

系列讲座

现代实用流行病学方法

第二讲 病例对照研究

章扬熙

通过探索性研究，建立假说；假说是否正确，需进行分析性流行病学研究与实验性流行病学研究来加以验证。病例对照研究 (Case-Control Study) 是常用的分析性流行病学方法。这种研究通常是通过一组患有某种病的人群 (病例) 和一组或几组未患该病的对照人群 (对照) 暴露于某个或某些可能危险因素 (或保护因素) 频度的比较，来考察这些因素是否与该病存在联系以及联系程度。危险因素可以是环境的，也可以是遗传的、心理的、或营养素的缺乏或过多等；保护因素可以是人为的干预措施，也可以是自然的。这种研究常用于探索与检验病因和评价防治措施的效果。从调查的方向性看，它是后向性的，是从果溯因的研究方法。正因为如此，这种研究可以一次同时探索与验证可能造成因果的多种因素，快速地获取较多的信息。显见，病例对照研究具有探索与检验两种功能，前者可归于探索性研究，后者则列入分析性研究。

病例对照研究的类型和对象的选择

一、病例的选择：病例力求严格符合公认的诊断标准。有些病因复杂的病，如癌症，临床表现类似的病例，其病因未必相同。所以，还要考虑病理细胞学分型等因素的一致，使病例在病因上尽可能相似。原则上应优先选用新发病例，因为新发病例回忆较可靠，病因与患病后果较易区分。其次，选用现患病例，但也宜选病期较短者，以减少回忆性偏倚。而死亡病例由于他人追忆暴露情况，偏倚较大，少有应用。

二、对照的选择：设立对照组的目的是提供一个作为比较用的暴露率。如果病例组和对照组对某一可疑因素的总体暴露率相同，则表明该因素与该病的发生无关。被选作对照的条件必须是不患有该病者。果因研究并不仅限于研究某病的发病因素，也可以研究其他果因的联系，比如，为什么有人得了某病能治愈，有人就死亡？对这种预后研究中，则以该病的死

亡者为“病例”，痊愈者为“对照”。应注意对照组与病例组在研究因素以外的主要因素方面的均衡性。

三、病例和对照的来源：以社区为基础的病例对照研究，通过普查或常规登记等方法掌握某一特定时间该社区的全部病例，可将所有病例或其中一随机样本进行研究。对照则从该社区内未患该病的人群中随机抽样产生。以医院为基础的病例对照研究，以某一特定时间一个或多个医院所有确诊病例为对象，进行总体或其中一个随机样本的研究。而对照则从同一医院不患所研究疾病，而患其他与研究疾病没有病因学联系的病人中选取。由于医院对病例及病例对照均有选择，所以容易产生进入率的偏倚，为了防止这种偏倚，可用几种入院率不同的其他疾病的病例作对照，最好既有医院其他病例的对照组，又有社区人群的对照组。如果各组的结论一致，将增加结论的可信性。

四、配比病例对照研究：病例对照研究有两种，一为组间的成组比较，一为配比的比较。配比是选择对照的一种方法，它使对照在一些基本特征 (如年龄、性别、民族、住地等) 与病例相同或基本相同，以使两组人群具有可比性。应注意配比的条件不宜过多，因为条件越多，对照越难找；更不要将与研究因子暴露有较强联系的因子作为配比条件，这样会低估了所研究的效应，造成配比过头。比如，研究产前使用己烯雌酚与其生女阴道癌的联系，若以病人母亲按先兆流产进行配对，就造成配比过头。因为先兆流产是己烯雌酚用药的指征，与暴露联系强，以先兆流产作为对照的配比条件，必会增加对照组己烯雌酚的暴露率，会导致过低估计或错误估计研究效应。当病例与对照对象都容易得到时，可用一个病例配比一个对照 (即1:1配比)，若病例较少时，可用一个病例配比2~4个对照。

病例对照研究的样本含量

病例对照研究的样本含量的计算与以下四个因素有关：①对照组中对研究因素的估计暴露率（ P_1 ）；②估计具有意义的比数比（OR）；③第一类错误发生的概率 α 应控制在什么水平；④第二类错误发生的概率 β 应控制在什么水平。

一、成组（非配比）病例对照研究的样本含量：可用下式计算：

$$N = \frac{(u_{\alpha}\sqrt{2\bar{P}\bar{Q}} + u_{\beta}\sqrt{P_1Q_1 + P_2Q_2})^2}{(P_2 - P_1)^2} \quad (1)$$

式中 α 、 β 分别为第一类错误与第二类错误发生的概率，第一类错误系指病例、对照两组总体暴露率本无差别错判成有差别的概率，第二类错误系指两组总体暴露率客观上本有差别而错判成无差别的概率。 u 为标准正态分布的横坐标， u_{α} 为双侧时，每侧尾面积为 $\alpha/2$ ， u_{α} 为单侧时，单尾面积为 α ， u_{β} 无论在单侧检验还是双侧检验时均为单侧，单尾面积为 β 。 P_1 、 P_2 分别为对照组与病例组研究因素的暴露率。 $Q_1 = 1 - P_1$ ， $Q_2 = 1 - P_2$ ， \bar{P} 为两组暴露率的平均值，即 $\bar{P} = (P_1 + P_2)/2$ ， $\bar{Q} = 1 - \bar{P}$ ， P_2 可依据 P_1 及OR值用下式求得：

$$P_2 = \frac{OR \times P_1}{1 - P_1 + OR \times P_1} \quad (2)$$

〔例1〕为了考察母亲孕期使用雌激素与所生子女患先天性心脏病的联系，拟进行病例对照研究，估计对照组使用雌激素的比率 $P_1 = 0.3$ ，比数比 $OR = 2$ ，设 $\alpha = 0.05$ ， $\beta = 0.10$ ，求样本含量 N 。

已知 $P_1 = 0.3$ ， $OR = 2$ ，代入公式2得

$$P_2 = \frac{2 \times 0.3}{1 - 0.3 + 2 \times 0.3} = 0.4615$$

$$Q_2 = 1 - P_2 = 0.5385$$

$$\text{再求 } \bar{P} = \frac{0.3 + 0.4615}{2} = 0.3808$$

$$\bar{Q} = 1 - 0.3808 = 0.6192$$

查标准正态曲线下的面积表（或 t 值表中自由度为 ∞ 时的 t 值），知双侧时 $u_{0.05} = 1.96$ ，单侧时 $u_{0.10} = 1.282$ ， $Q_1 = 1 - P_1 = 0.7$ ，将有关数据代入公式1，得

$$N = (1.96\sqrt{2 \times 0.3808 \times 0.6192} + 1.282 \times$$

$$\sqrt{0.3 \times 0.7 + 0.4615 \times 0.5385})^2 \div (0.4615 - 0.3)^2 = 188$$

即病例组与对照组各需调查约190人。

二、1:1配比病例对照研究的样本含量：可用下式计算：

$$K = \frac{[0.5u_{\alpha} + u_{\beta}\sqrt{P(1-P)}]^2}{(P-0.5)^2} \quad (3)$$

$$P = \frac{OR}{1 + OR} \quad (4)$$

$$N = \frac{K}{P_1Q_2 + P_2Q_1} \quad (5)$$

式中 K 为1:1配比四格表中 $b+c$ ， N 为样本含量。

〔例2〕若例1用1:1配对病例对照研究，求样本含量。

首先，用公式4求 P 得

$$P = \frac{2}{1 + 2} = 0.6667$$

将有关数据代入公式3求 K 得

$$K = \frac{[0.5 \times 1.96 + 1.282\sqrt{0.6667 \times 0.3333}]^2}{(0.6667 - 0.5)^2} = 90.33$$

再将有关数据代入公式5求 N 得

$$N = \frac{90.33}{0.3 \times 0.5485 + 0.4615 \times 0.7} = 185$$

即需1:1配比185对，每对1个病例1个对照。

三、1:M配比病例对照研究的样本含量：可用下式计算：

$$N' = \frac{N(1+M)}{2M} \quad (6)$$

式中 N 由公式5求得， N' 为病例的样本含量，对照的样本含量为 MN' 。

〔例3〕若例2用1:4配比病例对照研究，求样本含量。

从例2知，若用1:1配比病例对照研究，需病例185例，将其代入公式6得

$$N' = \frac{185 \times (1 + 4)}{2 \times 4} = 116$$

再将 N' 值乘以 M ，得

$$MN' = 4 \times 116 = 464$$

即需1:4配比116组, 每组1个病例, 4个对照, 共116个病例, 464个对照。

公式6也适用于非配比病例对照研究中病例与对照例数不等的场合, 其中N用公式1来求得。应当说明, 以上计算样本含量的诸公式都是基于查出不分层四格表暴露情况的差别经有关公式推导得到的, 所以适用于不作分层分析的场合。若须进行多因素分析, 样本含量则要增大, 具体含量依分析方法而异。若作分层分析, 形成L个子四格表, 理论上例数宜增加L倍。若进行Logistic回归分析, 一般认为样本含量N至少是因素数的10倍, N至少应达100例左右。这个样本含量比分层分析所需的样本含量减少很多, 这也正是为什么多因素分析, 尤其是因素数较多时, 常用Logistic回归分析而少用分层分析的缘由。

调查表与调查的质量控制

病例和对照对研究因素的暴露情况, 主要是根据被调查者的回忆和有关记录而提供。所以, 制定调查表是非常重要的, 应把能估计到的可疑的有关具体因素都反映在调查表中, 不要遗漏。因为病例对照研究的目的是从果溯因, 如果重要的有关因素被遗漏, 则不能弄清导致果的真正原因。为了得到一个好的有的放矢的调查表, 往往需要对所制定的调查表在一个小范围内试用, 了解其可行性和完整性, 根据试用结果进行修改, 最后制定出适用的调查表。提问题的方式, 可分为开放和闭锁式两种。在预查中, 可设些开放式的问题, 比如, 你认为你患某病是什么原因造成的? 有时会提供出宝贵的而你未想到的重要因素, 对完善调查表大有裨益。正式的调查表多用闭锁式询问, 但表中问题的答案要包括一切可能的结果。比如, 调查新生儿的喂养方式, 应包括母乳喂养、人工喂养和混合喂养三种可能结果。对暴露不仅要了解是否存在暴露, 当暴露存在时, 还要调查暴露的强度和持续时间。凡用电子计算机分析的调查, 在调查分析项目的右侧, 要设有方格, 备过录编码用, 将过录编码输入计算机进行统计分析。调查表中所列问题要紧扣研究目的, 易于提取你所需的资料, 又需通俗易懂, 使不同文化程度的被调查者均能理解和应答。

调查现场工作应有专门的班子, 设有若干名现场指导员、调查员、过录员及资料检查员等。对各种人员均应进行培训, 合格后上岗调查。对调查员的培训应采用扮演角色的方法来学习调查技术。在调查过程

中, 每个环节均应进行质量控制, 以防止偏误的发生。调查员应每日自查每份调查表有无漏问、记录错误并及时更正, 资料检查员要及时复检每份问卷, 发现问题也要及时更正。对调查员的调查实况录音, 抽查录音情况, 能更好了解调查员的工作质量。现场指导员是质量控制的核心人物。他必须切实掌握调查的进程、调查员的工作态度、能力和调查质量, 并及时发现问题、及时更正。现场指导员还必须抽取1%~5%样本亲自重查, 以评价调查的质量。

病例对照研究中的偏倚和控制方法

当某一研究结果与它的客观真值之间出现了方向性的差值, 这种系统误差叫作偏倚。在病例对照研究中可能出现的偏倚有选择性偏倚、信息偏倚和混杂性偏倚。对偏倚产生的原因必须有清醒的认识, 并采取必要措施来控制偏倚, 保证研究结果的真实性和重复性。

一、选择性偏倚(selection bias): 这种偏倚是由于在选择研究对象时(包括病例和对照)方法不正确或研究对象不合作造成的。

1. 伯克森(Berkson)偏倚: 在病例对照研究中, 如果病例来自医院, 由于患某病的人、无该病而患其他病的对照者及具有某特征的人有不同的入院率, 因而使两种无关的疾病或无关的某特征与某病出现了假联系, 反之亦然, 这种偏倚称为Berkson偏倚。

2. 奈曼(Neyman)偏倚: 是选用幸存者即现患病例作为研究病例。由于幸存者不是病例的总体, 其研究因素的暴露率若比已死去的病例过高或过低, 均可导出错误的结果。

3. 无应答: 在研究中可能遇到研究对象不认真合作, 如果参加应答者的暴露率有别于无应答者, 也会产生偏倚。

4. 检测偏倚(detection bias): 如果暴露状态影响疾病临床诊断的机会时, 可能出现非病因暴露与疾病的假联系, 产生检测偏倚。比如, 在某病例对照研究中, 发现子宫内膜癌患者发病前使用雌激素者比对照组高9倍, 推断雌激素可导致子宫内膜癌。但以后许多研究对此推断予以否定。这个错误推断是由于服用雌激素后阴道出血导致就诊机会增多而使无症状的子宫内膜癌的检测率提高造成的假象。

二、信息偏倚(information bias): 这种偏倚产生于对暴露和疾病状态的错误指定。有非区别性

错分 (non-differential misclassification) 和区别性错分 (differential misclassification) 两种类型的信息偏倚。前者是对一种状态 (如暴露状态) 的分类错误不依赖于其他状态 (如疾病状态) 的水平。比如, 在职业病研究中, 常不能对每个工人定量或定性地检测其某种暴露情况, 而用工种来代替暴露状态, 而这种代替可能不能代表真实的暴露情况而造成错分。又比如, 无论病例还是对照都不愿意对性生活、使用麻醉品等敏感问题如实回答, 结果得到的暴露情况是假的。后者是对一种状态 (如暴露状态) 的分类错误依赖于其他状态 (如疾病状态) 的水平。比如, 对暴露史的回忆受疾病状态的影响。病例常对可能是病因的任何事件记忆犹新, 而对照则不然, 结果造成回忆偏倚。又比如, 即使获得了正确暴露史, 在病例对照研究中有可能疾病的发生早于暴露, 这种原发性偏倚在疾病早期导致病例暴露类型的改变时可能发生。比如, 在某研究近视眼发病因素的病例对照研究中, 发现越作眼保健操的人越易患近视眼。其实, 这是假象, 因为患了近视眼才想起作眼保健操, 近视发生有一个过程, 何前何后很难记准, 实际它是果不是因。调查偏倚也是造成区别性错分的原因。比如, 调查员知道暴露-疾病的假设, 进行有意、无意的诱导提示而导致区别性错分偏倚。病例和对照知道暴露-疾病的假设, 为了迎合调查者的心理也可导致讲假话。病例和对照进行调查的环境、程序、方式不同也可导致错误的区分。

三、混杂偏倚 (confounding bias): 混杂是两种因素作用混在一起未能分开的情况, 它歪曲了暴露对疾病的影响, 这种歪曲是由于其他因素是疾病的有关因素并和研究因素的暴露又有联系所引起的, 这些其他因素称为混杂因素, 它们在暴露者和非暴露者

间分布不相等。比如, 在研究吸烟与肺癌的联系中, 性别是混杂因素。因为性别与吸烟有联系, 性别又在吸烟者与不吸烟者的分布不相同。如果不注意性别, 则性别因素会混杂或歪曲吸烟对肺癌的作用。

四、偏倚的控制:

1. 选择性偏倚的控制: 在以医院为基础的病例对照研究中, 选择多种疾病的、入院率不同的病人做对照则可减轻 Berkson 偏倚。若再从一般人群中选择一个随机样本为对照, 不仅可减少选择偏倚, 还可与从医院所选的对照组比较, 鉴定出该研究是否存在选择性偏倚。在设计时应尽量利用新发病例以减少奈曼偏倚。在以社区为基础的病例研究中, 以新发病例总体或从总体中随机抽取样本为病例, 在产生病例的社区人群中随机抽取样本为对照, 使上述两种偏倚可减少到最低程度。对无应答的原因要作分析, 设法补查, 或分析无应答的影响。

2. 信息偏倚的控制: 控制信息偏倚的主要方法是提高调查技巧和进行质量控制。调查者应对病例和对照同等对待, 两组的调查方法、程序要一样。采用盲法调查, 不应让调查员和研究对象知道所研究的某种暴露与疾病的联系。尽可能将研究对象回答的内容与其它来源的资料进行比较或采用两次调查是否一致来检验所得的资料是否真实、可靠。认真培训责任心强、科学态度好、社会工作能力强并有一定文化修养的调查员, 在调查时交叉分配任务来进行质量控制。每个调查员每天既调查病例又调查对照, 对调查情况进行录音, 然后抽查质量, 不合格者重新调查。

3. 混杂偏倚的控制: 在设计阶段可通过选择对象的限制法和匹配法来控制混杂作用。在分析阶段可应用标化率、分层分析和多因素 Logistic 回归模型分析把研究因素的作用与混杂因素的作用分离开来。

更 正

本刊1993年第14卷第6期登刊的“中华流行病学杂志1993年第14卷关键词索引和著者索引 (Author Index)”表内索引的第6期页码有误, 凡“索引”内第6期杂志的页码均应向后顺延2个页码, 如“Chen Xian-hui 陈贤辉 (6): 364”应改为“Chen Xian-hui 陈贤辉 (6): 366”, 特此更正。

中华流行病学杂志编辑部