临床意义 - 基线资料表

变量	Group1	Group2	p值	统计量	方法
n	209	130			
T.stage, n (%)			0.574	1.9903	Yates' correction
T2	59 (19%)	39 (12.6%)			
T3	96 (31%)	63 (20.3%)			
T4	31 (10%)	19 (6.1%)			
T1	3 (1%)	0 (0%)			
N.stage, n (%)			0.140	5.4789	Yates' correction
N2	43 (14.3%)	18 (6%)			
N0	117 (38.9%)	77 (25.6%)			
N1	19 (6.3%)	21 (7%)			
N3	4 (1.3%)	2 (0.7%)			
M.stage, n (%)			0.386	0.75109	Chisq test
M0	96 (54.9%)	65 (37.1%)			
M1	10 (5.7%)	4 (2.3%)			
Pathologic.stage, n (%)			0.706	1.3973	Yates' correction
Stage IV	71 (20.9%)	42 (12.4%)			
Stage III	69 (20.4%)	45 (13.3%)			
Stage II	67 (19.8%)	43 (12.7%)			
Stage I	2 (0.6%)	0 (0%)			
Age, median (IQR)	68 (60, 76)	70 (60, 77)	0.740		Wilcoxon
Weight, median (IQR)	77 (65, 91)	80 (66, 95)	0.338		Wilcoxon

网址：https://www.xiantao.love

更新时间：2023.10.19

目录

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

基线资料表用于描述每一个研究对象的基本情况，联合分组可以评估不同的分组之间不同临床变量的构成比是否有差别。

主要结果

列联表

	A	B	C	D	E	F
1	characteristics	Group1	Group2	pvalue	statistic	method
2	n	209	130			
3	T.stage, n (%)			0.5744166	1.990327	Yates' correction
4	T2	59 (19%)	39 (12.6%)			
5	T3	96 (31%)	63 (20.3%)			
6	T4	31 (10%)	19 (6.1%)			
7	T1	3 (1%)	0 (0%)			
8	N.stage, n (%)			0.1399033	5.478946	Yates' correction
9	N2	43 (14.3%)	18 (6%)			
10	N0	117 (38.9%)	77 (25.6%)			
11	N1	19 (6.3%)	21 (7%)			
12	N3	4 (1.3%)	2 (0.7%)			
13	M.stage, n (%)			0.3861310	0.751090	Chisq test
14	M0	96 (54.9%)	65 (37.1%)			
15	M1	10 (5.7%)	4 (2.3%)			
16	Pathologic.stage, n (%)			0.7061611	1.397328	Yates' correction
17	Stage IV	71 (20.9%)	42 (12.4%)			
18	Stage III	69 (20.4%)	45 (13.3%)			
19	Stage II	67 (19.8%)	43 (12.7%)			
20	Stage I	2 (0.6%)	0 (0%)			
21	Age, median (IQR)	68 (60, 76)	70 (60, 77)	0.7404427		Wilcoxon
22	Weight, median (IQR)	77 (65, 91)	80 (66, 95)	0.3379989		Wilcoxon

- characteristics：临床变量以及对应的分组
- Group1：Group1 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）
- Group2：Group2 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，该结果是对变量的整体评估，是比较组之间的显著性差异，非单个变量中等级之间的比较！

- 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	T.stage, n (%)	
3	T2	98 (31.6%)
4	T3	159 (51.3%)
5	T4	50 (16.1%)
6	T1	3 (1%)
7	N.stage, n (%)	
8	N2	61 (20.3%)
9	N0	194 (64.5%)
10	N1	40 (13.3%)
11	N3	6 (2%)
12	M.stage, n (%)	
13	M0	161 (92%)
14	M1	14 (8%)
15	Pathologic.stage, n (%)	
16	Stage IV	113 (33.3%)
17	Stage III	114 (33.6%)
18	Stage II	110 (32.4%)
19	Stage I	2 (0.6%)
20	Age, median (IQR)	69 (60, 76)
21	Weight, median (IQR)	78 (65.3, 92)

- 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）

数据格式

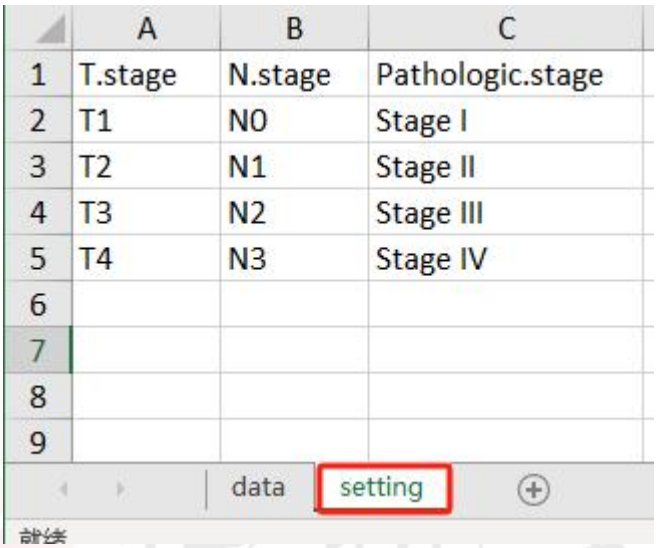
	A	B	C	D	E	F	G
1	group	T.stage	N.stage	M.stage	Pathologic.stage	Age	Weight
2	Group1	T2	N2		Stage IV	74	118
3	Group1	T3	N0	M0	Stage III	62	61
4	Group1	T4	N2	M0	Stage IV	56	83
5	Group1	T3	N2		Stage IV	83	95
6	Group2	T3	N0	M0	Stage III	45	81
7	Group2	T3	N0	M0	Stage III	79	104
8	Group2	T4	N1		Stage IV	90	
9	Group2	T3	N0	M1	Stage III	75	
10	Group2	T2	N0	M0	Stage II	70	95
11	Group2	T3	N1	M0	Stage IV	79	
12	Group2	T4	N0		Stage III	75	104
13	Group2	T3	N0	M0	Stage III	52	69
14	Group1	T3	N0		Stage III	82	62
15	Group1	T3	N2		Stage IV	61	93
16	Group1	T3	N0		Stage III	87	77

数据要求：表 1-分析数据

- 数据至少有 2 列以上，至少需要 10 行数据。当第一列为分类型，且分类个数为 2~4 时，默认统计 列联表，其他情况（数值型/单分类/多分类）统计 纯基线资料表。
- 每一列（第一列满足列联表要求除外）为变量，可以是数值型，也可以是分类型。
  - 基本要求：不应包含无法识别的字符、列缺失值的比例不应大于 85%。
  - 如果变量是数值型变量：应以数值纳入，只要包含非数值（除空值）或者是无穷值，则此列有可能没有办法纳入到分析中。
  - 数值型变量如果分类数<10 个（如 Stage 变量只有 0 1 2）则会按照分类变量来处理，否则会以数值变量纳入分析。
  - 如果变量是分类型变量：建议以具体的名字纳入，比如图中的 T.stage。如果分类数目超过 10 个将无法纳入分析。

- 分类变量可以在 excel 的表 2 中设置各分类的顺序（有序），默认以表 1 中变量的对应分类出现的顺序作为因子化顺序（无序）。注意，基线资料表中，该设置只影响输出结果的分类排列顺序，非指定参考组！

➢ 最多支持 20 列，20000 行。若验证数据时返回报错，需要在上传数据内进行相应的调整，然后再上传数据。



	A	B	C
1	T.stage	N.stage	Pathologic.stage
2	T1	N0	Stage I
3	T2	N1	Stage II
4	T3	N2	Stage III
5	T4	N3	Stage IV
6			
7			
8			
9			

data setting

数据要求:（表 2-可以不提供）

➢ 指定数据中分类变量的等级顺序，对应（表 1）变量（分类类型）中各分类的顺序。

- 比如 Pathologic.stage 想要设置 Stage I, Stage II, Stage III, Stage IV 的顺序，就可以如上图设置。注意，设置了等级顺序后，只影响输出结果的分类排列顺序，设置等级不对应时，将不会有变化。如果在表 1 中的分类变量没有设置等级顺序，则默认以在表 1 中各个分组出现的顺序作为等级顺序。

  【表 2 中设置正确时】

	A	B	C
1	T.stage	N.stage	Pathologic.stage
2	T1	N0	Stage I
3	T2	N1	Stage II
4	T3	N2	Stage III
5	T4	N3	Stage IV
6			
7			
8			
9			

data setting

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

· 基线资料表: 用来展示研究对象的基本信息情况

变量	Group1	Group2	p值	统计量	方法
T.stage, n (%)			0.574	1.9903	Yates' correction
T1	3 (1%)	0 (0%)			
T2	59 (19%)	39 (12.6%)			
T3	96 (31%)	63 (20.3%)			
T4	31 (10%)	19 (6.1%)			
N.stage, n (%)			0.140	5.4789	Yates' correction
N0	117 (38.9%)	77 (25.6%)			
N1	19 (6.3%)	21 (7%)			
N2	43 (14.3%)	18 (6%)			
N3	4 (1.3%)	2 (0.7%)			
M.stage, n (%)			0.386	0.75109	Chisq test
M0	96 (54.9%)	65 (37.1%)			
M1	10 (5.7%)	4 (2.3%)			

【表 2 中设置不正确时】

	A	B	C
1	T.stage		
2	T1		
3	T2	少了1个等级	
4	T3		
5			

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

· 基线资料表: 用来展示研究对象的基本信息情况

变量	Group1	Group2	p值	统计量	方法
n	209	130			
T.stage, n (%)			0.574	1.9903	Yates' correction
T2	59 (19%)	39 (12.6%)			
T3	96 (31%)	63 (20.3%)			
T4	31 (10%)	19 (6.1%)			
T1	3 (1%)	0 (0%)			
N.stage, n (%)			0.140	5.4789	Yates' correction
N2	43 (14.3%)	18 (6%)			
N0	117 (38.9%)	77 (25.6%)			
N1	19 (6.3%)	21 (7%)			

没有变化

- 比如以 0 1 2 编码的等级变量，如果没有在这个表中进行设置，且等级数>10 个，则会以数值类型纳入。

参数说明

（说明：标注了颜色的为常用参数。）

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，只有数据第 1 列为二分类或者是三分类的变量时才会有列联表类型提供，否则都是纯基线资料表。

- 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

主要结果格式为表格格式，提供 xlsx 和 docx 格式下载。

补充结果

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。

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

T.stage	Group1	Group2
T2	59 (59.7)	39 (38.3)
T3	96 (96.9)	63 (62.1)
T4	31 (30.5)	19 (19.5)
T1	3 (1.8)	0 (1.2)

T.stage中存在level满足 5>理论频数>=1 且 总样本数>=40 的条件，建议选用连续矫正卡方检验(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-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
Group1	Age	209	0.98342	0.0148
Group2	Age	130	0.97349	0.0119

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
Age	1	337	2.6716	0.1031

Age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）。

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. 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（第一列）的分类数不是 2-4 组。

② 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）。

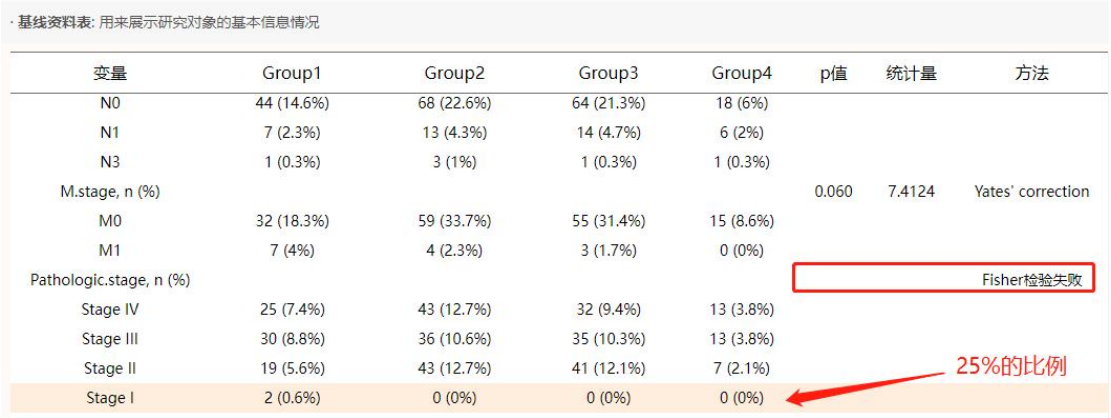
变量情况

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
分组	分类变量	4	0	不纳入	去掉任一纳入的变量有缺失的样本后，变成单...
T.stage	分类变量	4	49	纳入	
N.stage	分类变量	4	38	纳入	

③ 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（如下图，

Pathologic.stage 变量的 Stage I 分类，理论频数 T<1 的格子数占总格子数的 25%）。

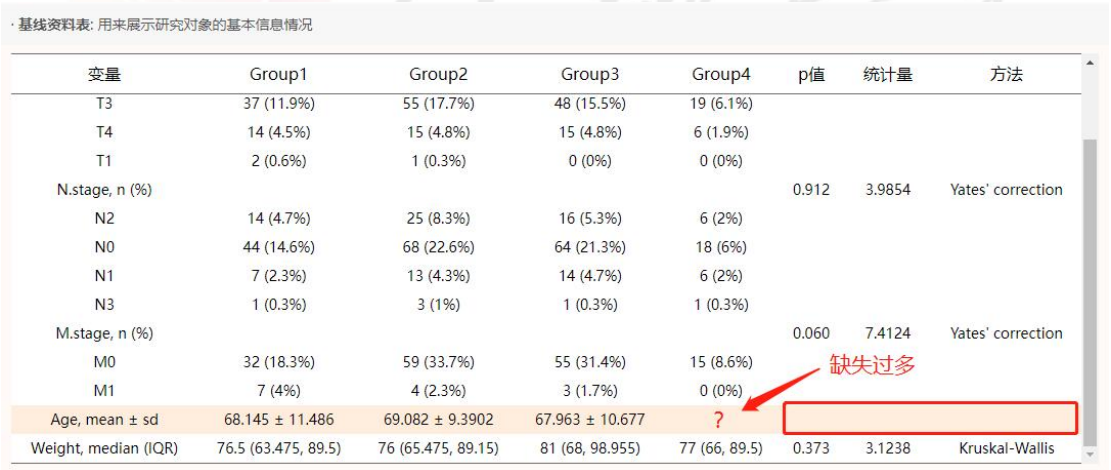


· 基线资料表: 用来展示研究对象的基本信息情况

变量	Group1	Group2	Group3	Group4	p值	统计量	方法
N0	44 (14.6%)	68 (22.6%)	64 (21.3%)	18 (6%)			
N1	7 (2.3%)	13 (4.3%)	14 (4.7%)	6 (2%)			
N3	1 (0.3%)	3 (1%)	1 (0.3%)	1 (0.3%)			
M.stage, n (%)					0.060	7.4124	Yates' correction
M0	32 (18.3%)	59 (33.7%)	55 (31.4%)	15 (8.6%)			
M1	7 (4%)	4 (2.3%)	3 (1.7%)	0 (0%)			
Pathologic.stage, n (%)							Fisher检验失败
Stage IV	25 (7.4%)	43 (12.7%)	32 (9.4%)	13 (3.8%)			
Stage III	30 (8.8%)	36 (10.6%)	35 (10.3%)	13 (3.8%)			
Stage II	19 (5.6%)	43 (12.7%)	41 (12.1%)	7 (2.1%)			
Stage I	2 (0.6%)	0 (0%)	0 (0%)	0 (0%)			

25%的比例

④ 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（如下图，Age 变量在分组 Group4 中缺失过多，无法统计均值等统计量）。



· 基线资料表: 用来展示研究对象的基本信息情况

变量	Group1	Group2	Group3	Group4	p值	统计量	方法
T3	37 (11.9%)	55 (17.7%)	48 (15.5%)	19 (6.1%)			
T4	14 (4.5%)	15 (4.8%)	15 (4.8%)	6 (1.9%)			
T1	2 (0.6%)	1 (0.3%)	0 (0%)	0 (0%)			
N.stage, n (%)					0.912	3.9854	Yates' correction
N2	14 (4.7%)	25 (8.3%)	16 (5.3%)	6 (2%)			
N0	44 (14.6%)	68 (22.6%)	64 (21.3%)	18 (6%)			
N1	7 (2.3%)	13 (4.3%)	14 (4.7%)	6 (2%)			
N3	1 (0.3%)	3 (1%)	1 (0.3%)	1 (0.3%)			
M.stage, n (%)					0.060	7.4124	Yates' correction
M0	32 (18.3%)	59 (33.7%)	55 (31.4%)	15 (8.6%)			
M1	7 (4%)	4 (2.3%)	3 (1.7%)	0 (0%)			
Age, mean ± sd	68.145 ± 11.486	69.082 ± 9.3902	67.963 ± 10.677	?			
Weight, median (IQR)	76.5 (63.475, 89.5)	76 (65.475, 89.15)	81 (68, 98.955)	77 (66, 89.5)	0.373	3.1238	Kruskal-Wallis

缺失过多

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。