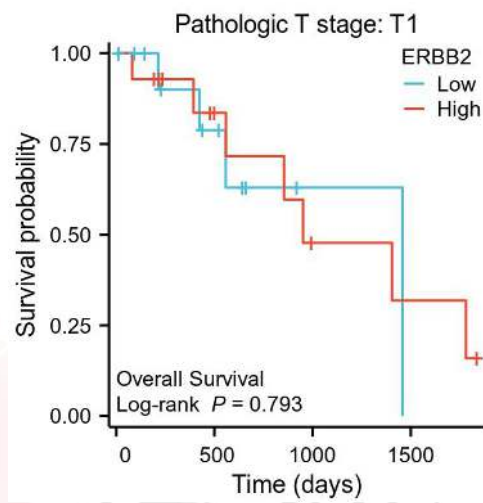


临床意义 - 亚组 KM 曲线[云]



网址: <https://www.xiantao.love>



更新时间: 2023.03.09

目录

基本概念	3
应用场景	5
结果解读	6
云端数据	7
参数说明	8
特殊参数	8
分子	8
临床变量&亚组	8
主要参数	9
预后参数	9
分组	10
统计	11
类型	12
线	13
删失数据	13
置信区间	14
风险表格	14
标题文本	15
图注(Legend)	16
坐标轴	17
风格	18
图片	18
结果说明	19
主要结果	19
补充结果	20
方法学	22
如何引用	23
常见问题	24

基本概念

- 生存曲线（也称 Kaplan-Meier 曲线）：可以描述各组患者的生存状况或者各组实验动物的存活情况
- 生存时间：从规定的起始事件开始到失效事件出现所持续的时间。对于失访者，是失访前最后一次随访的时间。
- 终点事件/终点结局：医学研究中可以是患者死亡，也可以是疾病的发生、某种治疗的反应、疾病的复发等。与之对应的起始事件可以是疾病的确诊、某种治疗的开始等。
- 删失/截尾（Censoring）：由于某些原因在随访中并没有观测到终点结局而不知道确切的生存时间，此部分数据即删失数据。常见原因有失访、患者退出试验、事件发生是由于非研究性疾病（如研究病人发生脑卒中后的生存时间，结果病人因为车祸死亡）、研究结束时研究对象仍未发生失效事件。删失数据的生存时间为起始事件到截尾点所经历的时间。
- 中位生存时间（Median Survival Time）：中位生存时间又称半数生存期，表示恰好一半个体未发生终点事件的时间，生存曲线上纵轴 50% 对应的时间。如果删失或者截尾数据较多，预后较好，则可能无法计算得到对应的中位生存时间。
- 生存分析的方法：
 - 非参数法：寿命表、Kaplan-Meier(乘积极限法 Product limit method/检验方法：对数秩 (Log rank)、Breslow、Tarone-Ware)等
 - 半参数法：Cox 回归（需要满足 Cox 比例风险假设）
 - 参数法

- 备注：Log rank 方法 和 Cox 方法都有很广泛的应用，一般两者选其一即可。
- 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。
- 风险比（Hazard Ratio, HR）：两个风险率的比值。当 $HR > 1$ 时，说明相对于对照组，实验组（研究对象）是一个危险因素；当 $HR < 1$ 时，说明相对于对照组，实验组（研究对象）是一个保护因素；当 $HR = 1$ 时，说明研究对象对生存时间不起作用。
- 统计类型：（可根据需要选择，一般选择展示累积生存）
 - 累积生存：时间段内生存概率的累积。
 - 累积风险：时间段内风险概率的累积
 - 累积事件：时间段内事件发生率。
- Overall Survival (OS)，总体生存期，指结局指标是死亡时间，这个死亡是任何原因导致的死亡都算进去，只关心是否死亡，不关心因为何种原因死亡。这个指标能比较方便的记录，因为患者死亡的日期确认没有困难。只要研究结果显示生存有提高，就可认为是临床又获益。但所需要的随访的时间较长。
- Disease Free Survival (DFS)，无病生存期，指经过治疗后未发现肿瘤，结局指标为疾病复发或死亡，同样不需要关心死亡原因。这一指标是临床获益的重要反映，随访时间可以缩短，因为增加了疾病复发这一节点。没有复发

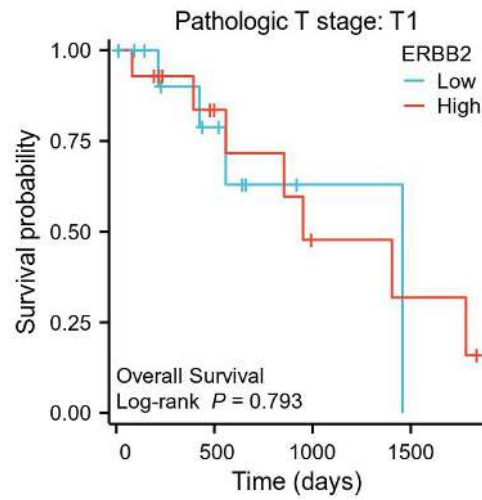
或没有死亡同样可以反映临床获益。这里也涉及到无疾病复发的一个定义，因此在临床资料纳入上比较困难。

- Disease-specific survival (DSS)疾病特异生存期： 结局指标为由特定疾病导致的死亡。如果不是特定疾病导致的则不计入结局指标。
- Progression-Free interval (PFI)无进展间隔： 从初次治疗的随机分组日期到疾病复发时间。

应用场景

生存曲线主要用于描述 受试或者研究对象 在一段时间内发生事件（存活 or 死亡 / 是否复发等）的情况。

结果解读



- 生存曲线的横坐标是观察时间，纵坐标一般是生存率。曲线上的每一个点代表了在该时间点上病人的生存率。
- 在坐标轴 (0, 1) 点上，由于才开始随访 (X 轴为 0)，此时没有患者死亡，所以两组患者的生存率都是 1 (100%)。
- 曲线上的“+”用于标识删失数据。曲线下降越缓慢，预后越好。

云端数据

提供预清洗好的云端数据，不同云端数据集的预后类型可能会有一定的差别。

（该样本数据：如下：）

数据参数

云端数据 ⓘ

食管癌 / TCGA / TCGA-ESCA / RNAseq / STAR / TPM @过滤:去除正常+去除无临床信息 @处理:log2(v...

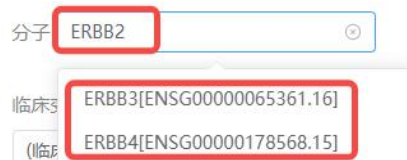


参数说明

(说明：标注了颜色的为常用参数。)

特殊参数

分子



- 分子：选择的云端数据中需要进行分析的变量/分子/基因，可以通过键盘输入内容进行搜索，[搜索出来的结果和当前所选择的云端数据有关](#)

临床变量&亚组



- 临床变量：选择清洗好的临床变量，可以进行输入搜索
- 亚组：选择清洗好的临床变量以及对应分组，[不要把所有变量都选上，都选上则等于对整体的分析而不是某个亚组](#)（[通过临床变量中的亚组，可以比较在亚组中 ERBB2 高低组之间的预后情况](#)）

主要参数

预后参数

预后参数

预后类型

OS[Overall S

时间单位

日

- **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。
- **时间单位**: 可以根据需要选择单位(影响可视化的 x 轴), 单位转换比例: 30.5 天=1 月 | 365 天=1 年



分组

➤ **分组**：可以选择分组的信息

- 其中 0-50 vs 50-100 是中位数分组。0-33 vs 66-100 是以数值按从小到大进行排序，从前往后的取前 33%作为低，取后 33%为高。其余分组以此类推。*p* 值最小分组为 *survminer* 包提供的一种方法，即将数值变量的不同数值作为 *cut-off*，每个数值均尝试作为分组的 *cut-off* 值，将每次尝试得到的 *p* 值进行排序，得到的最小的 *p* 值对应的 *cut-off* 值作为分组的 *cut-off*。这个类似于 *KMplot* 网站中的 *best-separate* 分割。（建议是先尝试中位数分组，其次是尝试 *p* 值最小。）

统计

统计

统计方法
Logrank检验

标注位置
左下

标注颜色

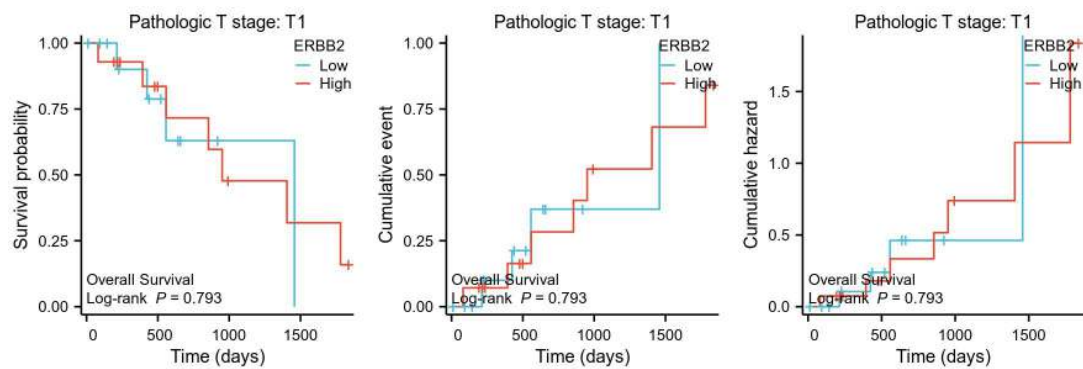
- **统计方法**: 在生存分析中 Kaplan-Meier(K-M)法可以估计生存概率, Log-rank 检验可以比较两条或多条生存曲线, 用 Cox 比例风险回归模型可以分析结局转归的影响因素。Cox 比例风险回归模型用于调查患者和一个或多个预测变量的存活时间之间的关联回归模型, 并且也能够同时评估多个因素对生存的影响。(两种方法均有很广的使用, Cox 需要满足 ph 假设)
- **标注位置**: 可以选择统计学标注的文字显示在图片的位置,
- **标注颜色**: 可以修改统计学标注文字的颜色

类型

类型

类型
生存概率(Su

- 类型：可以选择生存概率、累积事件、累积风险三种类型。最常用的为生存概率。



线



- **颜色**: 生存曲线对应的颜色，有几条线就会取几个颜色。这里的颜色受全局参数影响。
- **样式**: 可以选择实线或者虚线。
- **粗细**: 线的粗细

删失数据



- **展示**: 是否展示删失数据，图中标注了 的就是删失数据，一般默认展示
- **颜色**: 删失标注的颜色，默认是和线同色，可以根据需要进行修改
- **大小**: 的大小

置信区间

置信区间

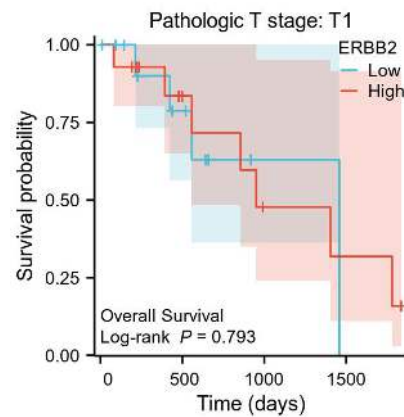
展示

样式

条带

不透明度

0.2



- 展示：是否展示每个分组每个时间点累积生存率的置信区间
- 样式：可选条带和虚线
- 透明度：设置置信区间的透明度，0 为完全透明，1 为完全不透明。

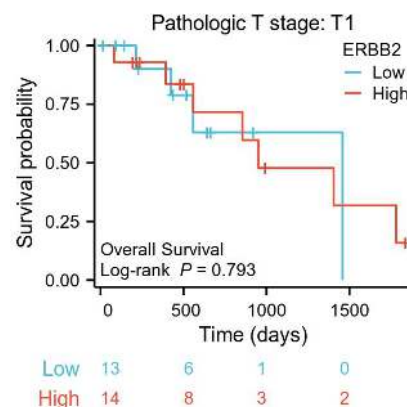
风险表格

风险表格

展示

边框

网格



风险表格记录了各个时间点上还在随访的人数

- 展示：是否展示风险表格，默认是不展示。开启后如右图下部分所示
- 外框：风险表格部分是否有外框，只有开启展示才有作用

- 网格：风险表格部分是否有网格，只有开启展示才有作用

标题文本

标题

大标题

大标题内容

x轴标题

x轴标题内容

y轴标题

y轴标题内容

- 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(Legend)

图注

▼

是否展示

☒

图注标题

图注标题内容

图注位置

默认

▼

- 展示：是否展示图注
- 图注位置：可选（图中）右上、右下、左上、左下，（图外）右、上，默认为右上。
- 图注标题：可以添加图注标题



坐标轴

坐标轴
▼

x轴范围+刻度
逗号隔开

y轴范围+刻度
逗号隔开

- X 轴范围+刻度：（注意：x 轴对应的范围必须是要在数据的随访时间内，可以根据需要修改成想要的时间刻度或者是修改范围）
 - 如果只是想要修改范围，可以只输入两个范围值，比如 0, 500。如果是想要同时修改范围和刻度，可以输入范围+刻度，比如 0, 100, 200, 300, 400, 500, 500.
- Y 轴范围+刻度：（注意：范围的修改不可以过大或者过小）
 - 如果只是想要修改范围，可以只输入两个范围值，比如 0.2, 1。如果是想要同时修改范围和刻度，可以输入范围+刻度，比如 0, 0, 0.1, 0.2, 0.5, 0.5.

风格

风格

边框

网格

文字大小 7pt

- 外框：是否添加主图外框
- 网格：是否添加网格
- 文字大小：控制整体文字大小，默认为 7pt

图片

图片

宽度 (cm) 5

高度 (cm) 5

字体 Arial

- 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

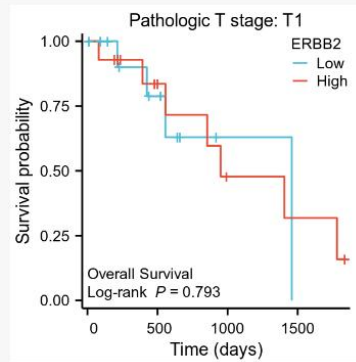
结果说明

主要结果

亚组KM

亚组KM图: 比较亚组(Pathologic T stage: T1)中ERBB2高低组之间的预后情况

统计方法: Logrank检验



亚组KM图.pdf

亚组KM图.tiff

亚组KM图.pptx

核心源码

补充说明: Cox回归需要满足比例风险假设(PH假设), 默认选择logrank检验



补充结果

统计描述

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	13	4	9	69.2%	1458	557-?
High	14	7	7	50.0%	951	558-?

累积生存率.xlsx

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

这里提供统计描述的表格, 包含删失情况和中位生存时间统计情况 (注意: 如果数据预后情况较好, 则对应的中位生存时间可能计算不出来, 包括中位生存时间置信区间)

这里提供每个分组每个时间点的累积生存率情况表格.xlsx 下载 (一个 sheet 代表一个分组情况): (可以查看某个时间节点的累积生存率, 比如 1 年、5 年)

	A	B	C	D	E	F	G
1	time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
2	213	10	1	0.9	0.09486833	0.732011643	1
3	424	8	1	0.7875	0.134032995	0.564125232	1
4	557	5	1	0.63	0.177038131	0.36319673	1
5	1458	1	1	0	#NUM!		

- time: 上传数据的每个时间
- n.risk: 每个时间点对应的人数
- n.event: 每个时间点上发生事件的人数
- survival: 当前时间点的累积生存率。
- std.err: 当前时间点的累积生存率的标准误
- lower 95% CI: 当前时间点的累积生存率的 95%置信区间下限
- upper 95% CI: 当前时间点的累积生存率的 95%置信区间上限

比例风险假设(PH)

Cox回归应用的前提是要求自变量满足比例风险假设($P > 0.05$)，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

Logrank检验没有要求满足比例风险假设，当不满足比例风险假设时可以临时选用Logrank检验，但是最严谨的是采用RMST(Restricted mean survival time)方法，当前模块无法兼容采用RMST方法

统计量(卡方值)	自由度	p值
0.17138	1	0.6789

如果p值小于0.05，则说明变量不满足比例风险假设，此时可以选用Logrank检验

这里提供风险比例假设检验情况，可以根据这个判断是否选择 cox 回归方法

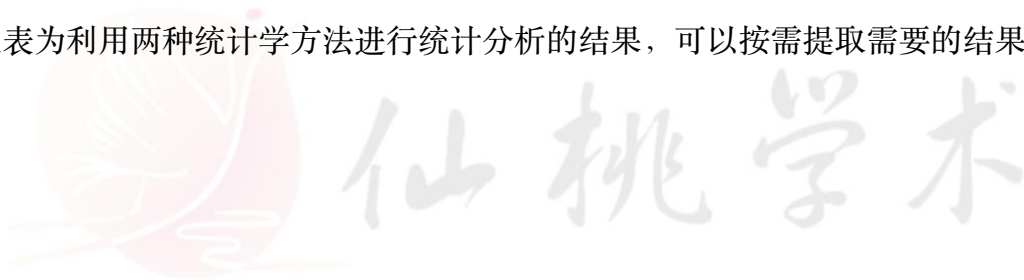
统计分析

同时提供Log-rank和Cox回归的检验结果，Cox回归应用需要满足风险比例假设(PH假设)

方法	统计量	HR	置信区间	p值
Log-rank	0.068538	0.855	0.243 - 3.011	0.7935
Cox回归	0.067639	0.841	0.230 - 3.076	0.7937

参考组(Reference): Low

此表为利用两种统计学方法进行统计分析的结果，可以按需提取需要的结果。



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：survminer 包（用于可视化），survival 包（用于生存资料的统计分析）

处理过程：

(1) 使用 survival 包进行比例风险假设检验 并 进行拟合生存回归，结果用 survminer 包以及 ggplot2 包进行可视化

(2) 如果选用了最佳分组方法(best)，则对应使用 survminer 包中 surv_cutpoint 函数进行最佳分组 cut-off 筛选



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. 为什么工具出来的结果 跟 GEPIA 或者 KMplot 数据库的不一样?

答:

① **表达数据可能不同**: 工具用的表达数据是直接 from TCGA 最新下载, GEPIA 未说明数据是什么时候从哪里下载的, KMplot 也未详细说明。数据过滤情况 GEPIA 和 KMplot 也未有进行说明。**不同的数据**、**数据格式**以及**样本例数**, 均会影响分组后两组的具体情况。

② **预后数据可能不同**: 工具使用的预后数据是从一篇 CELL 上整理好的预后数据, GEPIA 和 KMplot 未有说明。不同的预后数据也会对结果有很大的影响。

③ **统计方法**: 统计方法都是成熟的, 不存在方法学上的不同导致的。

另外, 下载区提供了“[分析数据.xlsx](#)”的下载(高级版), 对应的就是可视化的数据。同时, 方法学的最后部分(高级版)也提供了相应云端数据的原始数据, 也可以根据需要进行下载。

2. 在别的数据库上看到一个分子的趋势 跟 工具做出来的不一样?

答:

即便是同一个分子同一个疾病, 不同数据集得到的结果都可能会有差别, 甚至是存在趋势相反的情况。因为预后本身就是一个多因素综合作用的结果, 不同数据集之间**混杂因素**太多, 如果在这么多混在因素的情况下单看 1-2 个分子的表达, 难免是有可能出现 趋势不同或者相反的情况的。

所以, 如果只是单纯想要拿一些结果来充实自己的研究, 那么只放满足自己想要的趋势的数据即可。

3. 我有一个分子是高表达的，预后却提示是抑癌的？

答：

表达和预后是两个维度的内容，**表达和预后并不存在关联**，所以分子高表达而预后提示低表达组预后差，这种情况也是存在的。疾病的发生与否、是否进展、是否恶化而死亡都是很复杂的事件，都有可能是很多因素共同作用的结果，单个基因的作用是相对有限的。而且，KM 图只是单因素的结果，可能多因素后这个因素就没有作用了也说不准。

所以，当遇到这种情况时，建议换一个分子或者换一个数据集再看看是否存在想要的趋势就好。

4. 在云端数据框内看到的例数、分析时候的例数不同，这个是什么情况？

答：

云端数据的例数一般是对应组学所有的例数，分析时候可能会有剔除样本，具体需要看说明文本中对于数据的处理情况的说明。

有一些云端数据是存在有一个临床样本检测了多次的情况，去除重复检测的样本，能够降低同一份临床数据被同时纳入而影响结果。虽然存在有重复检测，但是一般这些重复检测的样本的数量很少。同样，也有一些云端数据对应的临床数据是只有临床数据，而没有对应的平台（组学）的检测的，一般这些没有检测的数据都是会被剔除的。

5. 应该选择 log rank 方法 还是 Cox 方法？

答：

两种方法均可，Cox 回归需要满足 PH 假设，当不满足 PH 假设时，建议是选择 Log rank 方法。

6. 如何查看某个分组 1 年生存率（或者某年生存率）？

答：

可以在补充结果下载“累积生存率.xlsx”表格：

统计描述						
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	349	93	256	73.4%	2835	1933-3571
High	349	179	170	48.7%	828	742-1152

[累积生存率.xlsx](#)

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
 · 数据中时间列或者结局列中含有缺失的数量: 1

	A	B	C	D	E	F	G
1	time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
2	213	10	1	0.9	0.09486833	0.732011643	1
3	424	8	1	0.7875	0.134032995	0.564125232	1
4	557	5	1	0.63	0.177038131	0.36319673	1
5	1458	1	1	0	#NUM!		

这里面一个 sheet 代表一个分组的情况，time 的单位与上次数据的时间的单位一致，这里示例数据的单位为天

1. 如何理解 p 值最小分组，如何在方法学中进行说明？

答：

p 值最小分组为 survminer 包提供的一种方法，即将数值变量的不同数值作为 cut-off，每个数值均尝试作为分组的 cut-off 值，将每次尝试得到的 p 值进行排序，得到的最小的 p 值对应的 cut-off 值作为分组的 cut-off。这个类似于 KMplot 网站中的 best-separate 分割。

方法学中在写作工具中限定“方法学”搜索“surv_cutpoint”：

方法 ▾

surv_cutpoint

× 筛选

检索

PTEN TPM optimal cutpoint to separate continuous variables was identified using the **surv_cutpoint** function from the R package survminer.

使用来自R包survminer的**surv_cutpoint**函数确定PTEN TPM分离连续变量的最佳切点。

所在语境 ▾

方法

来源

论著

2020

IF 11.5

Nucleic ACids Re...

写作区

或者以下这些例子（请勿直接复制）：

➤ <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7841432/>

- The “surv_cutpoint” function in the survminer R package was performed to search the best split by verifying all potential cut points.

➤ <https://www.frontiersin.org/articles/10.3389/fmolb.2021.608369/full>

- The “surv_cutpoint” function in the survminer package was used to determine the optimal cut-off value of the risk score.