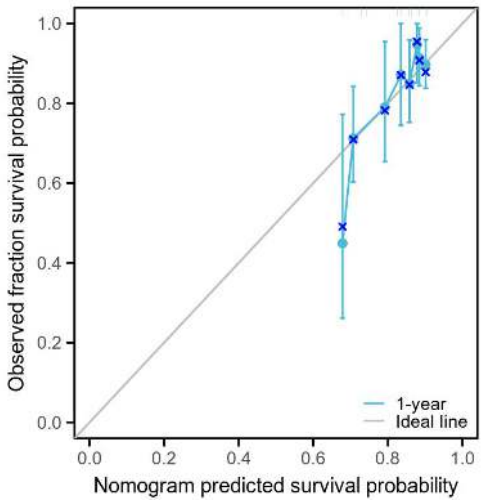
临床意义 - 预后 Calibration 校准曲线



网址：https://www.xiantao.love

更新时间：2023.03.06

目录

基本概念 . . . 3
应用场景 . . . 4
分析流程 . . . 5
结果解读 . . . 6
数据格式 . . . 7
参数说明 . . . 8
特殊参数 . . . 8
预后参数 . . . 10
预测时间 . . . 11
数据处理 . . . 12
分析参数 . . . 13
线 . . . 14
点 . . . 15
图注 . . . 17
坐标轴 . . . 18
风格 . . . 18
图片 . . . 19
结果说明 . . . 20
主要结果 . . . 20
补充结果：变量情况统计表 . . . 21
补充结果：中位生存时间表 . . . 22
补充结果：单因素 cox 回归分析表 . . . 23
补充结果：多因素 cox 回归分析表 . . . 24
补充结果：PH 比例风险假设检验表 . . . 25
补充结果：方差膨胀因子表 . . . 26
方法学 . . . 27
如何引用 . . . 28
常见问题 . . . 29

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

- Calibration 校准曲线：分为诊断类型和预后类型两类。在 Calibration 图中，通过在图中绘制不同情况下实际概率和模型预测的概率的拟合情况，判断模型对实际结果预测效果的评估。简单来说， Calibration 图只要看线是否能够很好的拟合到对角线上

应用场景

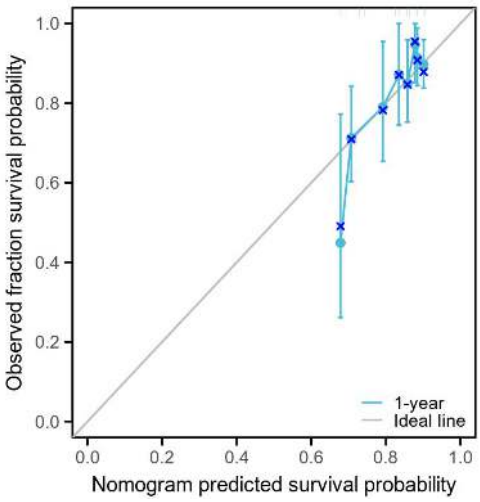
预后 Calibration 图：描绘模型在不同时间点对应的预测概率与实际概率之间的差异情况，主要用于对 Cox 回归方法建立的模型与实际情况的拟合分析。一般在 Nomogram 之后也都会带上一个 Calibration 图来说明模型的准度情况

分析流程：

云端数据 ⟶ 单因素多因素 Cox 分析 ⟶ Calibration 校准曲线可视化

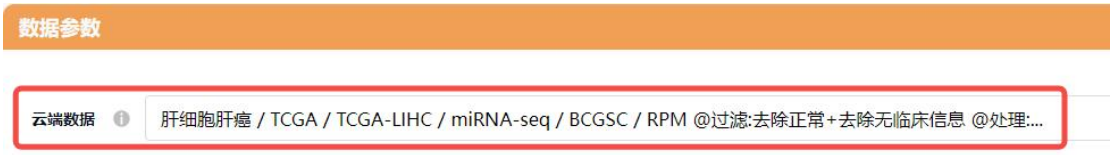
- 云端数据：提供预清洗好的云端数据，不同平台的云端数据集的分子和临床变量可能会有不同
  - 通过特殊参数[变量]选择临床变量或者分子，可以进行输入搜索
  - 通过特殊参数[分组]将选择的临床变量或分子进行分组，得到最终需要进行分析的数据（具体参数操作操作可看后面参数部分）
- 单因素多因素 Cox 回归分析：
  - 构建预后 cox 回归模型：选择好的数据进行 cox 模型构建
  - 通过模型得到模型所有变量的分析结果
  - 进行单因素多因素 Cox 回归分析
- Calibration 校准曲线可视化

结果解读



- 横坐标为模型预测的生存概率
- 纵坐标为实际观测到的生存概率
- 图中一共有 2 条线，每条线分别代表：模型预测 1 年（1-year）生存情况与实际情况的对比，以及最理想的线（对角线，灰色）；模型预测的线越贴近对角线说明拟合情况越好（如何简单判断结果好坏：可以就看每根线跟对角线的拟合情况，越贴近对角线说明模型拟合效果越好）
- 校准曲线上的点代表模型预测的生存概率和实际观测到的生存概率情况（类似 Nomogram 中的最下部分中不同得分对应的概率）
- 校准曲线上的点对应的竖线代表该位置的置信区间
- 校准曲线上蓝色的叉代表每个点经过分层 Kaplan-Meier 校正后的结果
- 顶部的竖线代表具体样本对应的生存概率（生存率的分布情况），越密集说明越多样本的生存概率在这个概率

数据格式


数据参数
云端数据
肝细胞肝癌 / TCGA / TCGA-LIHC / miRNA-seq / BCGSC / RPM @过滤:去除正常+去除无临床信息 @处理:...

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

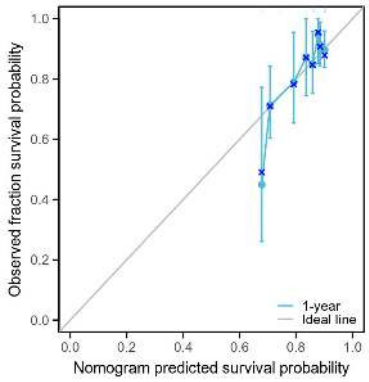
分组的数量，如果某个变量含有的分组对应的数量很少，则说明这个变量很可能信息缺失严重，建议是不纳入

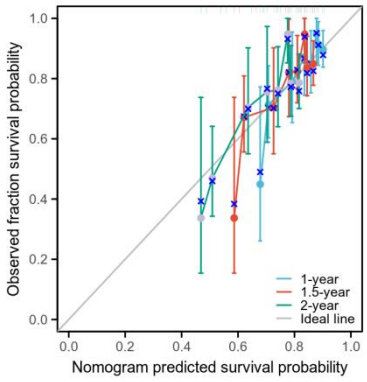
预后参数

- 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

分析参数

- 每次重复抽样的样本量：可以修改每次重复抽样的样本量
  - 每次抽样的样本量不能超过样本(行数)的一半
  - 每次抽样的样本量不能少于 20 个
- 抽样次数：可以选择抽样次数，默认为 200 次，还可以选择 400、600、800 次

线

- 颜色：可以修改不同预测时间点对应的校准曲线的颜色
- 线条类型：可以选择校准曲线的线条类型，默认为实线，还可以选择虚线类型
- 线条粗细：可以选择校准曲线的线条粗细，默认为 0.75pt
- 不透明度：可以修改校准曲线的不透明度，默认为 1，表示完全不透明

点

- 填充色：可以修改校准曲线上各点（模型预测的生存概率和实际观测到的生存概率情况）的填充色
- 描边颜色：可以修改校准曲线上各点的描边颜色
- 样式：可以选择校准曲线上各点的形状/样式，默认为圆形，还可以选择正方形、菱形、三角形、倒三角形
- 大小：可以修改校准曲线上各点的大小，默认为 1
- 不透明度：可以修改校准曲线上各点的不透明度，默认为 1，表示完全不透明

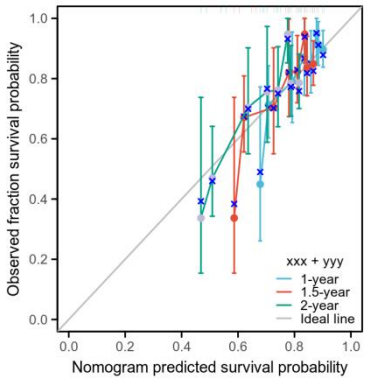
标题文本

- 大标题：大标题文本
- x 轴标题：x 轴标题文本
- y 轴标题：y 轴标题文本

补充：在要换行的中间插入\n。如果需要上标，可以用两个英文输入法下的大括号括住，比如 {{2}}；如果需要下标，可以用两个英文输入法下的中括号括住，比如 [[2]]

图注

- 是否展示：可以选择是否展示图注信息，默认展示（如左图），也可以不展示（如右图）
- 图注标题：可以修改图注对应的标题内容，如果需要换行可以在需要换行的位置插入“\n”，如下图：



坐标轴

- y 轴范围：可以控制 y 轴范围，需要提供 2 个值来控制范围。形如 0.1，0.3（最小值不能低于-0.5，最大值不能大于 1.5，如果调整过大可能会无作用）
- x 轴范围：可以控制 x 轴范围，需要提供 2 个值来控制范围。形如 0.1，0.3（最小值不能低于-0.2，最大值不能大于 1.2，如果调整过大可能会无作用）

风格

- 外框：是否添加外框，默认添加
- 网格：是否添加网格，默认不添加
- 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

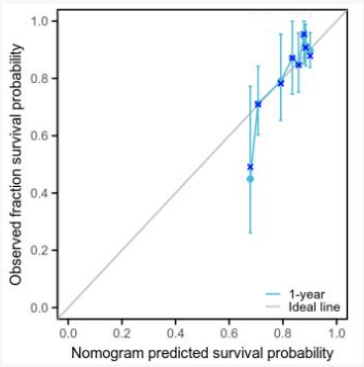
结果说明

主要结果

预后Calibration-云

预后Calibration: 描绘模型在不同时间点对应的预测概率与实际概率之间的差异情况

预后类型: OS[Overall Survival]



预后校准曲线.pdf　预后校准曲线.tiff　预后校准曲线.pptx

· 横坐标为模型预测的生存概率；纵坐标为实际观测到的生存概率

· 每条线代表对应各时间点的生存情况与实际情况的对比、以及最理想的线（对角线：灰色）；越贴近对角线说明拟合情况越好

· 每条线的点代表模型预测的生存概率和实际观测到的生存概率情况（类似 nomogram中的最下部分中不同得分对应的概率）

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	-	1	纳入	
Pathologic T stage	分类变量	3	3	纳入	
Pathologic N stage	分类变量(单分类)	1	115	不纳入	
hsa-let-7a-3p	分类变量	2	0	纳入	
hsa-let-7b-3p	分类变量	2	0	纳入	

总样本数: 375

· 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)(当前存在有1个样本缺失时间或者结局)

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	184	45	139	75.5%	2456	2116-?
T2	94	30	64	68.1%	1852	1149-?
T3&T4	93	52	41	44.1%	660	558-1490

hsa-let-7a-3p

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	186	59	127	68.3%	2486	1624-?
High	188	69	119	63.3%	1560	1149-2532

hsa-let-7b-3p

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	186	61	125	67.2%	2131	1490-?
High	188	67	121	64.4%	1624	1149-2542

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
Pathologic T stage	等级变量	371			7.35e-07
T1		184	Reference		
T2		94	1.464	0.922 - 2.325	0.1063
T3&T4		93	2.999	2.008 - 4.481	8.12e-08
hsa-let-7a-3p	等级变量	374			0.2047
Low		186	Reference		
High		188	1.253	0.884 - 1.777	0.2055
hsa-let-7b-3p	等级变量	374			0.4870
Low		186	Reference		
High		188	1.131	0.799 - 1.603	0.4873

表中所有变量都会纳入到多因素中

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 T stage				
T1		Reference		
T2	0.38533	1.470	0.925 - 2.337	0.1034
T3&T4	1.0943	2.987	1.995 - 4.472	1.07e-07
hsa-let-7a-3p				
Low		Reference		
High	0.20546	1.228	0.835 - 1.807	0.2971
hsa-let-7b-3p				
Low		Reference		
High	-0.057726	0.944	0.640 - 1.392	0.7707

多因素Cox.xlsx

模型常数/截距(Intercept): 0

原始数据一共有374个, 变量信息缺失的样本有3个, 最终纳入的样本数: 371

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:

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 T stage	0.00058853	2	0.9997
hsa-let-7a-3p	1.0678	1	0.3014
hsa-let-7b-3p	0.34758	1	0.5555
GLOBAL	1.1003	4	0.8942

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 T stage	等级变量	
T1		Reference
T2		1.2786
T3&T4		1.2861
hsa-let-7a-3p	等级变量	
Low		Reference
High		1.2136
hsa-let-7b-3p	等级变量	
Low		Reference
High		1.2393

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 Calibration 分析与可视化

补充说明：

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. 为什么不能改参数提交后直接查看结果?

答：如果每选择一次分析参数进行修改的时候，如预测时间、或数据处理、或分析参数修改，因为分析过程需要花费一定时间（比如当数据量比较大的时候），所以每修改一次都会在后台进行，完成后在“历史记录”查看结果。如果是可视化参数之类的话，可以在历史记录找到记录并选择修改

2. 如何理解结果（图）

答：简单来看，就是看每根线跟对角线的拟合情况，越贴近对角线说明模型拟合效果 越好。具体详细结果解读可以看本文档的结果解读部分

3. 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。