Original Article

Machine learning algorithms being an auxiliary tool to predict the overall survival of patients with renal cell carcinoma using the SEER database

Weixing Jiang[1#], Zhenghao Chen[1,2#], Cancan Chen[3], Lei Wang[1], Tiandong Han[1], Li Wen[4]

[1]Department of Urology, Beijing Friendship Hospital, Capital Medical University, Beijing, China; [2]Department of Urology, The Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China; [3]Digital Health China Technologies Co., LTD., Beijing, China; [4]Department of Urology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

Contributions: (I) Conception and design: L Wang, T Han, L Wen; (II) Administrative support: W Jiang, L Wen; (III) Provision of study materials or patients: T Han, L Wang; (IV) Collection and assembly of data: C Chen, Z Chen; (V) Data analysis and interpretation: W Jiang, Z Chen; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Lei Wang, MD; Tiandong Han, MD. Department of Urology, Beijing Friendship Hospital, Capital Medical University, No. 59 Yongan Road, Xicheng District, Beijing 100050, China. Email: sclare@163.com; cruiser412@163.com; Li Wen, MD. Department of Urology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Panjiayuan Nanli 17#, Chaoyang District, Beijing 100021, China. Email: wenli_cicams@163.com.

Background: The clinical prognosis assessment of renal cell carcinoma (RCC) still relies on nuclear grading and nuclear score by naked eye with microscope, which has defects long time, low efficiency, and uneven evaluation level criteria. There are few machine learning (ML) studies investigating the prognosis in the RCC literature which could also quantify the risk of postoperative recurrence of RCC patients and guide cancer patients to conduct individualized postoperative clinical management. This study evaluated the suitability of ML algorithms for survival prediction in patients with RCC.

Methods: A total of 192,912 RCC patients from the Surveillance, Epidemiology, and End Results (SEER) were obtained from 2004 to 2015. Six ML algorithms including support vector machine (SVM), Bayesian method, decision tree, random forest, neural network, and Extreme Gradient Boosting (XGBoost) were applied to predict overall survival (OS) of RCC.

Results: Patients from the SEER with a median age of 62 years and the pathological types were clear cell RCC (47.6%), papillary RCC (9.5%), chromophobe RCC (4.0%) and others (4.1%) were collected. In the deleting patients with missing data, the highest accurate model was XGBoost [area under the curve (AUC) 67.0%]. In the deleting patients with missing data and survival time <5 years, the accuracy of random forest, neural network and XGBoost were high, with AUC of 80.8%, 81.5% and 81.8%, respectively. In the only deleting the missing tumor diameter and filling the missing dataset with missForest, the highest accurate model was random forest (AUC: 71.9%). In this study, the overall accuracy of the SVM model was not high, apart from in the population of patients with deleting the missing tumor diameter and survival time <5 years, and filling the missing data with missForest. Random forest, neural network and XGBoost had high accuracy, with AUC of 84.1%, 84.7% and 84.8%, respectively.

Conclusions: ML algorithms could be used to predict the prognosis of RCC. It could quantify the recurrence possibility of patients and help more individualized postoperative clinical management. Given the limitations and complexity of datasets, ML may be used as an auxiliary tool to analyze and process larger datasets and complex data.

Keywords: Renal cell carcinoma (RCC); machine learning (ML); algorithm; prediction; prognosis

Submitted Aug 31, 2023. Accepted for publication Nov 22, 2023. Published online Jan 23, 2024.
doi: 10.21037/tau-23-319
View this article at: https://dx.doi.org/10.21037/tau-23-319

Introduction

In 2019, 73,820 new renal cancer cases were diagnosed, accounting for 4.2% of all tumors and 14,770 renal cancer-associated deaths in the United States (1). Renal cell carcinoma (RCC) originates from the epithelial cells of renal tubules and has a relatively favorable prognosis after surgery for localized diseases. However, approximately 20% of RCC patients are diagnosed with metastasis, which is incurable and associated with poor survival outcomes. Despite modest advances in treatment, the prognosis of patients with different risk factors is variable (2). Therefore, RCC survival outcomes differ significantly among individuals. Although prognostic factors and models have been developed for RCC, the limited data commonly restrict the accuracy of individualized accurate prediction (3-5). In a review conducted by Usher-Smith *et al.* (6), they provide the most comprehensive summary to date on the role of models in different populations. The authors argue that there is no clear single 'best' model for any of the populations being studied, namely recurrence free survival (RFS), cancer specific survival (CSS), and OS. Regarding RFS, the Sorbellini, Karakiewicz, Leibovich, and Kattan models have shown better performance, while University of California, Los Angeles Integrated Staging System (UISS) also demonstrates comparable performance in the European/American population. In terms of CSS, the Zisman, SSIGN, Karakiewicz, Leibovich, and Sorbellini models have performed better. For OS, the Leibovich, Karakiewicz, Sorbellini, and SSIGN models have demonstrated better predictive ability in terms of patient prognosis. One of the most interesting and challenging aspects is accurate survival time prediction.

Highlight box

Key findings

- Machine learning (ML) may be used as an auxiliary tool in renal cell carcinoma (RCC).

What is known and what is new?

- RCC prognosis differs significantly among individuals. Although prognostic factors and models for RCC. One of the most interesting and challenging aspects is accurate survival time prediction.
- In this study, we collected RCC patient data from the Surveillance, Epidemiology, and End Results database and aimed to investigate the utility of ML algorithms in predicting the 5-year overall survival outcomes of patients with RCC. And we proved all six algorithm models could predict the 5-year survival rate of RCC patients, but the prediction accuracy was different.

What is the implication, and what should change now?

- ML was one of the newly developed ways to quantify the risk of postoperative recurrence of RCC patients and guide cancer patients to conduct individualized postoperative clinical management.

Machine learning (ML), a major subbranch of artificial intelligence, is a promising statistical method that allows the analysis of heterogeneous and large-scale data. Within medicine, ML algorithms are built using data from large patient databases, with the intention to find patterns and make predictions (7,8). As a matter of fact, ML has been used for cancer prevention, diagnosis, and prognosis prediction (9-11). The main areas of ML research in RCC literature are the differentiation between benign and malignant small renal masses, Fuhrman nuclear grade prediction, and gene expression-based molecular signatures (12). However, few studies of ML techniques in RCC prognostication have focused on overall survival (OS) outcomes.

The Surveillance, Epidemiology, and End Results (SEER) database represents approximately 28% of the US population and consists of demographic, tumor characteristic, treatment, and outcome data. The SEER database could be a valuable source of research on large sample and multi-institutional patient pool. In the past, researchers have used the SEER database to conduct a large number of studies that are beneficial to the understanding of current diseases (13). In this study, we collected RCC patient data from the SEER database and aimed to investigate the utility of ML algorithms in predicting the 5-year OS outcomes of patients with RCC. We present this article in accordance with the TRIPOD reporting checklist (available at https://tau.amegroups.com/article/view/10.21037/tau-23-319/rc).

Methods

Data collection

The data used in this study were collected from the online, publicly available SEER database using SEER*Stat software (Version 8.3.6). A data agreement form was signed, and was submitted to the SEER administration. A data agreement form was submitted to the SEER administration. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). This study did not involve personal identifiers. Thus, the study was deemed exempt by the Domain-Specific Review Board. We extracted data from patients with RCC (International Classification of Disease for Oncology C64) in the SEER database from 2004 to 2015 (n=192,912). The variables consisted of race, sex, age at diagnosis, marital status, tumor location, primary tumor size, histological type, tumor grade, tumor-node-metastasis (TNM) stage information, surgical treatment, OS duration in months, and vital status. "Others" in the race category included American Indian/Alaska Native or Asian/Pacific Islander. Marriage was analyzed using marital status. "Single" included never married, separated/divorced, and widowed, according to SEER marital status categories.

Data preparation

To correct the loss of important feature information, such as tumor size, TNM stage and survival time, which may affect the judgment of prognosis information, we tested four groups to investigate the feasibility of six ML algorithms including support vector machine (SVM), Bayes point machine, decision tree, random forest, neural network and Extreme Gradient Boosting (XGBoost), which were applied to construct models based on prior work (14-19). Bayes is a flexible probability graph model that captures the dependencies between selected variables, with the advantage of using graphical representations. Other researchers have compared traditional logistic regression models with Bayesian networks for risk prediction in breast cancer recurrence, and found similar accuracy (20). In this study, similar to the SVM model, the accuracy of Bayesian network method was not high in general, which may also depend on the integrity of the data. And a decision tree is a model that follows a tree structure classification scheme that can process nonnumeric data and is easy to understand. Other scholars have successfully proven its usability (21,22). Random forest is an extension of a decision tree with multiple decision trees that use random variables. Zhao *et al.* (23) used a public database to investigate random forest for analyzing the differences in gene expression in the microenvironment of lung cancer, providing a new strategy for exploring prognostic biomarkers and immunotherapy.

First, we converted the category description of patient characteristics into a digital description scheme that the ML model can recognize. Then, three common data preprocessing methods in ML including standardization, normalization, and Min-Max Scaler were used to evaluate the feasibility of the algorithms. Finally, to investigate the influence of possible noise, such as missing data, completion and uncertain information on ML models, and to verify the effect of data completion methods on missing tumor information, all patients were divided into a training set and validation set, and data were analyzed in four different populations: (I) deleting patients with missing data (n=92,475); (II) deleting patients with missing data and survival time <5 years (n=55,334); (III) only deleting the missing tumor diameter and filling the missing dataset with missForest (n=120,202); (IV) deleting the missing tumor diameter and survival time <5 years, and filling the missing data with missForest (n=71,710).

Additionally, to obtain ML models with better generalization and stability, we divided the SEER dataset into 10 subsets and selected a 10-fold cross-validation method to debug the hyperparameters in the training process of all models.

Statistical analysis

Our outcome of interest was the 5-year OS outcome. We trained all ML models using Python, including Python version 3.6.8, Numpy version 1.20.1, Scikit-Learn version 0.19.2, Scipy version 1.7.3, and XGBoost version 1.6.2. The ML models were developed and evaluated using 10-fold cross-validation. The receiver operating characteristic (ROC) curve and the C index were also used to evaluate the accuracy of the model.

Results

Patient characteristics

A total of 192,912 patients from the SEER database were

Table 1 Patient characteristics at baseline

Characteristics	Values
Sex, n (%)	
Male	122,902 (63.7)
Female	70,010 (36.3)
Age (years), median [interquartile range]	62 [53–71]
Marital status, n (%)	
Married	159,183 (82.5)
Unmarried	25,366 (13.1)
Unknown	8,363 (4.3)
Race, n (%)	
White	160,280 (83.1)
Black	21,008 (10.9)
Others	10,630 (5.5)
Unknown	994 (0.5)
Tumor location, n (%)	
Left	93,975 (48.7)
Right	96,739 (50.1)
Bilateral	1,912 (1.0)
Uknown	286 (0.1)
Tumor size (cm), mean ± SD	3.4±3.2
Histological types, n (%)	
Clear cell	91,793 (47.6)
Papillary	18,337 (9.5)
Chromophobe	7,765 (4.0)
Others	7,837 (4.1)
Unknown	67,180 (34.8)
Tumor grade, n (%)	
G1/2	87,409 (45.3)
G3/4	45,170 (23.4)
Unknown	60,333 (31.3)
T stage, n (%)	
T1/2	93,149 (48.3)
T3/4	25,167 (13.0)
Unknown	74,596 (38.7)
N stage, n (%)	
N0	112,518 (58.3)
N1	6,055 (3.1)
Unknown	74,339 (38.5)
M stage, n (%)	
M0	107,455 (55.7)
M1	14,286 (7.4)
Unknown	71,171 (36.9)

SD, standard deviation.

collected, with a median age of 62 years [interquartile range (IQR), 53–71 years]. Among them, 122,902 (63.7%) were male and 70,010 (36.3%) were female. Most of the RCC patients were married, accounting for 82.5% (n=159,183). The population was mostly white (83.1%, n=160,280). The mean tumor diameter was 3.4±3.2 cm. At the initial diagnosis, 3.1% (n=6,055) and 7.4% (n=14,286) of patients had lymph node and organ metastasis respectively. Histologically, the common pathological types were clear cell RCC (47.6%, n=91,793), papillary RCC (9.5%, n=18,337), chromophobe RCC (4.0%, n=7,765) and others (4.1%, n=7,837). The baseline characteristics of RCC patients are summarized in *Table 1*.

Preprocessing result analysis

Min-Max Scaler was more beneficial for the training of SVM models than standardization and normalization preprocessing methods. Bayes, decision tree, random forest, and XGBoost were suitable for non-processing modes. The standard preprocessing method was appropriate for the neural network (*Figure 1*).

Validation analysis

The box-plot results of the area under the curve (AUC) with the 10-fold cross-validation showed that 6 ML algorithms were relatively stable and concentrated in this population (*Figure 2A*). The SVM (max AUC: 0.656), Bayes (max AUC: 0.636), decision tree (max AUC: 0.615) and random forest (max AUC: 0.643) models were less effective at predicting 5-year survival rates for RCC patients, while the models of neural network (max AUC: 0.666) and XGBoost (max AUC: 0.671) models performed relatively well (*Figure 2B*). Further calibration curves of the 5-year OS rates showed that the predicted and observed values of the neural network model (Brier score: 0.221) and XGBoost (Brier score: 0.220) were relatively consistent while the prediction and observations of the SVM (Brier score: 0.230), Bayes (Brier score: 0.316), decision tree (Brier score: 0.259) and random forest (Brier score: 0.232) models were relatively discrete (*Figure 2C*). Overall, all models revealed poor identification in this population.

Patients with deleting missing data and survival time <5 years

The box diagram results of the model test showed that

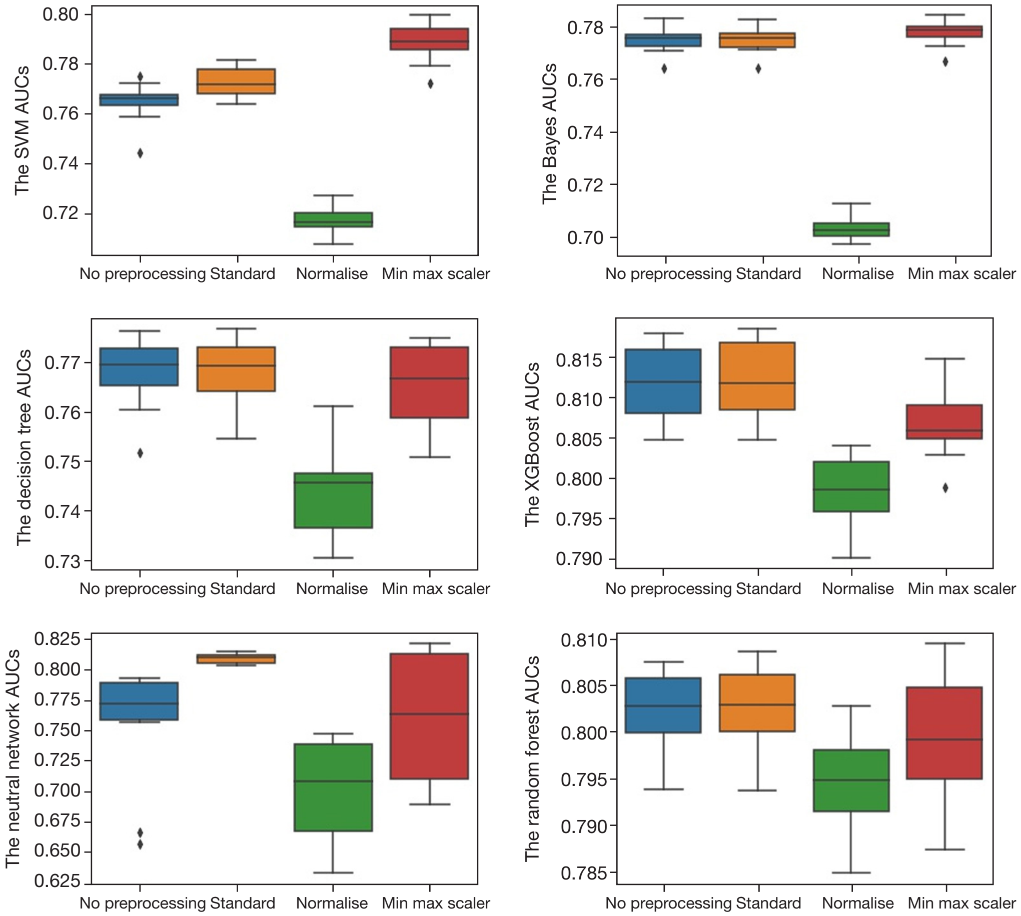

Figure 1 Preprocessing mode for different machine learning methods. SVM, support vector machine; AUC, area under the curve; XGBoost, Extreme Gradient Boosting.

random forest, neural network and XGBoost were more stable than SVM, Bayes and decision tree by the 10-fold cross-validation in this population (*Figure 3A*). The maximum AUC of the linear model was 0.797 (*Table 2*). The SVM (max AUC: 0.799), Bayes (max AUC: 0.783), decision tree (max AUC: 0.776), random forest (max AUC: 0.808), neural network (max AUC: 0.815) and XGBoost (max AUC: 0.818) models were all effective in predicting the 5-year survival rate of patients with RCC (*Figure 3B*). Further calibration curves also indicated that the SVM (Brier score: 0.162), Bayes (Brier score: 0.189), decision tree (Brier score: 0.170), random forest (Brier score: 0.148), neural network model (Brier score: 0.145) and XGBoost (Brier score: 0.144) models were relatively consistent (*Figure 3C*). In general, all models demonstrated good recognition results for this dataset.

Patients with only deleting the missing tumor diameter and filling the missing dataset with missForest

The model stability test of the box diagram showed that all models were relatively stable after 10-fold cross-validation in this population (*Figure 4A*). The SVM (max AUC: 0.675), Bayes (max AUC: 0.656) and decision tree (max AUC: 0.683) models were not effective for predicting the 5-year survival rate of RCC patients, while the random forest (max AUC: 0.719), neural network (max AUC: 0.698) and XGBoost (max AUC: 0.714) models performed relatively well (*Figure 4B*). Calibration curves also showed that SVM (Brier score: 0.223), Bayes (Brier score: 0.323) and decision tree (Brier score: 0.231) models were ineffective, and the predicted and observed values were relatively discrete, while random forest (Brier score: 0.208), neural

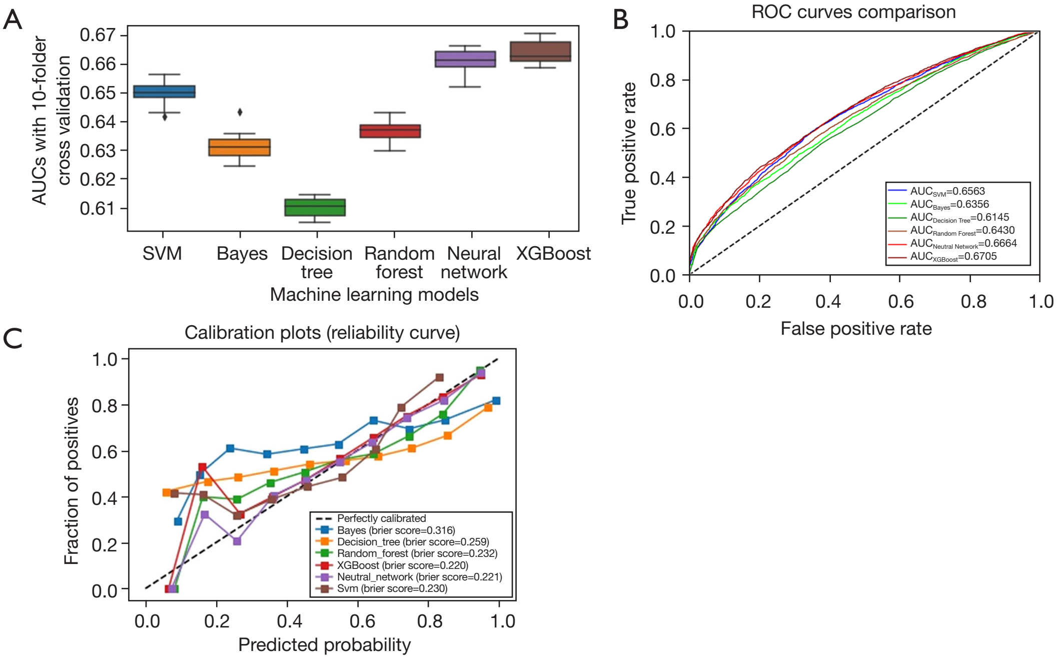

Figure 2 The characteristics of patients with deleting missing data. (A) The box-plot results of AUC with the 10-fold cross-validation of different machine learning algorithms. (B) ROC curve analysis of different models for predicting 5-year survival in patients with RCC. (C) Calibration curves of different models for predicting 5-year survival in patients with RCC. SVM, support vector machine; XGBoost, Extreme Gradient Boosting; AUC, area under the curve; ROC, receiver operating characteristic; RCC, renal cell carcinoma.

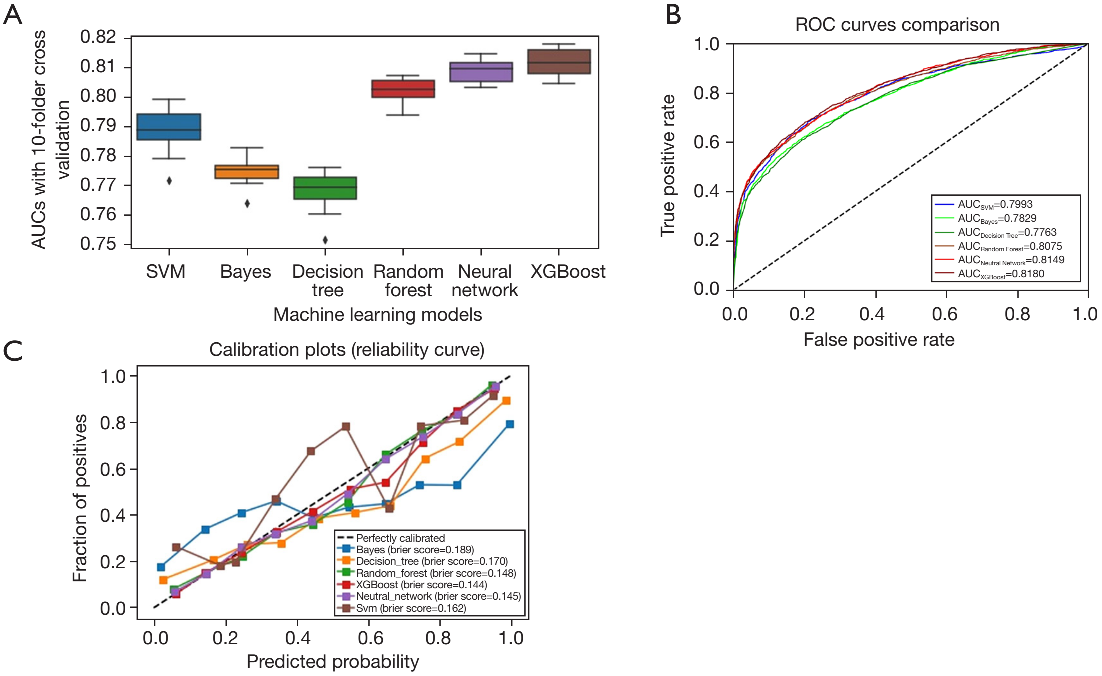

Figure 3 The characteristics of patients with deleting missing data and survival time <5 years. (A) The box-plot results of AUC with the 10-fold cross-validation of different machine learning algorithms. (B) ROC curve analysis of different models for predicting 5-year survival in patients with RCC. (C) Calibration curves of different models for predicting 5-year survival in patients with RCC. SVM, support vector machine; XGBoost, Extreme Gradient Boosting; AUC, area under the curve; ROC, receiver operating characteristic; RCC, renal cell carcinoma.

Table 2 An ensemble model based on ML models in patient population with those with missing data and a survival time <5 years removed

Validation	Bayes AUC	Dec. Tree AUC	Linear Reg. AUC	Random forest AUC	SVM AUC	XGBoost AUC	Ensemble AUC
Val 1	0.77939	0.76762	0.79458	0.80445	0.79654	0.81797	0.81109
Val 2	0.77853	0.77162	0.78755	0.80306	0.79022	0.81260	0.80974
Val 3	0.78024	0.773	0.79323	0.80627	0.77195	0.81577	0.81314
Val 4	0.77515	0.76795	0.78836	0.80101	0.78764	0.8068	0.8063
Val 5	0.77903	0.76459	0.78848	0.79628	0.78991	0.81106	0.80748
Val 6	0.7726	0.75161	0.7908	0.79962	0.78747	0.80726	0.80334
Val 7	0.76675	0.76023	0.7792	0.79381	0.77922	0.80473	0.80197
Val 8	0.77833	0.77088	0.79284	0.80252	0.78507	0.81031	0.80922
Val 9	0.7841	0.77389	0.79149	0.80750	0.79546	0.81793	0.81639
Val 10	0.78072	0.7763	0.79737	0.80674	0.79934	0.81602	0.81407

ML, machine learning; AUC, area under the curve; Dec. Tree, decision tree; Linear Reg., Linear Regression; SVM, support vector machine; XGBoost, Extreme Gradient Boosting.

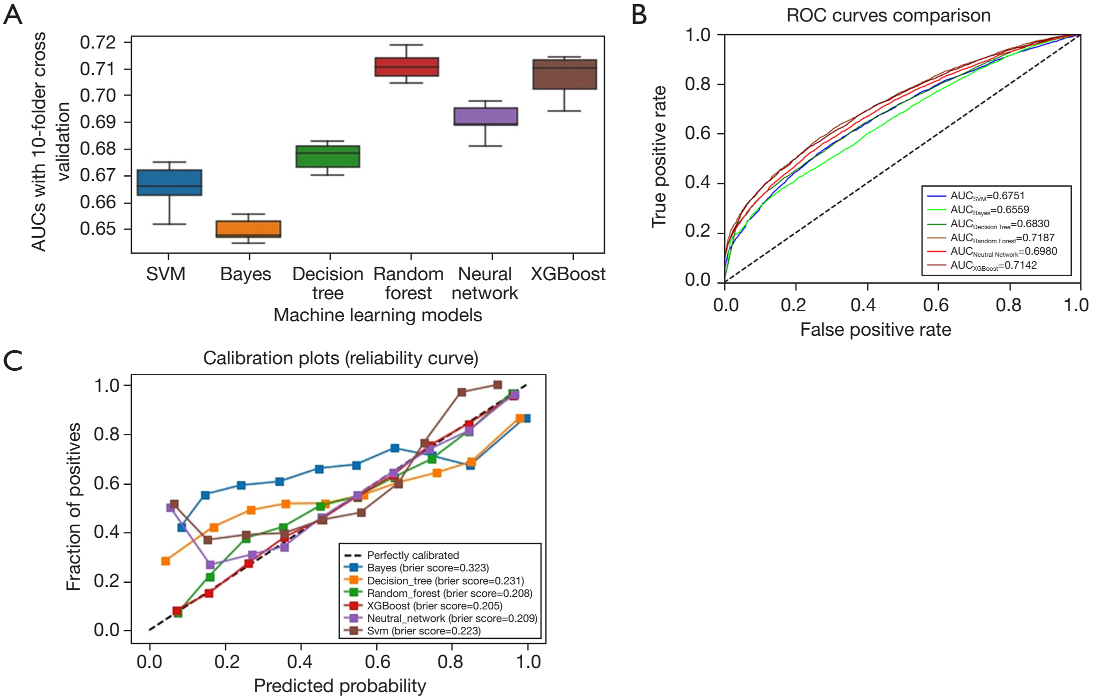

Figure 4 The characteristics of patients with only deleting the missing tumor diameter and filling the missing dataset with missForest. (A) The box-plot results of AUC with the 10-fold cross-validation of different machine learning algorithms. (B) ROC curve analysis of different models for predicting 5-year survival in patients with RCC. (C) Calibration curves of different models for predicting 5-year survival in patients with RCC. SVM, support vector machine; XGBoost, Extreme Gradient Boosting; AUC, area under the curve; ROC, receiver operating characteristic; RCC, renal cell carcinoma.

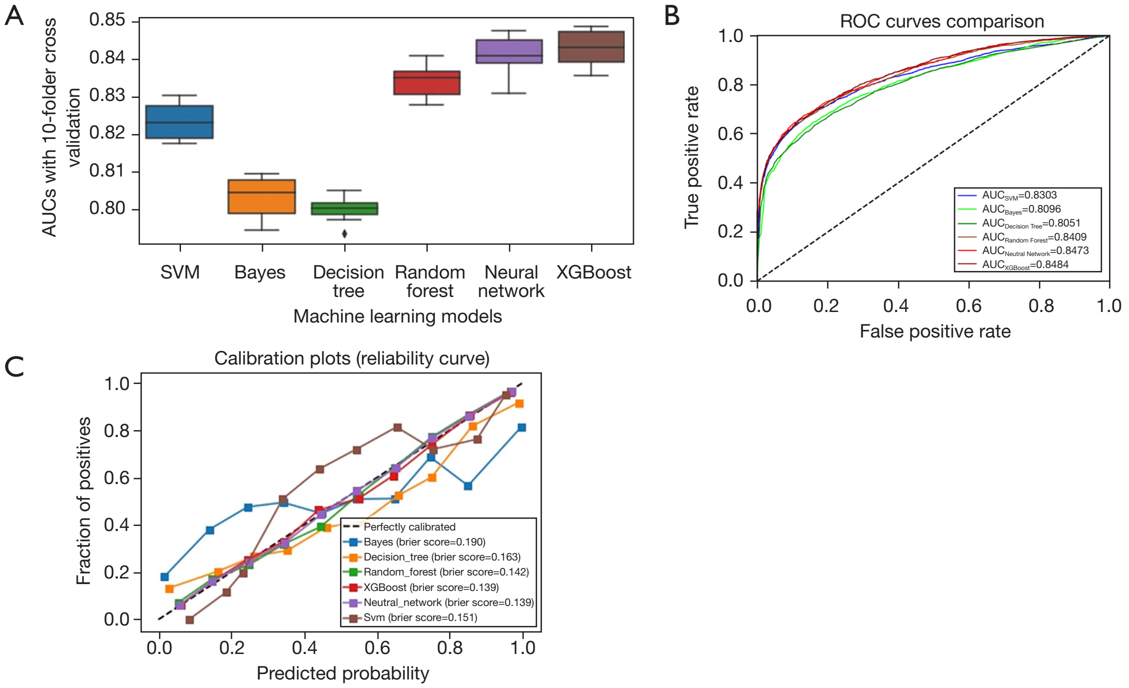

Figure 5 The characteristics of patients with deleting the missing tumor diameter and survival time <5 years, and filling the missing data with missForest. (A) The box-plot results of AUC with the 10-fold cross-validation of different machine learning algorithms. (B) ROC curve analysis of different models for predicting 5-year survival in patients with RCC. (C) Calibration curves of different models for predicting 5-year survival in patients with RCC. SVM, support vector machine; XGBoost, Extreme Gradient Boosting; AUC, area under the curve; ROC, receiver operating characteristic; RCC, renal cell carcinoma.

network (Brier score: 0.209) and XGBoost (Brier score: 0.205) performed relatively well (*Figure 4C*). Overall, all models demonstrated relatively poor recognition in this dataset.

Patients with deleting the missing tumor diameter and survival time <5 years, and filling the missing data with missForest

The box diagram showed that random forest, neural network and XGBoost were more stable than SVM, Bayes and decision tree in this population (*Figure 5A*). Further analysis showed that each model had a good recognition. The SVM (max AUC: 0.830), Bayes (max AUC: 0.810), decision tree (max AUC: 0.805), random forest (max AUC: 0.841), neural network (max AUC: 0.847) and XGBoost (max AUC: 0.848) models were all effective for predicting the 5-year survival rate of RCC patients (*Figure 5B*). Further calibration curves indicated that the SVM (Brier score: 0.151), Bayes (Brier score: 0.190), decision tree (Brier score: 0.163), random forest (Brier score: 0.142), neural network model (Brier score: 0.139) and XGBoost (Brier score: 0.139) models were almost consistent (*Figure 5C*).

Based on all ML models, we obtained the importance degree of all features of SEER, as shown in Figure S1. The most important feature affecting prognosis was M stage, followed by N stage, tumor grade, tumor size, T stage, surgery, age, pathology type, marital status, side, sex, and race.

Discussion

Cancer patients are often eager to know as much as possible about their disease and prognosis, and good predictive tools can help clinicians provide that information for patients. Several studies have sought to identify prognostic factors in kidney cancer, one of the most common urinary tumors (24-27). However, few researchers have attempted to provide clinicians with web-based tools. To the best of our knowledge, this study is the first research to use ML algorithms to predict the prognosis of RCC patients. Based on the primary investigation, the results suggested that ML algorithms could be an effective tool for predicting RCC patient. Additionally, given the limitations and complexity of some public datasets, the ML model as an auxiliary tool may be advantageous for processing large datasets.

Accurately predicting the prognosis of patients with RCC is one of the most interesting and challenging tasks for urologists. However, renal tumors are heterogeneous and associated with many factors, and traditional linear statistical models are not reliably accurate for predicting prognosis. Regarding nonlinear statistical models, various ML techniques (including SVM, Bayes, decision tree, random forest, neural networks, and XGBoost) have been widely explored to develop predictive models that can detect and identify patterns and nonlinear relationships between multidimensional factors (28). In addition, cancer research has entered the era of big data (29). Data from a single center often do not fully reflect the real situation. Research has been further deepened with the development and use of public databases. Multiple studies have shown that ML algorithms can be used to mine public databases and exhibit high accuracy (30-33). However, there is still a lack of research on ML algorithms in RCC.

SVM is a supervised ML algorithm mainly used to classify cases and has been proved successful in multiple tumors (34,35). In the dataset analyses in this study, the overall accuracy of the SVM model was not high, apart from in the population of patients with deleting the missing tumor diameter and survival time <5 years, and filling the missing data with missForest. Data integrity is likely beneficial for establishing the SVM model.

We also tried decision tree and random forest to predict the 5-year OS outcomes of RCC patients, and the accuracy was moderate, with the highest AUC at 80.5% and 84.1%, respectively.

A neural network is a group of interconnected nodes similar to the vast network of neurons in the human brain. Large quantities of information can be processed, and the network can output variables through a hidden layer, neural networks are developing rapidly and are widely used in many fields for disease prevention, diagnosis, treatment, and prognosis evaluation and for basic research (36,37). XGBoost is an optimized lift library with efficient and flexible features. XGBoost is also widely used, and several studies have confirmed that XGBoost has high accuracy and applicability (38,39). In this study, neural networks and XGBoost had better predictive power than other ML models in each dataset, with the highest AUC reaching 84.7% and 84.8%, respectively. These approaches may be potential ML algorithms to be further studied.

There were also some limitations in this study. First, this study is a retrospective study of a population from the SEER database, which has some defects. The algorithms developed have not been externally verified, and rigorous external verification is necessary to determine whether the model is suitable for a specific population. In addition, our data source is the SEER database, which is limited by the common limitations of all administrative dataset studies, including lack of data such as specific laboratory indicators, adjuvant therapy, and tumor recurrence.

Conclusions

All six algorithm models could predict the 5-year survival rate of RCC patients, but the prediction accuracy was different. ML algorithms may be used as an auxiliary tool to analyze and process large datasets in the future.

Acknowledgments

Funding: None.

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://tau.amegroups.com/article/view/10.21037/tau-23-319/rc

Peer Review File: Available at https://tau.amegroups.com/article/view/10.21037/tau-23-319/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://tau.amegroups.com/article/view/10.21037/tau-23-319/coif). C.C. is an employee of Digital Health China Technologies, however there is no conflicts of interest between this research and the company. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Open Access Statement: This is an Open Access article

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019;69:7-34.
2. Escudier B, Porta C, Schmidinger M, et al. Renal cell carcinoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†. Ann Oncol 2019;30:706-20.
3. Klatte T, Rossi SH, Stewart GD. Prognostic factors and prognostic models for renal cell carcinoma: a literature review. World J Urol 2018;36:1943-52.
4. Wang C, Wu B, Shen D, et al. Establishment and validation of a prognostic nomogram for patients with renal cell carcinoma based on SEER and TCGA database. Transl Cancer Res 2023;12:1411-21.
5. Yang T, Wu Y, Zuo Y, et al. Development and validation of prognostic nomograms and a web-based survival rate calculator for sarcomatoid renal cell carcinoma in pre- and post-treatment patients. Transl Androl Urol 2021;10:754-64.
6. Usher-Smith JA, Li L, Roberts L, et al. Risk models for recurrence and survival after kidney cancer: a systematic review. BJU Int 2022;130:562-79.
7. Ben-Israel D, Jacobs WB, Casha S, et al. The impact of machine learning on patient care: A systematic review. Artif Intell Med 2020;103:101785.
8. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med 2019;380:1347-58.
9. Misawa D, Fukuyoshi J, Sengoku S. Cancer Prevention Using Machine Learning, Nudge Theory and Social Impact Bond. Int J Environ Res Public Health 2020;17:790.
10. Adamson AS, Welch HG. Machine Learning and the Cancer-Diagnosis Problem – No Gold Standard. N Engl J Med 2019;381:2285-7.
11. Zhu W, Xie L, Han J, et al. The Application of Deep Learning in Cancer Prognosis Prediction. Cancers (Basel) 2020;12:603.
12. Suarez-Ibarrola R, Hein S, Reis G, et al. Current and future applications of machine and deep learning in urology: a review of the literature on urolithiasis, renal cell carcinoma, and bladder and prostate cancer. World J Urol 2020;38:2329-47.
13. Wang F, Gao SG, Xue Q, et al. Nomogram for predicting the overall survival of the patients with oesophageal signet ring cell carcinoma. J Thorac Dis 2021;13:1315-26.
14. Jiang Y, Xie J, Han Z, et al. Immunomarker Support Vector Machine Classifier for Prediction of Gastric Cancer Survival and Adjuvant Chemotherapeutic Benefit. Clin Cancer Res 2018;24:5574-84.
15. Santra T, Delatola EI. A Bayesian algorithm for detecting differentially expressed proteins and its application in breast cancer research. Sci Rep 2016;6:30159.
16. Lorenzo D, Ochoa M, Piulats JM, et al. Prognostic Factors and Decision Tree for Long-Term Survival in Metastatic Uveal Melanoma. Cancer Res Treat 2018;50:1130-9.
17. Wang W, Liu W. Integration of gene interaction information into a reweighted random survival forest approach for accurate survival prediction and survival biomarker discovery. Sci Rep 2018;8:13202.
18. Oh SE, Seo SW, Choi MG, et al. Prediction of Overall Survival and Novel Classification of Patients with Gastric Cancer Using the Survival Recurrent Network. Ann Surg Oncol 2018;25:1153-9.
19. Kilic A, Goyal A, Miller JK, et al. Predictive Utility of a Machine Learning Algorithm in Estimating Mortality Risk in Cardiac Surgery. Ann Thorac Surg 2020;109:1811-9.
20. Witteveen A, Nane GF, Vliegen IMH, et al. Comparison of Logistic Regression and Bayesian Networks for Risk Prediction of Breast Cancer Recurrence. Med Decis Making 2018;38:822-33.
21. de Melo NB, Bernardino ÍM, de Melo DP, et al. Head and neck cancer, quality of life, and determinant factors: a novel approach using decision tree analysis. Oral Surg Oral Med Oral Pathol Oral Radiol 2018;126:486-93.
22. Ghiasi MM, Zendehboudi S. Application of decision tree-based ensemble learning in the classification of breast cancer. Comput Biol Med 2021;128:104089.
23. Zhao M, Li M, Chen Z, et al. Identification of immune-related gene signature predicting survival in the tumor microenvironment of lung adenocarcinoma. Immunogenetics 2020;72:455-65.
24. Suárez C, Campayo M, Bastús R, et al. Prognostic and Predictive Factors for Renal Cell Carcinoma. Target Oncol 2018;13:309-31.
25. Kim JK, Kim SH, Song MK, et al. Survival and clinical prognostic factors in metastatic non-clear cell renal

cell carcinoma treated with targeted therapy: A multi-institutional, retrospective study using the Korean metastatic renal cell carcinoma registry. Cancer Med 2019;8:3401-10.
26. Teishima J, Inoue S, Hayashi T, et al. Current status of prognostic factors in patients with metastatic renal cell carcinoma. Int J Urol 2019;26:608-17.
27. Ren W, Gao X, Zhang X, et al. Prognostic factors for the survival of patients with papillary renal cell carcinoma after surgical management. Clin Transl Oncol 2020;22:725-33.
28. Kourou K, Exarchos TP, Exarchos KP, et al. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2014;13:8-17.
29. Brown JA, Ni Chonghaile T, Matchett KB, et al. Big Data-Led Cancer Research, Application, and Insights. Cancer Res 2016;76:6167-70.
30. Ryu SM, Lee SH, Kim ES, et al. Predicting Survival of Patients with Spinal Ependymoma Using Machine Learning Algorithms with the SEER Database. World Neurosurg 2018;S1878-8750(18)32914-0.
31. Thio QCBS, Karhade AV, Ogink PT, et al. Can Machine-learning Techniques Be Used for 5-year Survival Prediction of Patients With Chondrosarcoma? Clin Orthop Relat Res 2018;476:2040-8.
32. Ryu SM, Seo SW, Lee SH. Novel prognostication of patients with spinal and pelvic chondrosarcoma using deep survival neural networks. BMC Med Inform Decis Mak 2020;20:3.
33. Bhambhvani HP, Zamora A, Shkolyar E, et al. Development of robust artificial neural networks for prediction of 5-year survival in bladder cancer. Urol Oncol 2021;39:193.e7-193.e12.
34. Wang S, Cai Y. Identification of the functional alteration signatures across different cancer types with support vector machine and feature analysis. Biochim Biophys Acta Mol Basis Dis 2018;1864:2218-27.
35. Sun X, Liu L, Xu K, et al. Prediction of ISUP grading of clear cell renal cell carcinoma using support vector machine model based on CT images. Medicine (Baltimore) 2019;98:e15022.
36. Sherbet GV, Woo WL, Dlay S. Application of Artificial Intelligence-based Technology in Cancer Management: A Commentary on the Deployment of Artificial Neural Networks. Anticancer Res 2018;38:6607-13.
37. Abdelhafiz D, Yang C, Ammar R, et al. Deep convolutional neural networks for mammography: advances, challenges and applications. BMC Bioinformatics 2019;20:281.
38. Yu D, Liu Z, Su C, et al. Copy number variation in plasma as a tool for lung cancer prediction using Extreme Gradient Boosting (XGBoost) classifier. Thorac Cancer 2020;11:95-102.
39. Jiang YQ, Cao SE, Cao S, et al. Preoperative identification of microvascular invasion in hepatocellular carcinoma by XGBoost and deep learning. J Cancer Res Clin Oncol 2021;147:821-33.

Cite this article as: Jiang W, Chen Z, Chen C, Wang L, Han T, Wen L. Machine learning algorithms being an auxiliary tool to predict the overall survival of patients with renal cell carcinoma using the SEER database. Transl Androl Urol 2024;13(1):53-63. doi: 10.21037/tau-23-319

Supplementary

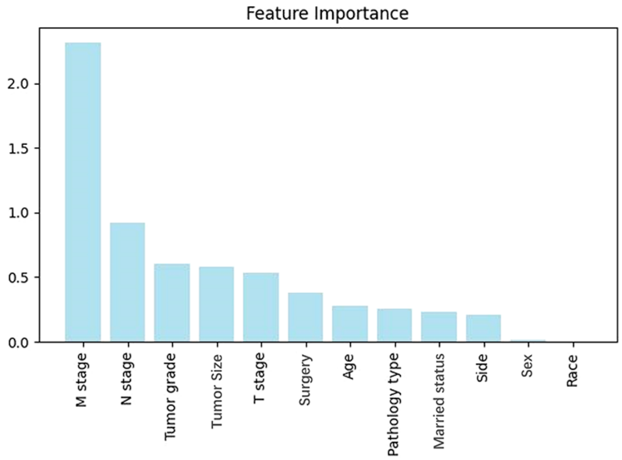

Figure S1 The importance of all features in ML model for prognostic prediction. ML, machine learning.

https://dx.doi.org/10.21037/tau-23-319