Exploration of an Integrative Prognostic Model of Radiogenomics Features With Underlying Gene Expression Patterns in Clear Cell Renal Cell Carcinoma

Yeqian Huang[1,2], Hao Zeng[1,3], Linyan Chen[1,3], Yuling Luo[1,3], Xuelei Ma[1,3]* and Ye Zhao[4]*

[1] Department of Biotherapy, Cancer Center, West China Hospital, Sichuan University, Chengdu, China, [2] West China School of Medicine, West China Hospital, Sichuan University, Chengdu, China, [3] State Key Laboratory of Biotherapy and Cancer Center, Collaborative Innovation Center for Biotherapy, West China Hospital, Sichuan University, Chengdu, China, [4] School of Bioscience and Technology, Chengdu Medical College, Chengdu, China

Background: Clear cell renal cell carcinoma (ccRCC) is one of the most common malignancies in urinary system, and radiomics has been adopted in tumor staging and prognostic evaluation in renal carcinomas. This study aimed to integrate image features of contrast-enhanced CT and underlying genomics features to predict the overall survival (OS) of ccRCC patients.

Method: We extracted 107 radiomics features out of 205 patients with available CT images obtained from TCIA database and corresponding clinical and genetic information from TCGA database. LASSO-COX and SVM-RFE were employed independently as machine-learning algorithms to select prognosis-related imaging features (PRIF). Afterwards, we identified prognosis-related gene signature through WGCNA. The random forest (RF) algorithm was then applied to integrate PRIF and the genes into a combined imaging-genomics prognostic factors (IGPF) model. Furthermore, we constructed a nomogram incorporating IGPF and clinical predictors as the integrative prognostic model for ccRCC patients.

Results: A total of four PRIF and four genes were identified as IGPF and were represented by corresponding risk score in RF model. The integrative IGPF model presented a better prediction performance than the PRIF model alone (average AUCs for 1-, 3-, and 5-year were 0.814 *vs.* 0.837, 0.74 *vs.* 0.806, and 0.689 *vs.* 0.751 in test set). Clinical characteristics including gender, TNM stage and IGPF were independent risk factors. The nomogram integrating clinical predictors and IGPF provided the best net benefit among the three models.

Conclusion: In this study we established an integrative prognosis-related nomogram model incorporating imaging-genomic features and clinical indicators. The results

indicated that IGPF may contribute to a comprehensive prognosis assessment for ccRCC patients.

Keywords: clear cell renal cell carcinoma, radiomics, genomics, machine learning, prognosis

INTRODUCTION

Renal cell carcinoma (RCC) is a common heterogeneous malignancy originated from renal tubular epithelial cells, with clear cell renal cell carcinoma (ccRCC) comprising about 80% of RCC cases (1, 2). Owing to the insufficient clinical symptoms and reliable diagnostic biomarkers at the early stage, about 30% of ccRCC patients had metastasis at the time of diagnosis, and about one-fifth of patients may experience metastasis or recurrence after radical treatment (3, 4). Imageological examinations such as conventional ultrasound, contrast-enhanced ultrasound, CT, contrast-enhanced CT and MRI have been applied to assess the overall profile of the tumor as noninvasive methods. However, there are limitations in these conventional imaging tests for differential diagnosis, preoperative pathological grading and prognosis of ccRCC, which also lack quantitative criteria.

Radiomics was first proposed by Lambin et al. (5) in 2012, which exploits high-throughput feature extraction algorithms to extract quantitative image features from standard medical images. Radiomics managed to perform the conversion from images into mineable data, which could then be applied to clinical decision support systems to achieve precise prediction, diagnosis, and prognostic evaluation of cancers (6, 7). A number of studies have reported that radiomics has been successfully applied in renal tumors researches, including Fuhrman staging of ccRCC (8–10), assessment of cancer phenotype and tumor microenvironment (11), differentiation of RCC and benign renal tumors (12, 13) and efficacy and prognosis evaluation (14, 15).

However, most studies regarding radiomics were primarily focused on the selection of image features and the quantitative analysis of tumors at the macroscopic level, and there has been little research into the mechanisms of deeper molecular biology. Combined with machine learning algorithms, we can further correlate the imaging data that reflects the quantitative phenotype of the disease with the genotype feature data which reveals the molecular activity. Correlation analysis between gene mutation, expression and imaging characteristics has been proved effective in the research of liver cancer (16), lung cancer (16–18), glioblastoma (19, 20) and Alzheimer's disease (21). Therefore, it is of vital importance to analyze the correlation and integration between imaging and genomic features of ccRCC, so as to understand the biological mechanism and furthermore obtain biomarkers for prognosis prediction, which will be more rewarding in personalized precision therapy.

Previous studies have proven that certain molecules and the activation of a series of signaling pathways are in close relation with the tumorigenesis and progression in ccRCC. For instance, the overexpression of vascular endothelial growth factor (VEGF) and platelet derived growth factor (PDGF) receptor tyrosine kinases are of great significance in promoting tumor angiogenesis and cell division. In addition, PI3K/AKT/mTOR pathway also results in affecting tumor cell growth and metabolism. Nevertheless, the associated gene expression profiles have not been thoroughly studied.

Standard treatments for ccRCC patients encompass surgery, radiotherapy and chemotherapy, and specific treatments including targeted therapy in combination with immune checkpoint inhibitors have shown efficacy in improving the overall survival (OS) of ccRCC patients (22, 23). However, the response of personalized therapy does vary and the prognosis is not as satisfactory. So far no routine genetic tests have been conducted, and these molecules concerning the mechanism of ccRCC development may provide opportunities to investigate potential biomarkers for diagnosis and prognosis. Therefore, it's essential to establish an effective model that conduce to risk stratification, treatment strategy support and prognostic prediction for patients with ccRCC.

In this study we concentrate on analyzing the radiomics features of contrast-enhanced CT and their association with genomics profiles of ccRCC samples, which has not been extensively researched. In order to select the imaging features significantly correlated to the prognosis of ccRCC, we applied several machine learning algorithms. Through machine-learning algorithms, we further estimated the correlation between prognosis-related image features (PRIF) and expressed genes profiles. Furthermore, the integration of radiomics and gene features was conducted to enhance the accuracy of prognostic evaluation. Eventually, we conducted validation of the imaging-genomic prognostic factors (IGPF) model, and the results suggested that these features may be of help in the prediction of prognosis in ccRCC patients. The potential connection and integration of macroscopic radiomics and genetic characteristics at the microscopic level needs further exploration.

MATERIALS AND METHODS

Data Source and Processing

The overall structure of our study was demonstrated in **Figure 1**. The detailed information of each section will be interpreted as follows. We downloaded the available enhanced CT images from the Cancer Imaging Archive (TCIA) portal (http://www.cancerimagingarchive.net/) and the information containing clinical features and mRNA sequencing data of corresponding ccRCC samples from the Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov). In total 205 available samples were gathered. For data normalization, we firstly acquired the raw count data of the ccRCC patients from the
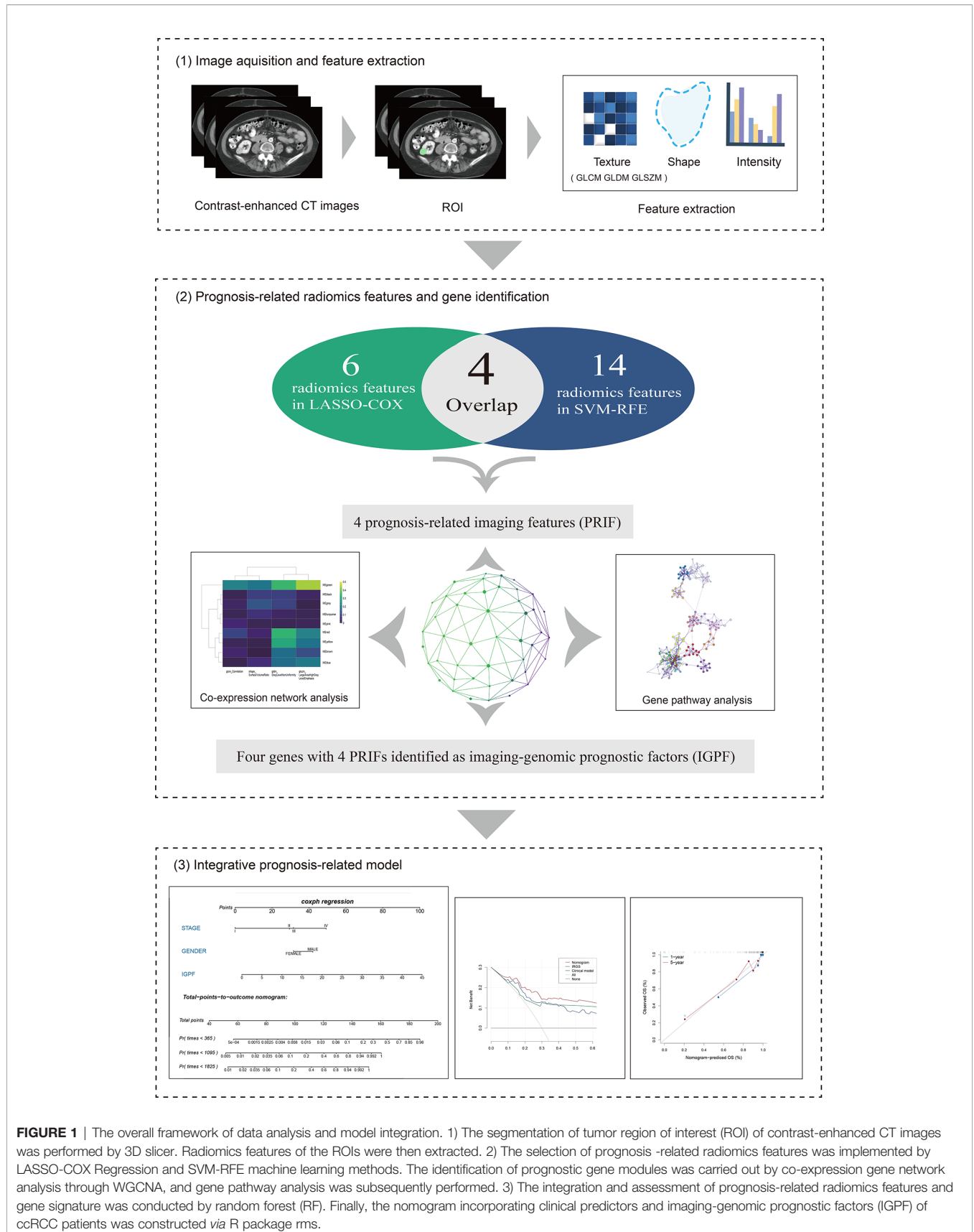
FIGURE 1 | The overall framework of data analysis and model integration. 1) The segmentation of tumor region of interest (ROI) of contrast-enhanced CT images was performed by 3D slicer. Radiomics features of the ROIs were then extracted. 2) The selection of prognosis -related radiomics features was implemented by LASSO-COX Regression and SVM-RFE machine learning methods. The identification of prognostic gene modules was carried out by co-expression gene network analysis through WGCNA, and gene pathway analysis was subsequently performed. 3) The integration and assessment of prognosis-related radiomics features and gene signature was conducted by random forest (RF). Finally, the nomogram incorporating clinical predictors and imaging-genomic prognostic factors (IGPF) of ccRCC patients was constructed via R package rms.

TCGA-KIRC project. Then we normalized the raw count data using variance stabilizing transformation through the vst function of DESeq2 package.

Extraction of CT Image Features

Tumor segmentation and feature extraction were performed using 3D slicer (Version 4.7) software. 3D slicer is an open source software platform which functions in medical image processing, analysis (including registration and interactive segmentation) and versatile visualization for image-guided therapy (24). We loaded deidentified transverse CT images (DICOM) of ccRCC into the software and conducted segmentation of area for each lesion with a paint function. The delineation of the region of interest (ROI) was firstly conducted by Xuelei Ma, an oncologist with experience in CT interpretation. To access the intra- and inter-rater feature stability against ROI delineation variations caused by human factors, Xuelei Ma and another experienced oncologist Ye Zhao conducted the delineation of the ROI again. Through the icc function of R package irr, we calculated the intraclass correlation coefficient and accessed the repeatability and stability of the radiomics features based on the ROI conducted by Xuelei Ma twice and that conducted by Ye Zhao (used for accessing the inter-rater stability of radiomics features).

Next we performed feature extraction calculations of ccRCC patients *via* pyradiomics package (https://pyradiomics. readthedocs.io/en/latest/), an extension *via* the 3D Slicer ExtensionManager. The pyradiomics is an open-source python package for the extraction of radiomics features from medical imaging, and most features are in compliance with feature definitions as described by the Imaging Biomarker Standardization Initiative (IBSI). Notes are added to specify the differences where the features vary in the website (25). Eventually, we obtained a total of 107 features in various classes. For instance, first order statistics describe the distribution of voxel intensities within the image region, including skewness, maximum, minimum, mean, range, and entropy etc. Shape-based category depicts shape eigenvalue of ROI and in 3-dimentional size. Gray Level Cooccurrence Matrix (GLCM) features and Gray Level Run Length Matrix (GLRLM) represent the eigenvalue of high-order texture characteristics. Other features extracted were contained in Gray Level Size Zone Matrix (GLSZM), Neighboring Gray Tone Difference Matrix (NGTDM) and Gray Level Dependence Matrix (GLDM).

Selection of Prognosis-Related Radiomics Features

All the ccRCC samples were randomly assigned to training and test cohorts on a scale of 1:1. Based on the training set, we applied the least absolute shrinkage and selection operator COX (LASSO-COX) and support vector machines-recursive feature elimination (SVM-RFE) algorithm in R package "glmnet" and "e1071" respectively using 5-fold cross-validation approach to filtrate prognosis-related imaging features (PRIF). LASSO-COX reduces feature space dimension and filters variables by performing a penalized function that compresses insignificant

coefficients to zero, and therefore contracts subsets and processes data with complex collinearity. The cv.glmnet function of glmnet package provides an argument for K-fold cross validation called "nfolds", and this argument was set at 0.04396 for 5-fold cross validation.

SVM arranges the extracted image features in descending order according to the variable importance and inputs them to the training model in sequence in each iteration of the cross-validation calculation, thus measuring the overall accuracy of the training sets during the accumulation course. SVM-RFE is a sequence backward selection algorithm based on the maximum interval principle of SVM. We applied the 5-fold cross validation algorithm as the resampling method for SVM-RFE. The final importance of features was based on the average importance of each feature in each iteration. Afterwards, we compared the features displayed in the outcome of two methods and selected those within the intersection of two subsets as PRIF for subsequent analyses.

Gene Co-Expression Network Analysis

To further explore the molecular biological mechanisms of the prognostic-related CT image features and obtain gene expression modules, we conducted weighted gene co-expression network analysis (WGCNA) based on training dataset. WGCNA is a systematic analytical tool which describes the correlation patterns among genes across microarray samples and clusters genes into modules, hence investigating the association between gene sets and clinical traits. The main workflow started with measuring adjacency coefficient which computes the joint strength between two nodes. Next we reduced the co-expression similarity to ensure a scale-free network. The topological overlap measure (TOM) was performed to eliminate false correlation, and then we conducted average linkage hierarchical clustering and classified functional gene modules in the co-expressed network. The module eigengenes (ME) was the first principal component of the expression matrix which represented the gene expression profile of the entire module. Afterwards we assessed the correlation between MEs and previously screened image features to identify the most relevant clinically significant module. Then to assess the preservation of the connectivity and density between each couple of modules (from the train and test networks), we carried out a permutation test through the function modulePreservation from the WGCNA package. This function provides a summary preservation Z-score for each module. Furthermore we applied Gene ontology (GO) enrichment analysis *via* Metascape (http://metascape.org) to evaluate the interlinkage between key modules.

Construction of Integrative Imaging-Genomic Prognostic Model

We utilized random forest (RF) algorithm with 1,000 decision trees (DTs) through "randomForestSRC" (rfsrc) in R to obtain optimal prognostic factors. RF algorithm constructs and assembles multiple decision trees based on data samples to attain a more precise prediction, which can reduce the over-

fitting by averaging the result. The default arguments of the rfsrc function contained a resampling method argument "bootstrap". The default value of the "bootstrap" argument was "by.root", which bootstraps the data by sampling with replacement at the root node before growing the tree. Based on training set we firstly constructed two prognostic models, one of which incorporated prognosis-related imaging features (PRIF) and the other integrated PRIF and the expressed genes profiles. The latter was defined as imaging-genomic prognostic factor (IGPF) model. Meanwhile we evaluated the prediction performance of the two models with test set using 5-fold cross-validation. Subsequently, we performed the discrimination of the signature by plotting the receiver operating curve (ROC) and calculating the corresponding area under curve (AUC) based on average accuracy of 5 iterations. ROC curve analysis obtained generalization abilities based on the means computed by all cross validation sets and the average 1-, 3-, and 5-year AUCs were then assessed. Furthermore, we calculated the risk scores for all ccRCC patients using RF, and patients were then separated into high-risk group and low-risk group based on the median cut-off value of risk scores. The overall survival (OS) of the two groups was acquired and displayed *via* Kaplan-Meier survival curve analysis and then compared by log rank test.

Univariate and multivariate Cox regression analyses were performed to further identify the predictive factors of survival outcome. Variables with $p < 0.05$ in univariate Cox regression analysis were considered statistically significant and selected for multivariate analysis. On the grounds of the results of Cox regression analysis we established a nomogram based on the training dataset, which comprised the IGPF and certain clinical factors including stage and gender through R package rms. Calibration plots were then applied based on training set to evaluate the predictive performance of the nomogram by illustrating the consistency between predicted OS and observed OS and model discrimination was estimated by the concordance index (C-index). Moreover we employed the decision curve analysis (DCA) based on training set to assess the clinical availability of the nomogram by calculating the net benefit under a range of threshold probabilities.

RESULTS

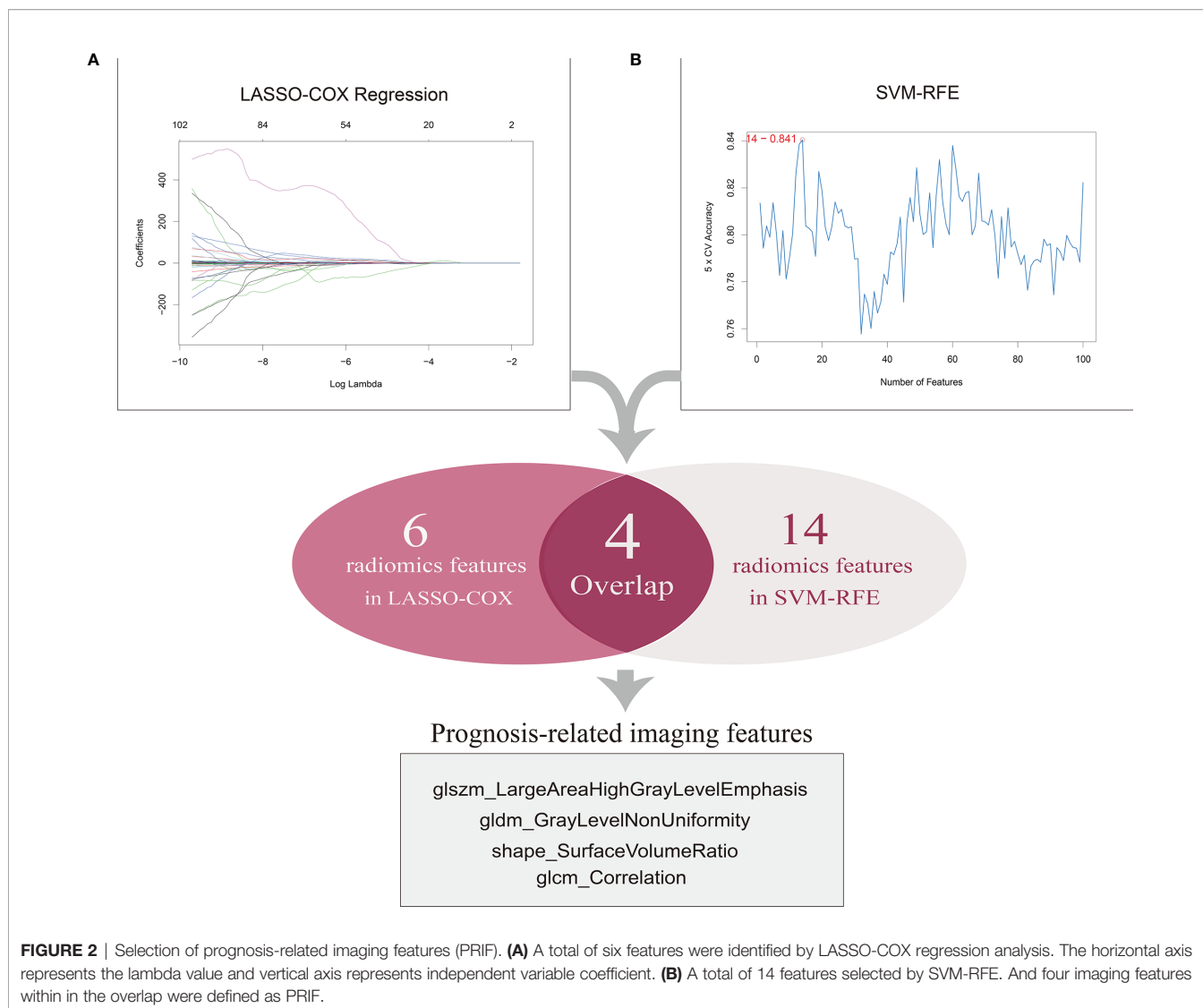### Acquisition of Prognosis-Related Radiomics Features

We initially obtained the patient data containing clinical features and mRNA sequencing data of 537 ccRCC samples from TCGA database and the matched CT images of 237 ccRCC patients from TCIA portal, among which 205 samples with available and complete data were enrolled for subsequent analyses. The patient clinical characteristics are listed in **Table 1**. The results of the repeatability and stability assessment showed that most of the radiomics features (104 of 107) were stable against ROI delineation variations caused by human factors (icc > 0.75 and $p < 0.05$). The raw data of the ROI delineation by two oncologists were presented in **Supplementary Material 1**. A total of 107

TABLE 1 | Demographic and clinical characteristics of patients.

Characteristics	Total (n=205)	Train (n=103)	Test (n=102)	P value
	NO.	NO. (%)	NO. (%)	
Age at diagnosis (years)				0.081
Mean ± SD	59.7 ± 12.2	58.4 ± 12.7	61.0 ± 11.7	
Gender				0.433
Male	134	70(68.0)	64(62.7)	
Female	71	33(32.0)	38(37.3)	
T classification				0.179
T1	109	56(54.4)	53(52.0)	
T2	22	15(14.6)	7(6.9)	
T3	70	31(30.1)	39(38.2)	
T4	4	1(0.9)	3(2.9)	
N classification				0.856
N0	83	43(41.7)	40(39.2)	
N1	5	2(1.9)	3(2.9)	
Unknown	117	58(56.4)	59(57.9)	
M classification				0.146
M0	176	93(90.3)	83(81.4)	
M1	28	10(9.7)	18(17.6)	
Unknown	1	0(0.0)	1(1.0)	
TMN stage				0.200
I	106	56(54.4)	50(49.0)	
II	18	12(11.6)	6(5.9)	
III	51	24(23.3)	27(26.5)	
IV	30	11(10.7)	19(18.6)	
Grade				0.227
G1	1	1(1.0)	0(0.0)	
G2	80	40(38.8)	40(39.2)	
G3	91	50(48.5)	41(40.2)	
G4	33	12(11.7)	21(20.6)	
OS (days)				0.090
Mean ± SD	1371.0 ± 925.1	1493.5 ± 996.7	1247.2 ± 833.5	

features of six categories were firstly extracted from original CT images from the ROIs using pyradiomics package, and the results adhered to the IBSI recommendations (**Supplementary Material 1**, icc data). To acquire a reliable and robust model, we randomly divided the ccRCC samples into a training set (n=103) and a test set (n=102) in a 1:1 ratio and proceeded to the further selection based on the training dataset. In an attempt to diminish the possibility of module overfitting by too many radiomics features and select the ones with higher prediction accuracy for OS, two machine-learning approaches including LASSO-Cox regression and SVM-RFE were employed for mutual authentication. The tuning parameter λ was settled at an optimal value of 0.04396 with the minimum criteria in LASSO regression, and 6 prognostic features were identified with nonzero coefficients out of 107 radiomics features (**Figure 2A**). As the extracted features ranked and excluded sequentially in SVM classifier during each iteration by contribution value, we found that the best prediction performance appeared when the first 14 radiomics features were included during the 5-fold cross validation (**Figure 2B**).

Therefore, the top 14 features in contribution value were filtrated as prognosis-related features for further module construction, covering six in GLCM, three in GLSZM, one in GLDM, two in shape, one in NGTDM and one in first order.
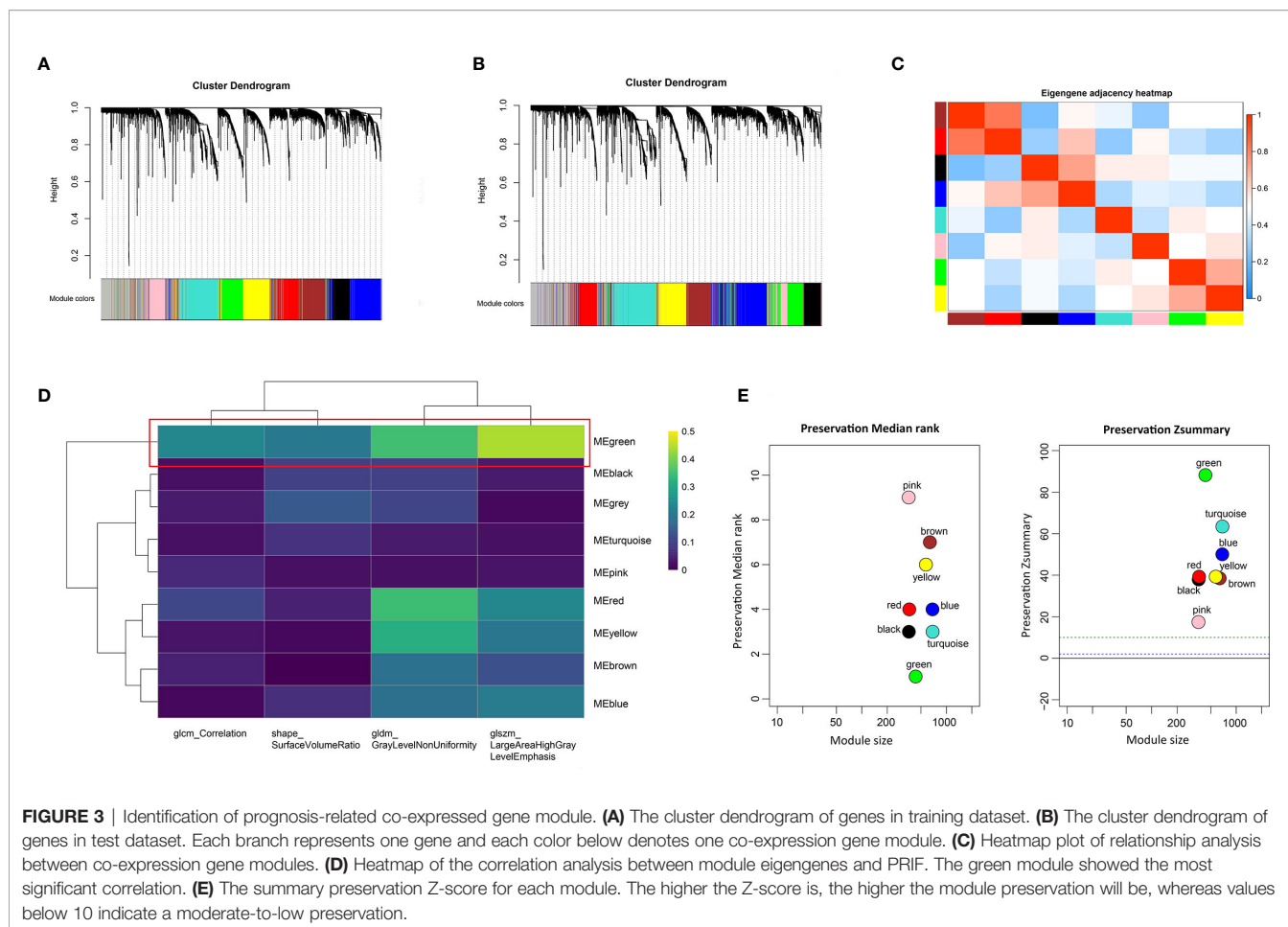
FIGURE 2 | Selection of prognosis-related imaging features (PRIF). **(A)** A total of six features were identified by LASSO-COX regression analysis. The horizontal axis represents the lambda value and vertical axis represents independent variable coefficient. **(B)** A total of 14 features selected by SVM-RFE. And four imaging features within in the overlap were defined as PRIF.

Eventually four features with predictive efficiency (glszm_Large AreaHighGrayLevelEmphasis, gldm_GrayLevelNonUniformity, shape_SurfaceVolumeRatio, glcm_Correlation) within the overlap of the results produced by the two methods were identified as prognosis-related imaging features (PRIF) (**Figure 2**).

Identification of Co-Expressed Gene Modules Related to Prognostic Image Features

To identify the gene modules highly correlated to PRIF in the ccRCC samples, we performed WGCNA to build a gene co-expression network based on training dataset. Threshold powers were set from 1 to 20 to choose an applicable soft-thresholding power, and the top 25% most variant genes (4,936 genes) ranked in descending order of SD sequence were included for subsequent analyses. A total of nine co-expressed gene modules were identified *via* the hierarchical clustering dendrogram (**Figures 3A, B**). Relationships of the modules

were illustrated in a heatmap drawn by adjacencies (**Figure 3C**). Afterwards, we conducted correlation analysis to estimate the association between nine MEs and image traits (**Figure 3D**). The correlation coefficients and FDR values between each of the nine gene modules and PRIF were displayed in **Supplementary Material 2**. Of all the nine gene co-expression modules, the green module (625 genes) displayed the most significant correlation with the prognosis-related image features of ccRCC, including glszm_LargeAreaHighGrayLevelEmphasis, gldm_GrayLevel NonUniformity, shape_SurfaceVolumeRatio and glcm_ Correlation. The module preservation analysis presented by the summary preservation Z-score showed that all the modules were rather stable and the green module was the most robust between training and test sets (**Figure 3E).** Thus we identified the green module as the key module of significant prognostic importance for continuous research.

Furthermore we carried out enrichment analysis to describe the biological interpretations of the genes in green module (**Supplementary Material 3**). As illustrated in **Figure 4**, the

FIGURE 3 | Identification of prognosis-related co-expressed gene module. **(A)** The cluster dendrogram of genes in training dataset. **(B)** The cluster dendrogram of genes in test dataset. Each branch represents one gene and each color below denotes one co-expression gene module. **(C)** Heatmap plot of relationship analysis between co-expression gene modules. **(D)** Heatmap of the correlation analysis between module eigengenes and PRIF. The green module showed the most significant correlation. **(E)** The summary preservation Z-score for each module. The higher the Z-score is, the higher the module preservation will be, whereas values below 10 indicate a moderate-to-low preservation.

genes were significantly related to certain biological processes such as blood vessel development, circulatory system process, cell morphogenesis involved in differentiation, cell-substrate adhesion, and extracellular structure organization. The results suggested that these genes may be involved in tumor angiogenesis and cell adhesion, which have been proved to be associated with tumorigenesis and progression.

Construction and Validation of Integrated Imaging-Genomic Prognostic Model

In order to establish an integrative model of PRIF and prognostic co-expressed gene profile, we applied RF algorithm based on training dataset, and furthermore performed model verification with the test dataset. Initially we presented PRIF as an independent variable to analyze its impact on prognosis and found a significant correlation. Then to explore the combined effect of genomics and imaging features, we assessed gene expression profiles in the prognostic-related green module and selected the top four genes with the highest module membership (MM) value (RPS6KA2, CYYR1, KDR, GIMAP6) (**Supplementary Material 4**, **Figure S1**).

Furthermore, we integrated the four genes with PRIF which were identified as imaging-genomic prognostic factors (IGPF) and calculated the risk score of each ccRCC patient. The patients

were divided into high-risk and low-risk groups in light of the median value of risk scores and then estimated with time-dependent ROC. To evaluate the statistical differences between different models, we applied the compare function of timeROC package in both training and test sets. The result showed that there were statistically significant differences between RPIF and IGPF models in 1-, 3-, and 5-year OS (P<0.05) (**Table 2**). The outcome illustrated a more satisfactory predictive performance of IGPF model compared to the RPIF model alone (**Table 3**). In the training set, the average AUCs for 1-, 3-, and 5-year OS were 0.845, 0.772, and 0.737 in PRIF model compared to 0.898, 0.849 and 0.808 in IGPF model respectively (**Figures 5C**, **6C**). In the test set, the average AUCs for 1-, 3-, and 5-year OS were 0.814, 0.74 and 0.689 of PRIF model compared to 0.837, 0.806 and 0.751 of the combined IGPF module (**Figures 5D**, **6D**).

Establishment and Evaluation of Nomogram Model

According to Kaplan-Meier survival curves, a distinctly significant difference of p < 0.0001 can be seen between the two groups in both test and train cohorts, and patients in the low-risk group showed a more promising OS than the high-risk group (**Figures 5A, B**, **6A, B**). In consideration of the relationship of IGPF and certain clinical predictors, we
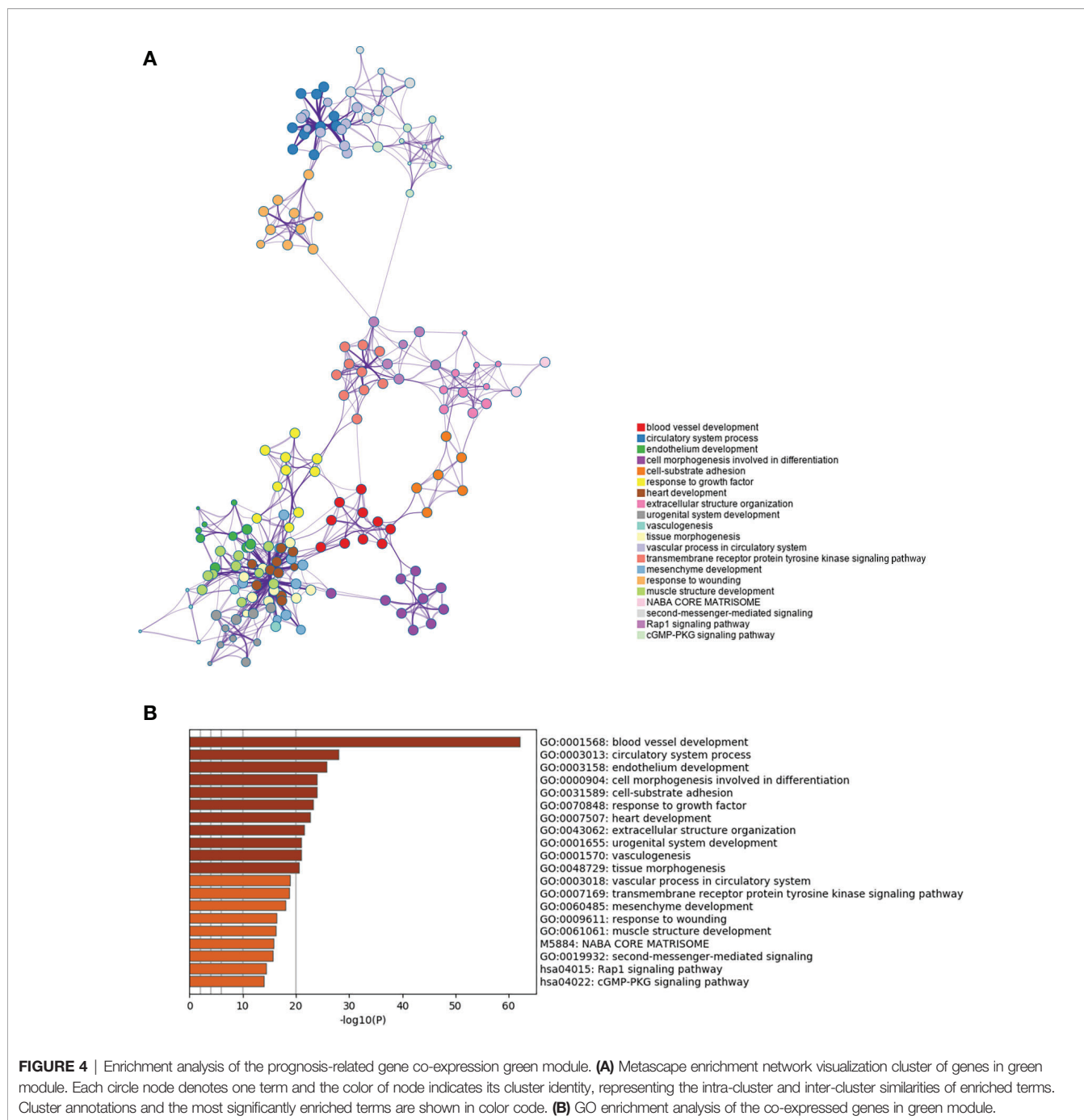
FIGURE 4 | Enrichment analysis of the prognosis-related gene co-expression green module. **(A)** Metascape enrichment network visualization cluster of genes in green module. Each circle node denotes one term and the color of node indicates its cluster identity, representing the intra-cluster and inter-cluster similarities of enriched terms. Cluster annotations and the most significantly enriched terms are shown in color code. **(B)** GO enrichment analysis of the co-expressed genes in green module.

TABLE 2 | Comparison of PRIF and IGPF models in training set and test set.

Dataset	Time (d)	P value
Train	t=365	0.294493777
(n=103)	t=1095	0.012522423
	t=1825	0.006498863
Test	t=365	0.048720526
(n=102)	t=1095	0.02381105
	t=1825	0.007957811

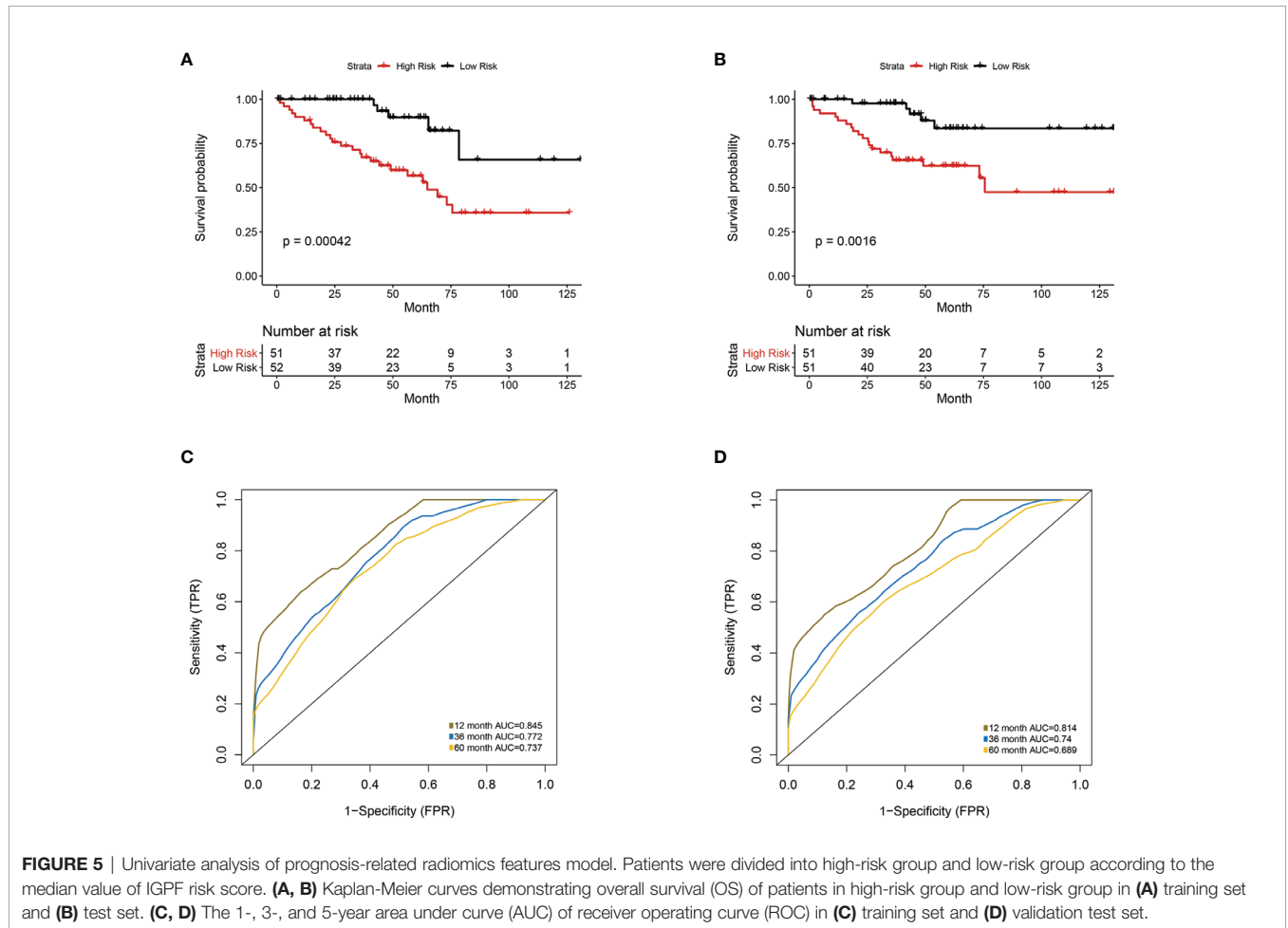*PRIF, prognosis-related imaging features; IGPF, imaging-genomic prognostic factors.*

performed univariate and multivariate Cox analysis. The results indicated that clinical characteristics including gender, TNM stage and IGPF were independent risk factors for OS of ccRCC patients. In order to acquire a quantitative prediction method for disease progression and survival probability of ccRCC, we established a nomogram on the basis of the independent predictors of OS (gender, TNM stage, and IGPF) identified earlier (**Figure 7A**). Calibration plots were then applied to assess the consistency between the nomogram-predicted values and actual values, and the calibration curves in **Figure 7B**

TABLE 3 | Survival models based on PRIF and IGPF in training set and test set.

Model		HR	z	P value	lower	upper	c-index	se(C-index)
Train	IGPF	9.555645221	3.663583845	0.000248711	2.856506317	31.96574608	0.7435393	0.0303249
(n=103)	PRIF	5.890757826	3.171763552	0.001215163	2.833927216	14.04278158	0.68764045	0.04130209
Test	IGPF	7.624785255	4.189170631	0.000027998	2.947573776	19.72379815	0.74161074	0.02908009
(n=102)	PRIF	4.461795265	3.522355167	0.000427731	1.941329799	10.25462907	0.68504314	0.03922711

PRIF, prognosis-related imaging features; IGPF, imaging-genomic prognostic factors.

FIGURE 5 | Univariate analysis of prognosis-related radiomics features model. Patients were divided into high-risk group and low-risk group according to the median value of IGPF risk score. **(A, B)** Kaplan-Meier curves demonstrating overall survival (OS) of patients in high-risk group and low-risk group in **(A)** training set and **(B)** test set. **(C, D)** The 1-, 3-, and 5-year area under curve (AUC) of receiver operating curve (ROC) in **(C)** training set and **(D)** validation test set.
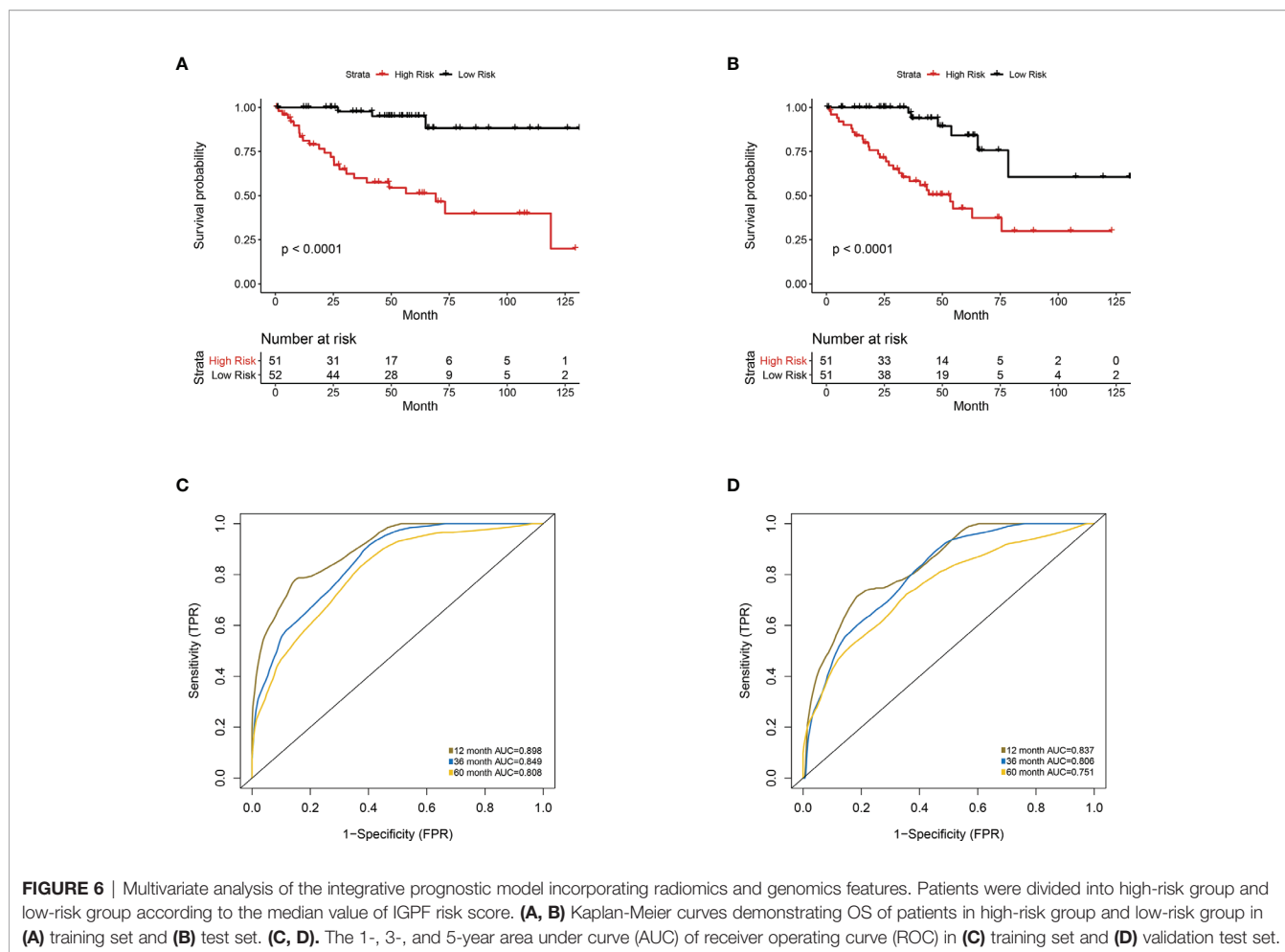
denoted good performance of 1- and 5-year nomogram model which showed a closer tendency to the 45-degree standard line. Meanwhile, the decision curves analysis evaluated the clinical utility of IGPF model containing radiomics and gene features, clinical model that involved TNM stage and gender and nomogram which integrated the former two models (**Figure 7C**). As depicted in the results, nomogram provided the best net benefit among most of the threshold probabilities range.

DISCUSSION

In this study, we extracted radiomics features from contrast-enhanced CT images of ccRCC, and subsequently selected

prognosis-related image features (PRIF) with significant prognostic value *via* several machine-learning algorithms. Furthermore we identified gene modules that are most relevant to PRIF through co-expression network. Based on the PRIF (screened by LASSO and SVM-RFE) and genes (screened by WGCNA and MM value), we constructed a robust imaging-genomic prognostic factors (IGPF) model incorporating prediction features in the two categories through random survival forest algorithm. The random survival forest algorithm acts as a bootstrap algorithm and can predict the overall survival. The OS prediction analysis demonstrated a notable performance of the integrative prognostic model, and thus the IGPF based risk score was considered as an independent prognostic factor. Afterwards, through nomogram we integrated the IGPF model

FIGURE 6 | Multivariate analysis of the integrative prognostic model incorporating radiomics and genomics features. Patients were divided into high-risk group and low-risk group according to the median value of IGPF risk score. **(A, B)** Kaplan-Meier curves demonstrating OS of patients in high-risk group and low-risk group in **(A)** training set and **(B)** test set. **(C, D).** The 1-, 3-, and 5-year area under curve (AUC) of receiver operating curve (ROC) in **(C)** training set and **(D)** validation test set.

and clinical predictor model, and then made comparisons of the three prognostic models. Ultimately, the prediction capability of the nomogram model outweighed the other two.

On the basis of the initially obtained 107 imaging features, we employed two machine-learning methods LASSO-Cox and SVM-RFE in combination aiming to achieve a group of prognostic radiomics features with more robust and accurate prediction abilities. Four conspicuous prognosis-related image features in our study were included in Gray Level Size Zone Matrix (GLSZM), Gray Level Dependence Matrix (GLDM), shape and Gray Level Cooccurrence Matrix (GLCM) respectively. As illustrated in the results, features based on intensity discretization were not screened out in the end. The results suggested that under these two unsupervised feature selection algorithms, the gray level-based features and shape-based features had a better prognostic performance than intensity discretization-based features in this cohort. However, considering the differences and limitations among multiple algorithms and cohorts, we cannot completely deny the importance of intensity discretization-based features.

A gray level zone is described as the number of connected voxels which show the same intensity. The texture feature Large Area High Gray Level Emphasis from GLSZM quantifies the

proportion in the image of the joint distribution of smaller size zones with higher gray-level values, which has been formerly adopted in the assessment of the robustness or patient response in different imageological examinations (26, 27). The GLDM-based textural feature Gray Level Non Uniformity (GLN) calculates the similarity of gray-level intensity values, where a lower GLN refers to a higher intensity value in the image (28). Surface Area to Volume Ratio is a shape feature that is not dimensionless and is partly dependent on the volume of the ROI. It has been utilized in differentiating the benign and malignant tumors based on shape and margin of the lesions (29, 30). GLCM conduces to reflecting the comprehensive information about pixel distribution containing direction, distance, gray value, and the pattern of gray level arrangement (28), and Correlation represents the linear dependency of gray level values to their respective voxels in the GLCM textural features. It has been applied previously in the evaluation of breast cancer, osteosarcoma, lung cancer and gliomas in imaging modalities such as CT, MRI, and PECT (31–35).

In our study, the predictive efficacy of the elected prognostic related radiomics features based on training set were found to be in accordance with some of the reference research above (30, 33,
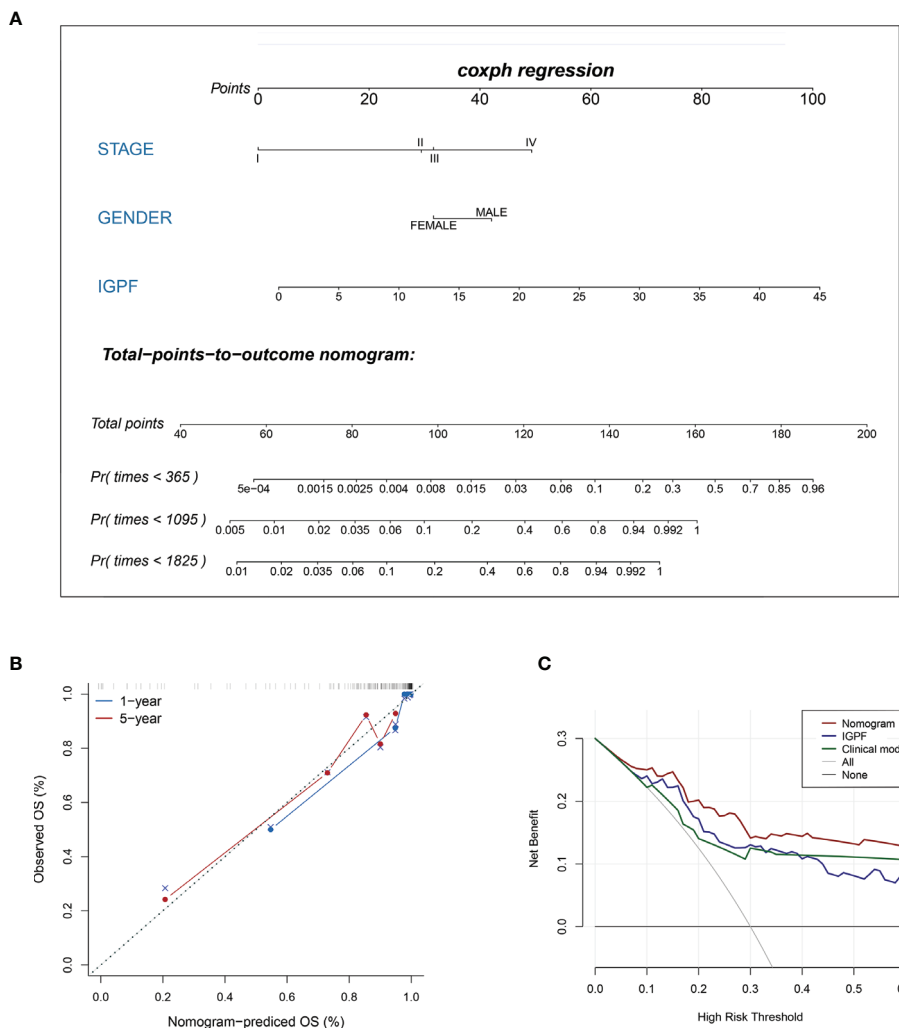
FIGURE 7 | Construction and validation of prognostic nomogram model. **(A)** The nomogram prediction of the 1-, 3-, and 5-year OS of ccRCC patients. **(B)** Calibration plots of the nomogram for 1- and 5-year OS prediction. The horizontal axis represents nomogram-predicted survival probability and the vertical axis represents actual survival. **(C)** Decision curve analyses of IGPF, clinical and nomogram model. The gray oblique line represents the net benefit of all intervening patients, and the horizontal gray line indicates the net benefit of no intervening patients. The nomogram model showed the best net benefits in the vast majority of the threshold probability range.

34, 36). However, a lot of former studies have concentrated on the performance of textural features of radiographic images, which may lack a comprehensive explanation of the biological mechanism and potential biomolecular features of the disease. While in our study, we conducted the identification of the prognostic gene co-expression module and then evaluated the association between the imaging phenotype and genomic characteristics. The results demonstrated that the green module was most related to all the PRIF, and gldm_gray level non uniformity feature could be mostly affected by gene expression pattern. In addition, the red and yellow modules also had a relatively high correlation with the gldm_gray level non uniformity feature. This may be related to the objective attributes of this feature, and further studies are still needed to explain the potential relevance and biological mechanism

between gene modules and radiomics features. Moreover, we implemented enrichment analysis in order to elaborate the latent molecular pathways relevant to the prognostic significant green gene module.

The results indicated that the most prominent enrichment leans towards pathways involved in tumor angiogenesis, cell adhesion and extracellular structure organization. Formation of new vascular networks is a pivotal step in tumor progression and also expedites the metastasis of cancer cells (37). At present, tumor microvessel density (MVD) and VEGF are important immunohistochemical indicators for tumor angiogenesis, and studies have reported that three-phase dynamic enhanced CT and MRI can be utilized as auxiliary evaluation methods for tumor angiogenesis, malignancy and prognosis in ccRCC (38–40). Cell-substrate adhesion has been widely demonstrated as an

indispensable process of metastasis *in vivo* (41). The modification of cell adhesion status has significant impact on biophysical patterns of tumor microenvironment (TME) and structure of extracellular matrix (ECM), which has been reported to be related to the prognosis of colorectal cancer, lung cancer and gastric cancer (42–45). In accordance with previous researches, the results may provide a chance to understand the upstream biological mechanisms of tumor development in ccRCC (46–48). RPS6KA2, CYYR1, KDR, and GIMAP6 were discovered to be most correlated with the prognostic-related module eigengene, which was also found relevant to blood vessel development and cell proliferation in existing researches. For instance, KDR has been reported to acts as an important mediator of VEGF-induced endothelial proliferation, tubular morphogenesis and sprouting and associate with signaling by GPCR pathway (49, 50). RPS6KA2 has been found to act downstream of EGFR, RAS, and ERK signaling, which mediates mitogenic and stress-induced activation of transcription factors and thus regulate the proliferation and differentiation of cells (51, 52).

Subsequently, we integrated the prognosis-related image features and gene profiles into an IGPF model and obtained corresponding risk scores. The clinical model took in gender and TNM stage as the common tumor assessment indicators for prognosis, but the predictive accuracy is still limited. The nomogram which integrated IGPF and clinical predictors was validated to outperform all the models with the best prediction performance.

There were several limitations to this study. First of all, the sample size was comparatively small because patients with available identified transverse CT images and gene expression profiles were limited. Secondly, the data of patients we enrolled may be incomplete, which might create discrepancies and lead to potential bias. To better promote the conclusions and understand the underlying biology molecular mechanism, a larger scale of multi-center data verification is necessarily needed. Thirdly, since we used random survival forest algorithm to build survival prognosis model in this study, the bootstrap step was a built-in process and the bootstrap corrected results could not be reported. Fourthly, more clinical trials and experimental researches are needed to assess the prove the

adaptability of the imaging-genomic prognostic model, and the molecular mechanisms remain to be further explored.

In conclusion, in this study we constructed an integrative prognosis-related model incorporating radiomics features, genomic profile and clinical indicators. The results illustrated that IGPF may improve the prognostic modalities on the basis of conventional clinical indexes, and the nomogram prediction model can serve as an advantageous measurement tool which may be conducive to personalized treatment and prognosis for ccRCC patients.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://www.cancerimagingarchive.net/; https://portal.gdc.cancer.gov.

AUTHOR CONTRIBUTIONS

YH: data curation, writing – original draft and submission. HZ: conceptualization, methodology, and software. LC: validation, writing – reviewing and editing. YL: writing – reviewing and editing. XM: conceptualization and supervision. YZ: conceptualization and supervision. All authors contributed to the article and approved the submitted version.

FUNDING

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2021. 640881/full#supplementary-material

REFERENCES

1. Hsieh JJ, Purdue MP, Signoretti S, Swanton C, Albiges L, Schmidinger M, et al. Renal cell carcinoma. *Nat Rev Dis Primers* (2017) 3:17009–9. doi: 10.1038/nrdp.2017.9
2. Jiang Y, Li W, Huang C, Tian C, Chen Q, Zeng X, et al. Preoperative CT Radiomics Predicting the SSIGN Risk Groups in Patients With Clear Cell Renal Cell Carcinoma: Development and Multicenter Validation. *Front Oncol* (2020) 10:909. doi: 10.3389/fonc.2020.00909
3. Siegel R, Miller K, Jemal A. Cancer statistics, 2019. *CA: Cancer J Clin* (2019) 69(1):7–34. doi: 10.3322/caac.21551
4. Cairns P. Renal cell carcinoma. *Cancer Biomarkers Section A Dis Markers* (2010) 9:461–73. doi: 10.3233/CBM-2011-0176
5. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical

6. images using advanced feature analysis. *Eur J Cancer* (2012) 48(4):441–6. doi: 10.1016/j.ejca.2011.11.036
6. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* (2017) 14(12):749–62. doi: 10.1038/nrclinonc.2017.141
7. Balagurunathan Y, Gu Y, Wang H, Kumar V, Grove O, Hawkins S, et al. Reproducibility and Prognosis of Quantitative Features Extracted from CT Images. *Trans Oncol* (2014) 7(1):72–87. doi: 10.1593/tlo.13844
8. Cui E, Li Z, Ma C, Li Q, Lei Y, Lan Y, et al. Predicting the ISUP grade of clear cell renal cell carcinoma with multiparametric MR and multiphase CT radiomics. *Eur Radiol* (2020) 30(5):2912–21. doi: 10.1007/s00330-019-06601-1
9. Shu J, Wen D, Xi Y, Xia Y, Cai Z, Xu W, et al. Clear cell renal cell carcinoma: Machine learning-based computed tomography radiomics analysis for the

prediction of WHO/ISUP grade. *Eur J Radiol* (2019) 121:108738. doi: 10.1016/j.ejrad.2019.108738

10. Goyal A, Razik A, Kandasamy D, Seth A, Das P, Ganeshan B, et al. Role of MR texture analysis in histological subtyping and grading of renal cell carcinoma: a preliminary study. *Abdom Radiol (NY)* (2019) 44(10):3336–49. doi: 10.1007/s00261-019-02122-z

11. Coy H, Young JR, Pantuck AJ, Douek ML, Sisk A, Magyar C, et al. Association of tumor grade, enhancement on multiphasic CT and microvessel density in patients with clear cell renal cell carcinoma. *Abdom Radiol (NY)* (2020) 45(10):3184–92. doi: 10.1007/s00261-019-02271-1

12. Ma Y, Cao F, Xu X, Ma W. Can whole-tumor radiomics-based CT analysis better differentiate fat-poor angiomyolipoma from clear cell renal cell caricinoma: compared with conventional CT analysis? *Abdom Radiol (NY)* (2020) 45(8):2500–7. doi: 10.1007/s00261-020-02414-9

13. Nie P, Yang G, Wang Z, Yan L, Miao W, Hao D, et al. A CT-based radiomics nomogram for differentiation of renal angiomyolipoma without visible fat from homogeneous clear cell renal cell carcinoma. *Eur Radiol* (2020) 30(2):1274–84. doi: 10.1007/s00330-019-06427-x

14. Goh V, Ganeshan B, Nathan P, Juttla JK, Vinayan A, Miles KA. Assessment of response to tyrosine kinase inhibitors in metastatic renal cell cancer: CT texture as a predictive biomarker. *Radiology* (2011) 261(1):165–71. doi: 10.1148/radiol.11110264

15. Schieda N, Thornhill RE, Al-Subhi M, McInnes MD, Shabana WM, van der Pol CB, et al. Diagnosis of Sarcomatoid Renal Cell Carcinoma With CT: Evaluation by Qualitative Imaging Features and Texture Analysis. *AJR Am J Roentgenol* (2015) 204(5):1013–23. doi: 10.2214/AJR.14.13279

16. Peng J, Zhang J, Zhang Q, Xu Y, Zhou J, Liu L. A radiomics nomogram for preoperative prediction of microvascular invasion risk in hepatitis B virus-related hepatocellular carcinoma. *Diagn Interv Radiol* (2018) 24(3):121–7. doi: 10.5152/dir.2018.17467

17. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* (2014) 5:4006. doi: 10.1038/ncomms5644

18. Lv J, Zhang H, Ma J, Ma Y, Gao G, Song Z, et al. Comparison of CT radiogenomic and clinical characteristics between EGFR and KRAS mutations in lung adenocarcinomas. *Clin Radiol* (2018) 73(6):590.e1–8. doi: 10.1016/j.crad.2018.01.009

19. Itakura H, Achrol AS, Mitchell LA, Loya JJ, Liu T, Westbroek EM, et al. Magnetic resonance image features identify glioblastoma phenotypic subtypes with distinct molecular pathway activities. *Sci Transl Med* (2015) 7(303):303ra138. doi: 10.1126/scitranslmed.aaa7582

20. Seow P, Wong JHD, Ahmad-Annuar A, Mahajan A, Abdullah NA, Ramli N. Quantitative magnetic resonance imaging and radiogenomic biomarkers for glioma characterisation: a systematic review. *Br J Radiol* (2018) 91(1092):20170930. doi: 10.1259/bjr.20170930

21. Kohannim O, Hibar DP, Jahanshad N, Stein JL, Hua X, Toga AW, et al. Predicting Temporal Lobe Volume On Mri From Genotypes Using L(1)-L(2) Regularized Regression. *Proc IEEE Int Symp BioMed Imaging* (2012) 1160–3. doi: 10.1109/ISBI.2012.6235766

22. Gill D, Hahn A, Hale P, Maughan B. Overview of Current and Future First-Line Systemic Therapy for Metastatic Clear Cell Renal Cell Carcinoma. *Curr Treat Options Oncol* (2018) 19(1):6. doi: 10.1007/s11864-018-0517-1

23. Allen E, Jabouille A, Rivera L, Lodewijckx I, Missiaen R, Steri V, et al. Combined antiangiogenic and anti-PD-L1 therapy stimulates tumor immunity through HEV formation. *Sci Trans Med* (2017) 9(385):eaak9679. doi: 10.1126/scitranslmed.aak9679

24. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J-C, Pujol S, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging* (2012) 30(9):1323–41. doi: 10.1016/j.mri.2012.05.001

25. van Griethuysen J, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res* (2017) 77(21):e104–7. doi: 10.1158/0008-5472.CAN-17-0339

26. Edalat-Javid M, Shiri I, Hajianfar G, Abdollahi H, Arabi H, Oveisi N, et al. Cardiac SPECT radiomic features repeatability and reproducibility: A multi-scanner phantom study. *J Nucl Cardiol Off Publ Am Soc Nucl Cardiol* (2020). doi: 10.1007/s12350-020-02109-0

27. Ekert K, Hinterleitner C, Baumgartner K, Fritz J, Horger M. Extended Texture Analysis of Non-Enhanced Whole-Body MRI Image Data for Response Assessment in Multiple Myeloma Patients Undergoing Systemic Therapy. *Cancers* (2020) 12(3):761. doi: 10.3390/cancers12030761

28. Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. *IEEE Trans Syst Man Cybern* (1973) SMC-3(6):610–21. doi: 10.1109/TSMC.1973.4309314

29. Cuocolo R, Stanzione A, Ponsiglione A, Romeo V, Verde F, Creta M, et al. Clinically significant prostate cancer detection on MRI: A radiomic shape features study. *Eur J Radiol* (2019) 116:144–9. doi: 10.1016/j.ejrad.2019.05.006

30. Kim J, Choi S, Lee S, Lee H, Park H. Predicting Survival Using Pretreatment CT for Patients With Hepatocellular Carcinoma Treated With Transarterial Chemoembolization: Comparison of Models Using Radiomics. *AJR Am J Roentgenol* (2018) 211(5):1026–34. doi: 10.2214/AJR.18.19507

31. Baidya Kayal E, Kandasamy D, Khare K, Bakhshi S, Sharma R, Mehndiratta A. Texture analysis for chemotherapy response evaluation in osteosarcoma using MR imaging. *NMR Biomed* (2020) 34(2):e4426. doi: 10.1002/nbm.4426

32. Chekouo T, Mohammed S, Rao A. A Bayesian 2D functional linear model for gray-level co-occurrence matrices in texture analysis of lower grade gliomas. *NeuroImage Clin* (2020) 28:102437. doi: 10.1016/j.nicl.2020.102437

33. Coroller T, Agrawal V, Narayan V, Hou Y, Grossmann P, Lee S, et al. Radiomic phenotype features predict pathological response in non-small cell lung cancer. *Radiother Oncol J Eur Soc Ther Radiol Oncol* (2016) 119(3):480–6. doi: 10.1016/j.radonc.2016.04.004

34. Pyka T, Bundschuh R, Andratschke N, Mayer B, Specht H, Papp L, et al. Textural features in pre-treatment [F18]-FDG-PET/CT are correlated with risk of local recurrence and disease-specific survival in early stage NSCLC patients receiving primary stereotactic radiation therapy. *Radiat Oncol (London England)* (2015) 10:100. doi: 10.1186/s13014-015-0407-7

35. Tsarouchi M, Vlachopoulos G, Karahaliou A, Vassiou K, Costaridou L. Multi-parametric MRI lesion heterogeneity biomarkers for breast cancer diagnosis. *Phys Med PM Int J Devoted Appl Phys Med Biol Off J Ital Assoc Biomed Phys (AIFB)* (2020) 80:101–10. doi: 10.1016/j.ejmp.2020.10.007

36. Xu F, Zhu W, Shen Y, Wang J, Xu R, Qutesh C, et al. Radiomic-Based Quantitative CT Analysis of Pure Ground-Glass Nodules to Predict the Invasiveness of Lung Adenocarcinoma. *Front Oncol* (2020) 10:872. doi: 10.3389/fonc.2020.00872

37. Weis SM, Cheresh DA. Tumor angiogenesis: molecular pathways and therapeutic targets. *Nat Med* (2011) 17(11):1359–70. doi: 10.1038/nm.2537

38. Huang J, Yao X, Zhang J, Dong B, Chen Q, Xue W, et al. Hypoxia-induced downregulation of miR-30c promotes epithelial-mesenchymal transition in human renal cell carcinoma. *Cancer Sci* (2013) 104(12):1609–17. doi: 10.1111/cas.12291

39. Xie C, Schwarz EM, Sampson ER, Dhillon RS, Li D, O'Keefe RJ, et al. Unique angiogenic and vasculogenic properties of renal cell carcinoma in a xenograft model of bone metastasis are associated with high levels of vegf-a and decreased ang-1 expression. *J Orthop Res* (2012) 30(2):325–33. doi: 10.1002/jor.21500

40. Gigli F, Zattoni F, Zamboni G, Valotto C, Bernardin L, Mucelli RP, et al. [Correlation between pathologic features and perfusion CT of renal cancer: a feasibility study]. *Urologia* (2010) 77(4):223–31. doi: 10.1177/039156031007700401

41. Langley R, Fidler I. Tumor cell-organ microenvironment interactions in the pathogenesis of cancer metastasis. *Endocr Rev* (2007) 28(3):297–321. doi: 10.1210/er.2006-0027

42. Ling B, Liao X, Huang Y, Liang L, Jiang Y, Pang Y, et al. Identification of prognostic markers of lung cancer through bioinformatics analysis and in vitro experiments. *Int J Oncol* (2020) 56(1):193–205. doi: 10.3892/ijo.2019.4926

43. Qiu X, Feng J, Qiu J, Liu L, Xie Y, Zhang Y, et al. ITGBL1 promotes migration, invasion and predicts a poor prognosis in colorectal cancer. *Biomed Pharmacother = Biomed Pharmacother* (2018) 104:172–80. doi: 10.1016/j.biopha.2018.05.033

44. Tampakis A, Tampaki E, Nonni A, Tsourouflis G, Posabella A, Patsouris E, et al. L1CAM expression in colorectal cancer identifies a high-risk group of patients with dismal prognosis already in early-stage disease. *Acta Oncol (Stockholm Sweden)* (2020) 59(1):55–9. doi: 10.1080/0284186X.2019.1667022

45. Wang Y, Li L, Zhao Z, Wang Y, Ye Z, Tao H. L1 and epithelial cell adhesion molecules associated with gastric cancer progression and prognosis in examination of specimens from 601 patients. *J Exp Clin Cancer Res CR* (2013) 32:66. doi: 10.1186/1756-9966-32-66

46. Ho T, Serie D, Parasramka M, Cheville J, Bot B, Tan W, et al. Differential gene expression profiling of matched primary renal cell carcinoma and metastases reveals upregulation of extracellular matrix genes. *Ann Oncol Off J Eur Soc Med Oncol* (2017) 28(3):604–10. doi: 10.1093/annonc/mdw652

47. Park J, Scherer P. Adipocyte-derived endotrophin promotes malignant tumor progression. *J Clin Invest* (2012) 122(11):4243–56. doi: 10.1172/JCI63930

48. Majo S, Courtois S, Souleyreau W, Bikfalvi A, Auguste P. Impact of Extracellular Matrix Components to Renal Cell Carcinoma Behavior. *Front Oncol* (2020) 10:625. doi: 10.3389/fonc.2020.00625

49. Chen J, Chen J, He F, Huang Y, Lu S, Fan H, et al. Design of a Targeted Sequencing Assay to Detect Rare Mutations in Circulating Tumor DNA. *Genet Test Mol Biomarkers* (2019) 23(4):264–9. doi: 10.1089/gtmb.2018.0173

50. Han Y, Wang L, Wang Y. Integrated Analysis of Three Publicly Available Gene Expression Profiles Identified Genes and Pathways Associated with Clear Cell Renal Cell Carcinoma. *Med Sci Monit Int Med J Exp Clin Res* (2020) 26:e919965. doi: 10.12659/MSM.919965

51. Milosevic N, Kühnemuth B, Mühlberg L, Ripka S, Griesmann H, Lölkes C, et al. Synthetic lethality screen identifies RPS6KA2 as modifier of epidermal growth factor receptor activity in pancreatic cancer. *Neoplasia (New York NY)* (2013) 15(12):1354–62. doi: 10.1593/neo.131660

52. Pancholi S, Lykkesfeldt A, Hilmi C, Banerjee S, Leary A, Drury S, et al. ERBB2 influences the subcellular localization of the estrogen receptor in tamoxifen-resistant MCF-7 cells leading to the activation of AKT and RPS6KA2. *Endocr Relat Cancer* (2008) 15(4):985–1002. doi: 10.1677/ERC-07-0240