

基于SEER数据库的老年乳腺癌患者预后的影响因素研究

陈浪¹, 雷杰¹, 韩元泉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7月7日; 录用日期: 2022年7月18日; 发布日期: 2022年7月28日

摘要

在全球老龄化加剧的背景下, 我们基于生存分析的方法对老年乳腺癌患者预后的影响因素进行分析, 能够更好地确定风险的分层并制定合适的治疗方式, 这对预测患者的生存概率以及降低死亡风险有着现实的意义和应用价值。本文主要分析了来自美国癌症数据库SEER的数据。首先用K-M估计刻画了总体、年龄、性别、种族、肿瘤学等级、T分期、M分期、N分期、是否接受放疗、是否接受化疗的生存曲线。其次通过单因素Cox模型分析确定独立的预后因素, 进而通过逐步回归的方式构建多因素Cox比例风险回归模型, 最终确定的影响变量为: 年龄、M分期、是否接受放疗; 同样在考虑竞争事件的情况下, 首先应用Nelson-Aalen累计风险曲线(CIF)分析各个变量不同组间的风险函数的差异性, 其次建立单因素竞争风险模型确定影响生存时间和结局的独立预后变量, 确定的独立预后因素为: 年龄、M分期、是否接受放疗, 据此构建多因素竞争风险模型通过比较这两个预测模型的一致性指标(C-index)、校准曲线来判断模型预测的准确性, 结果显示竞争风险模型相比较好, 最后通过构建列线图的方式可视化生存概率。

关键词

Cox比例风险回归模型, 竞争风险模型, SEER数据库, Kaplan-Meier估计

A Study of the Influencing Factors of Prognosis in Elderly Breast Cancer Patients Based on the SEER Database

Lang Chen¹, Jie Lei¹, Yuanquan Han¹, Meng Wang²

¹School of Science, Chongqing University of Technology, Chongqing

²Chongzhou Nankai (Rongqiao) Middle School, Chongqing

Received: Jul. 7th, 2022; accepted: Jul. 18th, 2022; published: Jul. 28th, 2022

文章引用: 陈浪, 雷杰, 韩元泉, 王猛. 基于 SEER 数据库的老年乳腺癌患者预后的影响因素研究[J]. 统计学与应用, 2022, 11(4): 703-721. DOI: 10.12677/sa.2022.114075

Abstract

In the context of increasing global aging, our survival-based method analyzes the influencing factors of prognosis of elderly breast cancer patients, which can better determine the risk stratification and formulate appropriate treatment methods, which has practical significance and application value for predicting the survival probability of patients and reducing the risk of death. This paper mainly analyzes the data from SEER, an American cancer database. First, the survival curves of population, age, sex, race, oncology grade, T stage, M stage, N stage, whether to receive radiotherapy and whether to receive chemotherapy were described by K-M estimation. Secondly, independent prognostic factors were determined by single factor Cox model analysis, and then a multi factor Cox proportional hazards regression model was constructed by stepwise regression. The final influencing variables were: Age, M stage, whether to receive radiotherapy; Similarly, in the case of competitive events, firstly, Nelson-Aalen cumulative risk curve (CIF) is used to analyze the difference of risk function between different groups of each variable. Secondly, a single factor competitive risk model is established to determine the independent prognostic variables that affect survival time and outcome. The independent prognostic factors determined are: Age, M stage, whether to receive radiotherapy or not, based on this, a multi factor competitive risk model is constructed. The accuracy of the model prediction is judged by comparing the consistency index (C-index) and calibration curve of the two prediction models. The results show that the competitive risk model is relatively good. Finally, the survival probability is visualized by constructing a nomogram.

Keywords

Cox Proportional Hazards Regression Model, Competitive Risk Model, SEER Database, Kaplan-Meier Estimation

Copyright © 2022 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

近年来乳腺癌(Breast Cancer, BC)已经成为了威胁人类健康的常见癌症,它的发病率呈逐年上升趋势,每年有30余万女性被诊断出乳腺癌[1]。在东部沿海地区及经济发达的大城市,乳腺癌发病率上升尤其明显。从发病年龄来看,我国乳腺癌发病率从20岁以后开始逐渐上升。针对年轻和中年乳腺癌患者的研究较多,但是对老年乳腺癌患者的研究比较缺乏。相比年轻和中年患者,老年乳腺癌患者的预后更差[2]。

为了提高患者的生存质量,明确BC老年患者的特点,以及影响预后的因素,本文选取来自于根据美国国立癌症研究监测、流行病学和最终结果(Surveillance, Epidemiology, and End Results, SEER)数据库资料中的478例BC老年患者进行生存分析研究,对患者预后的预测研究分别使用Kaplan-Meier估计、Cox比例风险回归模型以及竞争风险模型。乳腺癌是一种严重危害人类健康的癌症,它也是全球最常见的癌症之一,同时许多因素都会影响乳腺癌患者的生存时间,因此,本文的研究目的是明确BC老年患者的特点和影响其生存时间的因素对于提高患者的生存质量有重要意义[3]。

2. 建模

2.1. 数据说明与预处理

本文纳入分析的自变量为9个,变量赋值情况如表1所示。这里由于本文研究的是老年BC患者,

因此患者年龄较大, 较为集中在 74 岁以下和 74 岁以上这两个范围, 故这里对于年龄这个连续变量选取 74 岁作为截断点, 由此转变为 0-1 型因子型变量。其余剩余变量的赋值详见表 1, 所有的 9 个自变量均为因子型变量。本文的因变量为生存时间和生存状态。

Table 1. Variable assignment table

表 1. 变量赋值表

变量	变量名	英文名	变量赋值
X_1	年龄	Age	$\leq 74-0$; $>74-1$
X_2	性别	Sex	Male-0; Female-1
X_3	种族	Race	Black-0; White-1; Others-2
X_4	肿瘤学分级	Grade	I-0; II-1; III-2
X_5	T 分期	AJCC.T	T1-0; T2-1; T3-2; T4-3; TX-4
X_6	M 分期	AJCC.M	M0-0; M1-1
X_7	N 分期	AJCC.N	N0-0; N1-1; N2-2; N3-3; NX-4
X_8	放疗	Radiation	None-0; Yes-1
X_9	化疗	Chemotherapy	None-0; Yes-1

在对数据的整理过程中, 肿瘤组织学分级这一类别中还存在着未知的信息, 共计 55 例, 种族这一类别中未知人种共计 5 例, 比例较小, AJCC.N 的 N2, N3, NX 病例共六例, 占比过少, 考虑将这三类别的信息予以剔除, 方便后续数据的分析工作, 剔除后的数据总计 412 例。

同时, 将本文的数据随机按照 5:5 的比例分成训练集和测试集, 对训练集的数据进行相关模型的建立, 即 Cox 比例风险回归模型、竞争风险模型, 基于构建的模型于测试集中进行进一步的验证, 分析预测的准确性和拟合效果。

Table 2. Comparison between training set and test set data

表 2. 训练集与测试集数据的比较说明

	全集(n = 412)	训练集(n = 206)	测试集(n = 206)	P 值
Age, n (%)				0.482
≤ 74	242 (59)	123 (60)	119 (58)	
>74	170 (41)	83 (40)	87 (42)	
Sex, n (%)				1
Male	17 (4)	6 (3)	11 (5)	
Female	395 (96)	200 (97)	195 (95)	
Race, n (%)				0.327
Black	48 (12)	24 (12)	24 (12)	
White	317 (77)	159 (77)	158 (77)	
Others	47 (11)	23 (11)	24 (11)	
Grade, n (%)				0.502

Continued				
I	210 (51)	104 (50)	106 (51)	
II	167 (40)	90 (44)	77 (37)	
III	35 (9)	12 (6)	23 (11)	
AJCC.T, n (%)				0.722
T1	298 (72)	151 (73)	147 (72)	
T2	82 (20)	44 (22)	38 (18)	
T3	19 (5)	8 (4)	11 (5)	
T4	5 (1)	1 (1)	4 (2)	
TX	8 (2)	2 (1)	6 (3)	
AJCC.M, n (%)				1
M0	409 (99)	205 (99)	204 (99)	
M1	3 (1)	1 (1)	2 (1)	
AJCC.N, n (%)				1
N0	376 (91)	193 (94)	183 (89)	
N1	36 (9)	13 (6)	23 (11)	
Radiation				0.393
None	243 (59)	115 (56)	128 (62)	
Yes	169 (41)	91 (44)	78 (38)	
Chemotherapy				0.6361
None	383 (93)	196 (95)	187 (91)	
Yes	29 (7)	10 (5)	19 (9)	

从表 2 中可以看出：不同年龄段在总体的占比分别为 59%和 41%，较为均衡；在性别这个变量中，女性占比达到 96%，比较符合实际；在种族这个变量中，白色人种占比最高，高达 77%；在肿瘤学分级中，主要是级别 I 和级别 II 为主；在 T 分期中，以 T1 和 T2 为主；在 M 分期中，绝大多数为 M0；在 N 分期中，以 N0 为主；治疗方式中，选择放疗的有 40%的患者；选择化疗的仅有 7%的患者。

为了验证随机分配的训练集和验证集，是否具有合理性且能够一定程度地代表全集的总体特征，由于本文的自变量均为分类变量，因此这里采用卡方检验来比较训练集和验证集的 9 个自变量是否具有显著的差异性，根据上述表格结果，可以看到显著性检验得到所有的变量对应的 P 值均大于 0.05，这能够说明两组之间没有显著的差异性，随机抽取的数据是合理的，能够代表所有数据的信息。

2.2. Cox 比例风险回归模型

2.2.1. Kaplan-Meier 生存曲线

Kaplan-Meier 方法是一种统计描述方法，充分利用信息，给出准确的统计量。K-M 方法作为一种非参数统计方法，不要求总体的分布，因此常常用来进行生存分析。本文将分别对总体和各个自变量进行 K-M 方法生存曲线分析。

从图 1 可以看出，对于高龄乳腺癌患者而言，总体的生存预后的时间相较其它癌症而言较长。生存曲线在 60 个月时，生存概率下降到 80%左右，而在 120 个月时，生存概率下降到 60%左右。这表明，随

着年龄的增长，尤其是高龄患者，乳腺癌的生存率下降速度比较稳定，生存曲线几乎为一条直线，但在预后时间为 120 个月时，生存概率也只有 60%，在人口老龄化加剧的今天，如果能够对老年乳腺癌患者的研究有正确的治疗，会大大增加患者的生存概率。

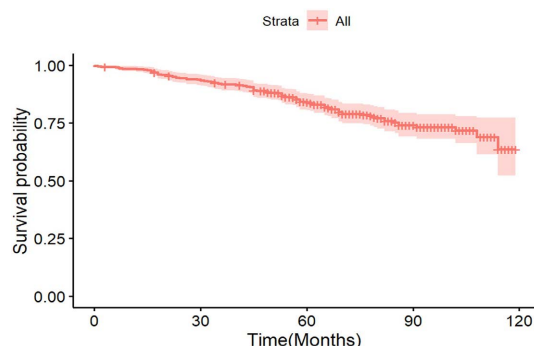


Figure 1. Overall K-M survival curve

图 1. 总体 K-M 生存曲线

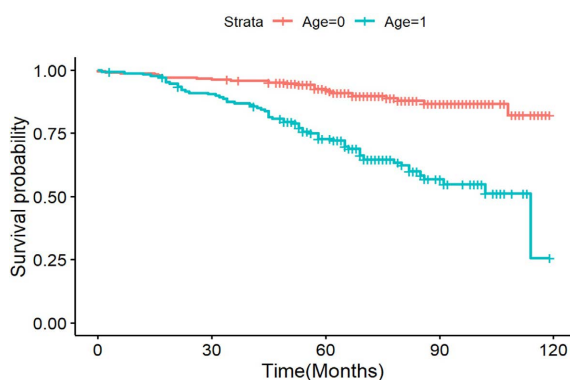


Figure 2. K-M survival curve at different ages

图 2. 不同年龄段的 K-M 生存曲线

对于不同年龄段的患者，本文是以 74 为阈值，从图 2 可以看出，年龄较小的一组的生存曲线远在年龄较大的一组的上方。这说明了，年龄越小，生存概率越大，且生存时间越长，区别越明显，到 120 个月时生存概率几乎相差三倍。

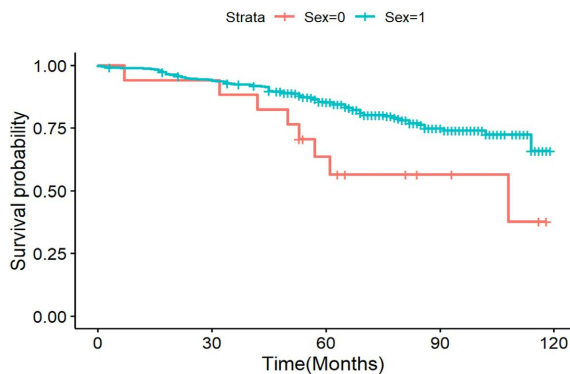


Figure 3. K-M survival curve of different sexes

图 3. 不同性别的 K-M 生存曲线

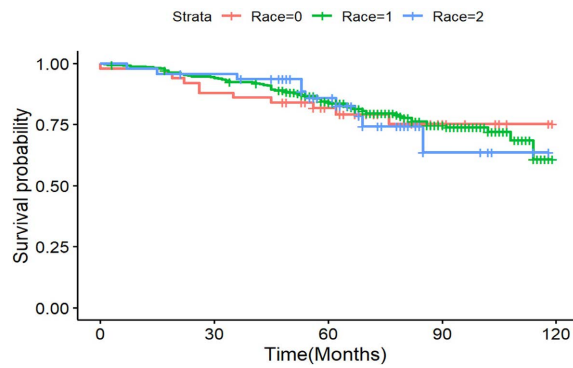


Figure 4. K-M survival curve of different races
图 4. 不同种族的 K-M 生存曲线

如图 3 所示，对于不同性别的患者，男性患者组的生存曲线在女性患者组的一组的下方。这表明男性的生存概率是低于女性的。在 30 个月前男性女性患者的生存概率差距不大，在 30 个月后差距变大，在 60 个月时差距达到 30%。

如图 4 所示，对于不同种族的患者，在 60 个月之前患者为非黑种人和非白种人的其它人种生存概率最高，白种人的生存概率次之，黑种人的生存概率最低，但总体相差都不大。当生存时间超过 60 个月时，其它人种的生存概率开始降低，成为生存概率最低的种族，黑种人的生存概率最高。在 120 个月时，白种人与其它人种的生存概率差不多，黑种人的生存概率相比之下高 10%左右。

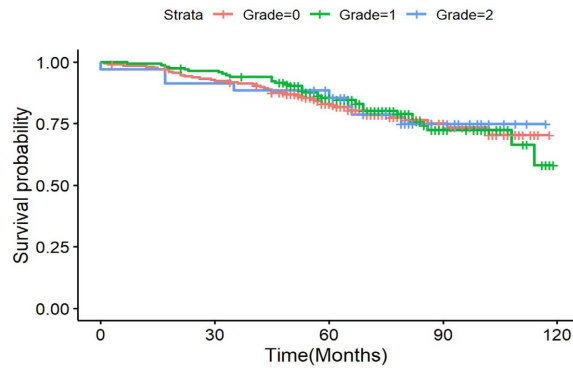


Figure 5. K-M survival curve of different oncology grades
图 5. 不同肿瘤学分级的 K-M 生存曲线

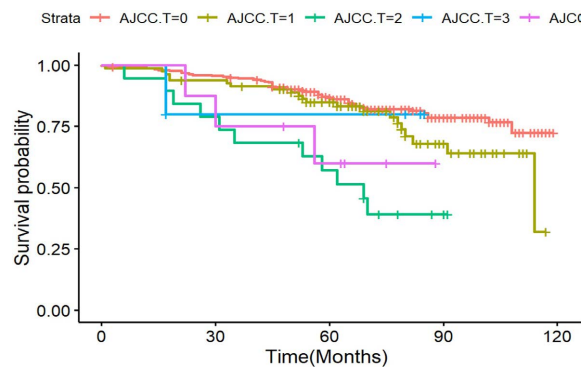


Figure 6. K-M survival curve at different T stages
图 6. 不同 T 分期的 K-M 生存曲线

如图 5 所示,对于不同肿瘤学等级的患者,在 60 个月之前,肿瘤学等级为 2 的患者的生存概率最高;在 60 个月到 90 个月之间,三个肿瘤学等级的患者生存概率相差不大;在 90 个月之后,肿瘤学等级为 2 的患者生存概率最低,其它两个等级的患者相差不大。

如图 6 所示,对于不同 T 分期的患者,总体而言,T1 分期的患者的生存概率最高,T3 分期的患者最低,但在 90 个月左右,数据中的 T3, T4, TX 分期的患者就已经死亡了,所以 90 个月后有 T1 分期和 T2 分期的患者的生存曲线。

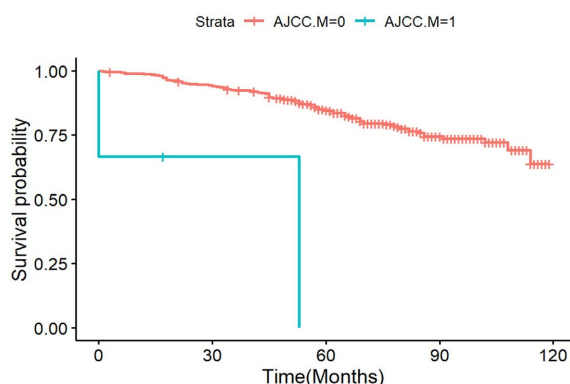


Figure 7. K-M survival curve of different M stages

图 7. 不同 M 分期的 K-M 生存曲线

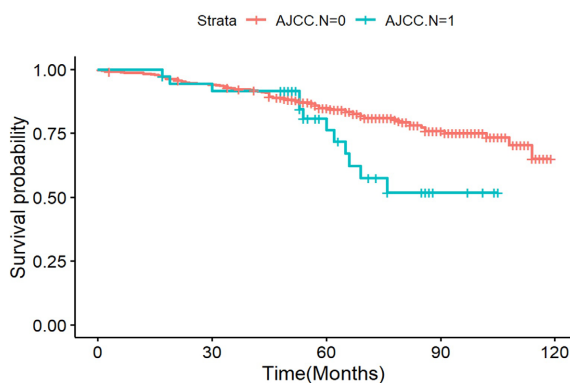


Figure 8. K-M survival curve at different N stages

图 8. 不同 N 分期的 K-M 生存曲线

如图 7 所示,对于不同 M 分期的患者,由于数据中 M 分期为 M1 的患者较少,则在生存时间为 50 个月左右就已经全部死亡。总体上 M 分期为 M0 的患者的生存概率高于 M 分期为 M1 的患者的生存概率。

如图 8 所示,对于不同 N 分期的患者,生存时间在 50 个月之前 N 分期为 N0 的患者和 N1 的患者的生存概率几乎相同,生存时间在 50 个月之后,N 分期为 N0 的患者的生存概率远高于 N 分期为 N1 的患者的生存概率。

对于是否接收放疗的患者,图 9 可以明显看出接收放疗的患者的生存曲线在未接收放疗的患者的生存曲线上方。这也就表明接收放疗的患者的生存概率是要明显高于未接收放疗的患者的。

对于是否接收化疗的患者,图 10 可以看到接受了化疗的患者的生存曲线基本上也是在未接受化疗的患者的上方,生存时间在 60 个月之前,接受化疗的患者生存概率还是有所下降,但在 60 个月后,接受化疗的患者的生存概率保持在 75% 稳定;未接受化疗的患者的生存概率基本上保持稳定的速率下降。

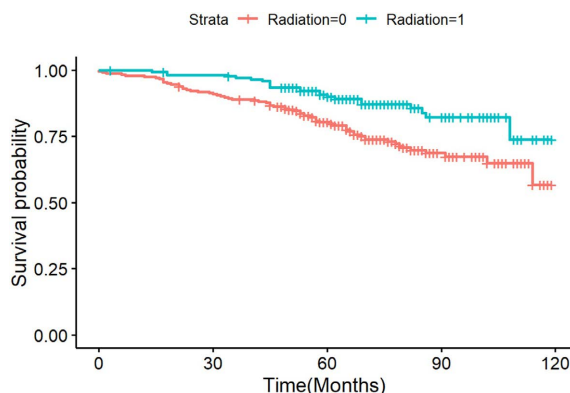


Figure 9. K-M survival curve with or without radiotherapy
图 9. 是否接收放疗的 K-M 生存曲线

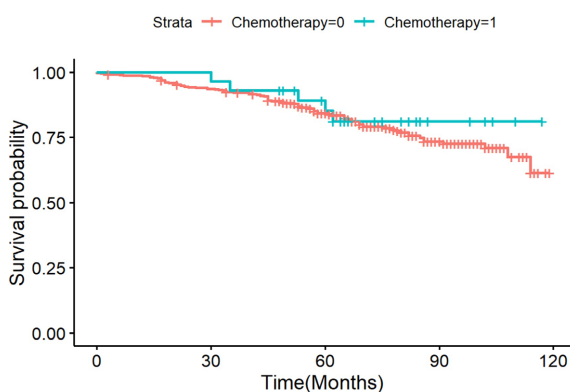


Figure 10. K-M survival curve with or without chemotherapy
图 10. 是否接受化疗的 K-M 生存曲线

2.2.2. 单因素分析与多因素分析

分析针对的是训练集的 206 例病例进行单因素和多因素分析，分析的自变量为：年龄、性别、肿瘤学分级、T 分期、M 分期、N 分期、是否放疗、是否化疗，因变量为生存时间和生存状态。

Table 3. Single factor analysis and multi factor analysis table

表 3. 单因素分析与多因素分析表

	单因素分析		多因素分析	
	HR (95% CI)	P 值	HR (95% CI)	P 值
Age				
≤74	参考			
>74	4.55 (2.30~9.02)	<0.01	6.1 (2.6~14)	<0.01
Sex				
Male	参考			
Female	0.55 (0.17~1.82)	0.33	0.43 (0.099~1.9)	0.26
Race				
Black	参考			

Continued

White	1.31 (0.47~3.71)	0.61	1 (0.3~3.6)	0.95
Others	2.04 (0.57~7.29)	0.27	2.7 (0.61~12)	0.19
Grade				
I	参考			
II	1.32 (0.45~3.90)	0.62	2.1 (0.27~17)	0.18
III	1.33 (0.46~3.90)	0.60	3 (0.39~23)	0.30
AJCC.T				
T1	参考			
T2	1.48 (0.69~3.18)	0.31	1.1 (0.47~2.7)	0.78
T3	4.42 (1.91~10.24)	<0.01	2.5 (0.63~9.8)	0.20
T4	0.08 (0.02~0.16)	0.87	0.03 (0.01~0.09)	0.97
TX	3.20 (0.76~13.17)	0.11	8.1 (1.4~46)	0.02
AJCC.M				
M0	参考			
M1	9.4 (1.3-70)	0.02	7.6 (3.60~16)	<0.01
AJCC.N				
N0	参考			
N1	1.15 (0.27~4.79)	0.85	0.41 (0.02~8.8)	0.57
Radiation				
None	参考			
Yes	0.33 (0.15~0.73)	<0.01	0.43 (0.18~1)	0.04
Chemotherapy				
None	参考			
Yes	0.94 (0.22~3.90)	0.93	0.68 (0.03~15)	0.81

从表 3 可以看出, 单因素分析中: 年龄、M 分期、是否接受放疗通过了显著性检验。从年龄来看, 年龄越大, 死亡的风险越(HR = 4.55); 从 M 分期来看, 相较于 M 分期为 M0 的患者来看, M1 的患者的死亡风险更大(HR = 9.4); 从患者是否接受放疗来看, 接受了放疗的患者的死亡风险会小很多(HR = 0.33)。

从性别来看, 相较于男性而言, 女性的死亡风险较低, 仅为男性的 0.55 倍, 但没有通过显著性检验 ($P = 0.33 > 0.05$), 说明性别与死亡风险没有较强的关系; 从种族来看, 白种人和其他人种相较于黑种人而言, 死亡风险较高, 但是未通过显著性检验, 则种族与死亡风险没有较强的关系; 从肿瘤学分级来看, 等级 II 和等级 I 相较于等级 III 而言, 死亡风险较高, 但未通过显著性检验, 则肿瘤学分级与死亡风险也没有较强的关系; T 分期中, 只有 T3 分期通过了显著性检验, 说明与预后存在着显著的相关关系, 其 T 分期与死亡风险没有较强的关系; 对于 N 分期而言, 没有通过显著性检验, 说明 N 分期与死亡风险中没有较强关系。对于是否接受化疗而言, 没有通过显著性检验, 说明是否接受化疗与死亡风险没有较强的相关性。

多因素分析中, 仍然是年龄、M 分期、是否接受放疗通过了显著性检验。但与单因素分析不同的是, 由原先的 T3 分期显著, 变为 TX 分期显著。

2.2.3. 模型构建与验证

在建立 Cox 比例风险模型之前，我们需要进行 PH 假定检验。只有满足了时间与自变量之间不存在交互作用时，才能建立 Cox 比例风险模型。检验结果如表 4 所示，可以看到所有变量和总体(GLOBAL)的 P 值都超过了 0.05 的显著性水平，这表明模型整体接受了原假设即认为时间与自变量之间不存在交互作用，满足了 PH 假定。

Table 4. PH assumption test result table

表 4. PH 假定检验结果表

变量	卡方值	P 值
Age	0.13	0.72
Sex	0.32	0.57
Race	3.60	0.17
Grade	1.86	0.39
AJCC.T	2.18	0.70
AJCC.M	2.24	0.13
AJCC.N	0.03	0.86
Radiation	1.32	0.25
Chemotherapy	0.66	0.42
GLOBAL	12.98	0.53

通过逐步回归的方法，即用 AIC 准则对纳入分析的变量进行筛选，最后筛选出的变量和系数如下表所示。

Table 5. Cox proportional hazard regression independent variable coefficient table

表 5. Cox 比例风险回归自变量系数表

	系数	HR	Z 统计量	P 值
Age	1.77	5.85	4.12	<0.01
AJCC.M	4.12	61.51	3.67	<0.01
Radiation	-0.95	0.39	-2.23	0.03

从表 5 可以看出最终纳入分析的变量有：年龄、M 分期、是否接受放疗。最终确定的模型为：

$$h(t|X) = h_0(t) \exp(1.77X_1 + 4.12X_6 - 0.95X_8)$$

其中三个变量均通过了显著性检验，随着年龄的增加，死亡的风险增大，在同等条件下，年龄大的患者死亡风险是其 4.55 倍；随着 M 分期的期数增加，死亡的风险增大，在同等条件下，M 分期中 M1 的死亡风险是 M0 的 9.4 倍；接受放疗的患者组比未接受放疗的患者组的死亡风险小，在同等条件下，接受放疗的患者组的死亡风险是未接受放疗的患者组的 0.33 倍。综上所述，年龄和 M 分期为危险预后因素，即数值越大，生存时间越短；是否接受放疗为保护预后因素。

对上述模型进行显著性检验，具体检验结果如表 6 所示。三种检验的 P 值远远小于 0.05，这说明上述模型通过了显著性检验，模型合理，同时这三个变量能够很好的反应生存函数的变化情况。

Table 6. Cox model significance test table
表 6. Cox 模型显著性检验表

检验类型	卡方值	自由度	P值
Likelihood ratio test	34.95	3	<0.01
Wald test	28.25	3	<0.01
Score (logrank) test	39.89	3	<0.01

根据上述的 Cox 比例风险回归模型的构建,可以预测患者 12 个月,24 个月,36 个月预后的列线图。这里选择的 12 个月,24 个月,36 个月时由于研究的对象为老年患者,预后效果较差。

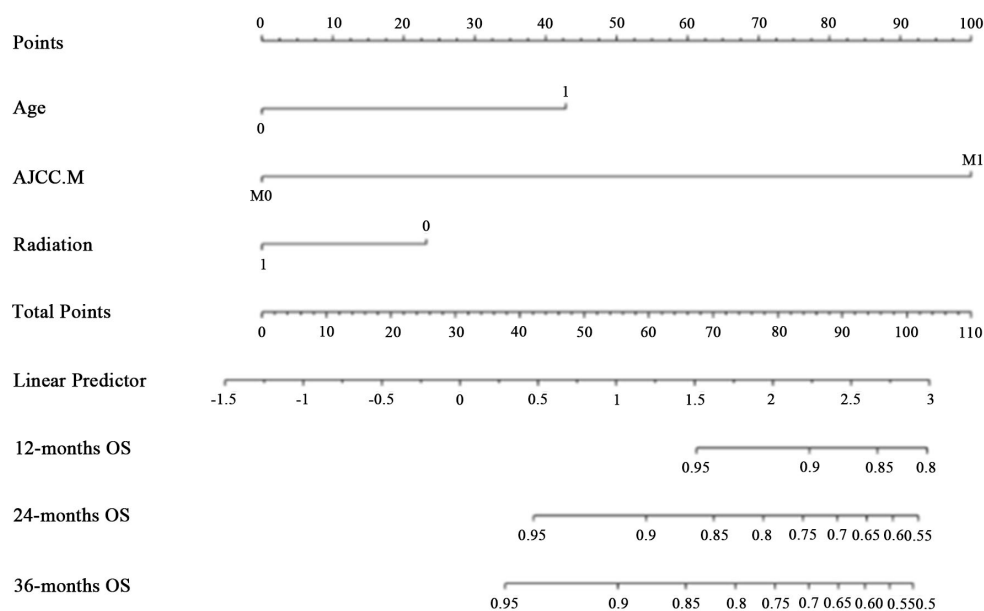


Figure 11. Nomogram for predicting patient survival
图 11. 患者生存率预测列线图

从图 11 可以直观地预测出具有不同临床特征的患者生存概率,具体表现在不同变量中的不同类别对应不同的刻度,而不同的刻度又对应不同的分值,根据这些具体分值预测出生存概率。例如:一位年龄超过 74 岁的男性乳腺癌患者接受了放疗,那么他的对应得分: $24 + 42 = 66$ 分,那么他的 12 个月后的生存概率为 0.95,24 个月后的生存概率为 0.88,36 个月后的生存概率为 0.83。

本文通过一致性指数和校准曲线图来验证模型的预测能力。

一致性指数(concordance index)又称为 C-index,它被用来评价模型预测能力。C-index 的计算公式如下:

$$C\text{-index} = \frac{\text{一致对子数}}{\text{有用对子数}}$$

根据 C-index 的计算公式,可以得知, C-index 在 0.5~1 之间,0.5 为完全随机,说明该模型没有预测作用,1 为完全一致,说明该模型预测结果与实际完全一致。在实际应用中,很难找到完全一致的预测模型,既往研究认为, C-index 在 0.50~0.70 为较低准确度;在 0.71~0.90 之间为中等准确度;而高于 0.90

则为高准确度，这与 ROC 曲线的 AUC 是类似的。根据上述的 Cox 比例风险回归模型得到的 $C\text{-index} = 0.769$ ，可以得知该模型预测较为准确，为中等准确度。

如图 12，描绘的是患者在 12 个月、24 个月和 36 个月的校准曲线图，通过将预测概率与实际概率作为比较进一步分析所构造的 Cox 比例风险回归模型预测的准确性。在时间为 12 个月时，图 12 中的红线分布在对角线左右，时间越长，偏差越大。综上所述，说明基于列线图的预测，短期内比较准确，但长期预测效果一般。

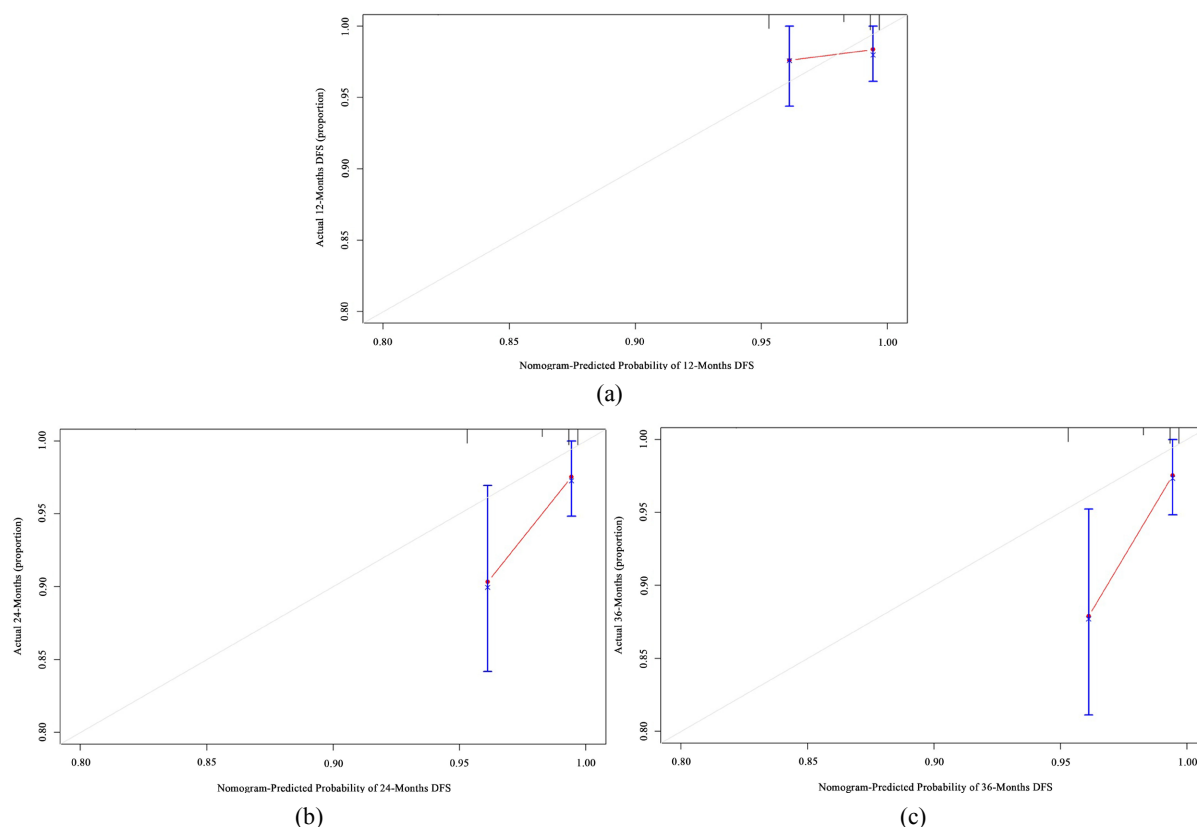


Figure 12. Calibration graph of patients in the Cox proportional risk model for 12, 24, 36 months

图 12. Cox 比例风险模型的患者 12, 24, 36 个月的校准曲线图

2.3. 竞争风险模型

2.3.1. Nelson-Aalen 累计风险曲线

与 Cox 比例风险回归分析不同的是，竞争风险模型考虑到了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。下文共展示了 9 幅累积生存曲线，分别代表的是：年龄、性别、种族、肿瘤学等级、T 分期、M 分期、N 分期、是否接受化疗、是否接受放疗，如图 13~21 所示，图中的实线表示的是变量中不同类别患者死亡的风险。

根据图 13~21 结果表明：年龄大一组的患者的累积死亡率高；男性患者组的累积死亡率高；在种族中 60 个月前三个种族的累积死亡率差不多，60 个月后其它人种的累计死亡率最高，60 到 80 个月中白种人累计死亡率最低，80 个月后黑种人累计死亡率最低；肿瘤学分级为 II 的患者的累积死亡率最高，肿瘤学分级为 III 的患者的累计死亡率最低；不同 T 分期中，累积死亡率最低的是 T4，累积死亡率最高的是

TX; 不同 M 分期中, M1 的累积死亡率较高; 不同 N 分期中, N1 的累积死亡率较高; 不接受放疗和化疗的患者的累积死亡率较高。

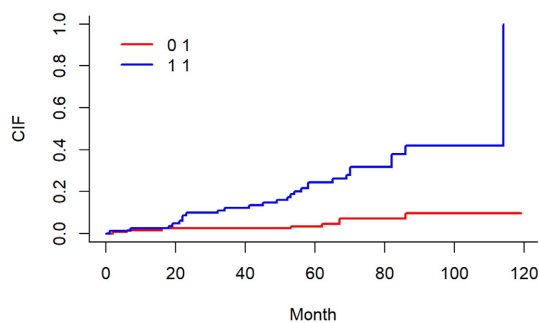


Figure 13. Cumulative risk map by different ages

图 13. 不同年龄的累积风险图

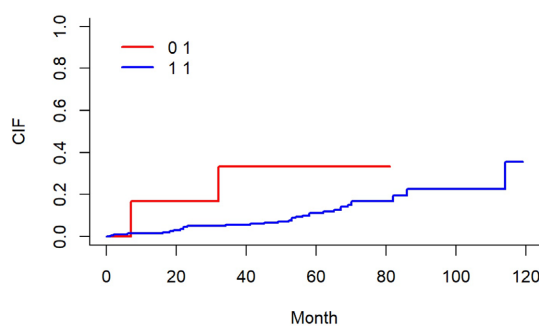


Figure 14. Cumulative risk map by different genders

图 14. 不同性别的累积风险图

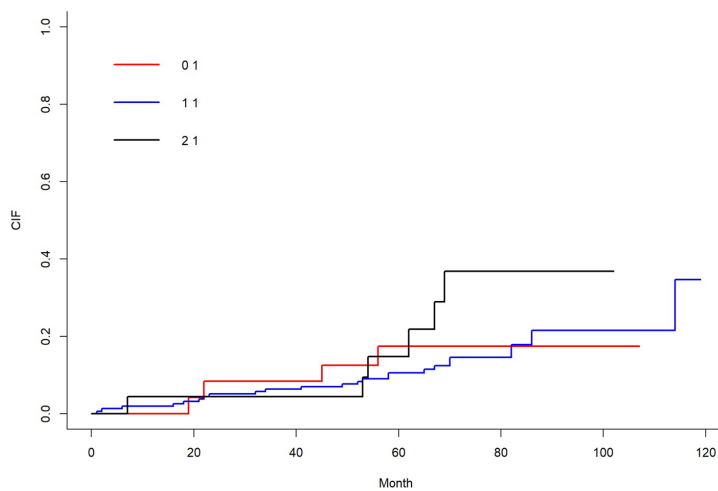


Figure 15. Cumulative risk map by different races

图 15. 不同种族的累积风险图

2.3.2. 单因素分析与多因素分析

表 7 所示是基于竞争风险模型构建的单因素与多因素分析表, 这里使用的数据集是随机获取的 206 例训练集。

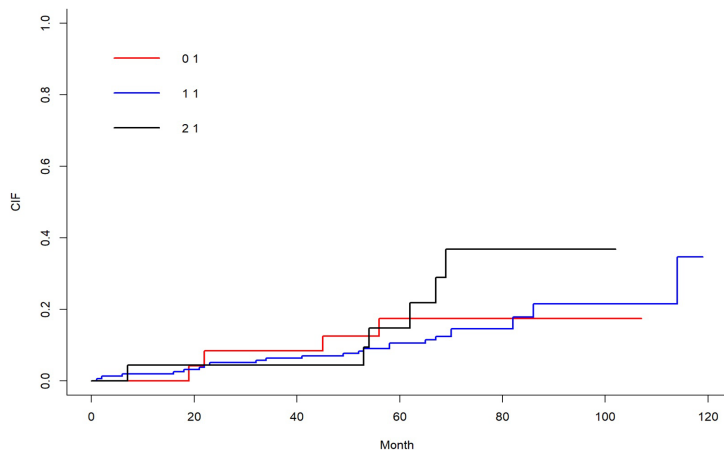


Figure 16. Cumulative risk maps for different oncology grades
图 16. 不同肿瘤学等级的累积风险图

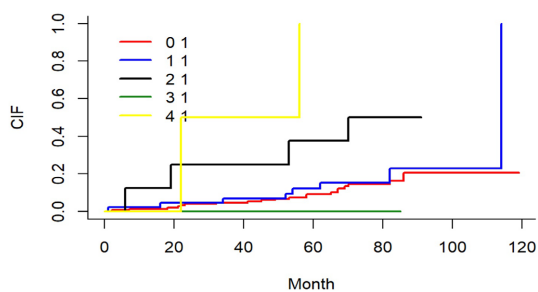


Figure 17. Cumulative risk map for different T stages
图 17. 不同 T 分期的累积风险图

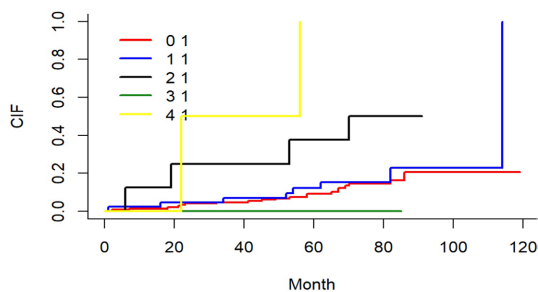


Figure 18. Cumulative risk map for different M stages
图 18. 不同 M 分期的累积风险图

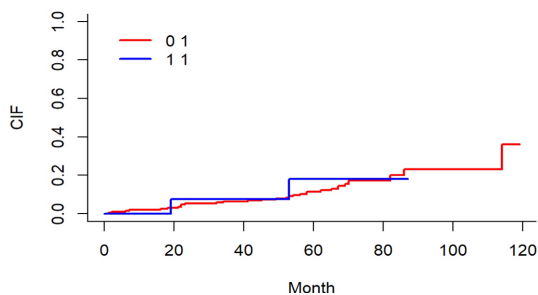


Figure 19. Cumulative risk map for different N stages
图 19. 不同 N 分期的累积风险图

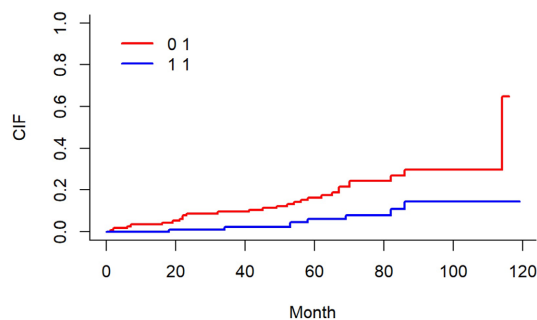


Figure 20. Cumulative risk map of radiotherapy or not
图 20. 是否接受放疗的累积风险图

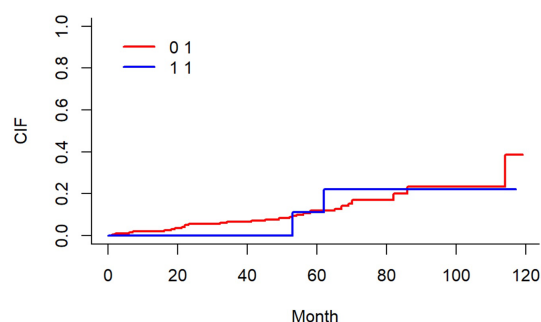


Figure 21. Cumulative risk map of chemotherapy or not
图 21. 是否接受化疗的累积风险图

在单因素分析表中，年龄，M 分期，是否接受放疗这三个变量都通过了显著性检验，这表明生存预后而言都是独立的预后因素。在性别变量中，女性相较于男性死亡风险高，但不显著；肿瘤学分级变量中，等级越高死亡风险越高，但不显著；T 分期变量中，只有 T4 分期是显著的且相较于 T1 分期，死亡风险更高；M 分期变量中，M1 的死亡风险相较于 M0 高；N 分期中，N0 的死亡风险更高；接受放疗和化疗的患者的死亡风险更低。

在多因素分析表中，年龄，M 分期，是否接受放疗这三个变量与单因素分析是一致的，都是显著的 ($P > 0.05$)。

Table 7. Single factor analysis and multi factor analysis table
表 7. 单因素与多因素分析表

	单因素分析		多因素分析	
	SHR	P 值	SHR	P 值
Age				
≤74	参考			
>74	24.09	<0.01	5.23	<0.01
Sex				
Male	参考			
Female	2.48	0.12	0.43	0.31
Race				

Continued

Black	参考			
White	2.54	0.28	1.37	0.46
Others	1.96	0.39	1.48	0.53
Grade				
I	参考			
II	2.51	0.62	0.95	0.67
III	2.63	0.6	0.91	0.64
AJCC.T				
T1	参考			
T2	3.03	0.33	2.65	0.21
T3	1.34	0.12	2.11	0.11
T4	1.21	<0.01	1.19	<0.01
TX	1.44	0.21	1.87	0.19
AJCC.M				
M0	参考			
M1	8.58	0.03	72.72	<0.01
AJCC.N				
N0	参考			
N1	0.03	0.85	0.41	0.76
Radiation				
None	参考			
Yes	0.27	<0.01	0.43	0.05
Chemotherapy				
None	参考			
Yes	0.01	0.93	0.69	0.89

2.3.3. 模型构建与验证

根据上述多因素分析，选取年龄、M分期，是否接受放疗这三个变量作为自变量，生存时间和生存结局作为因变量与Cox比例风险模型不同的是在因变量中，将竞争事件引入。

Table 8. Table of coefficients of independent variables

表 8. 自变量系数表

变量	SHR	P值
Age		
≤74	参考	
>74	5.83	<0.01
AJCC.M		

Continued

M0	参考	
M1	9.33	<0.01
Radiation		
None	参考	
Yes	0.39	0.03

从表 8 可以看出：通过逐步回归且满足了 AIC 最小准则，构建的竞争风险模型，于是竞争风险模型表达式如下：

$$\lambda(t|\omega) = \lambda_0(t) \exp(5.83X_1 + 9.33X_6 + 0.39X_8)$$

这里年龄，M 分期，是否接受放疗三个变量都通过了显著性检验，说明对于预后是显著相关的，年龄较大的患者的死亡风险是年龄较小的 5.88 倍，M 分期为 M1 的患者的死亡风险是 M0 的患者的 9.33 倍，在同等条件下，接受放疗的患者组的死亡风险是未接受放疗的患者组的 0.39 倍。综上所述，年龄和 M 分期为危险预后因素，即数值越大，生存时间越短；是否接受放疗为保护预后因素，与 Cox 比例风险回归模型一致。

对上述模型进行显著性检验，具体检验结果如表 9 所示。三种检验的 P 值远远小于 0.05，这说明上述模型通过了显著性检验，模型合理，同时这三个变量能够很好的反应生存函数的变化情况。

Table 9. Significance test table of competitive risk model

表 9. 竞争风险模型显著性检验表

检验类型	卡方值	自由度	P 值
Likelihood ratio test	33.92	3	<0.01
Wald test	27.23	3	<0.01
Score (logrank) test	38.83	3	<0.01

根据上述的竞争风险模型，可以预测患者 12 个月，24 个月，36 个月预后的列线图。这里选择的 12 个月，24 个月，36 个月时由于研究的对象为老年患者，预后效果较差。

图 22 是基于竞争风险模型所构建的生存概率的预测图，能够直观的计算出患者的生存概率。在图 22 中，我们可以看到随着时间的增加，生存概率下降。其中，年龄和 M 分期是危险预后因素且 M 分期的影响较大，接受放疗是保护预后因素，采用放疗能够很大程度的降低死亡的风险。

根据上述的竞争风险模型得到的 $C-index = 0.778$ ，可以得知该模型预测较为准确，为中等准确度。

如图 23 所示，描绘了患者在 12 个月、24 个月和 36 个月的校准曲线图，通过将预测概率与实际概率作为比较进一步分析所构造的竞争风险模型预测的准确性。在时间为 12 个月时，图 23 中的红线分布在对角线左右，时间越长，偏差越大。综上所述，说明基于列线图的预测，短期内比较准确，但长期预测效果一般。

2.4. 模型比较

这里模型的准确性主要依据 C-index 的值来进一步比较和说明，数值越大，表示预测的效果越好，反之亦然。两个模型中，Cox 比例风险模型的 C-index 的值为 0.769，竞争风险模型的 C-index 的值为 0.778。

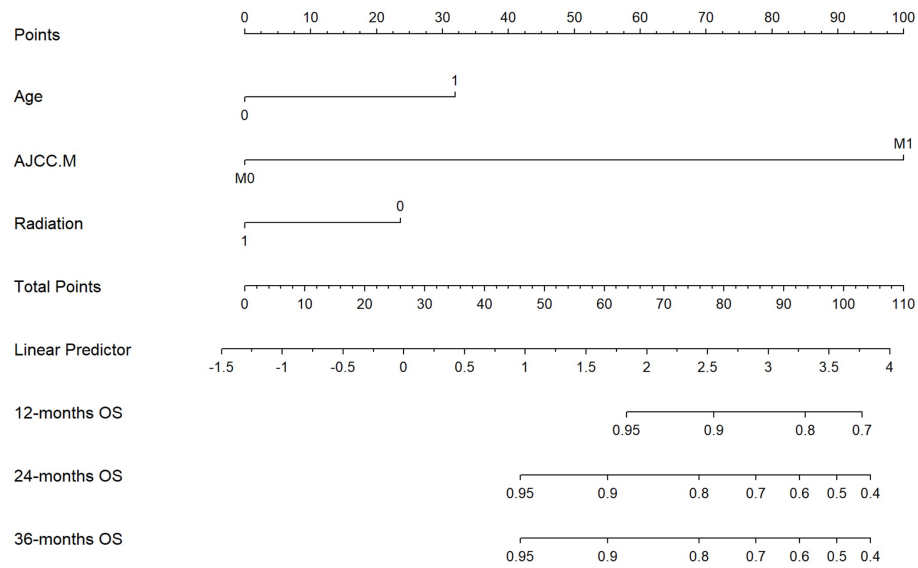


Figure 22. Nomogram for prediction of patient survival probability

图 22. 患者生存概率预测列线图

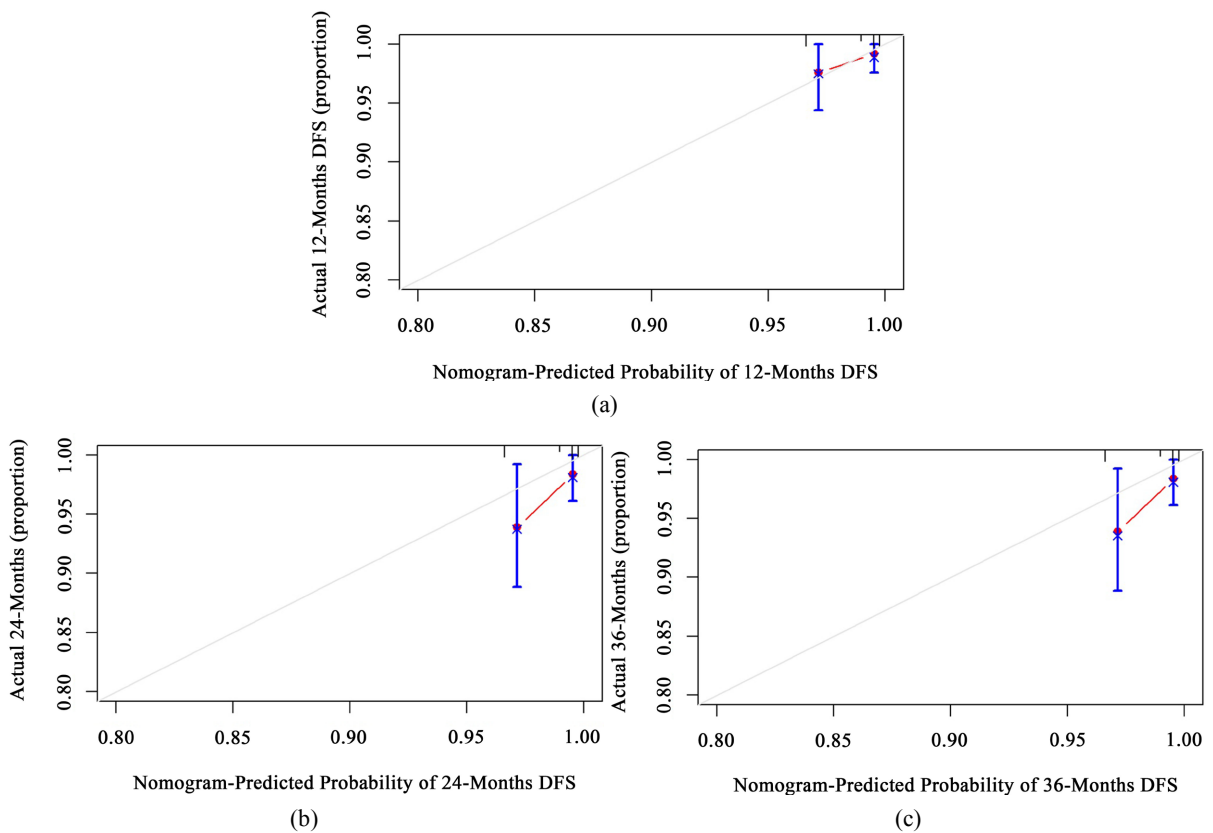


Figure 23. Calibration graph of patients competing for risk models of 12, 24, 36 months

图 23. 竞争风险模型的患者 12, 24, 36 个月的校准曲线图

可以看出竞争风险模型能够考虑到了生存结局中存在着竞争时间，从而导致了预测上有一定误差。校准曲线主也可以评价预测模型的校准度，它的横坐标表示的是对于患者预测的生存概率，而纵坐标表示

的是患者实际的生存概率，当绘制的曲线越接近 45 度，则说明模型预测能力较为准确。从校准曲线对比也可以看到，竞争风险模型的偏差比 Cox 比例风险模型要小。

综上所述，高龄乳腺癌患者确实存在着其它因素导致的竞争事件，因此竞争风险模型对本文的高龄乳腺癌患者预测有着更好的准确性。

3. 结论

本文主要通过研究来自于美国癌症数据库 SEER 中共计 412 例老年乳腺癌患者，得到以下结论：

1) 通过 K-M 曲线可以直观看出对于总体生存的特征：由于选用的患者数据年龄在 60 岁以上，总体的生存预后的时间相较其它癌症而言较长。生存曲线在 60 个月时，生存概率下降到 80% 左右，而在 120 个月时，生存概率下降到 60% 左右。这表明，高龄乳腺癌患者的生存率下降速度比较稳定，生存曲线几乎为一条直线，但在预后时间为 120 个月时，生存概率也只有 60%，这说明如果要提高总体的生存率，应该紧密关注患者的状态，对其进行积极的预后干预。

2) 使用 Cox 风险比例回归模型和竞争风险模型进一步研究对于患者生存预后影响因素的探讨，对于 Cox 风险比例模型和竞争风险模型构建纳入的变量均为年龄，M 分期，是否接受放疗三个变量。结果显示，Cox 模型的预测效果是较差，C-index 的值为 0.769，这主要是由于仅仅考虑传统的生存结局，忽略了对于高龄人群而言存在着竞争事件。竞争风险模型在对于存在竞争事件的生存分析中有着较好的表现，C-index 的值为 0.778，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。从校准曲线对比也可以看到，竞争风险模型的偏差比 Cox 比例风险模型要小，也就是说竞争风险模型的效果更好。

结合本文结论，我们提出以下建议：

1) 关注高龄人群，尽早治疗就医。乳腺癌作为全球最常见的癌症之一，特别是老年人，随着岁数的增长，预后效果会越来越差，但不可否认的是，随着全球老龄化的到来，老年人已逐步成为我国人口结构中重要组成的部分，所以要重点聚焦于老年人，定期体检，尽早发现问题，及时就医诊治，降低死亡率。

2) 注意对患者的分类，明晰风险程度。对于乳腺癌患者而言，一旦 M 分期为 M1，其死亡风险很高，那么针对死亡风险高的病人一定要重点照顾，积极采用治疗方式，尽可能地提高相关的预后效果，有的放矢的诊断治疗，能够极大程度地规避风险，提高诊治过程中的效率，避免盲目检测，延误病情。

3) 重点把握放疗的治疗手段。对于乳腺癌患者，化疗的治疗手段确实能够极大程度降低整体的死亡风险，在临床过程中，可以考虑将该治疗方式作为首选的方案，而化疗的疗效并不突出，可以视情况不予考虑。

参考文献

- [1] 任云, 张晋丹, 徐建忠, 杨海波. 年轻乳腺癌患者与老年乳腺癌患者临床病理学特征对比研究及预后分析[J]. 世界最新医学信息文摘, 2019, 19(92): 1-3. <https://doi.org/10.19613/j.cnki.1671-3141.2019.92.001>
- [2] 程梅莲, 孙闻, 李小琴, 庄秀芬. 基于 SEER 数据库分析老年乳腺癌患者预后相关因素[J]. 江苏大学学报(医学版), 2022, 32(3): 251-255. <https://doi.org/10.13312/j.issn.1671-7783.y210190>
- [3] 邱梅清, 佟仲生, 贾勇圣, 董国雷, 郝春芳. 老年乳腺癌临床特征预后多因素分析[J]. 实用肿瘤杂志, 2013, 28(2): 162-166. <https://doi.org/10.13267/j.cnki.syzlzz.2013.02.004>