临床研究数据分析样例集

Weituo Zhang

2025-06-08

# 代码生成要求

所有任务的共同上下文包括：研究背景和数据字典，  
默认数据已经载入完成, 大模型可以看到当前数据的前几行样例(如str) 每个任务独立执行，相互间没有关联  
代码生成时可以联网（或者给大模型提供R包的说明文档）

要求： 1. 代码正常运行，输出结果  
2. 代码中变量名与数据字典一致  
3. 统计分析方法与SAP要求一致（如果没有明确要求，可以自行发挥）

# 研究背景

本研究为高血压患者队列研究，2666名患者被分为试验组和对照组。入组时收集了人口学、血压、BMI、生活方式等特征。入组后随访患者的收缩压(SBP)和舒张压(DBP)。并且随访患者的心血管事件结局。本研究目标是研究不同干预方式能否影响患者的血压控制，以及降低心血管事件发生率。

# 数据载入

# 描述性统计(探索)

## Task 1-1

任务：计算两组患者各统计分析集人数  
输出：表格

## # A tibble: 2 × 3  
## arm ITT集人数 PP集人数  
## <chr> <dbl> <dbl>  
## 1 对照组 1114 890  
## 2 试验组 1132 890

## Task 1-2

任务：统计患者基线特征的缺失值数量  
输出：表格

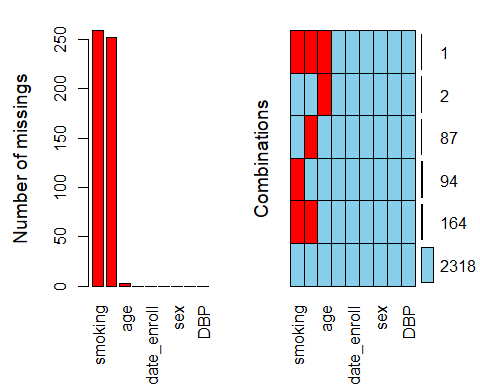
## ID date\_enroll arm age sex SBP   
## 0 0 0 3 0 0   
## DBP BMI smoking   
## 0 252 259

## Task 1-3

任务：统计患者基线特征的缺失值数量  
输出：图片pdf  
使用R包：VIM

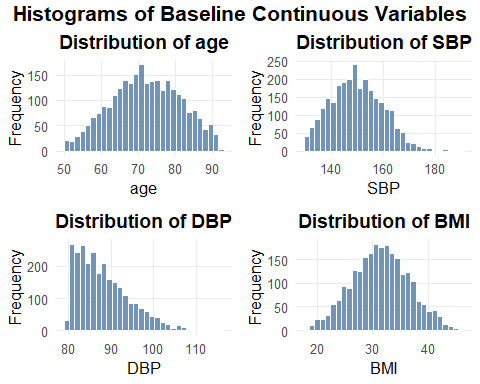
##   
## Variables sorted by number of missings:   
## Variable Count  
## smoking 259  
## BMI 252  
## age 3  
## ID 0  
## date\_enroll 0  
## arm 0  
## sex 0  
## SBP 0  
## DBP 0

## png   
## 2



##   
## Variables sorted by number of missings:   
## Variable Count  
## smoking 259  
## BMI 252  
## age 3  
## ID 0  
## date\_enroll 0  
## arm 0  
## sex 0  
## SBP 0  
## DBP 0

## Task-1-4

任务：展示所有基线连续变量的直方图  
变量：基线表中的连续变量  
（数据字典是一个meta数据框,包括表名 表中文名 字段名 字段中文名 字段类型 值域字典。选取baseline表中字段类型为numeric的变量）  
输出: 图片  
使用R包：ggplot2,gridExtra  


# 描述性统计(基线表)

## Task 1-5

任务：统计两组患者的基线特征  
变量：基线表中变量  
方法：收缩压、舒张压作为非正态分布处理  
输出：表格  
使用R包：tableone

## Stratified by arm  
## level 0 1   
## n 1308 1358   
## age (mean (SD)) 71.77 (9.53) 72.18 (9.20)   
## sex (%) 0 406 (31.0) 437 (32.2)   
## 1 902 (69.0) 921 (67.8)   
## SBP (median [IQR]) 149.50 [142.70, 157.33] 149.20 [142.90, 156.20]  
## DBP (median [IQR]) 87.20 [83.40, 91.70] 86.90 [83.40, 91.90]   
## BMI (mean (SD)) 30.29 (4.74) 32.48 (5.14)   
## smoking (%) 0 395 (33.7) 426 (34.5)   
## 1 560 (47.8) 584 (47.3)   
## 2 217 (18.5) 225 (18.2)   
## Stratified by arm  
## p test   
## n   
## age (mean (SD)) 0.256   
## sex (%) 0.554   
##   
## SBP (median [IQR]) 0.344 nonnorm  
## DBP (median [IQR]) 0.727 nonnorm  
## BMI (mean (SD)) <0.001   
## smoking (%) 0.919   
##   
##

## Task 1-6

任务：统计两组患者的基线特征  
人群：ITT  
基线变量：年龄、性别、BMI、吸烟  
输出：表格  
使用R包：gtsummary

| **变量** | **Overall** N = 2,246*1* | **对照组** N = 1,114*1* | **试验组** N = 1,132*1* | **P值***2* |
| --- | --- | --- | --- | --- |
| 年龄 | 72.1 (9.4) | 72.0 (9.5) | 72.2 (9.2) | 0.6 |
| 性别 |  |  |  | 0.8 |
| 女性 | 696 (31%) | 342 (31%) | 354 (31%) |  |
| 男性 | 1,550 (69%) | 772 (69%) | 778 (69%) |  |
| BMI | 31.6 (5.1) | 30.4 (4.8) | 32.8 (5.1) | <0.001 |
| 吸烟状态 |  |  |  | 0.9 |
| 从未吸烟 | 697 (35%) | 342 (34%) | 355 (35%) |  |
| 已经戒烟 | 368 (18%) | 177 (18%) | 191 (19%) |  |
| 曾经吸烟 | 955 (47%) | 475 (48%) | 480 (47%) |  |
| *1*Mean (SD); n (%) | | | | |
| *2*Wilcoxon rank sum test; Pearson's Chi-squared test | | | | |

## Task 1-7

任务：统计两组患者的基线特征  
人群：PP  
基线变量：年龄、性别、BMI、吸烟  
方法：连续变量采用中位数、上下四分位数进行描述;展示缺失值  
输出：表格  
使用R包：gtsummary

| **变量** | **Overall** N = 1,780*1* | **对照组** N = 890*1* | **试验组** N = 890*1* | **P值***2* |
| --- | --- | --- | --- | --- |
| 年龄 | 72.0 (65.2, 79.1) | 71.8 (65.1, 79.1) | 72.3 (65.6, 79.0) | 0.9 |
| Unknown | 2 | 2 | 0 |  |
| 性别 |  |  |  | 0.5 |
| 女性 | 552 (31%) | 270 (30%) | 282 (32%) |  |
| 男性 | 1,228 (69%) | 620 (70%) | 608 (68%) |  |
| BMI | 32.0 (28.5, 35.5) | 30.7 (27.1, 33.9) | 33.5 (30.0, 36.9) | <0.001 |
| Unknown | 171 | 86 | 85 |  |
| 吸烟状态 |  |  |  | 0.8 |
| 从未吸烟 | 556 (35%) | 273 (35%) | 283 (35%) |  |
| 已经戒烟 | 291 (18%) | 138 (18%) | 153 (19%) |  |
| 曾经吸烟 | 752 (47%) | 376 (48%) | 376 (46%) |  |
| Unknown | 181 | 103 | 78 |  |
| *1*Median (Q1, Q3); n (%) | | | | |
| *2*Wilcoxon rank sum test; Pearson's Chi-squared test | | | | |

## Task 1-8

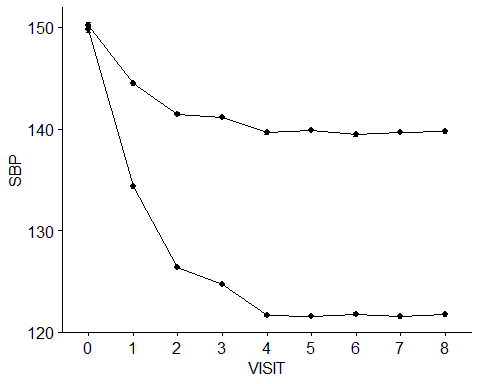
任务：统计两组患者的基线特征  
人群：年龄>60岁人群  
基线变量：年龄、性别、BMI>35、吸烟  
方法：连续变量采用中位数、上下四分位数进行描述;展示缺失值  
输出：表格  
使用R包：gtsummary

| **变量** | **Overall** N = 2,358*1* | **对照组** N = 1,145*1* | **试验组** N = 1,213*1* | **P值***2* |
| --- | --- | --- | --- | --- |
| 年龄 | 73.6 (67.7, 80.0) | 73.4 (67.6, 80.0) | 73.7 (67.8, 79.9) | 0.7 |
| 性别 |  |  |  | 0.7 |
| 女性 | 764 (32%) | 367 (32%) | 397 (33%) |  |
| 男性 | 1,594 (68%) | 778 (68%) | 816 (67%) |  |
| BMI |  |  |  | <0.001 |
| <=35 | 1,636 (77%) | 883 (85%) | 753 (69%) |  |
| >35 | 500 (23%) | 157 (15%) | 343 (31%) |  |
| Unknown | 222 | 105 | 117 |  |
| 吸烟状态 |  |  |  | 0.9 |
| 从未吸烟 | 700 (33%) | 332 (32%) | 368 (33%) |  |
| 已经戒烟 | 410 (19%) | 201 (20%) | 209 (19%) |  |
| 曾经吸烟 | 1,023 (48%) | 495 (48%) | 528 (48%) |  |
| Unknown | 225 | 117 | 108 |  |
| *1*Median (Q1, Q3); n (%) | | | | |
| *2*Wilcoxon rank sum test; Pearson's Chi-squared test | | | | |

# 纵向随访数据

## Task 2-1

任务：绘制两组患者血压随时间变化图  
变量：收缩压  
输出：折线图, 横轴为随访次数，纵轴为收缩压, 展示均值、标准误  
使用R包：ggplot2, ggpubr



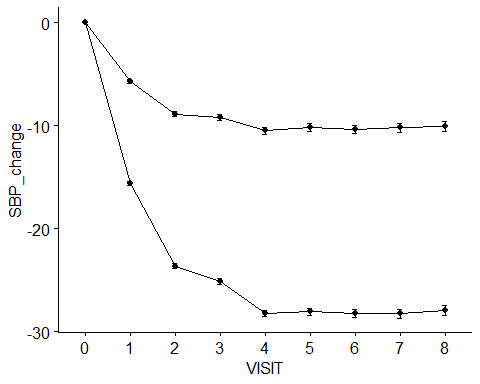
## Task 2-2

任务：统计两组患者各血压随访时间点的人数  
人群：PP  
变量：人数  
输出：表格

## # A tibble: 2 × 10  
## arm V0 V1 V2 V3 V4 V5 V6 V7 V8  
## <chr> <int> <int> <int> <int> <int> <int> <int> <int> <int>  
## 1 对照组 1308 1080 988 931 874 760 607 489 365  
## 2 试验组 1358 1176 1128 1020 1003 847 684 558 399

## Task 2-3

任务：绘制两组患者相对于基线的改变量随时间变化图  
变量：收缩压  
输出：折线图, 横轴为随访次数，纵轴为收缩压相对于基线的改变量, 展示均值、标准误 使用R包：ggplot2, ggpubr



# 血压结局比较

## Task 3-1

任务：比较两组患者6个月、1年时，血压相对于基线的变化量。  
人群：ITT  
变量：收缩压、舒张压在6个月+/-14天，1年+/-30天，相对于基线的改变量。  
方法：使用均值+-标准差描述各终点，使用t检验进行组间比较，计算组间差的mean difference 和95%置信区间  
缺失值：不插补  
输出：表格, 展示缺失值  
使用R包：gtsummary

| **Characteristic** | **对照组** N = 1,114*1* | **试验组** N = 1,132*1* | **Difference***2* | **95% CI***2* | **p-value***2* |
| --- | --- | --- | --- | --- | --- |
| 6个月收缩压变化(mmHg) | -8.8 ± 8.0 | -23.8 ± 8.3 | 15 | 14, 16 | <0.001 |
| Unknown | 272 | 194 |  |  |  |
| 6个月舒张压变化(mmHg) | -5.0 ± 4.8 | -13.8 ± 4.5 | 8.8 | 8.3, 9.2 | <0.001 |
| Unknown | 272 | 194 |  |  |  |
| 1年收缩压变化(mmHg) | -10.3 ± 9.7 | -28.4 ± 9.6 | 18 | 17, 19 | <0.001 |
| Unknown | 371 | 292 |  |  |  |
| 1年舒张压变化(mmHg) | -5.9 ± 5.7 | -16.4 ± 5.3 | 11 | 10, 11 | <0.001 |
| Unknown | 371 | 292 |  |  |  |
| *1*Mean ± SD | | | | | |
| *2*Welch Two Sample t-test | | | | | |
| Abbreviation: CI = Confidence Interval | | | | | |

## Task 3-2

任务：比较两组患者6个月、1年时，血压相对于基线的变化量。  
人群：ITT  
变量：收缩压、舒张压在6个月和1年时，相对于基线的改变量。  
方法：使用均值+-标准差描述各终点，使用t检验进行组间比较，计算组间差的mean difference 和95%置信区间  
缺失值处理: LOCF (对于6个月的终点，使用 182+14天前的最后一次观察值， 对于1年的终点，使用 365+14天前的最后一次观察值)  
输出：表格  
使用R包：gtsummary

| **Characteristic** | **对照组** N = 1,114*1* | **试验组** N = 1,132*1* | **Difference***2* | **95% CI***2* | **p-value***2* |
| --- | --- | --- | --- | --- | --- |
| 6个月收缩压变化(mmHg) | -7.5 ± 7.8 | -21.4 ± 9.9 | 14 | 13, 15 | <0.001 |
| 6个月舒张压变化(mmHg) | -4.2 ± 4.6 | -12.4 ± 5.5 | 8.2 | 7.7, 8.6 | <0.001 |
| 1年收缩压变化(mmHg) | -8.9 ± 9.3 | -25.7 ± 11.3 | 17 | 16, 18 | <0.001 |
| 1年舒张压变化(mmHg) | -5.0 ± 5.4 | -14.8 ± 6.4 | 9.8 | 9.3, 10 | <0.001 |
| *1*Mean ± SD | | | | | |
| *2*Welch Two Sample t-test | | | | | |
| Abbreviation: CI = Confidence Interval | | | | | |

## Task 3-3

任务：分析血压变化量的影响因素。  
变量：1年时，收缩压相对于基线的改变量。  
方法：多元线性回归，模型包含治疗组、年龄、性别、基线BMI、基线SBP  
缺失值处理: LOCF (对于1年的终点，使用 365+14天前的最后一次观察值)  
输出：多元线性回归表  
使用R包：gtsummary

| **Characteristic** | **Beta** | **95% CI** | **p-value** |
| --- | --- | --- | --- |
| 治疗组 |  |  |  |
| 对照组 | — | — |  |
| 试验组 | -17 | -17, -16 | **<0.001** |
| 年龄(岁) | 0.01 | -0.02, 0.04 | 0.5 |
| 性别 |  |  |  |
| 女性 | — | — |  |
| 男性 | -0.02 | -0.59, 0.55 | >0.9 |
| 基线BMI(kg/m²) | -0.03 | -0.08, 0.03 | 0.3 |
| 基线收缩压(mmHg) | -0.80 | -0.83, -0.77 | **<0.001** |
| Abbreviation: CI = Confidence Interval | | | |

# 生存事件结局

## Task 4-1

任务：derive数据。根据事件表，计算患者的生存结局。  
events表中包含患者的各类事件0: 结束随访, 1:心血管死亡, 2:卒中, 3:心梗, 4:心衰, 5:非心源性死亡, 6:失访  
根据这张表，计算患者的生存结局表survival\_data, 结构如下：  
包含字段  
ID: 患者ID  
days: 入组后天数  
status: 生存结局 1:死亡, 0:删失

输出：survival\_data.xlsx, 以及相应的数据字典。

## # A tibble: 6 × 3  
## ID days status  
## <chr> <dbl> <dbl>  
## 1 X0001 615 0  
## 2 X0002 1093 0  
## 3 X0003 1093 0  
## 4 X0004 1093 0  
## 5 X0005 3 0  
## 6 X0006 315 1

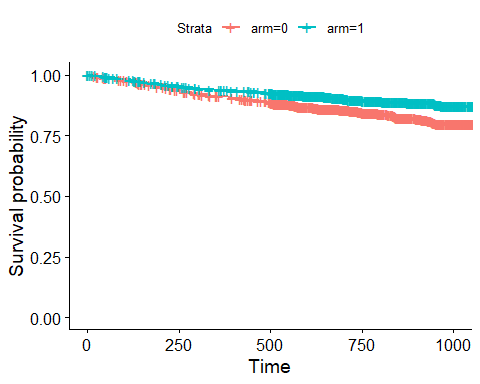
## Task 4-2

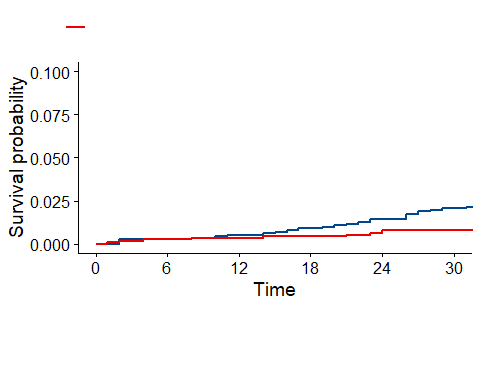
任务: derive数据。根据事件表，计算患者的MACE事件  
events表中包含患者的各类事件0: 结束随访, 1:心血管死亡, 2:卒中, 3:心梗, 4:心衰, 5:非心源性死亡, 6:失访  
变量：MACE事件包括心血管死亡、心梗、卒中、心衰；删失事件包括结束随访、失访、非心源性死亡 )  
计算患者的MACE事件表 MACE\_data, 结构如下：  
包含字段  
ID: 患者ID  
days: 入组后天数  
status: 生存结局 1:死亡, 0:删失  
输出：MACE\_data.xlsx, 以及相应的数据字典。

## # A tibble: 6 × 3  
## ID days status  
## <chr> <dbl> <dbl>  
## 1 X0001 615 0  
## 2 X0002 1093 0  
## 3 X0003 201 1  
## 4 X0004 1093 0  
## 5 X0005 3 0  
## 6 X0006 244 1

## Task 4-3

任务: 绘制两组患者的生存曲线  
人群：ITT  
变量：总生存OS (终点事件包括心血管死亡、非心源性死亡；删失事件包括结束随访、失访 )  
方法：KM曲线  
使用R包：survival, survminer

 ## Task 4-4 任务: 绘制两组患者的MACE事件累积风险曲线  
人群：ITT  
变量：MACE (终点事件包括心血管死亡、心梗、卒中、心衰；删失事件包括结束随访、失访、非心源性死亡 )  
方法: 绘制累积风险曲线, 使用lancet配色, 增加risk table, y轴范围为0-0.1  
输出：pdf图  
使用R包：survival, survminer



## Task 4-5

任务: 比较两组患者的MACE事件风险  
人群：ITT  
events表中包含患者的各类事件0: 结束随访, 1:心血管死亡, 2:卒中, 3:心梗, 4:心衰, 5:非心源性死亡, 6:失访 变量：MACE (终点事件包括心血管死亡、心梗、卒中、心衰；删失事件包括结束随访、失访、非心源性死亡 )  
方法: cox回归, 模型包括 治疗组、年龄、性别、基线SBP 输出：表格 使用R包：survival

## # A tibble: 4 × 5  
## Variable HR `95% CI Lower` `95% CI Upper` `P-value`  
## <chr> <dbl> <dbl> <dbl> <chr>   
## 1 arm 0.63 0.54 0.75 6.3e-08   
## 2 age 1.02 1.01 1.03 0.00015   
## 3 sex 1.09 0.91 1.3 0.37287   
## 4 SBP 1 1 1.01 0.25842