Preliminary Construction of Prognostic Risk Early Warning Model for Renal Cell Carcinoma with Type 2 Diabetes Mellitus Based on SMOTE Algorithm

YANG LIU, LINGLING MENG[1]*, JINAWEI LI, GUISONG QI AND MIAOMIAO SONG

Department of Urology, [1]Department of Endocrinology, Cangzhou Central Hospital of Hebei Province, Cangzhou, Hebei 150001, China

157 patients with renal cell carcinoma who underwent surgical treatment in our hospital from January 2019 to December 2020 were selected as the study subjects. Patients were divided into good prognosis group and poor prognosis group according to whether there were end-point events (death or distant metastasis or recurrence of tumor) during the follow-up period. The clinical data of the two groups of patients were collected, the independent risk factors for poor prognosis were screened by univariate and binary logistic regression analysis and the logistic regression model (P_1) were established. At the same time, the data set was improved based on synthetic minority oversampling technique algorithm and the early warning model (P_2) of the improved data set was constructed and the prediction efficiency of the model was compared and verified. There were 27 patients with end-point events during the follow-up period. The increased course of type 2 diabetes mellitus, high preoperative haemoglobin A1C, high body mass index and tumor–node–metastasis stage T3/T4 were independent risk factors for poor prognosis in patients with type 2 diabetes mellitus and renal cell carcinoma (p<0.05), while metformin was an independent protective factor for poor prognosis (p<0.05). Early warning model based on synthetic minority oversampling technique, oversampling algorithm $P_2=1/[1+e^{-(-13.084-0.438*X1+0.446*X2+0.096*X3-0.781*X4+1.155*X5)}]$, determination coefficient and receiver operating characteristic of P_2 model, the areas under the curve were significantly higher than those of the P_1 model. The course of type 2 diabetes mellitus, preoperative, body mass index level, tumor–node–metastasis stage and whether to receive metformin hypoglycemic therapy in patients with type 2 diabetes mellitus combined with renal cell carcinoma are closely related to poor prognosis. Based on this, the individualized early warning model established by synthetic minority oversampling technique oversampling algorithm is beneficial to patients with high risk of poor prognosis early identification.

Key words: Oversampling, renal cell carcinoma, diabetes mellitus, early warning model

Compared with non-Type 2 Diabetes Mellitus (T2DM) Renal Cell Carcinoma (RCC) patients, patients with T2DM have significantly shorter overall survival and higher tumor recurrence and mortality[1,2]. However, studies have also shown[3,4] that there is no statistical correlation between diabetes mellitus and the overall survival of renal cancer patients undergoing surgery. This suggests that the occurrence of RCC induced by T2DM and its influence on poor prognosis may be the result of the combined action of many complex factors. Therefore, it is necessary to analyze the risk factors of poor prognosis in patients with RCC complicated with T2DM and establish an individualized early warning model to take intervention measures to further improve the prognosis of patients and improve the survival rate of patients with renal cancer. In this study, the clinical data of patients with RCC complicated with T2DM were collected and an early warning model of poor prognosis of patients was established based on the Synthetic Minority Oversampling Technique (SMOTE) oversampling algorithm, in order to provide a reference for the prognosis analysis of patients with renal cancer.

***Address for correspondence**
E-mail: menglingling2015@163.com

MATERIALS AND METHODS

Research objects:

Inclusion criteria: All patients were diagnosed with RCC by histopathological examination; the patients were diagnosed with RCC for the first time and did not receive radiotherapy, chemotherapy or surgery before treatment; the patients had T2DM and the clinical diagnosis complied with the relevant standards in the guidelines for the prevention and treatment of type diabetes, and the clinical data of the patients are complete.

Exclusion criteria: Patients with metastatic renal cancer; patients with expected survival ≤2 y. A total of 157 renal cancer patients with T2DM who met the above criteria were included, including 105 males (66.9 %) and 52 females (33.1 %), with an age range of years, mean (±) years, and 80 patients with left renal cancer (51.0 %), 77 patients (49.0 %) with right renal cancer, 129 patients (82.2 %) underwent radical nephrectomy, 28 patients (17.8 %) underwent nephron-sparing surgery, and 144 patients (91.7 %) underwent open surgery, 13 patients (8.3 %) underwent laparoscopic surgery, 91 patients (58.0 %) were in T1 stage, 35 (22.3 %) in T2 stage, 20 (12.7 %) in T3 stage and 11 (7.0 %) in T4 stage.

Follow-up and grouping:

All patients were regularly reviewed and followed up after surgery. Follow-up was conducted once a month within 6 mo after surgery and every 3 mo after 6 mo, until June 2022. Endpoints were defined as tumor-specific death or postoperative tumor recurrence or postoperative distant metastasis.

Collection of clinical data:

Through the hospital electronic medical record system, the following clinical data of patients were collected, including gender, age, body mass index, duration of T2DM (time from the first diagnosis of T2DM to the time of surgery), preoperative fasting blood glucose, preoperative glycosylated Haemoglobin A1C (HbA1c), and blood glucose control methods, tumor location, tumor pathological type, Tumor–Node–Metastasis (TNM)-t stage, surgical method, surgical method and the occurrence of end-point events.

SMOTE oversampling algorithm:

SMOTE oversampling was implemented using Statistical Package for the Social Sciences (SPSS) Modeler 18.1. In this study, a small sample group is a poor prognosis group and the sample multiple (n) should be increased to represent the ratio of the number of patients in the good prognosis group to the number of patients in the poor prognosis group (rounded to the nearest integer). The specific process of SMOTE oversampling[5] is; calculate the k nearest neighbors of each sample in the infection group; randomly select a sample j from the k nearest neighbors of the sample point i in the infection group; calculate sample i and the difference Q of all variable attributes of sample j; randomly generate a value R between 0 and 1; generate a new sample=Sample i+R×Q; repeat steps to until the number of patients in the poor prognosis group reaches n times; repeat steps to until all the sample variables of the poor prognosis group have been processed. The data set expanded by this method is essentially to perform intra-class sample interpolation on the minority class samples, without changing the original spatial boundary of the samples, and has high reliability and validity.

RESULTS AND DISCUSSION

The average follow-up time was (24.13±10.57) mo. During the follow-up period, 27 patients had end-point events (17.2 %), of which 19 died. (12.1 %), 6 cases (3.8 %) had distant metastasis, and 2 cases (1.3 %) had tumor recurrence.

Univariate analysis of the prognosis of renal cancer patients with T2DM was shown in Table 1. In the univariate analysis, the variables with statistically significant differences in clinical data of the two groups of patients were used as independent variables and whether the patients had end point events as the dependent variables were used for binary logistic regression analysis (see Table 2 for the assignment of variables). The results suggest that the increased course of T2DM, high preoperative HbA1c, high Body Mass Index (BMI), and TNM stage T3/T4 are independent risk factors for poor prognosis in renal cancer patients with T2DM (p<0.05). Independent protective factor for prognosis (p<0.05). Probabilistic prediction model $P_1=1/[1+e^{-(-13.084-0.438*X1+0.446*X2+0.096*X3-0.781*X4+1.155*X5)}]$. The Hosmer-Lemeshow test was performed on the model, and the results indicated that the coefficient of determination was $R^2=0.692$ as shown in Table 3.

Based on the independent risk factors screened in 2.2, oversampling was performed by the SMOTE oversampling algorithm. In this study, a small sample group is a poor prognosis group (27 cases), the number to be expanded is n=good prognosis group/poor prognosis group=130/27≈5, and the sample size should be expanded to 27+27×5=162 cases, At this time, the ratio of the good prognosis group to the poor prognosis group was close to 1 (0.80). The logistic regression model was refitted to the oversampled data, and the results are shown in Table 4. Early warning model based on SMOTE oversampling algorithm $P_2=1/[1+e^{-(-13.084-0.438*\times1+0.446*\times2+0.096*\times3-0.781*X4+1.155*\times5)}]$ (the assignment of each variable is the same as before). The Hosmer-Lemeshow test was performed on the model, and the results indicated that the coefficient of determination was $R^2=0.833$, which was higher than that of the P_1 model as shown in Table 4. Receiver Operating Characteristic (ROC) curve analysis of prediction models P_1 and P_2 as shown in fig. 1.

TABLE 1: UNIVARIATE ANALYSIS OF CLINICAL DATA OF TWO GROUPS OF PATIENTS

Factor	Poor prognosis group (n=27)	Good prognosis group (n=130)	t/χ^2	p
Gender			0.226	0.635
Male	17 (63.0)	88 (67.7)		
Female	10 (37.0)	42 (32.3)		
Age (years, x±s)	61.24±12.72	60.49±10.46	0.326	0.745
BMI (kg/m^2)			8.204	0.004
≥28	13 (48.1)	28 (21.5)		
<28	14 (51.9)	102 (78.5)		
Duration of T2DM (years)	16.52±1.83	12.02±3.15		
Preoperative fasting blood glucose (mmol/l)	7.12±2.16	6.37±1.67	2.013	0.046
Preoperative HbA1c (%)	12.94±2.73	8.21±3.18		
History of hypertension	12 (44.4)	44 (33.8)	1.094	0.296
Antidiabetic drug regimen				
Acarbose	16 (64.0)	74 (56.9)	0.431	0.511
Metformin	13 (48.1)	101 (77.7)	9.812	0.002
Insulin	8 (29.6)	23 (17.7)	2.01	0.156
Sulfonate	4 (14.8)	20 (15.4)	0.006	0.94
Tumor site			0.01	0.918
Left	14 (51.9)	66 (50.8)		
Right	13 (48.1)	64 (49.2)		
Tumor pathological type			0.231	0.891
Clear cell carcinoma	24 (88.9)	111 (85.4)		
Chromophobe carcinoma	2 (7.4)	13 (10.0)		
Papillary RCC	1 (3.7)	6 (4.6)		
TNM-t staging			9.071	0.003
T1/T2	16 (59.3)	110 (84.6)		

T3/T4	11 (40.7)	20 (15.4)		
Surgical approach			0.033	0.856
Open surgery	25 (92.6)	119 (91.5)		
Laparoscopic surgery	2 (7.4)	11 (8.5)		
Surgical methods			1.006	0.316
Radical nephrectomy	24 (88.9)	105 (80.8)		
Nephron-sparing surgery	3 (11.1)	25 (19.2)		

TABLE 2: VARIABLE ASSIGNMENT METHODS

Type of data	Factor	Variable	Assignment method
Measurement data	Course of T2DM	x^1	Actual measurements before surgery
	Preoperative HbAc1	x^2	Actual measurements before surgery
	BMI	x^3	Actual measurements before surgery
Count data	Metformin	x^4	If used, the value is 1, if not used, the value is 0
	TNM staging	x^5	The stage is T3/T4, the value is 1 and the T1/T2 stage is 0

TABLE 3: BINARY LOGISTIC REGRESSION ANALYSIS RESULTS

Factor	Regression coefficients	Standard error	Wald value	p	Correlation	95 % confidence interval	
						Upper limit	Lower limit
Increased course of T2DM	0.765	0.265	8.304	0.004	2.148	1.277	3.613
High preoperative HbA1c level	0.635	0.226	7.921	0.005	1.887	1.213	2.936
High preoperative BMI	0.285	0.110	6.725	0.010	1.330	1.072	1.650
Using metformin	-2.624	1.287	4.158	0.041	0.073	0.006	0.903
TNM stage is T3/T4	2.283	1.149	3.944	0.047	9.802	1.030	93.255
Constant	-27.462	7.154	14.736	<0.001			

TABLE 4: LOGISTIC REGRESSION ANALYSIS BASED ON SMOTE OVERSAMPLING ALGORITHM

Factor	Regression coefficients	Standard error	Wald value	p	Correlation	95 % confidence interval	
						Upper limit	Upper limit
Increased course of T2DM	0.438	0.072	36.859	<0.001	1.549	1.345	1.785
High preoperative HbA1c level	0.446	0.067	44.134	<0.001	1.561	1.369	1.781
High preoperative BMI	0.096	0.034	7.852	0.005	1.101	1.029	1.177
Using metformin	-0.781	0.390	4.004	0.045	0.458	0.213	0.984
TNM stage is T3/T4	1.155	0.434	7.090	0.008	3.173	1.356	7.421
Constant	-13.084	1.817	51.851	<0.001	<0.001		

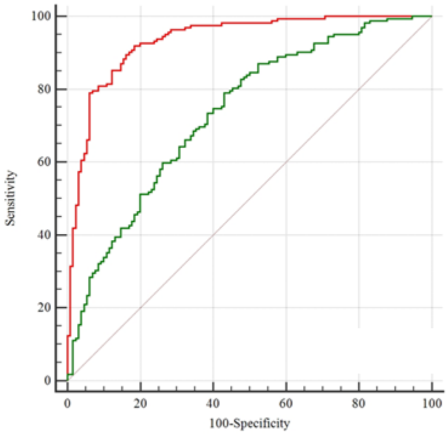

Fig. 1: ROC curve analysis of different early warning models (P_1 and P_2) for predicting poor prognosis in renal cancer patients with T2DM
Note: (🟥): P_2 and (🟩): P_1

RCC accounts for about 77.4 % of malignant tumors of the urinary system[6]. Although the incidence and mortality of renal cancer in my country are lower than the world average, they have an increasing trend year by year. According to the survey[7], the incidence rate of RCC in my country in 2015 increased by 21 % compared with 2003-2007; while the mortality rate of RCC has increased significantly since 1992, among which the mortality rate of renal cancer in male patients is increased every year. The increase is as high as 2.85 %. At present, the specific pathogenesis of renal cancer is not completely clear, but some studies believe that patient's genetic factors, environmental factors and comorbid chronic diseases play an important role in the occurrence and development of RCC. Among them, T2DM with metabolic syndrome as the main manifestation is considered to be an independent risk factor for renal cancer, and renal cancer patients with T2DM have a worse prognosis[8]. The study found that the risk of RCC in male diabetic renal cancer patients is as high as 86 %, which is an independent risk factor for tumor occurrence. It is believed that regular screening of renal disease in elderly diabetic patients is helpful for the early diagnosis and treatment of RCC. However, foreign scholars have found that there is no correlation between T2DM and overall survival of patients with RCC complicated with T2DM and without T2DM. It can be seen that the impact of T2DM on the occurrence and prognosis of RCC is complex and the correlation between the two is still controversial.

In this study, 157 patients with RCC complicated with T2DM were followed up for 2 to 36 mo. The results showed that 27 cases of end-point events occurred during the follow-up period, of which 19 cases were tumor-related deaths, and the patient survival rate was 87.9 % (138/157), lower than the 96.2 % reported in which suggests that RCC patients with T2DM have a shorter survival time[9]. A systematic review in China compared the overall survival, cancer-specific survival and tumor-free survival of RCC patients with and without T2DM. The results found that the above-mentioned survival time of patients with T2DM was significantly shortened, which was closely related to the poor prognosis of the patients significant correlation[10].

However, the specific clinical factors on how T2DM accelerates the development of poor outcomes in RCC patients are currently not fully understood. This study further explored the risk factors of patients with T2DM and RCC with poor prognosis through univariate and regression analysis. The results showed that increased course of T2DM, high preoperative HbA1c, high BMI, and TNM stage T3/T4 were independent risk factors for poor prognosis in patients with T2DM and RCC ($p<0.05$)[11]. The study also found that the duration of T2DM in female patients for more than 5 y can significantly increase the risk of RCC and it is closely related to all-cause mortality. This study showed that the risk of poor prognosis in RCC patients with T2DM increased approximately 2-fold for each additional year of T2DM duration. The longer duration of T2DM in patients means that their early exposure to hyperinsulinemia is longer and studies have confirmed that insulin cannot only promote the mitosis and proliferation of tumor cells by inducing insulin-like growth factor 1, but also up regulate epidermal growth[12]. The expression of growth factors promotes the growth of tumor angiogenesis, which leads to the progression of RCC. HbA1c is one of the main indicators of clinical diagnosis of diabetes. It reflects the blood sugar control level of the patient in the past 3 mo. An increase in the level indicates poor blood sugar control in the patient. BMI is an important indicator for measuring the degree of obesity in the human body. The patient may be in a state of metabolic syndrome. The tumor microenvironment is an important functional site for the proliferation and metastasis of malignant tumor cells and patients in the state of hyperglycemia and lipid metabolism disorder can drive the metabolism of per renal tissue to increase, produce a large number of metabolites and release them into the tumor microenvironment to promote RCC grows, invades and metastasizes[13], and in the state of hyperglycemia, immune dysfunction can also weaken the body's immune surveillance, resulting in immune escape of tumor cells and continued growth.

In addition, this study also found that in T2DM patients with RCC, receiving metformin hypoglycemic drug treatment was an independent protective factor for poor prognosis ($p<0.05$). Metformin is derived from traditional Chinese medicine goat bean and belongs to a biguanide derivative. It can inhibit hepatic gluconeogenesis and glycogenolysis, promote the uptake and utilization of glucose by peripheral target tissues and inhibit the absorption of intestinal glucose to reduce blood sugar. It is currently the first-line treatment for T2DM. Current studies suggest that metformin can have significant anti-tumor effects on various tumors such as lung cancer, liver cancer and breast cancer[14]. Conducted a follow-up study on the risk of RCC in 115 923 T2DM patients, and found that the risk of RCC in patients using metformin was significantly lower than that in patients who had never used metformin. Domestic scholars systematically evaluated the effect of metformin on the survival time of T2DM patients with RCC and found that receiving metformin therapy can significantly prolong the overall survival of patients, reduce the risk of death and help improve the prognosis of patients[15]. The above studies all suggest that metformin can significantly reduce the risk of RCC in T2DM patients and improve the prognosis of patients. This is more consistent with the results of this study. At present, the specific mechanism of metformin's anti-tumor action is not fully understood and it may be related to activating AMP-Activated Protein Kinase (AMPK) to induce tumor cell apoptosis and cell cycle arrest, promoting tumor cell autophagy, affecting tumor metabolism and enhancing the pharmacological effects of other chemotherapeutic drugs.

Unbalanced sample size is a common statistical problem in the medical field, which not only affects the specificity and sensitivity of prediction results, but also reduces the prediction accuracy of regression models. SMOTE is an oversampling algorithm, which is an effective processing method for imbalanced data. It cannot only expand the data set with low sample size, but also does not change the original spatial boundary of the sample and has high reliability and validity. Based on the SMOTE oversampling algorithm, this study effectively expanded the sample size of the poor prognosis group. After generating new data, the regression model was established again. According to the ROC analysis, the Area Under the Curve (AUC) and coefficient of determination of the P_2 model were 0.934 and 0.833, which were higher than those of the P_1 model of 0.734 and 0.692, suggesting that the regression model fitted after SMOTE oversampling had better predictive performance for poor prognosis in patients with T2DM and RCC and a higher proportion of the variance of the dependent variable is explained by the independent variable through the regression relationship.

In conclusion, the course of T2DM, preoperative HbA1c and BMI levels, TNM stage and the use of metformin for hypoglycemic therapy are closely related to the prognosis of patients with T2DM and RCC. Based on this, the individualized early warning model established by the SMOTE oversampling algorithm can significantly improve the predictive performance of poor prognosis. However, due to the small sample size and short follow-up time in this study and no external validation was carried out, this model is routinely used in preclinical studies and still needs to be validated by large-sample, multicenter and prospective studies.

Conflict of interests:

The authors declared no conflict of interests.

REFERENCES

1. Hsieh MH, Sun LM, Lin CL, Hsieh MJ, Hsu CY, Kao CH. The performance of different artificial intelligence models in predicting breast cancer among individuals having type 2 diabetes mellitus. Cancers 2019;11(11):1751.
2. Fregoso-Aparicio L, Noguez J, Montesinos L, García-García JA. Machine learning and deep learning predictive models

for type 2 diabetes: A systematic review. Diabetol Metab Syndrome 2021;13(1):1-22.

3. Talaei-Khoei A, Wilson JM. Identifying people at risk of developing type 2 diabetes: A comparison of predictive analytics techniques and predictor variables. Int J Med Inform 2018;119:22-38.
4. Nadimi-Shahraki MH, Mohammadi S, Zamani H, Gandomi M, Gandomi AH. A hybrid imputation method for multi-pattern missing data: A case study on type II diabetes diagnosis. Electronics 2021;10(24):3167.
5. Gharaibeh M, Alzu'bi D, Abdullah M, Hmeidi I, Al Nasar MR, Abualigah L, *et al.* Radiology imaging scans for early diagnosis of kidney tumors: A review of data analytics-based machine learning and deep learning approaches. Big Data Cogn Comput 2022;6(1):29.
6. Al-Hakeim HK, Hadi HH, Jawad GA, Maes M. Intersections between copper, β-arrestin-1, calcium, FBXW7, CD17, insulin resistance and atherogenicity mediate depression and anxiety due to type 2 diabetes mellitus: A nomothetic network approach. J Personalized Med 2022;12(1):23.
7. Peper KM, Guo B, Leann Long D, Howard G, Carson AP, Howard VJ, *et al.* C-reactive protein and racial differences in type 2 diabetes incidence: The REGARDS study. J Clin Endocrinol Metab 2022;107(6):e2523-31.
8. Park KH, Batbaatar E, Piao Y, Theera-Umpon N, Ryu KH. Deep learning feature extraction approach for hematopoietic cancer subtype classification. Int J Environ Res Public Health 2021;18(4):2197.
9. Ahlqvist E, Prasad RB, Groop L. 100 years of insulin: Towards improved precision and a new classification of diabetes mellitus. J Endocrinol 2022;252(3):R59-70.
10. Song X, Liu X, Liu F, Wang C. Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. Int J Med Inform 2021;151:104484.
11. Zhang X, Ardeshirrouhanifard S, Li J, Li M, Dai H, Song Y. Associations of nutritional, environmental and metabolic biomarkers with diabetes-related mortality in US adults: The third national health and nutrition examination surveys between 1988–1994 and 2016. Nutrients 2022;14(13):2629.
12. Prada M, Wittenbecher C, Eichelmann F, Wernitz A, Kuxhaus O, Kröger J, *et al.* Plasma industrial and ruminant Tran's fatty acids and incident type 2 diabetes in the EPIC-Potsdam cohort. Diabetes Care 2022;45(4):845-53.
13. Aljouie AF, Almazroa A, Bokhari Y, Alawad M, Mahmoud E, Alawad E, *et al.* Early prediction of COVID-19 ventilation requirement and mortality from routinely collected baseline chest radiographs, laboratory and clinical data with machine learning. J Multidiscip Healthc 2021;14:2017-33.
14. Thorpe LE, Adhikari S, Lopez P, Kanchi R, McClure LA, Hirsch AG, *et al.* Neighborhood socioeconomic environment and risk of type 2 diabetes: Associations and mediation through food environment pathways in three independent study samples. Diabetes Care 2022;45(4):798-810.
15. Hillary RF, Stevenson AJ, McCartney DL, Campbell A, Walker RM, Howard DM, *et al.* Epigenetic measures of ageing predict the prevalence and incidence of leading causes of death and disease burden. Clin Epigenetics 2020;12(1):1-2.

This article was originally published in a special issue, "Innovations in Biomedical Research and Drug Development" Indian J Pharm Sci 2023:85(3) Spl Issue "110-116"