·基础理论与方法·

有向无环图在因果推断控制混杂因素中的应用

向韧 戴文杰 熊元 吴鑫 杨艳芳 王玲 戴志辉 李娇 刘爱忠
410008 长沙，中南大学公共卫生学院流行病与卫生统计学系
通信作者：刘爱忠，Email：lazroy@live.cn
DOI：10.3760/cma.j.issn.0254-6450.2016.07.025

【摘要】 观察性研究是流行病学病因研究中最常用的方法之一，但在因果推断时混杂因素往往会歪曲暴露与结局的真实因果关联。为了消除混杂，选择变量调整是关键所在。有向无环图能够将复杂的因果关系可视化，提供识别混杂的直观方法，将识别混杂转变成识别最小充分调整集。一方面有向无环图可以选择调整更少的变量，增加分析的统计效率；另一方面有向无环图识别的最小充分调整集可以避开未被测量或有缺失值的变量。总之，有向无环图有助于充分揭示真实的因果关系。

【关键词】 病因学研究；有向无环图；混杂因素；因果关系

Application of directed acyclic graphs in control of confounding

Xiang Ren, Dai Wenjie, Xiong Yuan, Wu Xin, Yang Yanfang, Wang Ling, Dai Zhihui , Li Jiao , Liu Aizhong
Department of Epidemiology and Health Statistics, School of Public Health, Central South University, Changsha 410008, China
Corresponding author: Liu Aizhong, Email: lazroy@live.cn

【**Abstract**】 Observational study is a method most commonly used in the etiology study of epidemiology, but confounders, always distort the true causality between exposure and outcome when local inferencing. In order to eliminate these confounding, the determining of variables which need to be adjusted become a key issue. Directed acyclic graph (DAG) could visualize complex causality, provide a simple and intuitive way to identify the confounding, and convert it into the finding of the minimal sufficient adjustment for the control of confounding. On the one hand, directed acyclic graph can choose less variables, which increase statistical efficiency of the analysis. On the other hand, it could help avoiding variables that is not measured or with missing values. In a word, the directed acyclic graph could facilitate the reveal of the real causality effectively.

【**Key words**】 Etiology study; Directed acyclic graph; Confounder; Causality

病因研究是流行病学研究中主要内容之一。系以通过描述性研究，提出病因假设，然后以分析性研究检验该假设，最后通过实验性研究予以验证。其中随机对照试验被视为金标准，但在实际应用中往往难以实施。例如，将吸烟施加于人群违背伦理道德，因此在研究吸烟对肺癌的影响时就无法采用随机对照试验。所以此时病因研究更多采用描述性和分析性研究等观察性研究方法。但是在观察性研究中，经常测量不到或不能随机分配某些协变量，因而在因果推断时，这些因素就会导致混杂，歪曲暴露与结局的真实因果联系。因此有必要识别这些混杂因素，并在资料分析时采用多元回归、分层分析等方法对其调整，消除混杂的影响。

因果推断中如何识别混杂因素是其关键。目前流行病学研究中判别混杂因素的标准[1-3]：应是所研究结局的危险因素，并与研究的暴露因素有关，且不是暴露与结局因果链上的中间变量[4]。然而，对某一具体研究而言，当相关变量较多时，满足该混杂因素标准的变量可能有n个，那么为消除混杂，是否有必要调整所有的混杂因素？另外，由于变量间的关系复杂，是否漏掉某些混杂因素？

为解决此问题，有学者将有向无环图（directed acyclic graphs，DAG）与因果关系联系起来，并作为一门工具应用于流行病学[5-6]。特别是当存在多个混杂因素时，应用DAG可将各变量的关系用直观可视的图形表示，以梳理各变量间的关系[7-10]，并认为优于单纯使用混杂因素的传统标准。DAG研究变量间因果关系的假设是定性和非参数的[5-6]，其识别

混杂因素是利用变量间的因果关系，而不依赖于观察所得的统计学联系[11]。当涉及多个变量，即使存在无法测量的变量时，可将暴露、结局以及潜在的混杂因素等各变量之间的因果关系表示在图上，从而便于研究者运用一系列直观的、操作简单的规则识别混杂因素[6,10,12]，从而达到将复杂的关系可视化的效果[13]。

1. 基本原理：首先将箭头连接节点构成有方向的路径，但并不形成封闭的循环，由这些节点以及连接节点的箭头组成的图形，即DAG。其中节点表示变量，箭头表示变量间因果关系的方向（原因→结果），一个箭头代表一个变量对另一变量的直接因果效应。

若箭头从变量X出发指向Y，如中间无其他变量，则X是Y的父代（parent），Y是X的子代（child）（X→Y）；如果中间至少有一个其他变量M，则X是Y的祖先（ancestor），Y是X的孙代（descendant）（X→M→Y）[5-6,14]。箭头从变量X出发指向Z且中间无其他变量（X是Z的父代），同时另一箭头从变量Y出发也指向Z且中间无其他变量（Y是Z的父代），即Z是X、Y的共同（子代）效应，那么Z就称为经过X、Y、Z路径上的一个冲撞点（X→Z←Y）[6,15]。

构建DAG实质就是构建一个潜在的因果关系图，可以显示图中哪些变量之间存在统计学关联。如果X是Y的原因或Y是X的原因或它们有一个共同的原因P，那么X和Y可能具有统计学关联[14]。即使X、Y不相关，如果它们有一个共同的效应Z，那么在Z的某一层内，X和Y仍具有统计学关联[15]。这些关联可以通过开放路径传递，也随着路径的阻断而消失。即只有暴露与结局变量同时处于一条开放路径中，暴露与结局才会产生关联，如果二者间的路径是阻断路径，那么暴露与结局无关。DAG的冲撞点不能传递关联，至少包含一个冲撞点的路径是阻断路径，不包含冲撞点的路径就是开放路径[6]。

根据研究目的，原始DAG中的路径可分为因果路径和非因果路径，后者主要是指后门路径（backdoor path）。因果路径是指箭头从暴露出发，每一个箭头的方向不变，最后指向结局的路径[6,11]，此时，暴露与结局产生因果关系，如

吸烟→肺纤维化→肺癌

若暴露E到结局D的某条路径有一个变量箭头指向E，那么该路径是从暴露E到结局D的后门路径[5]。最简单的后门路径，如

路径“E←P→D”就是一条从暴露E到结局D的后门路径，在这条路径中，暴露E与结局D有关联。后门路径中没有冲撞点时就是一条开放路径，传递变量间的关联。

DAG中，暴露与结局产生联系是因为它们都在因果路径或者不含冲撞点的后门路径上。但是因果路径是代表暴露对结局因果效应的唯一路径，只有因果路径代表真正的因果关联，包含暴露与结局变量的开放的后门路径是非因果关联，可影响暴露与结局的真实因果关系，所以包含暴露与结局变量的开放的后门路径的存在即代表混杂的存在。作为识别混杂因素的工具，DAG主要优点之一就是有效地识别出开放的后门路径。

2. 识别混杂因素的步骤：

（1）识别是否存在混杂：可通过以下步骤判断[6]，①删除所有从暴露发出的箭头；②查看是否有任何从暴露到疾病的开放路径，即暴露的效应去除后，观察暴露与疾病是否通过开放的后门路径产生关联。如果步骤②结果为否，该研究中就不存在混杂；反之，如果步骤②中从暴露到疾病还存在开放路径，那么就代表存在混杂，需进一步识别混杂因素。

（2）判断某变量集是否足以控制混杂：在因果关系分析中，混杂因素通常可能不止一个。如果调控其中的某一个、几个或全部变量就足以控制所有混杂，那么这些需要控制的混杂变量组成的集合就称为控制混杂的充分调整集（sufficient adjustment sets）。理论上，在充分调整集的任意一阶层内，暴露与结局的因果关联都没有混杂[6]。

如果存在混杂，对于给定的S集（包括一个以上变量），Shrier和Platt[16]在Pearl充分性后门试验（the backdoor test for sufficiency）的基础上总结出判断该S集为充分调整集的六步法则（6-step process）[11]：①S集中的任意变量都不是暴露E的后代；②将既不是暴露E或结局D的祖先或父代，也不是S集的祖先或父代的变量删除；③删除所有从暴露E发出的箭头；④如果S集变量中有冲撞点，应连接冲撞点的父代；⑤删除连线的所有箭头；⑥删除所有与S集的变量相连的连线。如果E与结局D无关，那么调整步骤①所选择的协变量S集理论上就可以消除所有混杂。

（3）识别控制混杂的最小充分调整集（minimally sufficient adjustment sets）：确定充分调整集S后，删除S集的某些变量可能仍然足以控制混杂。如果调整Q集可以将混杂消除且其没有子集可以将混杂消除，那么Q集就为控制混杂的最小充分

调整集[6]。最小充分调整集Q通常是充分调整集S的子集或本身，DAG中可能有n个不同的最小充分调整集，且包含的变量个数及其测量的难易程度可能均有不同。为了找到一个最小充分调整集，可以循序地从充分集S中删除变量，直到没有新的能被删除的变量为止，即没有新的变量集能够满足后门试验或六步法则的要求。

例如为了解儿童抗组胺药治疗对哮喘发病率的影响，可利用DAG找出暴露与结局之间的混杂因素。图1表示各变量间的关系[6]，即空气污染水平(X)与性别(Y)无关；空气污染通过影响支气管反应(Z)或抗组胺药治疗(E)进而影响哮喘(D)发病；性别可直接影响哮喘发病或可首先通过影响支气管反应从而影响哮喘发病。

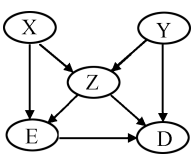

图1 抗组胺药对哮喘发病影响的DAG(原始图)

首先判断该研究是否存在混杂，删除所有从暴露E发出的单向头箭后，图1简化成图2，存在E到D的开放路径E–Z–Y–D，E–X–Z–D和E–Z–D。所以图1存在混杂，需要进一步识别混杂因素。利用六步法则识别能够控制图1中混杂的充分调整集S。

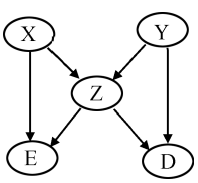

图2 由DAG原始图删除从暴露E发出的箭头

如果S={X,Y,Z}，图1经过六步法则后，得到图3。表明E与D无关，S={X,Y,Z}是一个充分调整集，即调整变量X、Y、Z足以控制混杂。

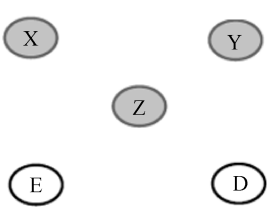

图3 将变量集{X,Y,Z}代入六步法则

如果S={Y,Z}，图1经六步法则后，简化成图4，显示E与D无关，说明S={Y,Z}是一个充分调整集。

如果S={Z}，图1经过六步法则后，简化成图5，此时E与D通过路径E–X–Y–D相关，说明S={Z}不是一个充分调整集，调整变量Z不能控制图1的所有混杂。

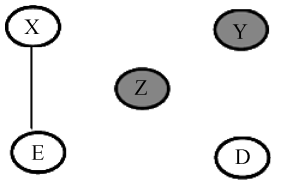

图4 将变量集{Y,Z}代入六步法则

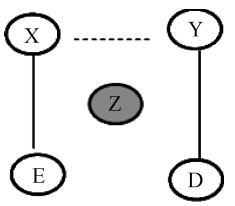

图5 将变量{Z}代入六步法则

如果S={X,Y}，图1经过六步法则后，简化成图6，E与D通过路径E–Z–D相关，说明S={X,Y}不是一个充分调整集，调整变量X，Y不能控制图1的所有混杂。

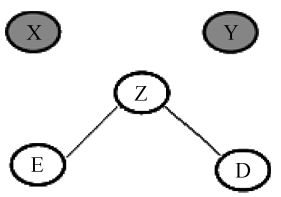

图6 将变量集{X,Y}代入六步法则

由图3、4可知，变量集S={X,Y,Z}、S={Y,Z}都是控制混杂的充分调整集，同样也能用与图4类似的图形得到S={X,Z}也是一个充分调整集。即调整以上3个S集中的任意一个S集均足以控制该研究的混杂。表明图1中空气污染水平、性别、支气管反应都是混杂因素。统计分析时同时调整三者或者同时调控支气管反应与性别或支气管反应与空气污染水平，就能得到真实的因果关联。以上实例中，{支气管反应，性别}、{支气管反应，空气污染水平}为最小充分调整集，而性别比空气污染水平更易测量，因此可以选择调整支气管反应和性别。

实际应用时，一般选取变量少且易于测量的S集，一方面可以减少工作量，另一方面可以增加统计分析效率(协变量减少，自由度增加)。但需要注意的是，由于DAG中未测量或不能测量的变量也有可能被选在充分调整集中，通过选取最适宜的最小充分调整集，在统计分析时也许可以避开这些变量。例如2013年Röhrig等[9]在观察心电图与老年人伤残状况的关联研究中，正是利用DAG别出了最适宜的最小充分调整集(图7)。该研究随机选取德国西部1 037名≥65岁老年人，收集性别、年龄、收入、饮酒、

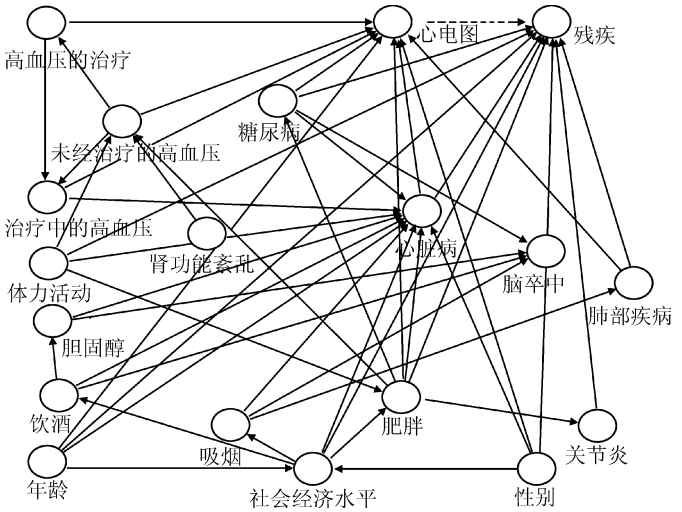

图7　心电图与残疾关联研究的DAG

心电图检查结果、伤残情况、脑血管病史，糖尿病、胆固醇水平、社会经济水平、BMI、体力活动、心脏病、关节炎、肺部疾病、高血压等资料。首先利用DAG识别出以下3个控制混杂的最小充分调整集。最小充分调整集1共12个变量：年龄、性别、体力活动、饮酒、胆固醇水平、糖尿病、心脏病、肺部疾病、肥胖、教育程度、收入、吸烟；最小充分调整集2共10个变量：年龄、性别、教育程度、收入、体力活动、肥胖、脑卒中、糖尿病、心脏病、肺部疾病；最小充分调整集3共9个变量：年龄、性别、体力活动、糖尿病、心脏病、未经治疗的高血压、治疗中的高血压、肺部疾病、肥胖。由于饮酒变量中有174人饮酒情况不明，而没有收集未经治疗的高血压资料，因此排除了第一个和第三个变量集，选择调整第二个最小充分调整集，避免缺失值和未测量变量的影响。

调整DAG选择的混杂因子，心电图检查结果异常与残疾的关联无统计学意义（OR=1.293，P>0.05）；而采用逐步回归法选择自变量，最后得出心电图检查结果异常与残疾有统计学关联（OR=1.523，P<0.05）。逐步回归法是在排除了174名饮酒情况不明者的数据基础上进行的，排除者和纳入者并非随机选择，与饮酒情况明确者相比，饮酒情况不明者残疾率更高，二者没有同质性，因此在排除这些缺失值的基础上得到的结论有偏倚。而DAG混杂因素的选择并不是根据现有收集到的数据得到的，而是根据先验知识，且调整的变量中没有包含饮酒这个变量，不涉及缺失值的情况。因此，调整DAG识别的混杂变量能够避开某些没有收集的或不完整的信息，从而更充分地消除混杂。

3. DAG在混杂控制中利弊：其优点首先可将复杂的因果关系可视化，为识别混杂及控制混杂的最小充分调整集提供了直观手段，可用于研究的设计和分析阶段；其次可识别和消除所有混杂的因素集，有助于选择所有的混杂因素，消除因果效应估计的混杂偏倚，充分有效地揭示尽可能真实的因果关系；此外还可定性识别混杂因素，其过程中不涉及统计学分析，数据分析时如与传统的流行病学分析方法（分层分析、回归模型等）结合更能真实地揭示因果关联。

但DAG也存在局限性。如使用因果关系图识别混杂因素的前提是根据先验知识（征询专家意见、查找文献等）绘制出正确的因果关系图，但出于病因研究的复杂性，有的因果关系假设可能不确定，从而得到几个DAG，且无法判断哪个更正确，此时将绘制出每一个可能的DAG，并识别各图的最小充分调整集，还需给予相应的解释[12,16]。

利益冲突　无

参 考 文 献

[1] Geng Z, Guo JH, Lau TS, et al. Confounding, homogeneity and collapsibility for causal effects in epidemiologic studies[J]. Stat Sinica, 2001, 11(1): 63–75.
[2] Geng Z, Guo JH, Fung WK. Criteria for confounders in epidemiological studies[J]. J Roy Stat Soc Ser B-Stat Methodol, 2002, 64(1): 3–15. DOI: 10.1111/1467–9868.00321.
[3] Jager KJ, Zoccali C, Macleod A, et al. Confounding: what it is and how to deal with it[J]. Kidney Int, 2008, 73(3): 256–260. DOI: 10.1038/sj.ki.5002650.
[4] Suttorp MM, Siegerink B, Jager KJ, et al. Graphical presentation of confounding in directed acyclic graphs[J]. Nephrol Dial Transplant, 2015, 30(9): 1418–1423. DOI: 10.1093/ndt/gfu325.
[5] Pearl J. Causal diagrams for empirical research[J]. Biometrika, 1995, 82(4): 669–688. DOI: 10.1093/biomet/82.4.669.
[6] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research[J]. Epidemiology, 1999, 10(1): 37–48. DOI: 10.1097/00001648–199901000–00008.
[7] Foster EM. Causal inference and developmental psychology[J]. Dev Psychol, 2010, 46(6): 1454–1480. DOI: 10.1037/a0020204.
[8] Risch EC, Owora A, Nandyal R, et al. Risk for child maltreatment among infants discharged from a neonatal intensive care unit: a sibling comparison[J]. Child Maltreat, 2014, 19(2): 92–100. DOI: 10.1177/1077559514539387.
[9] Röhrig N, Strobl R, Müller M, et al. Directed acyclic graphs helped to identify confounding in the association of disability and electrocardiographic findings: results from the KORA-Age study[J]. J Clin Epidemiol, 2014, 67(2): 199–206. DOI: 10.1016/j.jclinepi.2013.08.012.
[10] de Jonge HCC, Azad K, Seward N, et al. Determinants and consequences of short birth interval in rural Bangladesh: a cross-sectional study[J]. BMC Pregnancy Childbirth, 2014, 14(1): 427. DOI: 10.1186/s12884–014–0427–6.
[11] Williamson EJ, Aitken Z, Lawrie J, et al. Introduction to causal diagrams for confounder selection[J]. Respirology, 2014, 19(3): 303–311. DOI: 10.1111/resp.12238.
[12] Howards PP, Schisterman EF, Poole C, et al. "Toward a clearer definition of confounding" revisited with directed acyclic graphs[J]. Am J Epidemiol, 2012, 176(6): 506–511. DOI: 10.1093/aje/kws127.
[13] Vanderweele TJ, Tan ZQ. Directed acyclic graphs with edge-specific bounds[J]. Biometrika, 2012, 99(1): 115–126. DOI: 10.1093/biomet/asr059.
[14] VanderWeele TJ, Robins JM. Four types of effect modification: a classification based on directed acyclic graphs[J]. Epidemiology, 2007, 18(5): 561–568. DOI: 10.1097/EDE.0b013 e318127181b.
[15] VanderWeele TJ, Hernán MA, Robins JM. Causal directed acyclic graphs and the direction of unmeasured confounding bias[J]. Epidemiology, 2008, 19(5): 720–728. DOI: 10.1097/EDE.0b013e3181810e29.
[16] Shrier I, Platt RW. Reducing bias through directed acyclic graphs[J]. BMC Med Res Methodol, 2008, 8(1): 70. DOI: 10.1186/1471–2288–8–70.

（收稿日期：2015–11–18）
（本文编辑：张林东）