

病例对照研究 **case – control study**

王伟炳 副教授



复旦大学公共卫生学院流行病学教研室
Tel/Fax: 021-54237811
Email: wwb@fudan.edu.cn

研究简史

- ◆ 1843年Guy向统计学会报告-最早病例对照研究
分析职业暴露与肺结核发生的关系
 - ◆ 1844年Louis的著作
最早出现病例对照研究的概念
 - ◆ 1926年Lane Claypon 报告
生殖因素与乳腺癌关系的研究
-

研究简史

- ◆ 1947年Schreck和Lenowitz
包皮环切和性卫生与阴茎癌的关系
 - ◆ 1947年Hartwell
输血与肝炎关系的研究
 - ◆ 1950年Doll和Hill
吸烟与肺癌的研究
-

研究简史

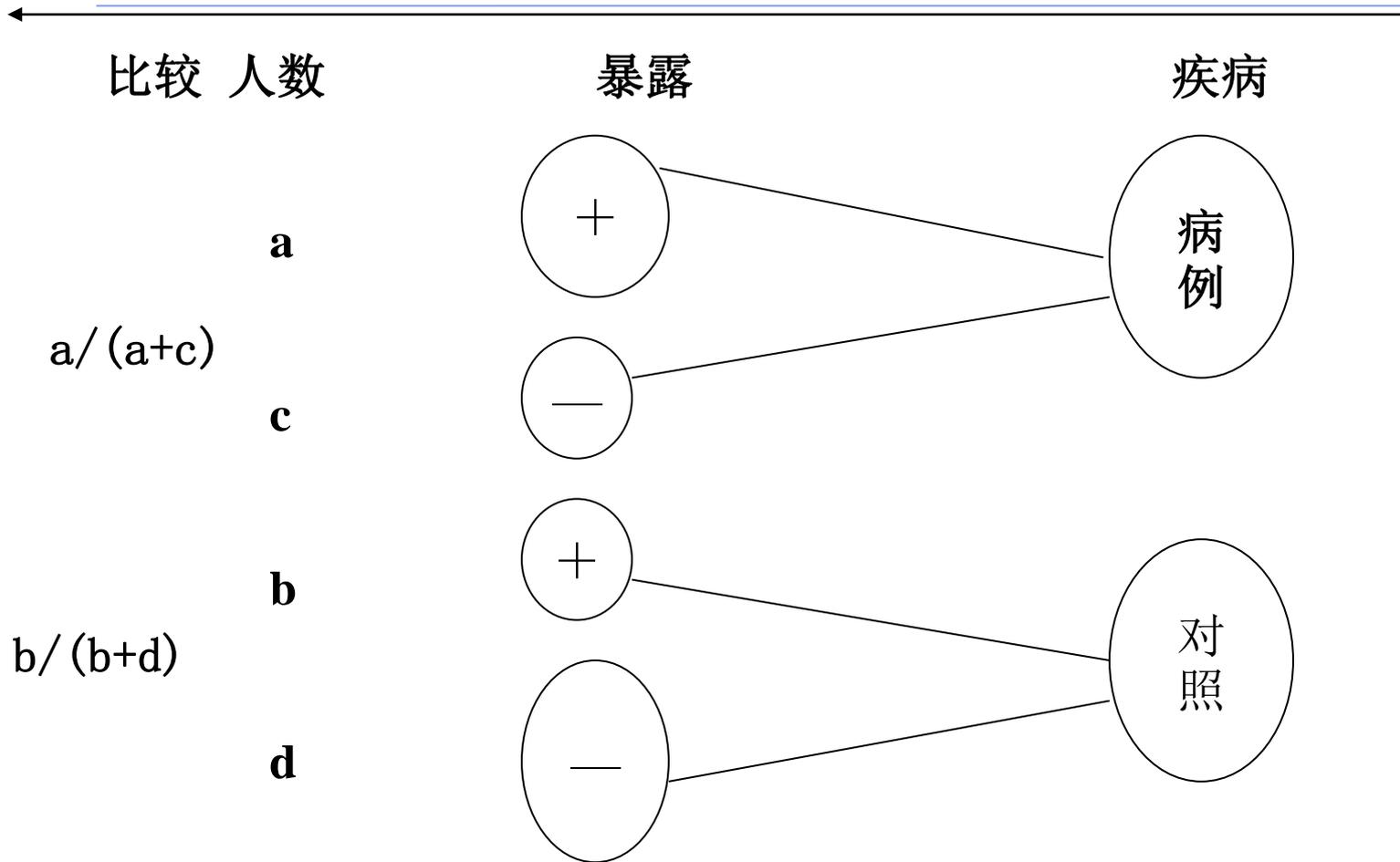
◆ 20世纪60年代以来

- 孕妇服用反应停（**thalidomide**）与婴儿短肢畸形
 - 母亲吸烟与先天性畸形
 - 早产儿吸入高浓度氧与晶体后纤维组织增生症
 - 经期使用月经棉与中毒性休克综合征
 - 小剂量电离辐射与白血病
 - 母亲早孕期服用雌激素与少女阴道腺癌等之间的关系
-

基本原理

- 以现在确诊的患有某特定疾病的病人作为病例，以不患有该病但具有可比性的个体作为对照
- 通过询问、实验室检查或复查病史，收集既往各种可能的危险因素的暴露史，测量并比较病例组与对照组中各因素的暴露比例
- 经统计学检验，若两组差别有意义，则可认为因素与疾病之间存在统计学关联。

回顾性收集暴露情况



暴露

- **暴露 (exposure)** 是指研究对象曾经接触过某些因素或具备某些特征或处于某种状态。这些因素、特征或状态即为暴露因素。
 - 暴露因素可以是机体特征，也可以是体外的；
 - 可以是先天的、人体固有的，也可以是后天获得的。
 - 暴露因素可以是疾病的危险因素，也可以是保护因素。

病例对照研究的设计类型

- 成组病例对照研究
- 匹配病例对照研究
- 巢式病例对照研究

病例与对照不匹配

- 按照一定的原则与标准，在规定的病例和对照人群中，分别抽取一定数量的研究对象，组成病例组与对照组进行研究。这种设计类型的研究称为**成组病例对照研究**。
- 在这种类型的病例对照研究中，一般对照组的人数应等于或多于病例组人数。此外没有其它任何特别的限制与规定。



病例和对照的差别

在进行一项病例对照研究时，需要重视的一个问题是病例和对照在某些非研究目的的特征或暴露上存在差异。

如果病例暴露的人数超过对照，我们可能会怀疑是否观察到的关联是由于病例和对照之间的差异引起，而不是暴露本身。

病例与对照匹配

- **匹配或配比 Δ (*matching*)** 即要求对照在某些因素或特征上与病例保持一致，目的是在两组比较时排除匹配因素的干扰。
- 匹配分成成组匹配和个体匹配

病例与对照匹配

- **成组匹配** 又称频数匹配

匹配因素所占的比例在两组相同。

- **个体匹配**

以病例和对照的个体为单位进行匹配。该匹配方法是指按照匹配因素为每一病例配上一个或一个以上适宜的对照。其中1:1匹配又称配对（pair matching）。

匹配过头△ (overmatching)

- 在以匹配方法选择对照时，匹配因素或称匹配变量必须是已知的混杂因素，或有充分理由怀疑其为混杂因素。否则，若**把不必要的变量列为匹配因素时**，不但达不到匹配的目的，降低研究效率，掩盖研究因素与疾病之间的真正联系，还可增加工作难度与研究成本。这种情况称为**匹配过头或匹配过度**。

两种情况不应使用匹配

- (1) 研究因素与疾病因果链中的中间变量不应匹配。如吸烟→血脂水平→心血管病。
- (2) 只与可疑因素有关而与疾病无关的因素不应匹配。如避孕药（宗教信仰）→某病

病例对照研究用途

- 探索疾病的可疑危险因素

在疾病病因不明时，可以广泛地筛选机体内外环境因素中的可疑危险因素。

- 检验病因假说

经过描述性研究或探索性病例对照研究，初步产生了病因假说后，可以应用设计精良的病例对照研究加以检验。

- 提供进一步研究的线索

病例对照研究特点

- ①观察性研究；
- ②设立对照组；
- ③纵向的、回顾性的、由果及因；
- ④该方法可探索疾病的危险因素，建立或初步检验病因假设，而一般不能确证暴露与疾病是否存在因果关系。

病例对照研究设计和实施

- 复习文献和提出假设
- 提出研究计划或设计书
- 执行调查研究计划
- 资料整理和分析
- 写出研究报告

实施要点 – 研究对象的选择

- 进行病例对照研究时，病例组与对照组的正确选择是该研究成败的关键。
- 病例组足以代表人群中所有的某病病例，而对照组则能代表其所出自的总体。

病例组的选择

- 确诊的病人 病例的正确诊断是防止错误分类偏倚的关键。
 - 要规定明确、统一的疾病诊断标准，所选病例组的成员均应符合规定的诊断标准。
 - 在制订诊断标准时，应采用国际通用或国内统一使用的标准。
 - 应尽可能采用灵敏度与特异度均较理想的检查方法，作为某种研究疾病诊断的执行标准。

病例组的选择△

- 新发病例（incidence case）在病例对照研究中最好选择此种病例。
- 优点：对危险因素的暴露经历回忆可靠，病案和有关疾病的各种信息易于获得。
 - 新发病例发病时间短，易于辨认病前的环境条件、生活行为等，回忆暴露史清晰、可靠
 - 各种医疗、职业史等记录资料也易获得
 - 新发病例尚未明显受到决定生存因素的影响
- 时间花费长。

病例组的选择△

- 现患病例（prevalence case） 由于发病时间长，在回忆过去的暴露情况时可能易出现偏差，并且此种病例是发病病例中的幸存者，已经过了影响生存因素的考验。如果影响新发病例与现患病例的因素不同，选用现患病例，则可能会导致错误的结论。
- 先天畸形、非致死性的慢性病（肥胖、高血压 - 不了解准确的发病时间）

病例组的选择

- 死亡病例

其暴露史的获得往往是由他人代述的，出现偏差的可能性会更大。因此，一般情况下不用死亡病例，但特殊情况例外。

病例组的选择

为了增加病例组和对照组在除研究因素以外的其它因素方面具有较好的均衡性或研究实施的可行性，在选择研究对象时亦可对病例和对照的其他特征进行限定。

病例组的来源 — 医院

- 从医院选病例，应明确规定某所或若干所医院在一定时期内确诊的全部病例，或其随机样本。
- 在医院内选病例，来源多，易收集，病例合作性好，资料较可靠。
- 但病人对医院，或医院对病人都可能存在一定的**选择性**，故易产生选择偏倚。
- 从医院中选病例时应特别注意在病情轻重、病期早晚等方面的代表性。

病例组的来源 – 人群

- 一般指某人群在一定时期发现的病例或其样本。
- 从人群中选病例其代表性较好
 - 但病例往往不易收集，合作性差，实际执行时往往困难较多。

病例组的来源

一 疾病（死亡）报告、登记系统

一个地区的疾病报告、死亡报告、或保险公司的病例、死亡记录等，亦均可作为病例对照研究的病例来源。

- 应用此类来源选择病例时，应注意资料的完整性，是否符合有关的标准与要求等。
- 可选择在特定时间内的所有某病病例、死亡病例或其样本进行研究。

对照组的選擇

对照组的選擇原則

1. **不患所研究的疾病，但有暴露于研究因素的
可能**
2. 不患与研究因素有关的其他疾病
3. 能以与病例相同的方法获取暴露资料
4. 与病例组的可比性
5. 尽可能设立多种对照

对照组的来源

1. 与病例同一医院的患其他疾病的患者。
2. 产生病例的一般人群的随机样本。
3. 与病例同一队列的非病例。
4. 病例的邻居亲属、朋友、同事、同学等。
5. 自以上两种或两种以上不同来源选择对照。

人群对照

- 优点：
 - 避免选择偏倚
 - 直接描述分布
- 不足
 - 花费高，耗时
 - 病例与对照有不同的回忆偏倚程度

医院对照

- 优点
 - 易寻找，集中
 - 应答率高
- 缺点
 - 对照所患疾病与研究疾病的病因可能相同
冠心病、口腔癌与饮酒——低估饮酒与口腔癌
 - 同样存在选择偏倚

样本大小估计 Δ

影响样本大小的因素

1. 研究因素的人群暴露率估计值，一般可用对照组的暴露率 (P_0) 来代替。
2. 估计的研究因素与研究疾病的相对危险度 (Relative Risk, RR) 或比值比 (Odds Ratio, OR)。
3. 所允许犯假阳性错误的概率 α 。
4. 所希望达到的检验把握度 ($1-\beta$)。

◆ 方 法

- 查表法

- 公式法：求病例对照研究
样本含量的公式

$$N = \frac{\left[Z_{\alpha} \times \sqrt{2\bar{P}(1-\bar{P})} + Z_{\beta} \sqrt{P_1(1-P_1) + P_0(1-P_0)} \right]^2}{(P_1 - P_0)^2}$$

- 或近似公式

$$N = \frac{(Z_{\alpha} + Z_{\beta})^2 \times \bar{P}(1 - \bar{P})}{(P_1 - P_0)^2}$$

暴露组患病率

对照组患病率

例：为研究某市肺癌与吸烟的关系，欲进行一次病例对照研究。

某市普通人群中吸烟率 $P_0=30\%$

OR=5.0 $\alpha=0.05$

把握度=0.90

问需要多少病例与对照？

$$P_i = RP_o / (1 - P_o + RP_o) = 5 \times 0.3 / (1 - 0.3 + 5 \times 0.3) = 0.68$$

$$\bar{p} = \frac{p_i + p_o}{2} = \frac{0.68 + 0.30}{2} = 0.49$$

α 、 β 值查表， Z_α 为1.64， Z_β 为1.28

代入公式得：

$$N = \frac{\left[1.64\sqrt{2 \times 0.49 \times (1 - 0.49)} + 1.28\sqrt{0.68(1 - 0.68) + 0.30(1 - 0.30)} \right]^2}{(0.68 - 0.30)^2} = 27.6$$

病例组与对照组，各需28人。

如代入近似公式计算，所得结果与原公式很接近

($n=29.5$)

病例组与对照组例数不等时的公式：

病例数：对照数=1：c

$$n = \frac{\left(1 + \frac{1}{c}\right) \times \bar{p}(1 - \bar{p})(Z_{\alpha} + Z_{\beta})^2}{(p_1 - p_0)}$$
$$\bar{p} = \frac{(p_1 + cp_0)}{1 + c}$$

对照组例数为c*n

资料收集

- 因素的选定：包括所研究的因素、其他可疑因素和可能的混杂因素。
- 因素的规定：调查前有明确的规定（定量、分级）
- 因素的收集：病例组与对照组的调查项目相同。
最好有记录或材料作依据。客观指标更好。

暴露信息的内容

- 暴露信息的正确收集是病例对照研究资料收集的重要任务。
- 流行病学研究中的暴露包括研究对象在外环境中接触的某些因素，如化学的、物理的、生物学的因素等，以及机体本身具有的生物学、社会学、心理学特征等等。
- 收集的信息包括所研究的危险因素、可疑危险因素，保护因素、潜在的混杂因素等等。

资料的分析

- 描述性统计
- 统计推断

资料分析

- 病例对照研究资料的基本分析方法是比较病例与对照组中研究因素暴露的比例，估计研究因素与疾病发生的危险度。
- 控制潜在混杂因素的混杂作用，从而对研究因素与疾病的关系做出正确的估计。

成组设计不分层资料

表4 病例对照研究资料整理表

研究因素 暴露	病例	对照	合 计
+	a	b	$a+b=n_1$
-	c	d	$c+d=n_0$
合计	$a+c=m_1$	$b+d=m_0$	$a+b+c+d=t$

成组设计不分层资料

- 比较病例组的暴露比例 $(\frac{a}{a+c})$ 与对照组的暴露比例 $(\frac{b}{b+d})$ 有无统计学差异。

$$\chi^2 = \frac{(ad - bc)^2 T}{N_1 N_0 M_1 M_0}$$

成组设计不分层资料

吸烟与肺癌关系的病例对照研究

吸烟史	病例	对照	合计
吸 烟	41(a)	28(b)	69(N ₁)
不吸烟	19(c)	32(d)	51(N ₀)
合 计	60(M ₁)	60(M ₀)	120(T)

$$\chi^2 = \frac{(ad - bc)^2 T}{N_1 N_0 M_1 M_0} = 5.76$$

$$p=0.016$$

成组设计不分层资料

- 联系强度测量 - 比值比(OR) :OR即指病例组研究因素的暴露比值与对照组的暴露比值之比。
- 比值 (odds) 是指某事物发生的概率与不发生的概率之比。

比值比(OR)

- 病例组的暴露比值为 $\frac{a}{a+c} / \frac{c}{a+c} = a/c$;
- 同理可获得对照组的暴露率比值为 b/d 。

$$\therefore OR = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{ad}{bc}$$

- 前例 $OR = ad/bc = 2.47$

$$OR = \frac{a/n_1 / c/n_1}{b/n_2 / d/n_2} = \frac{\frac{P_1}{1-P_1}}{\frac{P_2}{1-P_2}} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

在发病率低的情况下，OR可以用来估计相对危险度。如果暴露组和非暴露组发生某病的概率都很小，即： $1-P_1 \approx 1$ ， $1-P_2 \approx 1$

$$OR = \frac{a/n_1 / c/n_1}{b/n_2 / d/n_2} = \frac{\frac{P_1}{1-P_1}}{\frac{P_2}{1-P_2}} = \frac{P_1/1}{P_2/1} = \frac{P_1}{P_2} = RR$$

OR 与相对危险度 (RR)

- OR Δ 的含义与相对危险度 (RR) 相同，即暴露组发生疾病的危险性为非暴露组的多少倍。RR的本质为率比 (rate ratio)，即暴露组与非暴露组发病率 (或死亡率) 之比。但是在通常情况下，病例对照研究一般不能获得发病率资料，只可计算OR。
- 当所研究疾病的发生率小于5%时，OR是RR的极好近似值

表6 不同发病率和OR时的RR

OR	发病率			
	0.2	0.1	0.05	0.01
2	1.7	1.8	1.9	2
3	2.1	2.5	2.7	2.9
4	2.5	3.1	3.5	3
5	2.8	3.6	4.2	4.8
6	3	4	4.8	5.7
7	3.2	4.4	5.4	6.6
8	3.3	4.7	5.9	7.5
9	3.5	5	6.4	8.3
10	3.6	5.3	6.9	9.2

OR的可信区间估计

- Miettinen法:

$$OR_L, OR_U = OR^{(1 \pm 1.96 / \sqrt{\chi^2})} = (1.18 - 5.18)$$

- Woolf法:

$$OR_L, OR_U = \exp(\ln OR \pm 1.96 \sqrt{\text{Var}(\ln OR)}) \\ = (1.19 - 5.14)$$

匹配设计资料分析 (1:1)

表9 1:1匹配病例对照研究资料整理表

对照	病例		对子数
	暴露史(+)	暴露史(-)	
暴露史(+)	a	b	a+b
暴露史(-)	c	d	c+d
对子数	a+c	b+d	n

匹配设计资料分析 (1:1)

$$\chi^2 = \frac{(b - c)^2}{b + c}$$

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c}$$

$$OR = c/b$$

匹配设计资料分析 (1:1)

表10 外源性雌激素与子宫内膜癌关系配对资料

对照	病例		对子数
	暴露史(+)	暴露史(-)	
暴露史(+)	27(a)	3(b)	30(a+b)
暴露史(-)	29(c)	4(d)	33(c+d)
对子数	56(a+c)	7(b+d)	63(n)

匹配设计资料分析 (1:1)

$$\chi^2 = \frac{(|3 - 29| - 1)^2}{3 + 29} = 19.53, \quad p < 0.005$$

$$OR = 29/3 = 9.67$$

OR95% 可信区间：

$$OR_L, OR_U = \mathbf{OR}^{(1 \pm 1.96/\sqrt{\chi^2})} = 9.67^{(1 \pm 1.96/\sqrt{19.53})}$$

$$= 3.56 \sim 26.24$$

病例对照研究中的偏倚与控制

- 病例对照研究是一种回顾性的观察性研究，各类偏倚均易产生，如若不能有效防止，可严重影响研究结果的有效性（validity）。偏倚可发生在病例对照研究的设计、实施、分析乃至推论的各个阶段。

选择偏倚

- 入院率偏倚 (admission rate bias) 亦称**伯克森氏偏倚 (Berkson's bias)** ，是指利用医院就诊或住院病人作为研究对象时，由于**入院率的不同而导致的偏倚**。不同疾病在某一类医院的就诊或住院率各异，其原因是多方面的，如不同医院的技术专长，患者所患疾病的严重程度，患者的经济状况，以及就诊方便与否等等，均可影响入院率。因此，若在医院内选择研究对象进行病例对照研究时，这种偏倚出现的可能性较大。

选择偏倚

- 现患病例—新病例偏倚 (Prevalence-incidence bias) 又称**奈曼偏倚 (Neyman bias)** ▲。如果研究对象选自现患病例，可能会得到更多的信息，但其中很多信息**可能只与存活有关，而未必与该病的发病有关**，从而可高估某些暴露因素的病因作用。此外，某些疾病的幸存者可能已改变了原先的生活习惯，从而降低了某危险因素的水平，或当其被调查时夸大或缩小了病前生活习惯上的某些特征，从而导致对某因素与疾病关联的错误估计。

选择偏倚

- 检出偏倚 或称检出症候偏倚 (detection signal bias) Δ 。指某因素与某疾病在病因学上虽无关联，但由于该因素的存在而引起该病症状或体征的出现，从而使患者及早就医，使该群体有较高的检出率，以致得出该因素与该疾病相关联的错误结论。
- 在对一些慢性疾病如肿瘤、动脉硬化、结石等进行病因研究时，这种偏倚的意义特别重要。

选择偏倚

- 易感性偏倚 (susceptibility bias) 研究对象是否暴露于某一可疑致病因素，既与各种客观因素有关，也与研究对象的主观因素有关。这些因素可能直接或间接的影响观察人群或对照人群对所研究疾病的易感性，导致某因素与某疾病间的虚假联系，由此而产生的偏倚称为易感性偏倚。
- 如在对职业性疾病研究中的健康工人效应△ (healthy worker effect) 等。

选择偏倚的控制

主要应通过科学的研究设计和认真的实施，避免其发生。

- 研究者对在研究中可能会出现各种选择偏倚应有充分的了解、掌握。
- 严格掌握研究对象纳入与排除的标准。
- 尽量采取多种对照。

信息偏倚

- 信息偏倚又称观察偏倚（observational bias），是指在研究的实施阶段自研究对象获取研究所需要的信息时所产生的系统误差。
- 信息偏倚可来自于研究对象、研究者本身，也可来自于用于测量的仪器、设备、方法等。
- 信息偏倚的表现是使研究对象的某种特征被**错误分类**，如暴露于某因素者被错误的认为是非暴露者，某病的患者被错误的认为是非患者等等。

信息偏倚

- 回忆偏倚 (recall bias) 回忆偏倚是指研究对象在回忆某些因素的暴露史时，由于在准确性和完整性上的差异所导致的系统误差。
- 回忆偏倚在病例对照研究中最常见 Δ ，其产生与以下原因有关：
 - 调查的事件或因素发生的频率甚低，未给研究对象留下深刻的印象而被遗忘；
 - 调查事件是很久以前发生的事情，研究对象记忆不清；
 - 研究对象对调查的内容或事件关心程度不同，因而回忆的认真程度有异等。

信息偏倚

- 测量偏倚 (detection bias) 测量偏倚指对研究所需指标或数据进行测定或测量时产生的偏差。如所用仪器、设备校正不准确，试剂不符合要求。使用方法的标准或程序不统一，分析、测试条件不一致，以及操作人员的技术问题等等，均可导致测量结果的不正确，使测量结果偏离真值。
- 所用调查表设计的科学性、记录是否完整，调查人员对工作的认真程度以及访问方式、态度等等，亦可导致不准确的信息，产生测量偏倚。

信息偏倚控制方法

- 要制定明细的资料收集方法和严格的质量控制方法。
- 尽可能采用“盲法”收集资料。
- 尽量采用客观指标的信息。
- 通过一定的调查技巧。

混杂偏倚

- 混杂偏倚或称混杂（confounding），是指在病例对照研究中，由于一个或多个潜在的混杂因素（confounding factor）的影响，掩盖或夸大了研究因素与疾病（或事件）之间的联系，从而使两者之间的真正联系被错误地估计。

混杂偏倚控制

- 限制 (restriction)
- 匹配 (matching)
- 分层分析
- 多因素分析

病例对照研究的优点

- 简便、省力（人力、物力），省经费，效率高。
- 可同时研究一种疾病与多种因素间的关系。
- 可初步验证已明确的病因假设，亦可广泛探索众多因素与某种疾病的关系，用以建立病因假设。
- 特别适用于少见病的病因或危险因素研究，因其以较小的样本可得出有价值的结果。若以队列研究方法研究发病率很低的罕见病，则需要很大的样本，在实际工作中往往实行的可能性很小。

病例对照研究的缺点

- 不能直接计算疾病的率，只能估计相对危险度，虽可为进一步的研究提供线索，但不能直接得出因果关系的结论。
- 易发生各种偏倚，尤其是回忆偏倚
- 不适于研究人群中所研究因素暴露比例很低的因素
- 选择研究对象时常难以为病例选择到适宜的对照。

总结

- 是迄今为止最常用的流行病学研究方法之一。
- 广泛应用于探讨疾病病因、干预措施、项目评价以及公共卫生与医学实践的许多方面。
- 特别适用于潜伏期长的、少见病的研究，有时是识别其危险因素的唯一可行的方法。
- 也特别适用于研究分析多种危险因素与研究疾病的联系以及它们之间的相互作用等。