Mathematical Biosciences and Engineering

MBE, 20(7): 11676–11687.
DOI: 10.3934/mbe.2023519
Received: 06 March 2023
Revised: 20 April 2023
Accepted: 22 April 2023
Published: 08 May 2023

http://www.aimspress.com/journal/MBE

Research article

The transcriptional risk scores for kidney renal clear cell carcinoma using XGBoost and multiple omics data

Xiaoyu Hou[1], Baoshan Ma[1,*], Ming Liu[2], Yuxuan Zhao[1], Bingjie Chai[1], Jianqiao Pan[1], Pengcheng Wang[3], Di Li[4], Shuxin Liu[5,*] and Fengju Song[6,*]

[1] School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China
[2] Physical Department of Science and Technology, Dalian University, Dalian 116622, China
[3] Department of Mechanical Engineering, University of Houston, Houston 77204, USA
[4] Department of Neuro Intervention, Dalian Medical University affiliated Dalian Municipal Central Hospital, Dalian 116033, China
[5] Department of Nephrology, Dalian Medical University affiliated Dalian Municipal Central Hospital, Dalian 116033, China
[6] Department of Epidemiology and Biostatistics, Key Laboratory of Molecular Cancer Epidemiology, Tianjin, National Clinical Research Center of Cancer, Tianjin Medical University Cancer Institute and Hospital, Tianjin 300060, China

* **Correspondence:** Email: mabaoshan@dlmu.edu.cn, root8848@sina.com, songfengju@163.com; Tel: +8618624392829, +8613904088300, +8613164066716.

Abstract: Most kidney cancers are kidney renal clear cell carcinoma (KIRC) that is a main cause of cancer-related deaths. Polygenic risk score (PRS) is a weighted linear combination of phenotypic related alleles on the genome that can be used to assess KIRC risk. However, standalone SNP data as input to the PRS model may not provide satisfactory result. Therefore, Transcriptional risk scores (TRS) based on multi-omics data and machine learning models were proposed to assess the risk of KIRC. First, we collected four types of multi-omics data (DNA methylation, miRNA, mRNA and lncRNA) of KIRC patients from the TCGA database. Subsequently, a novel TRS method utilizing multiple omics data and XGBoost model was developed. Finally, we performed prevalence analysis and prognosis prediction to evaluate the utility of the TRS generated by our method. Our TRS methods exhibited better predictive performance than the linear models and other machine learning models. Furthermore, the prediction accuracy of combined TRS model was higher than that of single-omics TRS model. The KM curves showed that TRS was a valid prognostic indicator for cancer staging. Our proposed method extended the current definition of TRS from standalone SNP data to multi-omics data and was superior to the linear models and other machine learning models, which may provide a useful implement for diagnostic and prognostic prediction of KIRC.

Keywords: kidney renal clear cell carcinoma; diagnosis; transcriptional risk score; multi-omics data; XGBoost

1. Introduction

Kidney Renal Clear Cell Carcinoma (KIRC) is a highly frequent subtype of renal cancer with increasing incidence and death rates [1]. Effective prediction methods are of great significance in the prevention and treatment of KIRC. Recently, many genetic loci have been identified as markers of kidney cancer susceptibility through genome-wide association studies (GWAS). Some risk prediction methods based on GWAS have been proposed [2]. Polygenic risk scores (PRS), such as LDpred [3] showed great prospect in improving prediction for complicated disease risk. These studies suggest that PRS is effective in predicting the incidence of site-specific cancers and can be incorporated into mass screening and prevention strategies. For example, in coronary artery disease risk prediction, the PRS was able to identify more at-risk patients than standard single gene tests [4]. The PRS also showed good predictive performance for other diseases, such as type II diabetes and breast cancer.

However, the PRS only assessed the genetic risk of an individual without considering environmental exposures, while the phenotype may also change due to life-styles or external factors [5]. Besides, medical ethics and privacy hinder the public acquisition of SNP data. High-throughput sequencing technologies have been used to generate a large amount of publicly available omics data that can map the combined effects of genetic, environmental and lifestyle factors [6–9]. The application of multi-omics data may provide new insights into cancer risk prediction [10]. Linear statistical models are usually utilized in the existing PRS method for calculating the effect sizes of genetic variations [11–13]. However, the linear statistical models have certain limitations and can only be applied when specific conditions are satisfied [14]. Advanced machine learning models [15,16] can explain the nonlinear relationship between multiple variables and may improve the accuracy of risk prediction.

We employed multi-omics data and XGBoost algorithm to establish a transcriptional risk score (TRS) for KIRC. The results show that our proposed method surpasses the traditional linear models and other ML models. In the end, our proposed approach potentially may help to assess the risk of KIRC patients.

2. Materials and methods

2.1. Materials

The datasets in this study were downloaded from The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov/) project [17]. The Cancer Genome Atlas (TCGA) project provides multi-omics sequencing and microarray data for a variety of cancers, as well as clinical data for each corresponding sample. We finally selected four kinds of omics datasets which included 1895 cancer samples and 375 normal samples in total. The following four kinds of omics datasets on KIRC were included in our experiments: DNA methylation data, miRNA-seq data, mRNA expression and lncRNA expression. On the basis of previous study [18], the first and second stages are labeled as early-stage, and the third and fourth stages are labeled as late-stage. The patients used in our study ranged in age from 26 to 90 years.

For DNA methylation, the CpG sites with the highest negative correlation with gene expression were retained [19] and the CpG sites with missing values were removed in order to guarantee the high quality of datasets [20]. For miRNA, mRNA and lncRNA [21], we excluded samples with more than 20% of missing values and normalized the datasets using the minimum-maximum ratio method with a mapping range of 0 to 1. In addition, to develop TRS models, we matched the samples having all four kinds of omics data simultaneously (DNA methylation, miRNA, mRNA and lncRNA), and obtained a core dataset including 315 tumor samples and 24 normal samples.

2.2. Construction of TRS

2.2.1. Overview of TRS model

Multi-omics data and KIRC status were applied to establish two TRS models based on different phenotypes. In the first model, we labeled normal (control) and tumor (case) samples as 0 and 1, respectively. In the second model, we labeled normal, early tumor (stages I and II) and advanced tumor samples (stages III and IV) as 0, 1 and 2, respectively. The above-mentioned two TRS models were defined as case-control TRS and cancer stage TRS, respectively. The TRS can assess individual risk and improve the diagnosis of KIRC. In addition, as recent studies have found that cancer stage is highly correlated with prognosis [22], accurate construction of TRS with cancer-stage status is helpful to predict the prognosis of KIRC. The framework of TRS model is shown in Figure 1. We also provided a python program for this method (https://github.com/lab319/Muti_Omics_TRS).

2.2.2. TRS based on linear models and other machine learning models

In order to verify the effectiveness and reliability of XGBoost model, we compared the machine learning models using omics data, and did not compare the methods based on GWAS summary statistics such as LDpred [3], Lassosum [23] etc. The machine learning methods include: minimax concave penalty (MCP) [12], least absolute shrinkage and selection operator (LASSO) [24], elastic net [25] and support vector regression (SVR) [26]. Similar to the TRS approach based on the XGBoost model, each omics dataset is employed as an input to these models and the corresponding phenotype as output.

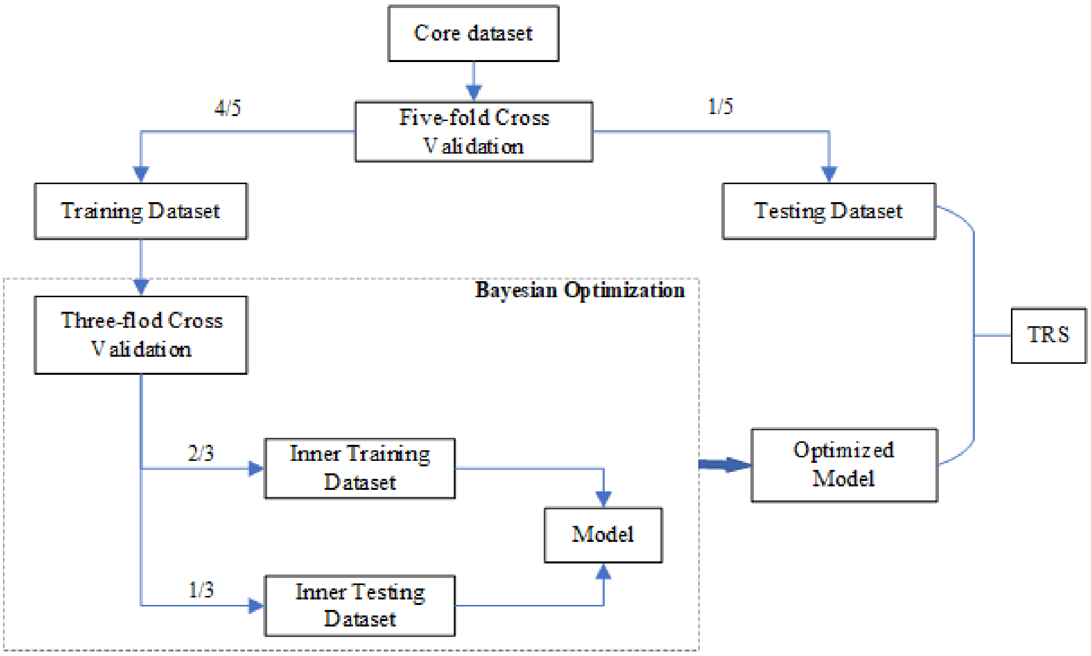

Figure 1. Schematic overview of the framework for constructing TRS model based on multiple omics data. The dataset of KIRC was split into two groups as training dataset and testing dataset based on 5-fold cross validation. We constructed TRS model by using MCP, LASSO, elastic net, SVM and XGBoost based on the training dataset. The hyper-parameters of five models were optimized using bayesian optimization and 3-fold cross validation. The TRS of testing dataset was predicted by the optimized model. The predictive performance of final models was evaluated with R^2 and AUC score.

2.2.3. Model training and evaluation

To ensure the robustness and stability of the model, this study used the 5-fold cross-validation method to conduct experiments. Cross validation can also be called rotation estimation. It is a common statistical analysis method to verify the performance of prediction models and can effectively avoid the occurrence of overfitting. In the cross-validation process, the omic dataset was randomly divided into five complementary folds. Each fold is taken as the test set, the remaining folds are taken as the training set to train the model. Bayesian optimization [27] and 3-fold inner cross-validation were applied to optimize the hyperparameters of TRS models in each training dataset. Finally, for cancer-stage trait, we evaluated the performance of different methods in the test dataset using Pearson correlation (R^2). For case-control trait, we assessed the performance of different methods using area under curve (AUC).

2.3. Combination model

In order to further upgrade the prediction performance of our method, we constructed a new model based on the TRS of each omics dataset [28]. After the processing and matching steps for raw data, we obtained a core dataset including 315 tumor samples and 24 normal samples. The TRS based on four kinds of omics datasets was used as a new biological variable for the combined model. Then, we constructed the combination model using the XGBoost model. Similarly, the above methods are used to perform hyper-parameters optimization and predictive performance evaluation.

3. Results

3.1. Data characteristics

In this study, we collected cancer tissue samples and normal tissue samples from the TCGA database. Table 1 lists the amounts of normal samples, cancer samples, biological variables and the sample size of each cancer stage in the original samples. After the processing and matching steps for raw data, we obtained a core dataset including 315 tumor samples and 24 normal samples. In addition, this paper also collected clinical information of patients including gender, race, age, tumor staging information, survival time and survival state. Patients whose tumor staging information is “stage I” and “stage II” are considered as early tumor patients. Patients with tumor staging information of “stage III” and “stage IV” are regarded as advanced tumor patients. Detailed clinical data of KIRC patients are shown in Table 2.

3.2. Predictive performance of TRS

In the case-control status, we use AUC score to evaluate the predictive accuracy of our model and other ML models. The histograms of AUC score for five different methods on four kinds of omics datasets were shown in Figure 2a. Compared with the baseline method SVR, the experiment results show that MCP performs poorly on four kinds of omics datasets, with a mean AUC score decreases of 13.91%. XGBoost performs well on four kinds of omics datasets with an average AUC score improvement of 5.85%. Elastic net is superior to SVR on 3 out of 4 datasets and the AUC score increases by 0.2%.

Identically, in the cancer-stage status, Figure 2b shows that MCP does not perform well on the four omics data sets and the average R^2 drops by 18.45% in comparison to the baseline method Lasso. SVR outperforms Lasso in three of the four data sets, with an average R^2 improvement of 4.83%. XGBoost performs well on four omics datasets with an average R^2 improvement of 6.13%. Elastic net is superior to MCP on 3 out of 4 datasets and the R^2 increases by 3.06%. Furthermore, the results show methylation data obtain better results than other omics data.

The results suggest that XGBoost surpasses other methods with an average AUC score of 0.928, which is 5.85%, 5.94%, 5.61% and 20.56% higher than SVR, Lasso, Elastic net and MCP for case-control status, respectively. Likewise, compared with the case-control status, XGBoost outperforms other assessment methods with an average R^2 of 0.494, which is 1.23%, 6.24%, 3.13% and 30.34% higher than SVR, Lasso, Elastic net and MCP for cancer-stage status, respectively. Overall, the results show that our model has acquired better accuracy when using multiple omics data to predict patients’ disease risk.

Table 1. The description of KIRC datasets from TCGA.

Omic type	Total of early-stage and late-stage tumor samples		Total of tumor samples	Total of normal samples	Total of biological variables
DNA methylation	Early-stage	190	319	160	15,837
	Late-stage	129			
miRNA	Early-stage	315	516	71	417
	Late-stage	201			
mRNA	Early-stage	351	530	72	17,630
	Late-stage	179			
lncRNA	Early-stage	339	530	72	8268
	Late-stage	191			
Common sample	Early-stage	186	315	24	
	Late-stage	129			

Note: The total of tumor samples is not equal to the sum of the early-stage and the late-stage samples, because some tumor samples have unknown kidney renal clear cell carcinoma stage.

Table 2. The detailed information of clinical data for KIRC.

Clinical data type		Amount	Percentage (%)
Gender	male	346	64.4
	female	191	35.6
Race	white man	466	86.8
	Black or African American	56	10.4
	Asian	8	1.5
	unknown	7	1.3
Age (years)	average value	61	—
	range	26–90	—
Tumor staging	early stage	326	61.0
	late stage	208	39.0
State of existence	survive	361	67.2
	death	176	32.8

3.3. Predictive performance of TRS based on combination model

We analyzed the predictive performance of the TRS combination model. The combination model shows the best predictive performance, outscoring all TRS methods mentioned in this article on the core datasets (Figure 2c,d). The results indicate that the combination model is superior to other prediction methods with AUC score of 0.976 for case-control status, which is 13.58%, 9.96%, 20.34% and 7.89% higher than the average predictive accuracy of DNA methylation, mRNA, miRNA and lncRNA, respectively. Further, the combination model largely enhances the predictive accuracy of the cancer-stage status with R^2 of 0.547 by increasing the average R^2 values 9.07%, 17.29%, 18.79% and 17.81%, respectively. In addition, compared to the model that achieves the best PA among other

TRS models, the R^2 and AUC score of combination model are improved by 3.01% and 4.05%, respectively. In summary, the combination model can achieve better prediction accuracy.

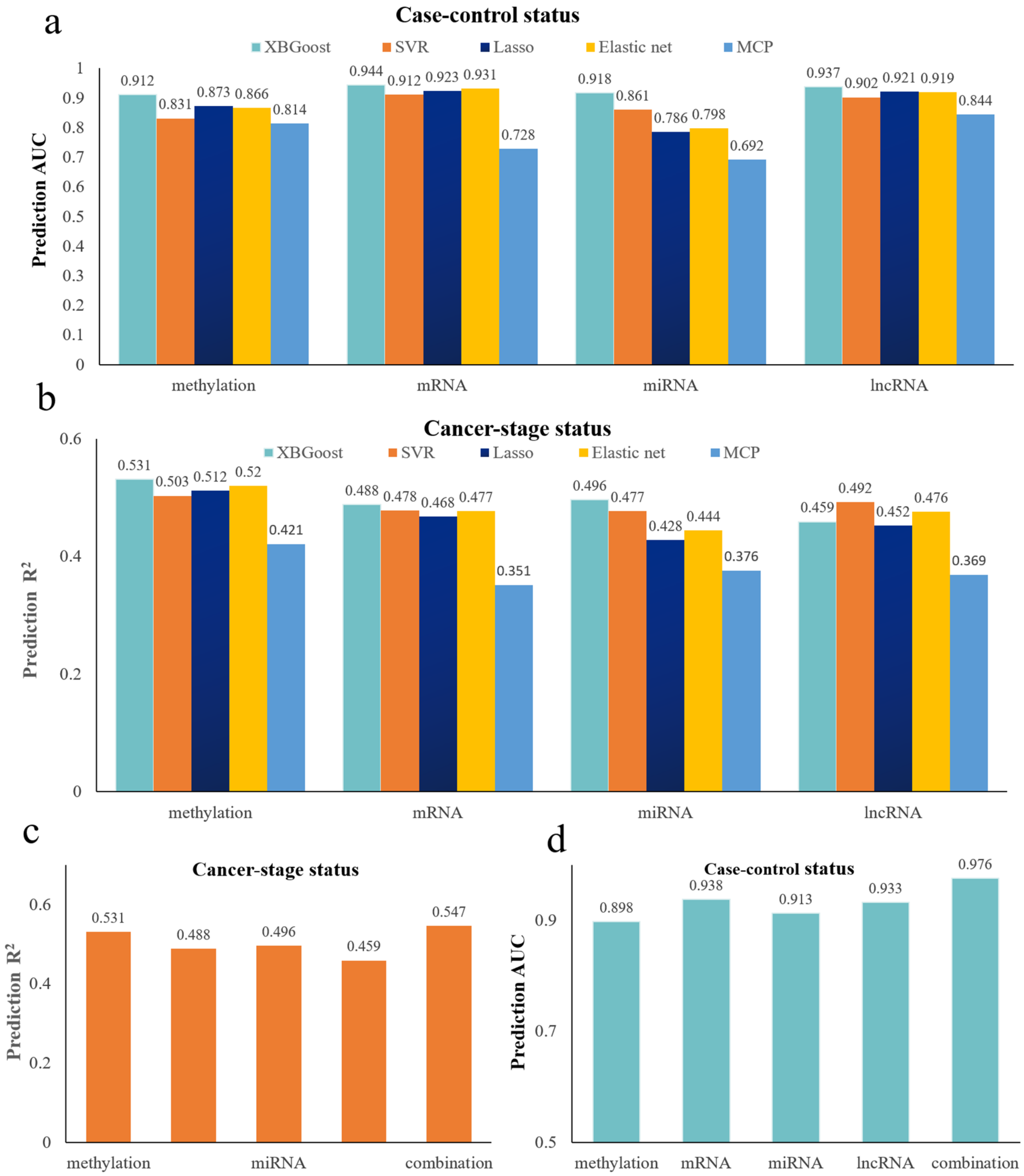

Figure 2. Predictive performance of MCP, LASSO, elastic net, SVR and XGBoost in four kinds of omics datasets. (a) Comparison results of multiple omics datasets for case-control status. (b) Comparison results of multiple omics datasets for cancer-stage status. (c) Comparison results of multiple omics datasets and combination model for cancer-stage status. (d) Comparison results of multiple omics datasets and combination model for case-control status.

3.4. Prevalence of KIRC

The main purpose of this section describes the risk stratification results for the case-control status and explores whether the prevalence of different groups has different effects on the prevention of KIRC [29]. Based on the combination model of TRS, KIRC patients were divided into 10 increasing strata and the prevalence of each stratum was calculated (Figure 3a).

In the core datasets, the prevalence is 30% in the first stratum, then reaches 100% in the other stratum. The large variation in a certain stratum is due to the relatively accurate prediction of KIRC risk in the case-control status by our proposed method. Prevalence trend diagram also shows that an individual of high TRS has greater KIRC risk than an individual of low TRS.

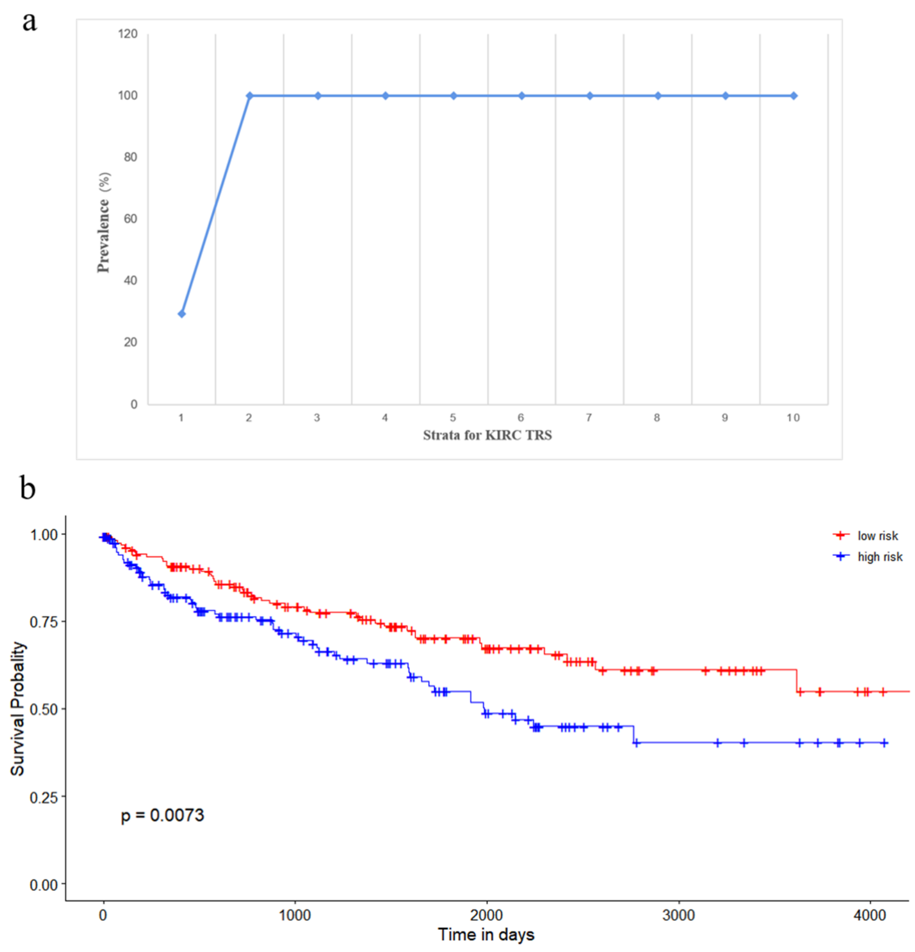

Figure 3. (a) The prevalence curve of TRS for case-control status. The sample size of 10 strata was equal and the prevalence of KIRC increased along with the TRS. The 1st stratum can be regarded as a low-risk TRS stratum and the 2nd to 10th stratum as a high-risk stratum. (b) The KM survival curve of KIRC patients in the high-risk and low-risk groups. We divided patients into high-risk and low-risk groups basing on the 50th TRS. The patients within low-risk group have better prognoses than those within high-risk group.

3.5. Prognosis prediction of kidney renal clear cell carcinoma

In this section, we aim to explore whether the TRS for cancer-stage status contributes to the prognosis of KIRC patients. First, 339 KIRC patients at the 50th percentile were divided into high-risk group and low-risk group according to TRS, which was calculated according to the combination model. Next, we generated a Kaplan-Meier curve (KM curve) [30] using each patient's survival time and status at the end of survival time. Obviously, the high-risk patients have statistically significantly worse prognoses (Figure 3b). The results indicate that the TRS on cancer stage forecasts may provide an effective prognostic tool for KIRC patients.

4. Discussion

In this paper, we proposed a novel TRS method for KIRC using multi-omics data and XGBoost model. The results based on single omics model indicate that our method has promising prediction performance than the existing linear models and other ML models. It is noticed that the combination of the four types of molecular data can obtain encouraging TRS results for both case-control and cancer-stage status. At the same time, the prediction results of 5-fold cross validation certify the robustness and dependability of the proposed method. In addition, by analyzing the prevalence trend between TRS and disease risk of KIRC, the results support the clinical understanding and application of TRS in KIRC. Eventually, we also found that our derived TRS can improve prognosis in patients with renal cancer.

Previous studies on the diagnosis and prognosis of KIRC focused mostly on the research of individual-level genotype data (SNPs) or gene expression profiling using traditional models. For example, Wei et al. used an SNPs-based approach to predict the recurrence risk of renal cell carcinoma, improving the accuracy of the prediction and investigating the factors influencing its accuracy [31]. Several studies have reported that gene expression patterns can distinguish histological subtypes of RCC, such as conventional ccRCCs, papillary types 1 and 2 carcinomas [32]. These studies obtained significant individual-level genetic risk of KIRC. Moreover, some studies have developed prognostic models to evaluate disease risk by utilizing gene expression data and clinical information. For instance, Wang et al. established an immune-related prognostic score based on 22 breast cancer cohorts consisting of a total of 6415 samples [33]. Similarly, Yang et al. investigated tumor-infiltrating lymphocytes in a large cohort of ovarian cancer patients and found that high expression levels of immune-related genes were associated with better prognosis in high-grade serous carcinomas [34]. In contrast to these studies, our investigation utilizes XGBoost algorithm and multi-omics data to construct a transcriptional risk score (TRS) model for estimating the risk of kidney renal clear cell carcinoma. In the first phenotype(case-control) analysis of our study, we obtain an AUC of 0.976 using the combination model. Therefore, the TRS based on multi-omics data and XGBoost improves the accuracy of KIRC prediction and extends the current definition of TRS from SNP data to genomics and transcriptomics data.

The present study has several strengths. Given the advantages of XGBoost model, this method may minimize the common noise and enhance the generalization ability of the model [35]. In addition, our study uses multiple omics data to construct KIRC TRS considering the interaction of genetic and environmental factors, which could provide higher prediction accuracy. Ultimately, other transcriptome data including lncRNA, mRNA and miRNA are important regulatory molecules for gene function, which play a vital role in the pathogenesis and treatment of diseases [36–38].

Although our present study uses multi-omics data for risk prediction with good predictive performance, it still has some limitations. First, the data used in the TRS method were mainly obtained from individuals of European ancestry (the white race accounting for 86.7% of the total sample size). Therefore, our findings may not be applicable to individuals of other descents. Second, the amount of KIRC data (339) used in the experiment is small, which may affect the prediction accuracy of TRS model. In our study, there are significantly more tumor samples than normal samples, which may lead to the decrease of TRS model performance. Third, this study lacks an independent validation dataset. Since it is hard to collect datasets that contain four kinds of omics data, as well as necessary clinical information. In addition, in the prediction results (Figure 2c), the prediction accuracy of the combined model is only slightly better than the methylation result, we elucidate it with the following reasons. The limited type of omics data (DNA methylation, miRNA, mRNA and lncRNA) can only generate four single-omics TRS features for the combined model, which may not provide sufficient information to greatly improve the prediction performance. The small sample size of the core dataset is also an important factor that may affect the prediction accuracy. In the future, one of our efforts is to apply our TRS model to analyze other phenotypes of KIRC using larger and balanced multiple omics datasets.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. C. D'Avella, P. Abbosh, S. K. Pal, D. M. Geynisman, Mutations in renal cell carcinoma, *Urol. Oncol. Semin. Orig. Invest.*, **38** (2020), 763–773. https://doi.org/10.1016/j.urolonc.2018.10.027
2. C. Kooperberg, M. LeBlanc, V. Obenchain, Risk prediction using genome-wide association studies, *Genet. Epidemiol.*, **34** (2010), 643–652. https://doi.org/10.1002/gepi.20509
3. B. Vilhjálmsson, J. Yang, H. Finucane, A. Gusev, S. Lindstrm, S. Ripke, et al., Modeling linkage disequilibrium increases accuracy of polygenic risk scores, *Am. J. Hum. Genet.*, **97** (2015), 576–592. https://doi.org/10.1016/j.ajhg.2015.09.001
4. A. Khera, M. Chaffin, K. Aragam, M. Haas, C. Roselli, S. Choi, et al., Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations, *Nat. Genet.*, **50** (2018), 1219–1224. https://doi.org/10.1038/s41588-018-0183-z
5. X. Chen, Z. Zhou, R. Hannan, K. Thomas, I. Pedrosa, P. Kapur, et al., Reliable gene mutation prediction in clear cell renal cell carcinoma through multi-classifier multi-objective radiogenomics model, *Phys. Med. Biol.*, **63** (2018), 215008. https://doi.org/10.1088/1361-6560/aae5cd
6. R. Lowe, N. Shirley, M. Bleackley, S. Dolan, T. Shafee, Transcriptomics technologies, *PLoS Comput. Biol.*, **13** (2017), e1005457. https://doi.org/10.1371/journal.pcbi.1005457
7. N. Rappoport, R. Shamir, Multi-omic and multi-view clustering algorithms: review and cancer benchmark, *Nucleic Acids Res.*, **46** (2018), 10546–10562. https://doi.org/10.1093/nar/gky889
8. C. P. Wild, Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology, *Cancer Epidemiol. Biomarkers Prev.*, **14** (2005), 1847–1850. https://doi.org/10.1158/1055-9965.EPI-05-0456

9. J. A. Alegría-Torres, A. Baccarelli, V. Bollati, Epigenetics and lifestyle, *Epigenomics*, **3** (2011), 267–277. https://doi.org/10.2217/epi.11.22
10. E. Zhao, L. Li, W. Zhang, W. Wang, Y. Chan, B. You, et al., Comprehensive characterization of immune- and inflammation-associated biomarkers based on multi-omics integration in kidney renal clear cell carcinoma, *J. Transl. Med.*, **17** (2019), 177. https://doi.org/10.1186/s12967-019-1927-y
11. D. Speed, D. J. Balding, MultiBLUP: improved SNP-based prediction for complex traits, *Genome. Res.*, **24** (2014), 1550–1557. https://doi.org/10.1101/gr.169375.113
12. J. Liu, K. Wang, S. Ma, J. Huang, Accounting for linkage disequilibrium in genome-wide association studies: A penalized regression method, *Stat. Interface*, **6** (2013), 99–115. https://doi.org/10.4310/SII.2013.v6.n1.a10
13. L. Lello, S. G. Avery, L. Tellier, A. I. Vazquez, G. de Los Campos, S. D. H. Hsu, Accurate genomic prediction of human height, *Genetics*, **210** (2018), 477–497. https://doi.org/10.1534/genetics.118.301267
14. S. W. Choi, T. S. Mak, P. F. O'Reilly, Tutorial: a guide to performing polygenic risk score analyses, *Nat. Protoc.*, **15** (2020), 2759–2772. https://doi.org/10.1038/s41596-020-0353-1
15. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, et al., Lightgbm: A highly efficient gradient boosting decision tree, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, (2017), 3149–3157.
16. T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2016), 785–794. https://doi.org/10.1145/2939672.2939785
17. K. Tomczak, P. Czerwińska, M. Wiznerowicz, The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge, *Contemp. Oncol.*, **19** (2015), A68–A77. https://doi.org/10.5114/wo.2014.47136
18. A. Rahimi, M. Gönen, Discriminating early- and late-stage cancers using multiple kernel learning on gene sets, *Bioinformatics*, **34** (2018), i412–i421. https://doi.org/10.1093/bioinformatics/bty239
19. Y. Yuan, E. M. V. Allen, L. Omberg, N. Wagle, A. Amin-Mansour, A. Sokolov, et al., Assessing the clinical utility of cancer genomic and proteomic data across tumor types, *Nat. Biotechnol.*, **32** (2014), 644–652. https://doi.org/10.1038/nbt.2940
20. B. Liu, Y. Liu, X. Pan, M. Li, S. Yang, S. C. Li, DNA methylation markers for pan-cancer prediction by deep learning, *Genes*, **10** (2019), 778. https://doi.org/10.3390/genes10100778
21. B. Ma, F. Meng, G. Yan, H. Yan, B. Chai, F. Song, Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data, *Comput. Biol. Med.*, **121** (2020), 103761. https://doi.org/10.1016/j.compbiomed.2020.103761
22. A. Weiss, M. Chavez-MacGregor, D. Y. Lichtensztajn, M. Yi, A. Tadros, G. N. Hortobagyi, et al., Validation study of the American Joint Committee on cancer eighth edition prognostic stage compared with the anatomic stage in breast cancer, *JAMA Oncol.*, **4** (2018), 203–209. https://doi.org/10.1001/jamaoncol.2017.4298
23. T. S. H. Mak, R. M. Porsch, S. W. Choi, X. Zhou, P. C. Sham, Polygenic scores via penalized regression on summary statistics, *Genet. Epidemiol.*, **41** (2017), 469–480. https://doi.org/10.1002/gepi.22050
24. R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. B*, **58** (1996), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
25. H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *J. R. Stat. Soc. B*, **67** (2005), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00527.x

26. A. J. Smola, B. Schölkopf, A tutorial on support vector regression, *Stat. Comput.*, **14** (2004), 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88
27. J. Snoek, H. Larochelle, R. P. Adams, Practical bayesian optimization of machine learning algorithms, *arXiv preprint*, (2012), arXiv:1206.2944. https://doi.org/10.48550/arXiv.1206.2944
28. B. Pavlyshenko, Using stacking approaches for machine learning models, in *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, (2018), 255–258. https://doi.org/10.1109/DSMP.2018.8478522
29. J. J. Barendregt, S. A. Doi, Y. Y. Lee, R. E. Norman, T. Vos, Meta-analysis of prevalence, *J. Epidemiol. Community Health*, **67** (2013), 974–978. https://doi.org/10.1136/jech-2013-203104
30. J. T. Rich, J. G. Neely, R. C. Paniello, C. C. Voelker, B. Nussenbaum, E. W. Wang, A practical guide to understanding Kaplan-Meier curves, *Otolaryngology-Head Neck Surg.*, **143** (2010), 331–336. https://doi.org/10.1016/j.otohns.2010.05.007
31. J. H. Wei, Z. H. Feng, Y. Cao, H. W. Zhao, Z. H. Chen, B. Liao, et al., Predictive value of single-nucleotide polymorphism signature for recurrence in localised renal cell carcinoma: a retrospective analysis and multicentre validation study, *Lancet Oncol.*, **20** (2019), 591–600. https://doi.org/10.1016/S1470-2045(18)30932-X
32. Y. Dor, H. Cedar, Principles of DNA methylation and their implications for biology and medicine, *Lancet*, **392** (2018), 777–786. https://doi.org/10.1016/S0140-6736(18)31268-6
33. S. Wang, Q. Zhang, C. Yu, Y. Cao, Y. Zuo, L. Yang, Immune cell infiltration-based signature for prognosis and immunogenomic analysis in breast cancer, *Briefings Bioinf.*, **22** (2021), 2020–2031. https://doi.org/10.1093/bib/bbaa026
34. L. Yang, S. Wang, Q. Zhang, Y. Pan, Y. Lv, X. Chen, et al., Clinical significance of the immune microenvironment in ovarian cancer patients, *Mol. Omics*, **14** (2018), 341–351. https://doi.org/10.1039/c8mo00128f
35. C. Zhang, Y. Ma, *Ensemble Machine Learning*, Springer, 2012. https://doi.org/10.1007/978-1-4419-9326-7
36. Y. Pan, G. Liu, F. Zhou, B. Su, Y. Li, DNA methylation profiles in cancer diagnosis and therapeutics, *Clin. Exp. Med.*, **18** (2018), 1–14. https://doi.org/10.1007/s10238-017-0467-0
37. J. Fan, K. Slowikowski, F. Zhang, Single-cell transcriptomics in cancer: computational challenges and opportunities, *Exp. Mol. Med.*, **52** (2020), 1452–1465. https://doi.org/10.1038/s12276-020-0422-0
38. T. Hou, H. Chang, H. Jiang, P. Wang, N. Li, Y. Song, et al., Smartphone based microfluidic lab-on-chip device for real-time detection, counting and sizing of living algae, *Measurement*, **187** (2022), 0263–2241. https://doi.org/10.1016/j.measurement.2021.110304