Sarcomatoid renal cell carcinoma prognosis prediction based on the machine learning algorithm

Rui Zhang[1], Xuefei Qin[2], Xuelin Gao[1], Yu Zheng[3], Guangdong Hou[1], Yueyue Zhang[1], Yuchen Tian[1], Yuliang Wang[1], Fuli Wang[4], and Shuaijun Ma[1]

[1]Xijing Hospital
[2]University of Edinburgh School of Arts Culture and Environment
[3]Xijing Hospital, Fourth Military Medical University
[4]Xijing Hospital, The Fourth Military Medical University

July 10, 2023

Abstract

Abstract Background There is currently no robust prognostic model for sarcomatous renal cell carcinoma (sRCC), which could help physicians make better decisions. Objectives To build an accurate predictive model for patients who have sRCC by investigating the important characteristics that influence the overall survival of patients. Design and Methods The Surveillance, Epidemiology and Results (SEER) database of the U.S. National Cancer Institute was used for gathering the dataset of sRCC patients. Following data preprocessing, the data was separated into the training set and the test set in an 8:2 ratio. Mann-Whitney U test and Chi-square test were used to verify whether the data set was evenly divided. Univariate Cox proportional hazard model, Kaplan-Meier analysis and machine learning (ML) algorithm were employed to identify the risk features on overall survival (OS). 10 reliable features were selected to construct six ML models. Model performance, predictive accuracy, and clinical benefits were evaluated by the receiver operating characteristic curves (ROC), calibration plots, and decision curve analysis (DCA) respectively. Results After data preprocessing, 692 patients with sRCC from 1975 to 2019 were included in this study. Ten variables including stage group, T stage, M stage, age, surgery, N stage, tumor size, chemotherapy, histological grade, and radiotherapy were selected as reliable features for machine learning model training. All the models show good prediction performance, among which XGBoost has the best prediction accuracy and stability. The DCA showed that all models except Adaboost could be used to support clinical decision-making with the 90-day, 1-, 2-, 3- and 5-year OS model. Conclusions Six machine learning models were developed to predict 90-day, 1-, 2-, 3- and 5-year overall survival in patients with sRCC. Model evaluations showed that the XGBoost model had the best predictive accuracy and clinical net benefit. These models can help make treatment decisions for patients with sRCC.

Hosted file

`Figures(full-size).rar` available at https://authorea.com/users/637521/articles/653791-sarcomatoid-renal-cell-carcinoma-prognosis-prediction-based-on-the-machine-learning-algorithm

Title

Sarcomatoid renal cell carcinoma prognosis prediction based on the machine learning algorithm

Contributor Information
Rui Zhang[1*], Xuefei Qin[2*], Xuelin Gao[1*], Yu Zheng[1,3*], Guangdong Hou[1], Yueyue Zhang[1], Yuchen Tian[1], Yuliang Wang[1]
Corresponding author
Shuaijun Ma[1], Fuli Wang[1]
1. Department of Urology, Xijing hospital, Air Force Medical University, Xi'an, China
2. Edinburgh College of Art, University of Edinburgh, Edinburgh, United Kingdom
3. Medical Innovation Center, Air Force Medical University, Xi'an, China
*. These authors contributed equally to this work and should be considered co-first authors.

Correspondence to:
Shuaijun Ma, Xijing hospital, Air Force Medical University, Xi'an, 710032, China
mashuaijun9@163.com
Fuli Wang, Xijing hospital, Air Force Medical University, Xi'an, 710032, China
wangfuli98@163.com

Abstract

Background

There is currently no robust prognostic model for sarcomatous renal cell carcinoma (sRCC), which could help physicians make better decisions.

Objectives

To build an accurate predictive model for patients who have sRCC by investigating the important characteristics that influence the overall survival of patients.

Design and Methods

The Surveillance, Epidemiology and Results (SEER) database of the U.S. National Cancer Institute was used for gathering the dataset of sRCC patients. Following data preprocessing,

the data was separated into the training set and the test set in an 8:2 ratio. Mann-Whitney U test and Chi-square test were used to verify whether the data set was evenly divided. Univariate Cox proportional hazard model, Kaplan-Meier analysis and machine learning (ML) algorithm were employed to identify the risk features on overall survival (OS). 10 reliable features were selected to construct six ML models. Model performance, predictive accuracy, and clinical benefits were evaluated by the receiver operating characteristic curves (ROC), calibration plots, and decision curve analysis (DCA) respectively.

Results

After data preprocessing, 692 patients with sRCC from 1975 to 2019 were included in this study. Ten variables including stage group, T stage, M stage, age, surgery, N stage, tumor size, chemotherapy, histological grade, and radiotherapy were selected as reliable features for machine learning model training. All the models show good prediction performance, among which XGBoost has the best prediction accuracy and stability. The DCA showed that all models except Adaboost could be used to support clinical decision-making with the 90-day, 1-, 2-, 3- and 5-year OS model.

Conclusions

Six machine learning models were developed to predict 90-day, 1-, 2-, 3- and 5-year overall survival in patients with sRCC. Model evaluations showed that the XGBoost model had the best predictive accuracy and clinical net benefit. These models can help make treatment decisions for patients with sRCC.

Keywords: sarcomatoid renal cell carcinoma, prognostic model, machine learning, predictive model, SEER

1. Introduction

Renal cell carcinoma (RCC) is a common malignancy of the genitourinary system, with global morbidity and mortality rates steadily increasing over the past decades. According to the estimates of Cancer Statistics, 81800 new cases of RCC will be diagnosed and cause 14890 deaths in the United States in 2023, representing the 8th most prevalent cancer[1]. Sarcomatoid RCC (sRCC) is an uncommon variant (4-5%) of RCC composed of atypical spindle cells and resembling any type of sarcoma[2]. Sarcomatoid transformation in RCC is distinguished by a transformational development pattern of the epithelial neoplasm into malignant spindle-shaped cells and is extremely invasive, and it can occur in any subtype of RCC[3]. sRCC often presents as an advanced or metastatic disease and carries a median survival of 6-13 months[3,4]. Compared with conventional RCC, the complexity of sRCC make it highly challenging to forecast and treat[3]. Thus, accurate prognostic models of sRCC could

aid patients understand their expected lifespan as well as help clinicians further guide appropriate treatment and care planning[5].

Previous studies on the prognosis of sRCC have mainly focused on the identification of risk factors that influence survival, like T2LLA volume and tumor size[6], the presence of distant metastases[7], and the clinic stage[8]. Traditional TNM staging is not effective in distinguishing the prognosis of sRCC patients. Consequently, a more effective and precise model is required for patient tracking and therapy options. Only a few studies have analyzed sRCC cases, and constructed and validated prognosis models based on nomogram[9,10]. However, these studies rely on conventional statistical methods, have a small number of samples, did not evaluate non-clinical factors, and have limited predictive accuracy.

Machine learning (ML), a new interdisciplinary technique, is effective in finding, analyzing, and summarising significant patterns from enormous biological datasets and reliably forecasting outcomes[11]. There are a growing number of research programs using ML algorithms to build prognosis models for their excellent performance at associating large amounts of datasets and high predictive accuracy[12,13].

For these reasons, this study compared the performance of classifiers using six ML models, including XGBoost (XGB), Light Gradient Boosting Machine (LGBM), Logistic Regression (LR), Gradient Boost Decision Tree (GBDT), Random Forest (RF), and AdaBoost (Ada), aimed to investigate the reliable predictors on prognosis and construct an accurate ML model to predict the overall survival (OS) rate of sRCC patients.

2. Materials and Methodology

2.1 Study Population

The patients' information was acquired from SEER Research Plus Data, 8 Registries, Nov 2021 Sub (1975-2019) using the SEER*Stat 8.4.0.1 software. The diagnosis of Sarcomatoid RCC was based on the histologic type code ICD-0-3 8318 (RCC, sarcomatoid). Exclusion criteria were (a) patients with missing values for the variables selected in 2.2. (b) patients who were alive at the end of the follow-up period but survived for less than 5 years. The original patient data included 722 cases, The raw dataset was processed, and 30 samples were removed. The final dataset used for model construction included 692 patients.

2.2 Variable Selection and Endpoints

Some demographic and clinical characteristics commonly used in cancer prognosis were selected as variables for preliminary analysis.14 variables were included in this study, including AJCC TNM stage, sex, race, stage group, tumor size, laterality, marital status,

histological grade, and whether patients have had surgery, radiotherapy, and chemotherapy. The 90-day, 1-, 2-, 3- and 5-year OS were defined as the endpoints for ML models.

2.3 Statistical Analysis and Model Building

This study was conducted in Python (version 3.9.12) and SPSS (version 26). First, data preprocessing: features extraction, multi-categorical variables transformation respectively, and target attribute data classification. The processed data set is randomly divided into a training set and a test set in a ratio of 8:2. Second, compared the difference between the training set and test set utilizing the Mann-Whitney U test and the chi-square test. Third, ML models and the univariate Cox proportional risk model were used to calculate the importance of each variable and make comparisons. Variables with high weight values were selected as reliable features and used for model construction. Finally, use the training set to build the ML models and evaluate them with the test set. The evaluation of performance included three parts. First, model discrimination was measured using a receiver operating characteristic (ROC) curve analysis, and the area under the ROC curve's (AUC) predictive accuracy was evaluated. Second, calibration plots were employed in this study to show the calibration and the degree to which the model's predictions differed from the actual event. Third, clinical applicability was evaluated using DCA, which could calculate the net benefit of a model by comparing true- and false-positive rates and weighting those results by the chances of the chosen threshold likelihood of associated hazards.

3. Results

3.1 Characteristics of the patients

Table 1 lists the baseline characteristics of the 692 sRCC patients that were included in this study. The overall survival curve for study patients fell substantially faster before the 1-year cut-off, compared to a slightly slower drop between 1 and 3 years and a moderate decline after 5 years (Figure 1). Thus, predicting 90-day, 1-, 2-, 3- and 5-year OS of sRCC patients is clinically useful for treatment planning. The average survival time of patients was 25.91 months (median, 7 months; range: 0-224 months). The patients ranged in age from 16 to 95 years old, with a mean and median age of 63.08 and 64 years respectively. There were 210 female patients (30.3%) and 482 male patients (69.7%). 573 (82.8%) patients were White Americans, 62 (9.0%) were African Americans, and 57 (8.2%) were Asian or Pacific Islanders. The majority of patients (65.5%) were married and 91 (13.2%) were never married. Among all patients, 3.9% (n = 27), 9.4% (n = 65), 28.6% (n = 198), and 58.1% (n = 402) had histological grade I, II, III, or IV, respectively. Up to 69.5% of patients had tumors with a pT3 or higher at the time of diagnosis, and 55.2% had metastatic disease. The distribution of patients in the stage group was as follows: group I (8.5%), group II (12.9%), group III (19.8%), group IV (58.8%). The proportion of patients receiving surgery, radiotherapy, and

chemotherapy was 64.9%, 19.2%, and 28.9% respectively. 69.2% of patients had tumors larger than 70 mm. Table 1 includes further information on the baseline data.

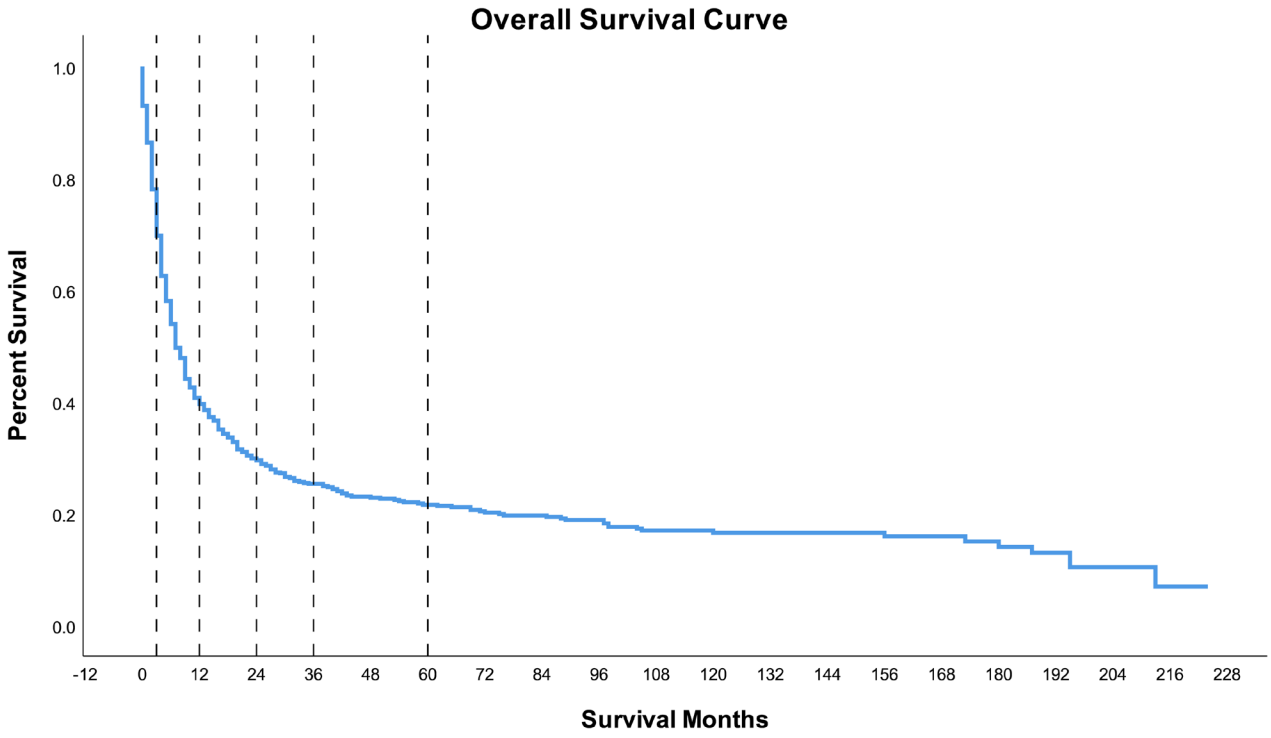


Figure 1: Overall survival curve of the 692 sRCC patients. The black dotted lines from left to right are the 90-day, 1-, 2-, 3- and 5-year cut-offs respectively.

Table 1:The baseline characteristics of the patients in this study

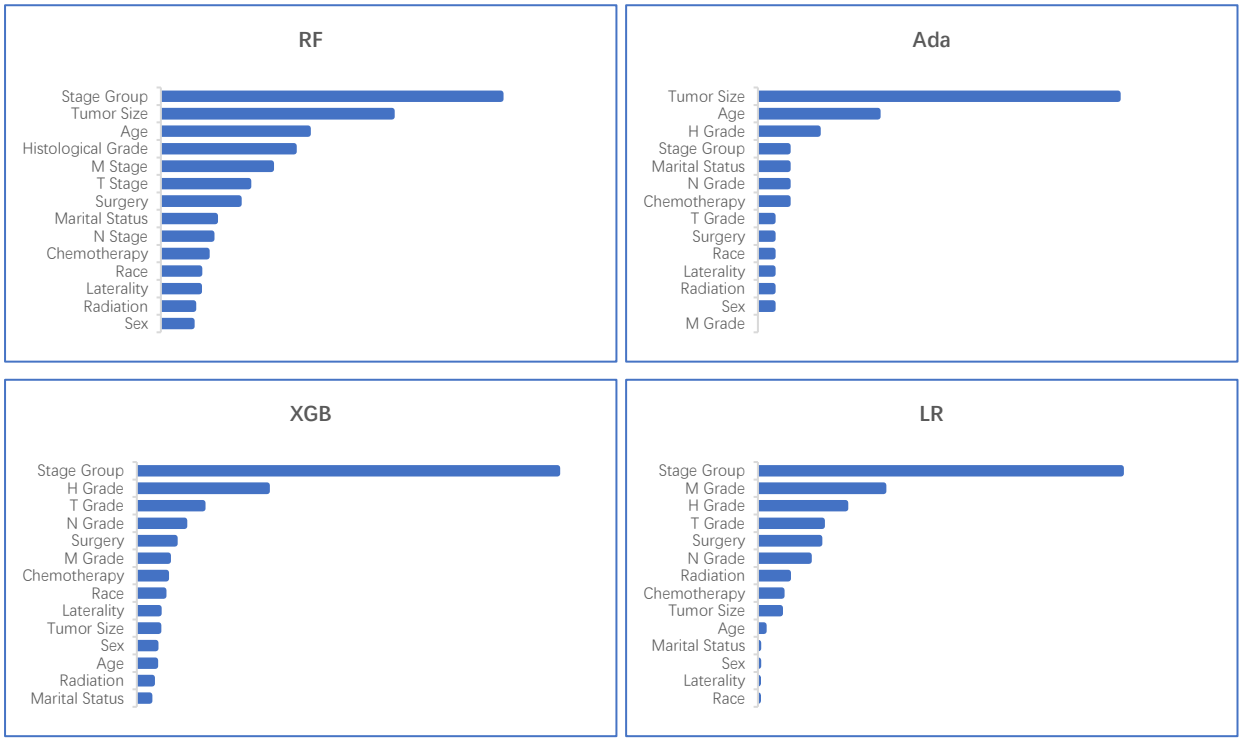
Characteristic	Total	Training set	Test set	p value
Total	692(100%)	553(80%)	139(20%)	
Survival month	25.91±43.485	24.79±41.700	30.38±49.876	0.725
Age at diagnosis				0.912
≥ 65	331(47.8%)	264(47.7%)	67(48.2%)	
< 65	361(52.2%)	289(52.3%)	72(51.8%)	
Race				0.332
White	573(82.8%)	448(81.0%)	125(89.9%)	
Black	62(9.0%)	53(9.6%)	9(6.5%)	
Other	57(8.2%)	52(9.4%)	5(3.6%)	
Sex				0.159
Female	210(30.3%)	161(29.1%)	49(35.3%)	
Male	482(69.7%)	392(70.9%)	90(64.7%)	
Marital status				0.812

Married	453(65.5%)	358(64.7%)	95(68.3%)	
Divorced	74(10.7%)	59(10.7%)	15(10.8%)	
Never Married	91(13.2%)	74(13.4%)	17(12.2%)	
Separated	12(1.7%)	11(2.0%)	1(0.7%)	
Widowed	62(9.0%)	51(9.2%)	11(7.9%)	
Histological grade				0.147
Grade I	27(3.9%)	21(3.8%)	6(4.3%)	
Grade II	65(9.4%)	45(8.1%)	20(14.4%)	
Grade III	198(28.6%)	162(29.3%)	36(25.9%)	
Grade IV	402(58.1%)	325(58.8%)	77(55.4%)	
T Stage				0.397
T1	92(13.3%)	68(12.3%)	24(17.3%)	
T2	119(17.2%)	94(17.0%)	25(18.0%)	
T3	311(44.9%)	255(46.1%)	56(40.3%)	
T4	170(24.6%)	136(24.6%)	34(24.5%)	
N Stage				0.166
N0	416(60.1%)	337(60.9%)	79(56.8%)	
N1	173(25.0%)	130(23.5%)	43(30.9%)	
N2	103(14.9%)	86(15.6%)	17(12.2%)	
M Stage				0.665
M0	310(44.8%)	250(45.2%)	60(43.2%)	
M1	382(55.2%)	303(54.8%)	79(56.8%)	
Stage Group				0.373
Group I	59(8.5%)	44(8.0%)	15(10.8%)	
Group II	89(12.9%)	68(12.3%)	21(15.1%)	
Group III	137(19.8%)	115(20.8%)	22(15.8%)	
Group IV	407(58.8%)	326(59.0%)	81(58.3%)	
Laterality				0.250
Left	371(53.6%)	303(54.8%)	68(48.9%)	
Right	317(45.8%)	247(44.7%)	70(50.4%)	
Bilateral	4(0.6%)	3(0.5%)	1(0.7%)	
Surgery				0.576
Yes	449(64.9%)	356(64.4%)	93(66.9%)	
No/Unknown	243(35.1%)	197(35.6%)	46(33.1%)	
Radiation				0.513

Yes	133(19.2%)	109(19.7%)	24(17.3%)	
No/Unknown	559(80.8%)	444(80.3%)	115(82.7%)	
Chemotherapy				0.649
Yes	200(28.9%)	162(29.3%)	38(27.3%)	
No/Unknown	492(71.1%)	391(70.7%)	101(72.7%)	
Tumor Size				0.821
0-40 mm	68(9.8%)	56(10.1%)	12(8.6%)	
40.1-70 mm	145(21%)	109(19.7%)	36(25.9%)	
70.1-100 mm	213(30.8%)	178(32.2%)	35(25.2%)	
>100 mm	266(38.4%)	210(38.0%)	56(40.3%)	

3.2 Reliable features and survival analysis

ML algorithms are used to determine the importance of the variables, and the larger the gain value obtained for each variable, the more important it is for predicting the target. Although the relative importance of each feature varies between different machine algorithms, a similar pattern emerges. Ten characteristics, including stage group, tumor size, T stage, N stage, M stage, age, histological stage, surgery, chemotherapy, and radiation, are ordinarily of greater significance (Figure 2). The risk variables defined by the univariate Cox proportional risk model include histological grade ($p < 0.001$), stage group ($p < 0.001$), surgery ($p < 0.001$), chemotherapy ($p < 0.001$), and tumor size ($p=0.001$).



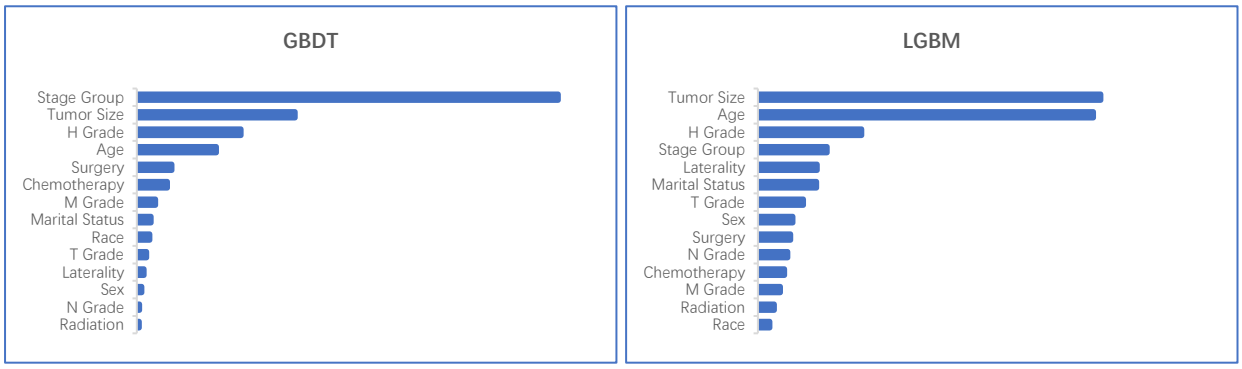


Figure 2: Relative importance for input features estimated by different machine learning algorithms

Table 2: Kaplan–Meier analysis and univariate Cox regression of overall survival for sRCC patients

Characteristic	90-day OS%	1-year OS%	2-year OS%	3-year OS%	5-year OS%	Kaplan-Meier		COX	
						Log Rank χ2 test	p value	HR (95% CI)	p value
Total	67.3	36.7	26.0	19.9	14.2				
Age at diagnosis						4.487	0.034		
≥ 65	63.4	34.1	21.8	22.7	11.8			0.872[0.726-1.047]	0.142
< 65	70.9	39.1	29.9	16.9	16.3			Reference	
Race						0.598	0.742		
White	68.4	36.1	25.7	20.1	14.3			Reference	
Black	56.5	38.7	32.3	21.0	17.7			0.887[0.642-1.226]	0.468
Other	68.4	40.4	22.8	17.5	8.8			0.934[0.598-1.458]	0.763
Sex						1.353	0.245		
Female	69.0	39.5	26.2	21.0	15.7			Reference	
Male	66.6	35.5	25.9	19.5	13.5			0.767[0.628-0.935]	0.009
Marital status						1.799	0.773		
Married	68.2	36.0	25.8	19.9	13.9			Reference	
Divorced	67.6	35.1	31.1	25.7	21.6			1.076[0.770-1.504]	0.667
Never Married	70.3	38.5	28.6	19.8	13.2			1.169[0.762-1.794]	0.475
Separated	58.3	41.6	0	0	0			0.982[0.650-1.483]	0.931
Widowed	58.1	40.3	22.6	17.7	11.3			1.178[0.562-2.473]	0.664
Histological grade						94.813	< 0.001		
Grade I	86.2	76.8	76.8	74.1	74.1			Reference	
Grade II	70.4	67.7	60.0	52.3	47.7			0.166[0.070-0.394]	< 0.001

Grade III	58.6	32.3	24.2	16.2	10.6			0.395[0.273-0.571]	< 0.001
Grade IV	66.2	30.1	16.9	11.7	5.2			0.791[0.644-0.971]	< 0.001
T Stage						128.335	< 0.001		
T1	85.9	67.4	58.7	53.3	43.5			Reference	
T2	80.7	54.6	40.3	29.4	16.8			0.954[0.627-1.452]	0.826
T3	67.8	34.7	21.5	14.8	10.3			0.880[0.652-1.188]	0.404
T4	47.1	11.2	6.5	4.7	3.5			0.941[0.747-1.185]	0.603
N Stage						157.176	< 0.001		
N0	79.3	51.9	38.5	30.0	20.7			Reference	
N1	52.0	17.3	9.8	5.8	5.8			0.715[0.552-0.926]	0.011
N2	44.7	7.8	2.9	2.9	2.9			0.931[0.710-1.222]	0.608
M Stage						201.905	< 0.001		
M0	90.0	61.3	53.5	37.7	27.4			Reference	
M1	49.0	16.8	9.4	5.5	3.4			1.223[0.945-1.582]	0.126
Stage Group						443.260	< 0.001		
I	98.9	94.9	89.8	81.4	62.7			Reference	
II	96.4	86.5	75.3	61.8	38.2			0.100[0.056-0.177]	< 0.001
III	91.5	57.6	36.5	24.8	19.0			0.130[0.086-0.197]	< 0.001
IV	45.7	9.1	2.5	2.5	2.5			0.285[0.201-0.403]	< 0.001
Laterality						0.829	0.661		
Left	67.1	38.0	26.4	20.2	13.2			Reference	
Right	67.5	35.3	25.9	19.9	15.5			1.458[0.506-4.204]	0.485
Bi	75.0	25.0	0	0	0			1.452[0.506-4.167]	0.488
Surgery						195.638	< 0.001		
Yes	83.1	50.6	37.6	29.0	20.9			Reference	
No/Unknown	73.2	11.1	4.5	3.3	3.3			0.390[0.318-0.479]	< 0.001
Radiation						32.991	< 0.001		
Yes	59.4	23.3	10.5	3.8	3.0			Reference	
No/Unknown	69.2	39.9	29.7	23.8	16.8			0.932[0.750-1.157]	0.522
Chemotherapy						6.313	0.012		
Yes	76.0	31.0	19.5	12.0	6.5			Reference	
No/Unknown	63.8	39.0	28.7	23.2	17.3			0.636[0.519-0.780]	< 0.001

Tumor Size						42.941	< 0.001		
0-40 mm	82.4	61.8	54.4	45.6	33.8			Reference	
40.1-70 mm	77.2	42.1	32.4	26.2	20.0			0.489[0.320-0.747]	0.001
70.1-100 mm	66.2	38.0	23.5	16.4	8.9			0.955[0.739-1.236]	0.726
>100 mm	59.0	26.3	17.3	12.8	10.2			0.852[0.688-1.056]	0.144

Kaplan–Meier survival curves indicate that tumor size, histological grade, stage group, radiation, and surgery all had a substantial impact on a patient's overall survival (OS). Age, race, marital status, sex, and laterality, however, were not important characteristics that affected OS (Figure 3 and Table 2).

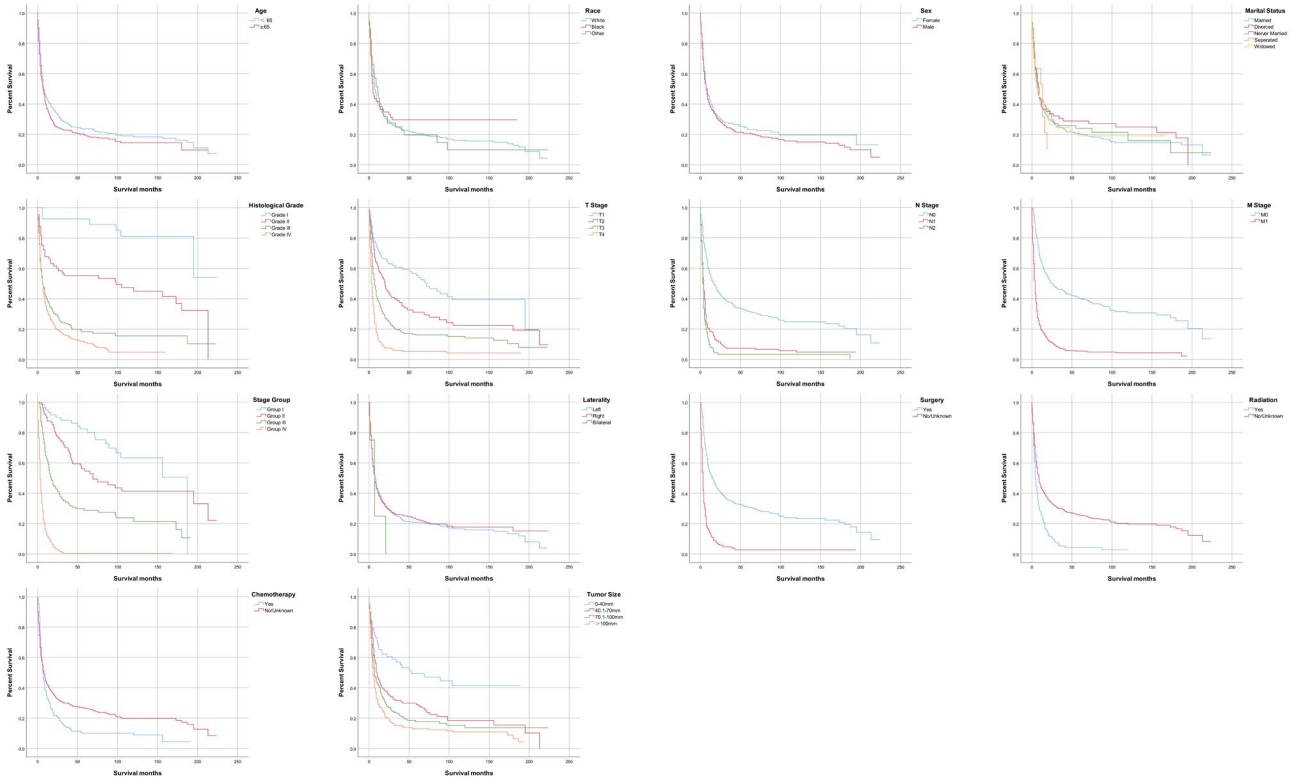


Figure 3: Kaplan–Meier survival curves of characteristics for patients in this study

3.3 Model performance

The model performance of the chosen six ML algorithms is summarized in Table 3. All six models have achieved high prediction accuracy, and the overall performance of the test set on 90 days was slightly lower than that of the other endpoints. XGB achieved the best accuracy on the 90-day, 1-, 2-, 3-, and 5-year OS test sets.

Table 3: Prediction accuracy of the chosen 6 ML algorithms

Endpoint	Dataset	Accuracy					
		LR	XGB	GBDT	RF	Ada	LGBM
90-day	Train	0.83	0.88	0.92	0.80	0.86	0.83

	Test	0.78	0.80	0.80	0.74	0.78	0.76
1-year	Train	0.84	0.87	0.89	0.86	0.86	0.92
	Test	0.81	0.83	0.82	0.82	0.78	0.81
2-year	Train	0.86	0.86	0.86	0.88	0.87	0.86
	Test	0.86	0.86	0.86	0.83	0.81	0.85
3-year	Train	0.85	0.84	0.93	0.96	0.86	0.90
	Test	0.86	0.86	0.83	0.84	0.85	0.86
5-year	Train	0.86	0.87	0.90	0.91	0.88	0.88
	Test	0.86	0.87	0.86	0.87	0.83	0.85

This study uses a hyperparameter search method to find the optimal parameters for each machine learning model to ensure the best performance of the model. ROC curves and calibration plots are established to evaluate the final performance of the model. All six models have achieved good performance in the training and test set (AUC>0.88, 0.81 respectively) (Figure 4). In the training set, the XGB, and LGBM models had the best performance of 90-day and 1-year OS prediction respectively (AUC=0.93902, 0.97460) (Figure 4 A, B), the RF model outperformed all other models of 2-, 3- and 5-year OS prediction (AUC=0.94650, 0.99502 and 0.96006 respectively) (Figure 4 C-E). In the test set, the XGB model had the highest performance of 90-day and 1-year OS prediction (AUC=0.87683, 0.88038 respectively) (Figure 4 F, G), the RF model was the top performer in the remaining predictions (AUC=0.88897, 0.89372 and 0.85895 respectively) (Figure 4 H-J).

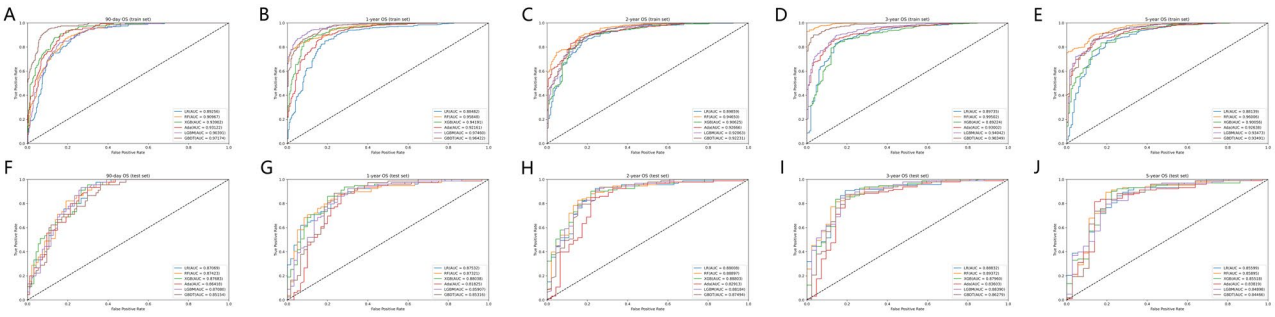


Figure 4: ROC curves of the six models. (A–E) training sets (F–J) test sets

The calibration plot is a scatter plot of the actual and predicted incidence of the event, with the 45-degree straight line in the plot indicating the optimal case, where all predicted values are equal to the true value. The better the calibration curve of the model fits the ideal curve, the better the model performs. If the predicted value is higher than the true value, the model calibration curve appears on the graph as higher than the ideal curve, indicating an overestimation of risk, and vice versa, indicating an underestimation of risk. In this study, all models have calibration curves around the 45-degree line on the training set, and XGB has the best consistency in the test set, with the other models tending to overestimate or underestimate risk across the entire range (Figure 5).

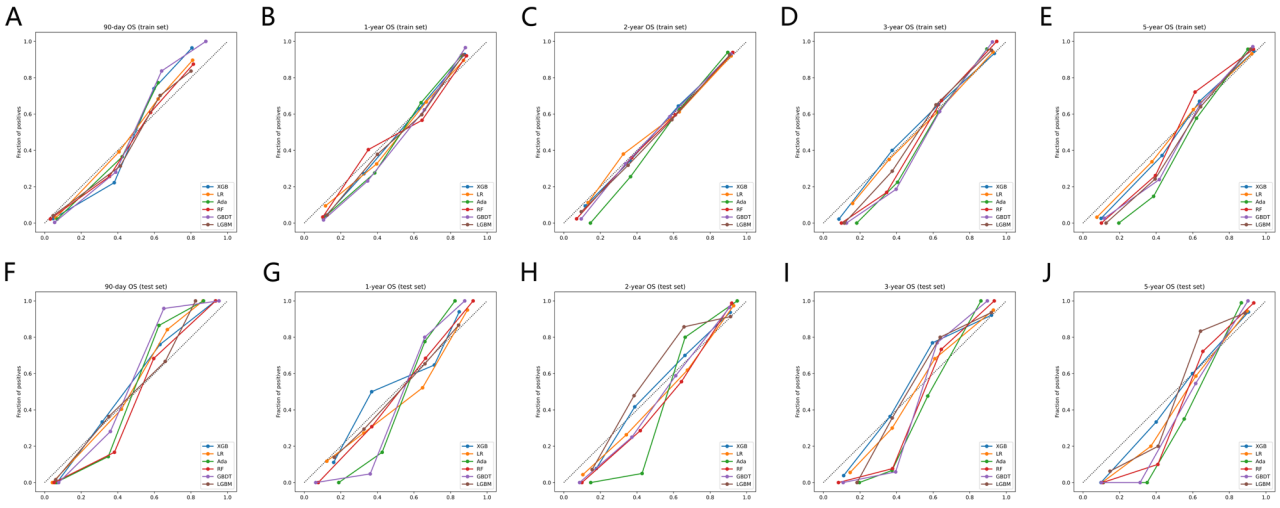


Figure 5: Calibration plots for ML models. (A–E) training sets (A-E), test sets (F-J)

In this study, the DCA of the six methods was subsequently built (Figure 6). The decision curve's y-axis shows the net benefit, a decision-analytic metric for determining if a given therapeutic option generates more benefits than damage. Each x-axis point corresponds to a threshold probability that distinguishes between patients who are dead and those who are still alive. All models, except Ada, achieved net therapeutic benefit, according to this research.

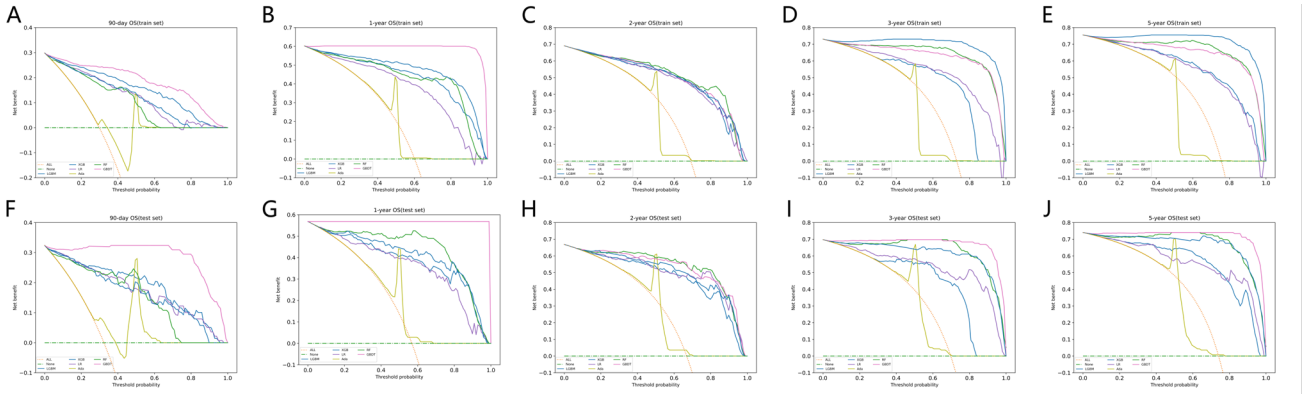


Figure 6: DCA curve analysis of ML models. All: a projection that all patients will die; None: a projection that all patients will be alive; (A–E) training sets; (F–J) test sets.

4. Discussion

sRCC is an aggressive, invasive tumor with a dismal prognosis and few effective treatment choices[3]. Accurate prediction of survival is difficult for this malignancy but has important implications for treatment planning and patient management. The machine learning algorithm provides a more efficient and reliable choice for the survival state prediction of patients with sRCC, and the trained and verified machine learning model can predict quickly and accurately. Therefore, it is of great clinical significance to establish a survival prediction model for sRCC patients based on machine learning.

Based on the analysis of clinicopathological characteristics and demographic information of sRCC patients, this study screened out important variables that could affect the OS of patients and built six ML prediction models based on these variables. With a median survival period of 7 months in this study, 537 (77.6%) patients died during follow-up, 338 (48.8%) of whom had the distant metastatic illness. Ten of the 14 variables selected, including age, TNM stage, staging group, tumor size, chemotherapy, histologic grade, surgery, and radiotherapy, were considered in the Kaplan-Meier survival analysis as variables that could affect patient survival. Surgery had a minimal impact on patient prognosis improvement. During the observation period, 29.3% of surgery patients died of the illness, with a median survival of 12 months. Because they are frequently in an advanced stage or have metastatic cancer, patients who get chemotherapy and radiation typically have a poor prognosis. For patients with advanced metastases, doctors may administer systemic treatment or radiotherapy to the metastatic areas. However, sRCC typically responds poorly to these kinds of therapies[3]. According to Stafford et al., compared to patients of all other races, black men and patients with RCC had a significantly greater incidence rate and shorter survival rate[14]. However, the sRCC study did not define race or sexual orientation as risk variables. This could be a result of the disease's exceedingly dismal prognosis, which sees a lot of people pass away quickly.

The most significant features determined by the ML models and Cox regression analysis were highly similar. To build six ML models for the 90-day, 1-, 2-, 3-, and 5-year OS for patients with sRCC, this study chose the 10 most crucial factors. Model performance was assessed using accuracy, ROC, calibration plots, and DCA. XGB model achieved the highest predictive accuracy over five different periods in test sets (90-day: 0.80, 1-year: 0.83, 2-year: 0.86, 3-year: 0.86, 5-year: 0.87). The accuracy of the model prediction in 90-day OS was slightly lower than in other periods, this may be because the overall survival rates of sRCC patients plummet and are less predictable within 90 days.

The consistency of the model predictions was further evaluated using the calibration plot. Six models achieved similar results on different endpoints of the training and test sets. XGB model showed the best agreement, while the other models tended to underestimate the risk of death across the entire prediction range, which may result in some patients missing out on due follow-up and treatment.

The DCA curve is used to help identify high-risk patients for intervention and low-risk patients to avoid intervention (avoiding overmedication), and it can assess the degree of patient benefit to determine whether a model is worth using[15]. In the model trained in this study, XGB, GBDT, RF, LGBM, and LR all achieved good clinical net benefits, while the Ada model performed poorly, which was related to its inability to adjust parameters. The results of the DCA show that our five models are good at balancing clinical decisions for maximum benefit.

Although the predictive models utilized in this investigation performed well, they did have certain drawbacks. First, the patient dataset was derived from SEER, a database that only collects information on the North American population. The prediction model built from this may not apply to other regions. Second, due to the limitation of geographical information of patients, geographically-related factors such as socioeconomic status, built environment, and air quality were not taken into account. Finally, the SEER database puts patients who did not undergo radiotherapy and chemotherapy and whose records are unknown in one directory, which can lead to some bias.

5. Conclusion

This paper analyzes the impact of 14 clinical and demographic characteristics of sRCC patients on their prognosis. The ML algorithm and the univariate Cox proportional risk model were used to filter the most important characteristics, including age, TNM stage, stage group, surgery, chemotherapy, radiotherapy, tumor size, and histological grade. these variables were then used as input data to build six ML models to predict 90-day, 1-, 2-, 3-, and 5-year survival rates in sRCC patients. All six models obtained good prediction accuracy, with XGB performing the best on the test set (AUCs of 0.87683, 0.88038, 0.88803, 0.87960 and 0.85518 on 90-day, 1-, 2-, 3- and 5-year OS, respectively) and the most consistent predictive ability. DCA showed that, except for the Ada model, all other models achieved good net clinical gains, which indicates that the predictive models developed in this study can help clinicians make treatment and follow-up decisions.

6. Declarations

6.1 Ethics approval and consent to participate

Not applicable.

6.2 Consent for publication

Not applicable.

6.3 Author contribution(s):

Rui Zhang: Conceptualization; Data curation; Methodology; Formal analysis; Investigation; Visualization; Writing – original draft; Writing – review & editing.
Xuefei Qin: Conceptualization; Methodology; Formal analysis; Investigation; Software; Visualization; Writing – original draft; Writing – review & editing.
Xuelin Gao: Methodology; Project administration; Validation; Writing – review & editing.
Yu Zheng: Methodology; Project administration; Validation; Writing – review & editing.
Guangdong Hou: Validation; Formal analysis
Yueyue Zhang: Data curation
Yuchen Tian: Formal analysis
Yuliang Wang: Data curation
Shuaijun Ma: Funding acquisition; Project administration; Supervision; Writing – review & editing.
Fuli Wang: Funding acquisition; Project administration; Supervision; Writing – review & editing.

6.4 Funding:

This study was supported by the Medical Research Project of Innovation Capability Enhancement Plan of Xi'an (21YXYJ0107) and the Boost Program Research Project of Xijing Hospital (XJZT21L09).

6.5 Competing interests:

The authors declare that there is no conflict of interest.

6.6 Availability of data and material:

The datasets and material used in this study is available online.

Reference

1. Siegel RL, Miller KD, Wagle NS, et al. Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians* 2023; 73: 17–48.

2. Delahunt B, Cheville JC, Martignoni G, et al. The International Society of Urological Pathology (ISUP) Grading System for Renal Cell Carcinoma and Other Prognostic Parameters. *The American Journal of Surgical Pathology* 2013; 37: 1490.

3. Mouallem NE, Smith SC, Paul AK. Sarcomatoid renal cell carcinoma: Biology and treatment advances. *Urologic Oncology: Seminars and Original Investigations* 2018; 36: 265–271.

4. Cheville JC, Lohse CM, Zincke H, et al. Sarcomatoid Renal Cell Carcinoma: An Examination of Underlying Histologic Subtype and an Analysis of Associations With Patient Outcome. *The American Journal of Surgical Pathology* 2004; 28: 435.

5. Stone PC, Lund S. Predicting prognosis in patients with advanced cancer. *Annals of Oncology* 2007; 18: 971–976.

6. Cheng M, Duzgol C, Kim T-H, et al. Sarcomatoid renal cell carcinoma: MRI features and their association with survival. *Cancer Imaging* 2023; 23: 16.

7. Zhang BY, Thompson RH, Lohse CM, et al. A novel prognostic model for patients with sarcomatoid renal cell carcinoma. *BJU International* 2015; 115: 405–411.

8. Yan Y, Liu L, Zhou J, et al. Clinicopathologic characteristics and prognostic factors of sarcomatoid renal cell carcinoma. *J Cancer Res Clin Oncol* 2015; 141: 345–352.

9. Gu L, Ma X, Li H, et al. Prognostic value of preoperative inflammatory response biomarkers in patients with sarcomatoid renal cell carcinoma and the establishment of a nomogram. *Sci Rep* 2016; 6: 23846.

10. Hou G, Li X, Zheng Y, et al. Construction and validation of a novel prognostic nomogram for patients with sarcomatoid renal cell carcinoma: a SEER-based study. *Int J Clin Oncol* 2020; 25: 1356–1363.

11. May M. Eight ways machine learning is assisting medicine. *Nature Medicine* 2021; 27: 2–3.

12. Goecks J, Jalili V, Heiser LM, et al. How Machine Learning Will Transform Biomedicine. *Cell* 2020; 181: 92–101.

13. Jin S, Yang X, Zhong Q, et al. A Predictive Model for the 10-year Overall Survival Status of Patients With Distant Metastases From Differentiated Thyroid Cancer Using XGBoost Algorithm-A Population-Based Analysis. *Frontiers in Genetics*; 13, https://www.frontiersin.org/articles/10.3389/fgene.2022.896805 (2022, accessed 12 May 2023).

14. Stafford HS, Saltzstein SL, Shimasaki S, et al. Racial/Ethnic and Gender Disparities in Renal Cell Carcinoma Incidence and Survival. *The Journal of Urology* 2008; 179: 1704–1708.

15. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006; 26: 565–574.