PART.01

数据分析前必须掌握的基础概念

$$Y=2X+3Z$$

X和Z就是自变量

Y就是应变量

连续变量和分类变量

连续变量：

可以用具体的数值去表示（身高，血压）

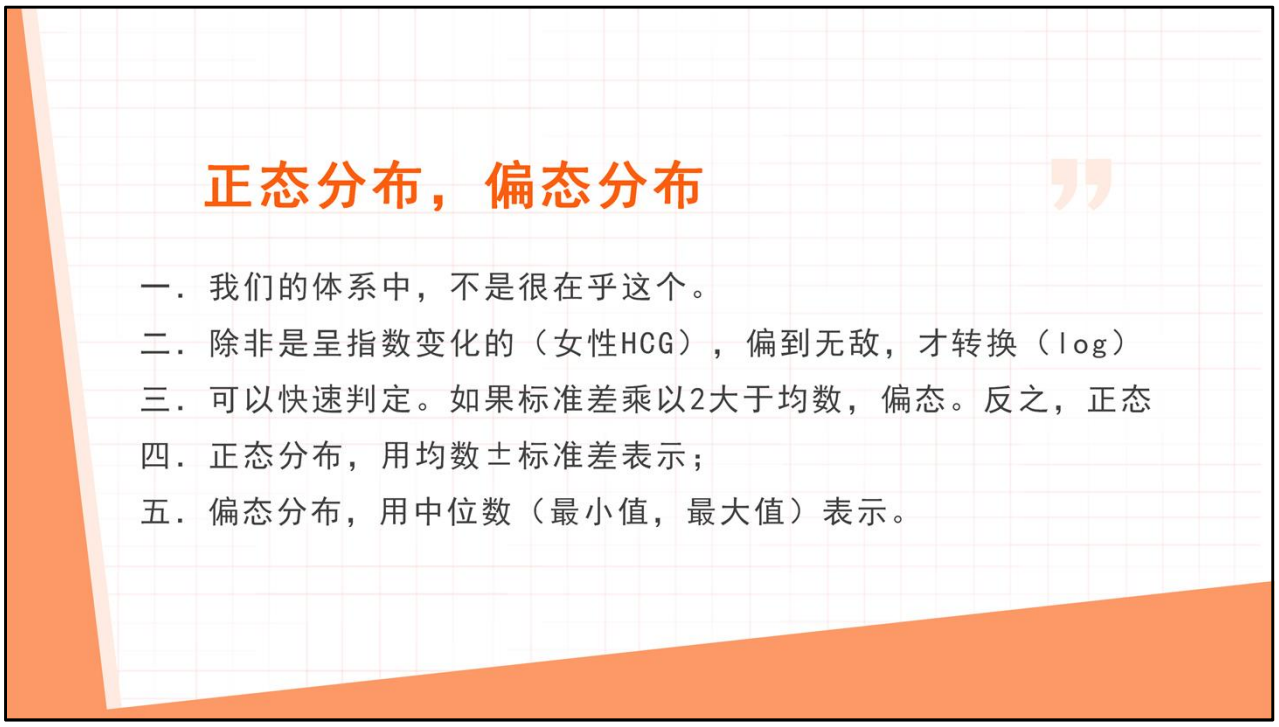
分类变量：

有序分类变量（好中差、轻中重）

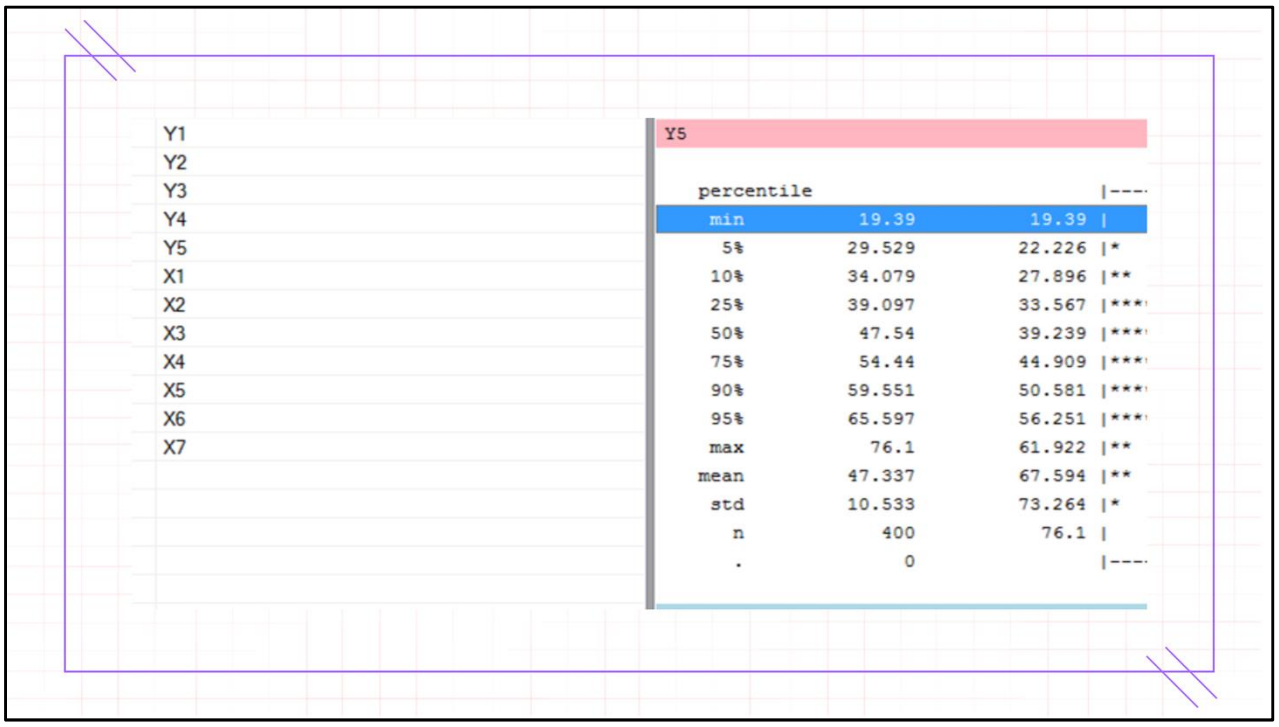
无序分类变量（血型，性别）

均数：全部加起来，除以例数。

- 标准差：反映一个数据的离散程度（数据的波动）
- 中位数：把数据从小到大排列，排第50位的。
- 三分组：排第33位的，66位的。
- 四分组：排第25位的 ，第50位的，第75位的，



易倆教材回归 分析基础中，对于回归方程XY正态、偏态分布要求并没有如此严格，因为实际上只需要方程残差正态分布即可。除非遇到指数性变化指标，譬如女性怀孕期间HCG的指数性变化才需要进行转换（Log10 转换）。此外，无需进行数据分析前的正态性检验。可以通过均数和标准差快速大致判定。这种判定无非就是在论文撰写时，对如何展示连续变量有一定帮助。因为两者的呈现形式不相同。

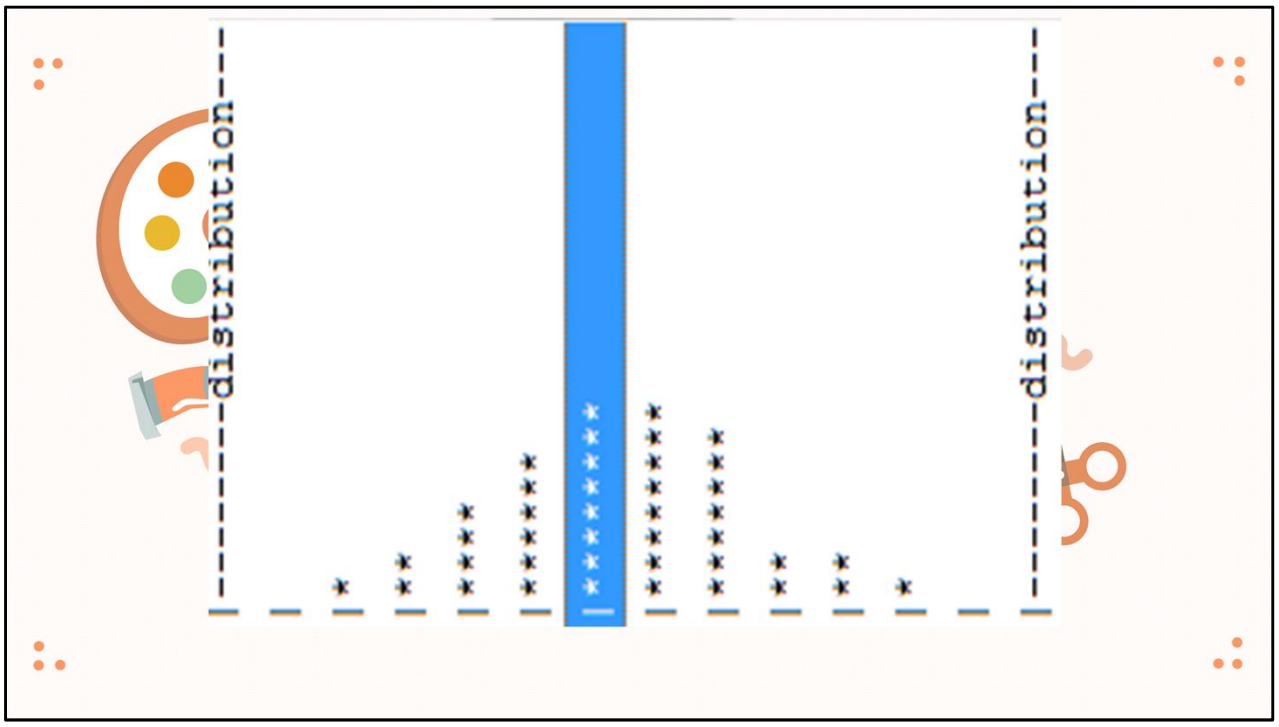


此图为易侕软件的数据展示，可轻松的获取最大值，最小值等关键信息。

Y5

percentile			--------distribution---
min	19.39	19.39	|
5%	29.529	22.226	|*
10%	34.079	27.896	|**
25%	39.097	33.567	|****
50%	47.54	39.239	|******
75%	54.44	44.909	|********
90%	59.551	50.581	|********
95%	65.597	56.251	|*******
max	76.1	61.922	|**
mean	47.337	67.594	|**
std	10.533	73.264	|*
n	400	76.1	|
.	0		--------distribution---

同样，右侧的星号组成图形，其实就是表示正偏态分布的示例直方图。



可知，最高点近似与接近中间，多为正态或近似正态分布

第一步

表1

我用了哪些变量

哪些和X有关

哪些和Y有关

表1是数据分析的第一步，其目的主要是：研究所涉变量包括哪些。根据分组变量，判定哪些与X有关，哪些有Y有关。

“

表1的名称：Baseline characteristics of participants

”

Table 1. Baseline Characteristics of WAFACS and Blood Substudy Participants

Characteristic	Active Group, No. (%) (n = 2721)[a]	Placebo Group, No. (%) (n = 2721)[a]
Age, mean (SD), y	62.8 (8.8)	62.8 (8.8)
Age, y		
40-54	582 (21.4)	584 (21.5)
55-64	990 (36.4)	970 (35.6)
≥ 65	1149 (42.2)	1167 (42.9)
Prior cardiovascular disease[b]	1764 (64.8)	1728 (63.5)
Risk factors		
Hypertension[c]	2360 (86.7)	2335 (85.8)
Elevated cholesterol[d]	2118 (77.8)	2150 (79.0)
Body mass index ≥30[e]	1341 (49.3)	1349 (49.6)
Parental history of myocardial infarction[f]	1056 (38.9)	1097 (40.5)
Alcohol intake, ≥1 glass/wk[g]	897 (33.0)	889 (32.7)
Diabetes	570 (21.0)	574 (21.1)
Current smoking	311 (11.4)	334 (12.3)
Current medication use		
Aspirin[h]	1446 (51.1)	1385 (48.9)

表1的分组变量，主要基于研究类型及一些特殊情况。

Table 1. Baseline Characteristics of WAFACS and Blood Substudy Participants

Characteristic	Active Group, No. (%) (n = 2721)[a]	Placebo Group, No. (%) (n = 2721)[a]

PLR (tertile)	Low	Middle	High	*P*-value
N	50	52	52	

如果研究类型为队列，以X作为分组。

X是分类变量，原来分几层，就是几个组

X是连续变量，根据样本量，3分组，4分组都可以

队列和RCT，主要以X作为分组变量。

如果是病例对照研究，以Y分组（病例组，对照组）

TABLE 1. Demographic and lifestyle characteristics of study participants: 15 women with spontaneous abortion (cases) and 15 women with live births (controls)

	Mean ± SD	
	Case (n = 15)	Control (n = 15)

如果是横断面研究，以Y、X分组均可

TABLE 1. Demographic and lifestyle characteristics of study participants: 15 women with spontaneous abortion (cases) and 15 women with live births (controls)

	Mean ± SD	
	Case (n = 15)	Control (n = 15)

下列情况可以不分组

01 前瞻性队列，为了体现盲法

02 有多个X，多个Y，可以不分组，简单罗列

Table 1| **Maternal, fetal, and childhood characteristics (n=1184). Values are numbers (percentages) unless stated otherwise**

Characteristics	Value
Maternal	
Median (90% range) age, years	31.3 (22.7-38.1)
Mean (SD) height, cm	168.8 (7.0)
Mean (SD) pre-pregnancy weight, kg	66.9 (11.8)
Mean (SD) pre-pregnancy body mass index, kg/m^2	23.4 (3.9)
Median (90% range) gestational age at intake, weeks	12.4 (10.5-13.9)
Mean (SD) systolic blood pressure, mm Hg	116.7 (12.4)
Mean (SD) diastolic blood pressure, mm Hg	69.1 (9.4)
Nulliparous	717/1179
Education:	
Primary or secondary school	507/1155 (43.9)
Higher education	648/1155 (56.1)
Race/ethnicity:	
Dutch, other European	855/1175 (72.8)
Non-European	320/1175 (27.2)
Smoking habits:	
Non-smoker	820/1060 (77.4)

前瞻性队列，未分组，以体现盲法

研究类型		分组变量
观察性	横断面	X或Y
	队列	无或X
	病例对照	Y
试验性	RCT	X

分组变量选定总结

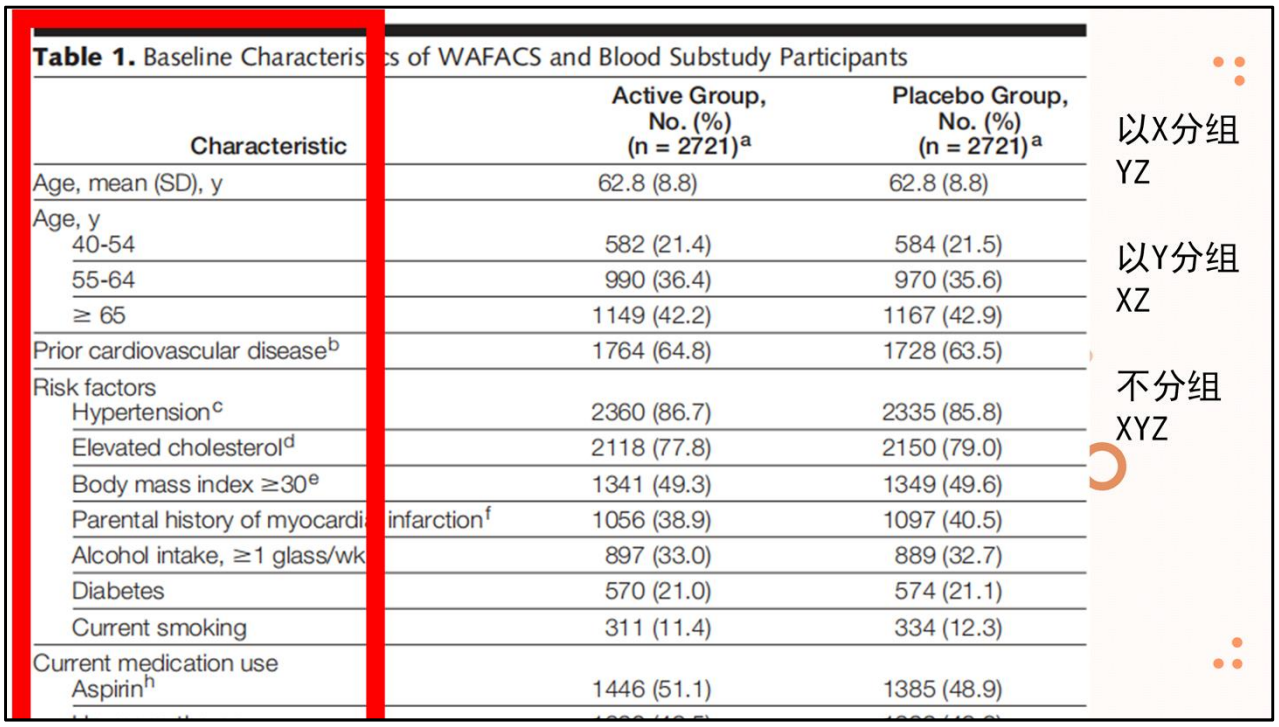


Table 1. Baseline Characteris cs of WAFACS and Blood Substudy Participants

Characteristic	Active Group, No. (%) (n = 2721)[a]	Placebo Group, No. (%) (n = 2721)[a]
Age, mean (SD), y	62.8 (8.8)	62.8 (8.8)
Age, y		
40-54	582 (21.4)	584 (21.5)
55-64	990 (36.4)	970 (35.6)
≥ 65	1149 (42.2)	1167 (42.9)
Prior cardiovascular disease[b]	1764 (64.8)	1728 (63.5)
Risk factors		
Hypertension[c]	2360 (86.7)	2335 (85.8)
Elevated cholesterol[d]	2118 (77.8)	2150 (79.0)
Body mass index ≥30[e]	1341 (49.3)	1349 (49.6)
Parental history of myocardi infarction[f]	1056 (38.9)	1097 (40.5)
Alcohol intake, ≥1 glass/wk	897 (33.0)	889 (32.7)
Diabetes	570 (21.0)	574 (21.1)
Current smoking	311 (11.4)	334 (12.3)
Current medication use		
Aspirin[h]	1446 (51.1)	1385 (48.9)

以X分组
YZ

以Y分组
XZ

不分组
XYZ

所涉及变量：以X分组，则包括YZ，以Y分组，包括XZ，不分组，包括XYZ

关于表1的说明：

一．观察性研究，表1的P小于0.05，太正常不过。

二．都大于0.05，有造假的嫌疑。

三．如果以X分组，P值小于0.05的变量为“与X有关”

四．如果以Y分组，P值小于0.05的变量为“与Y有关”

五．如果不分组，就没有P值了。就只剩下罗列。

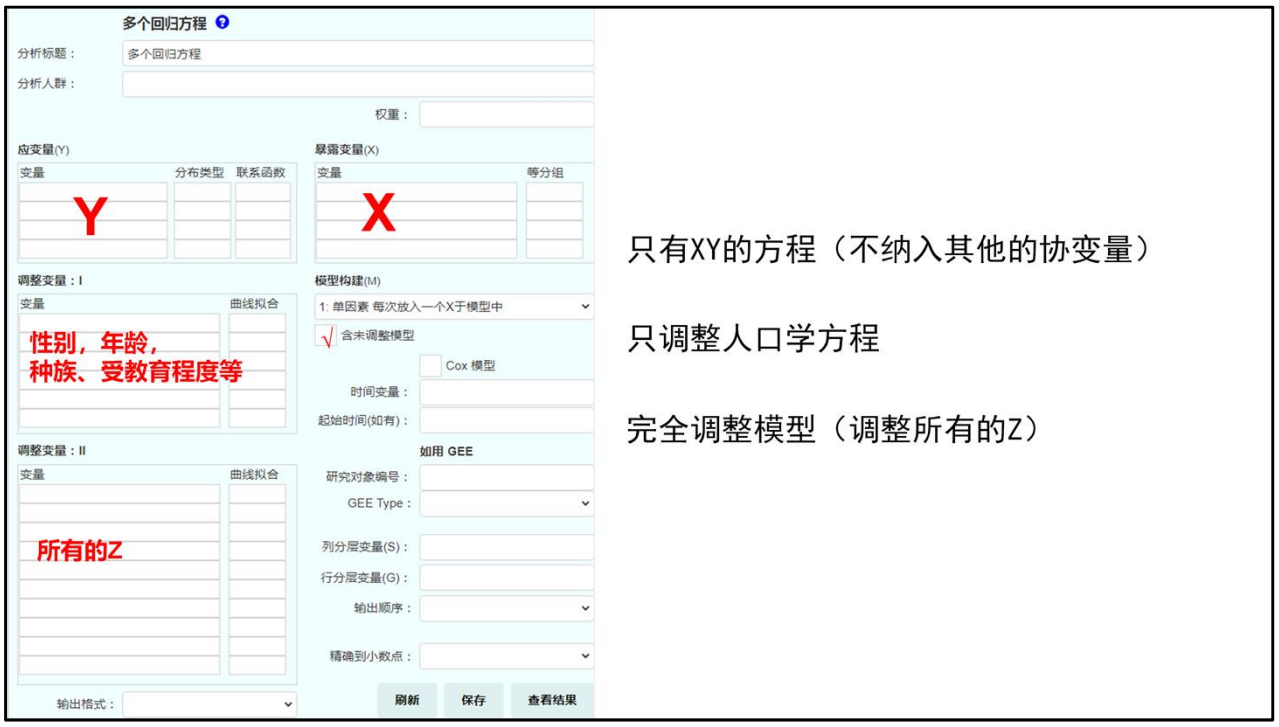
第二步

表2

排除Z的干扰

XY的直线关系

表2实际呈现的是排除混杂因素的干扰，XY的独立线性关系。其中，包括了多个回归方程，目的是观察不同调整策略下，XY的效应值趋势。



只有XY的方程（不纳入其他的协变量）

只调整人口学方程

完全调整模型（调整所有的Z）

操作如图所示。初学者对于不同模型调整策略可先照抄上述示例。等打好基础后，则可基于临床实际和数据分析思路，自行变化调整策略。

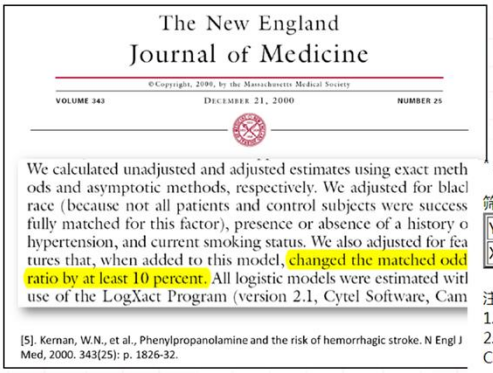
一．Y不选多分类，以二分类（且只能设置为0，1。其中，0为未发生，1为发生事件）及连续变量作为Y。

二．X可以是连续变量，分类变量，无区别。

三．Z的纳入主要基于既往文献。原则是同类研究，或者公认的可影响XY关系的变量必须纳入。

四．如果此前没有文献，或者是基于文献需要调整的变量太多，则使用协变量筛选模块

协变量筛选模块及方法学来源



The New England
Journal of Medicine

VOLUME 343 DECEMBER 21, 2000 NUMBER 25

We calculated unadjusted and adjusted estimates using exact meth
ods and asymptotic methods, respectively. We adjusted for blacl
race (because not all patients and control subjects were success
fully matched for this factor), presence or absence of a history o
hypertension, and current smoking status. We also adjusted for fea
tures that, when added to this model, changed the matched odd
ratio by at least 10 percent. All logistic models were estimated witl
use of the LogXact Program (version 2.1, Cytel Software, Cam

[5]. Kernan, W.N., et al., Phenylpropanolamine and the risk of hemorrhagic stroke. N Engl J Med, 2000. 343(25): p. 1826-32.

表示与起始回归系数相比变化超过 10%

筛选出来的协变量

Y	X	选出的协变量（标准1）	选出的协变量（标准2）
X2	X5	X6	X6 X10

注释：
1. 标准1：在基本模型中引进协变量或在完整模型中剔除协变量对X的回归系数的影响>10%
2. 标准2：标准1或协变量对Y的回归系数P值<0.1
Created by EmpowerStats (www.empowerstats.com) and R on 2021-07-03

Table 3

Cox proportional hazard regression models examining the association between HIV/MDD group and incident heart failure

Characteristic	Model 1 HR (95% CI)	Model 2 HR (95% CI)	Model 3 HR (95% CI)
HIV / MDD Group			
HIV -- | No MDD	1 [Reference]	1 [Reference]	1 [Reference]
HIV -- | MDD	1.13 (1.00 – 1.27)	1.30 (1.15 – 1.47)	1.19 (1.05 – 1.35)
HIV + | No MDD	1.25 (1.15 – 1.37)	1.32 (1.20 – 1.44)	1.28 (1.16 – 1.41)
HIV + | MDD	1.54 (1.34 – 1.77)	1.87 (1.62 – 2.15)	1.68 (1.45 – 1.95)

X=HIV*MDD

Y=是否心衰

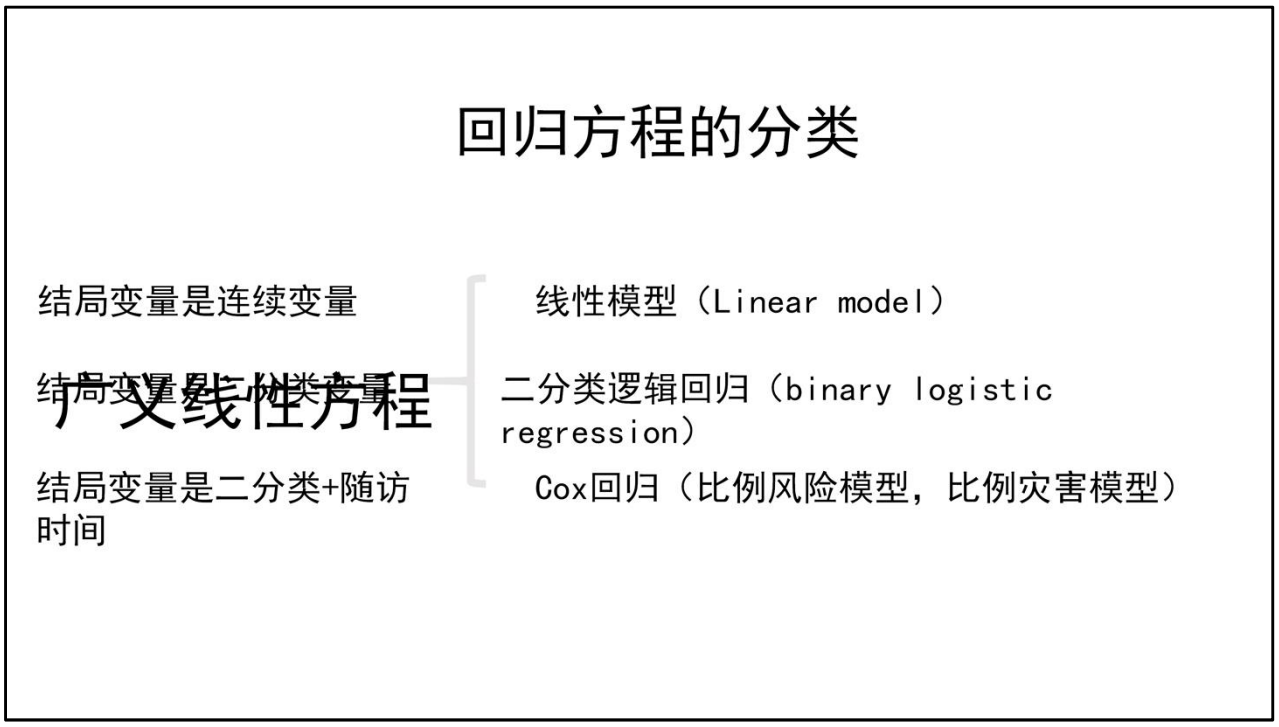
Model 1: HIV MDD only

Model 2: Model 1+ age, sex, race/ethnicity

Model 3: Model 2+ all covariates

Abbreviations: BMI, body mass index (calculated as kg/m^2); eGFR, estimated glomerular filtration rate; HCV, hepatitis C virus; HDL, high-density lipoprotein; HIV, human immunodeficiency virus; HR, hazard ratio; LDL, low-density lipoprotein; MDD, major depressive disorder

关联分析类临床研究SCI论文，无论影响因子高低，98%以上的核心结果是用回归分析得出来的。



让我们对回归分析进行一些基础的理解：最常见的回归方程分为三大类，其分类主要基于Y的变量类型。分别分为线性回归、二分类逻辑回归，比例风险模型（Cox回归模型）。三者统称为广义线性方程。因为模型中，所有X和Y的关系均为线性关系。换言之，处理非线性关系是广义线性方程的短板。

线性回归的效应值叫贝塔（β）

二分类逻辑就回归的效应值叫比值比(Odds ratio（OR）)

Cox回归的效应值叫风险比(Hazards ratio（HR）)

可信区间：=P值，且更为精确。

效应值结果解读第一步：

1. 根据Y定性
2. 如果Y是连续变量，效应值+0，正数就是增加，负数就是减少。
3. 如果Y是二分类变量（设定Y=0，没有发生，Y=1，发生）
4. （效应值-1）*100%=正数就是增加，负数就是减少。

效应值结果解读第二步：

根据X定量

如果X是连续变量，X每增加1个单位，Y的变化。

根据X定量

如果X是分类变量（设定X=1的为参照）

X=2的组与X=1的相比Y的变化

X=3的组与X=1的相比Y的变化

X=…… 与X=1的相比Y的变化

结果解读1

X=BMI（kg/m^2）　　Y=收缩压（mmhg）　　效应值：β=2.13

根据Y定性

如果Y是连续变量，效应值+0，正数就是增加，负数就是减少。

如果Y是二分类变量（设定Y=0，没有发生，Y=1，发生）

（效应值-1）*100%=Y=1发生的概率，正数就是增加，负数就是减少。

根据X定量

如果X是连续变量，X每增加1个单位，Y的变化。

如果X是分类变量（设定X=1的为参照）

X=2的组与X=1的相比Y的变化
X=3的组与X=1的相比Y的变化
X=…… 与X=1的相比Y的变化

BMI每增加1个kg/m^2，收缩压增加2.13个mmHg。

结果解读2

X=吸烟状态（1=不吸烟，2=吸烟，3=戒烟）　Y=收缩压（mmhg）　　效应值：β =2.13

根据Y定性

如果Y是连续变量，效应值+0，正数就是增加，负数就是减少。

如果Y是二分类变量（设定Y=0，没有发生，Y=1，发生）

（效应值-1）*100%=Y=1发生的概率，正数就是增加，负数就是减少。

根据X定量

如果X是连续变量，X每增加1个单位，Y的变化。

如果X是分类变量（设定X=1的为参照）

X=2的组与X=1的相比Y的变化
X=3的组与X=1的相比Y的变化
X=…… 与X=1的相比Y的变化

吸烟和不吸烟相比，收缩压增加2.13mmHg。
戒烟和不吸烟相比，收缩压增加1.13mmHg。

结果解读3

X=BMI（kg/m^2）　　Y=是否高血压　　效应值：OR=2.13

根据Y定性

如果Y是连续变量，效应值+0，正数就是增加，负数就是减少。

如果Y是二分类变量（设定Y=0，没有发生，Y=1，发生）

（效应值-1）*100%=Y=1发生的概率，正数就是增加，负数就是减少。

根据X定量

如果X是连续变量，X每增加1个单位，Y的变化。

如果X是分类变量（设定X=1的为参照）

X=2的组与X=1的相比Y的变化
X=3的组与X=1的相比Y的变化
X=…… 与X=1的相比Y的变化

BMI每增加1个kg/m^2，发生高血压的风险（Y=1高血压）增加113%。

第三步

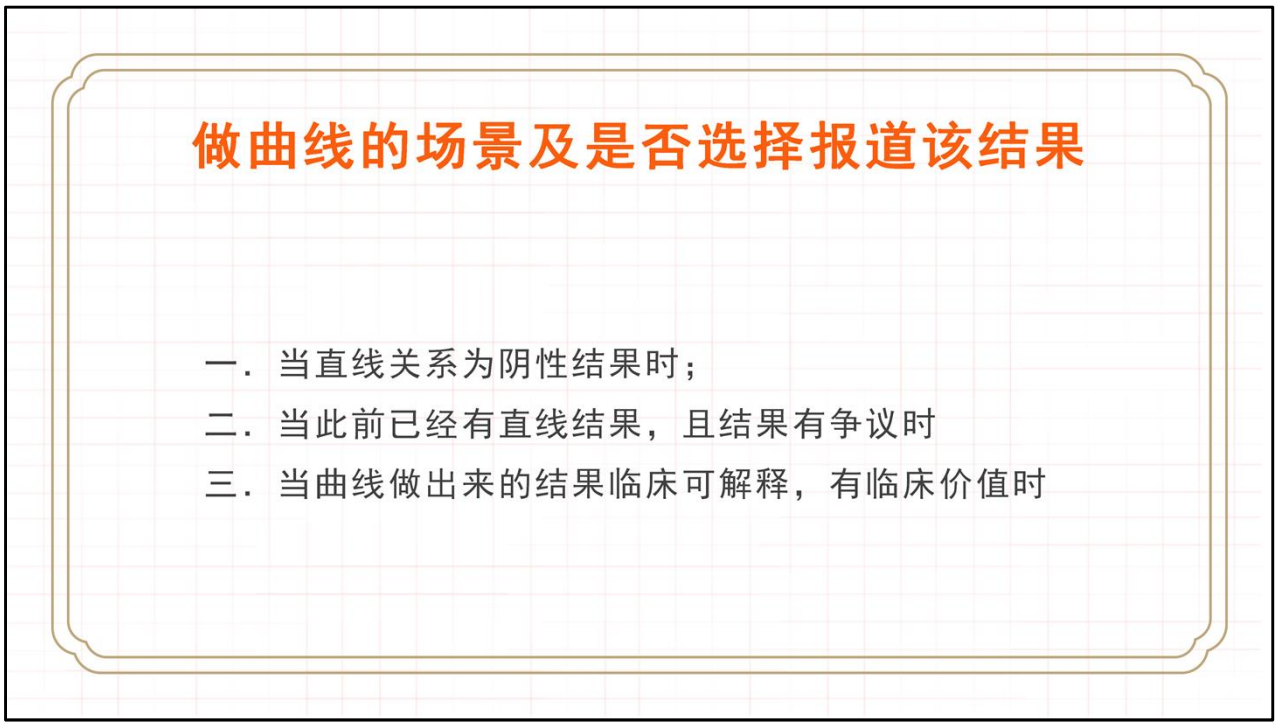
排除Z的干扰

XY的曲线关系

图1，表3

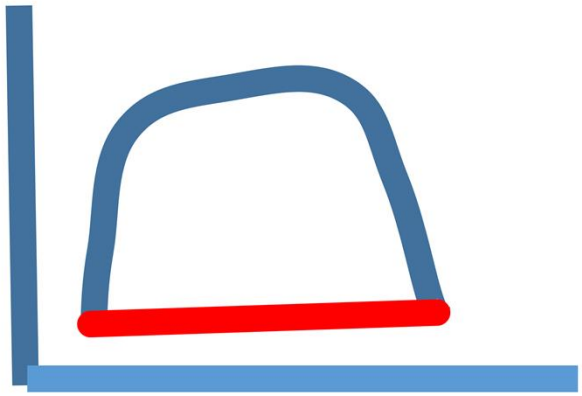
该部分结果承接的是XY的线性关系（表2）。很多时候，当X为连续变量时，此前所用的广义线性模型就无法发现XY之间可能存在的非线性关系，因此，需要使用新的数据分析思路及策略。其中，非线性关系的探讨即为后续备选方案。非线性关系的探讨包括了一图（曲线拟合图）和一表（对曲线拟合图进行解释）

首先，需明确：曲线的前提条件为X必须为连续变量。



此外，需注意非线性关系的应用时机。
时机1：在直线关系为阴性需要进一步探讨/验证时；，
时机2：此前XY结果有争议时，通过曲线拟合可对此前的结果进行验证/更新。
报道的原则：结果出来后，曲线结果临床是否可解释？是否具有临床价值？只有临床可解释，且有临床价值的曲线，才值得报道。

（1）为什么要做曲线



阴性结果？

补救

譬如：很多时候XY是阴性结果，即X变化，而Y无改变，这种情况很可能是因为其实际是一条曲线，如研究者忽略了可能存在的非线性关系，则可能得到错误的结果。故而，对XY的线性关系为阴性结果时，如X为连续变量，曲线拟合必做。

为什么要做曲线

张三 et al reported that X is positively associated with Y

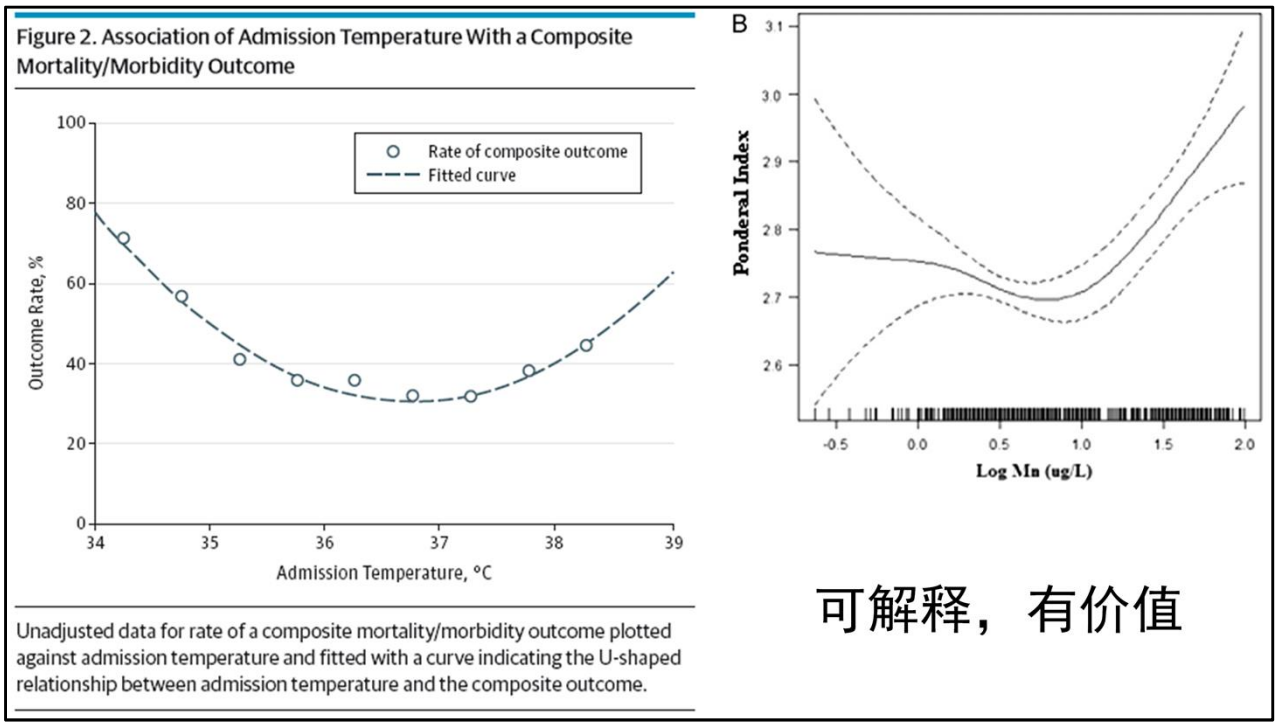
李四 et al reported that X is positively associated with Y

王五 et al reported that X is positively associated with Y

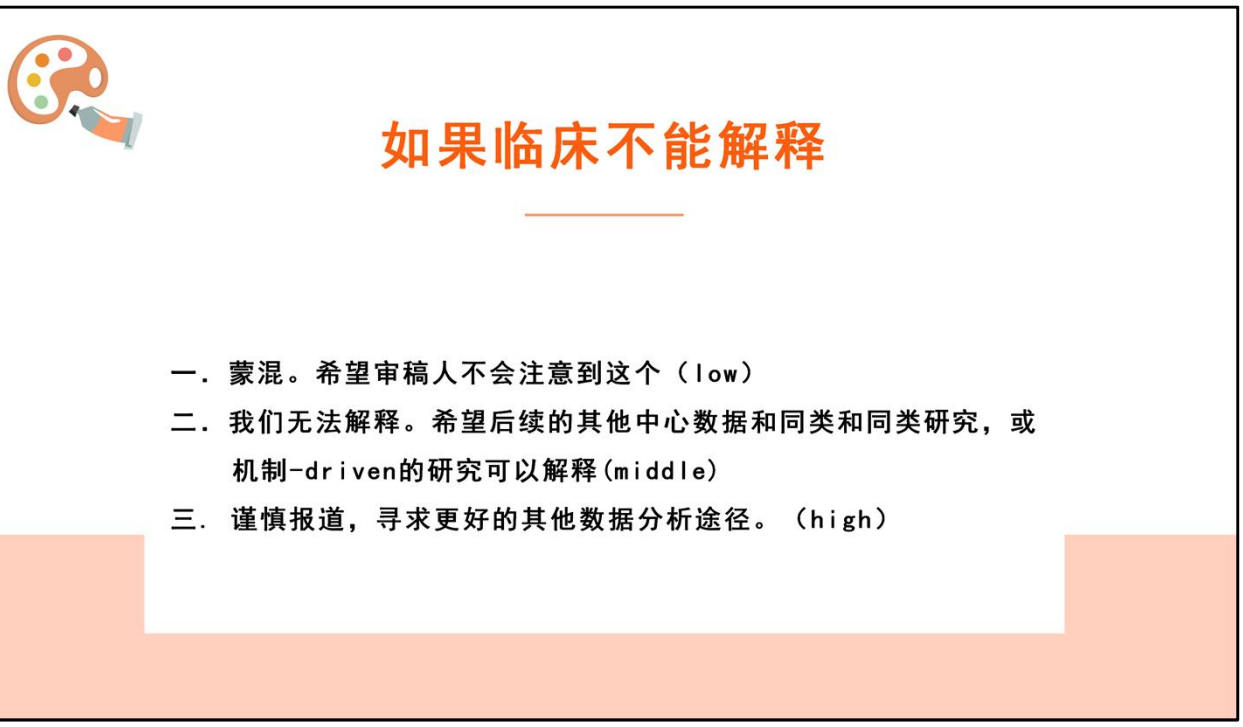
We found that X和Y的关系是非线性关系

创新

譬如，此前的三位研究者关于X和Y的关系存在争议，曲线关系的提出有可能对这种争议提出新的证据，这是一种创新和进步。



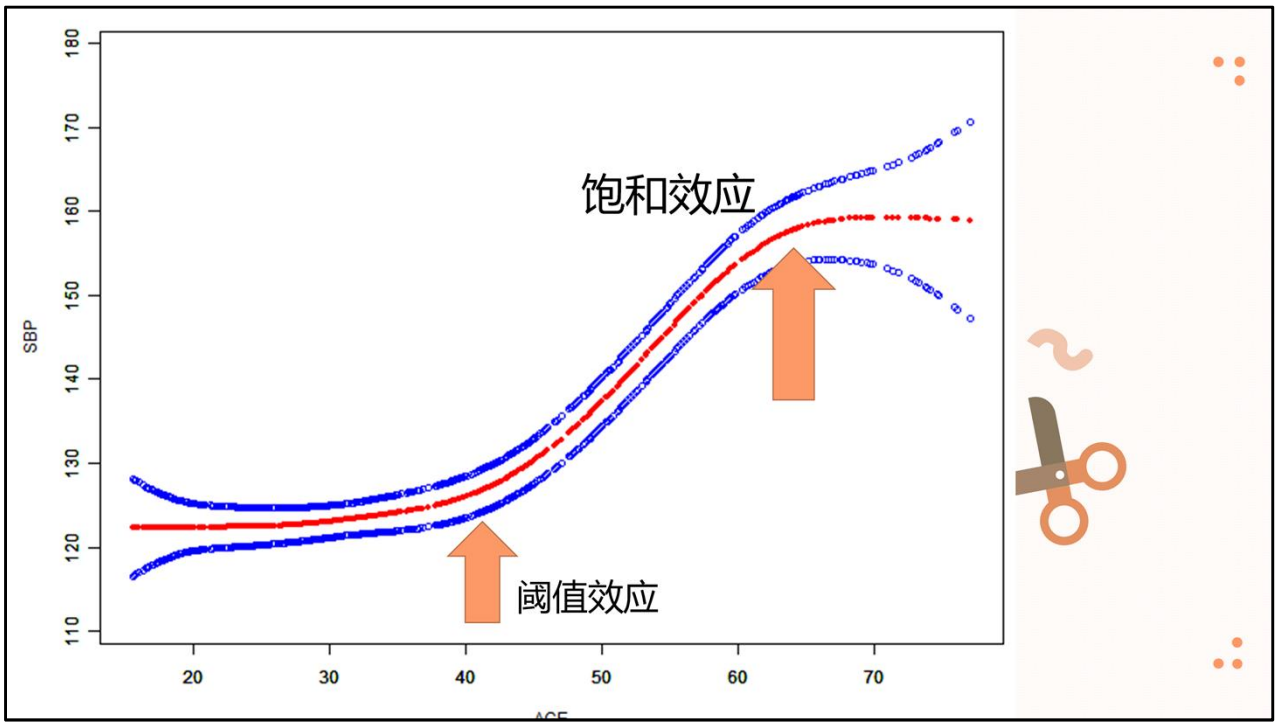
临床可解释且具有临床价值，是决定曲线结果是否报道的重要前提。试想，如果该曲线结果与临床完全违背，而研究者无法做出相关解释，如被审稿人所问询，则会非常被动。不能为了做曲线而做曲线。本例中，新生儿体温与不良事件发生为U型，即体温过高过低均不对（符合临床），其中间即为婴儿的最佳体温范围（临床意义）



譬如，我们在报道曲线时，如果不注意其临床解释或临床价值，很可能会被审稿人责难。此时，主动权将不在我们手上，应对的方式会为文章的接收增加未知的变数。

曲线关系的两个步骤：做出曲线图及饱和饱和阈值效应值（表）

曲线所用模块包括：平滑曲线拟合模块（作图）；饱和阈值 效应模块（做表）。两者相辅相成。后者是前者的精细化、数字化解读。



图中我们可以清晰的看到，阈值效应和饱和效应是曲线关系中最常见的两种表现形式。

For exposure: MPV

Outcome:	OUTCOME1
模型 I 一根	
一条直线回归系数	1.2 (1.1, 1.3) <0.001
模型 II 两根	
折点(K)	19
< K 段回归系数 1	1.3 (1.2, 1.5) <0.001
> K 段回归系数 2	1.0 (0.9, 1.1) 0.937
回归系数2与1的差	0.8 (0.6, 0.9) 0.001
对数似然比检验	<0.001

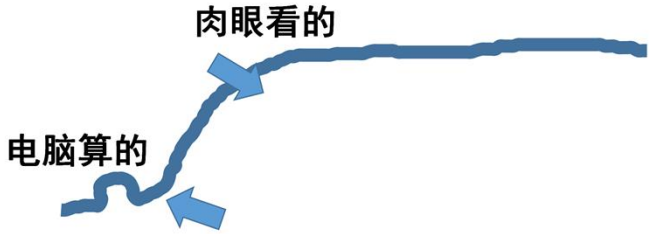
如果对数似然比<0.05，以模型2为主
反之，模型1

表中数据：HR (95% CI) Pvalue *P<0.05 **P<0.01 ***P<0.001
结果变量：OUTCOME1
暴露变量：MPV
调整变量：SEX; AGE; HEIGHT; WEIGHT; SMOKING; ALCOHOLIC; HP; DIABETES; HYACID; CVD; PLTBA

此图为易侕的饱和阈值效应模块输出，它同时提供了常规的广义线性方程结果（用一根直线拟合）及分段线性方程拟合结果，并提供了对数似然比检验。
首先我们先看对数似然比检验，如果它小于0.05，这意味着用分段线性模型去拟合XY的关系比用常规的线性模型拟合效果会更好。
此外还列出了拐点以及拐点左右两侧X和Y的线性关系，它的解读和此前的相同。

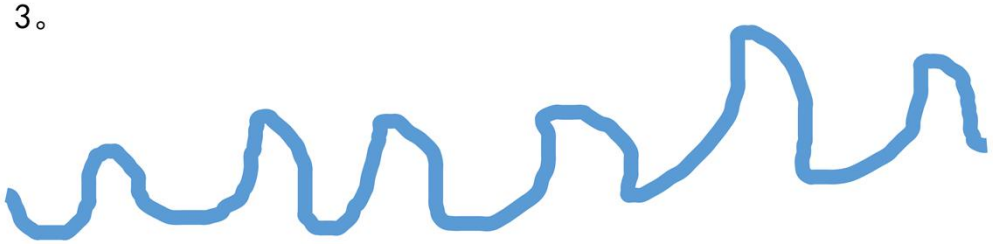
几种特殊情况

如果饱和阈值效应的结果和曲线对不上，以肉眼为准。肉眼选择拐点，再试一次。



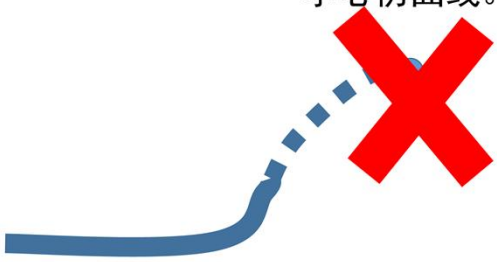
几种特殊情况：

如果曲线呈现波浪形，弯太多，将自由度调整至3。



几种特殊情况

小心伪曲线。将异常值去掉即可。



第四步

表4

排除Z的干扰

XY在不同层之间的直线关系

使用场景及报道的原则

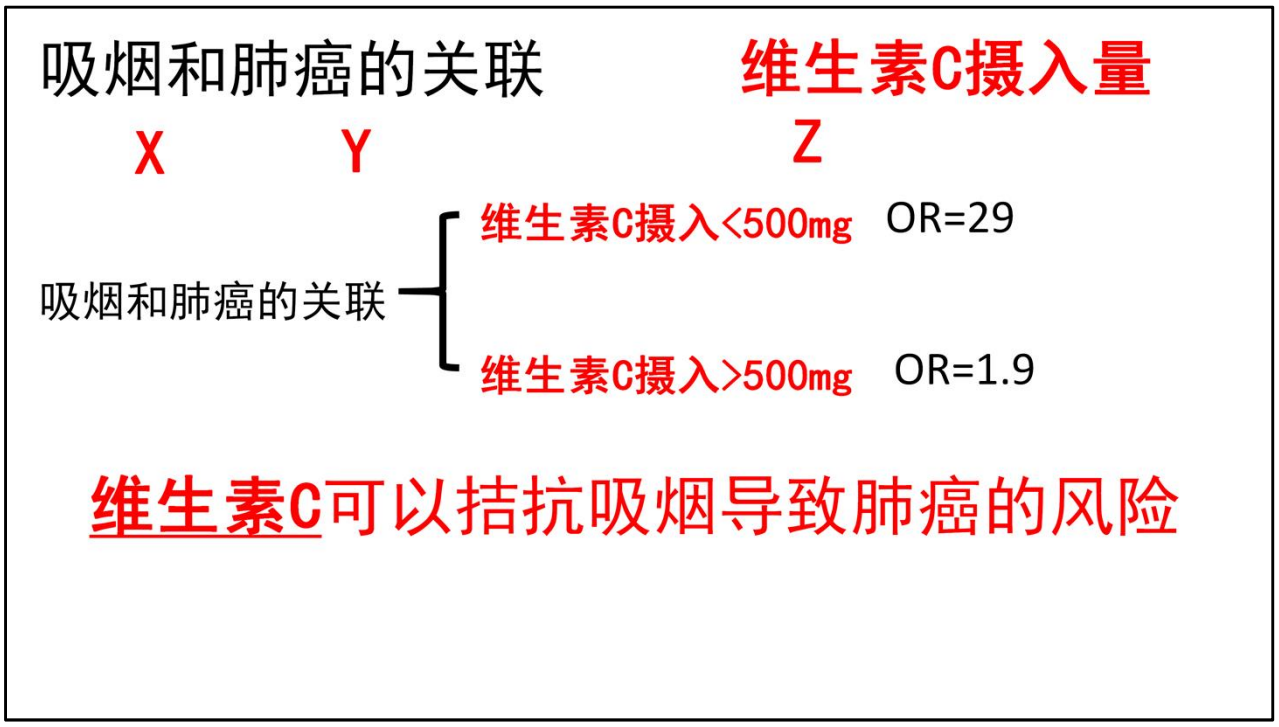
当直线和曲线均为阴性结果时，可用于补救。

报道时，需考虑临床是否可解释，是否有临床价值。

“

交互的第一种用法：突出交互本身

”



很多时候，我们的课题可能被斥为“无创新性”。好比研究吸烟与肺癌的关系。然而，当以维生素C摄入作为分层变量，可以惊奇的发现维生素C可拮抗吸烟导致肺癌的风险。此为交互的第一种做法：突出交互作用本身。即重点不再是吸烟和肺癌，而是这个效应修饰因子。

Table 3. Effect of Randomized Treatment Assignment on the Primary Outcome in Prespecified and Exploratory Subgroups

Characteristic	No. of Patients[a]		No. of Events (%)[a]		Relative Risk (95% Confidence Interval)	*P* Value for Interaction
	Active	Placebo	Active	Placebo		
Angiotensin-converting enzyme inhibitors						
Yes	627	668	102 (16.3)	125 (18.7)	0.81 (0.63-1.06)	.03
No	1957	1920	287 (14.7)	247 (12.9)	1.15 (0.97-1.36)	

X=叶酸　　Y=是否中风　　Z=是否吃ACEI

1、突出的不是叶酸是否能预防中风，而是：ACEI。

2、突出的这个焦点是否能成立的关键是交互作用的P值

3、对于每个层的效应值是否显著（P<0.05)非绝对需求

在一篇发表于JAMA的文章中，叶酸与中风在服用ACEI人群中的发现，成为这篇文章最大的亮点。

“

交互的第二种用法：舍阴取阳

”

X=缺氧诱导因子α　Y=结肠癌结局（死亡/alive）

Z=体重指数（以24作为分界）

调整体重指数后，HR=1.00（0.56-2.20）

BMI
- **BMI<24，X（缺氧诱导因子）和Y（死亡）HR=0.92（0.27-**
- **BMI>24，X（缺氧诱导因子）和Y（死亡）HR=1.27（1.04-**

避重就轻，取阳舍阴，不看交互P

第二种做法，即在不同人群中，找到“阳性结果”的人群，只对其进行报道。文章的实例为缺氧诱导因子与结肠癌患者死亡的关系。如果在全人群中，则为阴性结果。而在通过BMI分层后，则发现在胖子中，发现阳性结果。最终，该研究只需报道胖子即可。

科研假设：被动吸烟导致女性发生痛经

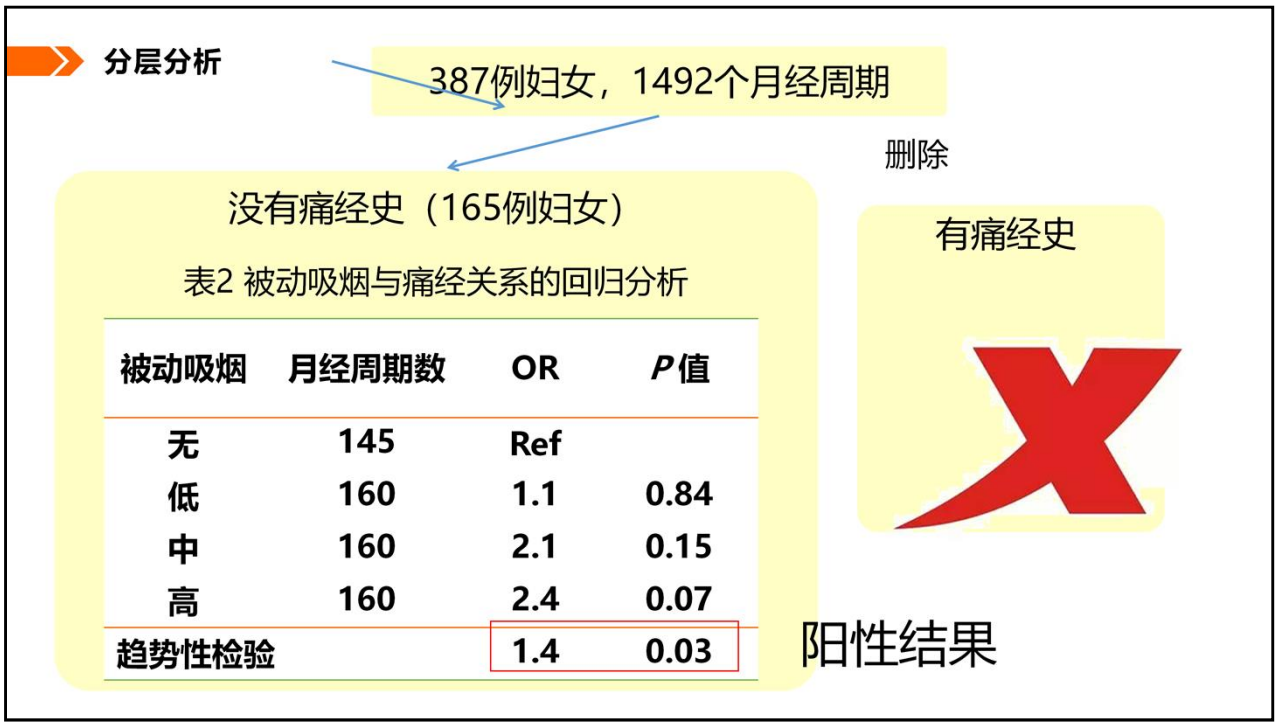
研究设计：跟踪387例妇女，1492个月经周期

表1 被动吸烟与痛经关系的回归分析

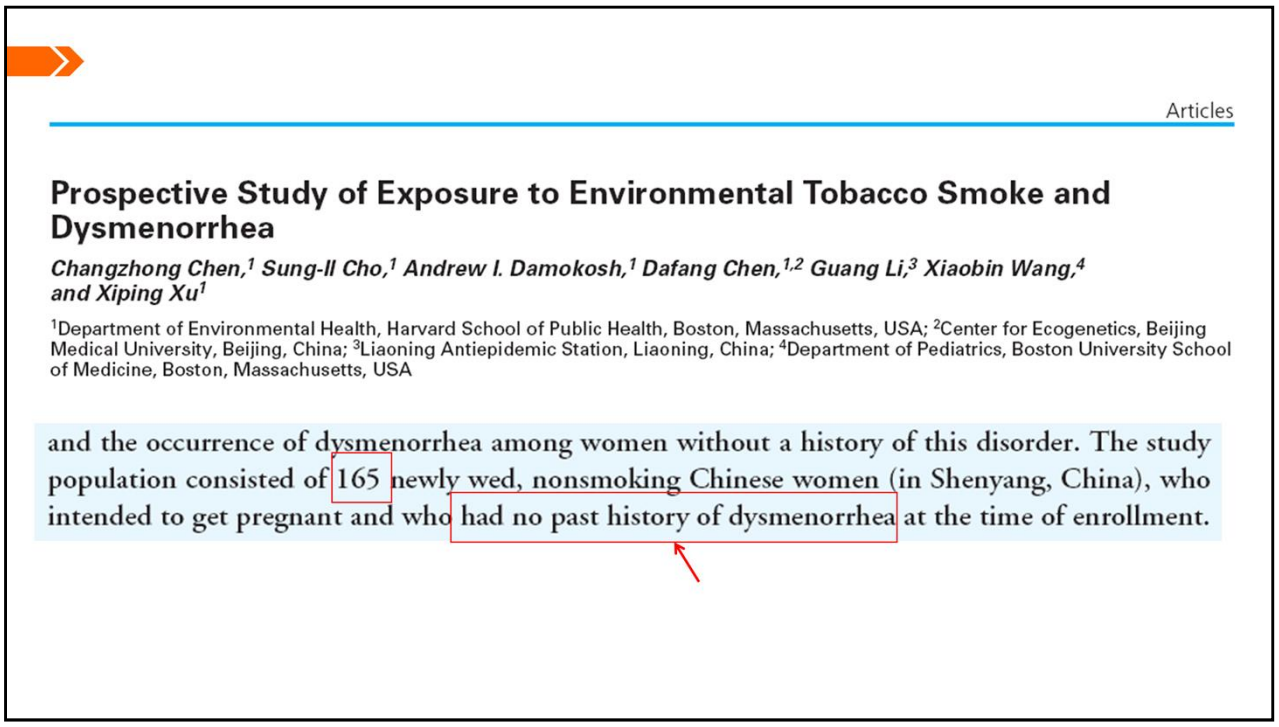
被动吸烟	月经周期数	OR	*P*值
无	370	Ref	
低	373	0.9	0.49
中	376	1.1	0.67
高	373	1.1	0.54

阴性结果

这篇文章的假设是被动吸烟导致女性发生痛经。研究设计很简单。跟踪了387例妇女，让她们记日记，记录被动吸烟和发生痛经的情况。随访到了1492个月经周期。这是最初的分析结果。这个表中的数字很多，大家可以只看红框里的，P值都大于0.05，是阴性结果。可以理解为，被动吸烟的人和没有被动吸烟的人相比，没有观察到痛经风险增加。



看到这个阴性结果之后，研究者运用分层分析，把研究人群劈成两半。是按照是否有痛经史劈开的。在没有痛经史的人中重新做一遍分析，得到阳性结果，大家只需要看红框例的两个数字。1.4表示，被动吸烟每增加一个等级，发生痛经的风险增加40%。P值是0.03，风险的增加是显著的。大家可能会问趋势检验是什么，这个问题有专题讲解，学习是循序渐进的过程，本次课大家只需掌握分层分析。删除了有痛经史的人，结果就变好了。下一步，我们要从临床意义角度解释原因。



Articles

Prospective Study of Exposure to Environmental Tobacco Smoke and Dysmenorrhea

Changzhong Chen,[1] Sung-Il Cho,[1] Andrew I. Damokosh,[1] Dafang Chen,[1,2] Guang Li,[3] Xiaobin Wang,[4] and Xiping Xu[1]

[1]Department of Environmental Health, Harvard School of Public Health, Boston, Massachusetts, USA; [2]Center for Ecogenetics, Beijing Medical University, Beijing, China; [3]Liaoning Antiepidemic Station, Liaoning, China; [4]Department of Pediatrics, Boston University School of Medicine, Boston, Massachusetts, USA

and the occurrence of dysmenorrhea among women without a history of this disorder. The study population consisted of 165 newly wed, nonsmoking Chinese women (in Shenyang, China), who intended to get pregnant and who had no past history of dysmenorrhea at the time of enrollment.

这篇文章发表的时候，经过了重新包装，没有用原始数据库的387例，而是在纳入标准中写清楚，只纳入没有痛经史的165例妇女。于是研究结果就很漂亮了。这篇文章是谁做的呢？我们看作者里有陈常中老师、陈大方老师现在北医公卫任教、徐希平老师我国的千人计划科学家。文章是2000年发表的，大牛们16年前在哈佛公卫学院时就这么玩儿了，我们也可以学他们合理删数据。这也做是完全可以的。并且好文章就是这样出来的。

交互的第三种用法：验证阴性结果

> 当XY的直线结果是阴性，曲线也未作出结果，因此，则需要将所有的Z都做一遍交互，证明这种阴性结果是稳定存在的。

举例

题目：

Statin Use and Survival After Colorectal Cancer: The Importance of Comprehensive Confounder Adjustment

X：他汀使用

Y：结肠癌的生存

特殊说明：本研究最大的亮点：confounder adjust

Table 3. Association of statin use at diagnosis with overall, colorectal cancer-specific, and recurrence-free survival stratified by age, sex, stage and location of colorectal cancer at diagnosis, and by conduct of chemotherapy

		Overall survival			CRC-specific survival		Recurrence-free survival*		
Subgroup	Statin use	No.	Events No. (%)	HR (95% CI)†	Events No. (%)	HR (95% CI)†	No.	Events No. (%)	HR (95% CI)†
Age at diagnosis, y									
<70	No	1257	290 (23)	1.00 (Ref.)	237 (19)	1.00 (Ref.)	1057	202 (19)	1.00 (Ref.)
	Yes	184	40 (22)	1.10 (0.70 to 1.74)	31 (17)	1.28 (0.77 to 2.15)	164	25 (15)	1.15 (0.65 to 2.03)
≥70	No	1028	361 (35)	1.00 (Ref.)	246 (24)	1.00 (Ref.)	880	211 (24)	1.00 (Ref.)
	Yes	228	78 (34)	1.19 (0.87 to 1.61)	49 (21)	1.13 (0.77 to 1.65)	204	39 (19)	0.85 (0.55 to 1.32)
Sex									
Male	No	1341	380 (28)	1.00 (Ref.)	276 (21)	1.00 (Ref.)	1125	237 (21)	1.00 (Ref.)
	Yes	272	74 (27)	1.00 (0.71 to 1.39)	46 (17)	0.93 (0.62 to 1.41)	243	37 (15)	0.80 (0.49 to 1.30)
Female	No	944	271 (29)	1.00 (Ref.)	207 (22)	1.00 (Ref.)	812	176 (22)	1.00 (Ref.)
	Yes	140	44 (31)	1.32 (0.89 to 1.96)	34 (24)	1.49 (0.94 to 2.35)	125	27 (22)	1.11 (0.66 to 1.84)

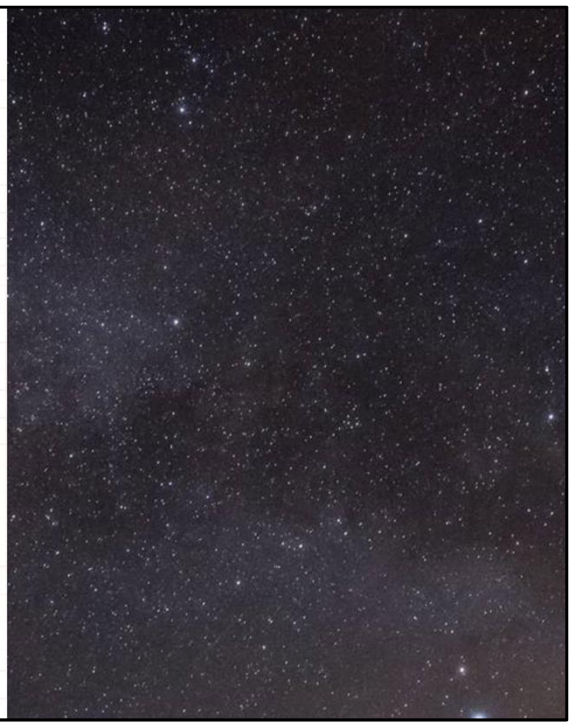
此表为作者在不同人群中对他汀使用与生存的关联进行验证。譬如在小于70岁和大于70岁的人群中，都没有发现他汀可提高生存，同理，在性别中亦没有发现。这种不同人群的检验，印证了该文章的“阴性”结论，即未能观察到他汀与结直肠癌患者的更长总体生存有关。

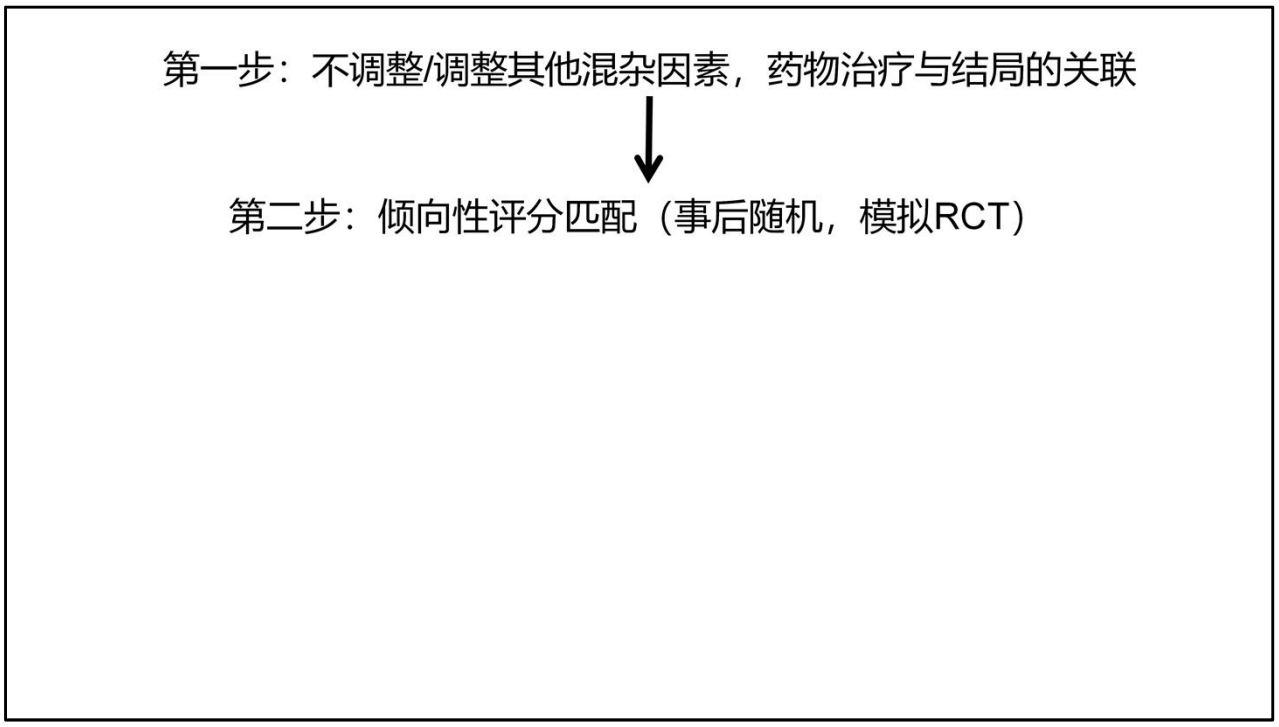
UICC stage									
Stage I + II	No	1165	174 (15)	1.00 (Ref.)	82 (7)	1.00 (Ref.)	1165	158 (14)	1.00 (Ref.)
	Yes	243	34 (14)	1.07 (0.66 to 1.72)	10 (4)	0.97 (0.43 to 2.19)	241	16 (7)	0.50 (0.26 to 0.95)
Stage III	No	771	210 (27)	1.00 (Ref.)	155 (20)	1.00 (Ref.)	771	255 (33)	1.00 (Ref.)
	Yes	126	48 (38)	1.14 (0.75 to 1.75)	35 (28)	1.22 (0.75 to 2.00)	126	48 (38)	1.25 (0.82 to 1.92)
Stage IV	No	348	267 (77)	1.00 (Ref.)	246 (71)	1.00 (Ref.)	--	--	--
	Yes	43	36 (84)	1.04 (0.67 to 1.63)	35 (81)	1.07 (0.68 to 1.70)	--	--	--
Location of CRC									
Proximal colon	No	701	205 (29)	1.00 (Ref.)	141 (20)	1.00 (Ref.)	604	110 (18)	1.00 (Ref.)
	Yes	145	47 (32)	1.08 (0.71 to 1.64)	32 (22)	1.01 (0.60 to 1.69)	124	18 (15)	0.78 (0.48 to 1.26)
Distal colon	No	631	175 (28)	1.00 (Ref.)	129 (20)	1.00 (Ref.)	523	103 (20)	1.00 (Ref.)
	Yes	110	27 (25)	0.95 (0. 57 to 1.59)	17 (15)	1.01 (0.53 to 1.95)	98	17 (17)	1.24 (0.71 to 2.14)
Rectum	No	946	267 (28)	1.00 (Ref.)	209 (22)	1.00 (Ref.)	805	198 (25)	1.00 (Ref.)
	Yes	156	44 (28)	1.25 (0.84 to 1.87)	31 (20)	1.35 (0.84 to 2.15)	145	29 (27)	1.16 (0.76 to 1.76)
Chemotherapy									
No	No	1325	330 (25)	1.00 (Ref.)	221 (17)	1.00 (Ref.)	1195	199 (17)	1.00 (Ref.)
	Yes	272	66 (24)	1.29 (0.93 to 1.79)	36 (13)	1.26 (0.82 to 1.93)	256	30 (12)	0.99 (0.67 to 1.47)
Yes	No	954	317 (33)	1.00 (Ref.)	258 (27)	1.00 (Ref.)	738	212 (29)	1.00 (Ref.)
	Yes	139	52 (37)	1.03 (0.69 to 1.52)	44 (32)	1.08 (0.70 to 1.67)	112	34 (30)	1.21 (0.82 to 1.78)

同上

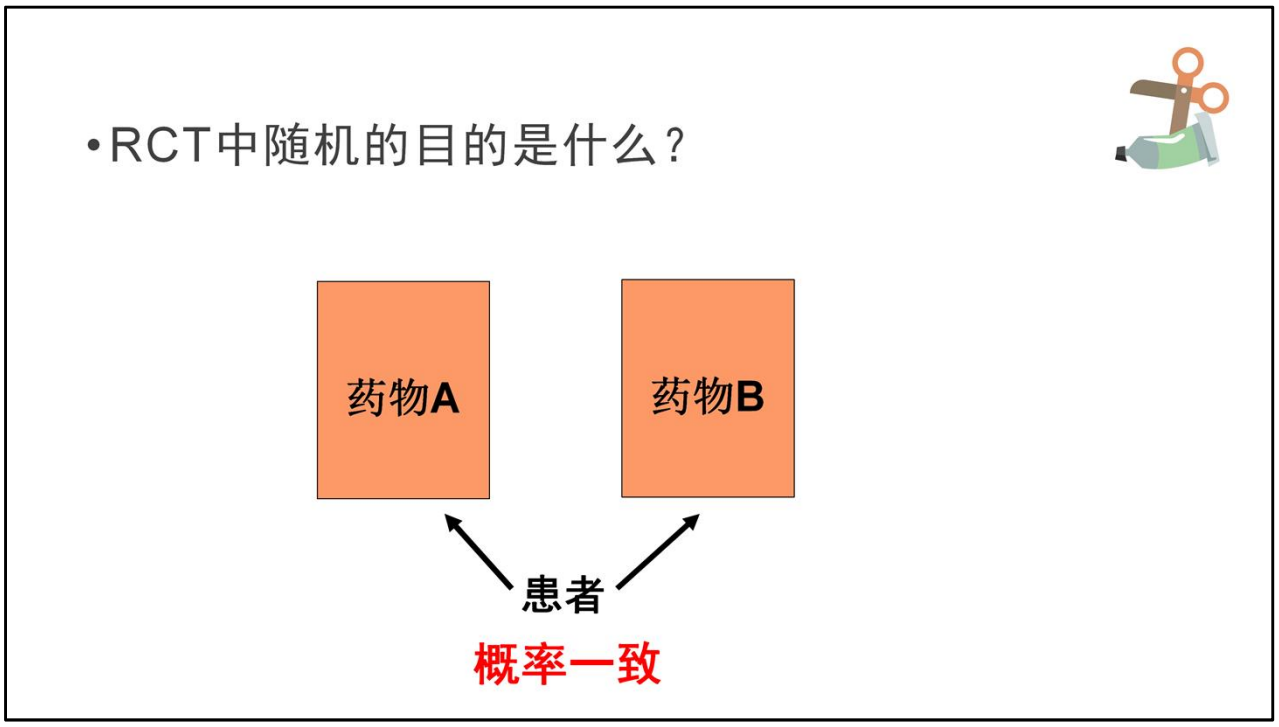
X：两种药物
Y：某种结局
Z：其他的协变量

真实世界研究的
数据分析

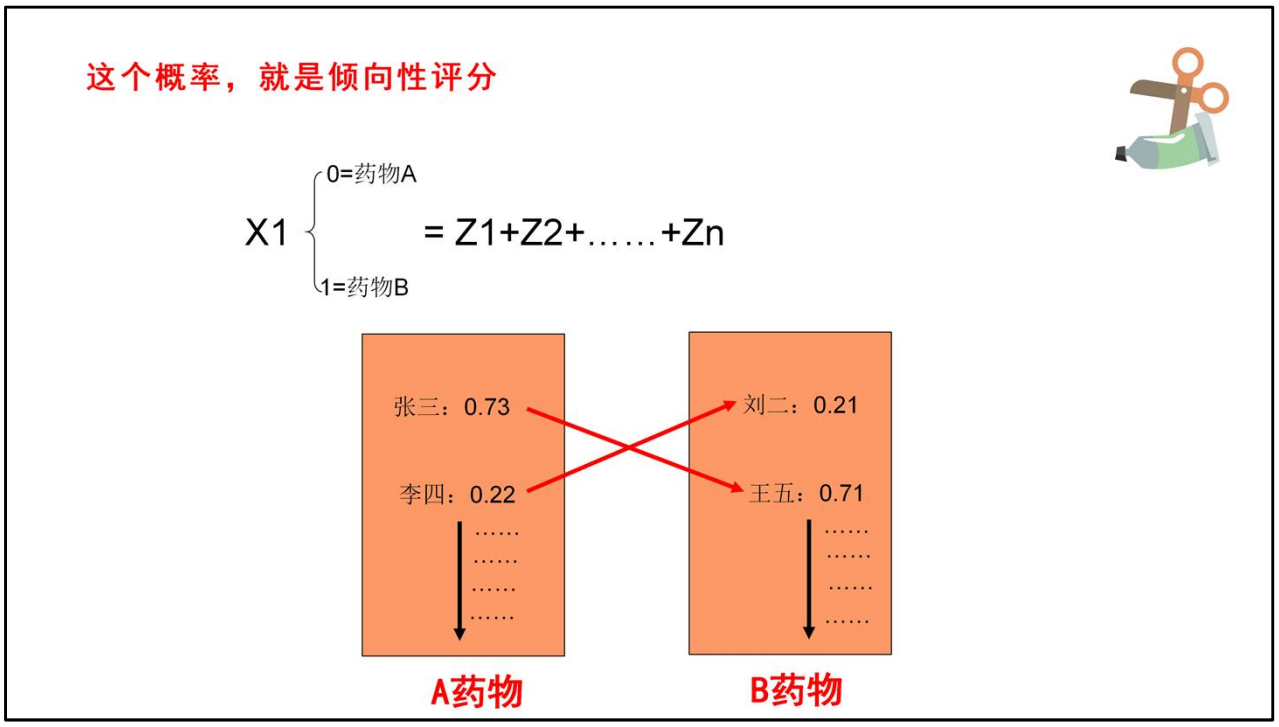




我们展示了真实世界研究最常用的套路。第一先用传统的回归方程调整混杂因素以后，得到X和Y的线性关系，第二步则是用倾向性评分匹配，他们目的是为了模拟随机化，又叫做事后随机

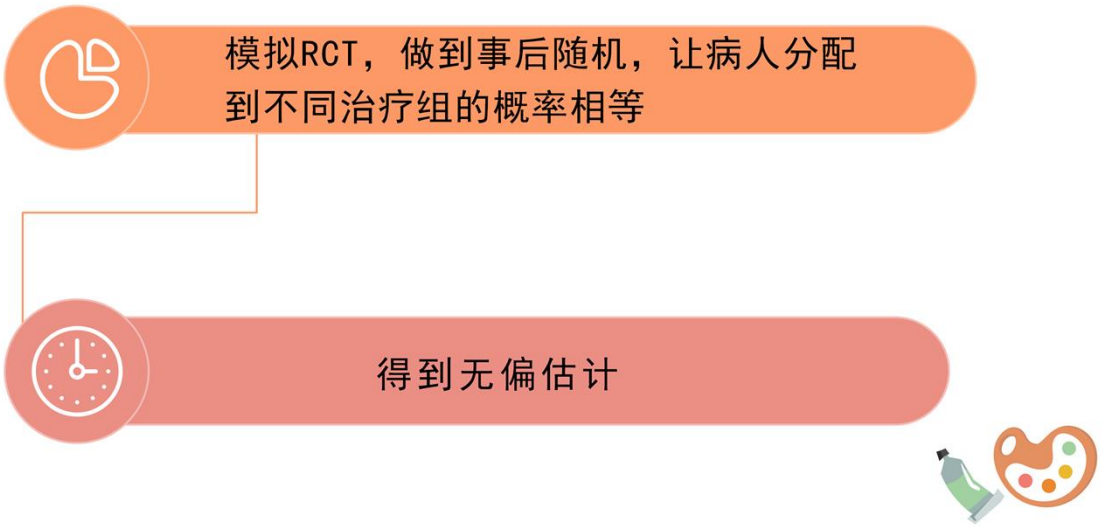


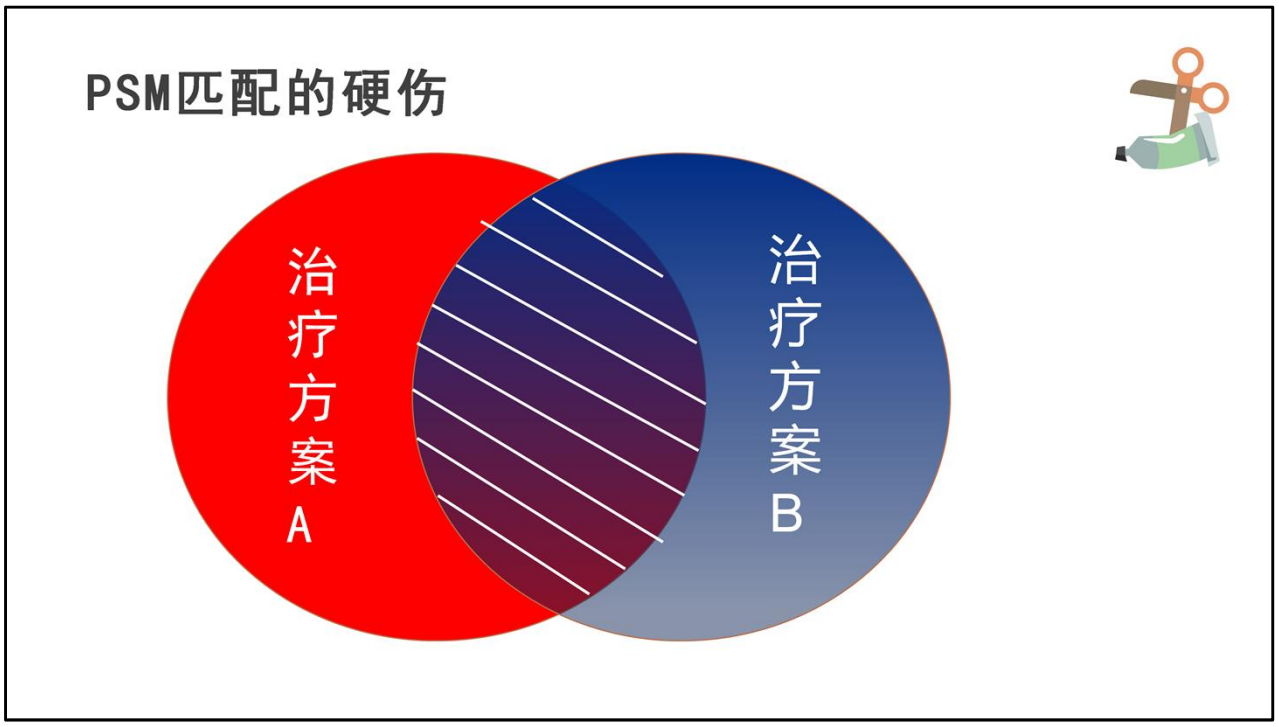
RCT随机的目的是让患者被分到两种药物组的概率相同。



既然在真实世界人群中，无法做到随机，那么，可否将每一位患者的概率算出来，最后，进行概率相同患者的匹配呢？即人为的模拟随机化，在一个真实世界人群中，创造出一个“人工随机”的数据集出来。这个概率，即为倾向性评分。

非随机化CER中，倾向性评分匹配的真正意义





当然，倾向性评分匹配用于模拟随机化，也是有其短板的。如果失匹配的人太多，则不可避免的会有选择偏倚的风险。

因此，我们采用另外一种方法，逆概率加权。该方法同样基于倾向性评分，然而，他与倾向性评分匹配不同的是，这种加权 并不会造成样本量的损失。

我们在易侕3.0的“真实世界处理效应分析”可轻松的一键完成上述所有操作。

研究人群描述

Distribution of baseline covariates by treatment:

Variables	手术入路 recoded: 0	手术入路 recoded: 1	Standardized diff.	P value
年龄	(452) 59.34 ± 12.21	(486) 57.24 ± 11.29	0.1792	0.0062
性别			0.0765	0.2691
1	235 (51.8)	233 (47.9)		
2	219 (48.2)	253 (52.1)		
CD117				0.5589
0	15 (3.3)	19 (3.9)	0.0325	
1	27 (5.9)	22 (4.5)	0.0638	
2	412 (90.7)	445 (91.6)	0.0287	
CD34				<0.0001
0	14 (3.1)	21 (4.3)	0.0656	

易倆输出结果中，包括了研究人群描述

表2：单因素分析（可放入文章附表）

Univariate analysis for treatment and covariates:

Exposure	Statistics	是否复发
手术入路 recoded		
0	454 (48.30%)	1.0
1	486 (51.70%)	0.07 (0.02, 0.23) <0.0001
性别		
1	468 (49.79%)	1.0
2	472 (50.21%)	0.41 (0.22, 0.74) 0.0035
年龄	58.25 ± 11.78	1.02 (1.00, 1.05) 0.0552
X5.GROUP		

单因素分析。即每一个自变量（X和Z）与Y的关联。但未调整任何混杂。

老回归结果

Associations of treatment with outcomes

Model	是否复发
1: Crude	0.07 (0.02, 0.23) <0.0001
2: Adjust for all covariates	0.56 (0.24, 1.34) 0.1936
3: Adjust for PS0	0.17 (0.05, 0.58) 0.0046
4: Adjust PS0(smooth)	0.16 (0.05, 0.55) 0.0037

最后，易倆的输出结果中，可看到不调增模型，调整所有协变量模型，调整倾向性评分的模型。

倾向性评分匹配结果

Estimate of treatment effect using PS Match

	是否复发
ATT	0.17 (0.05, 0.65) 0.0090
ATC	0.17 (0.03, 0.89) 0.0354
ATE	0.17 (0.04, 0.71) 0.0152

Results in table: HR (95%CI) Pvalue
Cluster-robust standard errors (Liang and Zeger 1986) were applied for calculating 95% CI.
Match for PS0
ATT: average treatment effect for treated
ATC: average treatment effect for control
ATE: average treatment effect for all

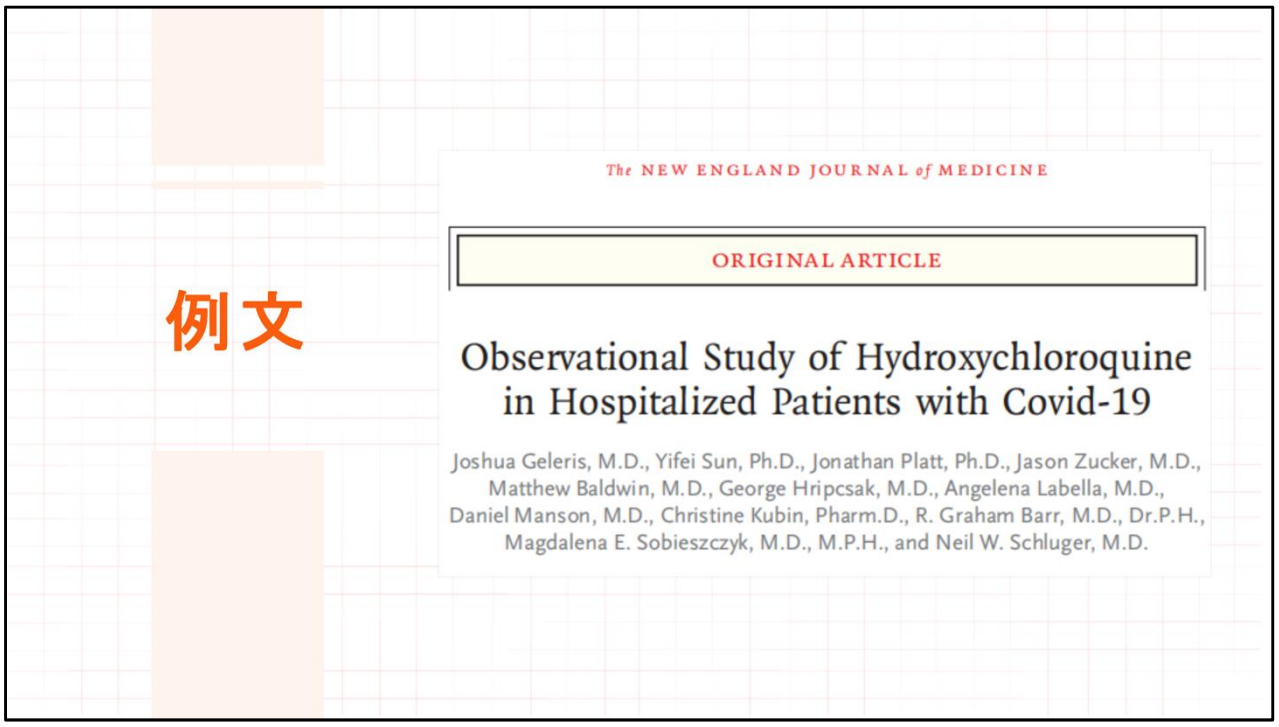
倾向性评分匹配后的模型结果。

Estimate of treatment effects using IPTW

IPW using PS0	是否复发 IPTW的结果
ATT	0.16 (0.04, 0.57) 0.0050
ATC	0.15 (0.03, 0.68) 0.0142
ATE	0.16 (0.04, 0.59) 0.0064

Results in table: HR (95%CI) Pvalue

逆概率加权后的模型



为了帮助大家更为直观的理解真实世界模块的用法，我们用一篇发表与新英格兰杂志的文章作为例文。该文章旨在观察用/不用羟氯喹能否降低COVID19感染者的死亡。

表1：研究人群描述

Table 1. Characteristics of Patients Receiving or Not Receiving Hydroxychloroquine, before and after Propensity-Score Matching.*

Characteristic	Unmatched Patients		Propensity-Score–Matched Patients†	
	Hydroxychloroquine (N=811)	No Hydroxychloroquine (N=565)	Hydroxychloroquine (N=811)	No Hydroxychloroquine (N=274)
Age — no. (%)				
<40 yr	80 (9.9)	105 (18.6)	80 (9.9)	28 (10.2)
40–59 yr	217 (26.8)	142 (25.1)	217 (26.8)	69 (25.2)
60–79 yr	367 (45.3)	220 (38.9)	367 (45.3)	118 (43.1)
≥80 yr	147 (18.1)	98 (17.3)	147 (18.1)	59 (21.5)
Female sex — no. (%)	337 (41.6)	258 (45.7)	337 (41.6)	113 (41.2)
Race and ethnic group — no. (%)‡				
Non-Hispanic white	74 (9.1)	57 (10.1)	97 (12.0)	36 (13.1)
Non-Hispanic black	89 (11.0)	92 (16.3)	120 (14.8)	40 (14.6)
Hispanic	412 (50.8)	286 (50.6)	530 (65.4)	172 (62.8)
Other	48 (5.9)	36 (6.4)	64 (7.9)	26 (9.5)

文章中，作者用了研究人群描述模块，其中，作者将匹配前和匹配后的结果同时列出。两者在易侕的文件中均有。研究者可选择合并报道或将匹配后的人群描述放在附表中。

Table 2. Associations between Hydroxychloroquine Use and the Composite End Point of Intubation or Death in the Crude Analysis, Multivariable Analysis, and Propensity-Score Analyses.

Analysis	Intubation or Death
No. of events/no. of patients at risk (%)	
Hydroxychloroquine	262/811 (32.3)
No hydroxychloroquine	84/565 (14.9)
Crude analysis — hazard ratio (95% CI)	2.37 (1.84–3.02)
Multivariable analysis — hazard ratio (95% CI)*	1.00 (0.76–1.32)
Propensity-score analyses — hazard ratio (95% CI)	
With inverse probability weighting†	1.04 (0.82–1.32)
With matching‡	0.98 (0.73–1.31)
Adjusted for propensity score§	0.97 (0.74–1.28)

例文中的表2，主要包括了未调整模型，全部调整协变量后的模型，倾向性评分匹配、加权、调整后的模型。这些，均在易倆的模块输出结果中。

老回归结果

Associations of treatment with outcomes

Model	是否复发
1: Crude	0.07 (0.02, 0.23) <0.0001
2: Adjust for all covariates	0.56 (0.24, 1.34) 0.1936
3: Adjust for PS0	0.17 (0.05, 0.58) 0.0046
4: Adjust PS0(smooth)	0.16 (0.05, 0.55) 0.0037

其中的crude模型，对应的就是未调整模型；

Adjust for all covariates对应的就是文章中的multivariate analysis

Adjust for PSO，即对应的是文章中的adjusted for propensity score。

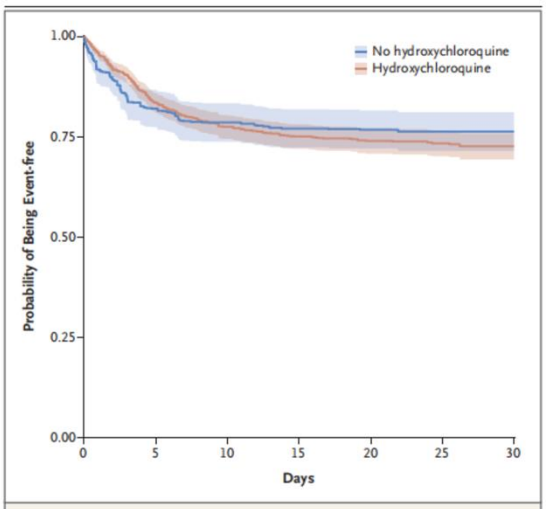


Figure 2. Freedom from Composite End Point of Intubation or Death.
The shaded areas represent pointwise 95% confidence intervals.