

基于数据挖掘开展临床科研的思路 与软件实现

魏晟 副教授

we2008cn@gmail.com

流行病学与卫生统计学系主任
华中科技大学同济医学院公共卫生学院

实例

临床问题 → 解决方案

实例

临床问题

有数据

上海第一人民医院
肝移植病例资料



有方向

结果变量 (Y)
术后急性肾功能衰竭



没头绪

因变量 (X)?
导致肾功衰的原因?



实例

临床问题



解决方案



数据挖掘

实例

临床问题 → 数据挖掘

整理原始数据

ID	GEN DER	AGE	ABO	GRA DE	ETIO LOG Y	ACU TERF	ANH EPAT IC	WAR MTIM E	COL DTIM E	OPTI ME	CRY OPR ECIPI TATE	WHO LEBL OOD	RED BLO OD	FRES HPL ASM A	PLAT ELET
1	1	53	1	1	2	1	60	3	420	450	20	0	8	10	0
2	1	40	0	1	2	1	60	2	360	540	8	0	12	0	0
3	1	45	1	1	2	1	55	3	480	425	0	0	12	0	0
4	1	41	1	1	2	1	45	5	600	420	0	0	20	3	1
5	1	45	1	1	2	1	60	4	600	420	5	0	0	0	1
6	2	30	1	1	4	1	65	2	540	450	16	0	14	0	0
7	1	38	1	1	2	0	55	5	540	420	0	0	22	10	2
8	1	32	1	1	2	0	60	4	420	420	0	0	6	0	2
9	1	42	1	1	2	0	55	3	540	390	0	6	0	0	2
10	1	40	1	1	2	0	60	4	600	450	0	0	12	0	0
11	1	44	1	1	2	0	60	4	600	480	5	0	10	12	1
12	1	49	1	1	2	0	50	4	600	420	0	0	6	0	0
13	1	47	1	1	2	0	50	4	540	405	0	0	0	3	0
14	1	55	0	1	2	0	60	4	480	415	0	4	12	0	2
15	1	38	1	1	2	0	70	4	720	600	8	10	20	20	2
16	1	41	1	1	2	0	120	3	540	470	0	0	6	2	1
17	2	56	0	1	2	0	85	3	420	470	0	0	14	2	1

实例

临床问题 → 数据挖掘

诊断数据关联关系






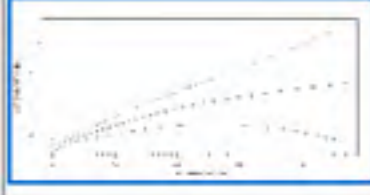

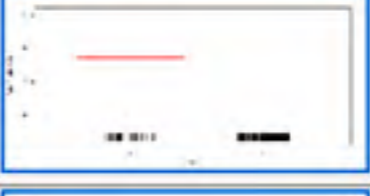


扫描数据



临床问题 → 数据挖掘

诊断数据关联关系

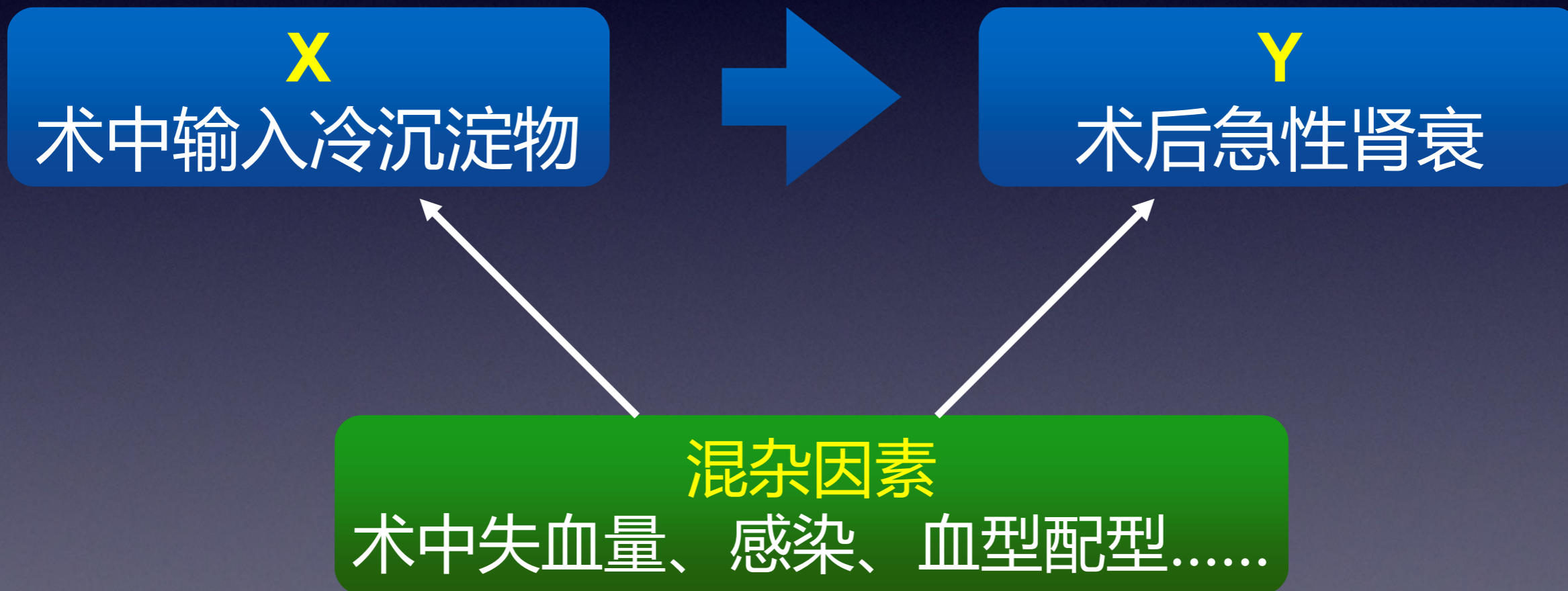
结果：扫描数据库中所有变量，按显著性水平排序

*** 0.0126 (0)	BLOODLOSS	BLOODLOSS s(BLOODLOSS)	OR 1.00 edf 4.04	95%low 1.00 Ref. df 4.97	95%upp 1.00 Chi. sq 33.54	p 0.0080 p-value 0		
** 0.0023 (0.0039)	CRYOPRECIPITATE	CRYOPRECIPITATE s(CRYOPRECIPITATE)	OR 1.00 edf 1.49	95%low 1.00 Ref. df 1.83	95%upp 1.00 Chi. sq 10.68	p 0.0015 p-value 0.0039		
** 0.0036	INFECTIONS	factor(INFECTIONS)	OR 3.00	95%low 2.00	95%upp 6.00	p 0.0020		
* 0.0118	ABO	factor(ABO)	OR 0.00	95%low 0.00	95%upp 1.00	p 0.0078		
* 0.046	STEROID	factor(STEROID)	OR 3.00	95%low 1.00	95%upp 15.00	p 0.0932		
0.2073	PLATELET	factor(PLATELET)1 factor(PLATELET)2 factor(PLATELET)3 factor(PLATELET)4	OR 2.00 1.00 0.00 1.00	95%low 1.00 0.00 0.00	95%upp 4.00 2.00 Inf 12.00	p 0.0993 0.3079 0.9893 0.7641		

实例

临床问题 → 数据挖掘

明确科研假设



临床问题 → 数据挖掘

分析思路

x与y是什么样的关系？
还有哪些因素与y有关系？

什么因素影响x与y的关系？
加强或减弱x对y的作用？

x对y有没有独立作用？
独立作用大小是多少？

实例

临床问题 → 数据挖掘 → 曲线拟合

x与y是什么样的关系？
还有哪些因素与y有关系？

曲线拟合

平滑曲线拟合

标题: Fig 3 Smoothing of Cryoprecipitate with ART

选择分析对象: : 所有数据记录

应变变量(Y)

变量名	分布	联系函数
ACUTERF	Binomial	Logit

危险因素(X)

变量	自由度
Cryoprecipitate amount	.

Cox 模型生存分析(事件=1)

选择时间变量:

或开始时间:

结束时间:

Cox模型条件logistic回归

黑白图

曲线拟合分层因子:

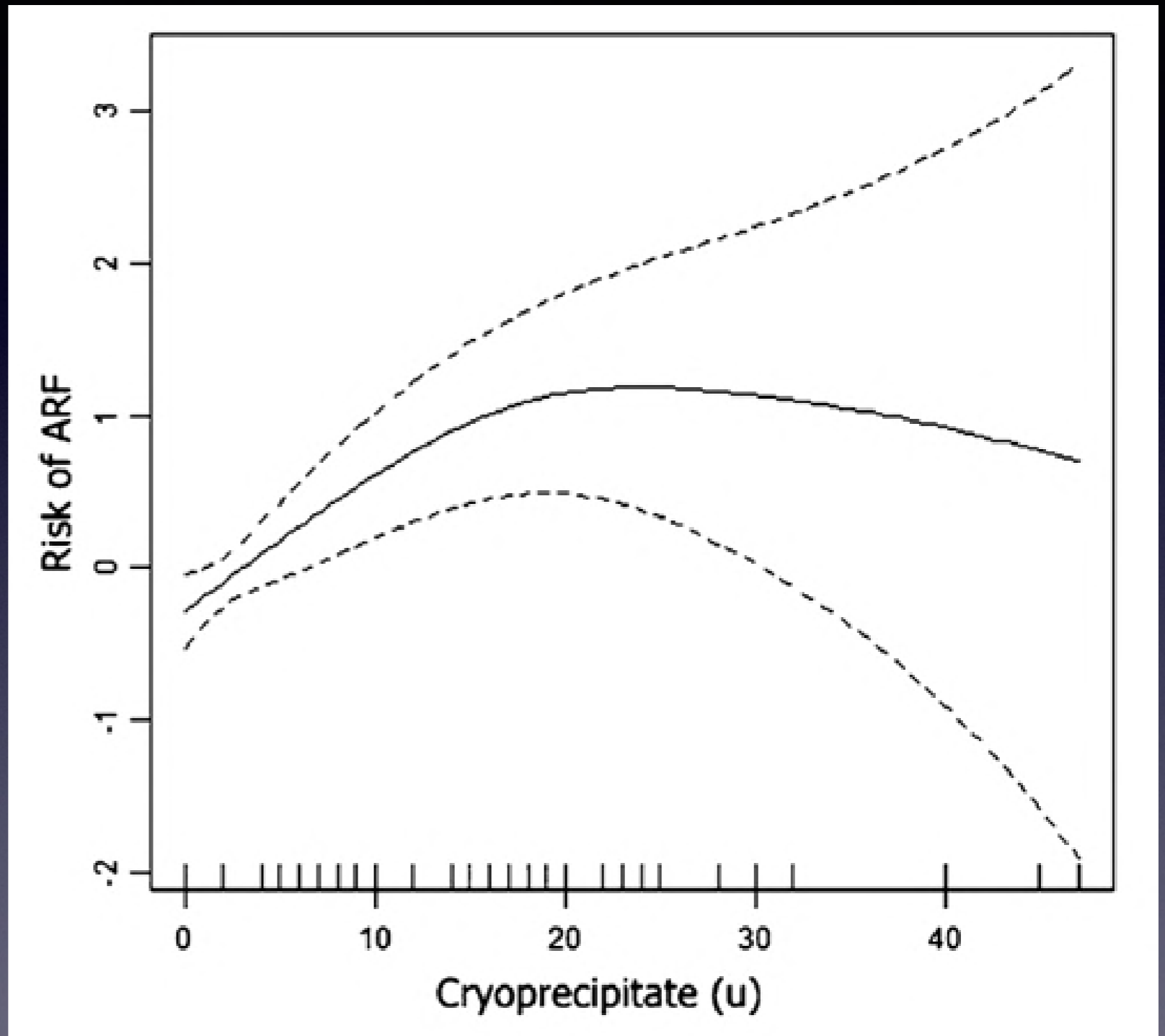
刷新 保存 查看结果

实例

临床问题 → 数据挖掘 → 曲线拟合

x与y是什么样的关系？
还有哪些因素与y有关系？

曲线拟合



实例

临床问题 → 数据挖掘 → 单因素分析

x与y是什么样的关系？
还有哪些因素与y有关系？

x与y是什么样的关系？
还有哪些因素与y有关系？

自动寻找曲线拐点

分段线性回归用于阈值效应或饱和效应分析

I. 标题: Table 4 Threshold effect

II. 选择分析对象: 所有数据记录

III. 选择结果变量 (Y): Cox 模型生存分析(事件=1)

变量	分布	联系函数
ACUTERF	Binomial	Logit

IV. 选择危险因素 (X):

选择时间变量:

或, 开始时间:

结束时间:

VI. 选择分层变量:

选择输出内容与格式: β (95%CI) Pvalue / OR (95%CI) Pvalue

精确到小数点: 0.1

如用GEE

研究对象编号:

内部相关类型:

选择或输入拐点(如3.8):

自动确定连接两条连线的拐点

Bootstrap 计算拐点可信区间

V. 选择调整变量:

变量
ETIOLOGY.NEW
Child-Pugh Grade A/B vs C
Blood loss
ABO
INFECTIONS

刷新 保存 查看结果

实例

临床问题 → 数据挖掘 → 单因素分析

x与y是什么样的关系？
还有哪些因素与y有关系？

x与y是什么样的关系？
还有哪些因素与y有关系？

单因素分析

逐个危险因素回归分析

I. 标题: Table 2 Effect of risk factors on ARF

II. 选择分析对象: 所有数据记录

III. 选择结果变量 (Y): Cox 模型生存分析(事件=1)

Variable	分布	联系函数
ACUTERF	Binomial	Logit

IV. 选择危险因素 (X):

- ETIOLOGY
- ABO
- Child-Pugh Grade A/B vs C
- Blood loss
- Whole blood
- Red blood cells
- FRESHPLASMA
- Platelet
- Cryoprecipitate amount
- Cell saver, ml
- Immunosuppressive protocol
- STEROID
- INFECTIONS

选择时间变量:

或, 开始时间:

结束时间:

V. 选择调整变量:

变量

VI. 选择列分层变量:

VII. 选择行分层变量:

VIII. 选择输出内容与格式: β (95%CI) Pvalue / OR (95%CI) Pvalue

如用GEE

研究对象编号:

内部相关类型:

X. 精确到小数点: 0.1

刷新 保存 查看结果

实例

临床问题 → 数据挖掘 → 单因素分析

x与y是什么样的关系？
还有哪些因素与y有关系？

x与y是什么样的关系？
还有哪些因素与y有关系？

单因素分析

Variables	Total	Odds ratio (95 % CI)	<i>p</i> value
Etiology, <i>n</i> (%)			
Others	12 (3.1 %)	0.74 (0.093, 6.0)	0.780
Fulminant hepatic failure	30 (7.7)	0.28 (0.037, 2.1)	0.221
Carcinoma	81 (20.8)	1.4 (0.69, 2.9)	0.341
Cirrhosis	266 (68.4)	1	
ABO blood group, <i>n</i> (%)			
Compatible	334 (85.9)	0.37 (0.18, 0.77)	0.008*
Incompatibility	55 (14.1)	1	
Child-Pugh grade, <i>n</i> (%)			
Grade C	302 (77.6)	0.94 (0.45, 2.0)	0.882
Grade A/B	87 (22.4)	1	
Blood loss, mean (SD), 500 ml	6.9 (4.8)	1.1 (1.0, 1.1)	0.008*
Transfusion of blood products			
Whole blood, mean (SD), U	1.6 (3.7)	1.0 (0.97, 1.1)	0.277
Red blood cells, mean (SD), U	9.9 (7.1)	1.0 (0.98, 1.1)	0.384
Fresh frozen plasma, mean (SD), U	1.7 (4.0)	0.99 (0.9, 1.1)	0.804
Platelet, mean (SD), U	0.6 (1.0)	0.92 (0.65, 1.3)	0.628
Cryoprecipitate, mean (SD), U	5.2 (7.9)	1.1 (1.0, 1.1)	0.002*
Cell saver, mean (SD), ml	554 (1,283)	1.0 (1.0, 1.0)	0.561
Immunosuppressive protocol, <i>n</i> (%)			
Including cyclosporine A	107 (27.5)	1.0 (0.51, 2.0)	0.950
Including FK506	273 (70.2)	1	
Steroid, <i>n</i> (%)			
Used after OLT	337 (86.6)	3.5 (0.81, 14.8)	0.093
Not used after OLT	52 (13.4)	1	
Infections, <i>n</i> (%)			

实例

临床问题 → 数据挖掘 → 扫描交互作用

x与y是什么样的关系？
还有哪些因素与y有关系？

x与y是什么样的关系？
还有哪些因素与y有关系？

什么因素影响x与y的关系？
加强或减弱x对y的作用？

扫描交互作用

快速扫描交互作用

标题: 快速扫描交互作用

选择分析对象: 所有数据记录

结局变量(Y)

变量名	分布	联系函数
ACUTERF	Binomial	

危险因素(X): cryoprecipitate 1

筛查交互作用变量

- GENDER
- AGE
- ABO
- Anhepatic phase
- ETIOLOGY
- Child-Pugh Grade A/B vs C
- Immunosuppressive protocol
- INFECTIONS
- FRESHPLASMA
- Whole blood
- Red blood cells
- Warm ischemia time
- Cold ischemia time
- Cell saver, ml
- STEROID

Cox 模型生存分析(事件=1)

选择时间变量: []

或开始时间: []

结束时间: []

如用GEE

研究对象编号: []

内部相关类型: []

调整变量

变量: []

精确到小数点: 0.1

刷新 保存 查看结果

扫描交互作用

实例

临床问题 → 数据挖掘 → 协变量筛选

x与y是什么样的关系？
还有哪些因素与y有关系？

x与y是什么样的关系？
还有哪些因素与y有关系？

什么因素影响x与y的关系？
加强或减弱x对y的作用？

调整 (协) 变量检查与筛选

I. 标题: 协变量检查与筛选

II. 选择分析对象: 所有数据记录

III. 选择结果变量 (Y): Cox 模型生存分析(事件=1)

Variable	分布	联系函数
ACUTERF	Binomial	Logit

IV. 选择危险因素 (X):

Variable
Cryoprecipitate amount

选择时间变量:

或, 开始时间:

结束时间:

V. 固定要调整的变量 (A):

Variable
AGE
GENDER

VII. 选择分层变量:

如用GEE

研究对象编号:

内部相关类型:

VIII. 协变量筛选标准

协变量与结果变量回归系数的 p 值

在单因素模型中 P<0.05

或 和

危险因素回归系数的变化

>= 10% 如加协变量入基本模型

刷新 保存 查看结果

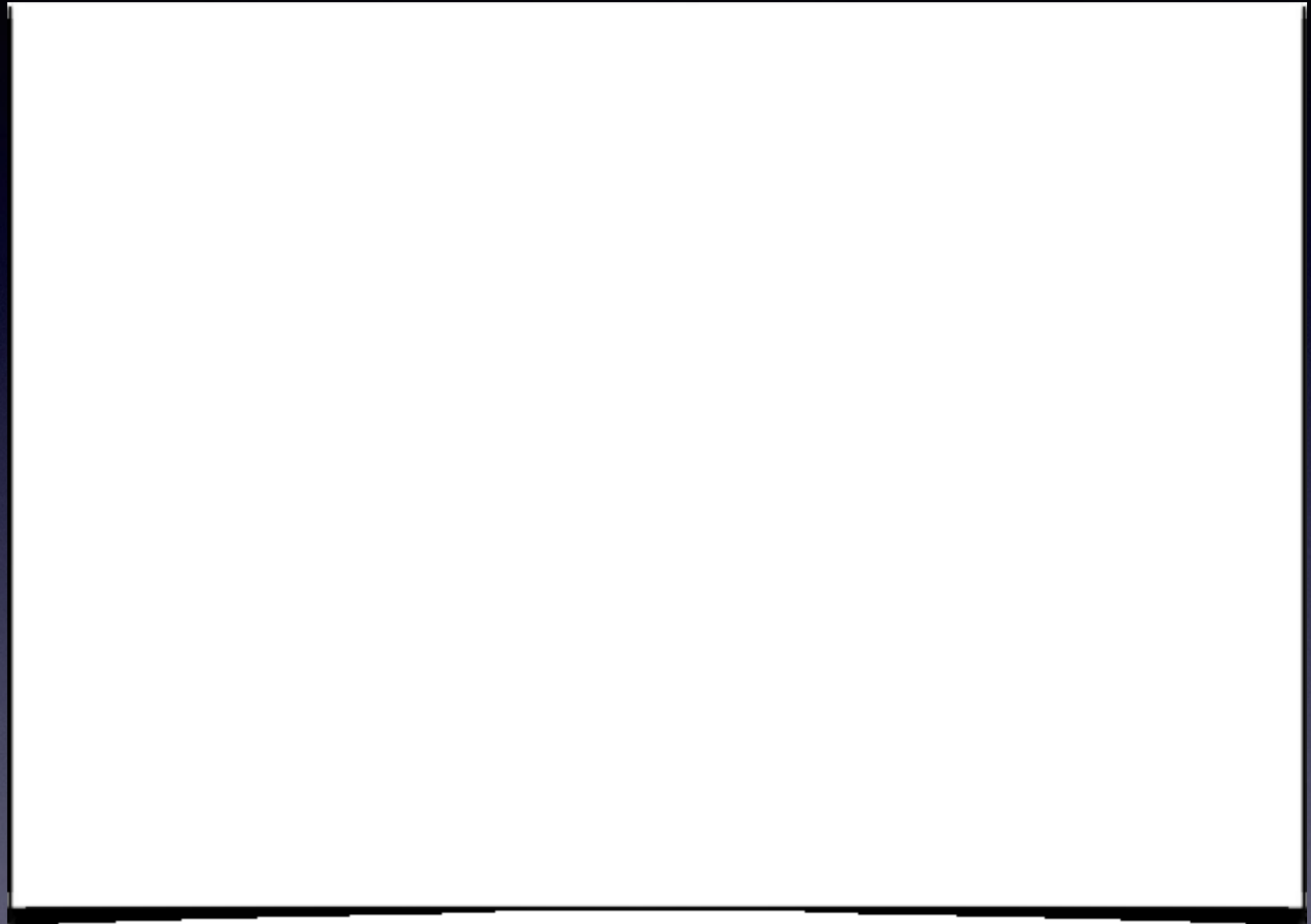
实例

临床问题 → 数据挖掘 → 协变量筛选

x与y是什么样的关系？
还有哪些因素与y有关系？

x与y是什么样的关系？
还有哪些因素与y有关系？

什么因素影响x与y的关系？
加强或减弱x对y的作用？



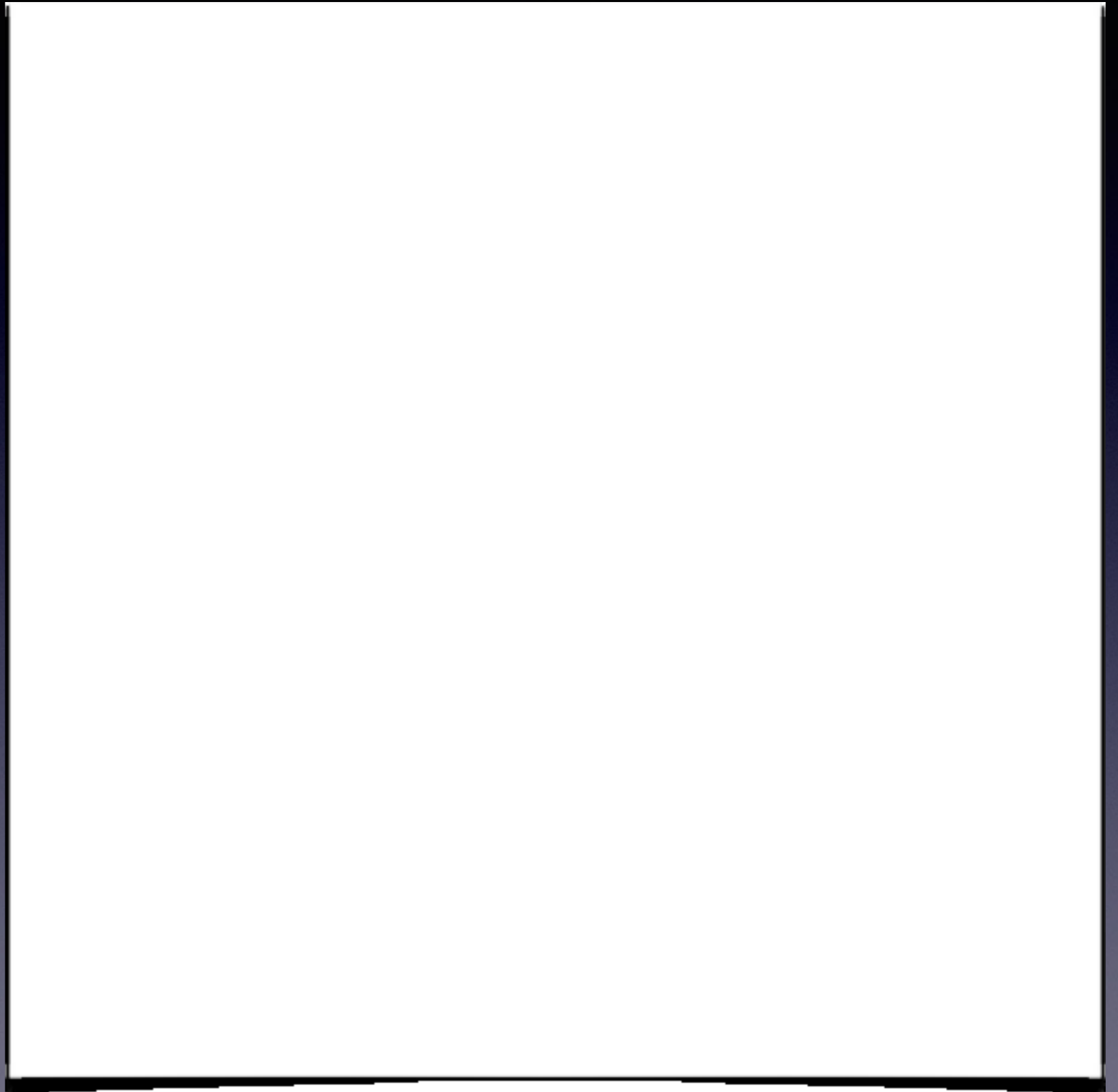
实例

临床问题 → 数据挖掘 → 多个回归方程

x与y是什么样的关系？
还有哪些因素与y有关系？

x与y是什么样的关系？
还有哪些因素与y有关系？

什么因素影响x与y的关系？
加强或减弱x对y的作用？



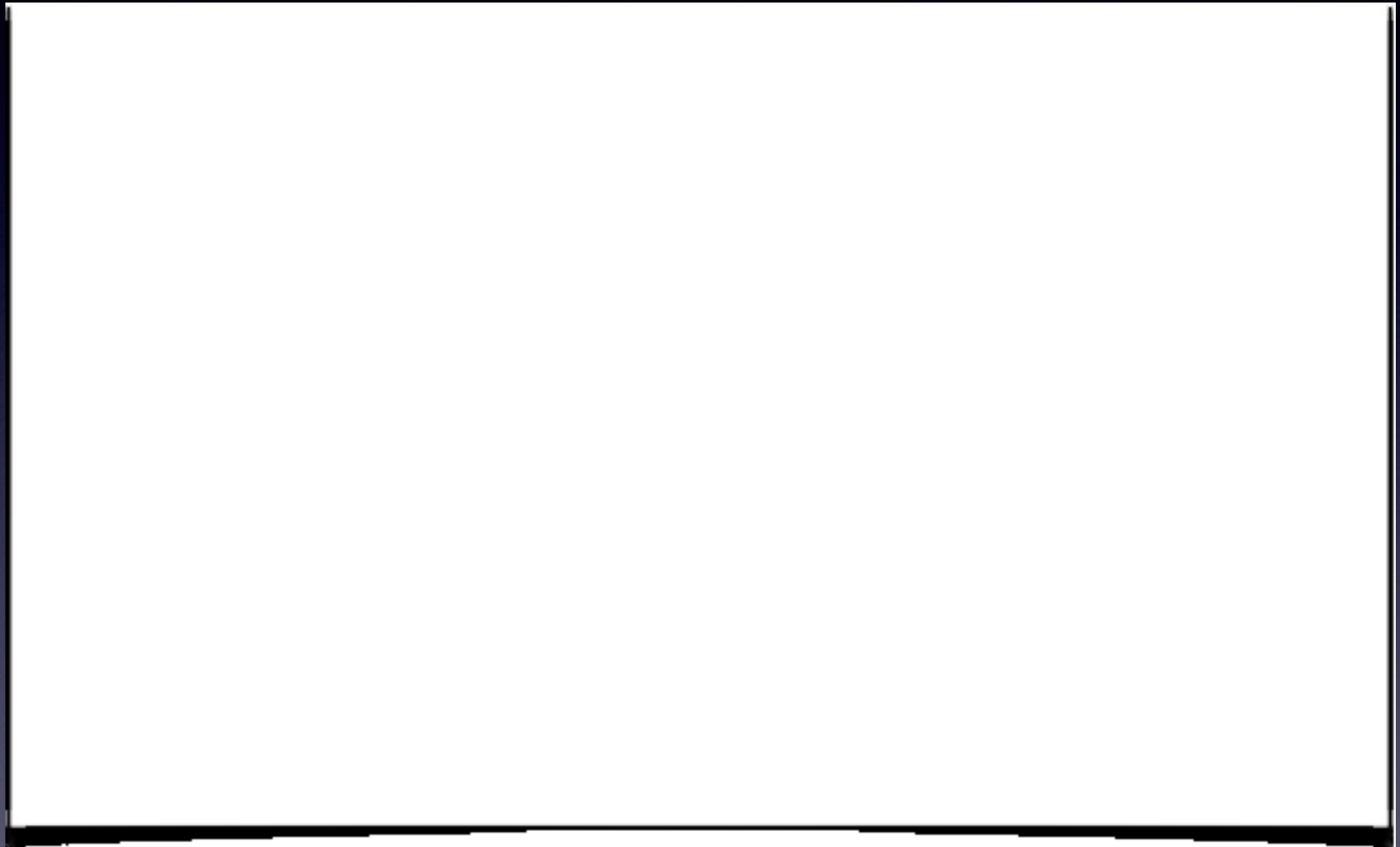
实例

临床问题 → 数据挖掘 → 协变量筛选

x与y是什么样的关系？
还有哪些因素与y有关系？

x与y是什么样的关系？
还有哪些因素与y有关系？

什么因素影响x与y的关系？
加强或减弱x对y的作用？



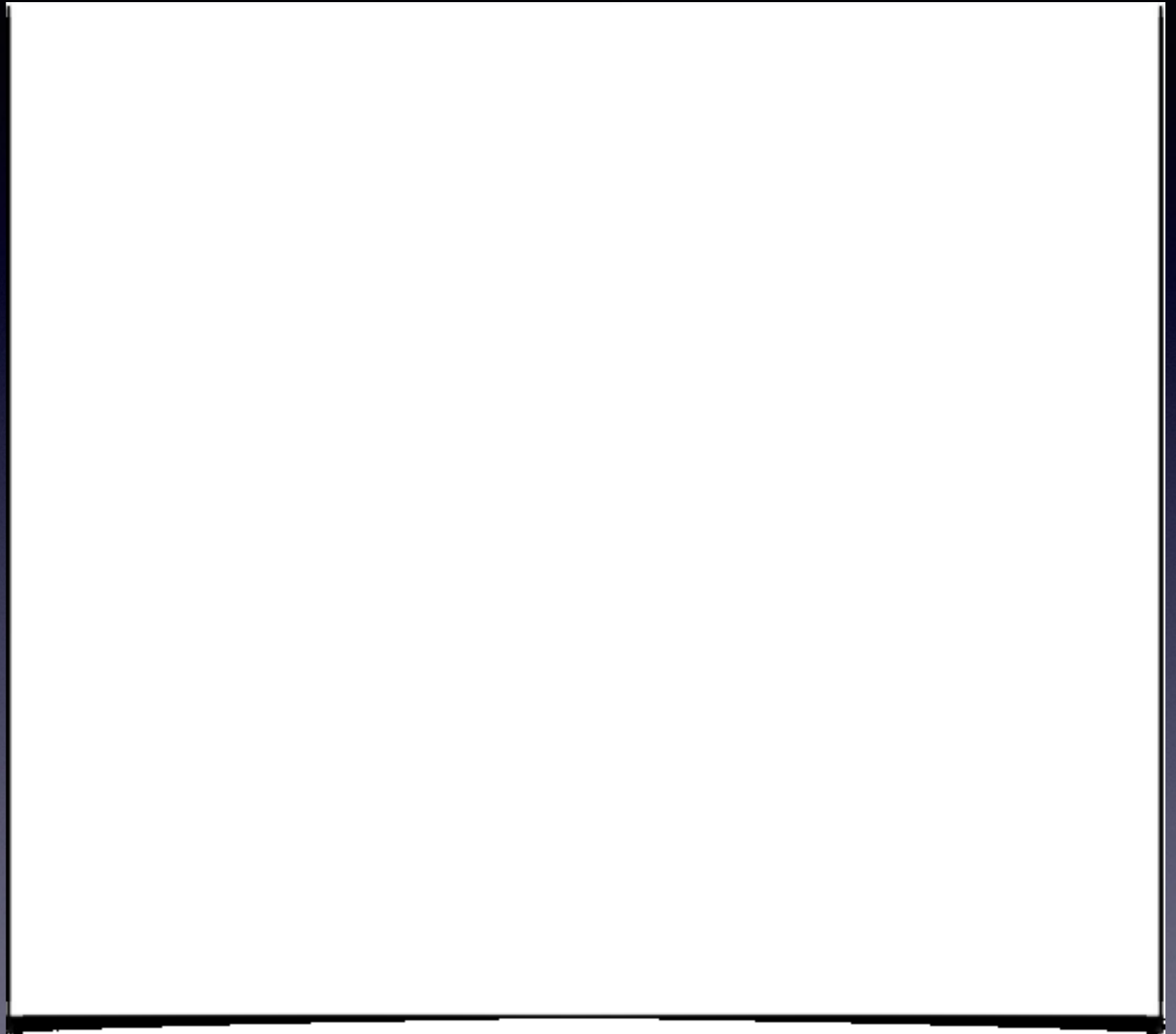
实例

临床问题 → 数据挖掘 → 发表论文

x与y是什么样的关系？
还有哪些因素与y有关系？

x与y是什么样的关系？
还有哪些因素与y有关系？

什么因素影响x与y的关系？
加强或减弱x对y的作用？



数据挖掘

整理原始数据

诊断数据关联关系

明确科研假设

分析思路

x与y是什么样的关系？
还有哪些因素与y有关系？

什么因素影响x与y的关系？
加强或减弱x对y的作用？

什么因素影响x与y的关系？
加强或减弱x对y的作用？

发表论文

科研流程

研究假设

现有数据

课题设计

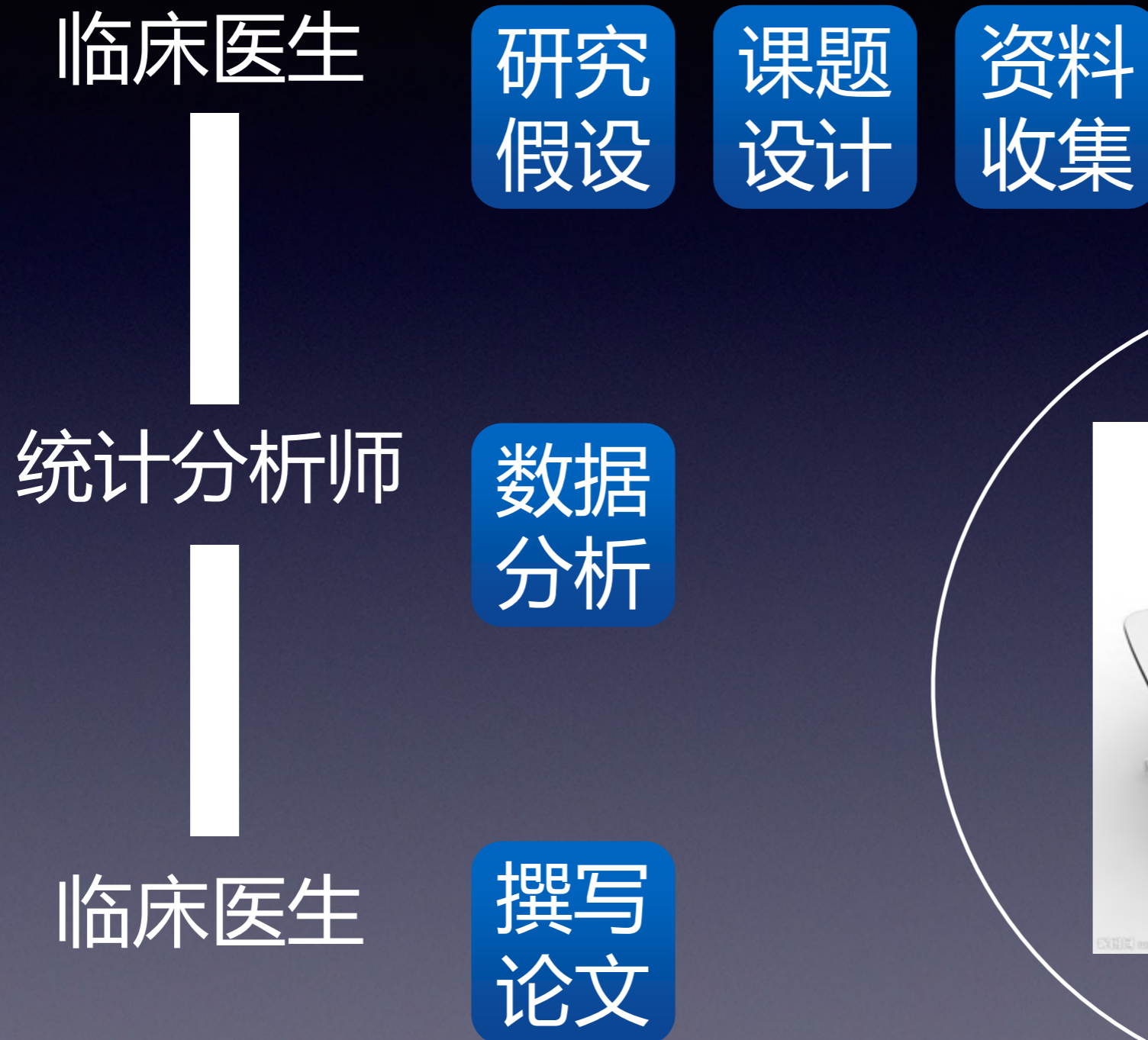
数据分析

研究假设

资料收集

资料提取

科研流程 国内外差异



流水线式



国内

国外

临床医生

流行病学家

统计分析师

研究
假设

课题
设计

资料
收集

数据
分析

编程
调试

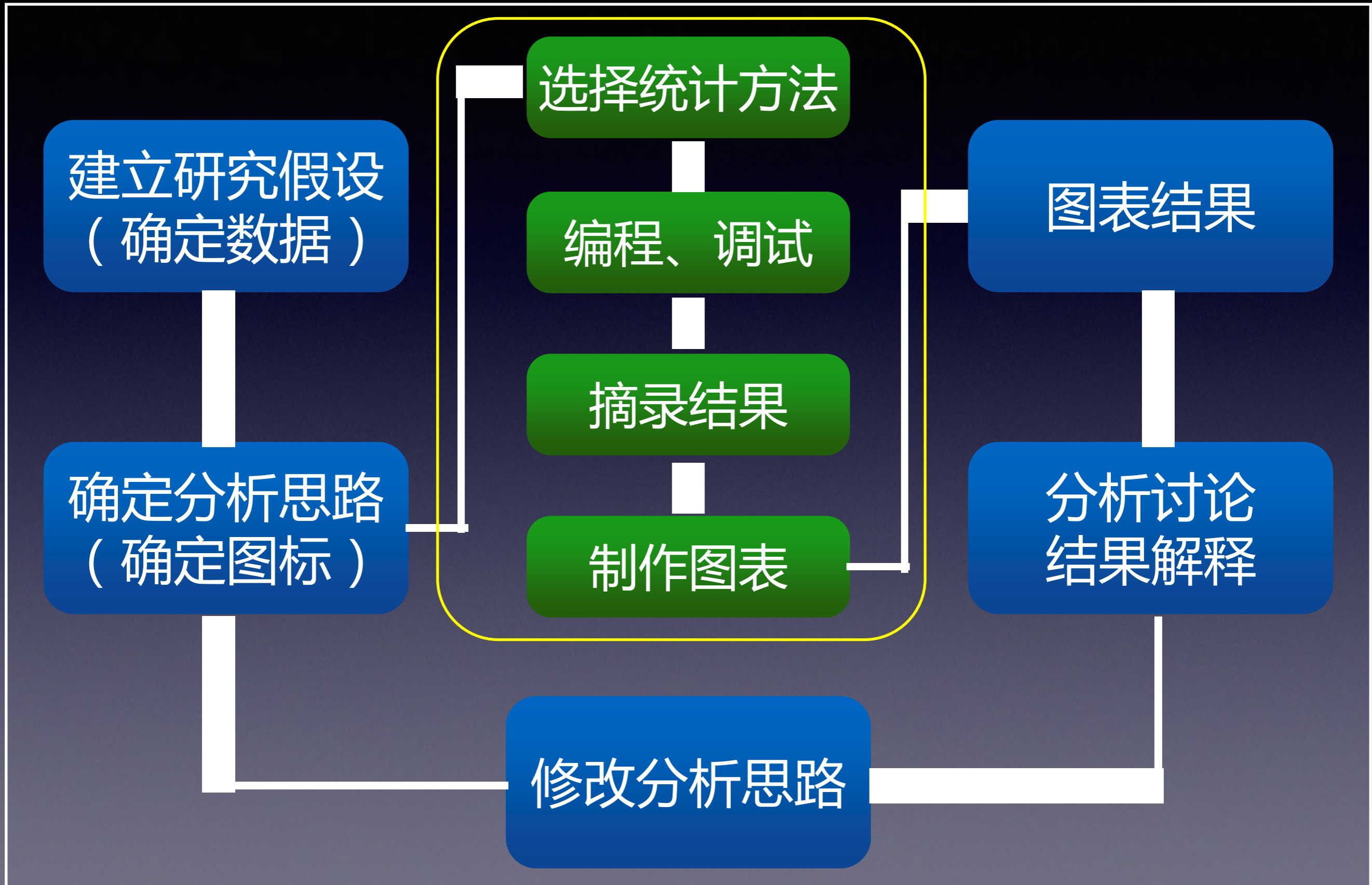
撰写
论文

计算机编程师

其他辅助人员

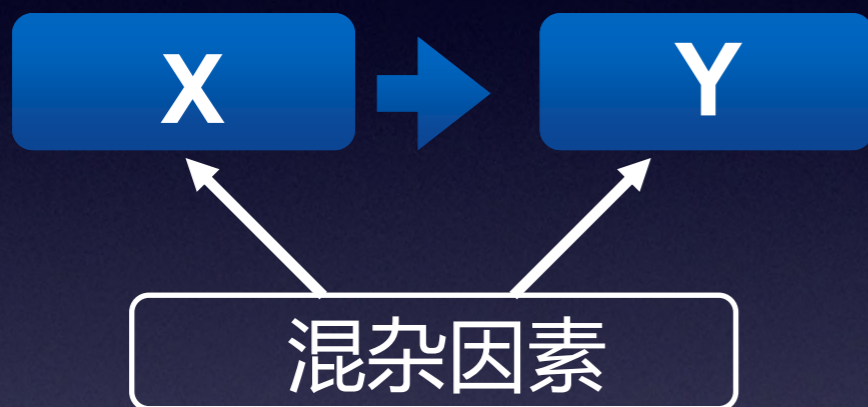
团队协作式

团队协作优势 - 循环往复出精品



1、明确的假设

可基于数据挖掘



2、分析思路

x与y是什么样的关系？
还有哪些因素与y有关系？

什么因素影响x与y的关系？
加强或减弱x对y的作用？

x对y有没有独立作用？
独立作用大小是多少？

3、通过合适的分析工具实现

Thank you for your attention !