

## 临床意义 - 预后 Cox 回归分析-分子[云]

Characteristics	Total(N)	HR(95% CI) Univariate analysis	P value Univariate analysis
FAM241B	82		0.607
Low	41	Reference	
High	41	1.225 (0.566 - 2.650)	0.607
TMEM37	82		0.319
Low	41	Reference	
High	41	1.489 (0.679 - 3.263)	0.320
UGT2B7	82		0.256
Low	41	Reference	
High	41	1.577 (0.712 - 3.494)	0.262

网址: <https://www.xiantao love>

更新时间: 2023.05.08

## 目录

基本概念 .....	3
应用场景 .....	4
分析流程: .....	4
主要结果/结果解读 .....	6
数据格式 .....	8
参数说明 .....	9
分子 .....	9
预后参数 .....	11
阈值控制 .....	11
数据处理 .....	12
结果说明 .....	13
主要结果: .....	13
补充结果: 变量情况统计表 .....	14
补充结果: 中位生存时间表 .....	15
补充结果: 单因素 cox 回归分析表 .....	16
方法学 .....	17
如何引用 .....	18
常见问题 .....	19

## 基本概念

- Cox 回归模型：又称为比例风险回归模型，是一种半参数回归模型。Cox 模型以生存结局和生存时间为因变量，分析众多自变量因素对生存期的影响

### ■ 数据要求

- ◆ 结局建议用数字编码（0/1，1/2），其中最好用 0 代表删失或者未发生事件，1 表发生事件

- ◆ 自变量（协变量）可以是数值或者分类变量。分类变量如果是含有等级的含义，则需要以等级资料纳入，需要设置参考组，其他组 and 这个参考组作对比；如果分类变量是无等级含义，一般是需要经过哑变量编码，但是经过哑变量编码后结果有可能不好解读，故无等级关系的分类变量也可以通过组合的方式形成二分类变量纳入。二分类的分类变量以等级或者非等级纳入的结果都是一致的（二分类分不分等级都一样）。数值变量可以直接以数值变量的形式纳入，亦可转换为等级资料或者二分类资料纳入

- 条件假设：观测值独立，风险比不随时间改变（比例风险假设）。（模块内默认是满足此条件）

- 对于回归模型的假设检验通常采用似然比检验、Wald 检验和记分检验




- PH 假设：比例风险 (Proportional hazards) 假定。Cox 模型应用的前提条件。基本假设为：协变量对生存率的影响不随时间的改变而改变，即风险比值  $h(t)/h_0(t)$  为固定值。而在实际进行生存分析的过程中，有些自变量对风险函数（事件发生概率）的影响会随时间的变化而变化，因此在构建 Cox 回归模型之前，必须对 PH 假定进行判定，只有 PH 假定得到满足时，Cox 回归模型的结果才有意义

- 中位生存时间（半数生存期）：即当累积生存率为 50%时所对应的生存时间，表示有且只有 50%的生病个体可以活过这个时间。只有当分组内最终累积生存率低于 50%才会有中位生存时间

## 应用场景

Cox 回归模块主要用于评估变量对于预后的影响，或者判断某个变量是否是独立预后因素（多因素中还有统计学意义）。一般在进行多因素 Cox 回归前，会先进行单因素 Cox 回归对每个变量逐个进行分析，将单因素有意义（ $p < 0.1$ ，这个一般不会设置 0.05）纳入到多因素中进行分析

## 分析流程：

云端数据  单因素 Cox 分析  (单因素分析 p 值是否满足设定的阈值)  
 多因素 Cox 分析

- 云端数据：提供预清洗好的云端数据，不同平台的云端数据集的分子和临床变量可能会有不同
  - 通过主要参数[分子]选择需要进行分析的分子(分子 ID)，云端数据中的分子
  - 通过主要参数[分子分组]将选择的分子进行分组，得到最终需要进行分析的数据（具体参数操作可看后面参数部分）

- 单因素 Cox 回归分析:
  - 构建预后 cox 回归模型: 选择好的数据进行 cox 模型构建
  - 通过模型得到模型所有变量的分析结果
- 多因素 Cox 回归分析:
  - **筛选变量:** 将单因素结果得到的 p 值, 没有达到参数“阈值控制”p 值大小的变量筛选出来, 不进行后续多因素 Cox 回归
  - 多因素 Cox 回归分析

## 主要结果/结果解读

Characteristics	Total(N)	HR(95% CI) Univariate analysis	P value Univariate analysis
FAM241B	82		0.607
Low	41	Reference	
High	41	1.225 (0.566 - 2.650)	0.607
TMEM37	82		0.319
Low	41	Reference	
High	41	1.489 (0.679 - 3.263)	0.320
UGT2B7	82		0.256
Low	41	Reference	
High	41	1.577 (0.712 - 3.494)	0.262

	A	B	C	D
1	Characteristics	Total(N)	HR(95% CI) Univariate analysis	P value Univariate analysis
2	FAM241B	82		0.607
3	Low	41	Reference	
4	High	41	1.225 (0.566 - 2.650)	0.607
5	TMEM37	82		0.319
6	Low	41	Reference	
7	High	41	1.489 (0.679 - 3.263)	0.320
8	UGT2B7	82		0.256
9	Low	41	Reference	
10	High	41	1.577 (0.712 - 3.494)	0.262

- Characteristic：变量以及分组
  - 如果变量是等级/分类变量，或者参数部分选择进行分组，则紧接其后的变量的分组，其中第一个分组为参考组
  - 如果变量是数值类型，或者参数部分选择数值纳入，则该变量在表格中只有有一行结果
- Total(N)：数量情况。对应变量总所选分组总的数量以及各组的数量，此样本数量为进行单因素分析时变量和对应分组的数目
  - 由于可能包含缺失信息，所以不同的变量之间的总数可能是不同。（可在参数中选择“进行单因素前先过滤缺失样本”，就能保证变量的总数是一致的）
  - 这一列在文章中并不是一定需要的，可以不提供

- HR(95% CI) Univariate analysis: 单因素分析得到的 HR (Hazard ratio, 风险比) 以及对应的置信区间。一般  $HR > 1$  说明变量是危险因素,  $HR < 1$  为保护因素
  - 如果是等级/分类变量的参考组, 则此分组 HR 为 Reference
  - 如果该行对应的是等级/分类变量的变量名 (非具体分组), 则不会有 HR 值
- P value Univariate analysis: 单因素分析得到的自变量对应的 p 值 (一般是满足  $< 0.1$  就会纳入到模型中)
  - 如果是等级/分类变量的参考组, 则此分组单因素 p 值为空
  - 如果该行对应的是等级/分类变量的变量名 (非具体分组), 则单因素 p 值为整个变量整体性检验的 p 值, 这个 p 值影响该变量是否纳入到多因素。(该 p 值在文章中不需要报告)
- HR(95% CI) Multivariate analysis: (只有变量满足进入多因素 Cox 模型的 p 值阈值才会有值) 多因素分析得到的 HR (Hazard ratio, 风险比) 以及对应的置信区间。一般  $HR > 1$  说明变量是危险因素,  $HR < 1$  为保护因素
- P value Multivariate analysis: (只有变量满足进入多因素 Cox 模型的 p 值阈值才会有值) 多因素分析得到的自变量对应的 p 值。当纳入一定的变量时, 多因素仍然有意义则说明该变量可能是独立预后因素。

## 数据格式

提供预清洗好的云端数据, 不同平台的云端数据集的分子和临床变量可能会有不同。如果有一些想要的临床变量不存在, 则可能是对应的数据集没有提供或者信息较少。

(此样本数据: 如下: )

数据参数

云端数据 ⓘ 食管鳞癌 / TCGA / TCGA-ESCC / RNAseq / STAR / TPM @过滤:去除正常+去除无临床信息 @处理:log2(...)





## 参数说明

(说明：标注了颜色的为常用参数。)

## 分子

分子

分子ID

FAM241B  
TMEM37  
UGT2B7

分子分组
0-50 vs 50-

- 分子 ID：选择云端数据中需要进行分析处理的分子/变量，一行一个 ID，可以是分子名，也可以是分子 ID，最多支持 40 个，如下：

分子

分子ID

FAM241B  
TMEM37  
UGT2B7

分子分组
0-50 vs 50-

Characteristics	Total(N)	HR(95% CI) Univariate analysis	P value Univariate analysis
FAM241B	82		0.607
Low	41	Reference	
High	41	1.225 (0.566 - 2.650)	0.607
TMEM37	82		0.319
Low	41	Reference	
High	41	1.489 (0.679 - 3.263)	0.320
UGT2B7	82		0.256
Low	41	Reference	
High	41	1.577 (0.712 - 3.494)	0.262

- **分子分组**：将选择的分子/变量进行相应的分组，后对各分组进行相关分析，

其中分组方式(0-50 vs 50-100 代表中位数分组)或者以数值纳入，如下：

分子

分子ID

FAM241B

TMEM37

UGT2B7

分子分组 0-50 vs 50-100

Characteristics	Total(N)	HR(95% CI) Univariate analysis	P value Univariate analysis
FAM241B	82		0.607
Low	41	Reference	
High	41	1.225 (0.566 - 2.650)	0.607
TMEM37	82		0.319
Low	41	Reference	
High	41	1.489 (0.679 - 3.263)	0.320
UGT2B7	82		0.256
Low	41	Reference	
High	41	1.577 (0.712 - 3.494)	0.262

分子

分子ID

FAM241B

TMEM37

UGT2B7

分子分组 数值纳入

Characteristics	Total(N)	HR(95% CI) Univariate analysis	P value Univariate analysis
FAM241B	82	1.278 (0.827 - 1.975)	0.269
TMEM37	82	1.115 (0.850 - 1.462)	0.432
UGT2B7	82	1.644 (0.657 - 4.112)	0.288

## 预后参数



- 预后类型：可选不同的预后类型。不同的数据集之间的预后类型可能不一样  
可以选择：

- OS[Overall Survival] (默认)：总体生存期
- DSS[Disease Specific Survival]：无病生存期
- PFI[Progress Free Interval]：无进展间隔

## 阈值控制



- p 值（单因素进入多因素）：可以控制变量是否进入到多因素 Cox 模型中，不同的数据集之间的预后类型可能不一样，常规阈值可选 0.1, 0.2，如果输入的是 1，则代表所有变量都纳入到多因素中
- 纳入多因素分析的 p 值阈值：默认为 0.1

## 数据处理



- 缺失值处理：可以选择对数据中缺失值进行处理
  - 默认为 单因素后多因素前处理变量缺失，表示在经过单因素分析之后，通过变量缺失处理在进行多因素分析
  - 还可以选择 单因素前统一处理缺失，则是在进行分析之前对全部的缺失值进行处理



## 结果说明

### 主要结果:

#### Cox分析-分子

- Cox回归: 比例回归风险模型, 用于预后资料的分析。条件假设: 观测值独立, 风险比不随时间改变 (比例风险假设)
- 预后类型: OS[Overall Survival]
- P值阈值: 0.1 (单因素变量均未达到p值阈值)

Characteristics	Total(N)	HR(95% CI) Univariate analysis	P value Univariate analysis
FAM241B	82		0.607
Low	41	Reference	
High	41	1.225 (0.566 - 2.650)	0.607
TMEM37	82		0.319
Low	41	Reference	
High	41	1.489 (0.679 - 3.263)	0.320
UGT2B7	82		0.256
Low	41	Reference	
High	41	1.577 (0.712 - 3.494)	0.262

Cox回归结果.xlsx

Cox回归结果.docx

- Characteristics: 变量以及分组
- Total(N): 数量情况。对应变量总所选分组总的数量以及各组的数量, 此样本数量为进行单因素分析时变量和对应分组的数目
- HR(95% CI) Univariate analysis: 单因素分析得到的 HR 值 以及对应的置信区间。其中Reference代表等级资料的参考组, 变量中其他组和其他组进行比较
- P value Univariate analysis: 单因素分析得到的自变量对应的 p 值 (一般是满足设定的p值阈值 就会纳入到多因素模型中)

- Characteristics: 变量以及分组
- Total(N): 数量情况。对应变量总所选分组总的数量以及各组的数量, 此样本数量为进行单因素分析时变量和对应分组的数目
- HR(95%CI) Univariate analysis: 单因素分析得到的 HR (Hazard ratio, 风险比) 以及对应的置信区间。其中 Reference 代表等级资料的参考组, 变量中其他组和其他组进行比较
- P value Univariate analysis: 单因素分析得到的自变量对应的 p 值 (一般是满足设定的 p 值阈值就会纳入到多因素模型中)
- HR(95% CI) Multivariate analysis: (只有变量满足进入多因素 Cox 模型的 p 值阈值才会有值) 多因素分析得到的 HR (Hazard ratio, 风险比) 以及对应的置信区间
- P value Multivariate analysis: (只有变量满足进入多因素 Cox 模型的 p 值阈值才会有值) 多因素分析得到的 p 值

## 补充结果：变量情况统计表

### 变量情况

各个变量识别出来的类型 以及 是否纳入 进行分析

变量	类型	分类数量	缺失数量	是否纳入分析	补充说明
event	数值变量	-	0	纳入	
time	数值变量	-	0	纳入	
FAM241B	分类变量	2	0	纳入	
TMEM37	分类变量	2	0	纳入	
UGT2B7	分类变量	2	0	纳入	

总样本数: 82

· 如果某个分类变量的分类 > 10, 将无法识别为分类变量/等级变量

· 如果变量的分组是以 0 1 2 此类进行编码, 如果分类数量 < 5, 会被识别为分类变量; 如果 > 5, 会被识别为数值变量

· 如果数据中含有无穷值, 无穷值会被当做缺失处理

补充说明: 单因素分析前, 会先去掉 结局和时间列 中的缺失的样本(时间或者结局缺失的样本是无法纳入进行分析的)

缺失处理策略: 单因素后多因素前处理变量缺失

这里提供变量情况统计表:

- 如果某个分类变量的分类 > 10, 将无法识别为分类变量/等级变量
- 如果变量的分组是以 0 1 2 此类进行编码, 如果分类数量 < 5, 会被识别为分类变量; 如果 > 5, 会被识别为数值变量
- 如果数据中含有无穷值, 无穷值会被当做缺失处理

补充说明:

- 单因素分析前, 会先去掉 结局和时间列 中的缺失的样本(时间或者结局缺失的样本是无法纳入进行分析的)
- 缺失处理策略: 单因素后多因素前处理变量缺失

## 补充结果：中位生存时间表

### 中位生存时间

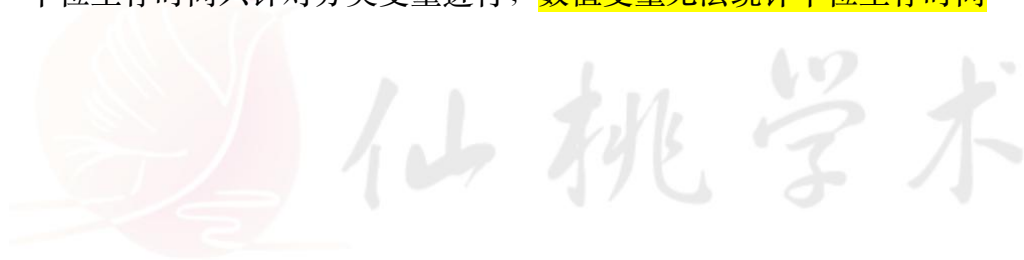
中位生存时间只针对分类变量进行，数值变量无法统计中位生存时间

FAM241B						
分组	数目	总事件数	总删失数	总删失比例	中位生存时间	中位生存时间置信区间
Low	41	13	28	68.3%	764	650-?
High	41	13	28	68.3%	1361	553-?
TMEM37						
分组	数目	总事件数	总删失数	总删失比例	中位生存时间	中位生存时间置信区间
Low	41	12	29	70.7%	1263	681-?
High	41	14	27	65.9%	763	553-?
UGT2B7						
分组	数目	总事件数	总删失数	总删失比例	中位生存时间	中位生存时间置信区间
Low	41	10	31	75.6%	1263	681-?
High	41	16	25	61.0%	763	553-?

备注：中位生存时间的置信区间如果有?，则代表 分组中样本较少 或者是 随访时间不足 或者是 预后相对较好无法计算出来对应的上限或者下限

这里提供分类变量中位生存时间表：

- 中位生存时间只针对分类变量进行，数值变量无法统计中位生存时间



## 补充结果：单因素 cox 回归分析表

单因素Cox					
变量	类型	数量	HR	置信区间	p值
FAM241B	等级变量	82			0.6074
Low		41	Reference		
High		41	1.225	0.566 - 2.650	0.6071
TMEM37	等级变量	82			0.3191
Low		41	Reference		
High		41	1.489	0.679 - 3.263	0.3202
UGT2B7	等级变量	82			0.2559
Low		41	Reference		
High		41	1.577	0.712 - 3.494	0.2620

单因素中满足  $p < 0.1$  就会纳入到多因素Cox回归中

这里提供单因素 cox 回归分析表





## 方法学

统计分析和可视化均在 R 4.2.1 版本中进行

涉及的 R 包: survival[3.4.0]

处理过程:

- (1) 使用 survival 包进行比例风险假设检验 并进行 Cox 回归分析
- (2) 变量筛选策略: 单因素中样本满足设定的  $p$  值阈值就会进入到多因素 Cox 中构建模型



## 如何引用

生信工具分析和可视化用的是 R 语言，可以直接写自己用 R 来进行分析和可视化即可，可以无需引用仙桃，如果想要引用仙桃，可以在致谢部分 (Acknowledge) 致谢仙桃学术 ([www.xiantao.love](http://www.xiantao love))。

方法学部分可以参考对应说明文本中的内容以及一些文献中的描述。



## 常见问题

### 1. 为什么不同变量的数量不同？

Characteristics	Total(N)	HR(95% CI) Univariate analysis	P value Univariate analysis	HR(95% CI) Multivariate analysis	I
Pathologic T stage	79		0.961		
T1	8	Reference			
T2	27	1.180 (0.303 - 4.588)	0.811		
T3&T4	44	1.058 (0.296 - 3.785)	0.931		
Pathologic N stage	78		0.063		
N0	46	Reference		Reference	
N1	26	1.960 (0.789 - 4.869)	0.147	1.960 (0.789 - 4.869)	
N2&N3	6	4.310 (1.301 - 14.279)	0.017	4.310 (1.301 - 14.279)	
ERBB2	82		0.407		
Low	41	Reference			
High	41	0.699 (0.300 - 1.625)	0.405		
ERBB3	82		0.229		

答：结果中的这个数量为进行单因素时的数量（如果是分组（而不是变量），则为对应分组的数量）。由于可能包含缺失信息，所以不同的变量之间的总数可能是不同的。（可在参数中选择“进行单因素前先过滤缺失样本”，就能保证变量的总数是一致的）。**这一列在文章中并不是一定需要的，可以不提供**

### 2. 为什么有一些变量在多因素中统计学数值为 NA？

答：有可能是这么几种情况：

- (1) 变量存在共线性
- (2) 去除任一变量信息缺失后，某个变量的某个分组变成了 0

### 3. 为什么多因素模型中某些变量值很大？或者为 Inf？

答：出现值很大的原因可能是因为纳入了缺失严重的变量，导致了某个变量的某些分组的数量变少了很多，最终导致数值异常了。需要检查是否纳入了信息缺失的变量。（可看说明文本中各个变量的数量）

#### 4. 为什么文章里面的多因素 Cox 模型没有常数？

答：由于多因素 Cox 模型是广义线性模型的一种，也是存在有常数项的。如果在文章中是有可能看到没有写这个常数项的情况，这种情况有这么几种可能：

（1）在 R 里面(print)输出 cox 模型的表是不带常数项的，只有变量和系数，所以就忽略了这个常数项的情况

（2）常数项本身在模型里面也不是关键的，因为是常数，不会对结果有什么影响

#### 5. 为什么文章里面的模型是放的多因素 p 值有意义的变量，而工具却给的是多因素纳入的变量？

答：首先可以明确的是，多因素模型中的自变量就是纳入到多因素模型的所有变量，包括进入多因素模型后 p 值没有意义的变量，肯定不是只要多因素 p 值有意义的变量才是多因素模型的变量

（1）多因素模型里面正是这些所有变量在模型里面经过变量之间和混杂因素分析后才得到的每个变量的校正后的情况

（2）如果是提取了多因素 p 值有意义的变量再构建一个新的多因素模型，那么这些变量的系数肯定不是用的之前的那个模型，肯定是来自一个这个新的模型，而且这些在上一个多因素中 p 值有意义的变量在新的多因素模型中未必还都是 p 值有意义的