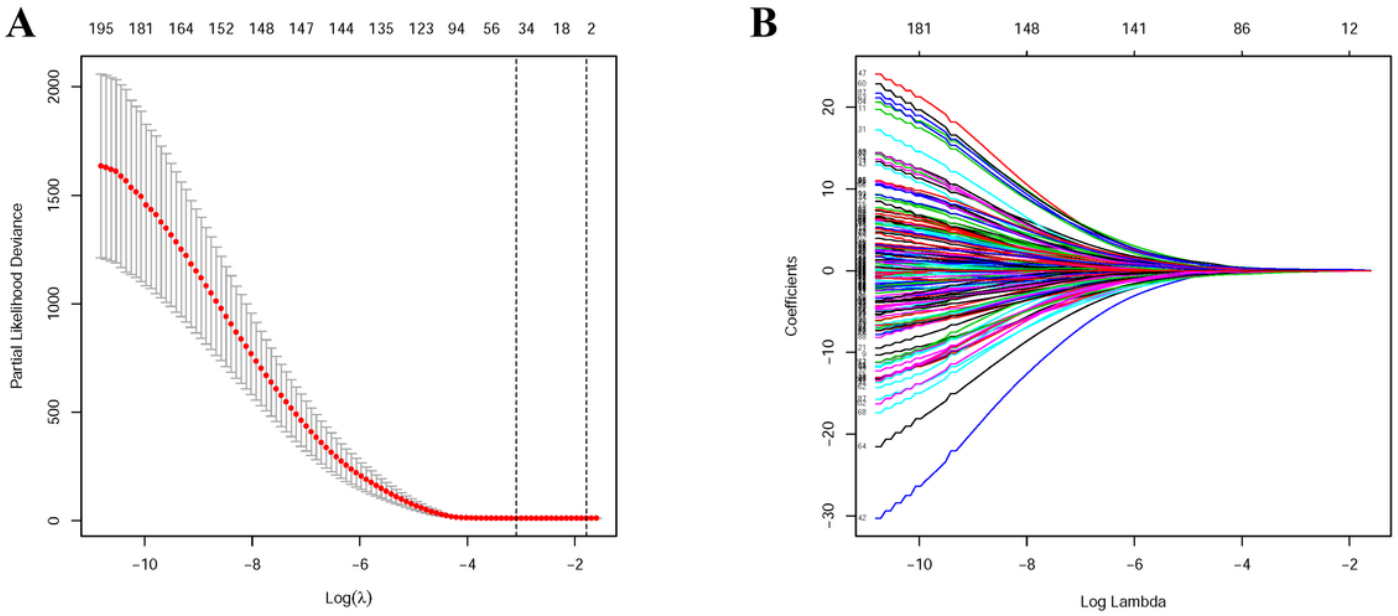

Figure 3

LASSO-Cox regression model. Notes: (A). The tuning parameter (λ) selection process in the LASSO-Cox regression model. On the top of the figure, the numbers represented the corresponding amount of the candidate lncRNAs according to different λ value. The dotted vertical lines are drawn at the optimum values by 1-SE and minimum criteria. The black solid vertical lines indicate the partial likelihood deviance ± SE. (B). LASSO-Cox coefficient profiles of the 217 KIRC-correlated lncRNAs. According to the optimal λ 0.046, 38 nonzero coefficients were left in the model. Abbreviations: LASSO, least absolute shrinkage and selection operator; KIRC, kidney clear cell carcinoma; SE□standard error□lncRNAs, long non-coding RNAs.

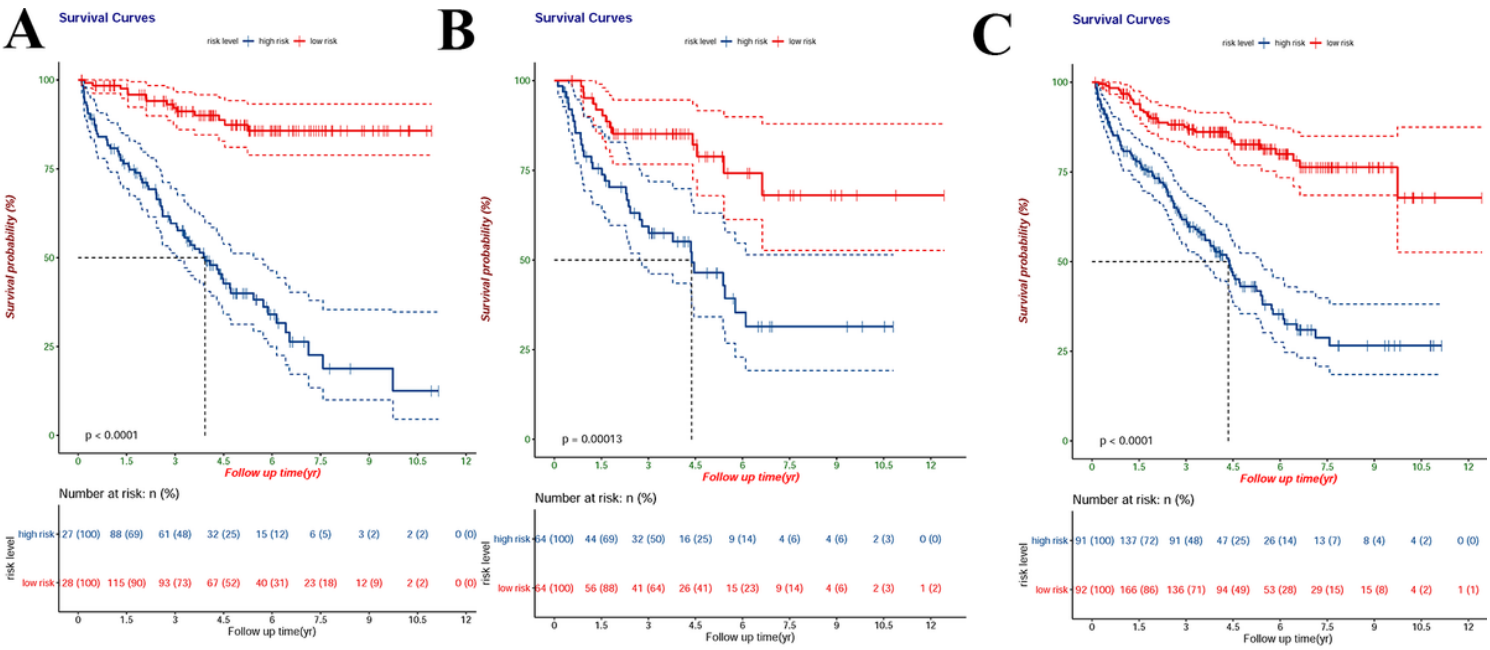

Figure 4

Kaplan-Meier plot for different datasets. Notes: Kaplan-Meier plot for (A). training cohort (n=255). (B). test cohort (n=128). (C). entire cohort (n=383). The red lines represent the low-risk group and the blue lines represent the high-risk group. The small vertical lines on the curves indicate censored data. The dotted blue and red lines are 95% confidence intervals of survival probability in two groups. The black dotted vertical lines are drawn at the 50% survival probability to show the median survival time of the high-risk group. The difference comparison results between the two groups were shown in each figure. The risk table was drawn at bottom in each figure to show the number and percentage of survival subjects in each follow-up time. Abbreviations: TNM, tumor node metastasis.

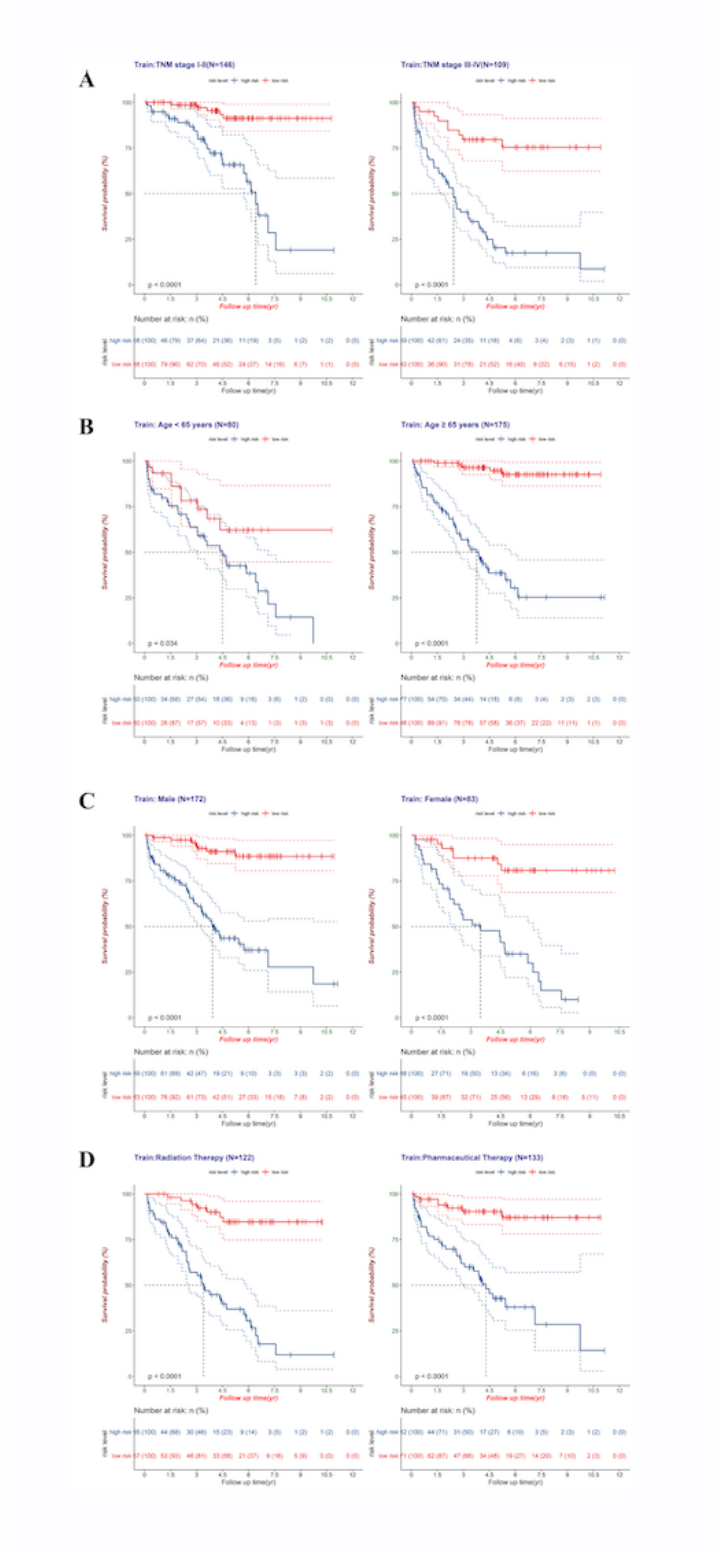

Figure 5

Subgroup analysis in the training cohort. Notes: The patients in each subgroup were divided into high risk group and low risk group. Kaplan-Meier plot for two groups in (A). TNM Ⅰ-Ⅱ patients (n=146) and TNM Ⅲ-Ⅳ patients (n=109). (B). age less than 65 years patients (n=80) and age more than 65 years patients (n=175). (C). male patients (n=172) and female patients (n=83). (D). radiation therapy patients (n=122) and no radiation therapy patients (n=133). Abbreviations: TNM, tumor node metastasis.

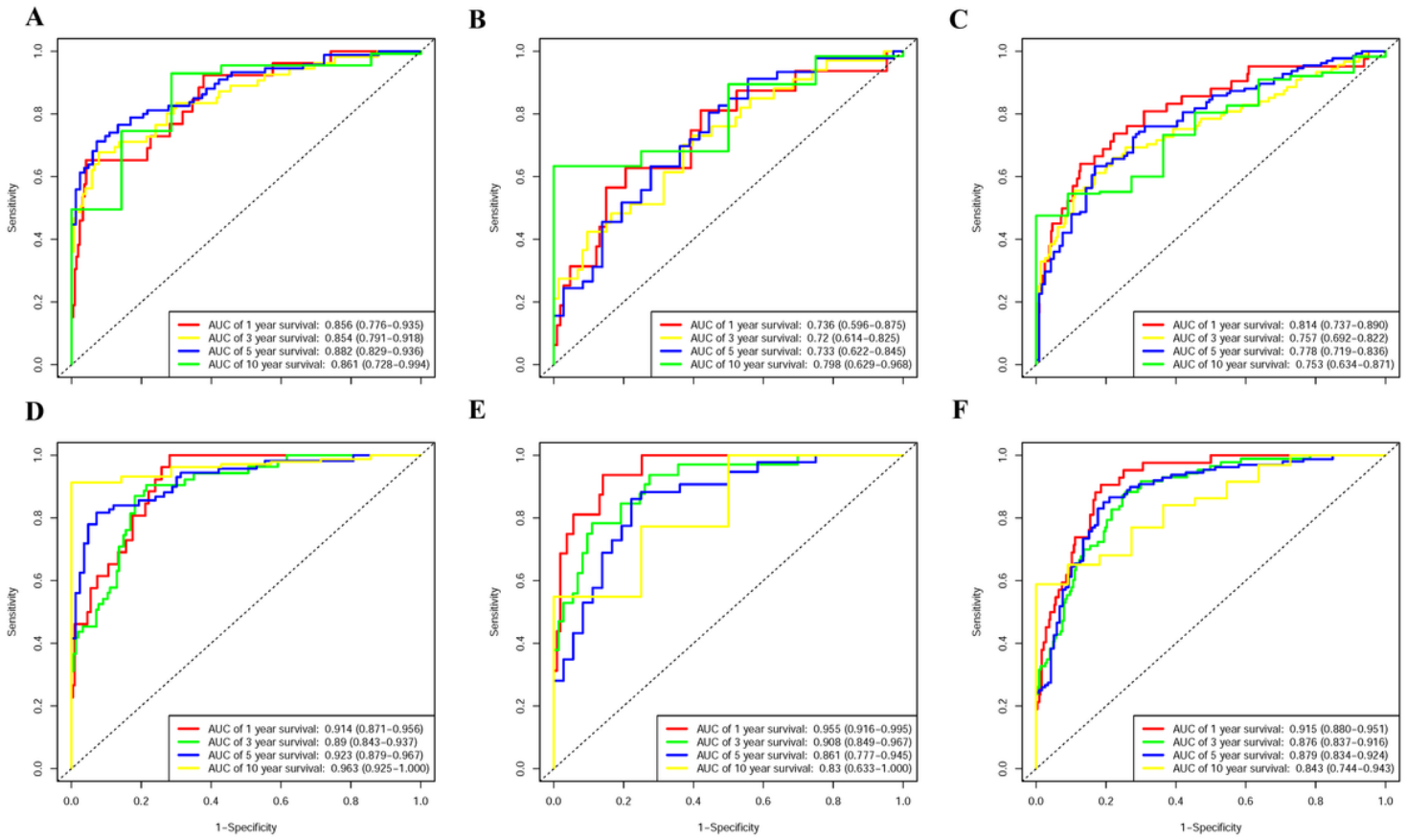

Figure 6

Evaluation of the prognostic model using (AUC(t)) (95%CI). Notes: The 1-year, 3-year, 5-year, and 10-year time-dependence ROC curves of 12-lncRNAs signature (A). in the training cohort. (B). in the test cohort. (C). in the entire cohort. The 1-year, 3-year, 5-year, and 10-year time-dependence ROC curves of the signature including lncRNAs and clinical factors (D). in the training cohort. (E). in the test cohort. (F). in the entire cohort. Abbreviations: AUC(t), area under time-dependent receiver operator characteristic curve; ROC, receiver operator characteristic lncRNAs, long non-coding RNAs; CI, confidence interval.

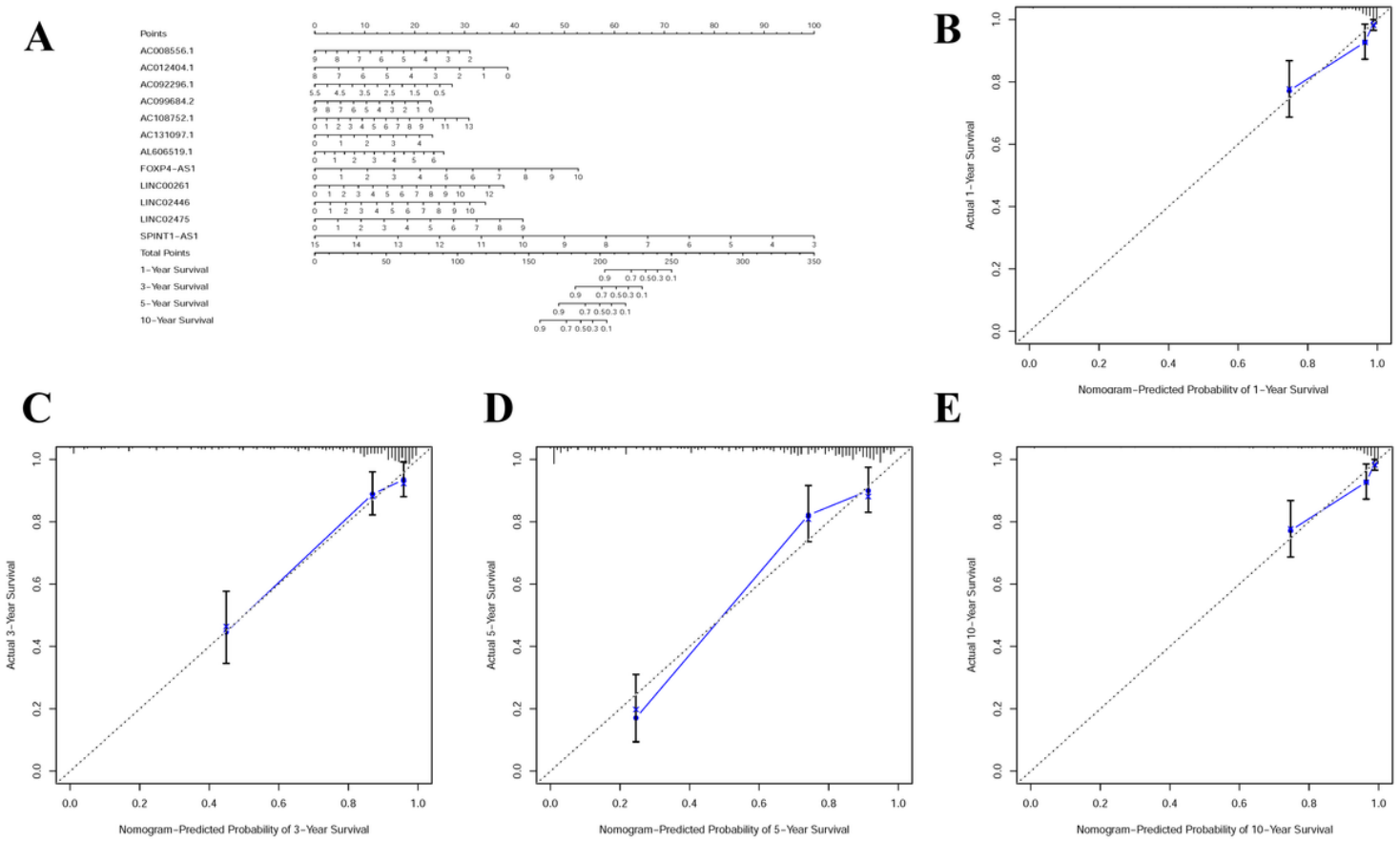

Figure 7

Prognostic nomogram and corresponding calibration curves. Notes: (A). Prognostic nomogram for the prediction of 1-year, 3-year, 5-year, and 10-year OS of patients with clear cell renal cell carcinoma in the training cohort based on 12 OS-related lncRNAs. Calibration curve of the nomogram prediction in the training cohort for (B). 1-year OS. (C). 3-year OS. (D). 5-year OS. (E). 10-year OS. Abbreviations: OS, overall survival

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalTable.docx