

·基础理论与方法·

“探针变量”:观察性研究中混杂因素识别的工具

洪翔 尹洁晨 王蓓

东南大学公共卫生学院流行病与卫生统计学系,南京 210009

通信作者:王蓓, Email:wangbeilxb@seu.edu.cn

【摘要】 观察性研究中往往存在未知或未测量的混杂因素,是流行病学因果关联研究中的重大挑战。本文介绍一种可以应用在观察性研究中的一种对未知/未测量混杂因素进行识别和效应评估的工具——“探针变量”。其主要可以分为暴露探针变量、结局探针变量以及中介探针 3 种形式,前 2 种不仅可以对未知/未测量混杂因素进行识别,也可以对其效应量进行估计,从而揭示真实的暴露与结局之间的关联。而中介探针则是针对“中介因子”进行控制,从而识别暴露和结局之间是否存在未测量混杂因素。该理论实践过程中最大的困难在于“探针变量”的选择和确定,不恰当的“探针变量”可能引入新的混杂,导致未测量混杂因素识别不准确。“探针变量”可以推荐作为观察性研究报告中的一项敏感性分析内容,有助于读者真实理解暴露与结局之间的关联,增加观察性流行病学研究中的证据力度。

【关键词】 观察性研究; 探针变量; 混杂因素

基金项目: 国家自然科学基金(8187121102);东南大学优秀博士培育基金(YBPY1983)

Probe variables: a tool for identification of unmeasured confounders in an observational study

Hong Xiang, Yin Jiechen, Wang Bei

Department of Epidemiology and Health Statistics, School of Public Health, Southeast University, Nanjing, 210009, China

Corresponding author: Wang Bei, Email: wangbeilxb@seu.edu.cn

【Abstract】 There are usually unknown or unmeasured confounders in the observational study, which is a significant challenge in epidemiological causal association research. This paper presents a tool for identification and effect assessment of unknown/unmeasured confounders in observational studies: probe variables. It can be divided into three forms: exposure probe variable, outcome probe variable, and mediation probe variable. The first two types can identify unknown/unmeasured confounding factors and estimate their size of effect to reveal the real correlation between exposure and outcome. The mediation probe variable controls for "mediating factors" to identify unmeasured confounders between exposure and results. The most significant difficulty in this theory's practice is selecting and determining "probe variables." Improper probe variables may introduce unknown confounders, which may lead to false identification of unmeasured confounders. Probe variables can be recommended as a sensitivity analysis in observational studies to help readers truly understand the association between exposure and outcomes and to increase the strength of evidence in observational epidemiological studies.

【Key words】 Observational study; Probe variables; Confounders

Fund programs: National Natural Science Foundation of China (8187121102); Scientific Research Foundation of Graduate School of Southeast University (YBPY1983)

DOI: 10.3760/cma.j.cn112338-20200315-00355

收稿日期 2020-03-15 本文编辑 万玉立

引用本文:洪翔,尹洁晨,王蓓.“探针变量”:观察性研究中混杂因素识别的工具[J].中华流行病学杂志,2021,42(4):735-739. DOI: 10.3760/cma.j.cn112338-20200315-00355.



流行病学研究手段主要包括观察性研究和实验性研究。虽然实验性研究能够提供最强级别的证据,是探索两个因素间因果关联的最佳手段,但在实际临床研究中往往存在一定的困难^[1]。而以队列研究、病例对照研究及横断面研究为主要手段的观察性研究,仍然是目前医学研究中最广泛使用的研究方法。近年来“真实世界研究”的兴起成为观察性研究的有力手段,其已被美国食品药品监督管理局认可,作为药物安全性评价的重要方法^[2]。相比于随机对照试验,观察性研究在探索暴露和结局之间因果关联时,往往受限于混杂因素识别和控制的困难,从而造成因果推断证据级别较低^[3]。混杂因素指的是一类与研究因素和研究结局均存在关联的变量,其不是暴露与结局之间的中介变量,但往往可以歪曲(掩盖或夸大)两者之间真正的关联^[3]。目前最常用的混杂因素控制方法包括限定研究对象、分层、多元回归(建模)、匹配、反概率加权等^[4]。但这些手段是在混杂因素“有效测量”的基础上开展的。而在现实研究中,研究者可能并不能全面识别并测量所有的混杂因素,而一些混杂因素也可能存在“测量偏倚”,这些未知/未测量混杂因素将使流行病学研究出现“虚假关联”。而近年来广泛研究的“工具变量”可以对未知混杂因素进行控制^[5-6]。如何识别并消除关联研究中未知/未测量混杂因素的作用,一直是流行病学关注的热点话题。本综述将介绍一种可以用在观察性研究中进行未知/未测量混杂因素识别的工具——“探针变量”(probe variables)的原理及应用。

探针变量被 Lipsitch 等^[4]总结为观察性流行病学研究中“阴性对照”(negative control)。但由于其译文与实验性研究中的“空白对照”相似,概念上容易混淆,故本综述将其统一称为“探针变量”予以阐述。常见的探针变量可以分为暴露探针变量(probe exposure, PE)、结局探针变量(probe outcome, PO)以及中介探针。

1. PE: 当我们试图研究暴露因素(exposure, E)与结局(outcome, O)之间的关联时,需要尽可能地识别并测量二者之间的混杂因素(confounders, C),并进行混杂因素的控制。然而,一些未知/未测量的混杂因素(unmeasured confounders, U)可能导致暴露-结局(E→O)之间的关联“扭曲”,呈现虚假关联。见图1。而探针变量的主要目的是验证E→O之间是否存在未知/未测量的混杂因素U,并探索U的效应有多大,以及寻找控制U的方法。

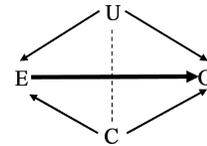


图1 一般“暴露-结局”关联模式

PE 又称阴性对照暴露(negative control exposure),旨在暴露-结局之间寻找一个新的已被测量的暴露变量,其满足以下2个假设(图2):从常理上或现有的研究证据上看,PE与结局之间不存在因果关联;从PE到结局的关联中,可能存在的混杂因素集(C+U)与所关注的暴露到结局之间存在的混杂因素集(C+U)重叠。

如果PE满足上述两种假设,则认为PE与暴露变量之间存在“U形可比”(U-comparable)关系。从模式图中可以得知,如果未测量的混杂因素U不存在,那么在控制已知混杂因素C之后,PE与结局之间将观测不到统计学关联。反之,如果在控制了混杂因素C之后,PE与结局之间仍然存在关联(PE-U-O),则证明未测量的混杂因素U存在,且混杂效应大小即为PE-U-O。

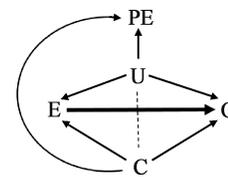


图2 “暴露探针变量”模式

Liew 等^[7]曾利用该原理证实孕期女性服用对乙酰氨基酚与后代儿童注意缺陷/多动障碍(ADHD)的关联。其通过队列数据初步分析发现,孕期对乙酰氨基酚暴露将使后代患 ADHD 的风险增加 1.34 倍(95%CI: 1.05~1.70)。为探索其中是否存在未知/未测量混杂因素,作者将怀孕前(4 年之前)及产后(4 年之后)服用对乙酰氨基酚的记录作为 PE,该 PE 在原理上与儿童 ADHD 患病风险无关。在调整了已测量的混杂因素之后,该 PE 与结局确实没有统计学关联,从而提示原假设中孕期对乙酰氨基酚暴露与后代患 ADHD 的风险的关联不存在未测量混杂因素,关联是可信的。

2. PO: 与 PE 类似,PO 又称阴性对照结局(negative control exposure outcome),旨在暴露-结局之间寻找一个新的已被测量的结局变量,其同样需要满足 2 个假设(图 3):从常理上或现有的研究证据上看,暴露与 PO 之间不存在因果关联;从暴露到 PO 之间的关联中,可能存在的混杂因素集(C+U)

与所关注的暴露与结局之间存在的混杂因素集 (C+U) 重叠。

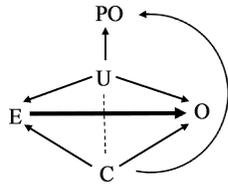


图3 “结局探针变量”模式

如果 PO 满足上述 2 个假设,则认为 PO 与结局变量之间存在“U 形可比”关系。从模式图中可以得知,如果未测量的混杂因素 U 不存在,那么在控制已知混杂因素 C 之后,暴露与 PO 之间将观测不到统计学关联。反之,如果在控制了混杂因素 C 之后,暴露与 PO 之间仍然存在关联 (E-U-PO),则证明未知混杂因素 U 存在,且混杂效应大小即为 E-U-PO。

Jackson 等^[8]用 PO 探索了老年人接种流感疫苗对肺炎住院的保护作用。研究初步分析提示,接种流感疫苗的老年人较之不接种者,肺炎住院风险明显降低。但是从常理分析,老年人对疫苗的免疫反应较弱,其在观察性研究中的“获益”远超过预期“获益”。为探索其中是否存在未测量的混杂因素 U,研究者提出 PO——因受伤而住院,该变量被认为与疫苗接种不存在因果关联。进一步分析后发现,接种疫苗的老年人发生“受伤住院”的风险也显著降低。该结果证实了 U 的存在,可能的解释是愿意主动接种疫苗的老年人身体状况及卫生习惯较好,故发生各类疾病的风险均更低,从而放大了疫苗接种与“肺炎住院”的负关联。上述可能的 U 变量均为难以准确测量的变量,而 PO 的引入使得原本难以测得的 U 效应得以估计。

3. 中介探针:是与上述两种“探针变量”不同的思路。在探索暴露-结局关联过程中,如果能够找到一个完全中介变量 M,即暴露要影响结局必须通过该中介变量,那么一旦中介变量消失,暴露与结局之间的关联将不存在。但如果存在未测量的混杂因素 U,那么在调整 C 的前提下,即使切断了中介变量,暴露-结局之间仍存在关联 (E-U-O)。见图 4。

遗憾的是,在观察性研究中,几乎很难找到完全中介变量。其中的典型例子还是 Jackson 等^[8]对老年人接种疫苗的效果研究。研究者提出假说,流感疫苗要发生保护作用,一定是减少流感的发生,

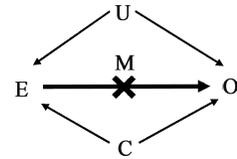


图4 “中介探针”模式

从而降低老年人因肺炎住院率。“确诊流感”即为流感疫苗和“因肺炎住院”的一个完全中介。注意到,因肺炎住院的人有可能是其他病毒、细菌等感染,并不完全是因为流感,而住院患者并不会全部进行病原体检测。所以没法将“确诊流感”作为中介变量。于是研究者采用“流感季节”作为“流感确诊”的替代。即认为:流感疫苗的保护作用应该在非流感季节中不存在,否则认为存在未控制的混杂因素。这种做法与“分层分析”类似,但是,此处的中介变量与协变量具有本质区别,在多元回归分析中,并不能简单地将其作为混杂因素之一进行调整或者控制。

4. “探针变量”使用的注意事项:在使用探针变量进行观察性研究未知/未测量混杂因素识别的过程中,应该时刻注意探针变量的前提假设。如果未能对探针变量进行很好的设定,将可能引入新的混杂因素,从而造成新的虚假关联。在探针变量的使用过程中,应注意以下几个方面:

(1) PE 设定应该遵循与主要关注暴露因素的“U 形可比”原则。如果 PE 与结局 O 之间存在除了“PE-U-O”“PE-C-O”以外的“PE-U2-O”的关联,那么 PE 和暴露将不满足“U 形可比”,即使暴露-结局中需要探讨的未知/未测量混杂因素 U 不存在,PE 与结局之间也会呈现关联,从而得出原关联中仍存在未控制的混杂因素的错误结论。见图 5。

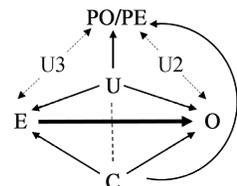


图5 不满足“U 形可比”的暴露/结局探针变量模式

(2)同理,如果 PO 与关注的暴露因素之间存在额外的混杂因素 U3,那么 PO 与结局变量之间不再“U 形可比”,会导致即使未知混杂因素 U 不存在时,暴露与 PO 之间也会呈现关联,从而做出认为原关联中仍存在未控制的混杂因素的错误结论。见图 5。在实际工作中,U2 和 U3 的识别是具有难度的,我们只能根据实际临床经验或既往流行病学资

料进行假设。

(3) 鉴于 U2 和 U3 的识别难度,“工具变量”的使用可以弥补这一缺陷。见图 6。工具变量仅与暴露存在因果关联,而与其他因素(包括结局、协变量等)都不相关。这样可以用工具变量与结局的关联来间接反映暴露与结局的关联,而不受任何可能的混杂因素的影响。典型的中介变量为基于孟德尔随机化的单核苷酸多态性^[5]。由于遗传信息是先于外界暴露而存在的,所以可以认为遗传因素是完美的工具变量。基于此,近年来流行病学家做了大量的因果探索研究^[9]。但是“工具变量”与“探针变量”是具有本质区别的,在实际应用过程中应该注意甄别。

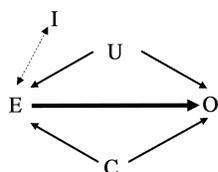


图 6 “工具变量”模式

(4) 在临床研究中,应该根据实际情况选择 PE 还是 PO 作为探针。一般来说,队列研究中会同时收集比较多的暴露和结局因素,选择 PE 和 PO 都是可行的。而在病例对照研究中,结局往往被确定,而可能会收集较多的暴露因素,故选择用 PE 作为探针变量较为合适。但“巢式病例对照”研究则可以弥补这个缺陷,其选用两种探针变量都是可行的。

(5) 除了对未知/未测量混杂因素的识别以外,暴露/PO 的设定还可以对观察性研究中的其他类型的偏倚进行控制,包括回忆偏倚、测量偏倚等^[10-11]。Zaadstra 等^[12]曾利用 PE 证实儿童期传染性单核细胞增多症与未来多发性硬化的风险存在关联。该研究通过对多发性硬化及对照组人群进行回顾性调查,了解其儿童时期曾患的疾病。在调整了年龄、性别、文化程度、居住地等协变量以后,儿童期传染性单核细胞增多症与多发性硬化之间存在关联 ($OR=2.22, 95\%CI: 1.73\sim 2.86$)。同时,其他感染性疾病,如儿童时期百日咳与多发性硬化也存在关联 ($OR=1.22, 95\%CI: 1.02\sim 1.47$)。由于回顾性调查过程中可能存在回忆偏倚,病例人群可能由于经常前往医院,对儿童时期的疾病史报告率比较高,导致儿童期各种疾病与多发性硬化存在虚假关联。为验证这种猜想,研究者设定了 PE——儿童期脑震荡,该变量被认为与多发性硬化的发生无

关。然而数据分析发现脑震荡的报告率与多发性硬化之间存在统计学关联 ($OR=1.23, 95\%CI: 1.07\sim 1.41$)。说明回忆偏倚确实存在。而百日咳与多发性硬化的关联效应与 U 形效应相似,可能是“虚假关联”。但单核细胞增多症与多发性硬化的关联强度明显大于 U 形效应,且作者进一步将 PE 作为调整因素在原模型中加以控制,从而确认儿童期感染性单核细胞增多症将增加多发性硬化的风险。

(6) 在观察性研究中的“中介探针”的选择难度较大。流行病学中“中介效应分析”(mediating effect analysis) 往往关注的是“部分中介”。即在暴露与结局之间,除了通过中介变量的“中介效应”以外,仍存在暴露与结局之间的“直接效应”。这种“直接效应”实际上可能是一系列未被识别的机制通路组成^[13]。部分中介变量并不适合用于作为“中介探针”,因为切断部分中介变量,且控制协变量以后,无论未知混杂因素是否存在,暴露和结局之间仍可能存在关联,从而做出认为原关联中仍存在未控制的混杂因素的错误结论。

(7) 所有的这些“探针变量”的目的是识别暴露与结局之间是否存在未知/未测量的混杂因素,而暴露/PO 还可以评估这种混杂因素的效应量。仅部分情况下可以借助探针变量对这未知混杂因素进行控制。但是应该注意这种方法并不能判定未知/未测量的混杂因素具体是什么。

5. 总结和展望:在观察性流行病学研究中,“探针变量”可以作为有效的工具对因果关联分析中冗余混杂因素进行识别和评估。虽然“探针变量”本身在应用过程中可能存在一些局限性,且有引入新的混杂因素的风险,但我们应该正视这个问题,不能否认其在流行病学因果关联研究中的潜在的巨大价值。国内外的一些流行病学家主张,在因果关联的研究报告中,应该在呈现关联效应量的同时,普遍报告可能存在的混杂因素作用效应量^[14],其中就包括设定“探针变量”^[4]、评估混杂因素 E 值^[15]、“混杂函数评估”^[16]等。我国流行病学研究中涉及“探针变量”的研究较少,关注还不够。建议在资料允许的情况下,增加对混杂因素作用的评估,即使“探针变量”分析发现原始因果关联中存在未知混杂因素,且我们无法控制,但该“混杂效应”已被识别,将有助于读者真实理解暴露与结局之间的关联。如果证实不存在未知混杂因素,也可增加观察性流行病学研究中的证据力度。

利益冲突 所有作者均声明不存在利益冲突

参 考 文 献

- [1] 何俏, 时景璞. 临床真实世界研究中的实验性研究设计[J]. 中华流行病学杂志, 2018, 39(4):519-523. DOI:10.3760/cma.j.issn.0254-6450.2018.04.027.
He Q, Shi JP. Realization of design regarding experimental research in the clinical real-world research[J]. Chin J Epidemiol, 2018, 39(4):519-523. DOI:10.3760/cma.j.issn.0254-6450.2018.04.027.
- [2] Corrigan-Curay J, Sacks L, Woodcock J. Real-world evidence and real-world data for evaluating drug safety and effectiveness[J]. JAMA, 2018, 320(9): 867-868. DOI: 10.1001/jama.2018.10136.
- [3] Sedgwick P. Bias in observational study designs: cross sectional studies[J]. BMJ, 2015, 350:h1286. DOI:10.1136/bmj.h1286.
- [4] Lipsitch M, Tchetgen TE, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies[J]. Epidemiology, 2010, 21(3): 383-388. DOI: 10.1097/EDE.0b013e3181d61eeb.
- [5] 王莉娜, Zhang ZF. 孟德尔随机化法在因果推断中的应用[J]. 中华流行病学杂志, 2017, 38(4): 547-552. DOI: 10.3760/cma.j.issn.0254-6450.2017.04.027.
Wang LN, Zhang ZF. Mendelian randomization approach, used for causal inferences[J]. Chin J Epidemiol, 2017, 38(4):547-552. DOI:10.3760/cma.j.issn.0254-6450.2017.04.027.
- [6] 黄丽红, 魏永越, 陈峰. 如何控制观察性疗效比较研究中的混杂因素:(二)未知或未测量混杂因素的统计学分析方法[J]. 中华流行病学杂志, 2019, 40(11): 1450-1455. DOI: 10.3760/cma.j.issn.0254-6450.2019.11.020.
Huang LH, Wei YY, Chen F. Confounder adjustment in observational comparative effectiveness researches: (2) statistical adjustment approaches for unmeasured confounders[J]. Chin J Epidemiol, 2019, 40(11): 1450-1455. DOI:10.3760/cma.j.issn.0254-6450.2019.11.020.
- [7] Liew Z, Kioumourtzoglou MA, Roberts AL, et al. Use of negative control exposure analysis to evaluate confounding: an example of acetaminophen exposure and attention-deficit/hyperactivity disorder in Nurses' health study II [J]. Am J Epidemiol, 2019, 188(4): 768-775. DOI: 10.1093/aje/kwy288.
- [8] Jackson LA, Jackson ML, Nelson JC, et al. Evidence of bias in estimates of influenza vaccine effectiveness in seniors [J]. Int J Epidemiol, 2006, 35(2): 337-344. DOI: 10.1093/ije/dyi274.
- [9] Emdin CA, Khera AV, Kathiresan S. Mendelian randomization[J]. JAMA, 2017, 318(19): 1925-1926. DOI: 10.1001/jama.2017.17219.
- [10] Sanderson E, Macdonald-Wallis C, Smith GD. Negative control exposure studies in the presence of measurement error: implications for attempted effect estimate calibration[J]. Int J Epidemiol, 2018, 47(2):587-596. DOI: 10.1093/ije/dyx213.
- [11] Miao W, Tchetgen ET. Invited commentary: bias attenuation and identification of causal effects with multiple negative controls[J]. Am J Epidemiol, 2017, 185(10):950-953. DOI:10.1093/aje/kwx012.
- [12] Zaadstra BM, Chorus AMJ, van Buuren S, et al. Selective association of multiple sclerosis with infectious mononucleosis[J]. Mult Scler, 2008, 14(3): 307-313. DOI: 10.1177/1352458507084265.
- [13] 郭楚豪, 吴诗蓝, 马淑娟, 等. 一种新的针对多个中介变量的中介分析方法[J]. 中华流行病学杂志, 2019, 40(9):1155-1158. DOI:10.3760/cma.j.issn.0254-6450.2019.09.026.
Guo CH, Wu SL, Ma SJ, et al. A new mediation analysis method for multiple mediators[J]. Chin J Epidemiol, 2019, 40(9): 1155-1158. DOI: 10.3760/cma. j. issn. 0254-6450. 2019.09.026.
- [14] Begum M, Pilkington RM, Chittleborough CR, et al. Effect of maternal smoking during pregnancy on childhood type 1 diabetes: a whole-of-population study[J]. Diabetologia, 2020, 63(6):1162-1173. DOI:10.1007/s00125-020-05111-w.
- [15] VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-Value[J]. Ann Intern Med, 2017, 167(4): 268-274. DOI: 10.7326/M16-2607.
- [16] 王丹华, 尤东方, 黄丽红, 等. 观察性研究中针对未测量混杂干扰的敏感性分析方法[J]. 中华流行病学杂志, 2019, 40(11): 1470-1475. DOI: 10.3760/cma. j. issn. 0254-6450. 2019.11.023.
Wang DH, You DF, Huang LH, et al. Sensitivity analysis method for unmeasured confounding interference in observational study[J]. Chin J Epidemiol, 2019, 40(11): 1470-1475. DOI:10.3760/cma.j.issn.0254-6450.2019.11.023.