



III、药物流行病学研究方法 - 分析性研究(续)

郑国民(上海新兴医药股份有限公司, 上海 200135)

中图分类号: R181.3⁺5

文献标识码 B

文章编号: 1006-0111(2003)06-0375-06

7 病例对照研究的数据分析和结果的评价

病例对照研究的评价标准十分强调严格的抽样结构的群体性、病例和对照组的可比性、药物暴露的真实性、药物不良反应或非预期的药物反应的确立过程及考虑混杂因素的影响等等。

数据分析主要是比较病例组与对照组对研究因素(或称暴露因素)的暴露史有无显著差别,包括暴露率/或暴露剂量的差别。同时应估计暴露与疾病(即药物不良反应)的联系程度。具体方法可以从简单到复杂,先把调查的暴露数据以 2 × 2 表列出,计算出相对危险度(relative risk, RR)或比值比(odds ratio, OR),用它来测量暴露与疾病之间有无联系及其联系强度。然后进一步用分层来消除混杂因素对暴露因素的影响。在证明暴露因素与疾病有联系的基础上可进行剂量效应的测定。非配对和配对设计的研究,分析方法有不同,分述如下:

7.1 频率或分层的成组配比(群组配比)的数据资料分析

暴露因素	病例组	对照组	合计
有	a	b	m ₁
无	c	d	m ₂
合计	a + c(n ₁)	b + d(n ₂)	N

7.1.1 RR 的估计值即 OR

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{a(c+d)}{c(a+b)} = \frac{\text{暴露的发病率}}{\text{非暴露的发病率}}$$

因为病例对照研究的数据资料不能直接计算暴露和非暴露人群的发病率,所以要用 RR 的估计值代替(即 OR)。

$$OR = \frac{ad}{bc}$$

OR 的计算公式,只有当调查疾病是人群中罕见病(某种药物不良反应在人群中的发生率低于 10%)和病例及对照来自相同的总体,调查人数又相当多时,发病率计算式中的 a 和 c 可忽略不计,则

RR = OR。所以病例对照研究中的 RR 是一个近似的相对危险度。

7.1.2 用 χ^2 检验测定样本的 OR 是否有显著性

$$\chi^2 = \frac{(|ad - bc| - \frac{N}{2})^2 N}{n_1 n_2 m_1 m_2}$$

χ^2 检验的目的是想通过 P 值的大小表明样本 OR 值是否明显偏离 1,因为 OR = 1 时,认为疾病与暴露因素间无联系,所以样本的 OR 如果不明显偏离于 1,则可能样本的 OR 值是抽样误差造成。

7.1.3 计算 OR 的 95% 可信限(Miettinen 氏法)

$$95\% \text{ OR 值可信限} = OR^{(1 \pm 1.96/\sqrt{\chi^2})}$$

95% 可信区间,表明 OR 值,有 95% 的可能在该可信区间内。当可信区间内包含 1 时,表示暴露与非暴露之间无差别或说明样本 OR 值无意义,例如可信区间内 0.8 ~ 4.0。另外,可信区间越窄则可靠性越大,精确度越高,例如 1.8 ~ 4.0 的可信区间比 1.2 ~ 36.0 的好。所以判断 OR 意义时要以 95% 可信限为准。

举例^[1]为研究子宫内膜癌患者与过去服用雌激素史的关系,用病例对照研究方法,共调查 188 例内膜癌患者,另调查 188 例同年出生未患内膜癌的妇女作对照,了解发病与过去使用雌激素史的关系,其结果如表 5。

表 5 子宫内膜癌与服用雌激素的关系

雌激素暴露史	病例组	对照组	合计
+	108	29	137
-	80	159	239
合计	188	188	376

OR = ad/bc = (108 × 159)/(80 × 29) = 7.4。OR 是 RR 的估计值或近似值,如 RR = 1,说明暴露因子与疾病无联系。如 RR > 1 时,表示存在正联系,RR 值越大表示联系强度大。本例 OR = 7.4,表示有雌激素暴露史的患子宫内膜癌危险性是无暴露史的 7.4 倍。如 RR < 1 时,说明存在负联系,提示暴露因素

是一种保护因素。例如, RR = 0.25, 则表示危险性降低了 75%。但是计算出来的 RR(或 OR)一定要经过卡方检验和 95% 可信限的测试才能判断其统计学上的显著性。

本例卡方检验根据上述卡方检验公式,

$$\chi^2 = \frac{(1108 \times 159 - 29 \times 801 - 376/2)^2 \times 376}{188 \times 188 \times 137 \times 239} = 69.86$$

自由度为 1, $\chi_2 > \chi_{0.01}^2$, $P < 0.01$, 说明 OR = 7.4 是抽样误差造成的可能性极小或者说明 7.4 的 OR 值极显著地偏离于 1。

95% OR 值可信限 = $OR^{(1 \pm 1.96/\sqrt{69.86})} = 4.6 \sim 11.8$

说明总体 OR 值有 95% 可能性在 4.6 ~ 11.8 区间内, 此区间内不包括 1, 说明有显著差异。

7.2 个体配比(配对调查)资料分析

7.2.1 一个病例配比一个对照病例的对照研究资料分析。其数据按下列四格表归纳之。

表 6 个体配比数据归纳表(1:1 配比)

	不良反应病例组某药暴露史	
	+	-
对照组 +	a	b
某药暴露史 -	c	d

表格内 a、b、c、d 的数值分别代表配比的病例和对照的对子数。例如 a 所代表数据是病例和其配比的对照都有某药暴露史的对子数, 计算 OR 和卡方检验时都用对子数。

计算公式

$$OR_{M-H} = \frac{c}{b}$$

卡方检验用 McNemar's 公式, 即

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

$$OR^{95\% \text{ 可信限}} = OR^{(1 \pm 1.96/\sqrt{\chi^2})}$$

举例^[1]: 1976 年 Mack 等报告了在洛杉矶所做的外源性雌激素与子宫内膜癌关系的 1:1 配比的病例对照研究, 作者从 1971 ~ 1975 年共收集退休妇女中 63 例子宫内膜癌的新发病例, 要求对照妇女与病例处于同一居住地, 具有同样的婚姻状况, 居住时间类似, 同时要求对照妇女在病例被诊断以前未做过子宫摘除术, 提示对照妇女如在危险因素的作用下仍可患子宫内膜癌。药物暴露史资料是从每个病例和对照的匹配医院病历中查出来的, 调查所获得的数据资料归纳如下表。

外源性雌激素与子宫内膜癌关联的 1:1 配比

的病例对照研究数据。

表 7 外源性雌激素与子宫内膜癌的关系

	病例组雌激素暴露史	
	+	-
对照组 +	27	3
雌激素暴露史 -	29	4

$$OR = \frac{c}{b} = \frac{29}{3} = 9.67$$

$$\chi^2 = \frac{(b - c - 1)^2}{b + c} = \frac{(3 - 29 - 1)^2}{3 + 29} = 19.53$$

$\therefore P = 0.00005$

说明 OR = 9.67 是机会造成的可能性极小。

95% OR 可信限 = $9.67^{(1 \pm 1.96/\sqrt{19.53})} = 3.53 \sim 26.48$

在此可信限内不包括 1, 而且说明 OR 值有 95% 可能性在此区间内。

根据上述结果说明用过雌激素的妇女患子宫内膜癌的危险性是未用的 9.67 倍。

7.2.2 1:M 配比资料分析在病例对照研究中, 如对照较容易选择, 为了识别和控制混杂因素的混杂作用, 可用一个病例选用多个对照(一般不超过 4 个), 如果对照超过 3 个, 其数据资料分析需用专门的方法。

\overline{OR} 的计算用 Mantel - Haenszel 方法, 其公式如下:

$$\overline{OR}_{M-H} = \frac{\sum_{j=0}^c (c-j)n_j(+)}{\sum_{j=0}^c jn_j(-)}$$

注: c = 配对比例即代表是 4 个或是 5 个对照

j 代表对照者有暴露史的个数

$n_j(+)$ 代表病例有暴露史的, 而同时对照者中不同暴露个数的实际“对子数”

$n_j(-)$ 代表病例是无暴露史而同时对照者中在不同暴露个数下的实际对子数

举例^[2]: 在上述例子中同一工作的著者 Mack 等(1976)报道了外源性雌激素与子宫内膜癌的关系, 是 1 个病例选择 4 个对照的 1:4 配比方法, 其研究结果归纳如下表。

表 8 雌激素与子宫内膜癌关系的 1:4 配对研究结果

病例	用雌激素的对照数					合计
	0	1	2	3	4	
用雌激素	3 ¹⁾ $n_0(+)$	17 $n_1(+)$	16 $n_2(+)$	15 $n_3(+)$	5 $n_4(+)$	56
不用雌激素	0 $n_0(-)$	4 $n_1(-)$	1 $n_2(-)$	1 $n_3(-)$	1 $n_4(-)$	7

¹⁾注: 表中的数值 3 是表示病例是暴露的, 而对照 4 个中无一个是暴露的对子数。

$$OR_{M-H} = \frac{(4-0) \times n_0(+)+ (4-1) \times n_1(+)+ (4-2) \times n_2(+)+ (4-3) \times n_3(+)+ (4-4) \times n_4(+)}{0 \times n_0(-)+ 1 \times n_1(-)+ 2 \times n_2(-)+ 3 \times n_3(-)+ 4 \times n_4(-)}$$

将表 8 中数据代入

$$= \frac{4 \times 3 + 3 \times 17 + 2 \times 16 + 1 \times 15 + 0 \times 5}{0 \times 0 + 1 \times 4 + 2 \times 1 + 3 \times 1 + 4 \times 1} = 8.46$$

显著性检验:作显著性检验前,先计算下列统计量 $m_j = n_j - 1(+)+ n_j(-)$ 。式中 m_1 表示 5 个中

$$T_1 = \sum_{j=0}^c n_j(+)= 3 + 17 + 16 + 15 + 5 = 56$$

$$T_2 = \sum_{j=0}^c j m_j / c + 1 = 1 \times 7 / (4 + 1) + 2 \times 18 / 5 + 3 \times 17 / 5 + 4 \times 16 / 5 = 31.6$$

$$T_3 = \sum_{j=0}^c j(c+1-j) m_j / (c+1)^2 = 1 \times 4 \times 7 / 5^2 + 2 \times 3 \times 18 / 5^2 + 3 \times 2 \times 17 / 5^2 + 4 \times 1 \times 16 / 5^2 = 12.08$$

卡方检验:

$$\chi^2_{M-H} = (|T_1 - T_2| - 1/2)^2 / T_3 = (|56 - 31.6| - 0.5)^2 / 12.08 = 47.286$$

非校正卡方:

$$\chi^2_{M-H} = (|T_1 - T_2|)^2 / T_3 = (|56 - 31.6|)^2 / 12.08 = 49.285$$

自由度为 1, $P < 0.0001$

$$\overline{OR}_{M-H} 95\% \text{ 可信限} = \overline{OR}^{(1 \pm 1.96/\sqrt{\chi^2})} = 8.46^{(1 \pm 1.96/\sqrt{47.286})} = 4.60 \sim 15.54$$

卡方检验证明差异极显著, $OR = 8.46$ 是由于偶然的机会造成的可能性很小很小。其 95% 可信限说明 OR 值有 95% 的可能性在 4.60 ~ 15.54 之间,在此区间不包括有 1,由此再次证明用雌激素与子宫内膜癌发生是有关联,用雌激素的妇女发生子宫内膜癌的危险性是不用者的 8.46 倍。

7.3 混杂的分析^[3]

7.3.1 混杂及混杂偏倚 当我们研究某个因素与某种疾病的关联时,由于某个既与疾病有制约关系,又与所研究的暴露因素有联系的外来因素(extraneous factor)的影响,掩盖或夸大了所研究的暴露因素与疾病的联系。这种现象或影响叫混杂(confounding),其所带来的偏倚叫混杂偏倚(confounding bias),该外来因素叫混杂因素(confounding factor)。混杂是暴露因素与疾病之间的一种非因果性联系。这种联系是由于引起混杂的因素在暴露组和非暴露组中的分布不均匀所致。

成为一种混杂因素必须具备的条件是:第一,是所研究疾病的危险因素;第二,与所研究的因素有联系;第三,不是所研究因素与疾病之间发病机制的中间步骤。

7.3.2 混杂因素的判断 目前尚无确切有效的方法,靠统计学显著性来判断,有时不切实际。因为混杂是有效性问题,有时即使统计学无显著性,仍需调整,以期获得联系的最佳估计。有时统计学上有显著性,但无实质性联系,则又不必调整。靠经验来判

断,常常又缺乏科学依据。因此判断混杂因素,除考虑其与疾病和所研究因素之间的实质性联系之外,还需视联系被歪曲的程度。

如何判断可疑的混杂因素是否需要调整?这可视 cOR/OR_i 的比值(cOR 为粗的、未经调整的 OR , OR_i 为第 i 层的 OR)而定,如 cOR/OR_i 的比值接近 1,可不调整。另外也可视混杂因素的分布差异有无统计学显著意义,调整后其他变量的作用是否看得更清楚等情况而定。例如,一项关于口服避孕药与心肌梗死的病例对照研究,结果如表 9

表 9 口服避孕药(OC)与心肌梗死(MI)关系的病例对照研究结果

	服 OC	未服 OC	合计
病例(MI)	39	114	153
对照	24	154	178
合计	63	268	331

$$cOR = 2.19 \text{ (公式: } \frac{ad}{bc} = \frac{39 \times 154}{114 \times 24} \text{)}$$

考虑到年龄与口服避孕药(OC)有关,也与 MI 有关,可能是个混杂因素。故可按年龄分层,分为 < 40 岁和 ≥ 40 岁两层,如表 10。

表 10 按年龄分层的结果

	<40 岁			≥ 40 岁		
	服 OC	未服 OC	合计	服 OC	未服 OC	合计
病例	21(a ₁)	26(b ₁)	47(m ₁₁)	18(a ₂)	88(b ₂)	106(m ₁₂)
对照	17(c ₁)	59(d ₁)	76(m ₀₁)	7(c ₂)	95(d ₂)	102(m ₀₂)
合计	38(n ₁₁)	85(n ₀₁)	123(t ₁)	25(n ₁₂)	183(n ₀₂)	208(t ₂)
	$OR_1 = 2.80$			$OR_2 = 2.78$		

计算各层的 OR。

$$OR_1 = (21 \times 59) / (17 \times 26) = 2.80$$

$$OR_2 = (18 \times 95) / (7 \times 88) = 2.78$$

两层的 OR_i 均较不分层时 OR 大。

进一步分析在非暴露组年龄与 MI 的关联, 见表 11。

表 11 年龄与 MI 的关联

	<40 岁	≥40 岁
MI	26	88
对照	59	95

$OR = 0.48, \chi^2 = 7.27$, 说明年龄与 MI 有联系(年龄小有保护)。

再分析对照组中年龄与口服避孕药的关联, 见表 12。

表 12 年龄与 OC 的关联

	<40 岁	≥40 岁
OC	17	7
对照	59	95

$OR = 3.91, \chi^2 = 8.89$, 说明年龄与口服避孕药也有联系。

年龄不是 OC 与 MI 联系的中间环节, 故可以认为年龄是研究 OC 与 MI 关系时的混杂因素。

7.3.3 混杂方法的控制 第一, 在设计时利用限制和配比的方法来选择研究对象; 第二, 在资料分析阶段采用分层分析, 如有多个因素混杂, 以致分层太多, 分层法不能胜任控制混杂的任务, 则须用多变量分析。如多元回归及 Logistic 回归法(利用计算机运算)。下面介绍在资料分析阶段控制混杂的分层调整法。

7.3.3.1 Mantel - Haensezed 法 计算总的 OR 值, 用 Mantel - Haensezed 提出的公式:

$$OR_{MH} = \sum (a_i d_i / t_i) / \sum (b_i c_i / t_i)$$

表 10 的数据 $OR_{MH} = \frac{21 \times 59}{17 \times 26} + \frac{18 \times 95}{7 \times 88} = \frac{123}{123} + \frac{208}{208} = 2.79$

OR_{MH} 即为调整后的 OR (adjusted OR, aOR)。

$\frac{cOR}{aOR}$ 是反映混杂程度的一种指标, 上例的 $\frac{cOR}{aOR}$

$= \frac{2.19}{2.79} = 78.49\%$, 它的意思是经过调整后的 aOR 比 cOR 增大了, cOR 对 aOR 来说低估了 $(1 - 78.49\%) = 21.51\%$ 。

对 OR_{MH} 的显著性检验公式为:

$$\chi_{MH}^2 = [\sum a_i - \sum E(a_i)]^2 / \sum V(a_i)$$

式中 $\sum E(a_i)$ 为 $\sum a_i$ 的理论值,

$$\sum E(a_i) = \sum m_{li} n_{li} / t_i$$

式中 $\sum V(a_i)$ 为 $\sum a_i$ 的方差,

$$\sum V(a_i) = \sum_{i=1}^I \frac{m_{li} m_{oi} n_{li} n_{oi}}{t_i^2 (t_i - 1)}$$

其中 I 为分层的总层数, i 为第几层。

表 10 数据计算结果如下, $\chi_{MH}^2 = 11.79$

自由度 $\nu =$ 处理组 - 1 = 4 - 1 = 3, 查 χ^2 界值表, $P < 0.01$ 。

估计总 OR 值 95% 的可信区间: 可用 Miettinen 法计算

$$OR_U, OR_L = OR_{MH}^{(1 \pm 1.96 / \sqrt{\chi_{MH}^2})}$$

即 OR_{MH} 95% CI 的上限为 3.18, OR_{MH} 下限为 2.22。如计算 99% CI, 将上式中的 1.96 换成 2.58 即可。可信区间中不包括 1.0, 即可认为该 OR 值在 0.05 或 0.01 水平上有显著性。

由以上分析可以看出, 分层后的 OR_{MH} 为 2.79, 如不分层分析, 则 OR 值为 2.19, 说明由于混杂因素年龄的作用, 使得到的暴露因素避孕药与心肌梗死的关联趋向于 1, 即减弱了它们的关联。

7.3.3.2 Hacuk 加权法

$$OR_w = e^{\sum (W_i \times \ln OR_i / \sum W_i)}$$

上式中 $W_i = \frac{1}{\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}}$

例如 $W_1 = \frac{1}{\frac{1}{21} + \frac{1}{26} + \frac{1}{17} + \frac{1}{59}} = 6.178 = 6.18$

$$\chi_w^2 = (\sum W_i) \times (\ln OR_w - \ln OR_0)^2$$

$$= (\sum W_i) \times (\ln OR_w - 1)^2$$

$$= (\sum W_i) \times (\ln OR_w)^2$$

上例 $W_1 = 6.18, W_2 = 4.54$

$$\ln OR_1 = 1.03, \ln OR_2 = 1.02, \ln OR_w = 1.03$$

$$\chi_w^2 = 11.37, \nu = 1, P < 0.001。$$

判断各层之间的 OR_i 是否一致, 以决定是否适用 Mantel - Haensezed 法计算调整 OR (aOR), 可作一致性检验 (test of homogeneity) 也称作不均衡检验 (test of heterogeneity), 二者检验结果的意义一样, 只是检验假设不同。

$$\chi_{k-1}^2 = \sum \left[\frac{(\ln OR_i - \ln OR_w)^2}{Var_{(\ln OR_i)}} \right]$$

上式中 $Var_{(\ln OR_i)} = 1/W_i = 1/a_i + 1/b_i + 1/c_i + 1/d_i。$

上例 $\chi^2_{k-1} = (1.03 - 1.03)^2 / (1/6.18) + (1.02 - 1.03)^2 / (1/4.54) \approx 0.0005$
 $v = 2 - 1 = 1, P > 0.5$ 。

故尚无证据证明二层之间的 OR_i 不是来自同一总体。因此,上述 OR_w 或 OR_{MH} 可以认为是 CO 与 MI 之间的经过调整后的联系强度指标。

如果检验结果 $P < 0.05$, 则说明层间不平衡,不宜计算 OR_w 或 OR_{MH}, 此时应该采用标准化死亡比,

以估计其相对危险度。同时还应分析不平衡的原因,注意是否还有其他混杂因素。

参考文献:

[1] 施倡元. 药物流行病学分析性研究方法(II)[J]. 药物流行病学杂志,1996,5(3):170.
 [2] 上海医科大学流行病学教研室编. 流行病学原理与方法(研究生班用)[M]. 1986:3~32.
 [3] 梁万年. 临床医学研究方法[M]. 北京:北京科学技术出版社,2002:47~70. 收稿日期:2003-10-28

III、药物流行病学研究方法——分析性研究(续) 试题

1. OR 的含义是
 A、暴露的发病率与非暴露的发病率之比 B、病例组的暴露比值与对照组的暴露比值之比
 C、暴露例数与非暴露例数之比 D、暴露人群中发病的例数 E、非暴露人群中发病的例数
2. RR 的估计值是
 A、ad/bc B、ab/cd C、ac/bd D、bc/ad E、cd/ab
3. 频率或分层的成组配比的数据资料分析中 χ^2 检验的目的是
 A、检查样本 OR 值是否明显偏离 1.0 B、检查样本 OR 值是否等于 1.0
 C、检查样本 RR 值是否明显偏离 2.0 D、检查样本 RR 值是否等于 2.0
 E、检查样本 RR 值是否等于 0
4. 病例对照研究的类型有(多选)
 A、分层的成组配比 B、个体配比 C、混杂分析 D、t 检验 E、 μ 检验
5. 公式 $\chi^2 = \frac{(|b-c|-1)^2}{b+c}$ 称为
 A、Miettinen 氏公式 B、McNemar 氏公式 C、Mill 氏公式 D、Mantel-Haenszel 公式 E、Hopkins 公式
6. Mantel-Haenszel 方法适用于
 A、一个病例有 1 个以上对照的数据资料 B、一个病例有 2 个以上对照的数据资料
 C、一个病例有 3 个以上对照的数据资料 D、一个病例有 4 个以上对照的数据资料
 E、一个病例有 5 个以上对照的数据资料
7. Mack 等报告在洛杉矶所做的外源性雌激素与子宫内膜癌关系的研究是在
 A、1975 年 B、1976 年 C、1977 年 D、1978 年 E、1979 年
8. 可信区间越窄则说明
 A、可比性越好 B、显著性越高 C、效度越高 D、信度越高 E、以上都不是
9. 在病例对照研究分析中,如果 $RR > 1$, 则表明暴露因子与疾病之间
 A、无联系 B、存在正联系 C、存在负联系 D、不能肯定 E、以上都不是
10. 混杂是暴露因素与疾病之间的一种
 A、异常联系 B、因果性联系 C、线性联系 D、非线性联系 E、非因果性联系
11. 所谓混杂是指
 A、某种对疾病有制约的因素 B、某种与暴露因子有联系的因素
 C、某种既对疾病有制约,又与暴露因子有联系的因素
 D、某种既对疾病无制约,又与暴露因子无联系的因素
 E、某种难以区分的因素
12. 成为混杂因素必须具备的条件是(多选)