·基础理论与方法·

有向无环图在混杂因素识别与控制中的应用及实例分析

刘慧鑫[1] 汪海波[2] 汪宁[3]

[1]北京大学人民医院 100044；[2]北京大学临床研究所 100191；[3]中国疾病预防控制中心性病艾滋病预防控制中心，北京 100026

通信作者：汪宁，Email：wangnbj@163.com

【摘要】 观察性研究是流行病学病因研究常用的研究设计，但应用观察性研究进行因果推断时，常由于未经识别、校正的混杂因素的存在，歪曲暴露因素与研究结局之间的真实因果关系。传统混杂因素判断标准在实际应用中不够直观，且有一定局限性，有时甚至出现混杂因素的误判。有向无环图（DAGs）可以直观识别观察性研究中存在的混杂因素，将复杂的因果关系可视化，判断研究中需要校正的最小校正子集，并可避免传统混杂因素判断标准的局限性，结合DAGs还可以指导混杂因素校正方法的选择，在观察性研究中因果推断具有重要指导价值，DAGs在未来的流行病学研究中将有更多的应用。

【关键词】 因果推断；混杂因素；有向无环图

基金项目：国家自然科学基金（81602939）

DOI：10.3760/cma.j.cn112338-20190729-00559

Application of directed acyclic graphs in identifying and controlling confounding bias

Liu Huixin[1], Wang Haibo[2], Wang Ning[3]

[1]Peking University People's Hospital, Beijing 100044, China; [2]Peking University Clinical Research Institute, Beijing 100191, China; [3]National Center for AIDS/STD Control and Prevention, Chinese Center for Disease Control and Preventionz, Beijing 100026, China

Corresponding author: Wang Ning, Email: wangnbj@163.com

【**Abstract**】 Observational study has been viewed as the most convenient method in designing etiological studies. However, the presence of confounders always challenge the researchers in study design, since unadjusted confounders may lead to biased results. The traditional definition of a confounder is not intuitional in application and sometimes leading to inappropriate adjustment of nonexistent "confounders" which might induce new bias to merge. The use of directed acyclic graphs (DAGs) may identify confounders easier and more intuitional, as well as avoiding superfluous adjustment. It can also contribute to the identification of adjustment methods, and be useful in causal inference of observational studies.

【**Key words**】 Causal inference; Confounder; Directed acyclic graphs

Fund program: National Natural Science Foundation of China (81602939)

DOI: 10.3760/cma.j.cn112338-20190729-00559

因果关系推断是流行病学研究的主要内容之一，随机对照试验被视为判断因果关系的金标准，但是在现实研究中，随机对照试验常由于伦理等原因无法实施，而更多采用观察性研究如队列研究进行因果关系推断。观察性研究在进行因果关系推断时，常由于未经识别、校正的混杂因素的存在，而歪曲暴露因素与研究结局之间的真实因果关系[1]，因此识别研究中存在的混杂因素尤为重要。传统流行病学研究中对混杂因素的判断在实际研究应用中不够直观，且有一定局限性，可能导致校正不存在的"混杂因素"而引入新的偏倚[2]。有向无环图（directed acyclic graphs，DAGs）在混杂因素识别中相对直观，并可避免传统混杂因素判断标准存在的局限性，在观察性研究因果推断中具有重要指导价值[3-5]。关于DAGs的语言、规则以及因果推断中的应用国内已有介绍[1,6-8]，本文主要介绍如何应用DAGs识别混杂因素、传统混杂因素判断标准的局限性及完善方法，并以实例说明如何结合DAGs进

行混杂偏倚的校正。

一、采用DAGs识别混杂因素的规则及举例

通过DAGs可以更加直观的识别研究中存在的混杂因素，其判断规则可归纳为[9]：①暴露因素与研究结局之间存在开放的后门路径；或者②暴露因素与研究结局存在共同的祖先变量或母变量，满足以上任意一条，即可判断存在混杂因素。下面结合DAGs图例说明混杂因素的判断：如图1A中，暴露因素E（阿司匹林）与结局变量O（脑卒中）存在共同的母变量C（心血管疾病），或者说E与O之间存在开放的后门路径，E←C→O，则该研究中，C（心血管疾病）为混杂因素；图1A中所举例子的因果关系也可能如图1B所示，E代表暴露因素阿司匹林，O代表研究结局脑卒中，C代表心血管疾病，U代表未测量的因素动脉粥样硬化，E和O之间存在一条开放的后门路径，即：E←C←U→O，C为研究中的混杂因素，或者说，E和O之间有共同的祖先变量U。

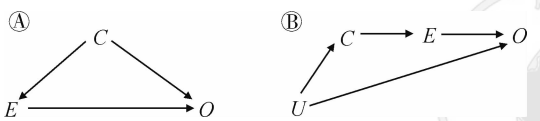

图1　混杂因素DAGs示意图

二、传统混杂因素判断标准的局限性及其完善

1. 应用DAGs说明传统混杂因素判断标准的局限性：传统流行病学研究中判断混杂因素的标准为：① 与研究暴露因素相关（关联）；②与研究结局相关（关联）；③不是暴露因素与研究结局因果链上的中间变量[10]。当研究中存在同时满足以上3点判断标准的因素即为混杂因素。但是这一判断标准是存在局限性的，通过DAGs的绘制可以发现，某些情况下，如果校正了符合传统混杂因素标准的“混杂因素”，可能会引入新的偏倚。

如图2A所示的一项子代肥胖与子代患糖尿病关系的研究中，E代表子代肥胖，O代表子代患糖尿病，U_2代表母亲携带糖尿病易患基因，L代表母亲患糖尿病，U_1代表母亲肥胖。U_1（母亲肥胖）和U_2（母亲携带糖尿病易患基因）都可导致L（母亲患糖尿病），U_1（母亲肥胖）可导致E（子代肥胖），U_2（母亲携带糖尿病易患基因）可导致O（子代患糖尿病），此时，暴露因素E到结局O之间并没有开放的后门路径（L为碰撞变量，阻断了E到O的后门路径），暴露因素E（子代肥胖）与研究结局O（子代患糖尿病）之间的关系为因果关系，E与O之间的路径为因果路径，其效应为E→O路径效应。但是，图2A中因素L满足传统混杂因素定义：L与暴露因素E相关（关联），即L←U_1→E，因素L与研究结局O相关L←U_2→O，而且因素L不是E与O因果链上的中间变量。如果在数据分析阶段校正因素L，那么暴露因素E到研究结局O之间的后门路径开启：E←U_1→[L]←U_2→O，暴露因素E与结局变量O之间的效应包含两部分：①E与O之间的关联效应加上②E 到O的后门路径效应，此时E与O的相关关系不再是因果关系，校正L因素后引入了偏倚（此处引入的为选择偏倚，因为校正L即在L某一取值水平上估计E与O之间的相关关系）。图2A所示的DAGs也称为“M-diagram”[2]。

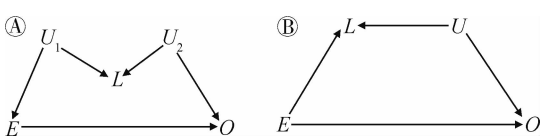

图2　因素L不是暴露因素E与结局O之间的混杂因素示意图

另外一种符合传统混杂因素判断标准，但并非研究中的混杂因素的例子：如图2B所示的宫颈癌癌前病变与宫颈癌研究中，E代表宫颈癌癌前病变，L代表宫颈癌筛查，U代表未测量的其他合并疾病或症状（如HPV感染），O代表宫颈癌。有癌前病变和HPV感染者，会增加宫颈癌筛查频率（E→L，U→L），癌前病变和HPV感染都可增加患宫颈癌风险（E→O，U→O），因素L与暴露因素E相关，同时，因素L与研究结局O相关（L←U→O），且L不是E与O因果链上的中间变量，L满足传统混杂因素判断标准。但是如图2B所示，L是碰撞变量（E→L←U），校正L使得暴露因素E与研究结局O之间阻断的路径开放，同样歪曲了暴露因素E与研究结局O之间的因果关系。

图2A、B的DAGs中，变量L满足传统混杂因素判断的3条规则，但校正L会歪曲暴露因素E与研究结局O之间的因果关系，即传统混杂因素判断的3条规则存在一定局限性。

2. 结合DAG对传统混杂因素判断规则进行完善：

（1）已有学者意识到传统混杂因素判断规则的局限性，并提出完善方案[11-12]：将规则1和规则2“变量C与暴露因素E和研究结局O相关”替换为：“U可拆分为互为补集的两个变量集合U_1和U_2（U是U_1和U_2的并集，且U_1和U_2的交集为空），且满足以下两个条件：① 按变量C分层后，U_1和暴露因素E不相关；②校正变量C、暴露因素E以及U_1后，U_2和研究结局O不相关”。这一规则可以避免将如图2A所示的研

究DAGs中的变量*L*误判为暴露因素*E*与结局*O*之间的混杂因素。

（2）将规则3“混杂因素*C*不是暴露因素*E*与研究结局*O*因果链上的中间变量”替换为：“存在变量*C*和*U*，使得校正变量*C*和*U*后，暴露因素*E*各水平发生结局事件*O*的风险相等”，当这一条件满足时，也满足*C*不是暴露因素*E*与研究结局*O*因果链上的中间变量这一条件。如图1B所示，校正变量*C*和变量*U*后，暴露因素*E*与结局事件*O*之间的后门路径*E*←*C*←*U*→*O*被阻断，暴露因素*E*与结局事件*O*之间的路径仅有直接路径*E*→*O*，即校正变量*C*和*U*后，暴露因素*E*各水平发生结局事件*O*的风险相等。按此规则判断图2B，变量*L*不满足此规则，因为校正变量*L*后，暴露因素*E*与结局事件*O*之间的路径除了直接路径*E*→*O*外，还存在开放的后门路径*E*→[*L*]←*U*→*O*，由此判断，图2B中，变量*L*非暴露因素*E*与结局*O*之间的混杂因素。

三、结合DAGs校正研究中的混杂因素及实例分析

为将研究中混杂偏倚的影响减少到最小，需要在研究的各个阶段对混杂偏倚进行控制。结合DAGs，可以识别出进行因果推断时需要校正的最小校正子集[1]。Pearl[9]给出了判断最小校正子集的两个标准：①集合中的变量阻断了从暴露因素到研究结局的每一条开放的后门路径；②集合中的变量阻断了因校正集合中的变量而产生的新的从暴露因素到结局的所有开放后门路径。结合DAGs，校正研究中的混杂因素，也就等同于阻断所有开放的后门路径。阻断开放的后门路径的方法有两种，一是基于分层分析的方法，即在DAGs中“框住”需要校正的混杂因素（图3A）；一是基于标准化和逆概率权重的方法，即在DAGs中“删除”混杂因素与暴露因素间的箭头（图3B）。

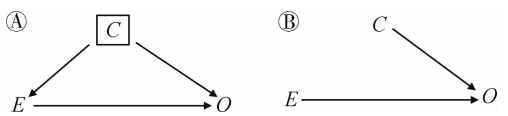

图3 DAGs中校正混杂因素*C*示意图

1.“框住”需要校正的混杂因素：在研究设计阶段采用限制或者匹配的方法来校正混杂偏倚，以及数据分析阶段采取分层分析方法控制混杂因素，都是图3A所示“框住”需要校正的变量来阻断开放的后门路径的方法。该方法是将研究人群按照混杂因素进行分层，使得每一层内暴露组和非暴露组混杂因素分布均衡可比，然后估计每一层内暴露因素与研究结局之间的关系，计算观察人群亚组中的条件概率。如图1A所示研究中，按照是否患心血管疾病C分层，然后计算每层内的*RR*值（或*OR*值），再将各层的*RR*（或*OR*）值综合起来（如采用Mantel-Haenszel分层分析方法），得到控制混杂因素后的调整*RR*（或*OR*）值[10]。

2.“删除”混杂因素与暴露因素间箭头：标准化和逆概率权重（inverse probability weighting，IPW）方法都是采用“删除”变量间箭头的方法（图3B），使得暴露因素与混杂因素不再相关。IPW通过权重的调整产生一个虚拟人群（pseudo-population），在该虚拟人群中，暴露因素E与研究结局*O*之间的相关关系不受混杂因素*C*影响，即“删除”了*C*与*E*之间的箭头，近些年使用的边际结构模型即基于此原理[13-14]。IPW就是基于过去的暴露和混杂因素水平，研究对象接受实际暴露因素处理的条件概率的倒数，估计公式：

$$W^{E}=1/f[\mathrm{E}|\mathrm{C}]$$ [13-14]

当暴露因素不随时间变化时，则以上两种阻断路径的方式都可以校正混杂因素，但是当暴露因素为依时协变量（time-dependent variable）时，则应采用基于逆概率权重（如边际结构模型、G-estimation等），即“删除”混杂因素与暴露因素之间箭头的方法进行因果推断[13]。

四、总结

不同于传统混杂因素判断标准，应用DAGs可以直观识别研究中存在的混杂因素，将复杂的因果关系可视化，避免传统混杂因素判断标准的局限性；同时，结合DAGs可以识别研究中需要控制的最小校正子集，选择混杂偏倚控制方法，可以预见，在未来的流行病学研究中，DAGs将更多的应用于因果推断中。

具有上述优点的同时，DAGs也存在一定局限性[1,7-8,15]：首先，DAGs的绘制依赖于研究者对于研究问题相关先验知识的掌握程度，当研究涉及的因果关系复杂时，通常会得到几个DAGs，无法判断哪一个更正确，此时需绘制所有可能的DAGs，识别每个DAGs的最小校正子集，才能得到真实可靠的结论；其次，由于未测量的混杂因素的存在，校正最小校正子集后，仍可能存在残余混杂，利用DAGs并不能总是使得因果推断的结果为无偏的估计，但是已尽可能使得因果推断接近最真实的因果效应[1,7-8,15]。流行病学专业人员在应用DAGs进行因果推断时，

应尽可能掌握足够多的研究相关先验知识，从而绘制最接近真实情况的DAGs；同时在研究设计阶段对识别出的混杂因素进行测量，以减小残余混杂，充分利用DAGs在因果推断中的优势，得到最接近真实的因果效应估计值。

利益冲突　所有作者均声明不存在利益冲突

参 考 文 献

[1] 向韧，戴文杰，熊元，等. 有向无环图在因果推断控制混杂因素中的应用[J]. 中华流行病学杂志，2016，37(7)：1035-1038. DOI：10.3760/cma.j.issn.0254-6450.2016.07.025.
Xiang R, Dai WJ, Xiong Y, et al. Application of directed acyclic graphs in control of confounding[J]. Chin J Epidemiol, 2016, 37(7): 1035-1038. DOI: 10.3760/cma.j.issn.0254-6450.2016.07.025.

[2] Rothman KJ, Greenland S, Lash TL. Modern epidemiology[M]. 3rd ed. Philadelphia: Lippincott Williams & Wilkins, 2008.

[3] Greenland S, Brumback B. An overview of relations among causal modelling methods[J]. Int J Epidemiol, 2002, 31(5): 1030-1037. DOI: 10.1093/ije/31.5.1030.

[4] Pearl J. Causality: models, reasoning and interference[M]. Cambridge: Cambridge University Press, 2009: 1-102.

[5] Pearl J. An introduction to causal inference[J]. Int J Biostat, 2010, 6(2): 7. DOI: 10.2202/1557-4679.1203.

[6] 郑英杰，赵耐青. 有向无环图：语言、规则及应用[J]. 中华流行病学杂志，2017，38(8)：1140-1144. DOI：10.3760/cma.j.issn.0254-6450.2017.08.029.
Zheng YJ, Zhao NQ. Directed acyclic graphs: languages, rules and applications[J]. Chin J Epidemiol, 2017, 38(8): 1140-1144. DOI: 10.3760/cma.j.issn.0254-6450.2017.08.029.

[7] 刘子言，吴小丽，解美秋，等. 有向无环图在循证医学中的应用[J]. 中国医师杂志，2018，20(2)：180-182. DOI：10.3760/cma.j.issn.1008-1372.2018.02.006.
Liu ZY, Wu XL, Xie MQ, et al. Application of directed acyclic graphs in evidence-based medicine[J]. J Chin Phys, 2018, 20(2): 180-182. DOI: 10.3760/cma.j.issn.1008-1372.2018.02.006.

[8] 刘子言，吴小丽，解美秋，等. 在因果推断中应用有向无环图识别和控制选择偏倚[J]. 中华疾病控制杂志，2019，23(3)：351-355. DOI：10.16462/j.cnki.zhjbkz.2019.03.022.
Liu ZY, Wu XL, Xie MQ, et al. Application of directed acyclic graphs in identification and control of selection bias in causal inference[J]. Chin J Dis Control Prev, 2019, 23(3): 351-355. DOI: 10.16462/j.cnki.zhjbkz.2019.03.022.

[9] Pearl J. Causal diagrams for empirical research[J]. Biometrika, 1995, 82(4): 669-710. DOI: 10.2307/2337338.

[10] 詹思延. 流行病学[M]. 7版. 北京：人民卫生出版社，2012.
Zhan SY. Epidemiology[M]. 7th ed. Beijing: People's Medical Publishing House, 2012.

[11] Robins JM. Causal inference from complex longitudinal data[M]//Berkane M. Latent Variable Modeling and Applications to Causality. New York, NY: Springer Verlag, 69-117.

[12] Greenland S, Pearl J, Robins JM. Confounding and collapsibility in causal inference[J]. Statist Sci, 1999, 14(1): 29-46. DOI: 10.1214/ss/1009211805.

[13] Hernán MA, Robins JM. Cousal Inference: What If[M]. Boca Raton: Chapman & Hall/CRC, 2020.

[14] 刘慧鑫，彭志行，苏迎盈，等. 应用边际结构模型控制依时混杂偏倚[J]. 中华流行病学杂志，2015，36(7)：759-761. DOI：10.3760/cma.j.issn.0254-6450.2015.07.020.
Liu HX, Peng ZH, Su YY, et al. Application of marginal structural models to control time-dependent confounding bias[J]. Chin J Epidemiol, 2015, 36(7): 759-761. DOI: 10.3760/cma.j.issn.0254-6450.2015.07.020.

[15] 谭红专. 现代流行病学[M]. 3版. 北京：人民卫生出版社，2019.
Tan HZ. Modern epidemiology[M]. 3rd ed. Beijing: People's Medical Publishing House, 2019.

（收稿日期：2019-07-29）
（本文编辑：李银鸽）

中华流行病学杂志第八届编辑委员会通讯编委组成人员名单

（按姓氏汉语拼音排序）

鲍倡俊	陈　曦	陈　勇	冯录召	高　培	高立冬	高文静	郭　巍	胡晓斌
黄　涛	贾存显	贾曼红	姜　海	金连梅	靳光付	荆春霞	寇长贵	李　曼
李　霓	李　希	李杏莉	林　玫	林华亮	刘　昆	刘　莉	刘　淼	马　超
毛宇嵘	潘　安	彭志行	秦　天	石菊芳	孙　凤	汤奋扬	汤后林	唐雪峰
王　波	王　娜	王　鑫	王海俊	王丽萍	席　波	谢　娟	闫笑梅	严卫丽
燕　虹	杨　鹏	杨祖耀	姚应水	余灿清	喻荣彬	张　本	张茂俊	张周斌
郑　莹	郑英杰	周　蕾	朱益民					