MEDS Public Health and Preventive Medicine (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/phpm.2022.020407
ISSN 2616-1915 Vol. 2 Num. 4

Multiple Factors Analysis of Cox Regression Model for Colon Cancer: Based on SEER Database

Yuxiao Fang

Faculty of Biotechnology and Food Science, Tianjin University of Commerce, Tianjin, China

Keywords: Cancer, SEER, colon cancer, regression analysis

Abstract: The mortality of colon cancer has been high and the cancer threatens people's health. It is necessary to acknowledge the risk factors of colon cancer and enhance the prognosis of colon cancer. **Method:** Surveillance, Epidemiology and End Results (SEER) database whose data of colon cancer has 44,052 samples from 1975 to 2019 were adopted for analysis, and their independent risk factors were selected by using cox multiple factors analysis, evaluating the model by using the concordance index, and finally presenting the results by using a nomogram. **Result:** The results of the analysis showed that age, race, T-stage, N-stage, and M-stage were independent factors affecting the incidence of the colon cancer. About racial factor, survival was higher in other races than in whites and in whites than in blacks. About age factor, the chance of survival was significantly lower for patients after 60 years of age. **Conclusion:** For colon cancer patients, age and race are important factors that threaten their survival. Black races and middle-aged and older adults after age 60 need to be screened regularly for their risk of developing colon cancer. The findings of this study will help patients' prognosis, help them screen for their own cancer risk, and work to reduce the incidence and mortality of colon cancer.

1. Introduction

Colon cancer (CC) is a highly prevalent gastrointestinal malignancy with high recurrence, lethality and metastasis rates. In the United States, colon cancer is the third most common cause of cancer death among men and women. More than half of these colon cancer patients have preventable causative factors, such as smoking, alcohol abuse, sedentary lifestyle, and overweight[1]. Currently, the treatment for colon cancer is mainly surgical, but also includes chemotherapy, radiotherapy, immunotherapy, cell therapy, gene therapy and targeted immunotherapy. Colon cancer staging is based on the TNM staging system of tumors defined by the American Joint Committee on Cancer (AJCC). The TNM system defines the spread of tumors from the primary segment to distant organs. T-staging identifies the depth of bowel wall invasion; N-staging is the degree of lymph node involvement, M-staging indicates the degree of distant organ metastasis, and TNM values together reflect the severity of the cancer[2].

2. Data and Methods

2.1. SEER Database

SEER, known as the Surveillance, Epidemiology, and End Results, was established in 1973 to reduce the burden of cancer in the U.S. by providing statistical information on cancer. SEER collects and publishes cancer incidence and survival data from cancer registry sites in the U.S.

These registry sites cover approximately 48% of the U.S. population and include regular follow-up data on basic patient demographics, primary tumor site, tumor form and stage at diagnosis, first treatment course, and vital status (survival) [3].

2.2. Survival Analysis

Survival analysis is a statistical analysis method that combines the study of an endpoint event with the time elapsed since that event occurred. The method has general applicability and requires less model assumptions and data than traditional classical statistics. Survival analysis, which can be applied in many different fields, is a basic idea for dealing with censored data to study the effect of experimental conditions on survival time[4].

2.2.1. Event

In the context of medicine, an event, generally refers to a death, the onset of a disease, a recurrence of a disease in remission, a recovery (i.e., return to work), or anything that may occur in an individual that is of research significance. In the same survival analysis, there may be more than one event, and when considering multiple events, such as death from different causes, the statistical problem may be expressed as a competing risk problem or a recurrent event.

2.2.2. Survival Time

Survival time, also called expiration time, is the time elapsed from a certain point to the occurrence of an endpoint event for the observed object.

Survival time can be recorded in time units such as days, weeks, months and years. In general, time units of short duration are more accurate, but in practice it is sometimes difficult to use finer time units to measure survival time.

2.2.3. Censored

Censoring of survival data means that the specific survival time of an individual cannot be determined if no expected change in the individual's status is observed during the observation period. Censoring can be classified as left censored, right censored, and interval censored. Right censored means that only the survival time is known to be longer than the observation time point. When the survival time is less than the observation time point, it is left censored; if the survival time of an individual is between two observation time points, the situation is called interval censored.

2.3. COX Regression Model

Cox proposed the proportional hazards model, or Cox model, in 1972 as a semi-parametric model. The Cox regression model (or proportional risk regression model) is a method to study the effect of several variables on the time to a specific event. This model uses survival outcome and survival time as dependent variables, allowing the simultaneous analysis of the effects of numerous factors on survival time, and enables the analysis of truncated data[5]. A total of 44,052 sample data were selected

at SEER. The assigned parts of the variables are shown in Table 1. In this table, the unit of survival time is years, and the age variable is assigned as 0 under 50 years old.

Table 1: Assignment table of influencing factor variables

Factors	Viarables	Assignments
Survival time/year	survival_time	/
Status	status	alive=0,dead=1
Age	age	age<50=0, age50-59=1, age60-69=2, age70-79=3, age>80=4
Race	race	black=0, white=1, others=2
Sex	sex	female=0, male=1
Primary focus	stage_T	T0=0, T1=1, T1a=2, T1b=3, T2=4, T3=5, T4=6, T4a=7, T4b=8, T4NOS=9, Tis=10, Tx=11
lymph gland	stage_N	N0=0, N1=1, N1a=2, N1b=3, N1c=4, N1NOS=5, N2=6, N2a=7, N2b=8, N2NOS=9, Nx=10
metastasis	stage_M	M0=0, M1=1, M1a=2, M1b=3, M1NOS=4

2.4. Nomogram

Nomogram is a graphical representation of a specific mathematical model that involves multiple predictors with a view to predicting a specific endpoint of a proportional risk model based on traditional statistical methods such as Cox survival analysis. The advantage of nomogram is that they represent an image of the relationship between three or more quantities of variables by arranging a series of graduated line segments, making the image intuitive and easy to understand, facilitating the assessment of the patient's condition and allowing the wide use of nomogram in medicine[6].

2.5. C-index

C-Index, also known as C-Statistic, or concordance index. in clinical research, the C-Index gives the probability that a patient who has randomly experienced an event has a higher risk score than a patient who has not experienced that event. The C-index is equal to the area under the ROC curve and ranges from 0.5 to 1. When the value of C-index is 0.5 means that the model does not predict the results better than random chance; a value over 0.7 means that it is a good model; a value over 0.8 means a model with good predictive power; and a value of 1 means that the model's data predicts the results perfectly[7].

2.6. R

R is a language and environment for statistical computing and graphics, free and open source software, and an excellent tool for statistical computing and statistical plotting. It is possible to extend the existing language by writing your own functions.

One of the advantages of R is the ease with which publication-quality graphs can be designed, including mathematical symbols and formulas where necessary. Very careful defaults have been taken for minor design choices in the graphs, but the user maintains full control[8].

3. Results

3.1. Multi-factor Cox Regression Analysis

The method can simultaneously detect the relationship between age, sex, race, T-stage, N-stage, and M-stage and survival. The following Table 2 are the results of a multi-element Cox regression analysis run on RGui. In Table 2, exp(coef) is the hazard ratio (HR). Table 3 shows the confidence interval factors for the cox regression analysis. The lower .95 and upper .95 in Table 3 indicate 95% confidence intervals.

The concordance index can be derived from Table 3 as follows: Concordance= 0.785 (se = 0.002). Further C-index analysis was performed and the results are shown in Table 2, which shows that the concordance index is 7.851416e-01. The error in the concordance index is small and the C-index is greater than 0.7, so the model is correct.

Call:

coxph(formula = fmla1, data = seer)

n= 30660, number of events= 15875

Table 2: Table of results of Cox regression analysis

Category	coef	exp(coef)	se(coef)	z	Pr(>|z|)	/
Age50-59	0.09285	1.0973	0.03865	2.403	0.01628	*
age60-69	0.37776	1.45902	0.03585	10.536	< 2e-16	***
age70-79	0.84572	2.32965	0.03481	24.293	< 2e-16	***
age>=80	1.60539	4.97983	0.03402	47.194	< 2e-16	***
sexMale	0.09809	1.10307	0.01604	6.115	9.66E-10	***
raceWhite	-0.21861	0.80364	0.02701	-8.092	5.85E-16	***
raceOther	-0.31427	0.73032	0.03492	-9	< 2e-16	***
stage_TT1	-1.00404	0.3664	0.11975	-8.384	< 2e-16	***
stage_TT1a	-2.34627	0.09573	0.24853	-9.441	<2 e-16	***
stage_TT1b	-2.46831	0.08473	0.46302	-5.331	9.78E-08	***
stage_TT2	-1.32439	0.26596	0.12177	-10.877	< 2e-16	***
stage_TT3	-1.03905	0.35379	0.11867	-8.756	< 2e-16	***
stage_TT4	-0.79082	0.45347	0.21396	-3.696	0.000219	***
stage_TT4a	-0.64262	0.52591	0.11992	-5.359	8.38E-08	***
stage_TT4b	-0.4713	0.62419	0.11992	-3.93	8.50E-05	***
stage_TT4NOS	-0.47326	0.62297	0.25221	-1.876	0.060595	.
stage_TTis	-1.17257	0.30957	0.13323	-8.801	< 2e-16	***
stage_TTX	-0.09124	0.9128	0.11876	-0.768	0.442306	
stage_NN1	0.24549	1.27825	0.08934	2.748	0.005999	**
stage_NN1a	0.02345	1.02373	0.03248	0.722	0.47036	
stage_NN1b	0.17355	1.18952	0.03059	5.674	1.40E-08	***
stage_NN1c	0.30558	1.35741	0.0705	4.335	1.46E-05	***
stage_NN1NOS	0.58385	1.79292	0.04154	14.055	< 2e-16	***
stage_NN2	0.47558	1.60895	0.13555	3.509	0.00045	***
stage_NN2a	0.38601	1.4711	0.03317	11.636	< 2e-16	***
stage_NN2b	0.68989	1.99349	0.03144	21.943	< 2e-16	***
stage_NN2NOS	0.83973	2.31574	0.15255	5.505	3.70E-08	***
stage_NNX	0.49489	1.64033	0.03382	14.634	< 2e-16	***
stage_MM1	1.36857	3.92972	0.12554	10.902	< 2e-16	***
stage_MM1a	1.17782	3.2473	0.02429	48.492	< 2e-16	***
stage_MM1b	1.42922	4.17546	0.02599	54.997	< 2e-16	***
stage_MM1NOS	1.18662	3.27599	0.06484	18.301	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3: Table of factors of confidence intervals for Cox regression analysis

Category	exp(coef)	exp(-coef)	lower .95	upper .95
age50-59	1.0973	0.9113	1.01725	1.1837
age60-69	1.45902	0.6854	1.36001	1.5652
age70-79	2.32965	0.4292	2.176	2.4942
age>=80	4.97983	0.2008	4.65864	5.3232
sexMale	1.10307	0.9066	1.06892	1.1383
raceWhite	0.80364	1.2443	0.7622	0.8473
raceOther	0.73032	1.3693	0.68201	0.7821
stage_TT1	0.3664	2.7293	0.28975	0.4633
stage_TT1a	0.09573	10.4465	0.05881	0.1558
stage_TT1b	0.08473	11.8024	0.03419	0.21
stage_TT2	0.26596	3.7599	0.2095	0.3377
stage_TT3	0.35379	2.8265	0.28037	0.4464
stage_TT4	0.45347	2.2052	0.29815	0.6897
stage_TT4a	0.52591	1.9015	0.41576	0.6653
stage_TT4b	0.62419	1.6021	0.49345	0.7896
stage_TT4NOS	0.62297	1.6052	0.38	1.0213
stage_TTis	0.30957	3.2303	0.23842	0.4019
stage_TTX	0.9128	1.0955	0.72325	1.152
stage_NN1	1.27825	0.7823	1.07293	1.5229
stage_NN1a	1.02373	0.9768	0.96058	1.091
stage_NN1b	1.18952	0.8407	1.1203	1.263
stage_NN1c	1.35741	0.7367	1.18224	1.5585
stage_NN1NOS	1.79292	0.5577	1.65273	1.945
stage_NN2	1.60895	0.6215	1.23357	2.0985
stage_NN2a	1.4711	0.6798	1.37849	1.5699
stage_NN2b	1.99349	0.5016	1.87436	2.1202
stage_NN2NOS	2.31574	0.4318	1.71728	3.1228
stage_NNX	1.64033	0.6096	1.53512	1.7527
stage_MM1	3.92972	0.2545	3.07259	5.026
stage_MM1a	3.2473	0.3079	3.09633	3.4056
stage_MM1b	4.17546	0.2395	3.96811	4.3936
stage_MM1NOS	3.27599	0.3053	2.88504	3.7199

Concordance= 0.785 (se = 0.002)

Likelihood ratio test= 16043 on 32 df, p=<2e-16

Wald test = 17614 on 32 df, p=<2e-16

Score (logrank) test = 21659 on 32 df, p=<2e-1

Table 4 shows the consistency index, which is used to evaluate the degree of differentiation between the results predicted by this cox model and the true situation.

Table 4: Consistency index analysis table

C Index	Dxy	S.D.	n	missing
7.851416e-01	1.570283e+00	9.964940e-01	-3.065900e+04	1.000000e+00
uncensored	Relevant Pairs	Concordant	Uncertain	
-1.587400e+04	-6.758555e+08	-1.452132e+08	-2.576983e+08	

The Figure 1 is the nomogram and the Figure 2 is the survival curves which are all derived from the multi-factor Cox regression analysis on colon cancer. Image(a) in Figure 2 is a risk control, with the green line representing low risk and the red line representing high risk. Image(b) shows the survival curve of age, and it is clear that the survival time of patients after 60 years age is significantly lower. Image(c) shows the survival curve of sex, and it is clear that sex has little effect on survival

time. Image(d) shows the survival curve of race, and it is clear that the survival time of black patients is less than that of white patients, and the survival time of white patients is less than that of other races.

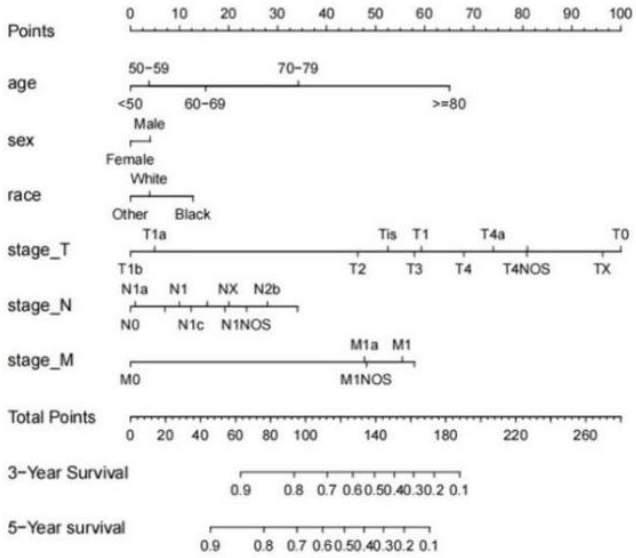

Figure 1: Multi-factor Cox regression nomogram for colon cancer

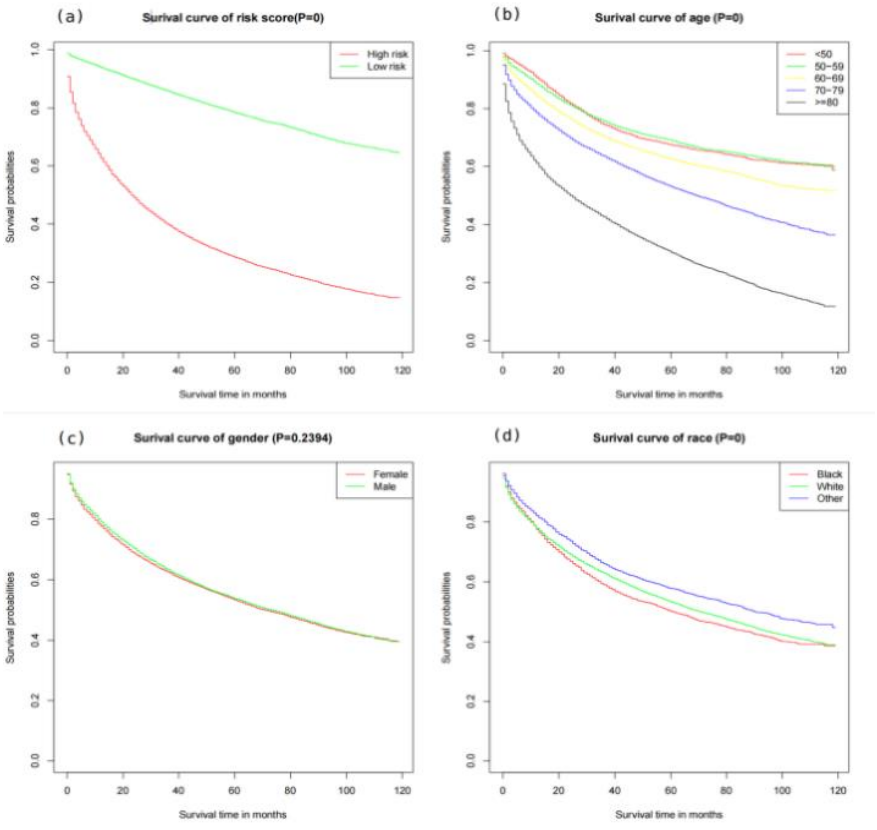

Figure 2: Survival curve graph

3.2. Analysis of results

By analyzing the risk factors for colon cancer with different clinical characteristics, the differences between independent risk factors for colon cancer patients were explored.

The analysis showed that age, sex, race, T-stage, N-stage, and M-stage are all risk factors that affect patients with colon cancer. And then, analysis was stratified by different age, and the results showed that different age had different effects on prognosis. Therefore, age are also independent factors affecting the prognosis of colon cancer patients. The risk factors affecting prognosis also varied by age.

4. Discussion

In China, colon cancer has been a serious threat to the health of the country's population, and the high incidence of colon cancer has been caused by increased environmental pollution and changes in social habits[9].

Studies have shown that the following effective measures can be taken to prevent the development of colon cancer:

(1) Reduce alcohol intake and quit smoking, as continuous alcohol intake and smoking can lead to a much higher risk of cancer.

(2) Reduce the intake of unhealthy foods and maintain a normal weight range. In one study, women with a body mass index (BMI) greater than 29 had a significantly increased risk of colon cancer.

(3) Family history of cancer greatly affects the risk of cancer, so family members with a history of colon cancer or middle-aged or older people over 50 should have regular cancer screenings.

(4) Good exercise habits can significantly reduce the risk of cancer.

References

[1] Siegel Rebecca L., Miller Kimberly D., Goding Sauer Ann, et al. Colorectal cancer statistics, 2020 [J]. CA: A cancer journal for clinicians, 2020, 70(3): 145-164.

[2] Xue SQ, He J, Tang Y. Advances in colon cancer treatment [J]. China Pharmaceutical Science, 2022, 12(09): 58-61.

[3] About the SEER Program [EB/OL] [2022-9-29]. https://seer.cancer.gov/about/

[4] Cai Meng. A Review of Survival Analysis Theory and Its Application Research [J]. Value Engineering, 2016, 35(10): 19-21.

[5] Stel V S, Dekker F W, Tripepi G, et al. Survival Analysis II: Cox Regression [J]. Nephron Clinical Practice, 2011, 119(3): c255-60.

[6] Nomogram [EB/OL] [2022-9-29]. https://www.britannica.com/science/nomogram

[7] C-Statistic: Definition, Examples, Weighting and Significance [EB/OL] [2022-9-29]. https://www.statisticshowto.com/c-statistic/

[8] About R[EB/OL] [2022-9-29]. https://www.r-project.org/about.html

[9] Shih, S.P. Clinicopathological and survival differences between stage II and stage IIIA colon cancer patients [D]. Fujian Medical University, 2019.