

临床意义 - [云]基线资料表

变量	Low expression of ERBB2	High expression of ERBB2	p值	统计量	方法
n	41	41			
Pathologic T stage, n (%)			0.062	5.5456	Yates' correction
T1	1 (1.3%)	7 (8.9%)			
T2	15 (19%)	12 (15.2%)			
T3&T4	25 (31.6%)	19 (24.1%)			
Age, n (%)			0.248	1.3338	Chisq test
<= 60	29 (35.4%)	24 (29.3%)			
> 60	12 (14.6%)	17 (20.7%)			
BMI, median (IQR)	22.058 (20.196, 23.243)	22.204 (20.303, 25.934)	0.110		Wilcoxon

网址: <https://www.xiantao.love>

更新时间: 2023.03.14

目录

基本概念	3
应用场景	3
分析流程	4
主要结果	5
云端数据	7
参数说明	8
特殊参数	8
数据处理	9
表格	10
变量	10
结果说明	11
主要结果	11
补充结果	11
方法学	13
如何引用	14
常见问题	15

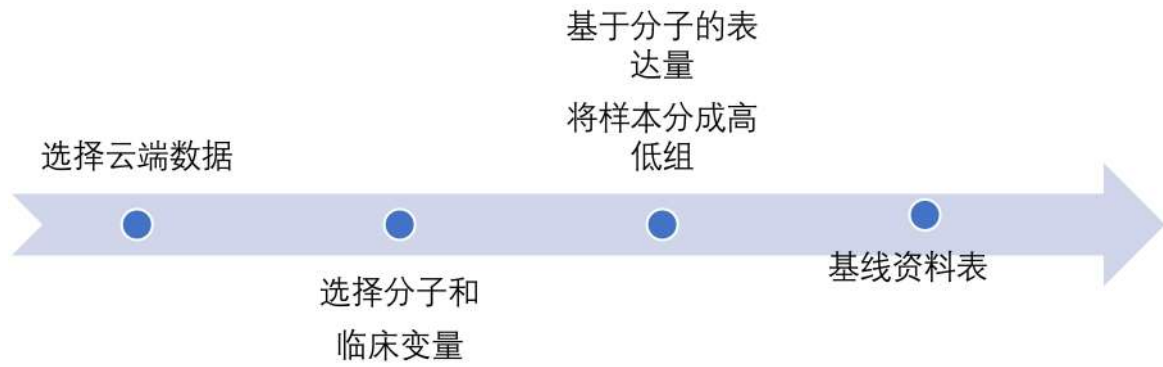
基本概念

- 基线资料表：展示每一格研究对象的基本信息情况。
- 卡方检验：比较不同组之间构成比（分类型资料）是否有差异，要求每个格子（level）中的理论频数 T 均大于 5 或 $1 < T < 5$ 的格子数不超过总格子数的 $1/5$ 。
- Fisher 精确检验：当不满足卡方检验的要求时，可以使用 Fisher 精确检验。
- T 检验：用于两组之间（数值型资料）的比较，需要满足两组正态性和方差齐性的要求。
- Welch t' test：又称为不等方差检验，即当两组仅满足正态性而不满足方差齐性的要求时，可以选择用该方法进行两组的比较。
- Wilcoxon rank sum test，也叫 Mann-Whitney U test（曼-惠特尼 U 检验），或者 Wilcoxon-Mann-Whitney test。秩和检验是一个非参的假设检验方法，一般用于两组不满足正态性的情况。
- One-way ANOVA：单因素方差分析，当比较组大于 2 时，可以使用该方法。
- Kruskal-Wallis test：克鲁斯卡尔-沃利斯检验，又称“K-W 检验”、“H 检验”等。本质也是一种秩和检验，用以检验多组（两组以上）不满足正态性的情况。

应用场景

基线资料表，是基于公共数据（云端数据）根据所选单个分子在所有样本中的中位数，将样本分为高低表达组，联合分组来评估不同的分组之间不同临床变量的构成比是否有差别。

分析流程



主要结果

列联表

	A	B	C	D	E	F
1	characteristics	Low expression of ERBB2	High expression of ERBB2	pvalue	statistic	method
2	n	41	41			
3	Pathologic T stage, n (%)			0.0624871632116509	5.54558826389699	Yates' correction
4	T1	1 (1.3%)	7 (8.9%)			
5	T2	15 (19%)	12 (15.2%)			
6	T3&T4	25 (31.6%)	19 (24.1%)			
7	Age, n (%)			0.248136154697613	1.33376707872479	Chisq test
8	<= 60	29 (35.4%)	24 (29.3%)			
9	> 60	12 (14.6%)	17 (20.7%)			
10	BMI, median (IQR)	22.058 (20.196, 23.243)	22.204 (20.303, 25.934)	0.109701172177491		Wilcoxon

- characteristics: 临床变量以及对应的分组
- Low expression of ERBB2: (低表达组) 对应的临床变量的构成比。当变量为分类型时，为不同水平 level 的计数和百分比；当变量为数值型时，或者是均值±标准差（样本数≤5000 且满足正态性或样本数>5000 时），或者是中位数(上下四分位)（样本数≤5000 且不满足正态性时）
- High expression of ERBB2: (高表达组) 对应的临床变量的构成比。当变量为分类型时，为不同水平 level 的计数和百分比；当变量为数值型时，或者是均值±标准差（样本数≤5000 且满足正态性或样本数>5000 时），或者是中位数(上下四分位)（样本数≤5000 且不满足正态性时）
- pvalue: 对应的列联表或者两组数值比较的统计学 p 值结果
- statistic: 统计量，只有卡方检验、t 检验以及 ANOVA 相关检验才会有，Fisher 精确检验是没有统计量的
- method: 所使用的统计学方法

纯基线资料表

	A	B
1	characteristics	overall
2	Pathologic T stage, n (%)	
3	T1	8 (10.1%)
4	T2	27 (34.2%)
5	T3&T4	44 (55.7%)
6	Age, n (%)	
7	<= 60	53 (64.6%)
8	> 60	29 (35.4%)
9	BMI, median (IQR)	22.175 (20.213, 24.687)

- characteristics: 临床变量以及对应的分组
- overall: 对应的临床变量的统计描述。当变量为分类型时，为不同水平 level 的计数和百分比；当变量为数值型时，或者是均值±标准差（样本数≤5000 且满足正态性或样本数>5000 时），或者是中位数(上下四分位)（样本数≤5000 且不满足正态性时）



云端数据

数据参数
重置参数

云端数据

食管鳞癌 / TCGA / TCGA-ESCC / RNAseq / STAR / TPM @过滤:去除正常 @处理:log2(value+1)

↓
↓

疾病名/来源/数据集/平台/分析流程/数据格式
@数据处理方式

云端数据 选择疾病
过滤数据: 默认 去除正常+去除无临床信息
×

疾病 请选择 ▼

数据过滤: 去除正常+去除无临床信息 ▼

数据格式: log2(value+1) ▼

	疾病系统	疾病名	疾病英文	来源	获取时间	数据集	平台	Wo
<input checked="" type="checkbox"/>	食管	食管鳞癌	Esophagus squamous cell carcinoma	TCGA	202208	TCGA-ESCC	RNAseq	STA
<input type="checkbox"/>	食管	食管癌	Esophageal carcinoma	TCGA	202208	TCGA-ESCA	RNAseq	STA
<input type="checkbox"/>	食管	食管癌	Esophageal carcinoma	TCGA	202208	TCGA-ESCA	RNAseq	STA
<input type="checkbox"/>	食管	食管腺癌	Esophagus adenocarcinoma	TCGA	202208	TCGA-ESAD	RNAseq	STA

共 115 条 上一页 1 2 3 4 5 6 ... 12 下一页

ⓘ 只有合适这个模块的云端数据才会展示

确认

本模块提供预清洗好的云端数据，不同平台的云端数据集的分子可能会有不同。注意查看当前数据参数选中的云端数据。



参数说明

(说明：标注了颜色的为常用参数。)

特殊参数

- 分组变量：下拉框将列出对应所选数据集分子，可以输入关键字搜索分子，基因 symbol 或 Ensembl ID，**只能选单个分析**。

- (临床)变量：下拉框将列出对应所选数据集的临床变量，**+** **-** 加减号可修改变量。选中变量后，**右侧**可选关联的分类信息，如 Pathologic_T_stage 对应 T1-T4 分类。

- **分组**：在变量对应的分类中自定义比较分组。   加减号修改分组，一个框内的分类组可以合成一个组，如 T3 和 T4 分类作为一组等等。**注意**，选择的分类（单个组）中样本数需要大于等于 3，且分组数大于等于 2，才能进行统计分析。根据具体情况可以自由选择参考组的分组。

数据处理



- **缺失值处理**：缺失处理是在**开始统计前**统一处理还是不处理。(如果想要保证所有的变量的总和加起来都是一个值，可以选择去除任一变量缺失的样本，但是这么操作需要关注变量的缺失情况，如果缺失很多，则最终会留下来的样本会少)

表格

- **表格类型**：可选列联表、纯基线资料表。
- **列联表百分比统计**：列联表中的分类变量的百分比统计方式，可选总数、按列、按行和无，默认以总数。只有列联表才起作用。

变量

- **强制正态的数值变量**：影响数值变量的展示方式以及对应的统计检验方法的选择。当通过经验判断该变量应该为正态分布（进行 t 检验）时，可以选择对应变量（程序自动返回选项，只有数值变量中选择了变量才会起作用）。此处选择后，对应的数值变量的汇总模式会更换成 均值±标准差
- **强制卡方的分类变量**：影响分类变量的统计检验方法的选择。当通过经验判断该变量应该进行卡方检验时，可以选择对应变量（程序自动返回选项，只有分类变量中选择了变量才会起作用）

结果说明

主要结果

主要结果		补充结果	方法学	保存结果	下载整份报告
基线资料表-云					
· 基线资料表: 用来展示研究对象的基本信息情况					
变量	Low expression of ERBB2	High expression of ERBB2	p值	统计量	方法
n	41	41			
Pathologic T stage, n (%)			0.062	5.5456	Yates' correction
T1	1 (1.3%)	7 (8.9%)			
T2	15 (19%)	12 (15.2%)			
T3&T4	25 (31.6%)	19 (24.1%)			
Age, n (%)			0.248	1.3338	Chisq test
<= 60	29 (35.4%)	24 (29.3%)			
> 60	12 (14.6%)	17 (20.7%)			
BMI, median (IQR)	22.058 (20.196, 23.243)	22.204 (20.303, 25.934)	0.110		Wilcoxon
基线资料表.xlsx 基线资料表.docx					
· 列表中包含变量以及分组、各组别统计描述、对应统计检验p值, 统计量以及对应统计检验方法					

主要结果格式为表格格式, 提供 [xlsx](#) 和 [docx](#) 格式下载。

补充结果

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
ERBB2	分类变量	2	0	纳入	
Pathologic T stage	分类变量	3	3	纳入	
Age	分类变量	2	0	纳入	
BMI	数值变量	-	4	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 缺失处理策略: 不统一处理变量缺失					

这里提供变量情况统计的表格, 包含数据类型、缺失情况、是否纳入分析 (纳入规则见 [数据格式](#)) 和补充说明。

Pathologic T stage-理论频数表

用于评估分类变量适合用什么统计检验方法

Pathologic T stage	Low expression of ERBB2	High expression of ERBB2
T1	1 (4.2)	7 (3.8)
T2	15 (14)	12 (13)
T3&T4	25 (22.8)	19 (21.2)

Pathologic T stage中存在level满足 $5 > \text{理论频数} \geq 1$ 且 总样本数 ≥ 40 的条件, 建议选用连续校正卡方检验(Yates' correction)。(备注: 括号内为各个level的理论频数)

分类变量（本例为 Pathologic T stage 分类）会提供对应的理论频数情况，以及给出选择统计方法的理由。

Age-理论频数表

用于评估分类变量适合用什么统计检验方法

Age	Low expression of ERBB2	High expression of ERBB2
≤ 60	29 (26.5)	24 (26.5)
> 60	12 (14.5)	17 (14.5)

Age中所有level满足 理论频数均 ≥ 5 且 总样本数 ≥ 40 的条件, 建议选用卡方检验(Chisq test)。(备注: 括号内为各个level的理论频数)

分类变量（本例为 Age 分类）会提供对应的理论频数情况，以及给出选择统计方法的理由。

BMI-正态性&方差齐性检验表

用于评估连续变量适合用什么统计检验方法

正态性检验 (Shapiro-Wilk Normality Test)				
分组	变量	自由度(df)	统计量	p值
Low expression of ERBB2	BMI	40	0.97362	0.4649
High expression of ERBB2	BMI	38	0.87899	0.0007
方差齐性检验 (Levene's test(Base on Mean))				
变量	自由度1(df1)	自由度2(df2)	统计量	p值
BMI	1	76	7.7727	0.0067

BMI不满足正态分布($P < 0.05$), 两组组间比较建议采用Wilcoxon检验。

数值变量（本例为 BMI 数值）会提供对应的正态性检验和方差齐性检验的结果，以及给出选择统计方法的理由。

方法学

所有分析和可视化均在 R 4.2.1 中进行

涉及的 R 包: stats



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. 为什么不同的临床变量的总数会不同?

答:

因为数据集可能会存在有缺失数据, 缺失数据在变量情况表中进行展示, 如果缺失值不在分析前统一处理, 则可能会存在有一些临床变量的总数和总的样本数对应不上的情况。变量最终是否纳入分析也是一个需要关注的问题。如果想要总数一样, 可以在参数中选择在分析前统一处理缺失。

2. 为什么结果中没有统计值?

答:

当变量不满足构成列联表条件或者表格类型选择列联表的, 均无统计检验相关的数据。反之, 根据分组内数据情况自动选择并生成统计检验结果, 具体统计方法的选择及给出的理由可参考结果中的 [补充说明](#)。

3. 在云端数据框内看到的例数、选择临床变量的数目 以及 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, 选择临床变量的数目为没有去除重复样本的例数, 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。

本模块, 可选 [去除正常](#) + [去除无临床信息](#)、[去除正常](#) + [去除无临床信息](#) + [去除重复](#)。

4. 为什么设置了分组信息，不显示统计检验结果？

答：

原因可能有下：

- ① 分组列的分类数目，可能因为其他任一变量的缺失过多，导致分组变成单分类（一组）。
- ② 在分组存在且满足条件时，任一分类型变量，如果存在有 level 的理论频数 < 1 占比大于百分之 20 的，则无法判断所使用的统计检验方法。
- ③ 在分组存在且满足条件时，任一数值型变量，如果存在有任一分组内的样本数小于 3 个的，则不做统计检验，且对应分组的统计描述缺失。

如果发现没有组间比较的统计检验结果时，可以先检查 补充结果 中的 变量情况 表，查看是否纳入分析和缺失数量的情况，尝试删除对应变量再进行分析。