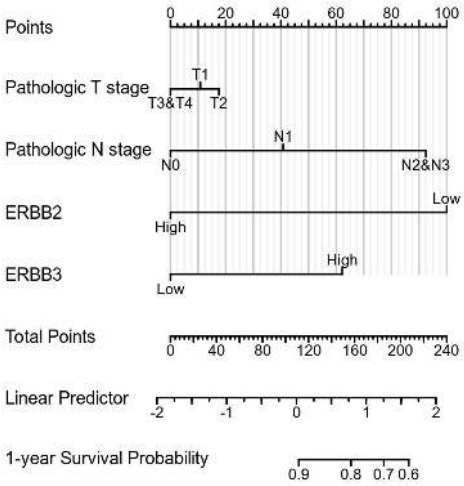
临床意义 - 预后 Nomogram 列线图[云]



网址：https://www.xiantao.love

更新时间：2023.03.06

目录

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，其他组和这个参考组作对比；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)/h0(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

- Nomogram 图（列线图/诺莫图）：在多因素回归分析基础上，通过设置标尺评分来表征多因素回归模型内各个变量的情况，最终计算出总的评分来进行预测事件发生的概率情况。列线图可以简单理解成 将模型可视化成一个预测事件发生的量表，用外部数据独立的个体在模型中各个指标的评分情况来预测该个体发生时间的概率

应用场景

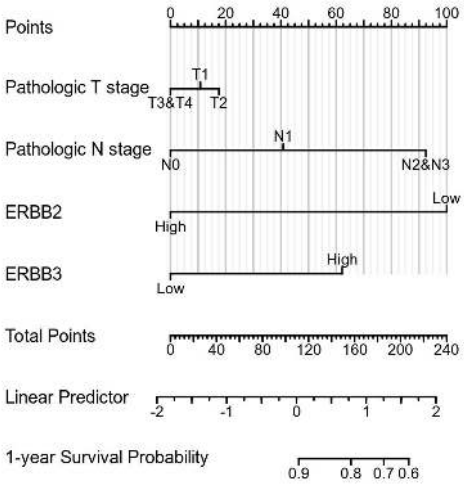
预后 Nomogram 图建立在多因素回归分析的基础上， 采用带有刻度的线段将多个预测指标进行整合，并按照一定的比例绘制在同一平面上，从而用以表达预测模型中各个预测变量之间的相互关系，得到的列线图可用于外部预后数据分析某个患者事件发生的概率情况。一般有用多因素 Cox 回归方法建立模型都会用 Nomogram 图进行可视化

分析流程：

云端数据 ⟶ 单因素多因素 Cox 分析 ⟶ Nomogram 列线图可视化

- 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
  - 通过特殊参数[变量]选择临床变量或者分子，可以进行输入搜索
  - 通过特殊参数[分组]将选择的临床变量或分子进行分组，得到最终需要进行分析的数据（具体参数操作操作可看后面参数部分）
- 单因素多因素 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
  - 进行单因素多因素 Cox 回归分析
- Nomogram 列线图可视化

结果解读



左侧标题文字：

- Points：表示模型中每个预测变量在不同分组/取值下所对应的单项评分情况，比如 Grade 分期中 1 组对应的单项评分为 32 分
- 第二行到 Total Points 的上一行（Pathologic T stage ~ ERBB3）：表示每一个上传的预测变量
  - 如果这个变量是数值变量，则认为该变量相同数值距离下的评分差值是一样的（多因素 Cox 回归中 HR 差值也是一样）。
  - 如果这个变量是等级变量，则认为该变量不同等级之间的评分差值是不同的，比如图中的 Pathologic T stage 变量，可以看到 T1、 T2、 T3&T4 之间的差异是不同的
    - 如果一个变量是二分类变量，则这个变量以分类变量或者以等级资料纳入的结果（评分/ HR 值）都是一样的
- Total Points：表示所有变量取值后对应的单项分数加起来合计的总得分，可用于（计算模型总的 y 值（Linear Predictor 行）对应的总的得分），也可用

于（计算某个总的得分下对应的事件发生概率（x-year Survival Probability 行））。比如总的得分是 100 分，此时对应的 1 年生存率的概率是 0.76 左右

- Linear Predictor：表示线性预测值，多因素模型总的 y 值情况，可以比对到 Total Points 行以确定对应不同 y 值对应的总的得分，也可以比对到下面对应的事件发生概率（x-year Survival Probability 行）
- x-year Survival Probability：表示在预测时间范围所对应的预测概率

右侧坐标：

- 代表左侧标题文字对应的坐标刻度与取值范围

补充：

- 此图不用于评估模型的好坏，评估模型的好坏是多个方面，其中一个可以看一致性指数（C-index）（在 0.5-1 之间，越大越好），在补充结果的多因素 Cox 回归模型中：

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：)

参数说明

（说明：标注了颜色的为常用参数。）

特殊参数

- 变量：第一个框为自变量变量，可以键盘输入进行搜索，下拉选择，可以搜索分子。
  - 输入“临床”关键字，可以搜到对应云端数据录入的所有临床数据
    - 临床变量有分类类型 和 数值类型
  - 直接输入分子名，也可以搜索分子

- 第一个框后面的 + 和 −，代表增加或者删去 一行临床变量。
- 第二个框以及以后的框，为选择对应的变量的分组组合。

- 如果是分类类型的变量，则第二个框为参考组，后面的框对应的分组组合和这个参考组进行对比。分类变量的分组中后的中括号为对应在临床资料中的分组的数量(未匹配对应的平台，仅仅只是临床数据中的，有 可能会和最终的结果不符 （因为存在有临床资料没有对应平台的检测结果））
- 如果是数值类型，则框内只有数值可以选择

- 设置某个变量的分组组合时：对应的分组的第 1 个框除了有 T1 外，还有一个变 量，这两个变量组合成一个整体作为参考组。点击×可以删除这个变量。下拉选 项框内显示无数据，说明这个临床变量的所有分组都已经选上了
- 当一个临床变量还有分组没有选上时，先机下拉框会显示这个还没有选上的分组

所有变量的选择 以及 对应的分组组合、参考组设置都是可以自定义选择的，请尽量保存所有的分组组合的数量比较平均和合适。 请注意查看每个变量对应的

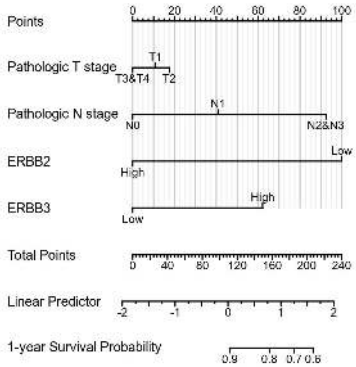
分组的数量，如果某个变量含有的分组对应的数量很 少，则说明这个变量很可能信息缺失严重，建议是不纳入

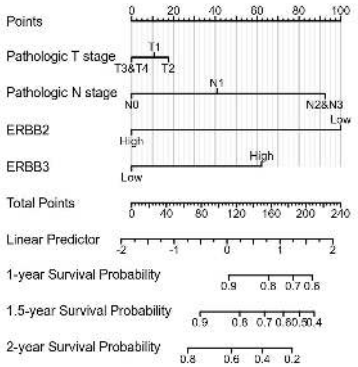
预后参数

- 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

预测时间

- 时间 1：第一个时间点，数字，单位根据选择的预测时间单位
- 时间 2：第二个时间点，数字，单位根据选择的预测时间单位
- 时间 3：第三个时间点，数字，单位根据选择的预测时间单位
- 时间单位：可以选择上传数据预测时间列的单位，默认以年为单位，可以选择月、天为单位
  - 如下所示：左侧为只设置一个预测时间的时候，右侧为设置了多个时间的情况



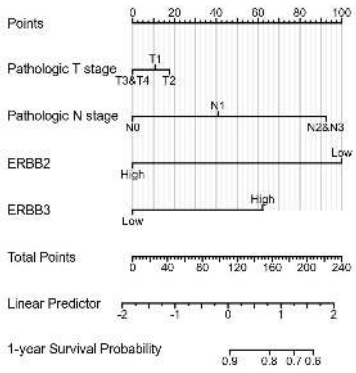


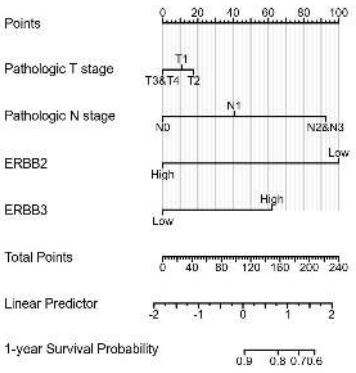
数据处理

- 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

坐标轴

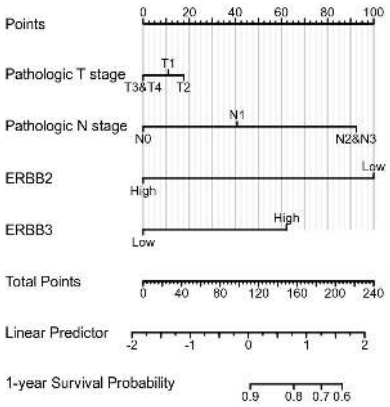
- 左侧标题与坐标轴中间的距离：可以调整左侧标题与坐标轴中间的距离，默认为 0.6（如下左图），还可以选择 0.9（如下右图）0.2/0.4/...

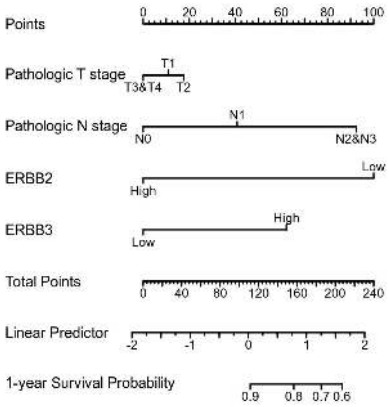




风格

- 坐标轴竖线：可以选择是否显示坐标轴竖线（坐标轴对应刻度线）
  - 默认不显示（如下左图）
  - 还可以选择显示（如下右图）





- 文字大小：控制整体文字大小，默认为 7pt

图片

- 宽度：图片横向长度，单位为 cm
- 高度：图片纵向长度，单位为 cm
- 字体：可以选择图片中文字的字体

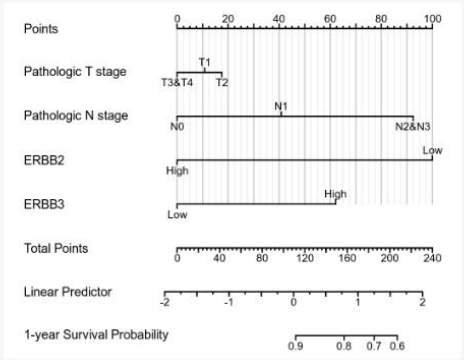
结果说明

主要结果

预后Nomogram-云

预后Nomogram: 建立在多因素回归分析的基础上，采用带有刻度的线段将多个预测指标进行整合，并按照一定的比例绘制在同一平面上，从而用以表达预测模型中各个预测变量之间的相互关系

预后类型: OS[Overall Survival]



预后列线图.pdf　预后列线图.tiff　预后列线图.pptx　Riskscore.xlsx

左侧标题文字：

· Points: 表示每个预测变量在不同取值下所对应的单项分数

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	
Pathologic T stage	分类变量	3	3	纳入	
Pathologic N stage	分类变量	3	4	纳入	
ERBB2	分类变量	2	0	纳入	
ERBB3	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

Pathologic T stage

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
T1	8	3	5	62.5%	855	557-?
T2	27	8	19	70.4%	763	650-?
T3&T4	44	12	32	72.7%	1263	553-?

Pathologic N stage

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
N0	46	9	37	80.4%	1458	764-?
N1	26	10	16	61.5%	763	557-?
N2&N3	6	4	2	33.3%	471.5	247-?

ERBB2

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	11	30	73.2%	764	553-?
High	41	15	26	63.4%	1263	681-?

ERBB3

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	557-?

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
Pathologic T stage	等级变量	79			0.9615
T1		8	Reference		
T2		27	1.180	0.303 - 4.588	0.8113
T3&T4		44	1.058	0.296 - 3.785	0.9308
Pathologic N stage	等级变量	78			0.0634
N0		46	Reference		
N1		26	1.960	0.789 - 4.869	0.1470
N2&N3		6	4.310	1.301 - 14.279	0.0168
ERBB2	等级变量	82			0.4066
Low		41	Reference		
High		41	0.699	0.300 - 1.625	0.4049
ERBB3	等级变量	82			0.2289
Low		41	Reference		

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：

- 表中所有变量都会纳入到多因素中

补充结果：多因素 cox 回归分析表

多因素Cox

变量	系数β	HR	置信区间	p值
Pathologic N stage				
N0		Reference		
N1	0.67311	1.960	0.789 - 4.869	0.1470
N2&N3	1.461	4.310	1.301 - 14.279	0.0168

模型常数/截距(Intercept): 0

原始数据一共有82个, 变量信息缺失的样本有4个, 最终纳入的样本数: 78

备注: 如果出现纳入了多因素但是对应的统计量为空的情况，说明(1)这个变量在去除变量信息缺失后某个分类数目过少(只有1个或者0个)或者是(2)存在严重共线性导致这个变量导致没办法计算。

备注: 当如果多因素中出现HR异常大或者异常小时，说明这个变量的这个分类数量过少或者是存在共线性问题导致

·(分类/等级)变量(非分组)对应的单因素p值为对应变量单因素模型全局性检验的p值，该变量是否纳入取决于此p值

▵ 模型全局性统计检验情况：

·· 一致性(Concordance, C-index): 0.624(0.555-0.693)

·· Likelihood ratio test= 5.52 on 2 df, p=0.0634

·· Wald test = 6.06 on 2 df, p=0.0483

·· Score (logrank) test = 6.81 on 2 df, p=0.0333

这里提供多因素 cox 回归分析表：

- 如果出现纳入了多因素但是对应的统计量为空的情况，说明(1)这个变量在去除变量信息缺失后某个分类数目过少(只有 1 个或者 0 个)或者是(2)存在严重共线性导致这个变量导致没办法计算
- 当多因素中出现 HR 异常大或者异常小时，说明这个变量的这个分类数量过少或者是存在共线性问题导致
- (分类/等级)变量(非分组)对应的单因素 p 值为对应变量单因素模型全局性检验的 p 值，该变量是否纳入取决于此 p 值

补充结果：PH 比例风险假设检验表

比例风险假设(PH)

Cox回归应用的前提是要求自变量满足等比例风险假设(P > 0.05)，即自变量的风险不会随着时间改变而改变，若不满足，则不适合用Cox回归进行检验。
这里只对多因素模型以及纳入的变量进行ph假设检验
备注: (1)单个变量直接PH假设和在模型里面这个变量的PH假设的结果是不一样的; (2)同一份数据不同Cox模型中同一个变量的PH假设的结果也是不一样的

变量	统计量(卡方值)	自由度(df)	p值
Pathologic N stage	0.26234	2	0.8771
GLOBAL	0.26234	2	0.8771

如果全局(GLOBAL)满足p > 0.05，可以认为多因素模型满足比例风险假设

这里提供 PH 假设检验表：

- Cox 回归应用的前提是要求自变量满足等比例风险假设(P ＞ 0.05)，即自变量的风险不会随着时间改变而改变，若不满足，则不适合用 Cox 回归进行检验
- 这里只对多因素模型以及纳入的变量进行 ph 假设检验
  - 单个变量直接 PH 假设和在模型里面这个变量的 PH 假设的结果是不一样的
  - 同一份数据不同 Cox 模型中同一个变量的 PH 假设的结果也是不一样的

补充结果：方差膨胀因子表

方差膨胀因子(VIF)

方差膨胀因子可用于分析模型中的变量是否存在多重共线性问题

变量	类型	VIF
Pathologic N stage	等级变量	
N0		Reference
N1		1.1681
N2&N3		1.1681

一般认为，当0 < VIF < 10，不存在多重共线性(补充: 也有认为VIF > 4就存在多重共线性); 当10 ≤ VIF < 100，存在较强的多重共线性; 当VIF >= 100或者是出现NaN，多重共线性非常严重

这里提供方差膨胀因子表：

- 方差膨胀因子可用于分析模型中的变量是否存在多重共线性问题
  - 当 1<VIF<10，不存在或存在较轻的多重共线性
  - 当 10<=VIF<100，存在较强的多重共线性
  - 当 VIF>=100 或者是出现 NaN，，多重共线性非常严重

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]，rms 包（6.3-0）

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) 使用 rms 包进行列线图分析与可视化

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）。

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. 为什么我设置了等级的顺序跟结果的顺序不一样?

答：模型中的等级资料不是设置了什么等级顺序就会在图中出现什么等级顺序，图中 变量的等级顺序是由多因素分析校正后得到的。

2. 为什么所有的变量都进行了单因素分析和多因素分析?

答：一般情况下，是通过对变量进行单因素分析，在对其结果进行筛选，选择单因素变量统计学 p 值大于 0.1（常用）作为筛选条件，满足则对这些变量进行多因素分析，不满足的这不分析。但是不能避免有时候上传的数据所有变量都不满足（或条件太过于苛刻）导致无法分析，所以就不进行筛选，直接通过单因素和多因素分析进行计较就行。

3. Nomogram 是基于 cox 回归分析的，能将 p 小于 0.05 的变量纳入到 nomogram 的绘制吗?

答：目前不支持将 p 值小于 0.05 或其他条件的变量纳入进行 nomogram 分析，而是将所有的变量作为一个整体的模型进行分析。如果有需要，可以自己过滤后再上传数据。

4．风险比例假设 ph < 0.05 该怎么办？是不是就不能进行 cox 回归分析了？还是说在模块中会进行矫正？

答：主要看整个模型和单个变量模型中的 ph 情况：

① 当整个模型(全局) ph < 0.05 时，所有的变量将不会被纳入进行多因素分析；

② 当单个变量模型 ph < 0.05，但整个模型(全局) ph > 0.05 时，则不会影响到整个模型的分析