·基础理论与方法·

基于虚拟事实理论的病因效应模型

陶秋山 李立明

在流行病学病因推断中关键的因素有两点:一是混杂因素的识别与控制,二是交互效应的定性与定量评价。而要研究混杂与交互因素对病因效应的影响,首先要了解无交互、无混杂时病因效应的实际情况,以此为基础才能对混杂和交互效应进行客观的评价。下面将以虚拟事实理论为基础,建立一个病因效应的理论模型,为进一步探讨混杂和交互效应奠定理论基础。

1. 因果关系中的虚拟事实理论:近年来,在因果理论中对因果关系的定义大多基于虚拟事实理论[1,2],这种哲学思想的主要内容可简单表述为:在观察到"事件 c 发生的条件下发生了事件 e"这一事实时,由于这一过程的不可逆性,在实际中无法观察到"事件 c 没有发生时事件 e 的情况"。因此,所谓的"事件 c 是事件 e 的原因"是基于以下的虚拟事实条件:"如果事件 c 不发生,则事件 e 也不会发生。"

现有的流行病学病因推断的哲学思想和推理逻辑基本符合上述虚拟事实理论,以吸烟与肺癌的队列研究为例,其推理逻辑可以概括为:当观察到吸烟组在某一时期内的肺癌发病率为 P_1 这一事实时,这组人群当初如果不吸烟(其他条件不变)的肺癌发病率 P_0 就不可能观察得到,而在理论上这个 P_0 是存在的,假定在虚拟的情况下 P_0 已知,且 $P_1 > P_0$,则称吸烟是肺癌的病因。这里吸烟相当于上述虚拟事实理论中的事件 c,而 $P_1 > P_0$ 则相当于上述事件 e。同理,当研究保护性因素时 $P_0 > P_1$。

由于上述 P_0 总是不可观察的,所以病因推断问题的本身就转化为如何对 P_0 进行有效的估计。在流行病学研究中,通常用对照组的发病率来近似代替 P_0,显然,这种方法同时也引入了可比性问题。

2. 基本概念与符号:下面将对队列研究中的虚拟事实模型进行建模分析,为了便于讨论,首先给出建模中需要用到的一些重要概念和符号。

(1)基线效应:设当某一目标人群不暴露于某一危险因素(E)时,某种疾病(D)的发病率称为基线效应(baseline effects),记为:$P(B)$,$P(B) \in [0, 1]$。这种病因效应可以看作是由一种或多种潜在的致病因素引起的。

(2)暴露效应:仅由暴露因素产生的病因效应称为暴露效应,记为:$P(E)$,$P(E) \in [0,1]$。它相当于在理想人群中测得的绝对病因效应;而所谓理想人群是指:如果不暴露就不会发病,并且不存在与暴露因素有交互作用的因素的人群。

(3)表观效应:由潜在的致病因素与暴露因素共同产生的效应称为表观效应,简记为:$P(B,E)$,并且其大小取决于基线效应和暴露效应的一个组合函数:

$$P(B,E) = f\{P(B), P(E)\} \quad (1)$$

(4)无交互效应模型:当基线效应与暴露效应无交互作用时,上述组合函数可写为:

$$P(B,E) = P(B) + P(E) - P(B)\cdot P(E) \quad (2)$$

(5)交互效应:如果暴露因素与潜在病因之间存在交互作用,表观效应不符合式(2),记为 $P(B, E)^*$。它又分为两种情况:当 $P(B,E)^* > P(B) + P(E) - P(B)\cdot P(E)$ 时,称正交互效应;当 $P(B,E)^* < P(B) + P(E) - P(B)\cdot P(E)$ 时,称负交互效应。

(6)相对危险度:表观效应与基线效应的比称为相对危险度(RR):

$$\text{非交互效应模型的相对危险度}(RR) = \frac{P(B,E)}{P(B)} \quad (3)$$

$$\text{交互效应模型的相对危险度}(RR^*) = \frac{P(B,E)^*}{P(B)} \quad (4)$$

(7)相对交互效应:交互效应模型的 RR^* 与非交互效应模型的 RR 之比称为交互比(interactive ratio, IR):

$$IR = \frac{RR^*}{RR} \quad (5)$$

3. 材料与方法:用统计分析软件 SAS8.01(SAS Institute Inc.)产生模拟数据集,其中的变量 x, y, z 分别代表上述的 $P(B)$, $P(E)$, RR, $P(B)$ 与 $P(E)$ 的取值范围为:[0.01, 1]。SAS 源程序如下:

```
Data DA;
DO x = 0.01 to 1 by 0.01;
  DO y = 0.01 to 1 by 0.01;
    z = (x + y - x * y)/x;
    output;
  END;
END;
Run;
```

用统计分析软件 S-Plus 2000 Professional(MathSoft Inc.)对上述的模拟数据集中的三个变量之间的关系进行 3D 绘图分析(图 1)。

4. 结果:

(1)无交互模型中基线效应、暴露效应与 RR 的关系:由

基金项目:国家自然科学基金资助项目(39930160)

作者单位:100083 北京大学公共卫生学院流行病学教研室(陶秋山);中国预防医学科学院(李立明)

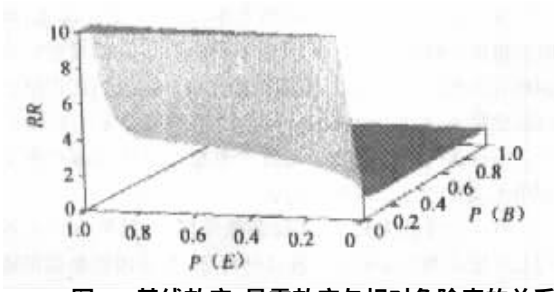

图 1 基线效应、暴露效应与相对危险度的关系（无交互效应）

式(2)与(4)得：

$$RR = \frac{P(B,E)}{P(B)} = 1 + (\frac{1}{P(B)} - 1) \cdot P(E) \quad (6)$$

可见当基线效应固定时，*RR* 与暴露效应之间呈正比；当暴露效应固定时，*RR* 与基线效应之间呈反比。由式(6)得：

当 $P(E) = P(B)$ 时，$RR = 2 - P(E)$，$RR \in [1,2]$；

当 $P(E) > P(B)$ 时，$RR > 2 - P(E)$，$RR \in (1,\infty)$；

当 $P(E) < P(B)$ 时，$RR < 2 - P(E)$，$RR \in (1,2)$；

上述变化趋势可以从 S-Plus 模拟结果中得到直观的反映(图 1)。

(2)交互效应对 *RR* 值的影响：由于正交互效应使表观效应增加，因此其 *RR* 值相应增加；同样，负交互效应使表观效应减小，其 *RR* 值也相应减小。

①当交互效应为正时，设交互效应为 $P(I)$，$P(I) \in (0,1)$。则：

$$P(B,E)^* = P(B,E,I) = P(B,E) + P(I) - P(B,E) \cdot P(I) > P(B,E) \quad (7)$$

②当交互效应为负时，交互效应为保护性因素，记为 $P(I)^*$，$P(I) \in (0,1)$，则：

$$P(B,E)^* = P(B,E) \cdot [1 - P(I)^*] < P(B,E) \quad (8)$$

(3) *IR* 与 *RR* 的关系：由式(3)、(4)、(5)可得：

$$IR = \frac{RR^*}{RR} = \frac{P(B,E)^*}{P(B)} \times \frac{P(B)}{P(B,E)} = \frac{P(B,E)^*}{P(B,E)} \quad (9)$$

即从本质上 *IR* 相当于交互模型与非交互模型之间相比，即交互因素的 *RR*。因此，虽然 *IR* 与 *RR* 的病因学意义不同，但它们具有相同的统计特征。

5. 应用举例：假设有一项关于基因与环境的交互作用的研究得到了表 1 中主要研究结果。

表 1 基因与环境的交互作用的模拟数据

组别		观察总人数	观察期内发病数	发病率(%)
环境暴露组(E)	E_1 组：目的基因(+)	1 000	200	20
	E_2 组：目的基因(−)	2 000	300	15
环境非暴露组(N)	N_1 组：目的基因(+)	1 000	150	15
	N_2 组：目的基因(−)	2 000	200	10

表 1 数据分析：由于缺乏相应的虚拟条件，显然此研究无法直接应用上述的虚拟事实模型直接求解。此问题虽然无法求得真值，但在给定的前提条件下是可估的。

前提假设①：假定此研究设计良好，并且根据专业知识可基本排除混杂因素的干扰，则就可以用 E_2 组的信息来估计 E_1 组的虚拟事实信息，用 N_2 组的信息来估计 N_1 组的虚拟事实信息。在上述前提假设下环境暴露组和非暴露组的效应分别可估，此前提假设与现有的流行病学队列研究的设计思想完全一致。

前提假设②：为了估计交互效应，就必须有无交互效应的前提，上述研究没有给出这个前提，在没有更多信息的情况下，为了简化计算，可以假定研究在不存在除研究因素与目的基因外的其他交互效应，或虽然存在也可忽略不计。

基于以上前提假设，就可以用虚拟事实模型对上述问题进行以下求解：

(1)在环境非暴露组(N)中，根据假设②知其无交互效应，则：

设：基线效应：$P(B) = 0.10$，表观效应：$P(B,E) = 0.15$

解：$RR = P(B,E)/P(B) = 0.15 \div 0.10 = 1.5$

$$P(E) = [P(B,E) - P(B)]/[1 - P(B)] = (0.15 - 0.10) \div (1 - 0.10) = 0.06$$

(2)在环境暴露组(E)中，根据虚拟事实模型其暴露效应与上述计算结果相同，则：

设：基线效应：$P(B) = 0.15$，暴露效应：$P(E) = 0.06$，表观效应：$P(B,E)^* = 0.20$

解：若环境因素与目的基因之间无交互效应，则

$$P(B,E) = P(B) + P(E) - P(B) \cdot P(E) = 0.15 + 0.06 - 0.15 \times 0.06 = 0.20 = P(B,E)^*$$

$$RR = P(B,E)/P(B) = 0.20 \div 0.15 = 1.33$$

由于 $P(B,E)^* = P(B,E)$，因此可以认为环境因素与目的基因之间没有交互作用，上述环境非暴露组(N)与环境暴露组(E)之间相对危险的差异(1.5 − 1.33 = 0.17)主要是由于基线效应的不同而造成的。

6. 讨论：上述虚拟事实模型中各种概念的理论意义在定义中已经明确，它们的流行病学意义需要进一步阐明。首先，按照流行病学队列研究的设计思想，它是用对照组来代替暴露组的虚拟事实情况，即：假如暴露组当初也不暴露于研究因素，在可比情况下它与对照组的发病情况应该相同。因此，对照组的发病率就相当于基线效应，暴露组的发病率相当于表观效应，而暴露效应和交互效应是无法直接测量的理论值，是病因学研究的核心内容。

长期以来，*RR* 被认为是暴露效应大小的重要指标之一，其优点是直观、计算简便，但是 *RR* 计算依据的并不是真实的暴露效应 $P(E)$，而是表观的暴露效应 $P(B,E)$。如前所述，表观效应是暴露效应、基线效应和潜在的混杂和交互效应的综合表现。因此，相对危险度是一个比较粗糙的病因效应指标。在以往的许多研究中，研究资料有以下几个特点：①暴露效应较大，基线效应相对较小；②混杂因素基本上能

够控制;③无明显的交互效应。此时,暴露组的表观效应与暴露的理论效应相差不大,因此长期以来流行病学研究并没有严格区分这两个概念。

在慢性病的病因学研究中,研究资料往往具有同上述内容相反的几个特点:①暴露效应相对较小,基线效应相对较大;②混杂因素不易控制;③交互效应不易识别。以上这些特点就构成了弱效应病因推断中的主要障碍,这也是当前流行病学研究面临的一个比较难以突破的极限[3]。在此情况下,只有深入研究基线效应、交互效应和混杂效应等因素与暴露效应的内在关系,才能够对弱效应病因进行科学的定性定量评价。

交互效应的评定一直是流行病学研究中的热点问题,但目前尚缺乏对交互效应的共识。问题的根本原因在于对非交互效应没有统一的认识,只有清楚了什么是无交互,才有可能对交互进行研究和评价。上述虚拟事实模型中对无交互效应的定义是基于下述的流行病学实际意义的:假如研究吸烟与肺癌的关系,分别进行两次调查:第一次调查在理想人群中进行,由于没有基线效应和交互效应的干扰,因此可以测得吸烟的暴露效应,假设为 10%。然后进行第二次调查,假设调查1 000人,其中有 100 人即使不吸烟也会患肺癌,即基线效应为 10%;假设不存在交互作用,则剩下的 900 人中会有 90 人(10%)因吸烟而发病,因此在整个暴露人群中总的发病率为 19%(190/1 000),可以证明这恰好符合式(2)。

另外,在流行病学研究中有两个重要的概念,即:归因危险度(AR),其流行病学意义是指危险特异地归因于暴露因素的程度[4]。按照归因危险的定义和式(2),则:

$$AR = I_e - i_0 = P(B,E) - P(B)$$
$$= [1 - P(B)] \cdot P(E) \qquad (10)$$

在上述的吸烟与肺癌的例子中,证明了上式成立,即说明虚拟事实模型与现有病因理论有兼容性,它并没有完全推翻现有的病因理论,而是使其更加完善。式(10)还说明由于归因危险还包括基线效应,因此它还是一个较为粗糙的指标。不过这一指标所包含的哲学思想恰好证明虚拟事实模型中无交互效应公式可以成立。

有了上述的无交互、无混杂效应模型,就可以进一步探讨交互效应和混杂效应的统计学特征,为病因推断提供理论依据。需要说明的是,现实中的流行病学资料中混杂效应和交互效应的机制可能十分复杂,并且还存在其他各种各样的偏倚情况,因此,在应用上述虚拟事实模型进行分析和做出结论时要十分慎重。同样,虚拟模型也有待进一步完善和发展。

流行病学与数理统计的最大区别在于它能够使用更多的先验或后验知识进行推理分析,并不完全依赖于统计数据,这也是流行病学在病因学研究中的灵魂所在。例如在上述举例分析中的数据信息对虚拟事实模型来说是不完备的,但根据一定的先验知识给定相应的前提条件后,从一定程度上解决了病因推断的问题。虽然这种方法得出的结论与真实情况有一定的差距,但它是一个科学的结论,并且随着先验知识的完备性不断增加,这种方法得出的结论就会不断接近真值,最终达到真理。

参 考 文 献

1 Roese NJ. Counterfactual thinking. Psychol Bull, 1997, 121: 133-148.
2 Sledge W. Counterfactual thoughts about counterfactual thinking. Psychiatry, 2000, 63: 336-338.
3 Taubes G. Epidemiology faces its limits. Science, 1995, 269: 164-169.
4 王天根. 队列研究. 见:连志浩,主编. 流行病学. 第 3 版. 北京:人民卫生出版社, 1994. 105.

(收稿日期:2001-08-29)
(本文编辑:张林东)

·出版信息·

《循证医学与临床实践》现已出版

《循证医学与临床实践》是由复旦大学中山医院王吉耀教授主编的一本关于循证医学和临床实践的专著。该书共分四篇。第一篇着重介绍循证医学的概念和方法、实践的原则和步骤。第二篇介绍得到证据的具体方法,包括文献检索的方法,Meta-分析的步骤和统计方法,系统综述示例等。第三篇则从诊断、治疗、预防、预后、不良反应、经济分析、生命质量、决策分析、卫生技术和医疗质量评估、教学等方面介绍在临床实践中如何开展循证医学。第四篇举实例指导各专科医师如何应用循证医学的原理和方法解决临床问题。

本书主编王吉耀教授早年从师于国际循证医学的创始人和奠基人——加拿大 McMaster 大学 Sackett 教授。本书其他作者绝大多数经过国内或国际临床流行病学培训,部分为丹麦专家。他们当中绝大部分从事临床医疗工作,并将循证医学理论和临床实践融合在一起,陈述循证医学的实施,方法具体,内容丰富,实用性强,可作为广大临床医师和临床流行病学工作者学习和开展循证医学的参考教材。本书已于 2001 年 11 月由科学出版社出版,每本定价 38.00 元,各新华书店均有现书,欢迎订购。欲购者可直接与科学出版社黄敏编审联系(电话:010-64033542)。