Hans Journal of Data Mining 数据挖掘, 2022, 12(1), 43-48
Published Online January 2022 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2022.121006

Hans汉斯

基于Kaplan-Meier与Cox的口腔鳞状细胞癌患者数据分析

黄钰凝

云南民族大学，云南 昆明

收稿日期：2021年12月13日；录用日期：2022年1月13日；发布日期：2022年1月25日

摘 要

临床医疗事业是大数据发展及应用中重要的一个方面。本文以口腔鳞状癌患者的数据作为基础，通过对生存分析(Kaplan-Meier)曲线和对患者观察的数据进行挖掘，探究出影响因素与生存时间和结局的关系，再通过Cox回归模型对风险率进行预测，实现了研究各个因素对生存概率的重要性。该模型可用于临床医疗对口腔鳞状癌症的研究。

关键词

医疗大数据，Kaplan-Meier，Cox模型，比例风险假设

Data Analysis of Patients with Oral Squamous Cell Carcinoma Based on Kaplan-Meier and Cox

Yuning Huang

Yunnan Minzu University, Kunming Yunnan

Received: Dec. 13th, 2021; accepted: Jan. 13th, 2022; published: Jan. 25th, 2022

Abstract

Clinical medical career is an important aspect of the development and application of big data. In this paper, using data from patients with oral squamous carcinoma as a basis, we explored the relationship between the three of influencing factors, survival time and outcome by mining survival analysis (Kaplan-Meier) curves and data from patient observation, and then predicted the risk rate

文章引用: 黄钰凝. 基于 Kaplan-Meier 与 Cox 的口腔鳞状细胞癌患者数据分析[J]. 数据挖掘, 2022, 12(1): 43-48.
DOI: 10.12677/hjdm.2022.121006

by Cox regression model, and achieved to study the importance of each factor on survival probability. This model can be used in clinical care for the study of oral squamous cancer.

Keywords

Big Data in Healthcare, Kaplan-Meier, Cox Model, Proportional Risk Assumption

1. 引言

癌症对于世界医疗事业一直以来都是一个很大的挑战，大数据时代的到来为此做出了一定的贡献。医疗大数据目前存在四个问题：数据量剧增，大量类型复杂的数据给数据的存储、分析、处理带来很大的挑战；信息采集网不完善，导致采集到的信息与实际应用存在一定程度的脱离；大数据分析能力欠缺，医务人员不擅长大数据分析，容易导致大量数据的潜质挖掘不出来；数据安全问题，互联网发达容易造成病人信息泄露[1]。

本文以芬兰北部两个省的 338 名口腔鳞状癌患者的数据为例，对数据进行整合利用，在客观的角度上分析预测病患情况，致力于用 Kaplan-Meier 与 Cox 模型解决目前医疗界大数据分析能力欠缺的问题，通过对数据的建模和分析，往往能够对医学事业起到指导性的作用，不仅可以了解到什么因素对口腔鳞状癌影响最大，也可以预测患者的生命期，以及通过图像形象地了解影响因素与生存时间和结局的关系。同时在病人的病情预测和治疗方面起到关键作用，以便医务人员更早采取预防措施或救治行动。

2. 数据介绍和数据预处理

2.1. 数据介绍

本文研究的数据包括自 1985 年 1 月 1 日起至 2005 年 12 月 31 日，共计 30 年的患者数据。其中包括病人序号、病人性别、诊断出癌症时的年龄、确诊时肿瘤的分期、从诊断到死亡或截止调查的随访时间、结束随访时病人的生存状况，共计 6 个属性特征值。

此数据有两个较明显的特征：右偏分布和删失。在这种存在很多删失的情况下，采用生存分析方法才是合理的选择。

2.2. 数据预处理

对数据观察后发现，无缺失值、无重复值、无非结构化数值。

为了方便研究，将病患的性别分别替换为男：1，女：0；将肿瘤分期的罗马数字分别替换为阿拉伯数字 1 至 5；新添加一列数据为在结束随访时病患是否仍然存活，存活：1，死亡：0；新添加一列数据为病患是否死于口腔鳞状癌症，是：1，否：0。

3. 数据初步观察

病患的此段患病时间与肿瘤分期的关系如图 1 所示，大体上是一个反比例，stage 越大，time 越小，即癌症分期越是后期，存活时间越短。

年龄和生存状况的关系，年龄越大存活几率越小，且较年轻人存活率高为明显，年龄是影响病症的一大重要因素。

死于口腔鳞状癌症的病患大多是中后期，即 3、4、5 期的病患，死于其他原因以及存活下来的病患大多处于早期，即 1、2 期。

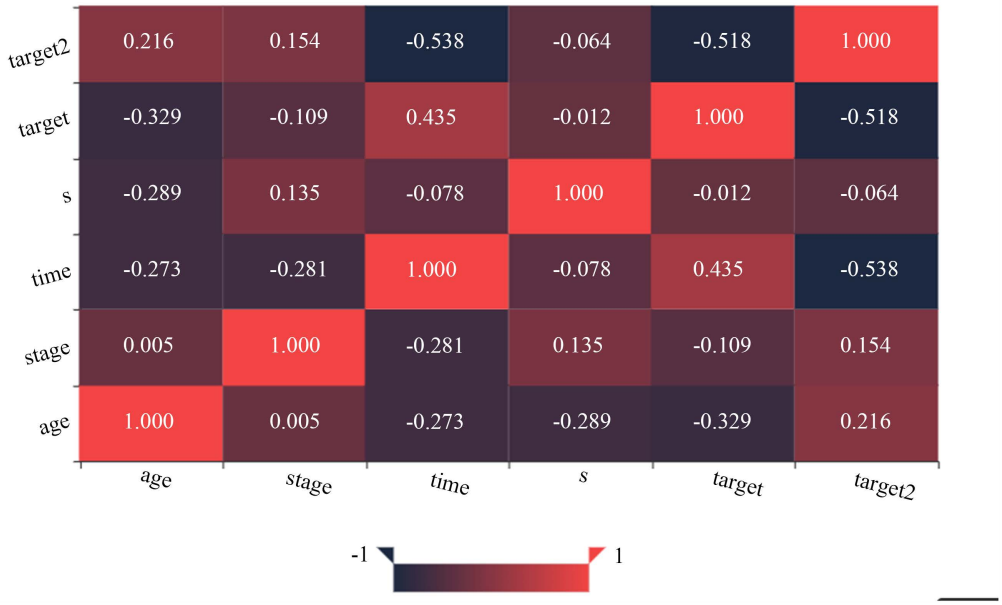

Figure 1. Heat map of data feature relationships
图 1. 数据特征关系热力图

患者年龄最大的 92 岁，最小的仅为 15 岁，且平均年龄和年龄中位数均为 64 岁左右。对于重点关注的生存时间，平均存活时间为 5.662 年，因为数据有删失的特征，存活期应该要比平均存活期长。截至统计数据的 2005 年 12 月 31 日，男性存活 59 人，女性存活 50 人；男性死于口腔鳞状细胞癌 62 人，女性死于此癌症 60 人；男性死于其他病因 65 人，女性死于其他病因 42 人。跟查时间较长的病人大多数死于此癌症。男性存活率低于女性。

4. Kaplan-Meier 及 Cox 原理

4.1. Kaplan-Meier 原理

Kaplan-Meier 方法是帮助我们描述生存结局发生情况的有效手段。曲线为我们描画了患者生存率随时间变化的特征，它完美的将时间因素考虑在内，各个时间点的生存率值也被称为时点生存率[2]。

$$\begin{cases} f(t)=\lim\limits_{\Delta t\to 0}\dfrac{1}{\Delta t}Pr\left(t\le T\le t+\Delta t\right) \\ F(t)=P\left(T\le t\right)=\int_0^t f(u)\mathrm{d}u \\ S(t)=P\left(T>t\right)=\int_0^{\infty} f(u)\mathrm{d}u \\ \lambda(t)=\lim\limits_{\Delta t\to 0}\dfrac{1}{\Delta t}Pr\left(t\le T\le t+\Delta t\,|\,\aleph\ge T\right)=\dfrac{f(t)}{S(t)} \\ A(\tau)=\int_0^t \lambda(u)\mathrm{d}u \end{cases} \tag{1}$$

分别为密度函数、积累密度函数、生存函数风险函数以及风险积累函数。

4.2. Cox 建模原理

提出了可加函数 Cox 模型来灵活量化函数协变量与事件数据时间之间的关联[3]。

Cox 的比例风险模型定义为：

$$h(t,x)=h_0(t)\times\exp\left(b_1x_1+b_2x_2+\cdots+b_px_p\right) \tag{2}$$

$h_0(t)$被称为基准风险函数，t 代表生存时间，$h_0(t)$也为当所有协变量取值为 0 时的风险函数。exp(b)为预后指数，若 exp(b)的值越大，则其风险函数 $h(t,x)$ 越大，进而预后越差。$X=\left(x_1,x_2,\cdots,x_p\right)$ 是协变量，协变量是固定值，且协变量的效应不随时间改变而改变[4]。由于此回归模型只对参数 b 进行估计，所以是一个半参数模型。

5. Kaplan-Meier 曲线及 Cox 建模

5.1. Kaplan-Meier 曲线

生存曲线由美国生物学家雷蒙·普尔首次提出，是反映种群个体在各年龄级的存活状况曲线，是借助于存活个体数量来描述特定年龄死亡率。它是通过把特定年龄组的个体数量相对时间作图而得到的[5]。

本文使用非参数的方法绘制 Kaplan-Meier 曲线，以生存时间为横坐标，生存概率为纵坐标，绘制的总体生命函数曲线图如图 2 所示，在图下有三个参数，At risk：被调查的生命期限超过时间点的患者人数；Censored：生命期限小于等于时间点的未去世患者人数；Events：生命期限小于等于时间点的去世患者人数。整体曲线是生命函数曲线，随着时间变化患者的存活几率也在变化。

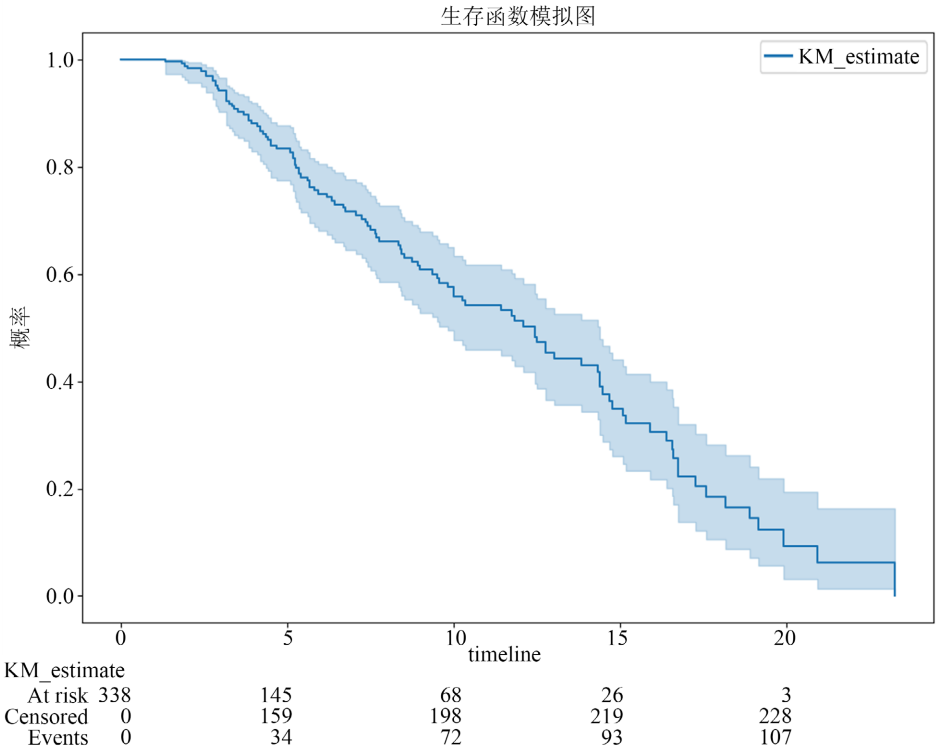

Figure 2. Survival analysis curve
图 2. 生存分析曲线

根据上图可以得出以下结论：

存活率从在 0~1 年内下降的最快；超过 20 年后的生存几率几乎为 0；有 26 个病患患癌 15 年以上；有 29 个患者患癌未超过 15 年还存活；有 93 个患者未患癌 15 年后死亡。

5.2. Cox 模型

用两个上述的 Cox 建模式相比，即为两个个体的风险函数比，称为风险比。当风险比大于 0 时，变量的增加将加大事件发生的概率，即死亡率加大；当风险比小于 0 时，变量的增加将减小事件发生的概率，即死亡率减小；当风险比等于 0 时，变量与事件的发生无关。

进行 Cox 回归模型的建立之前，先绘制了 Kaplan-Meier 生存曲线用来检测所有因变量与自变量之间存在的关系，再通过多因素分析，以确保结果更加精准[3]。

筛选变量时也利用逐步向前回归法，筛选出变量后的模型为：将死亡时间作为因变量，性别、年龄、肿瘤分期作为自变量的模型。

部分参数变量示例得出如下表 1 数据：

Table 1. Examples of variables and parameters
表 1. 变量和参数示例

参数 变量	Coef	Exp(cof)	Exp(coef) lower 95%	Exp(coef) upper 95%
Age	0.416	1.516	1.357	1.693
Stage	0.463	1.666	1.029	2.700
Sex	0.351	1.421	1.077	1.875

从结果分析来看，风险比体现为 exp(cof)，三个变量分别为年龄、肿瘤分期和性别，其中年龄和肿瘤分期在 5%的显著水平下显著，风险较大，性别的相关风险略小。

Table 2. Table of model evaluation parameters
表 2. 模型评价参数表

参数	参数值
Concordance	0.83
Partial AIC	1274.90
Log-likelihood ratio test	103.81
−Log2(p) of ll-ratio test	69.19

模型评价参数表如表 2 所示，Concordance 值为 0.83，证明模型拟合效果不错，继而用残差验证风险比例假设的准确性。

Table 3. Residual verification table
表 3. 残差验证表

变量	p
Age	0.6002
Stage	0.1916
Sex	0.4557
Global	0.5296

如表 3 所示，所有 p 值均 > 0.05，所建立的模型良好，没有违背风险比例假设，说明实验的真实性以及可靠性较好。

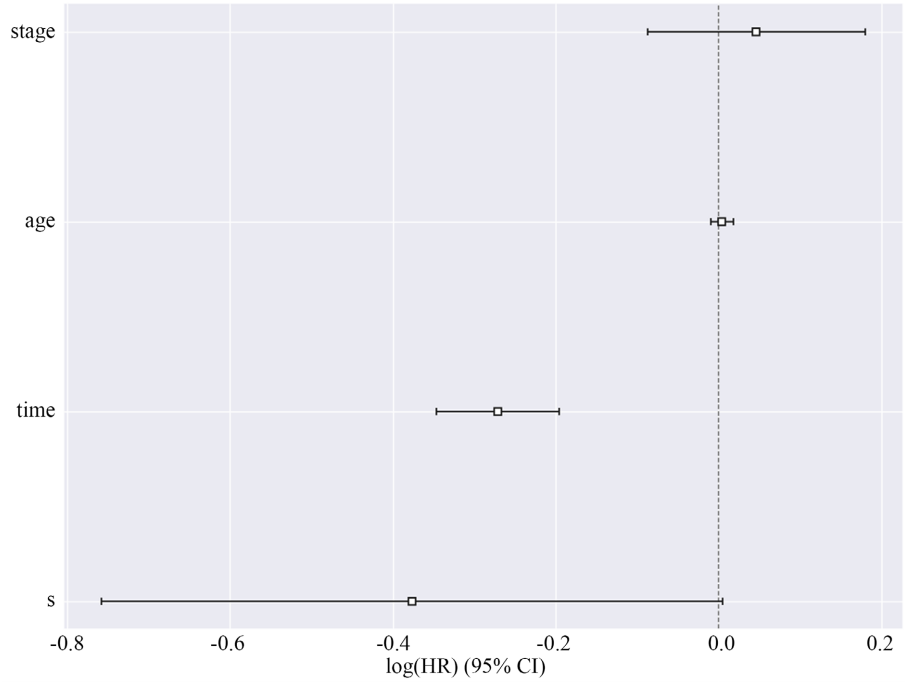

Figure 3. Important factors chart
图 3. 重要因素图

由图 3 可知，肿瘤分期和年龄的特征重要性在大于 0 的部分居多，所以主要加速病人死亡的影响因素有：肿瘤分期和年龄，最后挑选出对口腔鳞状癌症影响最大的因素排序为：肿瘤分期、年龄、性别。

6. 结语

对口腔鳞状癌症影响最大的因素为肿瘤分期，其中最严重的为 5 期，即患者患肿瘤的分期越靠近末期，死亡机率越大，且生存时间越短；5 期患者的存活率几乎为 0，存活时间大约在 2 年；其次为年龄，年龄越大，生存率越低，且生存时间越短；年龄在 67 岁以上的患者死于该癌症的死亡率最高，随着时间的增加，存活率降低；最后为性别，在收集到的患者信息中，男性死于该癌症的风险为女性的 1.42 倍，时间与该变量关系较小。

对于口腔鳞状癌症的治疗目前可能没有较好的救治方法，在此模型的基础上可以增加计算剩余价值以及潜在可提升的价值，以通过模型手段达到最好的预防和防治效果。

参考文献

[1] 医疗大数据存在的问题[EB/OL]. https://www.docin.com/p-1966055788.html, 2021-10-01.

[2] 李雪迎. 画说统计|生存分析之 Kaplan-Meier 曲线都告诉我们什么[EB/OL]. http://www.360doc.com/content/17/0626/11/6175644_666623573.shtml, 2021-10-01.

[3] Cui, E., Leroux, A., Smirnova, E. and Crainiceanu, C.M. (2021) Fast Univariate Inference for Longitudinal Functional Models. *Journal of Computational and Graphical Statistics*, 1-12.

[4] 王定坤, 杨杉. 基于 COX 比例风险模型分析心力衰竭影响因素[J]. 电脑知识与技术, 2021, 17(24): 33-35.

[5] 李清河, 江泽平. 白刺研究(Research on Plant Species of *Genus Nitraria L*)[M]. 北京: 中国林业出版社, 2011: 102.