

Multivariate Analysis of Cox Regression Model for Oral Squamous Cell Carcinoma Based on SEER Database

Binying Huang^{1,*}, Lixin Wang²

¹Second Clinical Medical School, Guangdong Medical University, Dongguan, Guangdong, 523109, China

²Dentistry, University of Baguio, Baguio, 2600, Philippines

*Corresponding author: 13829720448@163.com

Keywords: Oral cancer; oral squamous cell carcinoma; regression analysis; prediction model; influencing factors

Abstract: In this experiment, the National Cancer Institute SEER database was used and 13922 sample data were selected. Based on the basic demographic, clinicopathological and treatment modalities of OCSS patients, survival curves were plotted using the Kaplan-Meier method, the Cox survival analysis model was used to quantitatively assess the risk of death in patients with oral cancer, and the model was evaluated by the consistency index, and the results were presented in columns and lines. The results were presented in a line graph to provide a basis for early judgment and screening of oral cancer patients with poor prognosis. The results showed that age, race, gender, tumor site, radiotherapy, chemotherapy, T-stage, N-stage and M-stage were independent risk factors affecting oral cancer. Clinically, it is recommended to check the status of the cancer cells regularly, focusing on T, N and M stages, and to assess the patient's prognosis.

1. Introduction

Oral Cavity Cancer (OCC) refers to malignant tumors occurring in the lips, the anterior two-thirds of the tongue, the floor of the mouth, the buccal and gingival mucosa, the posterior region of the teeth, and the hard palate. Its most common clinical symptom is pain, which is also the main reason why most patients seek medical treatment. OCC is the sixth most common cancer worldwide. The latest statistics show that about 500,000 patients suffer from oral cancer every year worldwide, and 2/3 occur in developing countries. There are about 40,000 new cases of oral malignant tumors in my country every year. The mortality rate is about 1.9%. According to the GLOBCAN report, the global incidence of oral cancer increased from 185,976 cases in 1990 to 389,760 cases in 2017, a 109.6% increase[1]. The United Arab Emirates had the largest increase in the number of cases, followed by Qatar and Taiwan. OCC accounts for more than 90% of oral malignant tumors, among which squamous cell carcinoma (OSCC) is the most common, with high malignancy and poor prognosis[2]. The incidence rate of oral cancer in the region is relatively low, and the sample size is small, which cannot fully reflect the survival prognosis of patients. At present, there are few studies on the analysis of prognostic factors of OSCC at home and abroad, and the sample sizes of each study are small, and the results are also different. Large database analysis has the advantages of sufficient sample size,

complete follow-up information, and unified standards for data processing. Therefore, this study uses clinical data from the SEER database to analyze clinical pathological parameters related to OSCC, aiming to reduce the incidence of oral cancer, reduce economic costs, reduce the family burden.

2. Materials and Methods

2.1 SEER database

SEER (Surveillance, Epidemiology and End Results) database is funded by the National Cancer Institute (NCI) in 1973, after decades of accumulation. After decades of accumulation, it is now one of the most authoritative cancer statistical databases in the United States; it contains data on more than 30% of American malignant tumor patients, and collects a large amount of data related to evidence-based medicine, which provides more complete and precious case information and open and systematic evidence-based medical support for the evidence-based practice of medical practitioners as well as for basic and clinical medical research[3]. Reflecting advances in oncology research and practice, methods of controlling cancer are evolving from simply listing the development of cancer in a population by organ site to monitoring the development of cancer through histopathology and molecular subtyping (through conduction mutations and other alterations). SEER is an important demographic resource for studying the impact of diagnostic pathology across populations, geographic regions, and time, and has become a unique research resource for the practice of oncology in the United States (Note: Data from the United States). It provides incidence, survival, and mortality data for different histopathologic cancer subtypes, and data on molecular typing are expanding. The database is currently under further development to capture additional biomarker data and results from special populations, and to expand the biospecimen pool to support cutting-edge cancer research that can improve the practice of oncology.

2.2 COX regression model

Table 1: Table of influencing factor variable assignments

Factor	Variable Name	Assignment Description
Survival time/year	survival_time	/
Status	status	alive=0,dead=1
Age (years)	age	age<50=1 age50-59=2 age60-69=3 age70-79=4 age>80=5
Race	race	black=1,white=2,others=3
Sex	sex	male=1,female=0
Tumor Site	site	Buccal Mucosa=0,Floor Of Mouth=1,Gum=2,Mouth Other=3,Palate=4,Tongue5
Chemotherapy	chemotherapy	No/Unknown=0,Yes=1
Radiotherapy	radiotherapy	No/Unknown=0,Yes=1
Primary Focus	stage_T	T1=0 T2=1 T3=2 T4=3
Lymph nodes	stage_N	N0=0 N1=1 N2=2 N3=3
Distant Metastasis	stage_M	M0=0 M1=1

The COX regression model, also known as the "proportional hazards model," is a semiparametric regression model proposed by British statistician D.R. Cox (1972). Mainly used in survival data, it can evaluate the survival status of patients using truncated data without considering the distribution

of survival time, and predict the survival time as well as analyze the risk factors affecting the survival time, and it is the most important analytical method for evaluating the prognostic effect of diseases. The model has been widely used in medical research since its inception, and is by far the most widely used multifactorial survival analysis method[4]. Cox regression model also has conditions in its application, and the data need to meet the proportional hazard (PH) assumption, which means that the effect of covariates on survival does not change over time. Therefore, Cox regression should be noted in practical application due to some data constraints, when the PH assumption is not satisfied, then the method cannot be used for modeling.

In total 13922 specimen data were selected in SEER this time. The assignment section of the variable is shown in Table 1.

2.3 Line diagrams

The line diagrams, also called Nomogram Plot, is a method based on a multifactor Cox regression model (or other multifactor regression models), in which scoring criteria are developed based on the magnitude of the regression coefficients of all the predictors, and a score is assigned to each level of each value of each predictor; a total score can be calculated for multiple predictors for each patient, and a score can be calculated by the A total score is calculated for multiple predictors for each patient, and the probability of each patient's clinical outcome occurring is calculated from the score. Bar charts transform complex regression equations into visual graphs, making the results of predictive models more readable and easier for patients to assess. It is the intuitive and easy-to-understand nature of column-line diagrams that has led to their increasing interest and application in medical theory and medical practice. The contents of a column-line diagram include: variable name, score, and predicted probability. Variable names: such as age, race, patient's details in the graph, each variable corresponds to a certain scale. Length of the line reflects the magnitude of the factor's contribution to the final event. Score: includes individual and total scores, with the individual score representing the single score for each variable at different values and the total score representing the sum of the individual scores for all variables at different values. Predicted probability: generally, the predicted probability of survival is three or five years.

2.4 C-index

The C-index (Concordance Index) was first proposed by Frank E Harrell Jr, Professor of Biostatistics at Vanderbilt University in 1996, and is primarily used to calculate the difference between the COX model predictions and the true value of the survival analysis, also known as the Harrell Consistency Index; at this stage, it is most commonly used for the prediction accuracy of prognostic models for tumor patients. Discrimination, also known as Harrell's concordance index; at this stage, it is most commonly used for the predictive accuracy of prognostic models for tumor patients. The C index was calculated by randomly matching all study subjects in the data. C index is calculated by randomly pairing up all study subjects in the data being studied. Subjects are randomly paired two by two. Taking survival analysis as an example, for a patient, if the predicted survival time of the party with the longer survival time is also longer than the predicted survival time of the other party, or the predicted survival time of the one with high survival probability is longer than the other with low survival probability, it is said that the predicted results are consistent with the actual results.

In general, there are two main aspects to assessing the fit of a model, one is the fit of the model (goodness of fit), and the common assessment indexes are R-square, $-2\log L$, AIC, BIC, etc. The other is the prediction accuracy of the model, i.e., the size of the difference between the actual and predicted values of the model, the mean square error and the relative error, etc. From a clinical application perspective, we must consider the predictive accuracy of the model and also the survival time of the

prediction is longer than the survival time predicted by another model. From the perspective of clinical application, we focus more on the latter, i.e., statistical modeling is mainly used for prediction.

Generally speaking, the study considers C-index between 0.50-0.70 as low precision: between 0.71-0.90 as medium precision, and above 0.90 as high precision.

2.5 ROC Curve

The full name of ROC is "Receiver Operating Characteristic curve" (Receiver Operating Characteristic curve), which was first invented by electronic engineers and radar engineers in World War II for the detection of enemy aircraft carriers (airplanes, ships) on the battlefield, which is also known as Signal Detection Theory[5]. It was quickly followed by the introduction of psychology for perceptual detection of signals. Since then, it has been introduced into the field of machine learning for judging classification and detection results. Therefore, ROC curve is very important and commonly used for statistical analysis. The idea of "ROC curve" is to sort the samples according to the prediction results of the learner, and then predict the samples as positive examples one by one in this order, and calculate the values of two important quantities (TPR and FPR) each time, and make a graph with their horizontal and vertical coordinates respectively. Area under ROC curve, between 0.1 and 1, as a numerical value can be visualized to evaluate the classifier, the larger the value the better.

In general, $0.5 < \text{AUC} < 1$ is better than random guessing, and if the thresholds are set properly, the model will have predictive value; $\text{AUC} = 0.5$, the model has no predictive value; $\text{AUC} < 0.5$ is worse than random guessing.

2.6 Calibration curve

A calibration curve is a scatter plot of actual and predicted incidence. In essence, the calibration plot curve is a visualization of the results of the Hosmer-Lemeshow goodness-of-fit test. Currently calibration curves are commonly used to evaluate logistic regression and cox regression models[6]. This approach involves calculating the predicted and true values and then plotting them with the plotCalibration function. Ideally, the calibration curve is a diagonal line (predicted probability is equal to empirical probability). The calibration curve is not necessarily monotonically increasing. Typically, the calibration curve for Logistic Regression is very close to the diagonal line, and the calibration curve for lackadaisical models is sigmoid shaped. The calibration curve is not necessarily monotonically increasing. The calibration curve for Logistic Regression is very close to the diagonal line.

3. Results

3.1 Clinicopathologic characteristics of OSCC in SEER database

A total of 54,260 cases of OSCC were screened from the SEER database, and 40,338 cases with incomplete follow-up were excluded, so that a total of 13,922 cases of OSCC were finally included in this study, including 11,757 cases of Caucasians, 909 cases of Blacks, and 1,256 cases of other races (Table 2). The analysis of patient information from the SEER database showed that 35.20% of ACC patients were under 60 years of age, and 64.70% were over 60 years of age, with 59.30% of male patients and 40.60% of female patients. From the analysis of the site of onset of the disease, it was found that the oral mucosa accounted for 8.20%, the floor of the mouth accounted for 16.40%, the dental bed accounted for 18.80%, the palate accounted for 3.10%, the tongue accounted for 50.40%, and others accounted for 2.9%. From the analysis of chemotherapy and radiotherapy

treatment, 21.80% of the patients received chemotherapy and 78.10% did not receive chemotherapy; 40.80% of the patients received radiotherapy and 59.1% did not receive radiotherapy.

Table 2: Clinicopathologic parameters of OSCC patients included in this study in the SEER database

Clinicopathologic parameters		rate(%)
Race	White	11757(84.40)
	Black	909(6.50)
	Other	1256(9.00)
Age	≥60age	9017(64.70)
	<60age	4905(35.20)
Sex	Male	8256(59.30)
	Female	5666(40.60)
Tumor Site	Buccal Mucosa	1143(8.20)
	Floor Of Mouth	2296(16.40)
	Gum	262(18.80)
	Palate	432(3.10)
	Tongue	7025(50.40)
	Other	404(2.90)
Chemotherapy treatment status	Accepted	3035(21.80)
	Not accepted	10887(78.10)
Radiotherapy treatment status	Accepted	5683(40.80)
	Not accepted	8239(59.10)
stage_T	T1	6035(43.30)
	T2	3681(26.40)
	T3	1315(9.40)
	T4	2891(20.70)
stage_N	N0	9401(67.50)
	N1	1742(12.50)
	N2	2653(19.00)
	N3	126(0.90)
stage_M	M0	13682(98.20)
	M1	240(1.70)

3.2 Multi-element Cox regression analysis

The following (Table 3 and Table 4) are the experimental results of the multi-element Cox regression analysis:

Table 3: Table of results of Cox regression analysis

form	coef	exp(coef)	se(coef)	z	Pr(> z)
age>=60	0.64273	1.90166	0.02599	24.725	<2e-16
sexMale	0.03412	1.03470	0.02341	1.457	0.145077
raceWhite	-0.19851	0.81995	0.04170	-4.761	1.93e-06
raceOther	-0.38099	0.68318	0.05720	-6.661	2.72e-11
stage_TT2	0.52597	1.69210	0.03042	17.288	<2e-16
stage_TT3	0.92442	2.52040	0.04059	22.774	<2e-16
stage_TT4	0.97320	2.64640	0.03533	27.549	<2e-16
stage_NN1	0.53989	1.71582	0.03524	15.322	<2e-16
stage_NN2	0.86463	2.37413	0.03328	25.981	<2e-16
stage_NN3	1.02109	2.77621	0.10157	10.053	<2e-16
stage_MM1	0.83236	2.29873	0.07037	11.829	<2e-16
siteFloor Of Mouth	-0.02142	0.97881	0.04590	-0.467	0.640836
siteGum	-0.16063	0.85161	0.04537	-3.541	0.000399
siteMouth Other	-0.07925	0.92381	0.07127	-1.112	0.266167
sitePalate	0.01979	1.01998	0.06915	0.286	0.774761
siteTongue	-0.22003	0.80249	0.04154	-5.297	1.18e-07
radiationYes	-0.29097	0.74754	0.02941	-9.894	<2e-16
chemotherYes	-0.04879	0.95239	0.03288	-1.484	0.137916

Table 4: Table of factors for confidence intervals for Cox regression analysis

form	exp(coef)	exp(-coef)	lower.95	upper.95
age>=60	1.9017	0.5259	1.8072	2.0011
sexMale	1.0347	0.9665	0.9883	1.0833
raceWhite	0.8199	1.2196	0.7556	0.8898
raceOther	0.6832	1.4637	0.6107	0.7642
stage_TT2	1.6921	0.5910	1.5941	1.7961
stage_TT3	2.5204	0.3968	2.3277	2.7291
stage_TT4	2.6464	0.3779	2.4694	2.8361
stage_NN1	1.7158	0.5828	1.6013	1.8385
stage_NN2	2.3741	0.4212	2.2242	2.5341
stage_NN3	2.7762	0.3602	2.2751	3.3878
stage_MM1	2.2987	0.4350	2.0026	2.6387
siteFloor Of Mouth	0.9788	1.0216	0.8946	1.0710
siteGum	0.8516	1.1743	0.7791	0.9308
siteMouth Other	0.9238	1.0825	0.8034	1.0623
sitePalate	1.0200	0.9804	0.8907	1.1680
siteTongue	0.8025	1.2461	0.7398	0.8706
radiationYes	0.7475	1.3377	0.7057	0.7919
chemotherYes	0.9524	1.0500	0.8929	1.0158

Concordance= 0.7 (se = 0.003)

Likelihood ratio test= 3504 on 18 df, p=<2e-16

Wald test= 3697 on 18 df, p=<2e-16

Score (logrank) test = 4066 on 18 df, p=<2e-16

A higher risk factor exp (coef) indicates a higher risk of occurrence. The consistency index can be derived from a multifactorial Cox analysis as: Concordance= 0.7 (se = 0.003)

Further C-index analysis is carried out below and the results are shown in Table 5. From the table

it can be concluded that the consistency index is: 7.003111e-01. The error in the consistency index is very small compared to that obtained in the multifactorial Cox analysis above and the C-index is in the range of 0.70-0.90, which is correct for the model.

Table 5: Consistency index analysis table

C Index	Dxy	S.D.	n	missing	Uncensored
7.003111e-01	1.400622e+00	9.940180e-01	-	1.000000e+00	-7.946000e+03
Relevant Pairs	Concordant	Uncertain			
-1.498233e+08	-4.490037e+07	-4.287244e+07			

A column-line graph derived from multifactorial Cox regression analysis of oral squamous cell carcinoma is shown in Figure 1

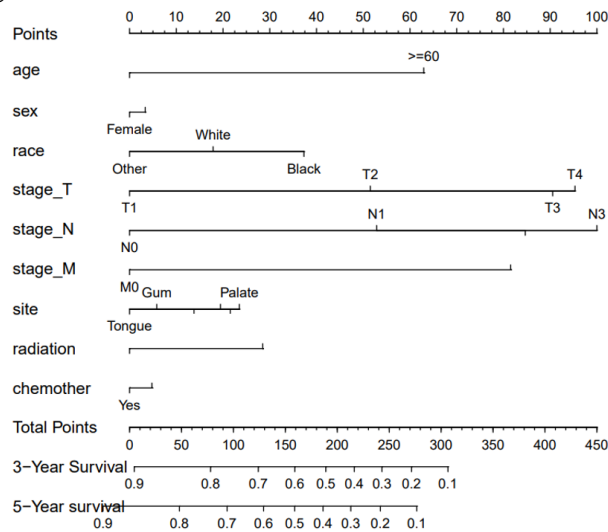


Figure 1: Lung cancer multifactorial Cox regression column line plot

3.3 ROC curve of patients with oral squamous cell carcinoma

The patient's 3- and 5-year AUC values were 0.742 and 0.81 respectively. The ROC curves are shown in Figure 2 and Figure 3.

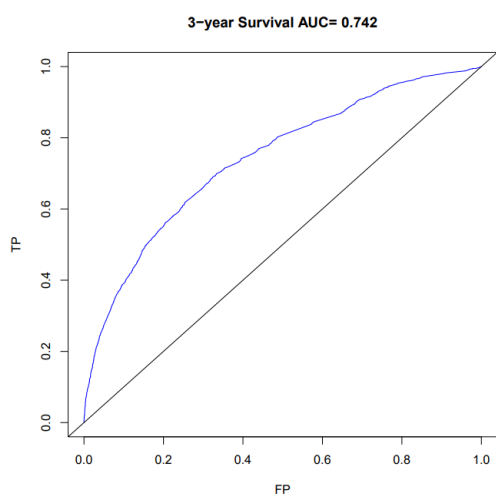


Figure 2: Three-year ROC curve

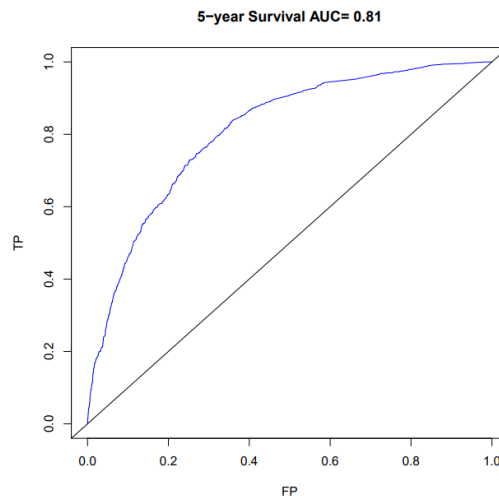


Figure 3: Five-year ROC curve

As can be seen from the results in the figure, the AUC values are all in the range of 0.70-0.90, which is better than the random guess. The correct model can have predictive value.

3.4 Calibration curve calibration chart for patients with oral squamous cell carcinoma

The 3- and 5-year calibration curves for the patients are shown in Figures 4 and 5, respectively.

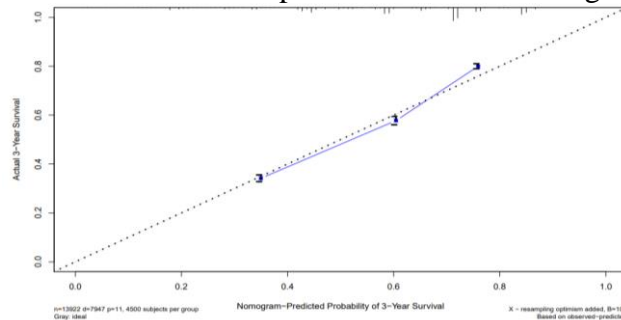


Figure 4: Three-year calibration curve

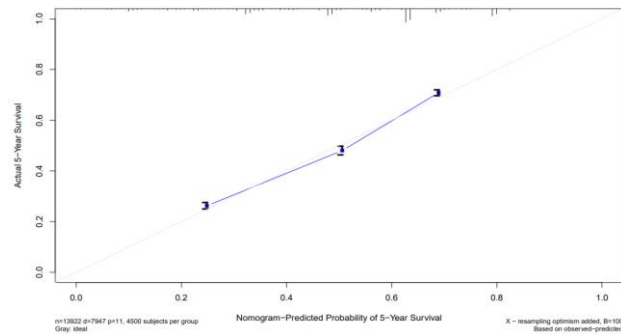


Figure 5: Five-year calibration curve

Both calibration curves show an increasing trend. The calibration curve of Logistic Regression is very close to the diagonal line, which proves that the predicted probability is very close to the empirical probability. The model is correct and has research value.

3.5 The survival curves

The survival curves of patients with OSCC are shown in Figure 6, Figure 7, Figure 8, and Figure 9.

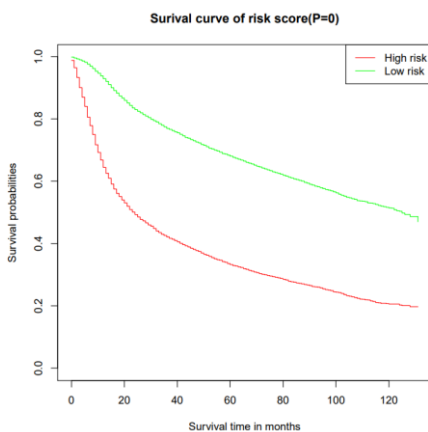


Figure 6: Risk score survival curves

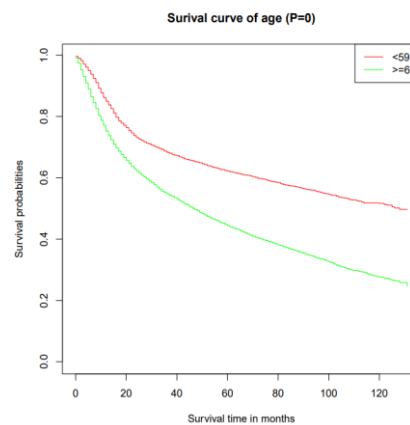


Figure 7: Age Survival Curve

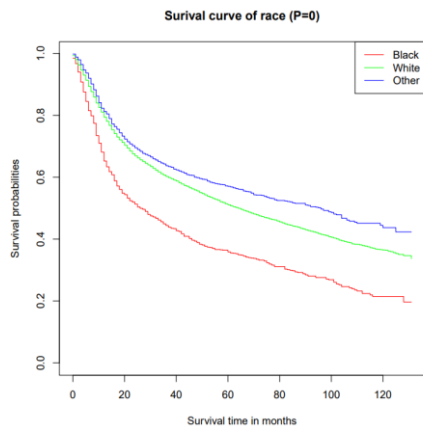


Figure 8: Race Survival Curve

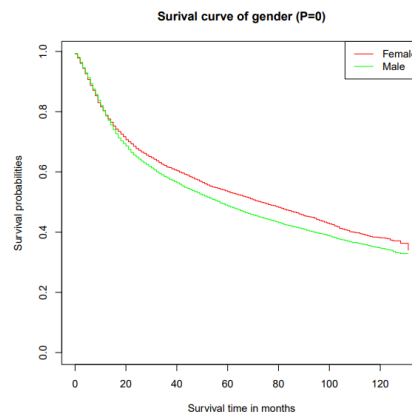


Figure 9: Gender Survival Curve

4. Experimental results

In this paper, it firstly conducted a one-way survival analysis using Kaplan-Meier on the possible prognostic factors of oral squamous cell carcinoma patients, such as age, gender, race, tumor site, T, N, and M stages, surgical methods, and other factors. The results of univariate analysis showed that gender, age, race, T-stage, N-stage, and M-stage were the relevant factors affecting the prognosis of patients with oral squamous cell carcinoma ($P < 0.05$). Based on the univariate analysis, a multifactorial Cox regression model was further constructed, and the results showed that age, gender, race, tumor location, whether receiving chemotherapy and radiotherapy, T-stage, N-stage, and M-stage were the independent risk factors affecting the prognosis of patients with oral squamous cell carcinoma (C-index ranged from 0.70 to 0.90). Oral squamous cell carcinoma was more common in men than in women ($P < 0.05$), which was mainly due to men's higher exposure to risk factors, including hammer betel, tobacco and alcohol. The risk of death from oral squamous cell carcinoma was significantly higher in whites than in blacks and other races ($P < 0.05$), which was analyzed as possibly related to factors such as dietary habits, living environment, and economic and educational levels[7]. In terms of the age distribution of the patients, the majority of patients were over 60 years old, which is similar to the results of related studies, the quality of life of middle-aged and elderly patients over 60 years old is relatively poor, and the larger tumor load leads to a higher mortality rate of the patients[8]. However, with the emergence of bad habits such as smoking and drinking in adolescents, the incidence of the disease has tended to be younger.

Clinicopathological features are important factors in tumor treatment and prognostic assessment in recent years, including T-stage, M-stage and N-stage. The results showed that compared with T1 stage, T2~T4 stages would increase the risk of death of patients. The risk of death in M1 stage was higher compared with M0 stage, which was about 1.647 times higher than that in M0 stage. This suggests that the T, N, M stages of oral cancer patients are of great significance in evaluating the prognosis of patients, which influences the choice of treatment and the prognosis of patients, and the later the T, N, M stage, the higher the risk of local recurrence and cervical lymph node metastasis. It can be seen that age and ethnicity are force majeure factors affecting the prognosis of patients with oral squamous cell carcinoma, and it is recommended that clinical attention should be paid to the T, N, and M staging in order to assess the prognosis of patients in a timely manner[9].

5. Conclusion

Most of the domestic and international studies on the prognosis of OSCC are based on small samples with limited data, and it is difficult to draw reliable conclusions. Recently, with the

continuous updating of tumor public database information, it is gradually becoming possible to use large databases for clinical analysis. The SEER public database includes information on a variety of tumors, including basic clinical data on patients with oral squamous cell carcinoma. It has a wide coverage, large sample size, and high data quality.

With the combined application of surgery, radiotherapy and chemotherapy, the local control of oral cancer has been significantly improved. However, survival of patients with oral squamous cell carcinoma has not improved significantly over the past five years. This is because the survival period of patients with oral squamous cell carcinoma is not only related to the T, N, and M stages, but also the location of the tumor determines the patient survival's important reasons. Once it occurs in the tongue, radical cure of the tumor will become difficult, and the prognosis of tumor patients will become worse and even life-threatening. To date, the most common treatment options for oral cancer patients are surgery and chemoradiotherapy. Numerous studies have shown that the following effective measures can be taken to prevent lung cancer. Stop smoking, drinking, and drinking; develop the habit of brushing and cleaning teeth; avoid bad dental restorations; avoid infection with HPV virus; conduct regular examinations, detect oral abnormalities in time, and treat them as early as possible to achieve a good prognosis. In summary, age, race, gender, tumor location, radiotherapy, chemotherapy, T stage, N stage and M stage are independent risk factors for oral squamous cell carcinoma. Therefore, the results of this study have guiding significance for the formulation of clinical treatment plans.

References

- [1] Avraham Z, Rakefet C, D H S. Oral cancer over four decades: epidemiology, trends, histology, and survival by anatomical sites.[J]. *Journal of oral pathology & medicine : official publication of the International Association of Oral Pathologists and the American Academy of Oral Pathology*, 2010, 39(4).
- [2] Vasileios Z, Dimitrios A, Anastasios I, et al. Oral Squamous Cell Carcinoma (OSCC) Imitates Denosumab-Induced Osteonecrosis of the Mandibular Alveolus: A Diagnostic Challenge.[J]. *Cureus*, 2023, 15(7).
- [3] Junmiao W, Jiayan C, Donglai C, et al. Evaluation of the prognostic value of surgery and postoperative radiotherapy for patients with thymic neuroendocrine tumors: A propensity-matched study based on the SEER database.[J]. *Thoracic cancer*, 2018, 9(12).
- [4] Y C, H M, F T, et al. Instrumental variable estimation of the marginal structural Cox model for time-varying treatments.[J]. *Biometrika*, 2023, 110(1).
- [5] Sundaresan S. Prevention, Detection and Management of Oral Cancer[M]. *IntechOpen*: 2019-12-11.
- [6] Sabry A S. Beware the IBM SPSS statistics® in multiple ROC curves analysis.[J]. *Internal and emergency medicine*, 2023, 18(4).
- [7] K-Y K, X Z, S-M K, et al. A combined prognostic factor for improved risk stratification of patients with oral cancer.[J]. *Oral diseases*, 2017, 23(1).
- [8] Fu Panfeng. Analysis of the influence of clinicopathological factors on the prognosis of patients with oral cavity squamous carcinoma[J]. *Modern Pharmaceutical Applications in China*. 2016, 10(03):64-65. DOI:10.14164/j.cnki.cn11-5581/r.2016.03.049.
- [9] Brian T, Margaret S, Louise D. Grade as a prognostic factor in oral squamous cell carcinoma: a population-based analysis of the data.[J]. *The Laryngoscope*, 2014, 124(3).