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

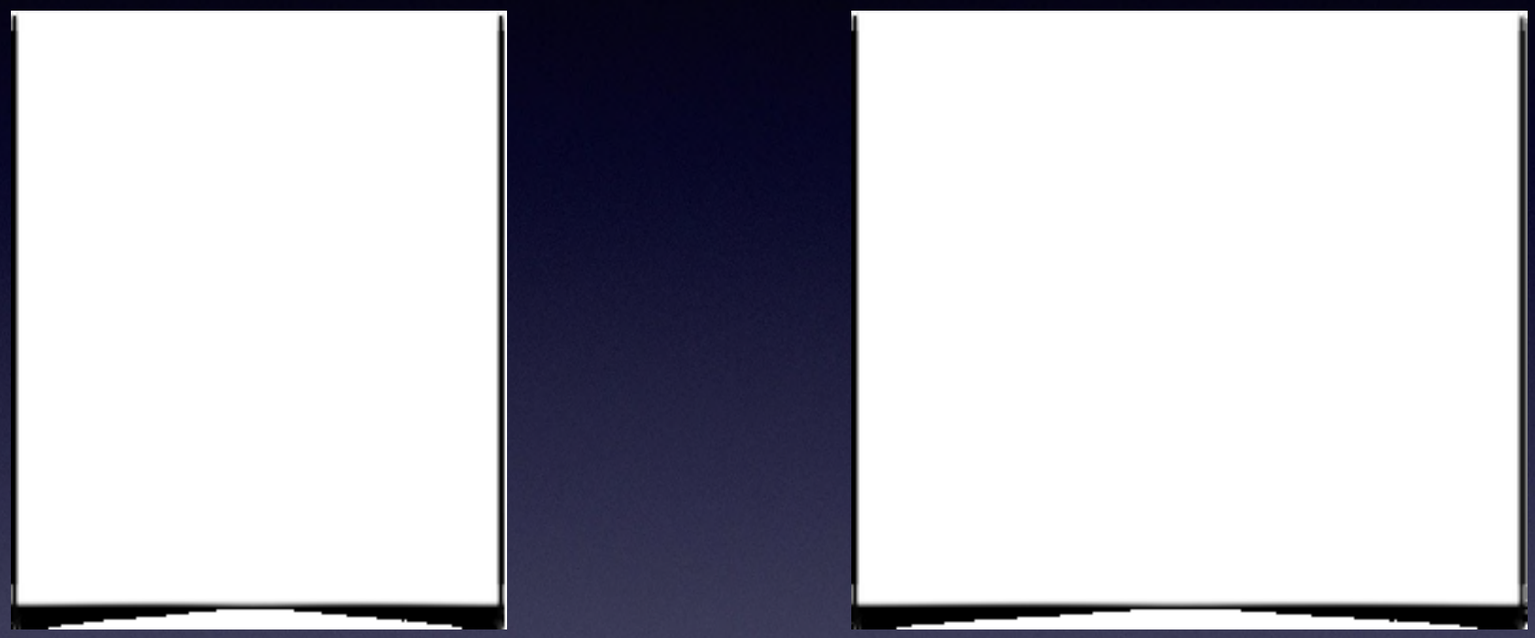
结果变量（Y）
术后急性肾功能衰竭

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	GENDER	AGE	ABO	GRADE	ETIOLOGY	ACUTERF	ANHEPATIC	WARMTIME	COLDTIME	OPTIME	CRYOPRECIPITATE	WHOLEBLOOD	REDBLOOD	FRESHPLASMA	PLATELET
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

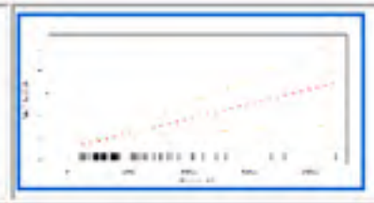
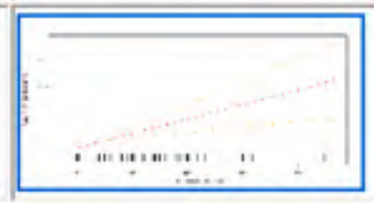
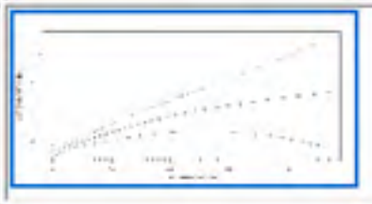
扫描数据

实例

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	OR 1.00	95%low 1.00	95%upp 1.00	p 0.0080		
			edf	Ref.df	Chi.sq	p-value		
		s(BLOODLOSS)	4.04	4.97	33.54	0		
** 0.0023 (0.0039)	CRYOPRECIPITATE	CRYOPRECIPITATE	OR 1.00	95%low 1.00	95%upp 1.00	p 0.0015		
			edf	Ref.df	Chi.sq	p-value		
		s(CRYOPRECIPITATE)	1.49	1.83	10.68	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		OR	95%low	95%upp	p		
		factor(PLATELET)1	2.00	1.00	4.00	0.0993		
		factor(PLATELET)2	1.00	0.00	2.00	0.3079		
		factor(PLATELET)3	0.00	0.00	Inf	0.9893		
		factor(PLATELET)4	1.00	0.00	12.00	0.7641		

实例

临床问题 → 数据挖掘

明确科研假设

X
术中输入冷沉淀物

Y
术后急性肾衰

混杂因素
术中失血量、感染、血型配型……

实例

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？

↓

什么因素影响x与y的关系？
加强或减弱x对y的作用？

↓

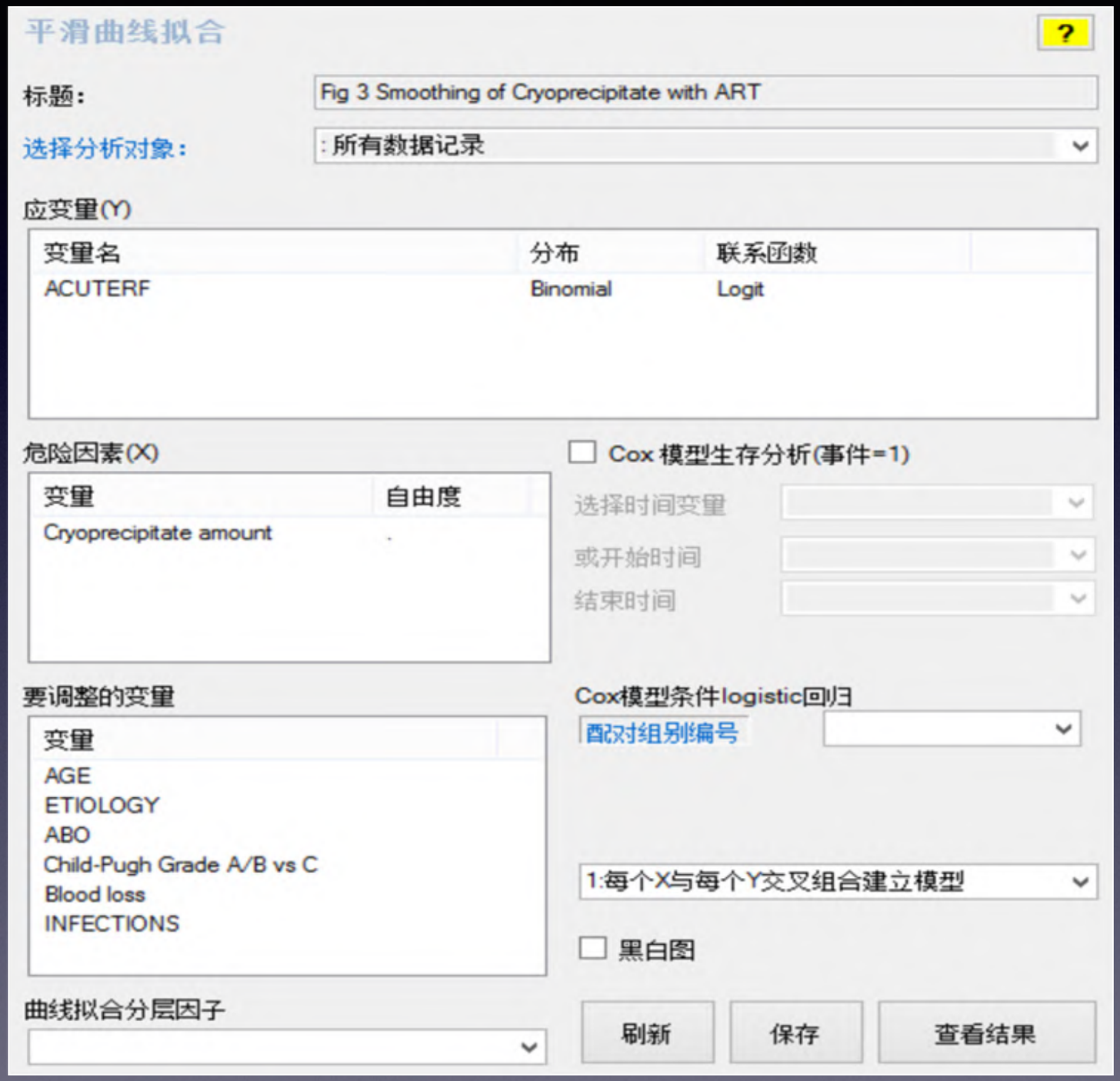
x对y有没有独立作用？
独立作用大小是多少？

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

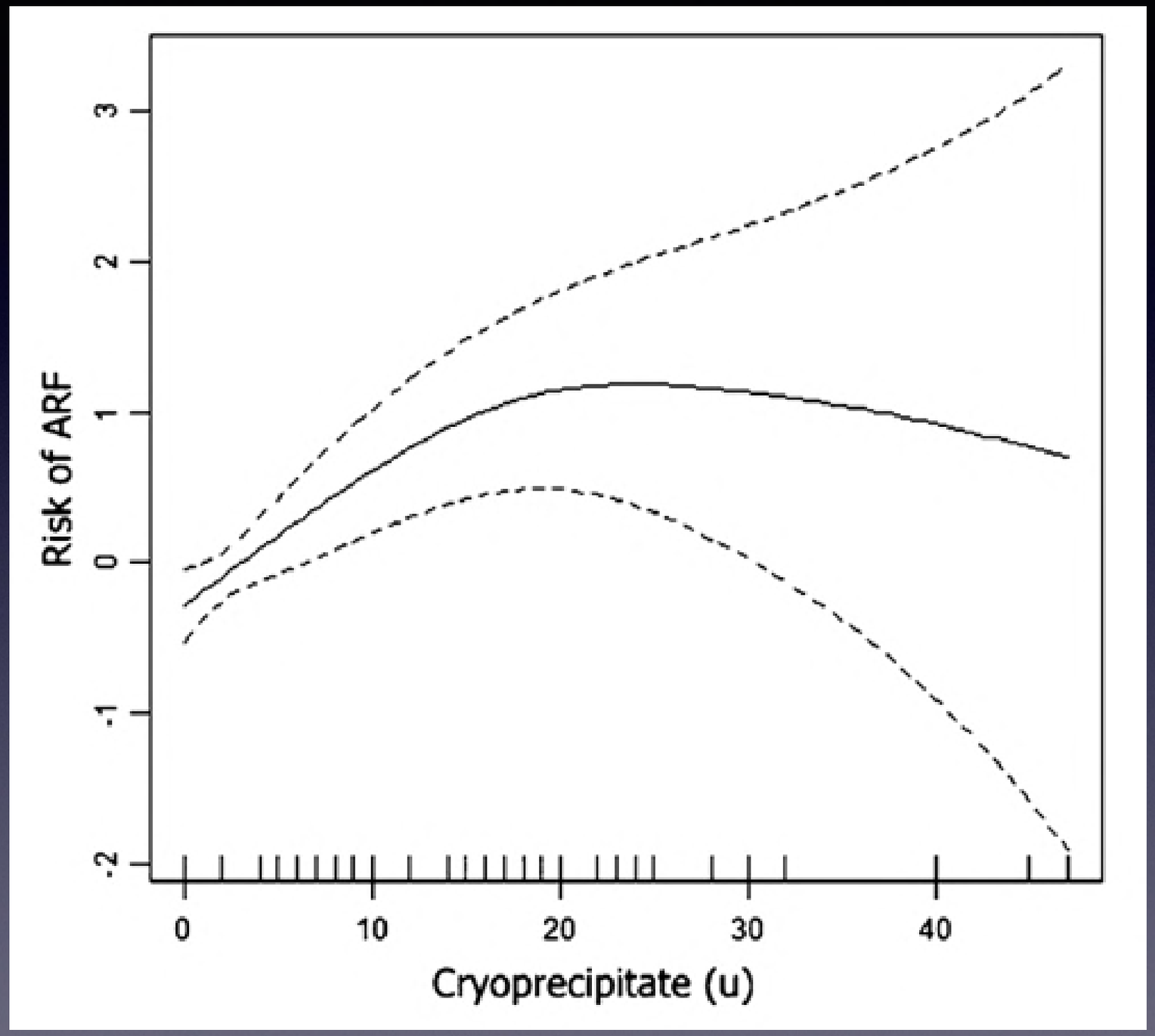


实例

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

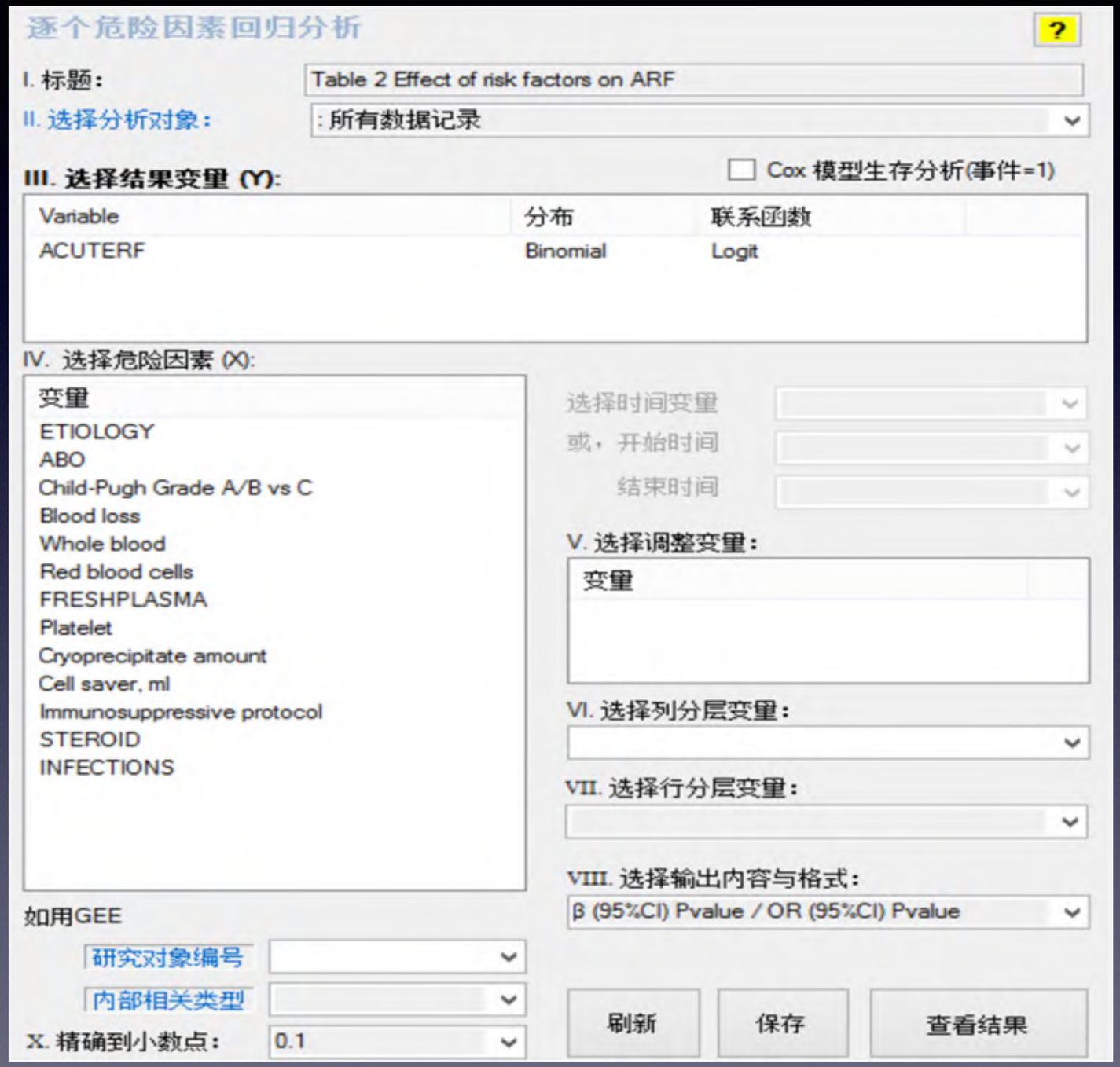
实例

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



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)	*p* value
Etiology, *n* (%)			
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, *n* (%)			
Compatible	334 (85.9)	0.37 (0.18, 0.77)	0.008*
Incompatibility	55 (14.1)	1	
Child-Pugh grade, *n* (%)			
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, *n* (%)			
Including cyclosporine A	107 (27.5)	1.0 (0.51, 2.0)	0.950
Including FK506	273 (70.2)	1	
Steroid, *n* (%)			
Used after OLT	337 (86.6)	3.5 (0.81, 14.8)	0.093
Not used after OLT	52 (13.4)	1	
Infections, *n* (%)			

实例

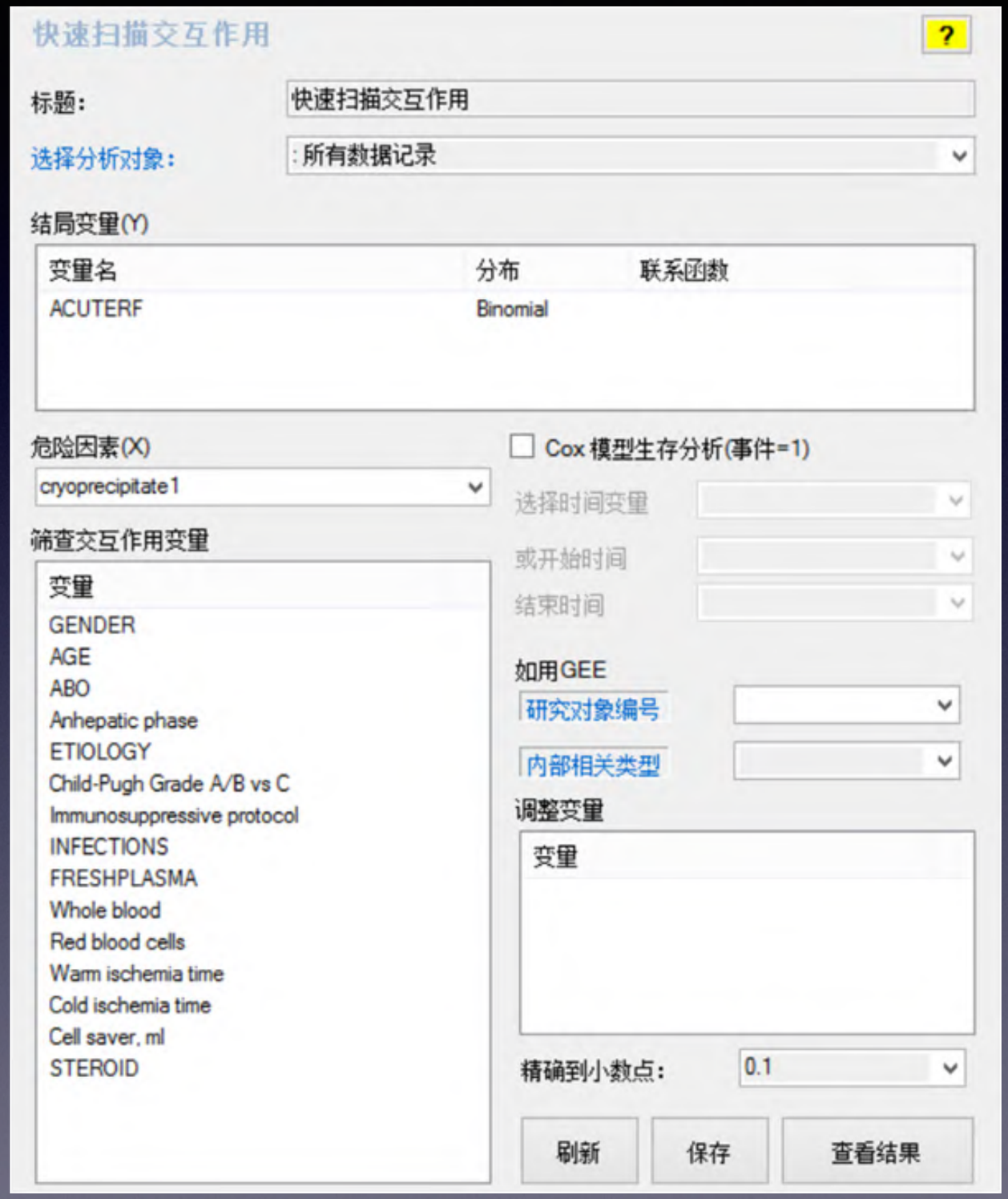
临床问题 → 数据挖掘 → 扫描交互作用

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



扫描交互作用

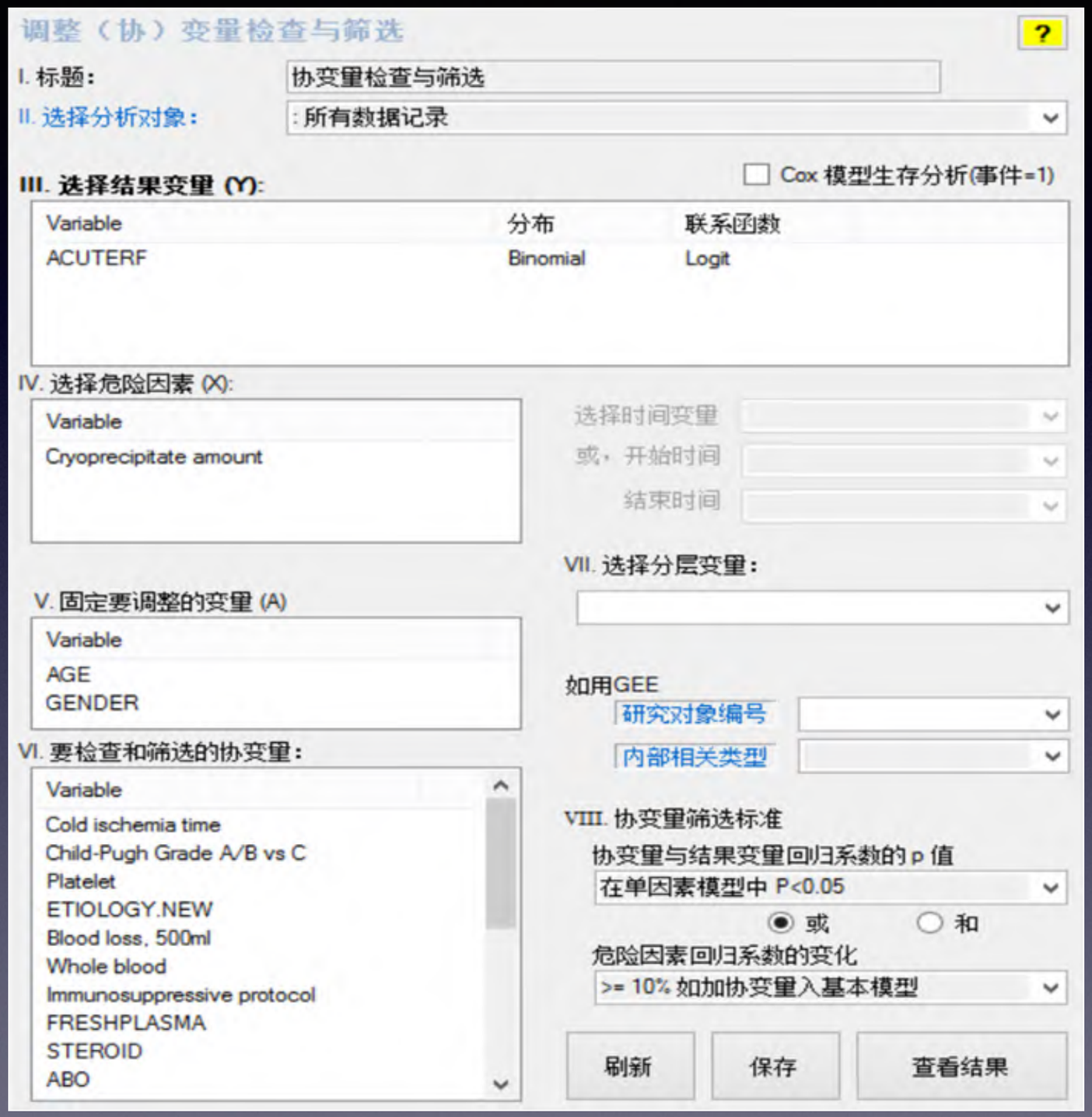
实例

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？



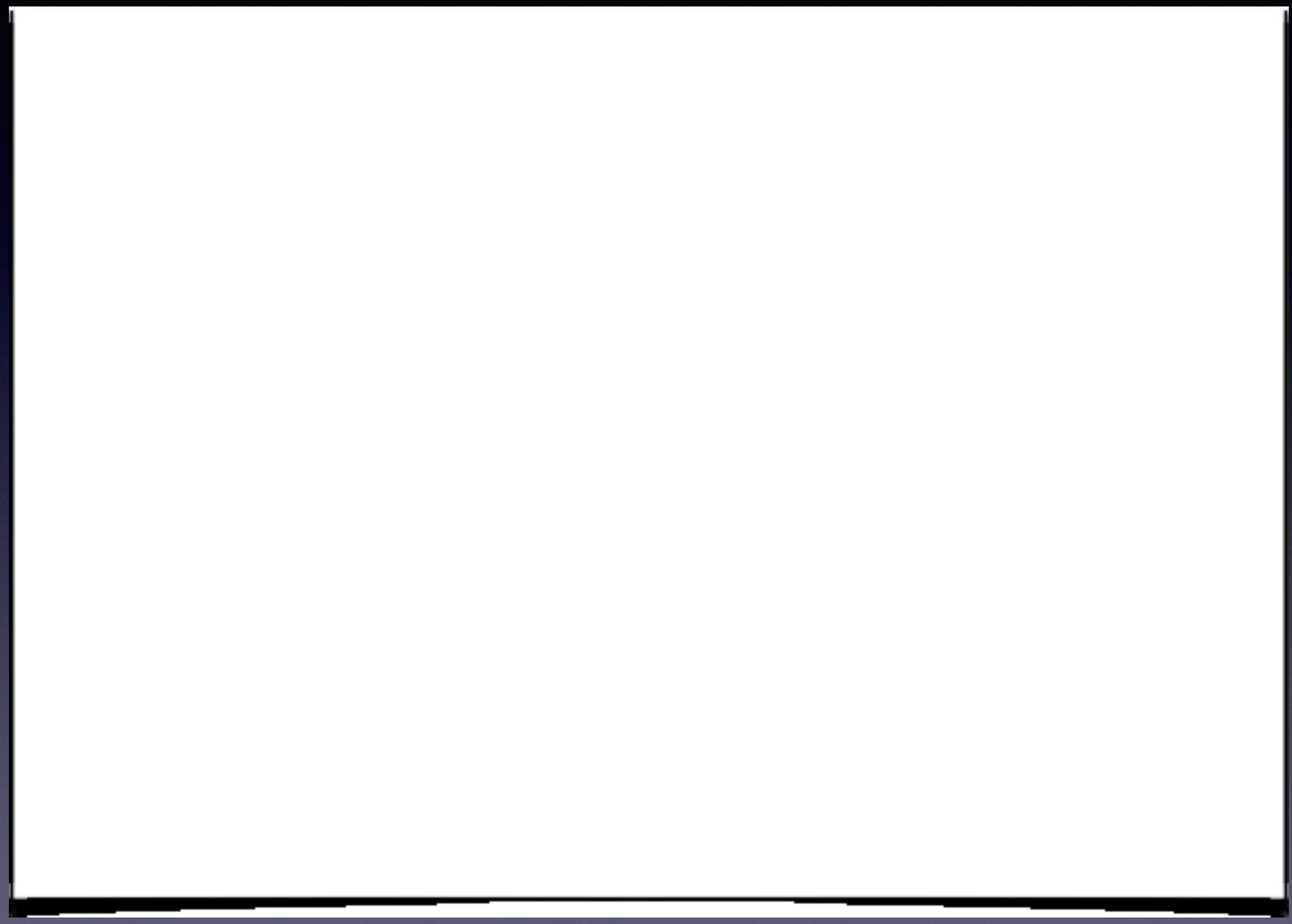
实例

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？

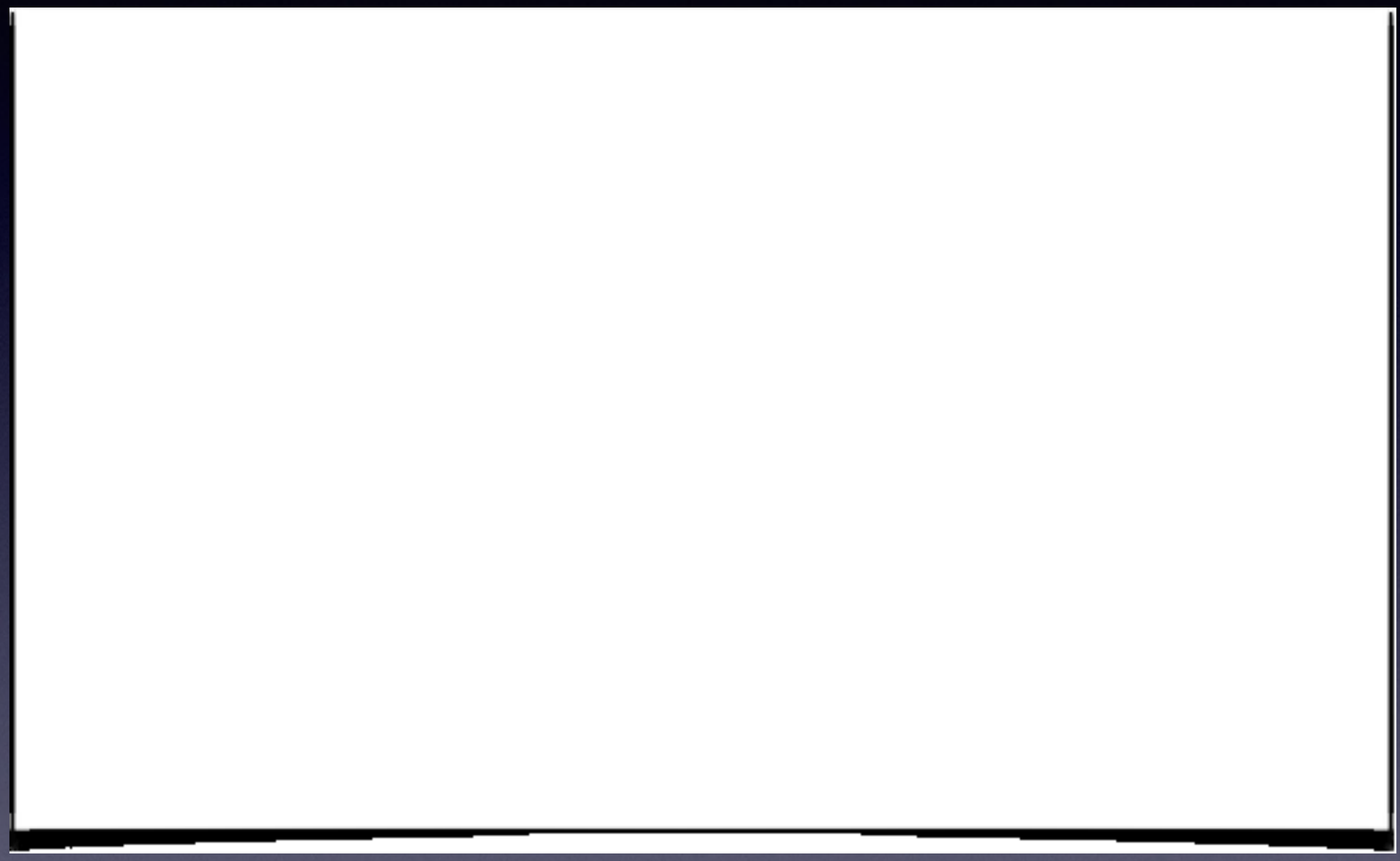
实例

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？



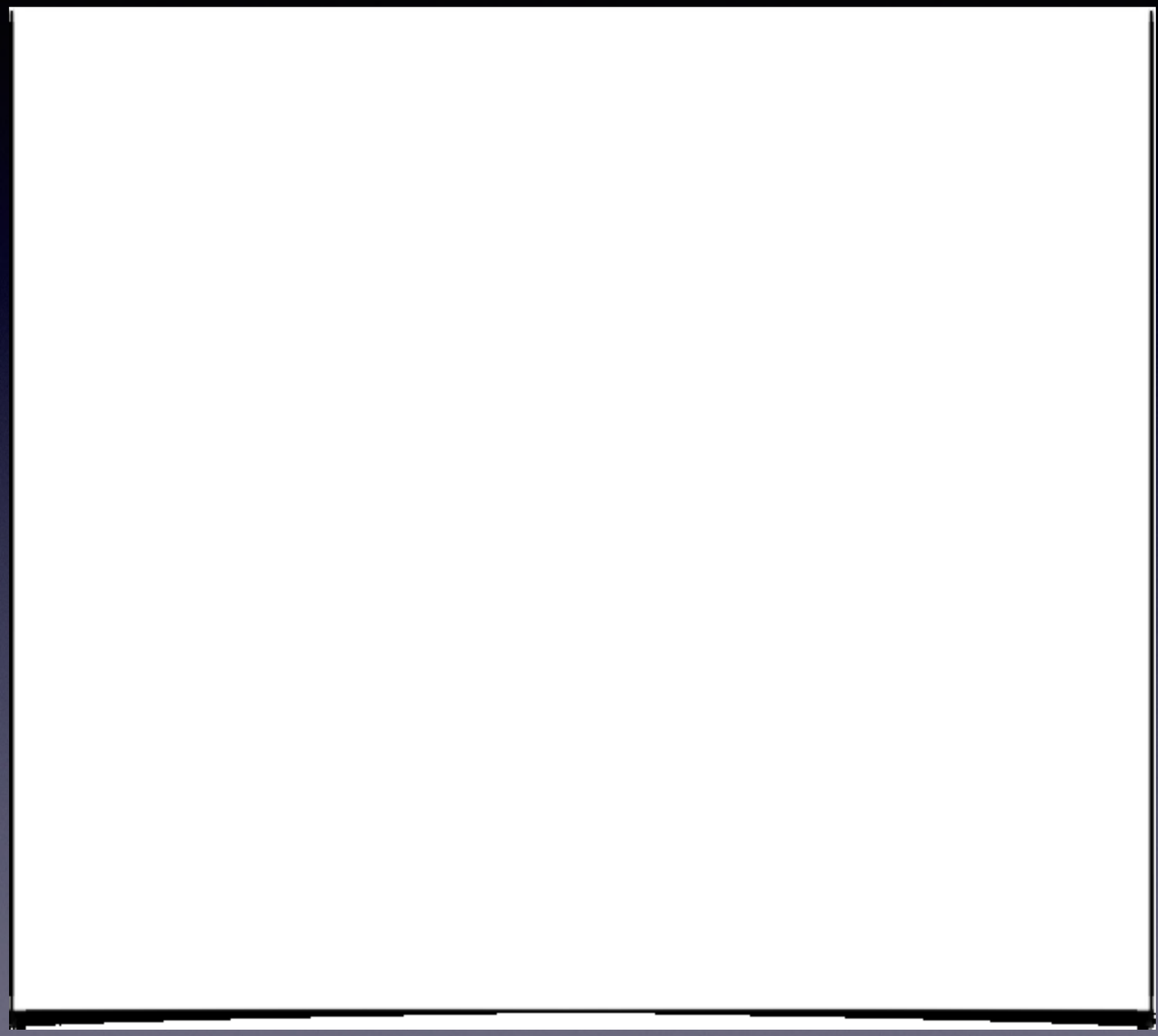
实例

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

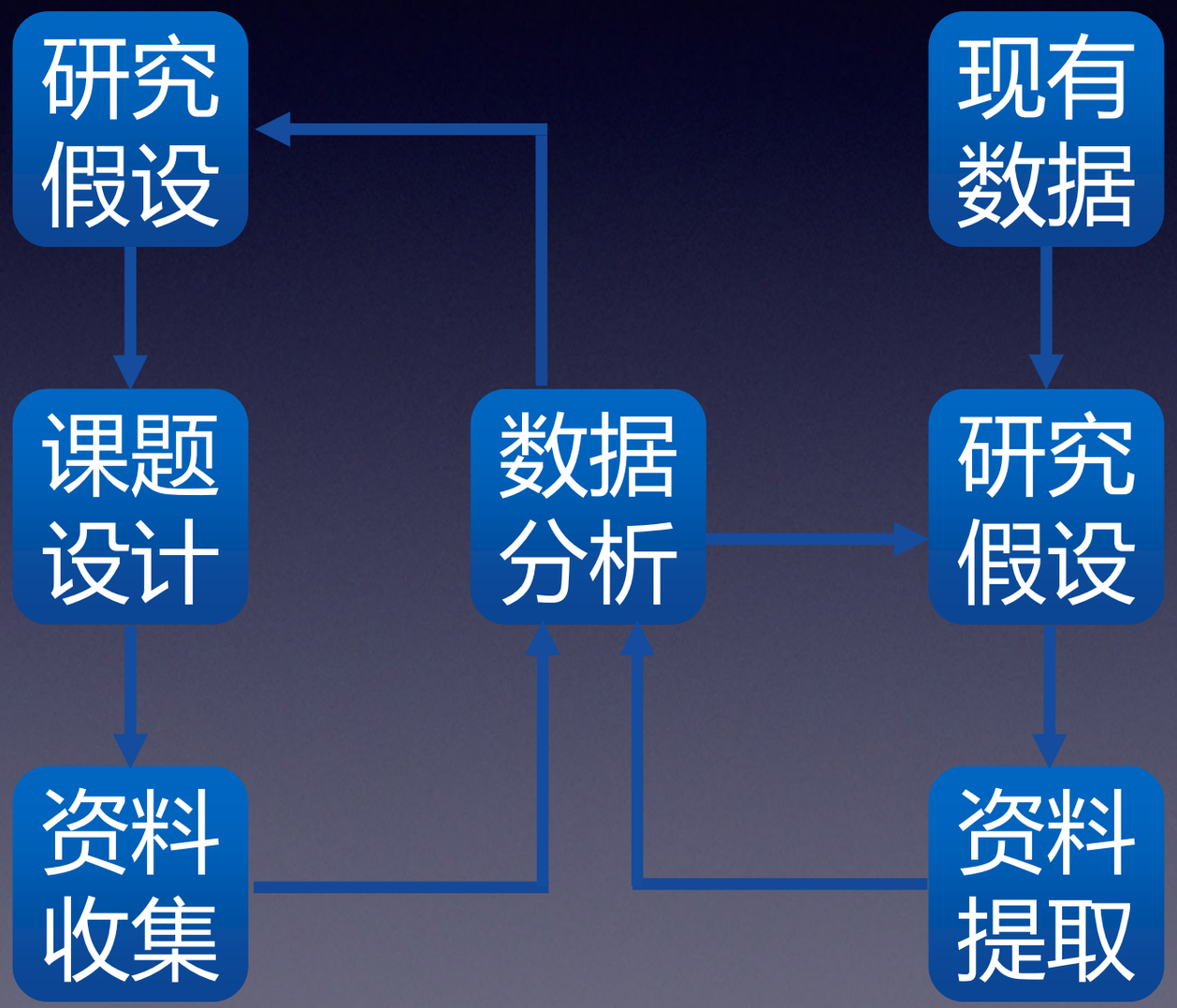
x与y是什么样的关系？
还有哪些因素与y有关系？

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



科研流程　国内外差异

临床医生

研究假设

课题设计

资料收集

统计分析师

数据分析

临床医生

撰写论文

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

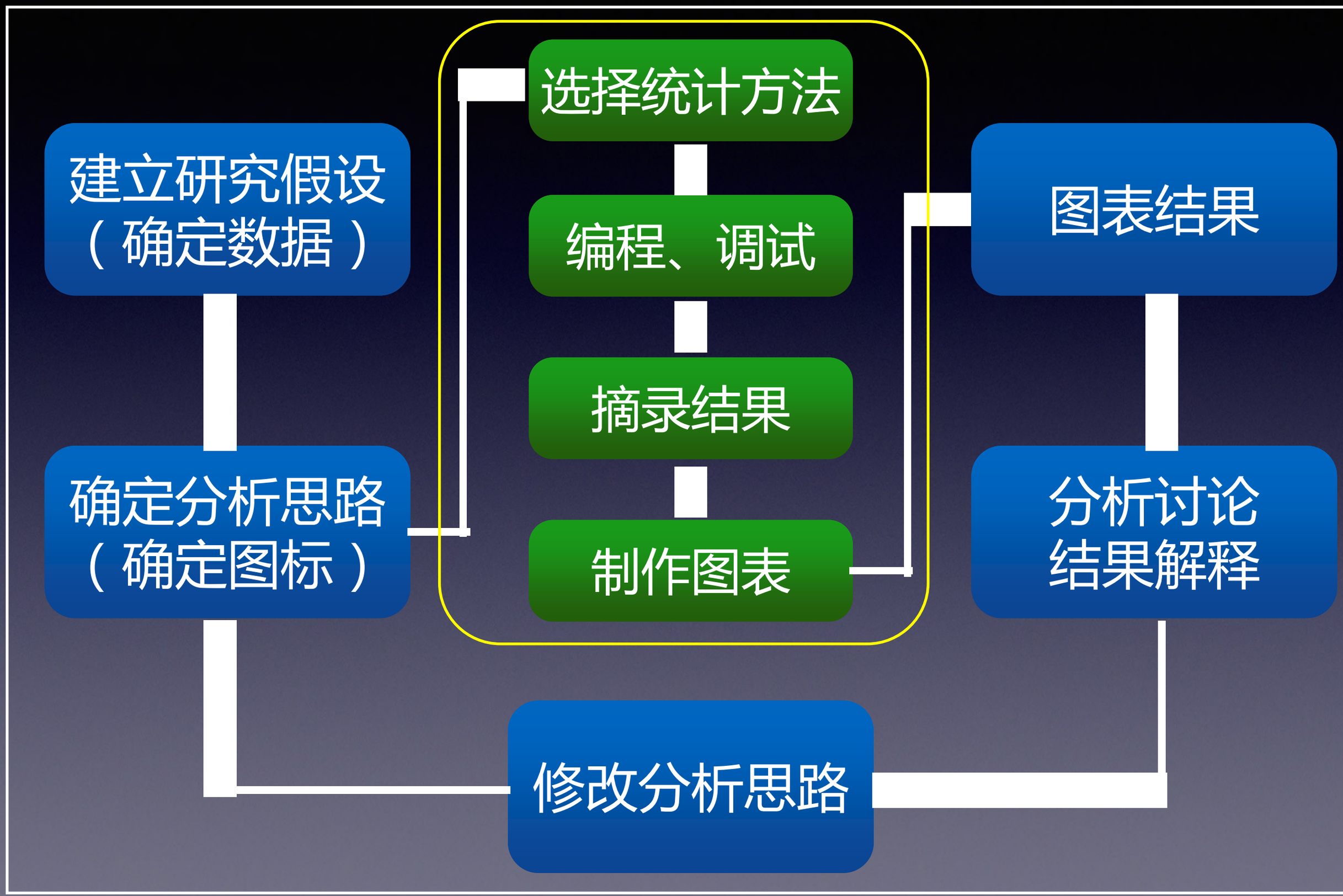
撰写论文

计算机编程师

其他辅助人员

团队协作式

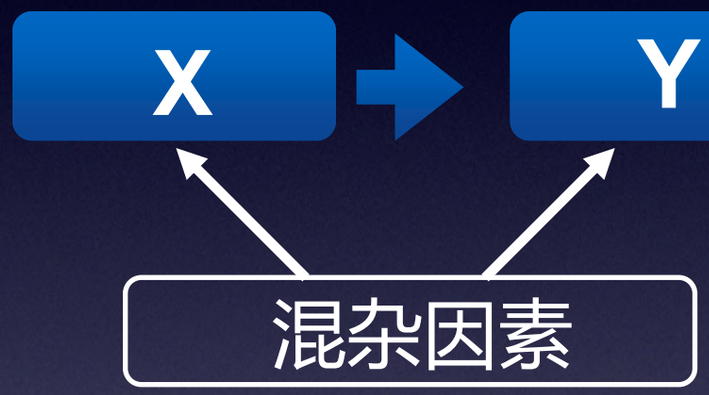
团队协作优势 - 循环往复出精品



总结

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 !