Statistics and Application 统计学与应用, 2022, 11(4), 943-960
Published Online August 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.114098

Hans汉斯

基于SEER数据库盲肠癌患者数据的生存预后分析

雷　杰[1]，陈　浪[1]，韩元全[1]，王　猛[2]

[1]重庆理工大学理学院，重庆
[2]重庆南开(融侨)中学，重庆

收稿日期：2022年8月2日；录用日期：2022年8月12日；发布日期：2022年8月24日

摘　要

本文主要分析了来自美国癌症数据库SEER的盲肠癌数据。首先通过随机生存森林模型(Random survival forest)进行盲肠癌的独立预后因素初步筛选，筛选出来的变量为：AJCC (American Joint Committee on Cancer)分期、肿瘤大小、年龄、婚姻状况、组织学分级、化疗状况、种族、放疗状况。然后通过筛选出来的变量分别建立了多因素Cox比例风险回归模型和多因素竞争风险模型。结果表明：Cox比例风险回归模型中，化疗治疗、已婚、肿瘤直径大小在1 cm以上的为盲肠癌患者生存预后的保护因素，年龄大于65岁患者、放疗治疗、AJCC分期大于I、组织学等级高于一级、婚姻状况为其它的因素为危险因素；在竞争风险模型中，化疗治疗、肿瘤直径大小在1 cm以上变量为盲肠癌患者生存预后的保护因素，年龄大于65岁患者、AJCC分期大于I、组织学等级高于I级、放疗治疗都为危险因素。在模型的比较中，竞争风险模型更胜一筹，在对于存在竞争事件的生存分析中，选择基于竞争风险构建的预测模型不仅准确度高，而且更具合理性。

关键词

盲肠癌，Cox比例风险模型，随机生存森林，竞争风险模型，SEER数据库

Survival Prognosis Analysis in Patients Data with Cecum Cancer: Based on the SEER Database

Jie Lei[1], Lang Chen[1], Yuanquan Han[1], Meng Wang[2]

[1]School of Science, Chongqing University of Technology, Chongqing
[2]Chongqing Nankai (Rongqiao) Secondary School, Chongqing

Received: Aug. 2nd, 2022; accepted: Aug. 12th, 2022; published: Aug. 24th, 2022

文章引用: 雷杰, 陈浪, 韩元全, 王猛. 基于 SEER 数据库盲肠癌患者数据的生存预后分析[J]. 统计学与应用, 2022, 11(4): 943-960. DOI: 10.12677/sa.2022.114098

Abstract

This article mainly analyzes the cecum cancer data from the US cancer database SEER. Firstly, the independent prognostic factors of cecum cancer were preliminarily screened by the Random survival forest, and the variables screened out were: AJCC Stage, Tumor Size, Age, Marital status, Grade, Chemotherapy status, Race, and Radiotherapy status. Then, the multi-factor Cox proportional risk regression model and the multi-factor competitive risk model were established by the filtered variables. The results showed that in the Cox proportional risk regression model, chemotherapy treatment, marriage, and tumor diameter size of more than 1 cm were the protective factors for survival and prognosis of patients with cecum cancer, and patients with age greater than 65 years old, radiotherapy treatment, AJCC stage was greater than I, Grade was higher than grade I, and marital status was other factors as risk factors. In the competitive risk model, chemotherapy therapy and tumor diameter size of more than 1 cm were the protective factors for survival prognosis in patients with cecum cancer, and patients older than 65 years old, AJCC stage greater than I, Grade higher than grade I, and radiotherapy therapy were all risk factors. In the comparison of models, the competitive risk model is superior, and in the survival analysis of the existence of competitive events, the selection of a prediction model based on competitive risk is not only more accurate, but also more reasonable.

Keywords

Cecum Cancer, Cox Proportional Risk Model, Random Survival Forest, Competing-Risks Model, SEER Database

1. 引言

结直肠癌(Colorectal cancer)是全球第三大常见恶性肿瘤和第四大癌症相关死亡原因[1]，因此是一种全球关注的癌症[2]。根据最近的一项研究，被诊断为盲肠癌的患者的预后仍然比上升结肠癌患者更差，被诊断为盲肠癌的患者需要承担更多的疾病负担[3]。因此，探索影响盲肠癌患者预后的风险因素将有助于临床医生制定对这些患者有利的个性化诊断和治疗方案[4]。

本文选取来自于 SEER 数据库中的 5240 例盲肠癌患者数据进行生存分析研究，考虑到机器学习等大数据分析手段目前充分应用到统计学学科中，它能够适应各类复杂条件，本文首先引用了随机生存森林模型(Random survival forests)对影响盲肠癌生存预后因素进行初步筛选[5]，其次基于筛选后的因素分别使用 Cox 比例风险模型[6]和竞争风险模型[7]对盲肠癌患者的生存预后进行预测研究。

2. 数据说明与预处理

本文纳入分析的变量为 9 个，具体的变量说明如表 1 所示。因变量为生存时间和生存结局，特别这里的生存结局不单只有存活和死亡两种情况，还包括了另一类结局事件即由于出现了竞争事件导致死亡结局的出现，相比较为原来的死亡结局的二分类变量，这里扩充为三分类变量。

处理完数据后剩余 2551 例，将数据处理并重新进行编码，并按照训练集：测试集为 5:5 划分，其中

训练集为 1275 例，测试集为 1276 例。对训练集的数据进行相关模型的建立，基于构建的模型于测试集中进一步检验说明，分析预测的准确性和拟合效果。

Table 1. Variable descriptions and assignment representations
表 1. 变量说明及赋值表示

变量名	英文名	变量赋值	
年龄	Age	≤64-0	>65-1
性别	Sex	Male-0	Female-1
种族	Race	Black-1	White-2
		Asian or Pacific Islander-3	Others-4
AJCC 分期	AJCC Stage	I、IE、IEA、IEB-1	II、IIA、IIB、IIC、IIE、IIEA、IIEB、IIES、IIESA、IIESB、IISB-2
		IIIA、IIIB、IIIC、IIIE、IIIEA、IIIEB、IIIESA-3	IV、IVA、IVB-4
放疗状态	Radiotherapy	None-0	Yes-1
化疗状态	Chemotherapy	None-0	Yes-1
组织学分级	Grade	Grade I-1	Grade II-2
		Grade III-3	Grade IV-4
婚姻状况	Marital status	Unmarried-1	Married-2
		Others-3	
肿瘤大小	Tumor Size	最大直径 ≤ 1cm-1	1 cm < 最大直径 ≤ 3 cm-2
		3 cm < 最大直径 ≤ 5 cm-3	最大直径 > 5 cm-4
结束随访事件	status	Alive-0	dead due to Cecum cancer-1
		dead of other cause-2	

3. 模型预测实证分析

本章首先构建随机生存森林预测模型，并用机器学习中的 VIMP (variable importance)法和最小深度法结合对影响盲肠癌生存预后因素进行初步筛选，然后分别构建 Cox 比例风险模型和竞争风险模型对影响盲肠癌患者预后的因素进行分析确定。

3.1. 随机生存森林模型

用所有纳入分析的 9 个变量构建随机生存森林模型，表 2 展示出该模型在训练集和测试集上的结果，该模型默认生成 500 个二元生存树，平均每个生存树有 15 个终端节点，模型在训练集上的错误率为 27.1%，而在测试集上的错误率为 27.8%，测试集和训练集相差不大，效果较好。

Table 2. Random forest model training set and test set comparison
表 2. 随机森林模型训练集和测试集对比

	训练集	测试集
Sample size (样本量)	1275	1276
Number of deaths (深度)	659	696
Number of trees (树的数量)	500	500

Continued

Forest terminal node size (节点大小)	15	15
Average no. of terminal nodes (终端节点)	53.548	54.062
No. of variables tried at each split (mtry)	3	3
Total no. of variables (变量数)	9	9
Resampling used to grow trees (重抽样方法)	swor	swor
Resample size used to grow trees (重抽样次数)	806	806
Analysis	RSF	RSF
Family	surv	surv
Splitting rule	logrank*random*	logrank*random*
Number of random split points (随机分割点数)	10	10
(OOB) CRPS	0.16697752	0.16599417
(OOB) Requested performance error (预测错误率)	0.27111016	0.2782446

图 1 表示随机生存森林模型生成的生存树的数量与模型预测错误率大小的关系图，随着生存树数量的增加，其预测错误率明显降低；当生存树增加到一定数量后，预测错误率曲线趋于平稳(27.1%)，所以选择该随机生存森林模型树的数量选择合适。

随机生存森林模型可以对变量的重要度进行排名，因此利用该模型对影响盲肠癌生存预后因素进行初步筛选。图 2 表示 VIMP 法和最小深度法相结合的散点图。其中，蓝色点代表 VIMP 值大于 0，红色则代表 VIMP 值小于 0；主对角线以上的点表示 VIMP 排名更高，主对角线以下的点表示最小深度法排名更高。根据综合排名，决定去掉性别变量，则初步筛选剩下影响盲肠癌生存预后的因素为 AJCC 分期、肿瘤大小、年龄、婚姻状况、组织学分级、化疗状况、种族、放疗状况。

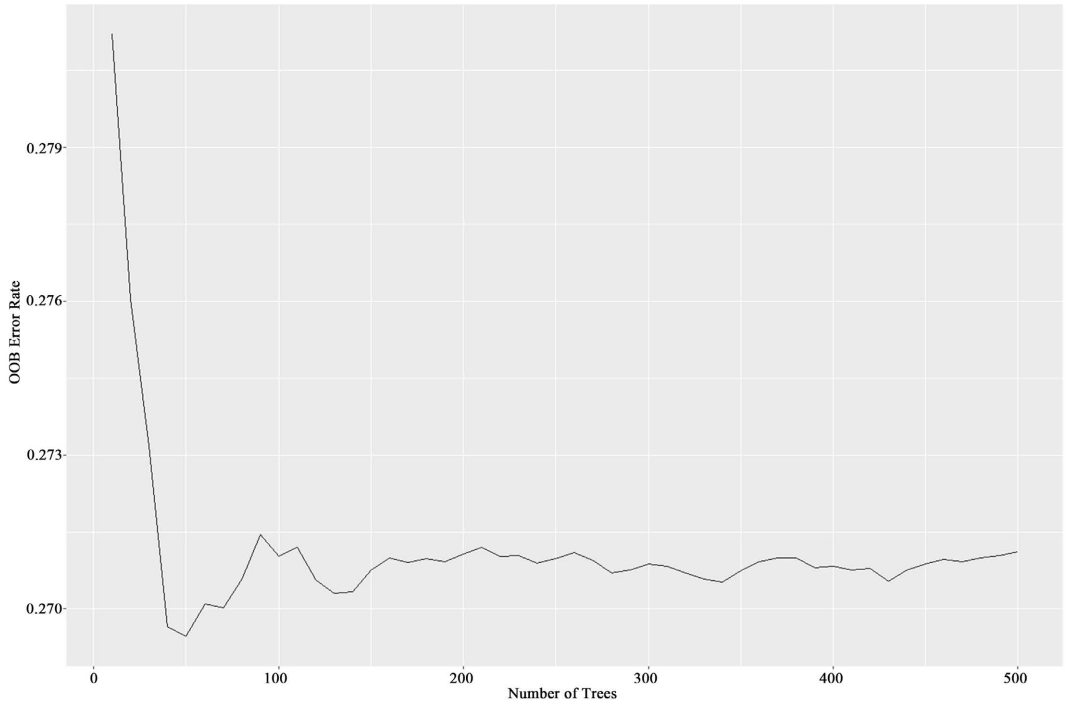

Figure 1. Models with different numbers of survival trees predict error rates
图 1. 不同数量生存树下的模型预测错误率

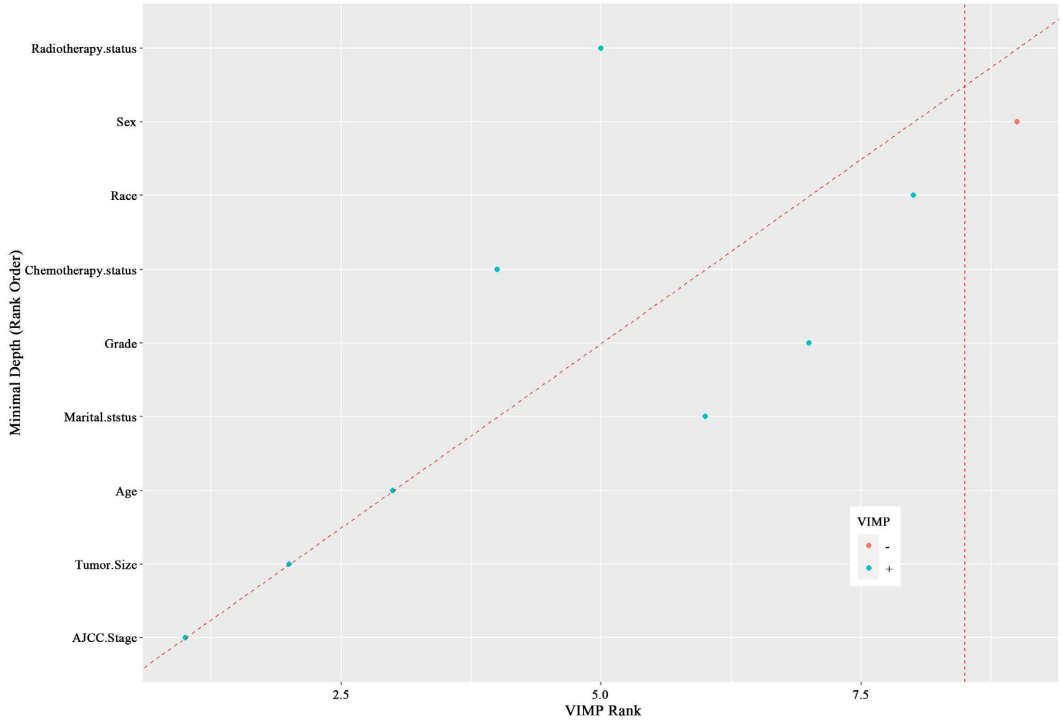

Figure 2. The VIMP method and the minimum depth method combined variable selection
图 2. VIMP 法和最小深度法结合变量筛选

3.2. Cox 比例风险模型

3.2.1. K-M 生存曲线

本小节考虑基于机器学习方法初步筛选出来的变量构建 Cox 比例风险模型。图 3~6 分别为不同年龄、种族、AJCC 分期、化疗状况、放疗状况的 K-M 生存曲线，并给出对数秩检验，由于篇幅限制，文中只给出前 4 个变量的生存曲线。除了种族变量没有通过对数秩检验，其它 7 个变量都通过了对数秩检验。如图 3，年龄大于 65 岁的患者整体生存概率要低于年龄小于 65 岁的患者；图 5 中，AJCC 分期为 IV 期的生存概率要比其它分期生存概率低的多，在生存时间为 15 个月时，生存概率就降到 50%；值得注意的是图 6 中，选择放疗治疗的患者要比选择不放疗治疗的患者生存概率低很多，因为放疗治疗的患者很少，只有当盲肠癌病情更加严重的患者会选择进行放疗治疗，因此其生存概率低于不放疗治疗的患者。

3.2.2. 基于 Cox 比例风险模型的单因素与多因素分析

为了更进一步分析影响盲肠癌患者生存预后影响因素。表 3 为 Cox 比例风险模型的单因素与多因素分析表，单因素和多因素分析结果基本一致。在单因素分析表中，只有种族变量对应的 p 值均大于 0.05，这说明这些变量对于生存预后而言不是独立的预后因素。年龄、AJCC 分期、放疗状况、化疗状况、组织学分级、肿瘤大小、婚姻状况这些变量都大致通过了显著性检验，这些都是盲肠癌生存预后独立的预后因素。其中，对于年龄来说，HR 危险比为 1.72，表示明在其他协变量不变的情况下，年龄大于 65 岁患者死亡风险率为年龄小于 65 岁患者的 1.72 倍。

在多因素分析表中，种族对应的 p 值均大于 0.05，这说明该变量对于生存预后而言不是独立的预后因素。组织学分级、肿瘤大小、婚姻状况这些变量大致都通过了显著性检验。其中，对于年龄来说，HR 危险比为 2.02，表示在其他协变量不变的情况下，年龄大于 65 岁患者死亡风险率为年龄小于 65 岁患者的 2.04 倍。

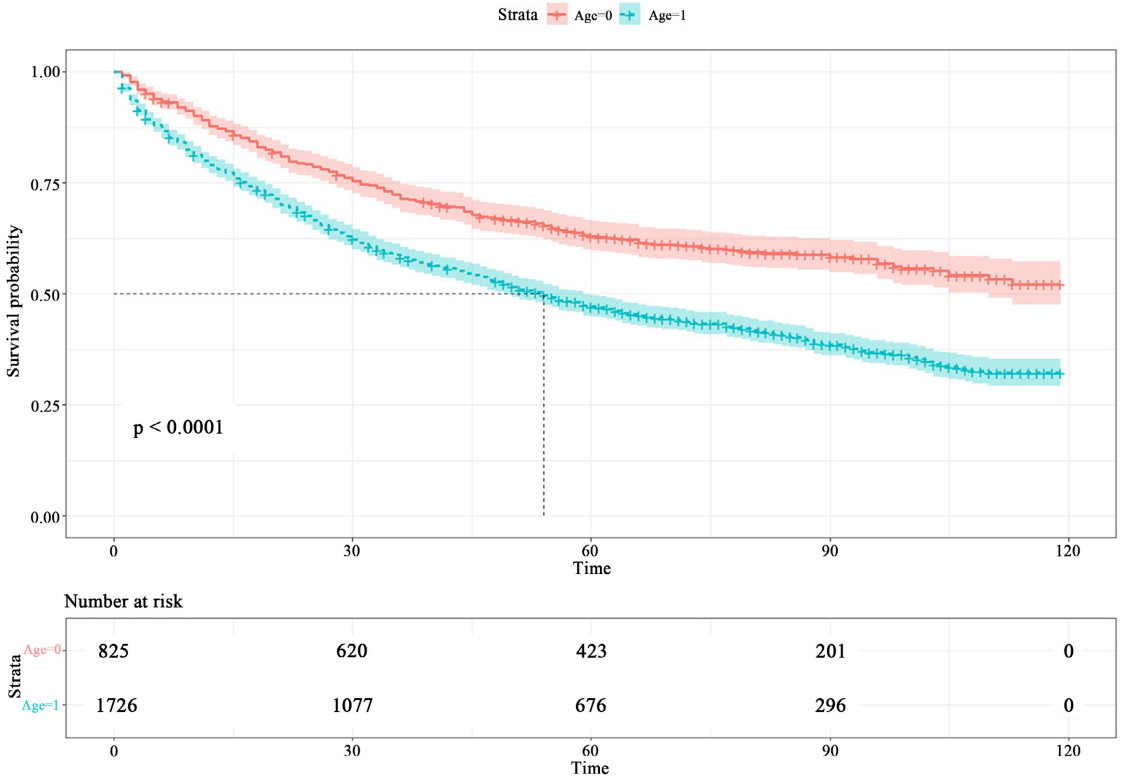

Figure 3. K-M survival curves for different Age
图 3. 不同年龄的 K-M 生存曲线

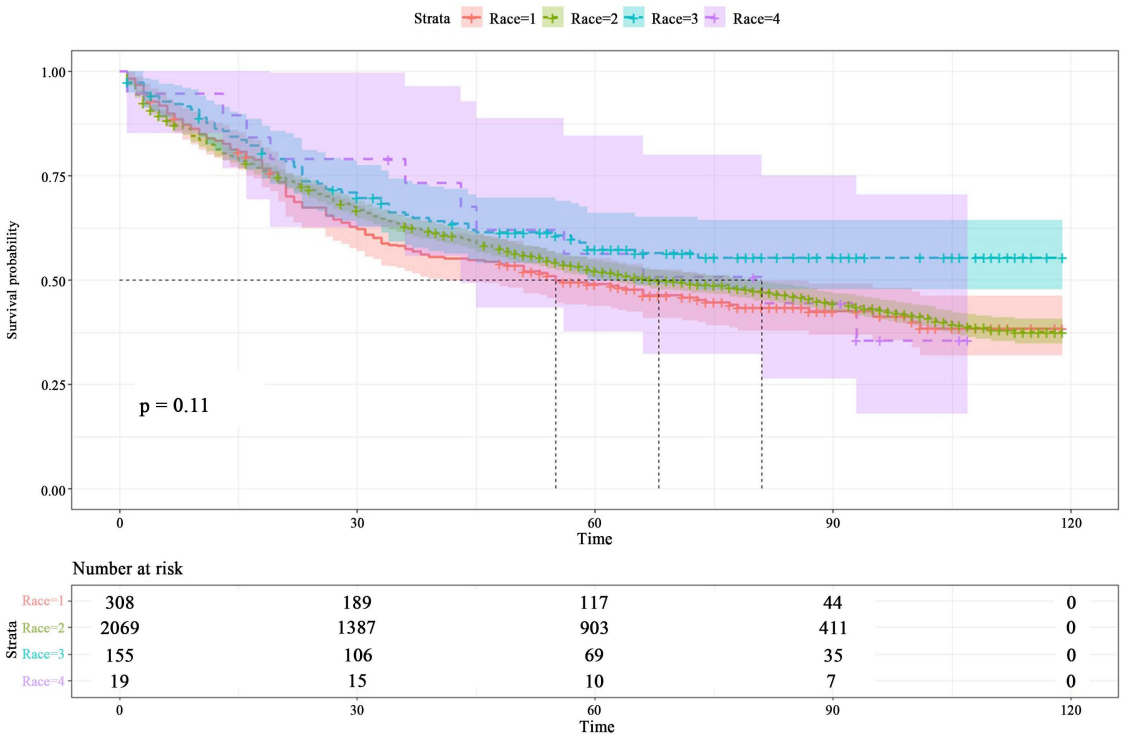

Figure 4. K-M survival curves for different Race
图 4. 不同种族的 K-M 生存曲线

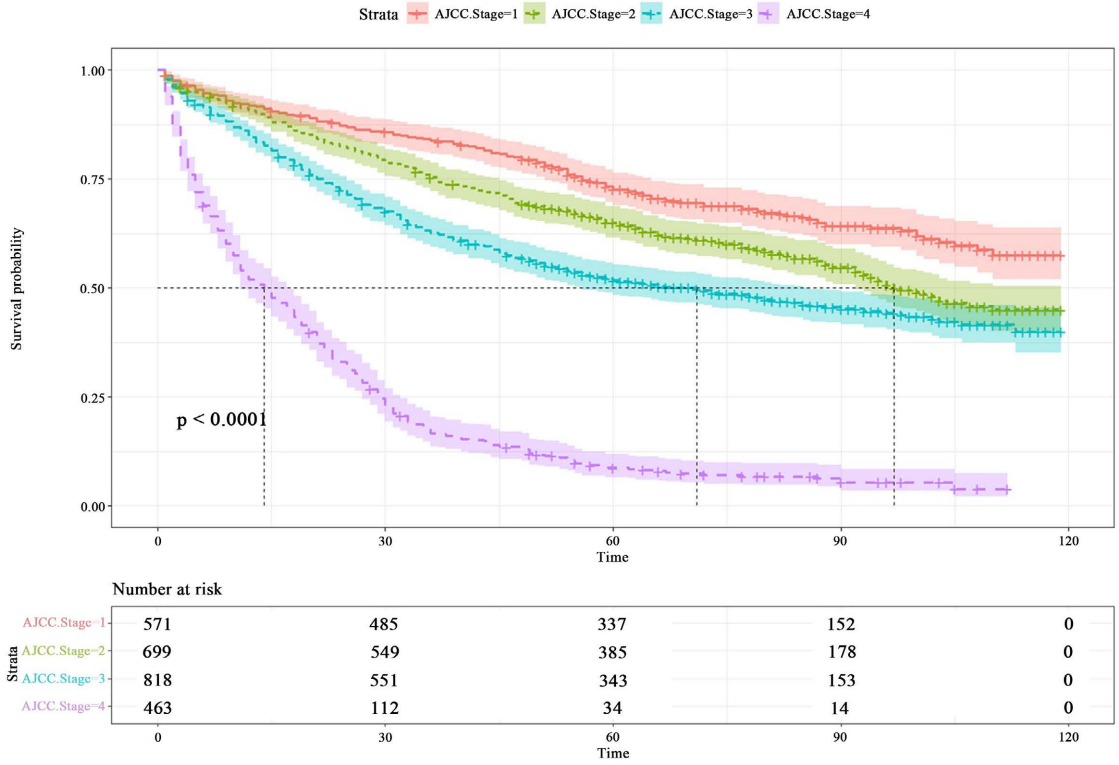

Figure 5. K-M survival curves for different AJCC stage
图 5. 不同 AJCC 分期的 K-M 生存曲线

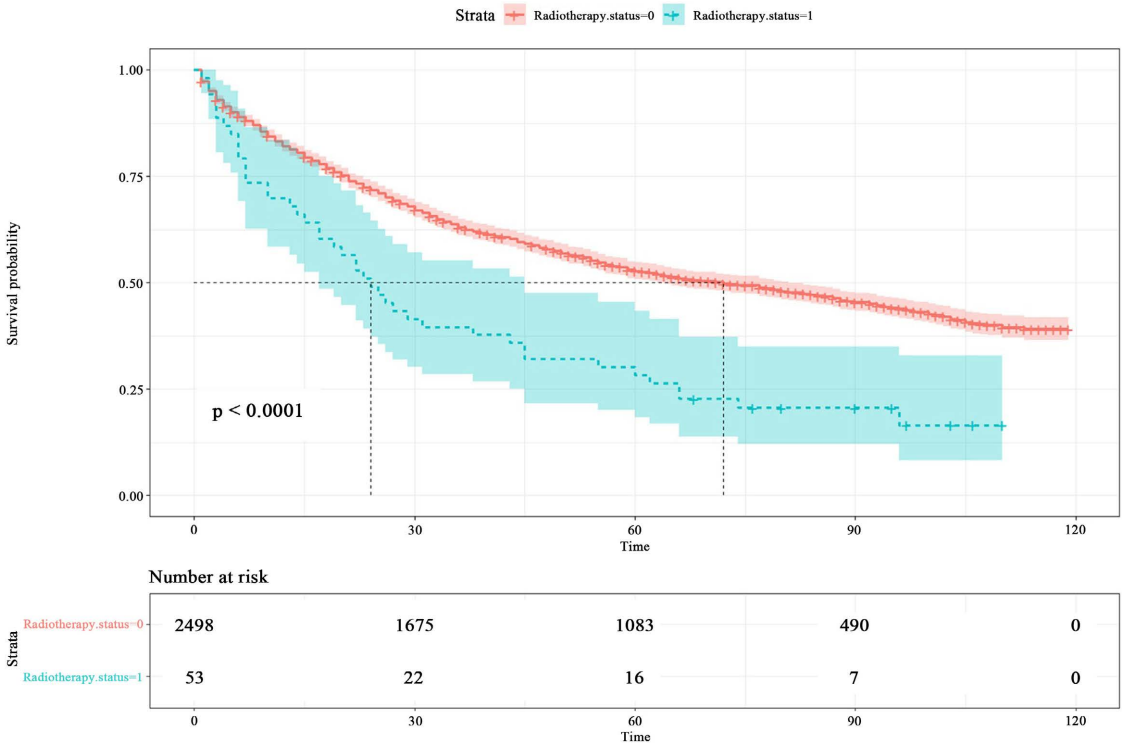

Figure 6. K-M survival curves for different Radiotherapy conditions
图 6. 不同放疗状况的 K-M 生存曲线

Table 3. Single-factor and multivariate analysis table for Cox proportional risk models
表 3. Cox 比例风险模型单因素与多因素分析表

	单因素分析		多因素分析	
	HR (95% CI)	p 值	HR (95% CI)	p 值
年龄				
≤64	参考		参考	
>65	1.72 (1.44~2.05)	<0.001	2.02 (1.67~2.45)	<0.001
种族				
黑种人	参考		参考	
白种人	1.07 (0.84~1.37)	0.596	1.00 (0.78~1.28)	0.992
亚洲或环太平洋岛民	0.86 (0.58~1.29)	0.469	0.80 (0.53~1.21)	0.293
其他	0.97 (0.42~2.24)	0.947	1.21 (0.52~2.79)	0.659
AJCC 分期				
I	参考		参考	
II	1.39 (1.07~1.8)	0.013	1.27 (0.97~1.65)	0.083
III	1.91 (1.49~2.46)	<0.001	2.62 (2.00~3.44)	<0.001
IV	7.14 (5.54~9.2)	<0.001	10.77 (8.00~14.49)	<0.001
化疗状况				
None	参考		参考	
Yes	1.21 (1.03~1.41)	0.019	0.50 (0.41~0.61)	<0.001
放疗状况				
None	参考		参考	
Yes	1.97 (1.21~3.18)	0.006	2.03 (1.23~3.33)	0.005
组织学分级				
I	参考		参考	
II	1.45 (1.06~1.98)	0.021	1.52 (1.11~2.10)	0.01
III	2.46 (1.76~3.44)	<0.001	2.33 (1.64~3.29)	<0.001
IV	2.01 (1.32~3.05)	0.001	1.35 (0.87~2.09)	0.174
婚姻状况				
未婚	参考		参考	
已婚	0.78 (0.65~0.94)	0.007	0.76 (0.63~0.91)	0.003
其他	1.24 (1.01~1.54)	0.044	1.14 (0.91~1.42)	0.243
肿瘤大小				
最大直径 ≤ 1 cm	参考		参考	
1 cm < 最大直径 ≤ 3 cm	0.26 (0.2~0.33)	<0.001	0.46 (0.35~0.62)	<0.001
3 cm < 最大直径 ≤ 5 cm	0.31 (0.08~1.26)	0.101	0.25 (0.06~1.04)	0.057
最大直径 > 5 cm	0.57 (0.29~1.11)	0.098	0.84 (0.42~1.65)	0.607

3.2.3. 预测模型的构建

1) 比例风险假定检验

在使用 Cox 比例风险回归模型时，其模型存在一个前提假设即时间与协变量之间不存在交互作用(PH

假定) [8]。本文主要采用时间依赖变量的 Cox 模型对前提假设条件进行检验。检验结果主要如表 4 所示，不难看出，所有变量和总体的 p 值均超过了显著性水平，这说明模型整体接受了原假设即认为时间与协变量之间不存在交互作用，满足了构建模型的前提。

Table 4. PH test results
表 4. PH 检验结果

变量	卡方值	自由度	p 值
年龄	3.4786	1	0.062
组织学分级	5.4346	3	0.143
肿瘤大小	4.5565	2	0.207
AJCC 分期	6.6713	3	0.083
化疗状况	4.1636	1	0.091
放疗状况	2.4859	1	0.115
婚姻状况	0.0598	2	0.971
GLOBAL	32.2471	13	0.071

2) Cox 比例风险回归模型的建立

通过逐步回归的方式，即满足 AIC 最小的准则，对纳入分析的变量进行筛选和确定，使得最终模型满足该准则。所筛选出来的变量及系数等最终结果可见表 5。

Table 5. Table of Cox proportional risk regression covariate coefficients
表 5. Cox 比例风险回归协变量系数表

变量	系数	z 统计量	p 值
年龄 > 65 (X_{12})	0.70741	7.19	<0.01
AJCC II (X_{22})	0.23703	1.75	0.08012
AJCC III (X_{23})	0.96264	6.975	<0.01
AJCC IV (X_{24})	2.37689	15.702	<0.01
化疗(X_{32})	−0.6895	−6.864	<0.01
放疗(X_{42})	0.70771	2.796	<0.01
组织学 II 级(X_{52})	0.41204	2.532	<0.01
组织学 III 级(X_{53})	0.82859	4.688	<0.01
组织学 IV 级(X_{54})	0.29918	1.346	0.1784
已婚(X_{62})	−0.28185	−3.032	<0.01
婚姻状况其他(X_{63})	0.13101	1.16	0.24599
1 cm < 肿瘤直径 ≤ 3 cm (X_{72})	−0.76409	−5.219	<0.01
肿瘤直径 > 5 cm (X_{73})	−0.22531	−0.652	0.51453

从表 5 中可以看出最终纳入的分析变量有：年龄、组织学分级、肿瘤大小、AJCC 分期、婚姻状况、放疗状况、化疗状况。最终确定的模型为：

$$h(t|X)=h_0(t)\exp\{0.71X_{12}+0.96X_{23}+2.37X_{24}-0.69X_{32}+0.70X_{42}$$
$$+0.41X_{52}+0.82X_{53}-0.28X_{62}-0.76X_{72}\}$$

其中，系数为正表示该因素为危险因素；系数为负表示保护因素。在该多因素比例风险回归模型中，化疗治疗、已婚、肿瘤直径大小在 1 cm 以上的为盲肠癌患者生存预后的保护因素，年龄大于 65 岁患者、放疗治疗、AJCC 分期大于 I、组织学等级高于一级、婚姻状况为其它的因素都为危险因素。根据查询，患者中选择放疗治疗的患者数量极少，只有病情特别严重的患者选择放疗治疗，但效果不佳。

3) 显著性检验

基于所构建的模型进行显著检验，具体检验结果可见表 6。三种情况下的检验 p 值远远小于 0.05，这说明，上述 Cox 多因素回归模型通过了显著性检验，模型总体效果不错并且具有合理性，同时这 7 个变量能较好的刻画出生存函数总的变化情况。

Table 6. Significance test table
表 6. 显著性检验表

检验方式	卡方值	自由度	p 值
Likelihood ratio test	496.1	13	p ≤ 2e−16
Wald test	549.6	13	p ≤ 2e−16
Score (logrank) test	646.7	13	p ≤ 2e−16

3.3. 竞争风险模型

3.3.1. Nelson-Aalen 累计风险曲线

与 Cox 比例风险模型不同的是，考虑到导致出现感兴趣终点事件的发生存在着竞争事件。故这里使用通过计算每个结局的累积发生率函数(CIF)绘制累积生存曲线，并使用 Fine-Gray's 检验来比较不同变量的组间之间的风险函数是否存在显著的差异。下文基于随机生存森林模型筛选出的 8 个变量绘制累积生存曲线，由于篇幅限制，本文给出前 4 个变量的累积生存曲线，图 7~10 分别代表的是：年龄、种族、AJCC 分期、化疗状况。图中的实线表示的是变量中不同类别盲肠癌患者死亡的发生率，而虚线表示的是在死于竞争事件发生率。例如：图 7 在年龄变量中，年龄大于 65 岁的患者的死于盲肠癌累计发生率要远高于竞争事件的死亡发生率，年龄小于 65 岁的患者死于盲肠癌发生率也远高于竞争事件的死亡发生率。

结果表明：在控制竞争事件后，年龄($p < 0.01$)、种族($p = 0.2$)、组织学分级($p < 0.01$)、肿瘤大小($p < 0.01$)、AJCC 分期($p < 0.01$)、婚姻状况($p < 0.01$)、化疗状况($p < 0.01$)、放疗状况($p < 0.01$)，其中除了种族没有通过显著性检验之外，其余变量都通过了显著性检验，同一种死亡原因的不同组之间死亡率存在差别。

3.3.2. 基于竞争风险模型的单因素与多因素分析

为了更加准确确定盲肠癌患者生存预后影响因素以及比较 Cox 比例风险模型确定的盲肠癌患者生存预后影响因素，本节考虑控制死亡原因为盲肠癌的竞争风险模型进行建模分析。单因素与多因素分析结果也基本一致。

表 7 为基于竞争风险模型的单因素与多因素分析表。在单因素分析表中，种族、婚姻状况对应的 p 值均大于 0.05，这说明这些变量对于生存预后而言不是独立的预后因素。年龄、AJCC 分期、放疗状况、化疗状况、组织学分级、肿瘤大小这些变量均通过了显著性检验。其中，对于年龄来说，HR 危险比为 1.17，表示明在其他协变量不变的情况下，年龄大于 65 岁患者死亡风险率为年龄小于 65 岁患者的 1.17 倍。

在多因素分析表中，种族、婚姻状况对应的 p 值均大于 0.05，这说明该变量对于生存预后而言不是独立的预后因素。年龄、AJCC 分期、放疗状况、化疗状况、组织学分级、肿瘤大小大致通过了显著性检验。其中，对于年龄来说，HR 危险比为 1.34，表示明在其他协变量不变的情况下，年龄大于 65 岁患者死亡风险率为年龄小于 65 岁患者的 1.34 倍。

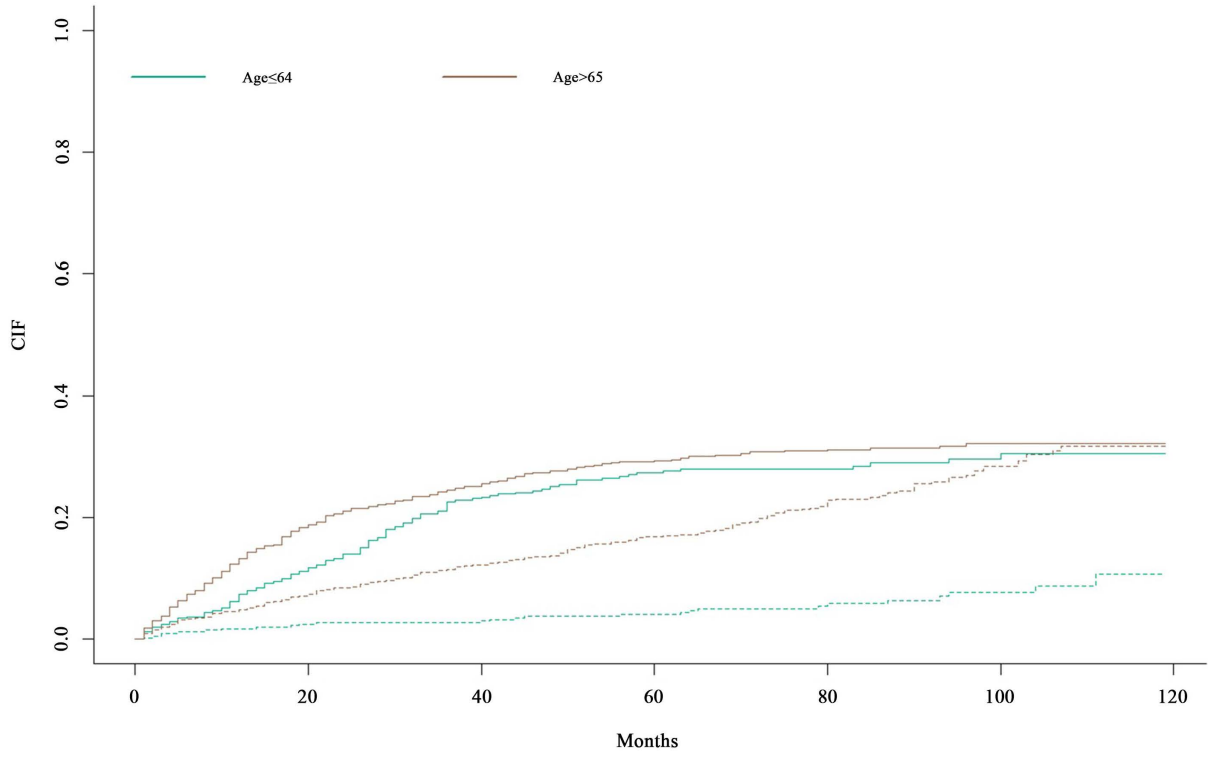

Figure 7. CIF cumulative survival curves at different Age
图 7. 不同年龄的 CIF 累积生存曲线

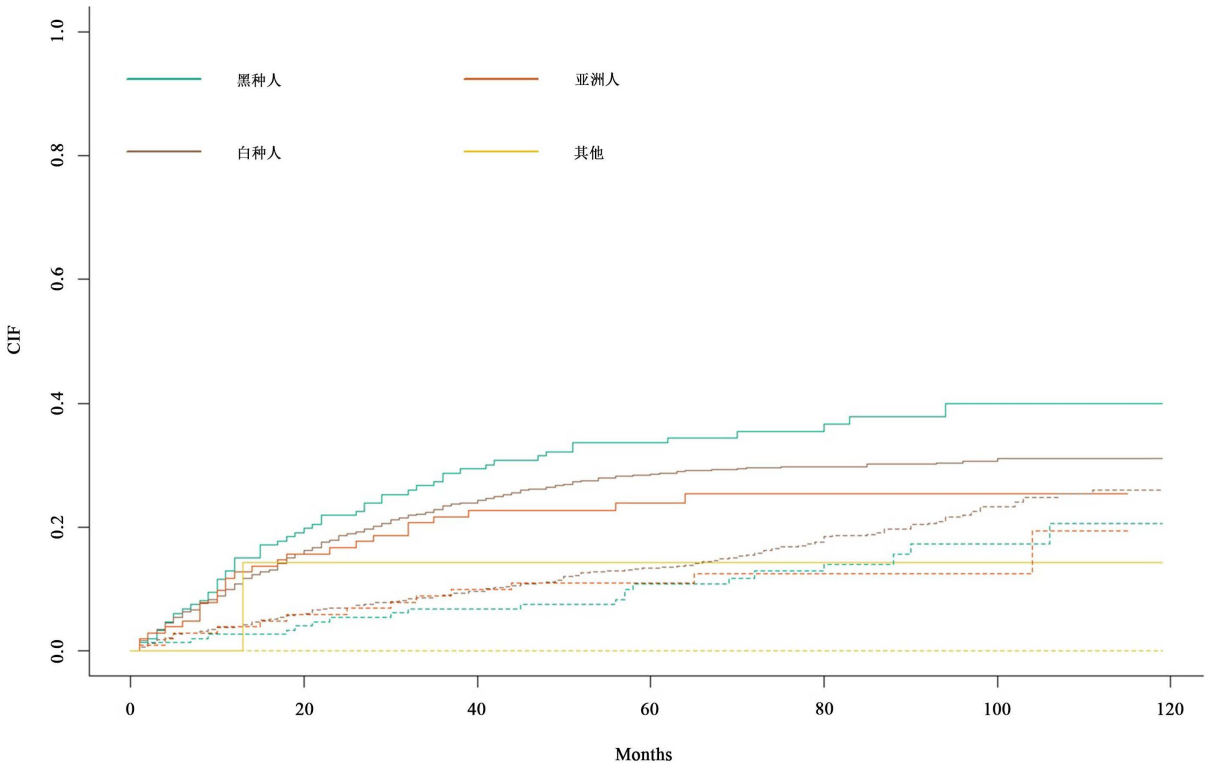

Figure 8. CIF cumulative survival curves at different Race
图 8. 不同种族的 CIF 累积生存曲线

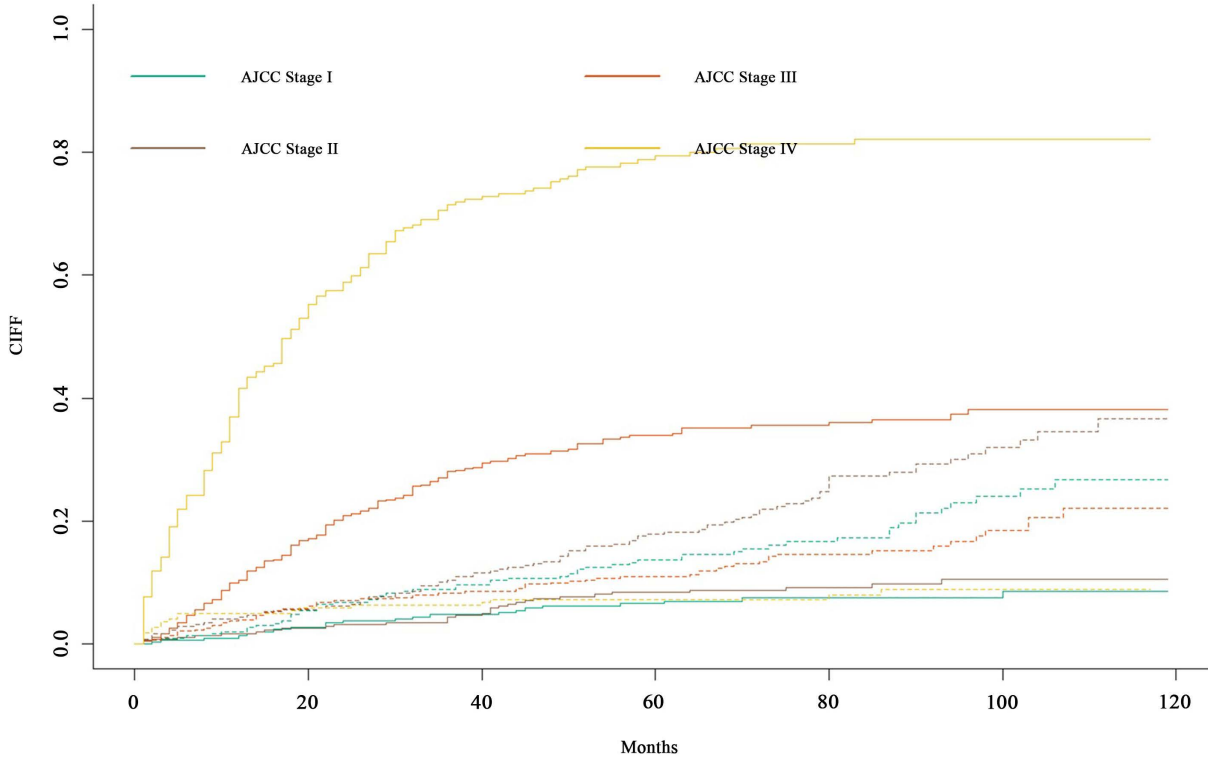

Figure 9. CIF cumulative survival curves at different AJCC stage
图 9. 不同 AJCC 分期的 CIF 累积生存曲线

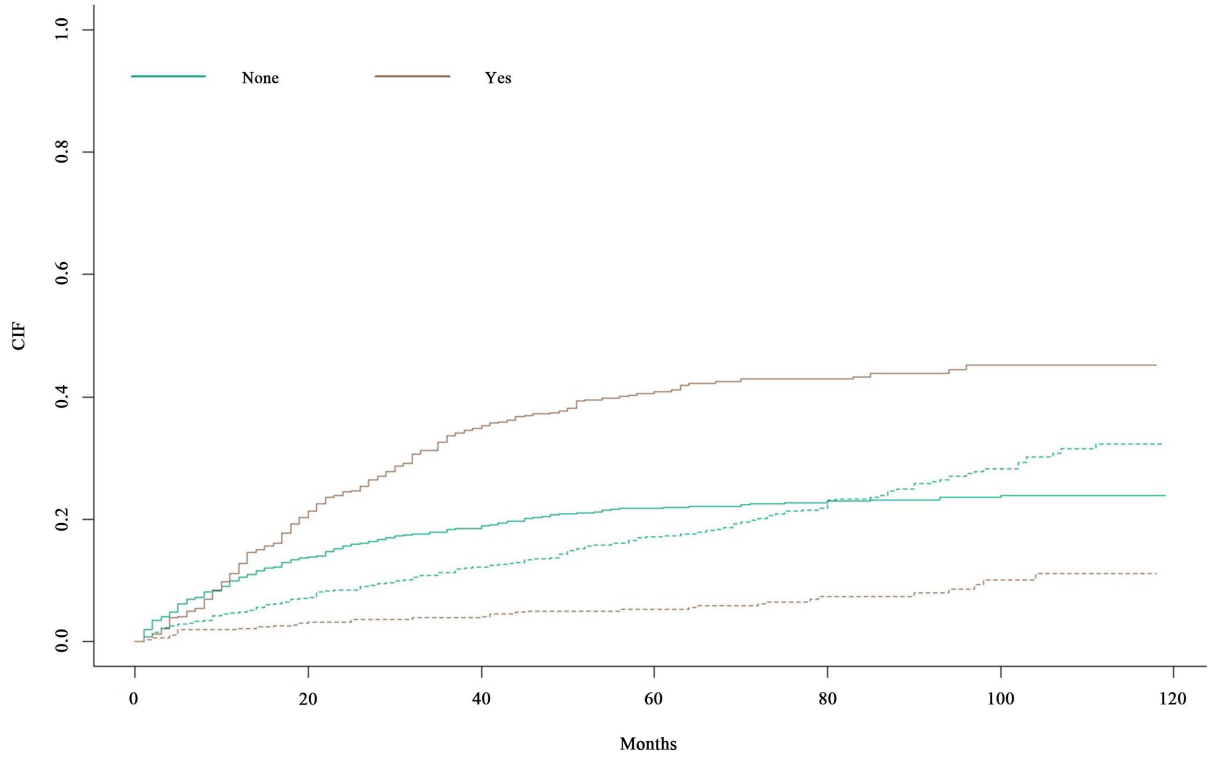

Figure10. CIF cumulative survival curves at different Chemotherapy conditions
图 10. 不同化疗状况的 CIF 累积生存曲线

Table 7. Competitive risk model univariate and multivariate analysis table
表 7. 竞争风险模型单因素与多因素分析表

	单因素分析		多因素分析	
	HR (95% CI)	p 值	HR (95% CI)	p 值
年龄				
≤64	参考		参考	
>65	1.17 (1.01~1.34)	0.043	1.34 (1.06~1.70)	0.016
种族				
黑种人	参考		参考	
白种人	0.76 (0.57~1.01)	0.061	0.83 (0.62~1.13)	0.239
亚洲或环太平洋岛民	0.64 (0.4~1.03)	0.069	0.65 (0.40~1.06)	0.084
其他	0.36 (0.05~2.57)	0.306	0.45 (0.06~3.29)	0.431
AJCC 分期				
I	参考		参考	
II	1.25 (0.72~2.15)	0.425	1.22 (0.71~2.12)	0.468
III	5.88 (3.76~9.2)	<0.001	7.03 (4.40~11.24)	<0.001
IV	25.54 (16.35~39.89)	<0.001	33.41 (20.61~54.16)	<0.001
化疗状况				
None	参考		参考	
Yes	2.23 (1.82~2.73)	<0.001	0.59 (0.47~0.76)	<0.001
放疗状况				
None	参考		参考	
Yes	2.9 (1.5~5.61)	0.002	3.67 (1.78~7.59)	<0.001
组织学分级				
I	参考		参考	
II	1.34 (0.84~2.15)	0.222	1.56 (0.96~2.55)	0.073
III	2.75 (1.7~4.46)	<0.001	2.39 (1.44~3.96)	0.001
IV	2.55 (1.41~4.61)	0.002	2.94 (1.58~5.45)	0.001
婚姻状况				
未婚	参考		参考	
已婚	0.86 (0.68~1.09)	0.203	0.97 (0.76~1.24)	0.808
其他	1.01 (0.76~1.35)	0.926	0.97 (0.71~1.33)	0.862
肿瘤大小				
最大直径 ≤ 1 cm	参考		参考	
1 cm < 最大直径 ≤ 3 cm	0.21 (0.15~0.27)	<0.001	0.42 (0.31~0.58)	<0.001
3 cm < 最大直径 ≤ 5 cm	0.46 (0.06~3.35)	0.446	0.20 (0.03~1.47)	0.114
最大直径 > 5 cm	0.61 (0.28~0.98)	0.033	0.17 (0.07~0.39)	<0.001

3.3.3. 竞争风险预测模型的构建

通过上述单因素与多因素分析结果，选取年龄、AJCC 分期、放疗状况、化疗状况、组织学分级、肿瘤大小这六个变量作为自变量，而因变量则是生存时间和生存结局(DSS)，区别于前面的 Cox 比例风险回归模型，这里的因变量发生了改变，考虑将竞争事件引入。

Table 8. Multi-factor competitive risk model coefficient table
表 8. 多因素竞争风险模型系数表

变量	系数	z 统计量	p 值
年龄 > 65 (X_{12})	0.2793	2.39	0.0172
组织学 II 级(X_{22})	0.4193	1.7	0.0388
组织学 III 级(X_{23})	0.825	3.27	0.0011
组织学 IV 级(X_{24})	1.0507	3.35	0.0008
1 cm < 最大直径 ≤ 3 cm (X_{32})	−0.818	−5.06	<0.0001
3 cm < 最大直径 ≤ 5 cm (X_{33})	−1.598	−1.57	0.1158
最大直径 > 5 cm (X_{34})	−1.755	−4.13	<0.0001
AJCC II (X_{42})	0.2017	0.72	0.0468
AJCC III (X_{43})	1.9514	8.2	<0.0001
AJCC IV (X_{44})	3.5391	14.46	<0.0001
化疗(X_{52})	−0.527	−4.49	<0.0001
放疗(X_{62})	1.2538	3.42	0.0006

基于逐步回归以及满足 AIC 最小的准则，构建了竞争风险模型，从表 8 中可以看出：除了肿瘤大小个别变量之外，其余变量均通过了显著性检验，说明对于生存预后而言是显著相关的。因此构建的竞争风险模型为：

$$h_c\left(t|X\right)=h_{0c}\left(t\right)\exp\{0.28X_{12}+0.42X_{22}+0.42X_{23}+1.05X_{24}-0.82X_{32}-1.76X_{34}$$
$$+0.2X_{42}+1.95X_{43}+3.54X_{44}-0.53X_{52}+1.25X_{62}\}$$

其中，系数为正表示该因素为危险因素；系数为负表示保护因素。在该多因素竞争风险模型中，化疗治疗、肿瘤直径大小在 1 cm 以上变量为盲肠癌患者生存预后的保护因素，年龄大于 65 岁患者、AJCC 分期大于 I、组织学等级高于 I 级、放疗治疗等都为危险因素。分析结果基本与 Cox 比例风险模型一致。

4. 模型预测与评估

4.1. 生存预测列线图

图 11 和图 12 是根据 Cox 比例风险模型和竞争风险模型构建的预测患者一年、三年、五年生存概率列线图。根据列线图[8]可以诊断病人的在一年、三年、五年时的生存概率。例如在 Cox 比例风险模型图 11 预测中：某盲肠癌患者年龄大于 65 岁则年龄得分记为 30；组织学评级为 III 级，得分为 35 分；肿瘤大小为小于 1 cm，得分为 40 分；AJCC 分期为 III，得分为 40 分；没有放疗治疗，得分为 0 分，已婚得分为 10 分；没有化疗治疗，得分为 30，则总分为 185 分，对应的一年生存概率为 0.45、三年生存概率为 0.1、五年的生存概率基本为 0。而在竞争风险模型图 12 中患者年龄大于 65 岁则年龄得分记为 8；组织学评级为 III 级，得分为 25 分；肿瘤大小为小于 1 cm，得分为 50 分；AJCC 分期为 III，得分为 55 分；没有放疗治疗，得分为 0 分；没有化疗治疗，得分为 15，则总分为 153 分，对应的一年生存概率为 0.7 左右，三年生存概率 0.2，五年的生存概率为 0.1。两种模型在生存概率预测上表现不同主要是因为竞争风险模型考虑的死亡原因除盲肠癌之外还有其它原因，所以竞争风险模型中表现的生存概率要比 Cox 比例风险模型生存概率高。

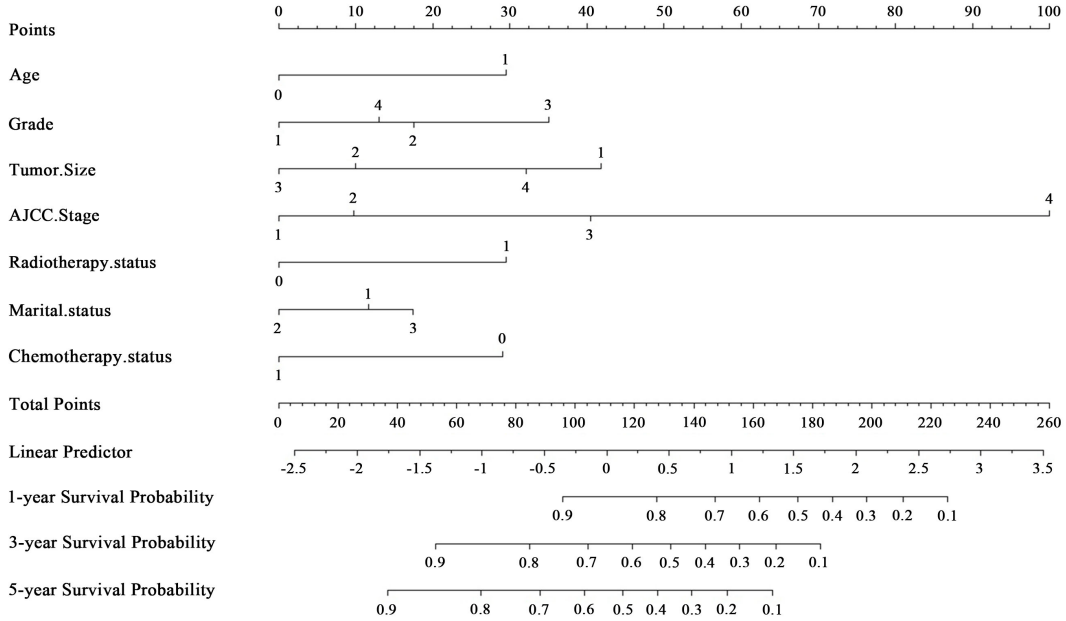

Figure 11. Cox proportional hazards model survival nomogram
图 11. Cox 比例风险模型生存列线图

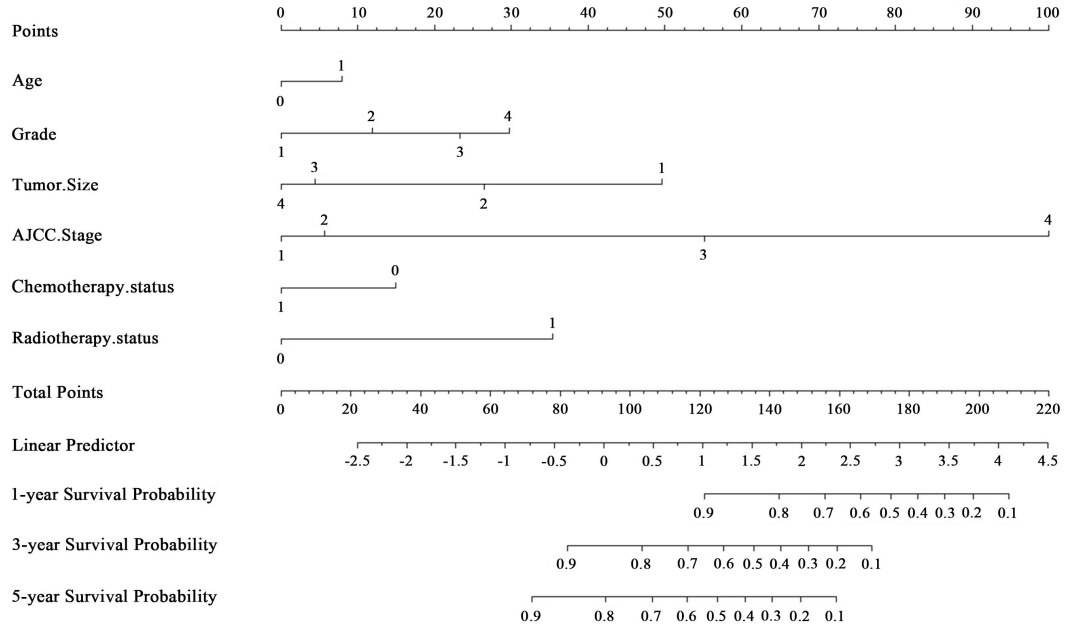

Figure 12. Competing risk model survival nomogram
图 12. 竞争风险模型生存列线图

4.2. 模型评估

图 13 和图 14 分别表示 Cox 比例风险模型和竞争风险模型在训练集上的 ROC 曲线[9]，在训练集中，Cox 比例风险模型预测的一年、三年、五年的 ROC 曲线对应的 AUC 值分别是 0.799、0.790 和 0.792，而竞争风险模型对应的 AUC 值为 0.848、0.841 和 0.830。图 15 和图 16 分别表示 Cox 比例风险模型和竞争风险模型在测试集上的 ROC 曲线，在测试集中，Cox 比例风险模型三个预测时间点对应的 AUC 值分别

为 0.809、0.779、0.752，竞争风险模型三个预测时间点对应的 AUC 值分别为 0.850、0.840 和 0.825。不论是在训练集还是测试集上，竞争风险模型比 Cox 比例风险模型效果都要好，预测准确率更高。

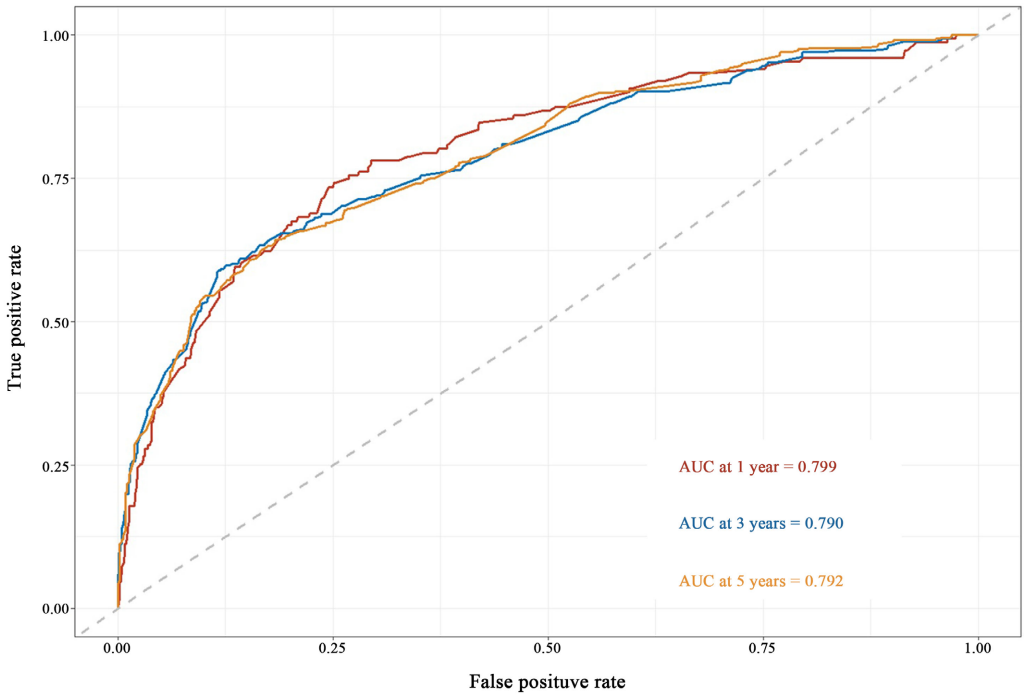

Figure13. Cox proportional hazards model training set ROC curve
图 13. Cox 比例风险模型训练集 ROC 曲线

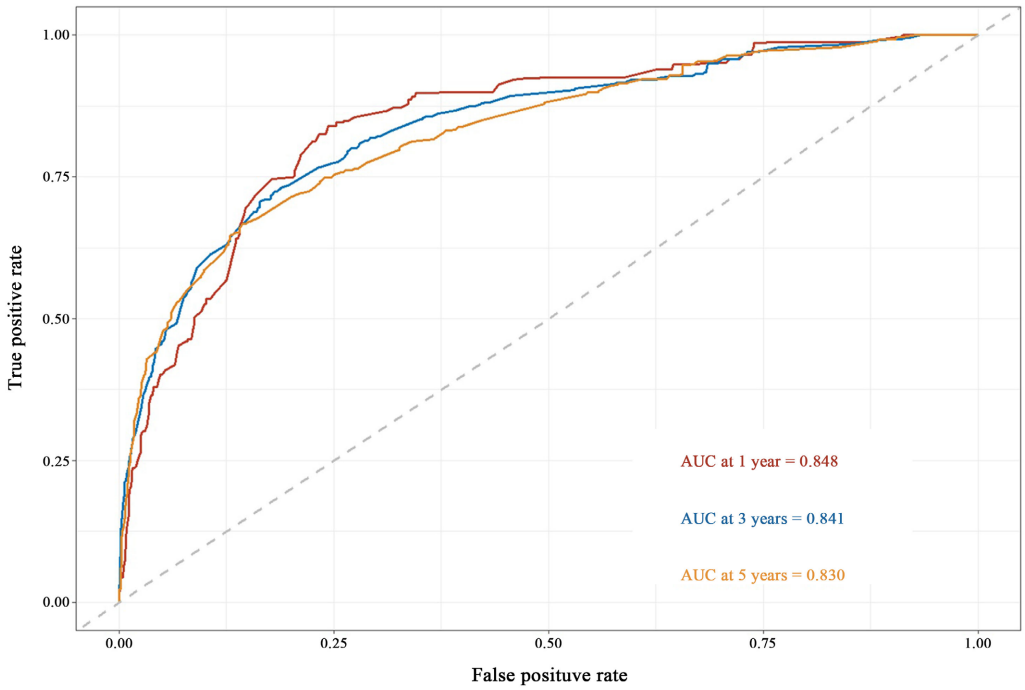

Figure14. Competitive risk model training set ROC curve
图 14. 竞争风险模型训练集 ROC 曲线

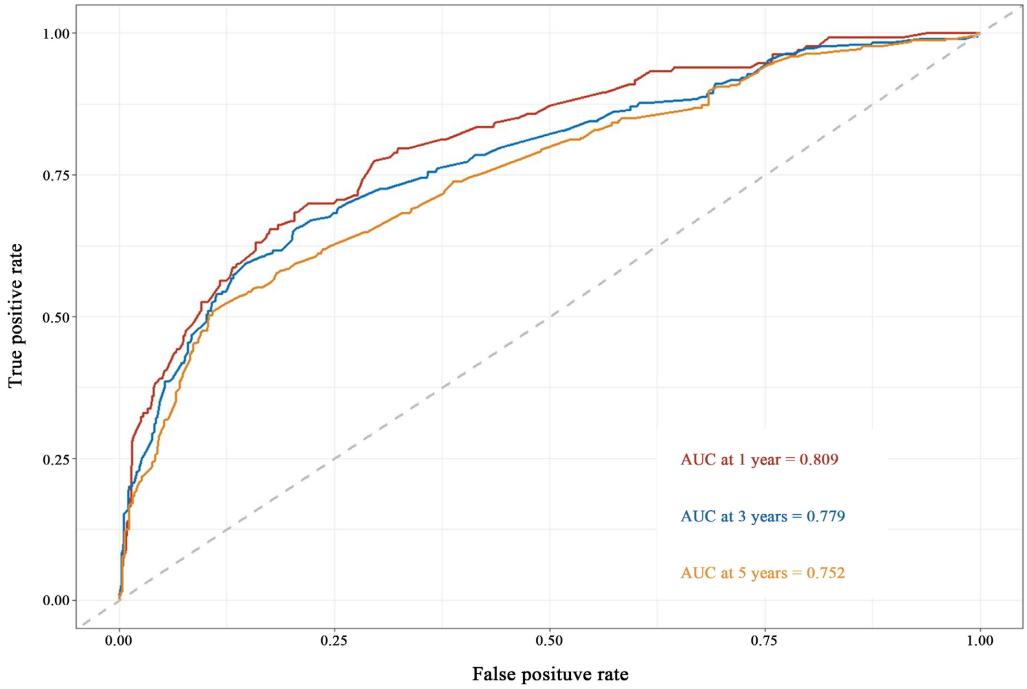

Figure 15. Cox proportional hazards model test set ROC curve
图 15. Cox 比例风险模型测试集 ROC 曲线

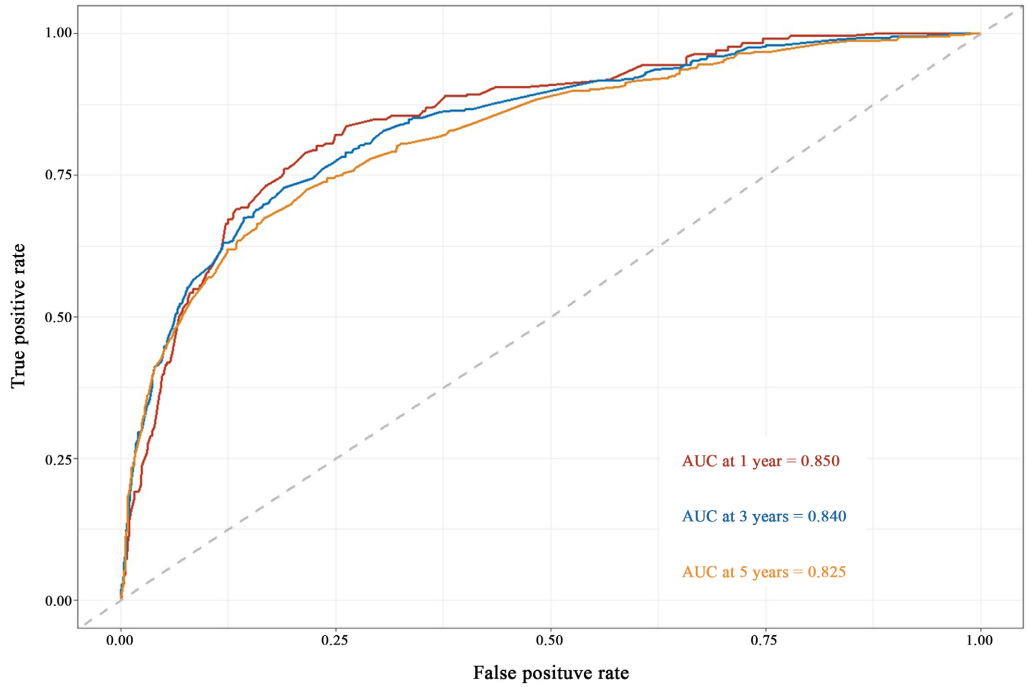

Figure 16. Competitive risk model test set ROC curve
图 16. 竞争风险模型测试集 ROC 曲线

5. 结论

通过建立随机生存森林模型、Cox 比例风险模型、竞争风险模型进一步研究对于盲肠癌患者生存预

后影响因素的探讨。结果表明：Cox 比例风险回归模型中，化疗治疗、已婚、肿瘤直径大小在 1 cm 以上的为盲肠癌患者生存预后的保护因素，年龄大于 65 岁患者、放疗治疗、AJCC 分期大于 I、组织学等级高于一级、婚姻状况为其它的因素为危险因素；在竞争风险模型中，化疗治疗、肿瘤直径大小在 1 cm 以上变量为盲肠癌患者生存预后的保护因素，年龄大于 65 岁患者、AJCC 分期大于 I、组织学等级高于 I 级、放疗治疗等都为危险因素。竞争风险模型剔除了婚姻状况这一变量，其余分析结果与 Cox 比例风险模型基本一致。

随机生存森林模型在训练集和测试集上的错误率为 0.271 和 0.278；Cox 比例风险模型三年预测结果在训练集和测试集上的 AUC 值为 0.790 和 0.779；竞争风险模型三年预测结果在训练集和测试集上的 AUC 值为 0.841 和 0.840。模型的比较中，竞争风险模型更胜一筹。在对于存在竞争事件的生存分析中，选择基于竞争风险构建的预测模型不仅准确度高，而且更具合理性。

参考文献

[1] Sung, H., Ferlay, J., Siegel, R., *et al.* (2021) Global Cancer Statistics 2020: Globocan Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA*: *A Cancer Journal for Clinicians*, **71**, 209-249. https://doi.org/10.3322/caac.21660

[2] Favoriti, P., Carbone, G. and Greco, M. (2016) Worldwide Burden of Colorectal Cancer: A Review. *Updates in Surgery*, **68**, 7-11. https://doi.org/10.1007/s13304-016-0359-y

[3] Magaji, B.A., Moy, F.M., Roslani, A.C., *et al.* (2017) Survival Rates and Predictors of Survival among Colorectal Cancer Patients in a Malaysian Tertiary Hospital. *BMC Cancer*, **17**, Article No. 339. https://doi.org/10.1186/s12885-017-3336-z

[4] Hermann, J., Karmelita-Katulska, K., Paszkowski, J., *et al.* (2011) Diagnosis of a Cecal Tumour with Virtual Colonoscopy. *Polish Journal of Radiology*, **76**, 25.

[5] Ishwaran, H., Kogalur, U.B., Blackstone, E.H., *et al.* (2008) Random Survival Forests. *The Annals of Applied Statistics*, **2**, 841-860. https://doi.org/10.1214/08-AOAS169

[6] 吴喜之. 应用回归及分类[M]. 北京: 中国人民大学出版社, 2016: 154-161.

[7] Liu, M., Yang, P., Mao, G., *et al.* (2019) Long Non-Coding RNA MALAT1 as a Valuable Biomarker for Prognosis in Osteosarcoma: A Systematic Review and Meta-Analysis. *International Journal of Surgery*, **72**, 206-213. https://doi.org/10.1016/j.ijsu.2019.11.004

[8] Varadhan, R., Weiss, C.O., Segal, J.B., *et al.* (2010) Evaluating Health Outcomes in the Presence of Competing Risks: A Review of Statistical Methods and Clinical Applications. *Medical Care*, **48**, S96-S105. https://doi.org/10.1097/MLR.0b013e3181d99107

[9] 陈卫中, 潘晓平, 宋兴勃, 等. ROC 曲线中最佳工作点的选择[J]. 中国卫生统计, 2006, 23(2): 157-158.