Statistics and Application 统计学与应用, 2021, 10(5), 794-804
Published Online October 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.105082

Hans汉斯

基于生存分析的口腔鳞状细胞癌患者研究

曾　莎*，李　超，瞿颖秋，张晓良

重庆理工大学理学院，重庆

收稿日期：2021年9月2日；录用日期：2021年9月17日；发布日期：2021年10月8日

摘　要

本文基于生存分析方法对口腔鳞状细胞癌患者特点及预后影响因素研究。首先利用参数估计和非参数估计对生存函数进行估计，了解其特点；然后基于COX模型，以死亡时间为因变量，分别以年龄、肿瘤阶段及性别为因变量建模，分析生存时间影响因素。研究结果表明将死亡的时间和年龄、肿瘤阶段以及性别建立的模型有效，得出影响口腔鳞状细胞癌患者死亡时间的主要因素为年龄、肿瘤阶段以及性别。

关键词

参数估计，非参数估计，Cox模型，生存分析

Oral Squamous Cell Carcinoma Patient Study Based on Survival Analysis

Sha Zeng*, Chao Li, Yinqiu Qu, Xiaoliang Zhang

College of Science, Chongqing University of Technology, Chongqing

Received: Sep. 2nd, 2021; accepted: Sep. 17th, 2021; published: Oct. 8th, 2021

Abstract

Based on the survival analysis method, the characteristics and prognostic factors of patients with oral squamous cell carcinoma were studied. Firstly, the survival function is estimated by parametric estimation and nonparametric estimation to understand its characteristics. Then, based on COX model, time of death was used as the dependent variable, and age, tumor stage and gender were used as the dependent variable to analyze the factors affecting survival time. The results showed that the model of time of death and age, tumor stage and gender was effective, and the

*通讯作者。

文章引用: 曾莎, 李超, 瞿颖秋, 张晓良. 基于生存分析的口腔鳞状细胞癌患者研究[J]. 统计学与应用, 2021, 10(5): 794-804. DOI: 10.12677/sa.2021.105082

main factors influencing the time of death of patients with oral squamous cell carcinoma were age, tumor stage and gender.

Keywords

Parameter Estimation, Non-Parameter Estimation, Cox Model, Survival Analysis

1. 引言

口腔鳞状细胞癌(OSCC)是目前为止非常常见的恶性肿瘤，其每年的发病率逐渐升高，同时许多因素会影响口腔鳞状细胞癌患者的生存时间，因此，明确口腔鳞状细胞癌患者的特点和影响其生存时间的因素对于提高患者生存质量有重要意义。

许多学者对影响口腔鳞状细胞癌患者预后的因素进行了分析，其中，张华、杨蓉[1]等人利用SPSS20.0 及 STATA 软件对 389 例初诊 OSCC 患者的临床病理及随访资料进行了统计学分析，发现 OSCC 5 年生存率相对较低。中晚期患者淋巴结外侵犯明显，严重影响预后。早期发现和诊治可降低 OSCC 复发率，从而提高患者生存率；王立业[2]等人对口腔鳞状细胞癌患者预后相关基因标志物的生物信息学进行分析；张泽军、南欣荣、闫星泉[3]等人探讨临床 III~IV 期口腔鳞状细胞癌患者术后复发的危险因素和复发患者的预后状况，采用单因素和多因素 Logistic 回归分析研究影响患者术后复发的危险因素，以及单因素和多因素生存分析研究复发患者的预后状况，发现 III~IV 期患者复发率较高且复发后预后不佳，淋巴结比率和神经侵犯时影响复发的不良特征；霍玉荣、康鹏[4]等人分析了口腔鳞状细胞癌患者生存及复发的临床影响因素，采用 Logistic 回归分析了 88 例口腔鳞状细胞癌患者的临床资料，发现口腔鳞状细胞癌患者预后并不理想；以及 Y.K [5]等人研究了在台湾南部 703 人口腔鳞状细胞癌预后因素。在这些研究中，大多数学者采用单因素分析、多因素分析等方法对患者预后因素进行研究，而对于口腔鳞状细胞癌患者生存时间及其影响因素的研究还比较少，因此本文考虑采用生存分析的方法对患者进行研究。

2. 口腔鳞状细胞癌患者的生存分析

2.1. 研究数据

本文研究的数据为 orca 数据集，我们对 1985.1.1 到 2005.12.31 这个时间段中芬兰最北部省份诊断结果为口腔鳞状细胞癌的 338 名患者进行研究。本次研究的患者的随访开始日期为癌症诊断当天，并且患者于 2008.12.31 死亡，迁移或随访截止日期结束。造成患者死亡原因主要有两种：1) OSCC 死亡；2) 其他原因造成的死亡。

2.2. OSCC 患者生存数据分析

生存分析主要关注于事件数据的时间，在本文中，为诊断后的死亡时间。

为了明白该数据的特点以及其主要形式，我们首先对 OSCC 患者进行生存数据分析，如图 1 所示，其中图 1 左半部分为受试者随访时间与事件的关系，右半部分为受试者随访时间与年龄的关系：

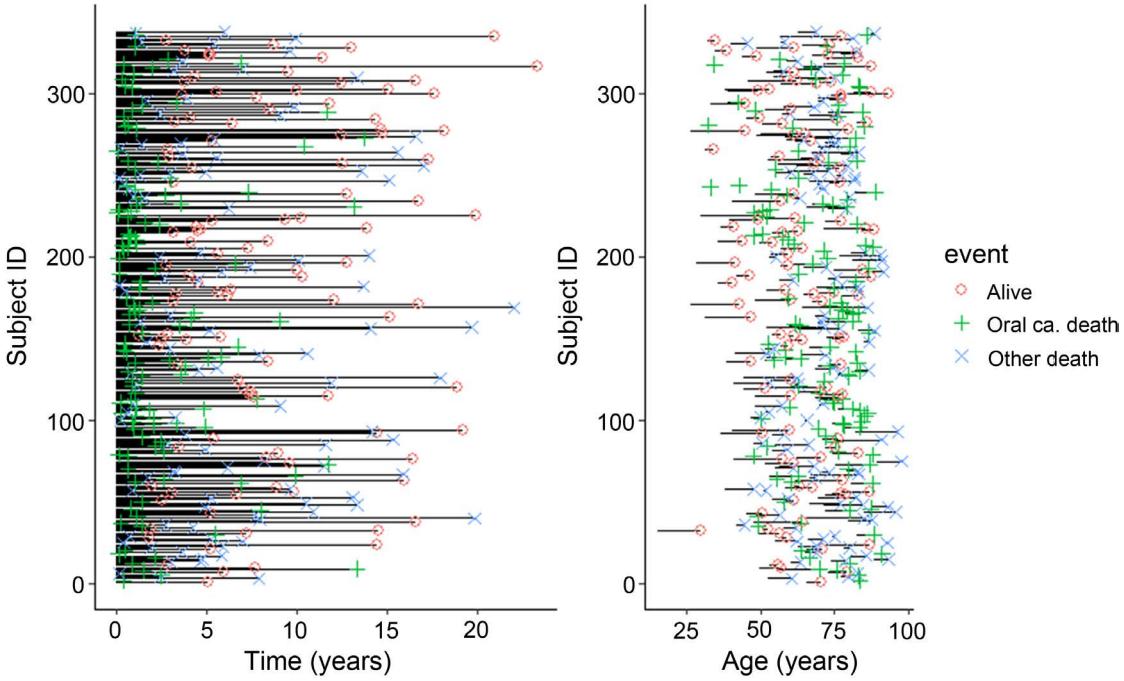


Figure 1. Relationship between follow-up time and event and age
图 1. 受试者随访时间与事件和年龄的关系

从上述左图我们可以看出，在随访事件内，因口腔鳞状细胞癌死亡的受试者大多发生在早期，即早期 OSCC 疾病引起的死亡率较高，而不是由于其他原因引起的死亡。从上述右图可以看出，因口腔鳞状细胞癌死亡的患者大多集中在中老年段和老年段，中老年人和老年人抵抗力弱，产生抗体的能力弱，所以更容易死亡。

2.3. OSCC 患者生存函数估计

2.3.1. 非参数估计

1) Kaplan-Meier 估计：

K-M 方法即乘积极限法，是一种统计描述方法，充分的利用了信息，给出了准确的统计量。假设在 r 个生存时间中有 J 个死亡时间：$\tau_1 < \cdots < \tau_j < \cdots < \tau_J$.令 $\tau_0 = 0$，用 τ_{J+1} 表示最大存活时间。对删失生存数据的 Kaplan-Meier 方法首先以死亡时间为切入点讲随访期划分为 $J+1$ 个区间：$[\tau_0, \tau_1), [\tau_1, \tau_2), \cdots, [\tau_j, \tau_{j+1}), \cdots, [\tau_{J-1}, \tau_J), [\tau_J, \tau_{J+1}]$。设 a_j 表示在第 j 个区间 $(j = 0,1,\cdots,J)$ 中的死亡人数，根据定义 $a_0 = 0$；设 r_j 表示第 j 个风险集合 $(j = 0,1,\cdots,J)$ 中的受试者数，并用 r_{J+1} 表示存活到 τ_{J+1} 的受试者数；设 p_j 表示在给定存活到 $\tau_j - \varepsilon (j = 0,1,\cdots,J)$ 时，存活到 $\tau_j + \varepsilon$ 的条件概率。

基于上述划分和某些条件概率推导出生存概率 $S(T)$ 的估计如下：

$$\hat{S}_j = \hat{p}_1 \hat{p}_2 \cdots \hat{p}_j$$

生存函数的 K-M 估计见图 2。

从下述的 K-M 生存曲线图中，可以得到每个时间点的生存概率。在早期时，陡峭的 K-M 曲线，说明在这一段时间内，死亡人数多，也证实了生存数据分析板块中，早期口腔鳞状细胞癌患者的高死亡率；在中期时，K-M 曲线逐渐平稳，生存率下降趋势渐渐变缓；在后期时，K-M 曲线基本平稳，生存率不会有大幅度的下降。

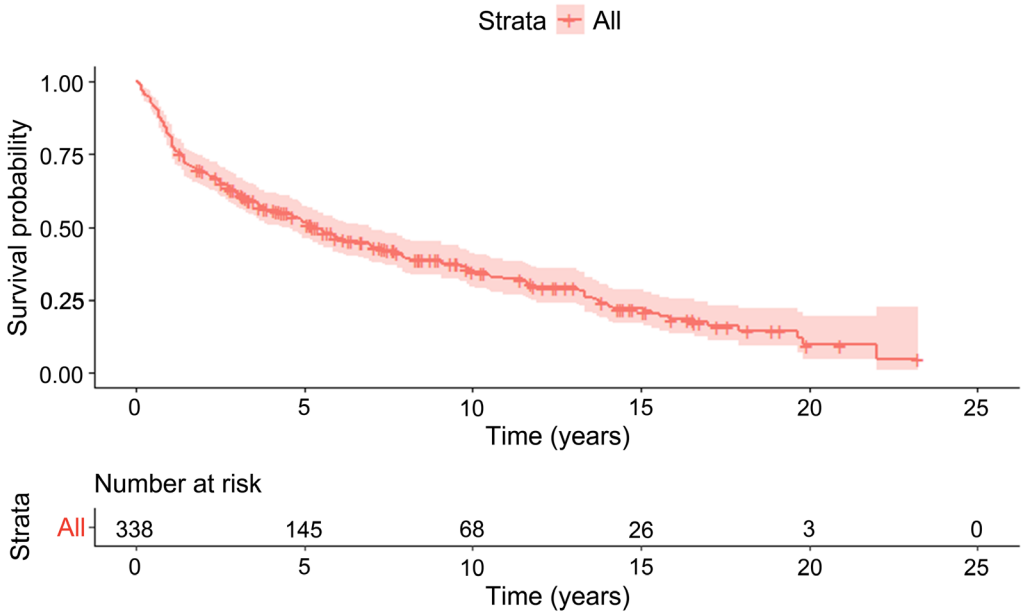


Figure 2. K-M survival curve
图 2. K-M 生存曲线图

2) Nelson-Aalen 估计：

Nelson-Aalen 估计是一种应用于生物统计学中对生存性概率分析的非参数估计方法，其根本的思想为依据累计的失效率函数对累计死亡率进行估计。本文研究数据的 N-A 估计表见表 1：

Table 1. N-A estimation table
表 1. N-A 估计表

	time	n.risk	n.event	n.censor	surv	std.err	upper	lower
1	0.085	338	2	0	0.9941	0.004184	0.998521	0.976618
2	0.162	336	2	0	0.988201	0.005935	0.995555	0.968869
3	0.167	334	4	0	0.976436	0.008431	0.988146	0.953437
4	0.17	330	2	0	0.970537	0.009457	0.984038	0.945933
5	0.246	328	1	0	0.967582	0.009937	0.981916	0.942227
6	0.249	327	1	0	0.964628	0.010397	0.979756	0.938553
7	0.252	326	3	0	0.955791	0.011676	0.97311	0.92774
8	0.329	323	1	0	0.952837	0.012079	0.970838	0.924167
9	0.334	322	1	0	0.949882	0.012472	0.968543	0.920615
10	0.413	321	1	0	0.946928	0.012855	0.966228	0.917082
…	…	…	…	…	…	…	…	…

3) Life-table：

生命表又称“死亡率表”，其根据年龄划分的死亡率进行编制，描述了一批人从出生后陆续死亡的

全部过程的一种统计表，其主要的内容有：① 当年生存者的年龄；② 在划分年龄组中的死亡人数；③ 在划分年龄区间的条件死亡概率；④ 在划分年龄区间的生存条件概率；⑤ 生存到年龄为 x 的人数。

本文研究数据的生命表结果见表 2：

Table 2. Life table
表 2. 生命表

	nsubs	nlost	nrisk	nevent	surv
0~1	338	0	338	64	1
1~2	274	4	272	41	0.810651
2~3	229	9	224.5	21	0.688457
3~4	199	12	193	20	0.624058
4~5	167	9	162.5	13	0.559389
5~6	145	14	138	13	0.514638
6~7	118	5	115.5	8	0.466157
7~8	105	8	101	9	0.433869
8~9	88	7	84.5	1	0.395208
9~10	80	4	78	8	0.390531
10~11	68	4	66	5	0.350476
11~12	59	3	57.5	5	0.323925
12~13	51	6	48	0	0.295758
13~14	45	2	44	8	0.295758
14~15	35	6	32	3	0.241984
15~16	26	3	24.5	4	0.219298
16~17	19	5	16.5	2	0.183494
17~18	12	2	11	1	0.161252
18~19	9	2	8	0	0.146593
19~20	7	2	6	2	0.146593
20~21	3	1	2.5	0	0.097729
21~22	2	0	2	1	0.097729
22~23	1	1	0.5	0	0.048864

2.3.2. 参数估算器

1) 指数模型：

设数据来自指数分布，其概率密度函数为：

$$f(t)=\begin{cases}\lambda e^{-\lambda t} & t\geq 0,\lambda>0\\ 0 & t<0\end{cases}$$

分布函数：$F(t)=1-e^{-\lambda t}(t\geq 0)$

生存函数：$S(t)=1-F(t)=e^{-\lambda t}$

危险函数：$h(t)=\frac{f(t)}{S(t)}=\lambda,\lambda>0,t\geq 0$

2) Weibull 模型：

Weibull 模型的相关函数如下。

生存时间 T 的概率密度函数：$f(t)=\lambda\gamma(t)^{\gamma-1}\exp\left[-\lambda(t)^{\gamma}\right]$

分布函数：$F(t)=1-\exp\left[-\lambda(t)^{\gamma}\right]$

生存函数：$S(t)=\exp\left[-\lambda(t)^{r}\right]$

危险函数：$h(t)=\lambda\gamma(t)^{\gamma-1}$

3) log-logistic 模型：

log-logistic 模型称为双对数模型。三种模型的曲线对比图见图 3：

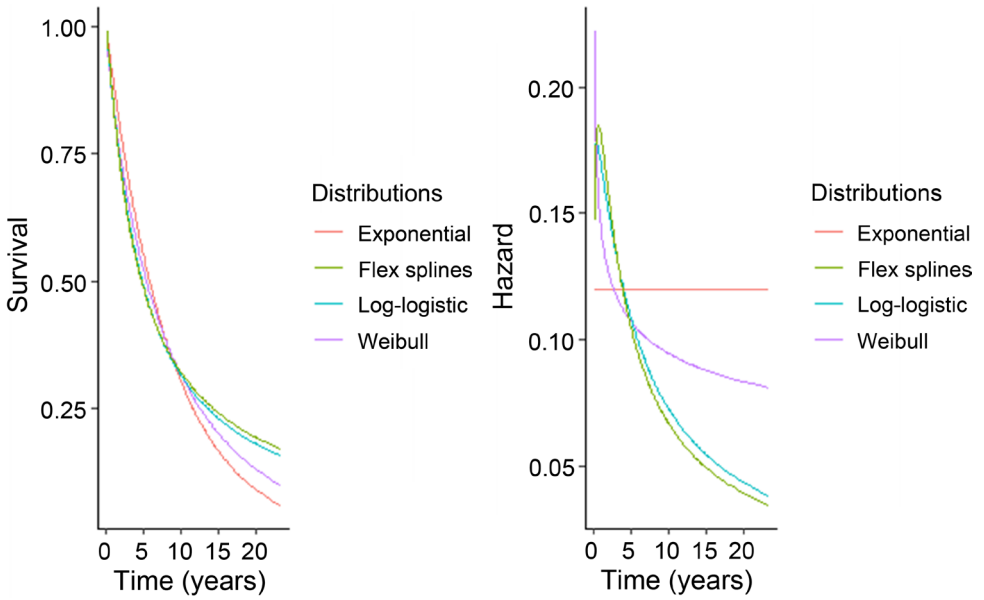


Figure 3. Comparison diagram of model curves
图 3. 模型曲线对比图

从图 3 右半部分图形可以看出，三种模型的生存曲线在随访时间 0~10 年期间的趋势比较接近，但是在随访时间 10 年之后的曲线有明显差异。从上述左图可以看出，指数分布的危险函数曲线为一条平行于 x 轴的线；而 log-logistic 模型的危险函数曲线，具有非常明显的下降趋势；Weibull 模型的危险函数曲线较 log-logistic 模型更为平缓，但两种模型都在随访时间 7 年左右下降趋势明显。

2.4. 生存曲线比较

2.4.1. 肿瘤阶段生存曲线比较

肿瘤阶段是常见的医学指标，在此我们对不同肿瘤阶段的生存曲线进行比较，判断肿瘤阶段是否为癌症存活研究中的重要影响因素。

从图 4 中我们可以看出，不同肿瘤阶段的生存曲线图有明显的差异，几乎没有重叠的部分，可能为癌症存活研究中的重要影响因素。

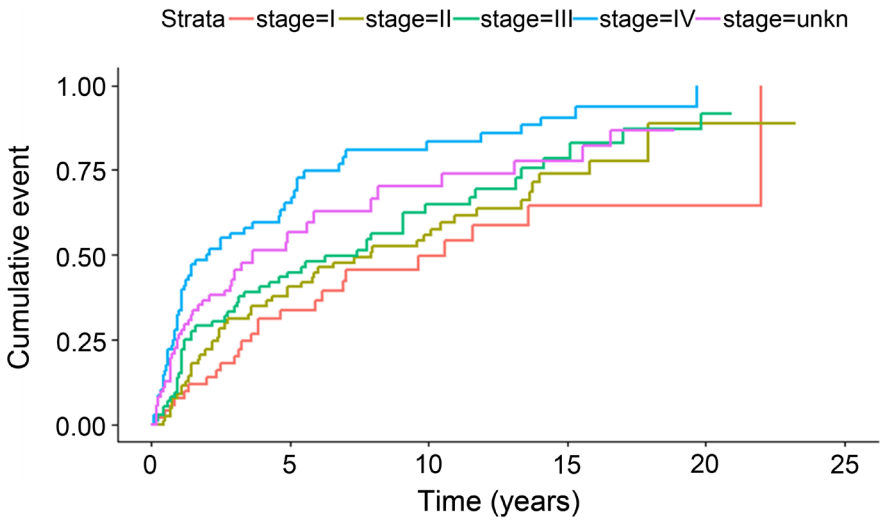


Figure 4. Comparison of survival curves at different tumor stages
图 4. 不同肿瘤阶段的生存曲线对比图

2.4.2. Mantel-Haenszel Logrank 检验

肿瘤阶段的 M-H logrank 检验结果见表 3：

Table 3. M-H test results
表 3. M-H 检验结果表

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
stage = I	50	25	39.9	5.573	6.813
stage = II	77	51	63.9	2.606	3.662
stage = III	72	51	54.1	0.174	0.231
stage = IV	68	57	33.2	16.966	20.103
stage = unkn	71	45	37.9	1.346	1.642
Chi sq = 27.2 on 4 degrees of freedom P = 0.00002					

从检验的结果来看，检验的 P 值为 0.00002，在 $\alpha = 0.05$ 的检验水平下，显著的拒绝原假设，所以没有充分的理由拒绝肿瘤阶段是癌症存活研究中的重要影响因素。从上述肿瘤阶段的生存曲线图也可以看出，低肿瘤阶段的癌症患者生存率要明显低于高肿瘤阶段的癌症患者，说明口腔鳞状细胞癌患者的早期死亡率高。

3. 模型

3.1. 建立模型

我们比较单个的因子水平的生存函数时，利用非参数检验非常有效。但是当我们需要检验因素的数量增大时，上述方法将会变得难以进行。所以，在此我们可以利用回归模型来发现生存和预测因子之间的关系。在此我们选择 CoxPH 模型[6]进行建模。我们考虑将死亡的时间和性别、年龄和肿瘤阶段进行建模。建模结果见表 4：

Table 4. CoxPH modeling results
表 4. CoxPH 建模结果

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
sexMale	0.35139	1.42104	0.14139	2.485	0.012947*
I((age-65)/10)	0.41603	1.51593	0.05641	7.375	0.000000000000165***
stageII	0.03492	1.03554	0.24667	0.142	0.887421
stageIII	0.34545	1.41262	0.24568	1.406	0.159708
stageIV	0.88542	2.42399	0.24273	3.648	0.000265***
stageunkn	0.58441	1.79393	0.25125	2.326	0.020016*

从 Cox 模型的结果可知，性别，年龄和阶段对模型具有显著影响。我们从估计中发现，第一阶段和第二阶段的差异非常微小；而对于未知阶段的群体来说，可能是来自不同阶段患者的混合，所以，我们可以将第一阶段和第二阶段进行组合。

3.2. 模型检验

3.2.1. 似然比检验、Wald 检验、Score 检验

对 CoxPH 建模结果，即死亡时间建模为性别功能、年龄和肿瘤阶段是否合理进行似然比检验、Wald 检验、Score 检验，检验结果见表 5：

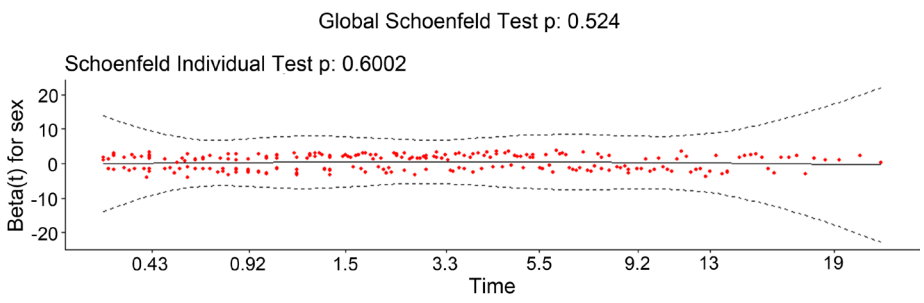
Table 5. Test results
表 5. 检验结果

Likelihood ratio test = 86.76 on 6 df	P =< 0.0000000000000002
Wald test =80.5 on 6 df	P = 0.000000000000003
Score (logrank) test = 82.86 on 6 df	P = 0.0000000000000009

分别从三种检验结果的 P 值可知，其 P 值都远小于显著性水平 0.05，所以该模型显著。选择年龄、肿瘤阶段以及性别建立模型是有效的。

3.2.2. 变量检验

除了对模型进行检验，还需要使用函数分别检查数据与功能性别、年龄和肿瘤阶段变量的比例风险假设是否和全局一致。检验结果见图 5：



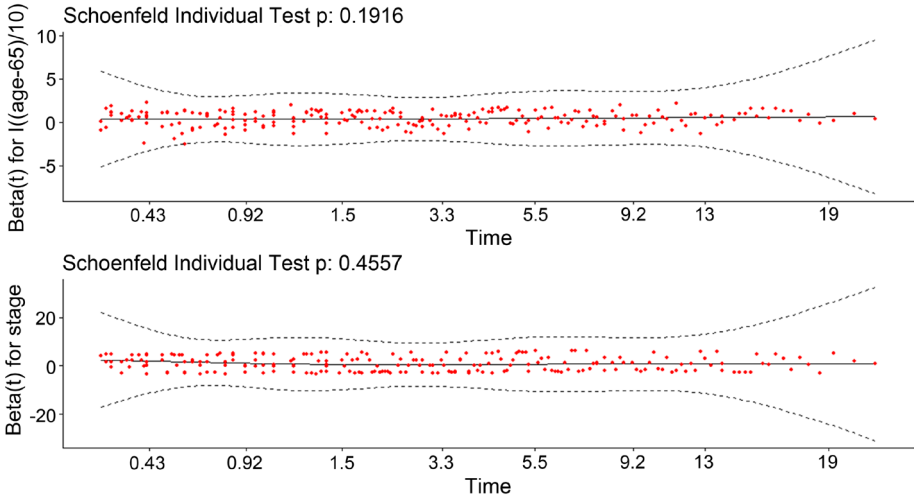


Figure 5. Test results
图 5. 检验结果

从上图中，我们可以看见其 P 值都明显大于显著性水平 0.05，没有充分的理由拒绝原假设，所以每个变量的比例风险假设分别和全局一致。

3.2.3. 图形化比较多变量

由于在 CoxPH 结果中，第一阶段和第二阶段没有明显变化，在此选取森林图，将第一阶段和第二阶段合为一个变量进行绘图与第一阶段和第二阶段分别进行绘图进行对比，图 6 为对比结果：

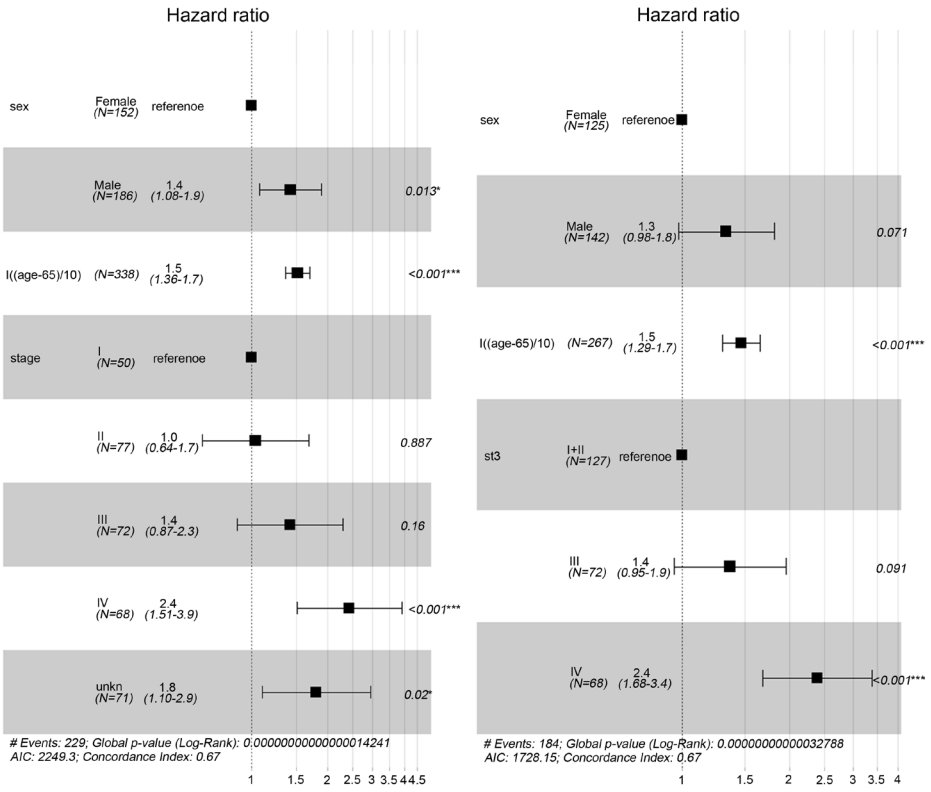


Figure 6. Comparison of variables
图 6. 变量对比

我们可以从上图发现，在图 6 左半部分图形中，第一阶段和第二阶段的图形非常接近，区别非常微小；在图 6 右半部分图形中，将第一阶段和第二阶段合为一个变量，其与其他变量的区别更为明显。所以可以将第一阶段和第二阶段合为一个变量进行建模。

3.3. 模型预测

逐步绘制预测的生存曲线，根据拟合的模型确定性别和年龄的值，图 7 为生存曲线拟合图：

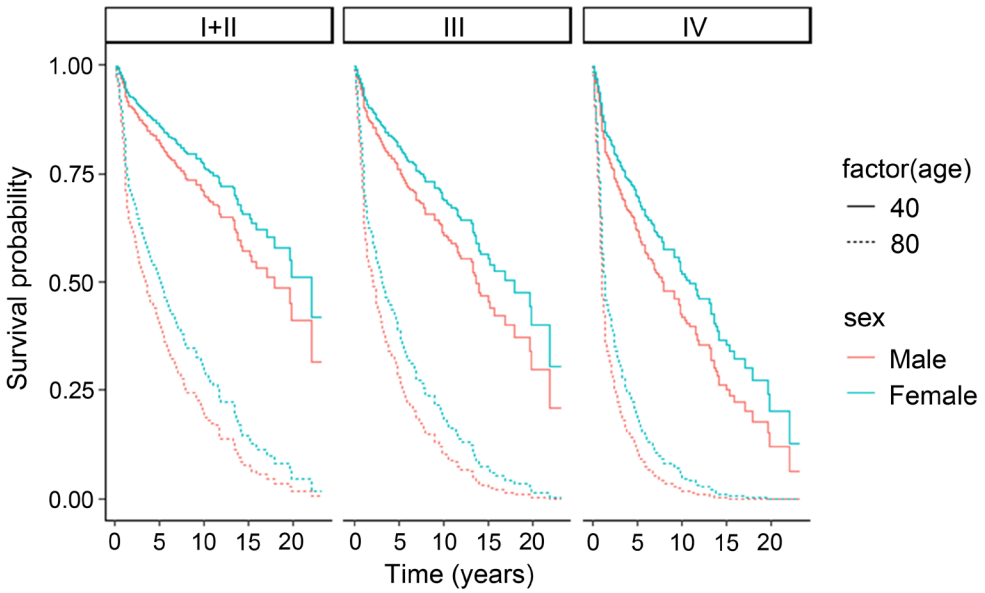


Figure 7. Survival curve fitting
图 7. 生存曲线拟合

在这里分别确定年龄为 40 和 80，性别为女性和男性。年龄为 80 岁的患者，生存率明显低于年龄为 40 岁的患者，且年老者生存函数陡峭，下降比中年者更为迅速。而对于性别来说，女性的死亡率比男性更低。

4. 结论

本文通过采用生存分析方法对口腔鳞状细胞癌患者进行研究，建立了Cox回归模型，得出了以下结论：

1) 口腔鳞状细胞癌患者的特点

在上述的分析结果中，我们可以知道。在随访事件内，因口腔鳞状细胞癌死亡的受试者大多发生在早期，即早期口腔鳞状细胞癌疾病引起的死亡率较高，而不是由于其他原因引起的死亡。而因口腔鳞状细胞癌死亡的患者大多集中在中老年段和老年段，中老年人和老年人抵抗力弱，产生抗体的能力弱，所以更容易死亡。

2) 影响口腔鳞状细胞癌患者死亡时间的因素

本文利用 Cox 模型将死亡时间建模和性别、年龄以及肿瘤阶段进行建模，从检验的结果可以得出，其 P 值都远小于显著性水平 0.05，所以该模型显著。由于其肿瘤第一阶段和第二阶段从图形分析上非常接近，于是便将两个阶段合并，采用随机森林图对合并前后进行对比，结果发现，合并后的效果更加优良，于是采用合并后的 Cox 模型，最后对模型进行预测，发现年龄、肿瘤阶段以及性别是影响死亡时间的重要因素。

参考文献

[1] 张华, 杨蓉, 叶贝贝, 张文超. 389 例口腔鳞状细胞癌预后影响因素分析[J]. 天津医科大学学报, 2018, 24(4): 315-322.

[2] 王力业, 高莺, 田淳. 口腔鳞状细胞癌患者预后相关基因标志物的生物信息学分析[J]. 口腔预防疾病, 2021, 29(21): 27-33.

[3] 张泽君, 南欣荣, 闫星泉. III~IV 期口腔癌患者术后复发的危险因素及复发患者的预后分析[J]. 医学研究杂志, 2021, 50(2): 102-106.

[4] 霍玉荣, 康鹏, 宋伟霞. 口腔鳞状细胞癌患者生存及复发的临床影响因素分析[J]. 实用癌症杂志, 2021, 36(7): 1175-1177.

[5] Chen, Y.K., Huang, H.C., Lin, L.M. and Lin, C.C. (1999) Primary Oral Squamous Cell Carcinoma: An Analysis of 703 Cases in Southern Taiwan. *Oral Oncology*, **35**, 173-179. https://doi.org/10.1016/S1368-8375(98)00101-8

[6] 路文馨. 基于比例风险模型的生存分析研究[J]: [硕士学位论文]. 广州: 华南理工大学, 2019.