因果挖掘统计方法论：

井底之蛙、替罪羔羊、盲人摸象

耿　　直

北京大学数学科学学院

‘*I would rather discover a single causal relationship than be king of Persia.*’(Democritus, B.C.460-B.C.370?)

[译：我宁肯找到一个因果关系的说明，不愿获得一个波斯王位.(德谟克利特,古希腊哲学家)]

‘*Problems involving causal inference have dogged at the heels of statistics since earliest days.*’ (Holland, 1986)

[译:涉及因果推断的问题自始就缠住了统计学的脚后跟]

1 因果推断

自古以来，探讨事物之间的因果关系就是哲学、自然科学、社会科学、医学等几乎所有科学研究的最终目的。

亚里士多德(Aristotle, 公元前384-前322)
科学知识是关于原因的知识
因果知识为科学知识之本

休谟(Hume, 1740, *Atreatise on Human Nature*)
Hume问题: 原因产生结果不是经验归纳可证实的.
仅凭观察性研究得不出因果关系。

穆勒(Mill, 1843, *A System of Logic*)
Mill的四种方法:
契合法、差异法、共变法、剩余法.

因果与相关是两个不同的重要概念，

无因果关系也可能会表现出虚假相关性；
相反地，有因果关系也可能表现出虚假的独立性。

Freedman（1991）：小学生的阅读能力与鞋的尺寸有相关性; 人为地改变鞋的尺寸，不会提高他们的阅读能力。

曾经不少统计学者和医学研究者问我：
有因果关系的话，总应该表现有相关性吧？！

打太极拳可以强壮身体，延长寿命，
但是，打太极拳的人的寿命可能会与不打太极拳的人的寿命没有什么差异。因为打太极拳的人都是体弱多病的人，而表现出虚假的独立性。

铀矿工作的工人与其它人的寿命一样长（或更长），这不能说明铀矿不会影响寿命，而可能是因为铀矿工人是经挑选出来的身体健壮的人，假若当年他们不暴露于铀矿的话，寿命可能会更长一些。这种现象称为健康工人效应。

2 统计因果推断

Galton(1888, Correlation) 相关与回归

Person, K.(1911, *The Grammar of Science*, 3Ed.)

一旦读者认识了一个列联表的性质，他将掌握了原因与结果之间相关概念的本质

Wright(1921, Path analysis) 路经分析

Lewis(1973, *Counterfactuals*)

如果以前事件E 没出现的话, 那么,现在事件D 就不会出现了.

我们说: 事件E 是事件D 的原因.

Neyman (1923)和Rubin(1974, Counterfactuals)引入潜在结果变量：

$Y(1)$: 暴露情况下的结果，

$Y(0)$: 未暴露情况下的结果；

Y: 观测到的结果。

暴露的个体因果作用: $ICE = Y(1) - Y(0)$

'*You can't step into the same river twice.*' (Heraclitus, 东罗马皇帝)

[译:你不可能两次踏入相同的河]

暴露的平均因果作用: $ACE = E[Y(1) - Y(0)]$

Pearl(1995, Causal diagrams)

因果网络图、外部干预

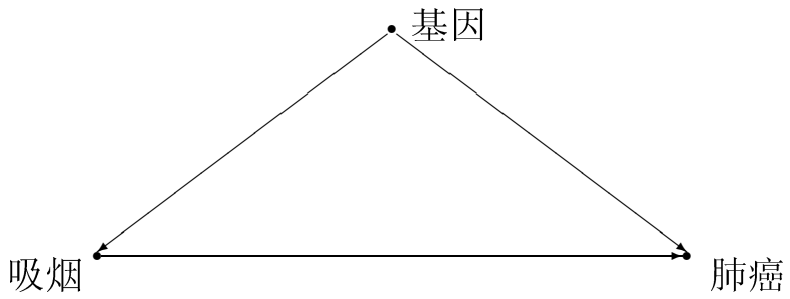
3　井底之蛙

Simpson 悖论

	有效	无效	行总和
新药	80	120	200
安慰剂	100	100	200
	$RD = \frac{80}{200} - \frac{100}{200} = -0.10$		

按性别分层

	男性		女性	
	有效	无效	有效	无效
新药	35	15	45	105
安慰剂	90	60	10	40
	$RD_1 = 0.10$		$RD_2 = 0.10$	



在随机化试验，或可忽略性假定: $[Y(0), Y(1)] \perp\!\!\!\perp T$下，

$$\begin{aligned} ACE &= E[Y(1)] - E[Y(0)] = E[Y(1)|T=1] - E[Y(0)=1|T=0] \\ &= E(Y=1|T=1) - E(Y=1|T=0) = RD. \end{aligned}$$

消除混杂偏倚：混杂变量U可观测.

假定$[Y(0), Y(1)] \perp\!\!\!\perp T|U$.

$$\begin{aligned} & ACE(T \to Y) \\ = \; & E[Y(1) - Y(0)] = E\{E[Y(1) - Y(0)|u]\} \\ = \; & E\{E[Y(1)|u] - E[Y(0)|u]\} \\ = \; & E\{E[Y(1)|T = 1, u] - E[Y(0)|T = 0, u]\} \\ = \; & E\{E[Y|T = 1, u] - E[Y|T = 0, u]\}. \end{aligned}$$

未知的混杂因素是否存在?

按性别和年龄分层

年龄	≤ 40岁				> 40岁			
性别	女		男		女		男	
	有效	无效	有效	无效	有效	无效	有效	无效
新药	5	5	40	50	30	10	5	55
安慰剂	60	55	5	5	30	5	5	35
	$RD_{11} = -0.02$		$RD_{12} = -0.06$		$RD_{21} = -0.11$		$RD_{22} = -0.04$	

Holland (1988, JASA),

Greenland, Robins & Pearl (Statist. Sci., 1999),

Geng, Guo & Fung (2002, JRSS B).

不可检验的假定$[Y(0), Y(1)] \perp\!\!\!\perp T|U$.

T	$Y(0)$	$Y(1)$	U
1	?	1	...
1	?	0	...
⋮			...
1	?	1	...
0	1	?	...
0	1	?	...
⋮			...
0	0	?	...

去掉不可检验的假定?

4 替罪羔羊: 替代指标(surrogate, biomarker, intermediate)

替罪羔羊

在医学研究中，真正的结局指标有时很难得到。

治疗→ CD4 → AIDS病人的生存时间，常用CD4作为替代指标。

如何确定替代指标?

例: (J Am Med Asso, 2002) 用雌激素和黄体酮进行绝经后的激素补充治疗(HRT)曾被认为能降低患心脏病的风险，其理由是

1. 激素治疗降低血清胆固醇，

2. 胆固醇低的人一般患心脏病的风险低。

可是，后来的用安慰剂对照的随机化研究表明，

HRT实际增加心脏突发事件。

例 HIV感染和AIDS的研究。

常常采用CD4作为治疗AIDS药物的替代指标。

治疗对CD4的作用不能预测治疗对临床结果（AIDS的发展或死亡时间）的作用。

例 绝经妇女的骨质疏松的研究。

氟化钠组增加骨密度。

处理组比安慰剂组有更高的骨折率。

结论，氟处理增加骨密度，但使得骨骼变脆，因此导致骨折脆弱。

例 关于心脏病学的替代指标的问题。

关于替代指标不可信的经典例子:

用“减少心室异常”作为降低心血管死亡率的替代指标。

心律失常抑制理论：

轻微心律失常会导致心脏骤停。抑制心律失常能减少死亡率。

关于心律失常抑制试验研究(CAST)评价三种药:

Enkaid(别名：encainide),

Tambocor (化学名：flecainide acetate),

Ethmozine(别名：moricizine)).

它们都可以有效地抑制心律失常，得到FDA批准。

美国每年有20多万人服用这些药。

有超过5万人死于抗心律失常药。

这个数字与越南战争以及朝鲜战争中死亡的人数相当。

是美国经历的最大一次药害事件。

4.1 目前使用的几种替代指标准则

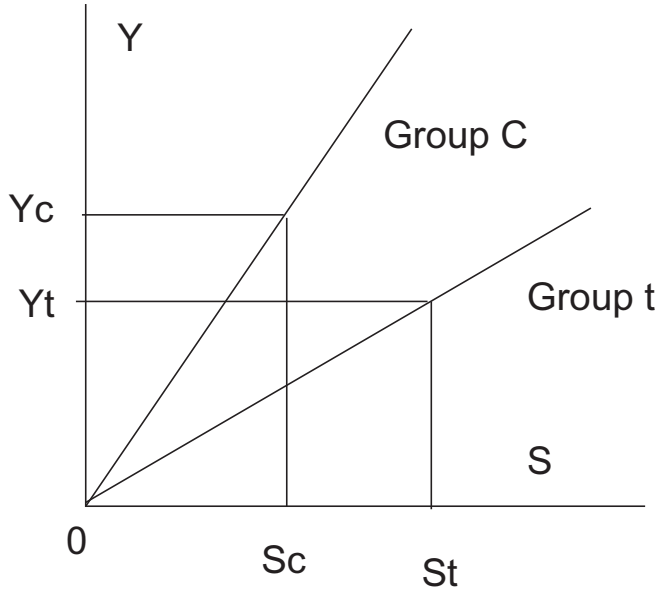
a. 强相关准则：替代指标X必须与结果Y有强相关性。

但是，强相关准则不能保证处理T对替代指标S有正相关作用，则处理T对结果Y一定有正相关作用。

反例见下图(Baker and Kramer, 2003)。

设AIDS病人的生存时间Y，CD4计数S.

CD4计数S与生存时间Y完全线性相关。



替代指标S与真结果Y完全相关

b. 条件独立准则(Prentice, 1989)：

给定替代指标条件下，结果与处理独立($Y \perp\!\!\!\perp T|S$)，

那么，S是一个统计替代指标。

由($S \perp\!\!\!\perp T$) 能推出($Y \perp\!\!\!\perp T$).

c. 主分层准则：(Frangakis and Rubin, 2002)

因果必要性： 处理T对替代S无因果作用，则对结果Y无作用。

即，平均因果作用(the average causal effects - ACE) 满足:

$$ACE(T \to S) = 0 \Longrightarrow ACE(T \to Y) = 0.$$

统计替代指标不满足因果必要性。

主代理准则： 如果对所有的s, 比较集合

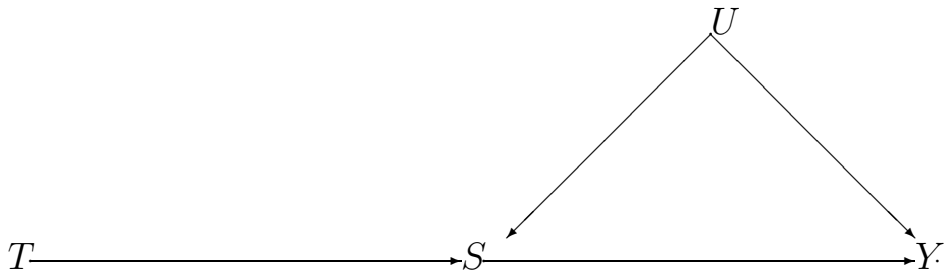
$$\{Y_i(1) : S_i(1) = S_i(0) = s\} \ \& \ \{Y_i(0) : S_i(1) = S_i(0) = s\},$$

产生相等的结果，那么S 是一个主代理(a principal surrogate)，用于比较处理$T = 1$和$T = 0$对结果Y的作用，

即，处理T对替代S无因果作用，则对结果Y无作用。

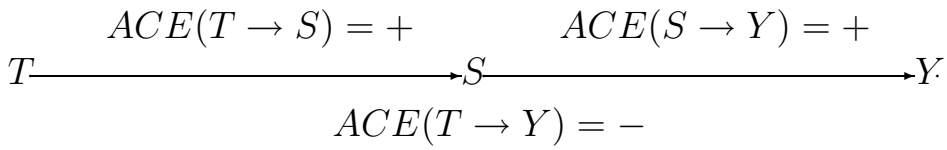
d. 强替代指标的准则(Lauritzen, 2004)：

强替代指标S是处理T至结果Y因果路径上的中间变量。



强替代指标: 保证处理T对替代指标S无因果作用的话，

处理T对结果Y就一定无因果作用。

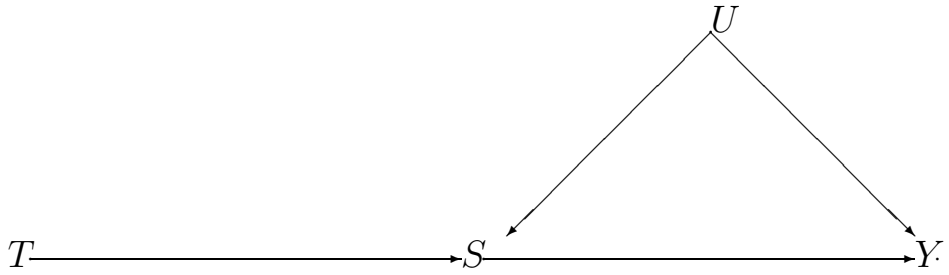


我们(JRSS B, 2007)称这种现象为**替代指标悖论(a surrogate paradox)**.

4.2 替代指标悖论的例子

例. 实际的例子(Fleming and DeMets, 1996; Moore, 1995).

在一个治疗心律失常的药物研究中, 心律失常会引起早期死亡, 但是抑制心律失常不仅不能延长寿命，反而增加了死亡率.



$T = 1$ 处理, $T = 0$ 对照；

$S = 1$ 抑制心律失常;

$Y = 0$ 猝死；

$U = 0$ 心脏损伤，或基因高表达.

假设$P(U=0)=0.3$, $P(T=1)=0.5$, 其他如Tab. 3所示.

Table 1: 假设的概率分布

	$p(S=1\|u,t)$		$p(Y=1\|u,s)$	
	$T=0$	$T=1$	$S=0$	$S=1$
$U=0$	0.98	0.79	0.00	0.98
$U=1$	0.02	0.99	0.98	0.99

处理可以3倍有效地抑制心律失常：:

$$P(S=1|T=1)/P(S=1|T=0)\approx 3.02;$$

但是，处理增加了死亡率3倍:

$$P(Y=0|T=1)/P(Y=0|T=0)\approx 2.91;$$

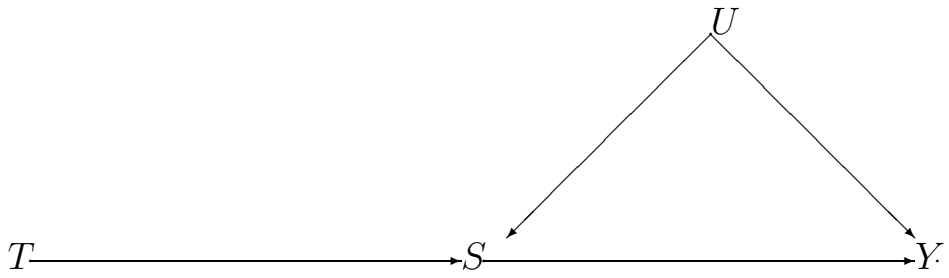
由概率分布，可以得到：

$ACE[T\rightarrow S|do(T=1),do(T=0)]=0.6220>0$，

$ACE[S\rightarrow Y|do(S=1),do(S=0)]=0.3010>0$，

但是，$ACE[T\rightarrow Y|do(T=1),do(T=0)]=-0.0491<0.$

关于该悖论名称的解释:



- 变量S是从T到Y因果路径中的中间变量，

  该悖论的一个更合适的名称也许是

  '**中间变量悖论(the intermediate variable paradox)**'

  可以表示更一般的含义.

- 可以将变量T看为一个工具变量.

  该悖论也称为'**工具变量悖论(the instrumental paradox)**'。

- 该悖论可以看为**Yule-Simpson 悖论**，

  可能存在一个未知的混杂因素U，它可以戏剧性地改变统计结论.

4.3 一致替代指标，严格一致替代指标(Chen, Geng & Jia, 2007, JRSS B)

为了避免替代指标悖论，

并使得可以推测$ACE(T \to Y)$的符号,

我们建议替代指标S应该有下面的性质：

一致性和严格一致性.

定义1. 一致替代指标(Consistent surrogate)

一个强替代指标S 是真正终点指标Y的一致替代指标需要满足条件：

1. 当定义S使得$ACE(S \to Y) > 0$时, 有

   $ACE(T \to S) \leq 0 \Longrightarrow ACE(T \to Y) \leq 0$ 和

   $ACE(T \to S) \geq 0 \Longrightarrow ACE(T \to Y) \geq 0,$

2. $ACE(T \to S) = 0 \Longrightarrow ACE(T \to Y) = 0.$

定义2. 严格一致替代指标(Strictly consistent surrogate)

一个强代理S 是真正终点指标Y的严格一致替代指标需要满足条件：

1. 当定义S使得$ACE(S \to Y) > 0$时, 有

   $ACE(T \to S) > 0 \Longrightarrow ACE(T \to Y) > 0$ 和

   $ACE(T \to S) < 0 \Longrightarrow ACE(T \to Y) < 0,$

2. $ACE(T \to S) = 0 \Longrightarrow ACE(T \to Y) = 0.$

定理1. S 是一致替代指标的条件是

1. Y的条件期望$E(Y|s,u)$是s的单调函数

   即$\partial E(Y|s,u)/\partial s \geq 0$ or ≤ 0 $\forall u$, 并且

2. 对于S，T是一个危险因素(即$F(s|t'',u) \geq F(s|t',u)$, $t' > t''$, $\forall s \& u$)

   或者T是一个保护因素(即$F(s|t'',u) \leq F(s|t',u)$, $t' > t''$, $\forall s \& u$).

注解:

- 因为没有观测到U, 定理1 中的条件是不可检验的.

  需要根据专业知识来判断条件的合理性.

- 条件1, 期望的单调性意味着替代指标S 是一个危险因素.

  例如，对于相同背景的病人，肺中的焦油量S越大，患肺癌的概率或期望就有大。即$p(Y=1|u,s') \geq p(Y=1|u,s'')$ $\forall s' > s''$.

  在线性模型$E(Y|s,u) = bs + g(u)$下, 条件1 自然成立.

- 条件2 意味着分布的单调性,

  它比个体单调性的假定更弱，(Imbens and Angrist, 1994).

  例如，研究食用脂肪对心脏病的作用，血脂作为替代指标。

  同一总体食用大量脂肪比食用少量脂肪有较大概率有高血脂。

  但不要求每一位个体食用脂肪

  一定比他本人食用少量脂肪更容易有高血脂。

5 网络图模型，结构学习

$T = \{t_1, t_2, \cdots, t_S\}$ — 观测数据模式

t_s — 组s中个体的观测变量集合.

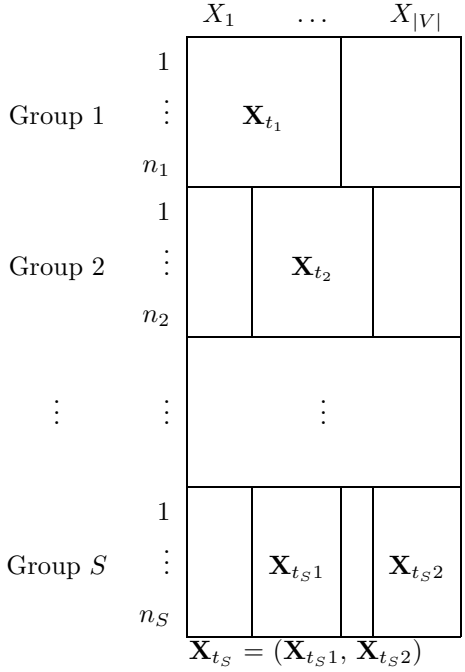


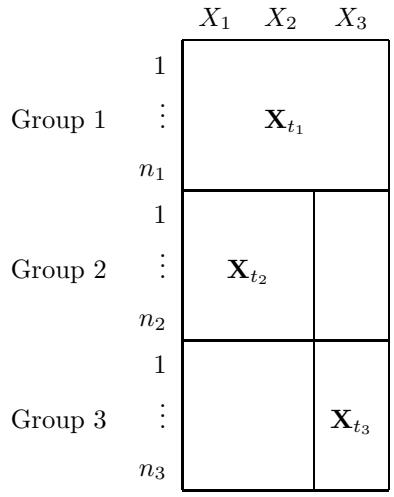
Fig. 3. An observed data pattern.

超图表示观测数据模式.

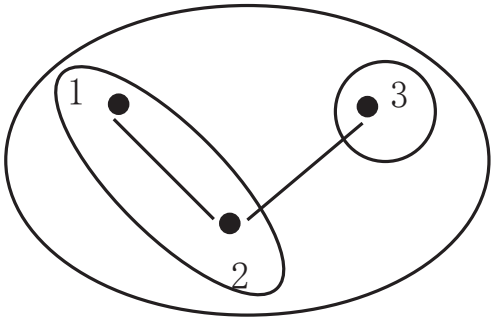
超边 — 一个节点的集合, 表示观测变量的集合。

Example 1. (Continued)

Observed data pattern is $T = \{\{1, 2, 3\}, \{1, 2\}, \{3\}\}$.

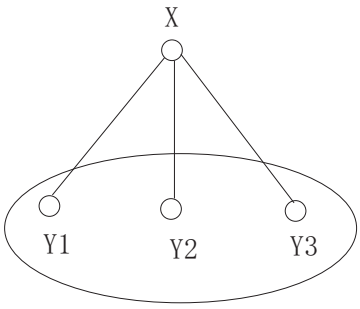


Observed data pattern $T = \{\{1, 2, 3\}, \{1, 2\}, \{3\}\}$.

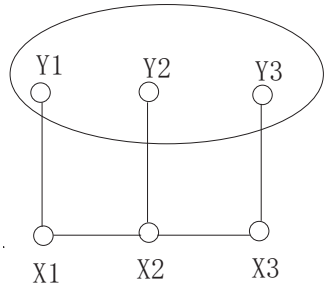


A graph with hyperedges.

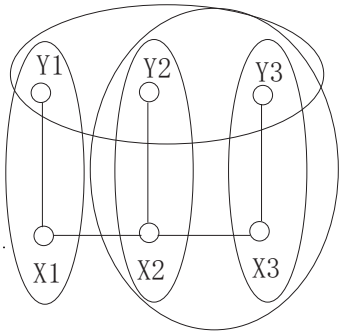
A Naive Bayesian Model:



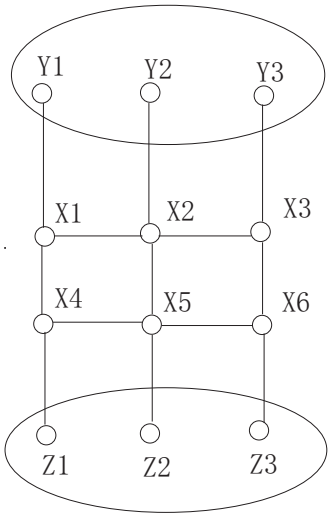
A Hidden Markov Model (HMM):



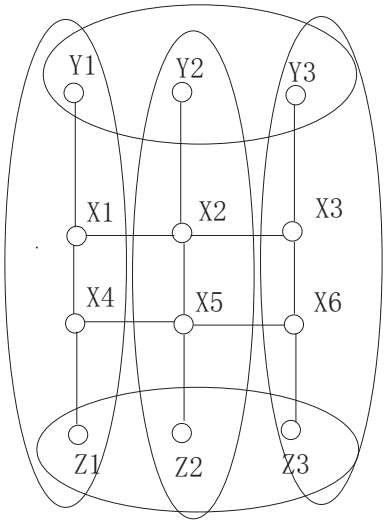
Training data patterns:



A Coupled Hidden Markov Model:

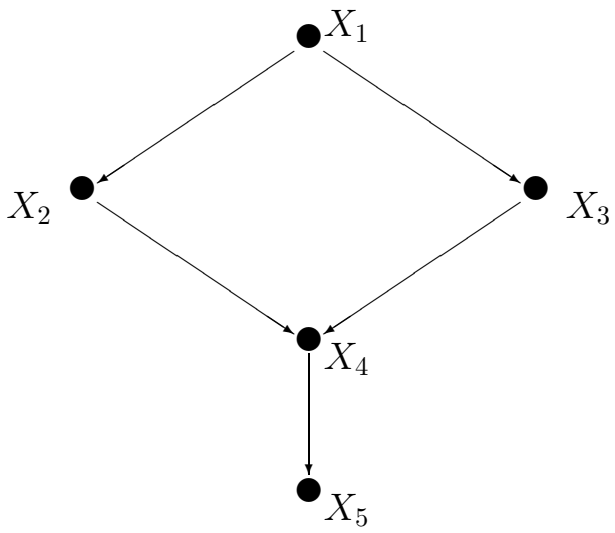


Training data patterns:



5.1 因果网络的学习

网络中边的方向表示因果方向。



Pa_i：随机变量X_i的父结点集合.

每个节点由父节点确定：

$$X_1 = f(\varepsilon_1), \quad X_2 = f(X_1, \varepsilon_2),$$
$$X_3 = f(X_1, \varepsilon_3), \quad X_4 = f(X_2, X_3, \varepsilon_4),$$
$$X_5 = f(X_4, \varepsilon_5), \quad X_6 = f(X_5, \varepsilon_6).$$

联合概率分布:

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2, x_3)P(x_5|x_4).$$

根据观测数据，构造Bayesian网络。

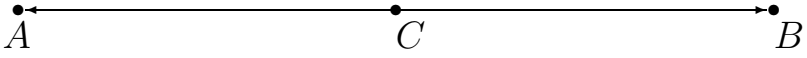
Table 2: 数据矩阵

	基因1	基因2	...	基因K
Microarray 1	X_{11}	X_{12}	...	X_{1K}
Microarray 2	X_{21}	X_{22}	...	X_{2K}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Microarray n	X_{n1}	X_{n2}	...	X_{nK}

例: 如果两个变量相关$A \not\perp\!\!\!\perp B$, 不能确定哪个是因，哪个是果。

变量A与变量B也许虚假相关，

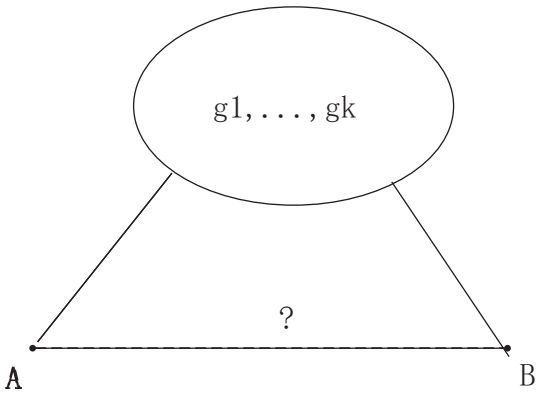
另一个变量C是A和B的公共原因：



因果网络学习：

1. 如何确定两点之间是否有边？

   对于一对相关变量A和B，是否存在一组变量$S = \{g_1, \ldots, g_k\}$，能够解释A与B的相关是虚假的？



   IC算法，IP算法：

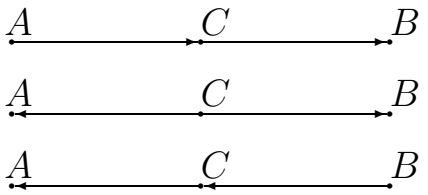
   对于任意两个变量A和B，

   寻找所有可能的变量集合S,

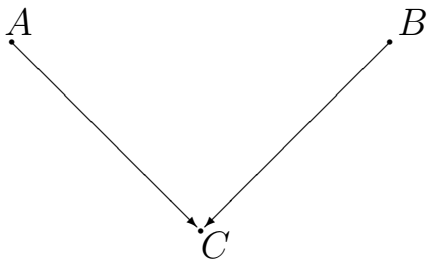
   判断是否$A \perp\!\!\!\perp B | S$。

2. 如何确定因果的方向?

例：如果A与B相关，给定C下A与B独立，则可能是下面情况：



如果A与B不相关，但是给定C下A与B相关的话，可以确定方向：



通过外部干预，进行因果方向的学习。

例如，干预某些变量，观察其他变量的变化。

1. 根据条件独立的假设检验，建立网络，

   Spirtes, Glymour & Scheines (1993), Pearl (2000)；

2. 采用得分方法，选择最佳得分网络；

3. Bayesian方法，Heckerman, D. (1997)

4. 选择邻居的方法；

5. 连续正态分布的Lasso方法，

   Meinshausen & Buhlmann (Ann. Stat., 2006)；

6. 基于Markov等价类的因果学习，He, Geng & Liang (ALT 2005)。

5.2 盲人摸象

数据库1

	X_1	X_2	X_3	X_4	X_5
患者1	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}
患者2	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}
$\vdots$	$\vdots$	缺失X_{ij} =?	$\vdots$	...	...
患者n	X_{n1}	X_{n2}	X_{n3}	X_{n4}	X_{n5}

数据库2

	X_1	X_6	X_7	X_8	X_9
患者$n+1$	$X_{n+1,1}$	$X_{n+1,6}$	$X_{n+1,7}$	$X_{n+1,8}$	$X_{n+1,9}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
患者$n+m$	$X_{n+m,1}$	$X_{n+m,6}$	$X_{n+m,7}$	$X_{n+m,8}$	$X_{n+m,9}$

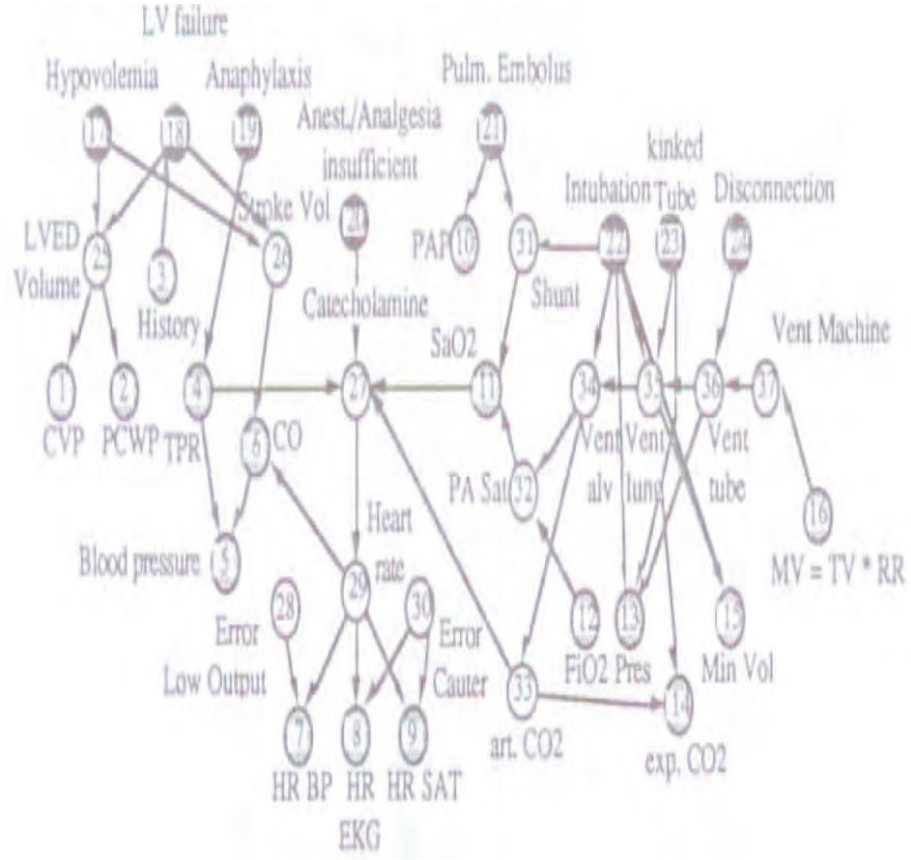
数据库J

	X_8	...	X_p
患者k	$X_{k,8}$	...	$X_{k,p}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
患者N	$X_{N,8}$	...	$X_{N,p}$

5.3 有向因果网络的分解学习算法(Xie, Geng & Zhao, AI, 2006)

医学诊断系统，病人监控，

有37各变量的ALARM网络：



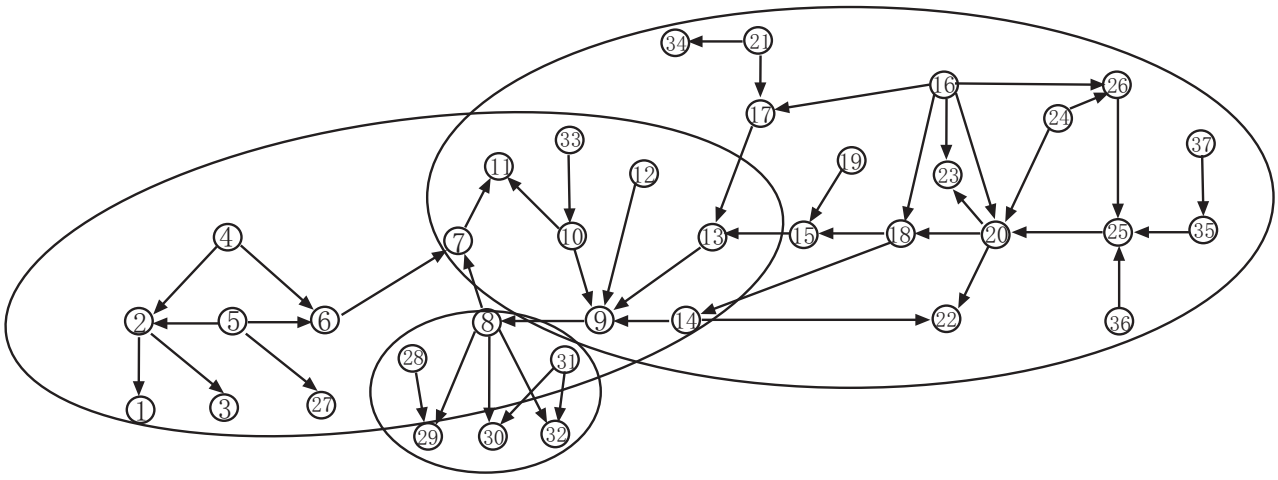
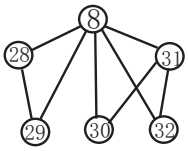
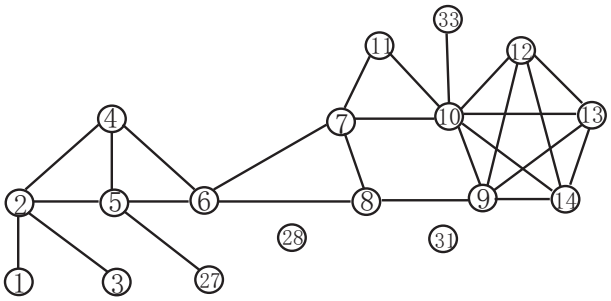


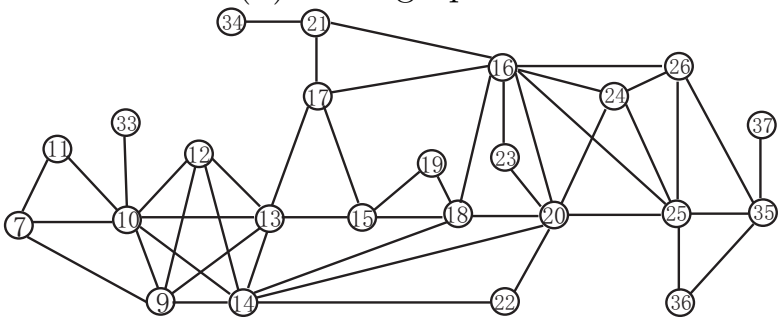
Fig. 10. Three databases.



(a) A subgraph for the left node of T



(b) A subgraph for the middle node of T



(c) A subgraph for the right node of T

Fig. 12. Undirected independence graphs for nodes of T.

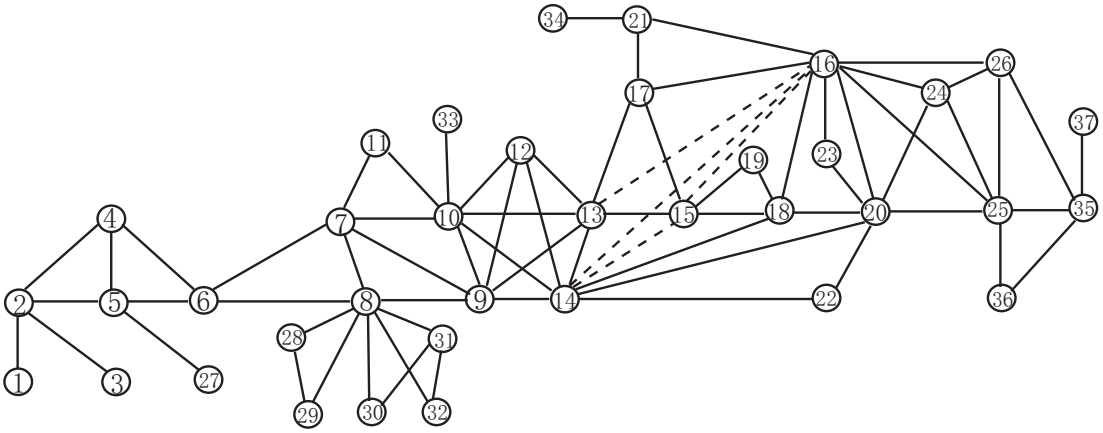


Fig. 13. The global triangulated graph.

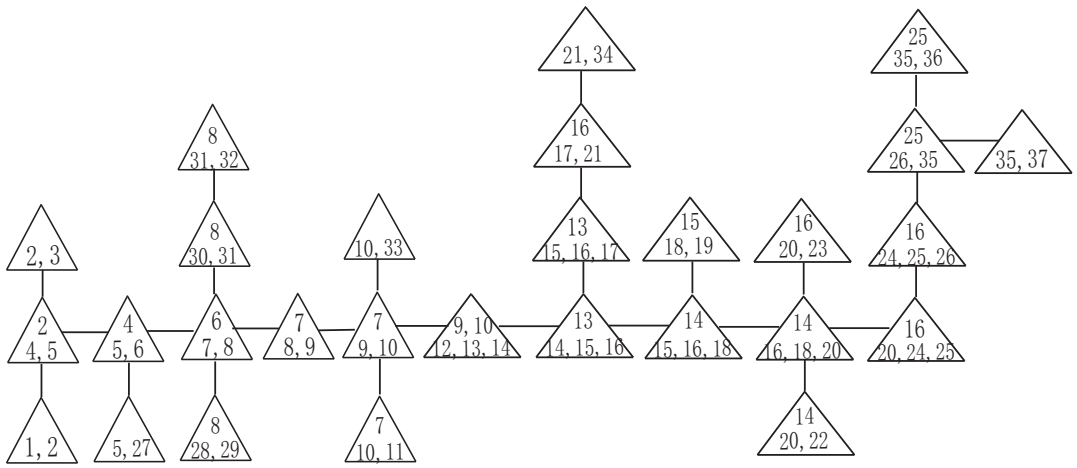


Fig. 14. The d-separation tree T'.

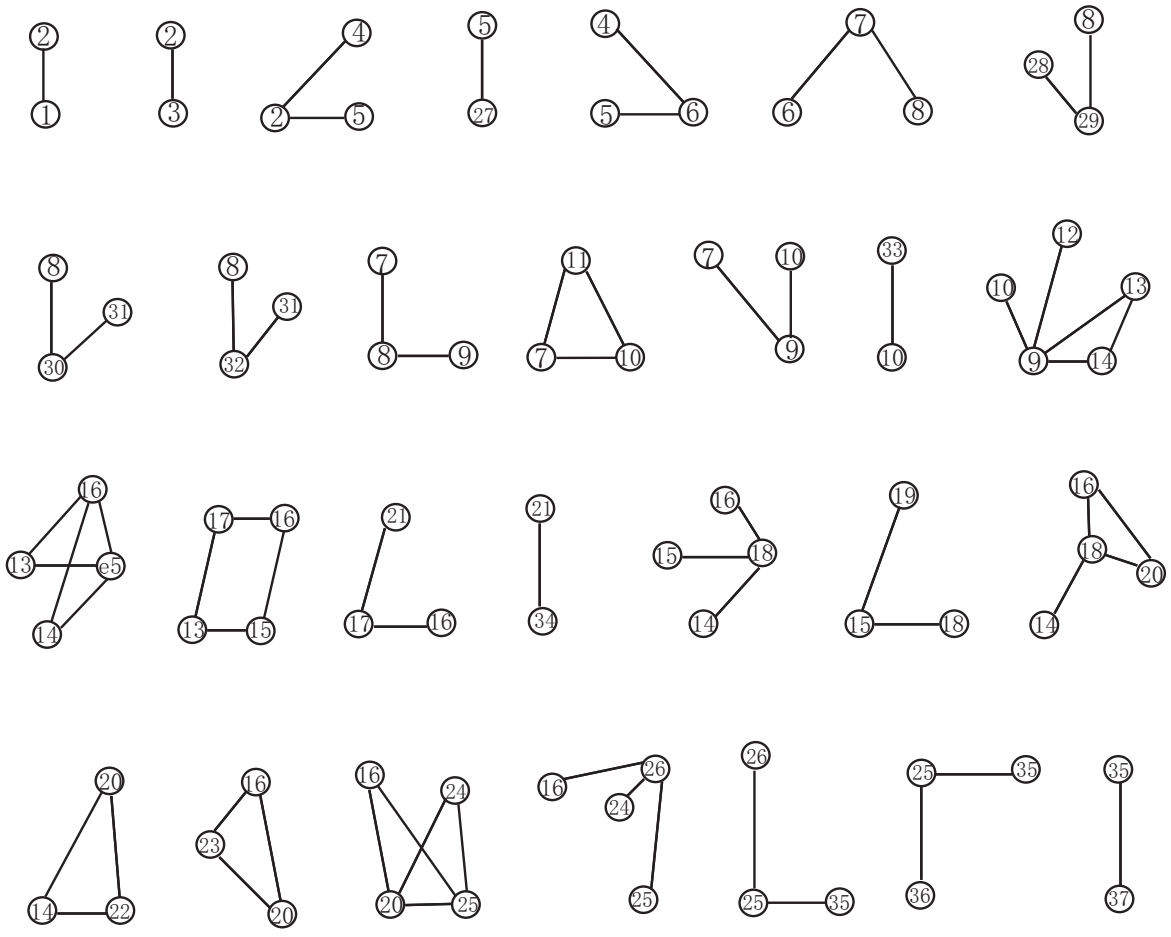


Fig. 15. Skeletons for all nodes of T'.

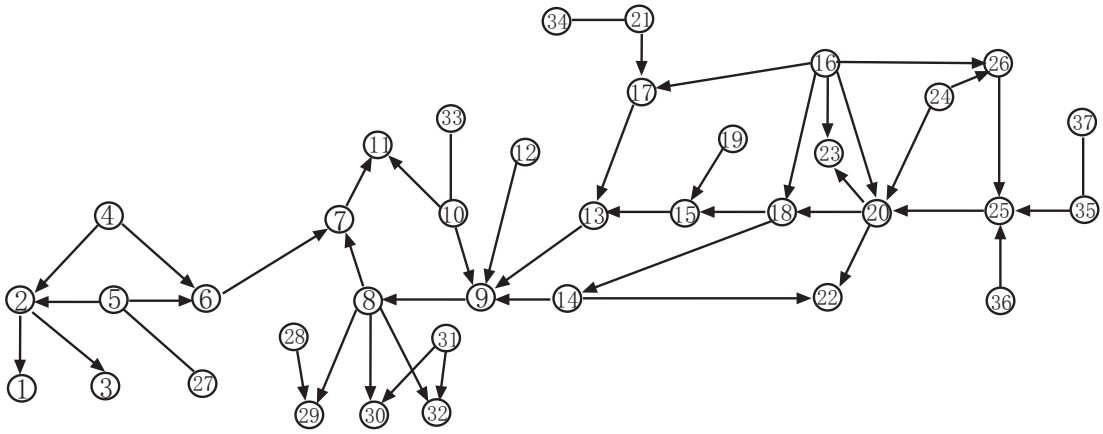
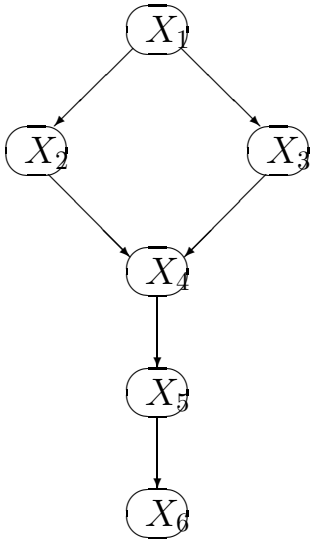


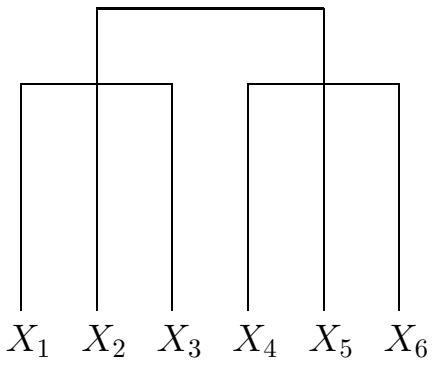
Fig. 16. The partially directed acyclic graph.

6 网络结构学习与聚类分析的结合(Wang, Geng, Wang, ISNN 2006)

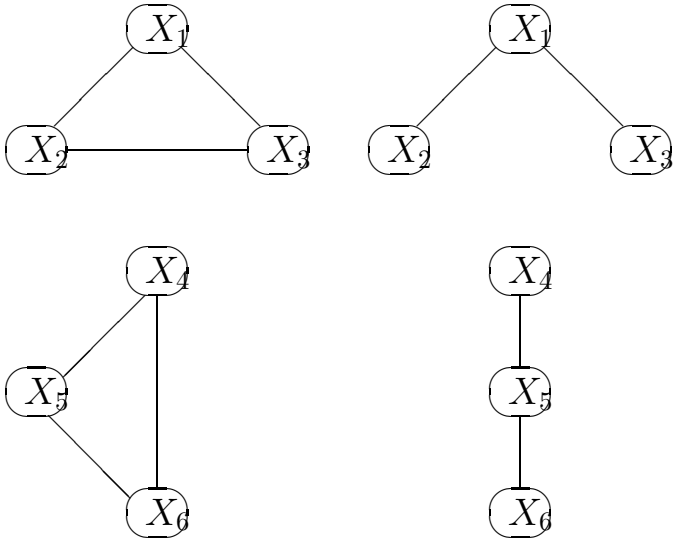
根据数据，将类似的变量分为一大类。



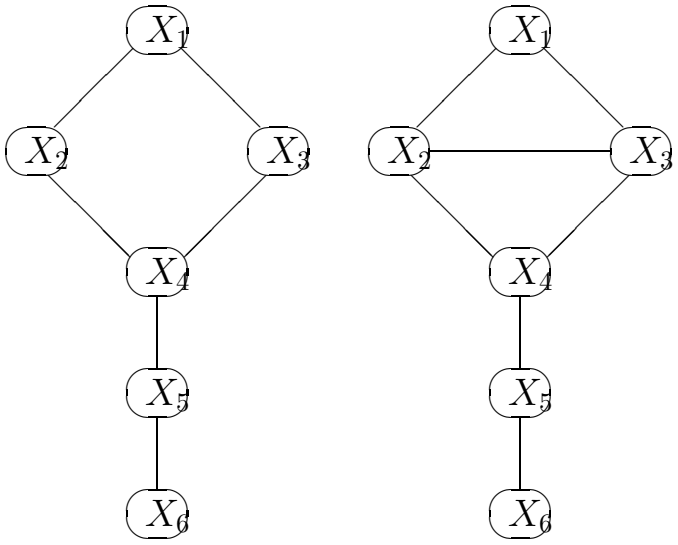
假设的真网络



(a) 聚类结果图



(b)无向子网络



(c)合并后的整体网络

(d)增加道德边后的网络

表5. 吸烟与肺癌的流行病学研究

	(a) 病例对照研究			(b) 随访研究		(c) 概率分布的估计	
	Smoker $X=1$	Non-smoker $X=0$	Total	Smoker $X=1$	Non-smoker $X=0$	Smoker $X=1$	Non-smoker $X=0$
Cancer $Y=1$	688	21	709	8	2	0.0067	0.0002
Control $Y=0$	650	59	709	1992	1998	0.9082	0.0849
Total				2000	2000		

7 多数据库及条件抽样数据(Jia, Geng & Wang, SSPR 2006)

若干个流行病学调查研究得到的多个不同的数据库，每个调查研究有不同的选择样本的标准。

例1 一个病例对照研究是根据疾病状态选定病人，

然后调查他们是否曾经暴露过某种危险因素。

表5(a)关于吸烟与肺癌的病例对照研究。根据该数据得不到肺癌病例在人群中的比率，也得不到相对风险的估计，只能得到优势比的估计。

假设还有一个关于吸烟与肺癌的随访研究，得到的数据如表5(b)所示。根据这个随访研究可以估计相对风险:

$$\frac{\hat{P}(Y=1|X=1)}{\hat{P}(Y=1|X=0)}=\frac{8/2000}{2/2000}=4.$$

因为患病人数非常少，这个估计精度不高。

将病例对照研究与随访研究相结合，得到相对风险的估计：

$$\frac{\hat{P}(Y=1|X=1)}{\hat{P}(Y=1|X=0)}=\frac{0.0067/(0.0067+0.9082)}{0.0002/(0.0002+0.0849)}=3.1160.$$

例2 考虑有两个研究，一个病例对照研究：选择癌症病人和对照病人，看他们是否吸烟，如表6(a)；

另一个研究调查了10000个人中吸烟的频数，如表6(b)。

仅根据病例对照研究得不到患病的分布。根据表6(b)的调查数据也得不到患病的分布。将表6(a)和(b)的两个数据综合，可以得到表6(c)的概率分布的估计，由此可以得到该人群患病的分布0.59%。

表6. 两个调查数据集合和概率分布

(a) 病例对照研究

	吸烟	不吸烟	
癌症病人	688	21	709
对照病人	650	59	709

(b) 吸烟调查

吸烟	不吸烟	总和
9171	829	10000

(c) 概率分布的估计

	吸烟	不吸烟	
癌症病人	0.0057	0.0002	0.0059
对照病人	0.9114	0.0827	0.9941
	0.9171	0.0829	1.0000

进一步，多个研究各有不同的调查项目，如表7所示。可以将多个数据库综合进行统计分析。

表7. 多研究数据库

研究1

病人	血压	年龄	心律	$\cdots$
1	110/70	25	75	$\cdots$
2	120/80	40	60	$\cdots$
$\vdots$	$\vdots$			

研究2

病人	血压	年龄	体温	$\cdots$
1	110/70	25	37	$\cdots$
2	120/80	40	38	$\cdots$
$\vdots$	$\vdots$			

K个条件数据库：$[B_1|A_1], \ldots, [B_K|A_K]$, 其中$\bigcup_k [A_k \cup B_k] = V$。
是否可以识别联合分布？

定理 $P(V)$可由数据库$[B_1|A_1], \ldots, [B_K|A_K]$识别的充分条件：
下面算法得到空集合：

- 初始化$t = 0$, $[B_k^{(0)}|A_k^{(0)}] = [B_k|A_k]\ \forall\ k$, $V^{(0)} = V$.
- 寻找数据库$[B_i^{(t)}|A_i^{(t)}]$ 使得$B_i^{(t)} \cup A_i^{(t)} = V^{(t)}$.
  如果没有这样的数据库, 则$P(V)$不可识别.
- 令$V^{(t+1)} = A_i^{(t)}$, $[B_k^{(t+1)}|A_k^{(t+1)}] = [B_k^{(t)} \setminus B_i^{(t)}|A_k^{(t)} \setminus B_i^{(t)}]\ \forall\ k$, 令$t = t+1$.
- 重复上步，直至$V^{(t)} = \emptyset$.

例： 数据库[456|1237], [147|23], [45|6], [2|1347], [3|45].

- $V^{(0)} = \{1, 2, 3, 4, 5, 6, 7\}$.
- $V^{(1)} = \{1, 2, 3, 7\}$. 数据库化为: [17|23], [2|137], [3].
- $V^{(2)} = \{2, 3\}$. 数据库: [2|3], [3].
- $V^{(3)} = \{3\}$. 数据库: [3].
- $V^{(4)} = \emptyset$.