·基础理论与方法·

匹配在观察性研究中的作用——有向无环图视角

罗涛[1]　王璐[1]　田恬[1]　符文慧[1]　裴华莲[1]　郑英杰[2]　戴江红[1]

[1]新疆医科大学公共卫生学院流行病学与卫生统计学教研室，乌鲁木齐 830011；[2]复旦大学公共卫生学院流行病学教研室，国家卫生健康委员会卫生技术评估重点实验室，公共卫生安全教育部重点实验室，上海 200032

通信作者：戴江红，Email：epi102@sina.com

【摘要】 匹配是观察性研究中选择研究对象的一种常用方法，具有控制混杂因素、提高统计效率等作用，但其控制混杂因素的作用在不同观察性研究中并不一致，匹配在队列研究中能够消除匹配变量的混杂偏倚，但在病例对照研究中匹配本身并不能消除混杂偏倚。在匹配性病例对照研究选择匹配变量时，研究者可能并不能准确判断该变量是否为混杂变量，若误将真实情况为非混杂因素的变量进行匹配，则会形成过度匹配，造成统计效率下降或引入难以避免的偏倚或增加工作量等；若将真实情况为混杂因素的变量遗漏不予匹配，则会造成混杂偏倚。有向无环图是一种直观的展示不同流行病学研究设计、变量间复杂因果关系的可视化图形语言。本文从有向无环图视角分析匹配在不同观察性研究设计中的作用、匹配性病例对照研究中欲匹配变量的选择标准制定，为今后流行病学研究设计提供一定的参考建议。

【关键词】 匹配；　有向无环图；　观察性研究

基金项目：国家重点研发计划（2017YFC0907203，SQ2017YFSF090013）；国家自然科学基金（81560539）

Matching in observational research: from the directed acyclic graph perspective

Luo Tao[1], Wang Lu[1], Tian Tian[1], Fu Wenhui[1], Pei Hualian[1], Zheng Yingjie[2], Dai Jianghong[1]

[1]Department of Epidemiology and Health Statistics, School of Public Health, Xinjiang Medical University, Urumqi 830011, China; [2]Department of Epidemiology, Key Laboratory for Health Technology Assessment, National Health Commission, Key Laboratory of Public Health Safety, Ministry of Education, School of Public Health, Fudan University, Shanghai 200032, China

Corresponding author: Dai Jianghong, Email: epi102@sina.com

【Abstract】 Matching is a standard method for selecting research objects regarding the observational research, which controls confounding factors and improves statistical efficiency. However, its role in controlling confounding is not consistent in different observational studies. Matching can eliminate the confounding bias of matching variables in cohort studies, but checking on itself cannot eliminate confounding bias in case-control studies. In matched case-control studies, researchers may not accurately judge whether the variable is a confounder. Sometimes the variables that are not confounders are mistakenly matched. In that case, it will result in overmatching, which will lead to the decline of statistical efficiency or the introduction of unavoidable bias or increase of workload. If the real confounding factors are omitted, it will cause confounding bias. Therefore, researchers should consider what kind of matching variable selection criteria should be formulated. A directed acyclic graph is a visual graphic language that can show the complicated causality among

DOI：10.3760/cma.j.cn112338-20200601-00793

收稿日期　2020-06-01　　**本文编辑**　万玉立

引用本文：罗涛，王璐，田恬，等. 匹配在观察性研究中的作用——有向无环图视角[J]. 中华流行病学杂志，2021，42(4)：740-744. DOI：10.3760/cma.j.cn112338-20200601-00793.

different epidemiological research designs. This article analyzes the role of Matching in different observational research designs from the perspective of the directed acyclic graph, formulates the selection criteria for matching variables in matched case-control studies, and provides some reference suggestions for future epidemiological research design.

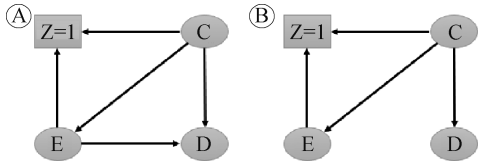
【Key words】 Matching;　Directed acyclic graph;　Observational research

Fund programs: National Key Research and Development Program of China (2017YFC0907203, SQ2017YFSF090013); National Natural Science Foundation of China (81560539)

观察性研究不可避免地存在着不同程度的选择偏倚、信息偏倚和混杂偏倚[1]，由于上述偏倚的存在，研究者可能得到的是被扭曲的暴露和结局之间的因果关联[2]。对混杂因素进行匹配是观察性研究设计中常用的方法之一，在队列研究设计及病例对照研究设计中广泛使用，尤其在病例对照研究设计中[3]。对某因素进行匹配，可采用个体匹配、频数匹配以及倾向性评分等方法[4]，确保该因素在队列研究设计中各暴露组间或病例对照研究设计中各结局组间的分布相同（或尽可能接近）[5]。对混杂因素进行正确的匹配可消除混杂因素对研究结果造成的偏倚以及提高研究效率，但目前研究人员对匹配在不同观察性研究设计中的作用以及匹配性病例对照研究中欲匹配变量的选择等问题上认识并不充分[6-7]。有向无环图（directed acyclic graph，DAG）是因果关系研究图像工具，其由节点（表示变量）以及连接各个节点的有向边（表示变量之间的因果关系）组成[8-9]，DAG 在混杂因素的识别、指导数据分析以及复杂变量之间因果关系可视化等方面得到了广泛应用[10-11]。因此本文使用 DAG 对不同观察性研究中匹配的作用以及匹配性病例对照研究中欲匹配变量的选择标准进行分析。

1. 匹配在不同观察性研究中的作用：

（1）匹配性队列研究即使用频数匹配或个体匹配的方法确保匹配的混杂变量在不同暴露组之间大致相同，以达到控制混杂因素的目的。本文图中以椭圆表示变量，方框表示变量取某一固定值，有向箭头表示变量之间存在因果关系。图 1A 表示 E 与 D 存在因果关联时的 DAG，设图中 E 代表是否吸烟（1：吸烟，0：不吸烟），D 代表是否发生肺癌（1：发生，0：不发生），C 代表性别（1：男性，0：女性），Z 代表是否纳入研究（1：纳入，0：不纳入）。从 E→Z=1 表示暴露者（吸烟者）相对于非暴露者（非吸烟者）更容易被纳入队列研究中（假设人群中吸烟率为 20%，当暴露组与非暴露组纳入人数相等，则吸烟者被选入研究的概率为非吸烟者的 4 倍）；C→Z=1 表示不同性别人群吸烟情况不同，因人群中男性吸烟比例高于女性，故吸烟组中男性所占比例高，为达到性别匹配的目的，在非吸烟组中男性将更容易被选入研究。C 与 E 之间存在两条开放路径 C→E 以及 E→Z=1←C，对变量 C 进行匹配，C 在 E 的每个组间分布一致，此时 C 与 E 条件独立，意味着 C 与 E 之间存在的两条开放路径 C→E 以及 E→Z=1←C 刚好大小相等、方向相反、相互抵消。E 与 D 之间存在 E→D、E←C→D、E→Z=1←C→D 3 条开放路径，其中 E→D 为因果路径，E←C→D、E→Z=1←C→D 为偏倚路径，但根据 C→E 以及 E→Z=1←C 刚好相互抵消这一条件，可知 E←C→D、E→Z=1←C→D 这两条路径也能相互抵消，因此 E、D 之间仅剩欲研究的因果路径 E→D[12]，此时匹配本身就能实现控制混杂因素的目的。图 1B 表示 E 与 D 无因果关系时，其基本分析过程与图 1A 一致，研究者也可通过匹配获得 E 与 D 之间的真实零关联。

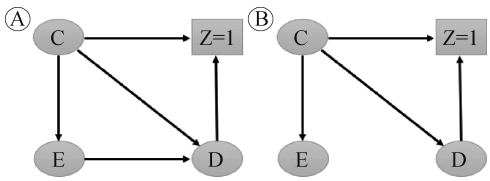


注：匹配因素为混杂因素

图 1　匹配性队列研究

（2）匹配性病例对照研究即使用频数匹配或个体匹配的方法确保匹配变量在病例组与对照组之间的分布大致相同。如图 2A 所示，当 E 与 D 存在因果关联，设图 2 中 E 代表有无口服避孕药服药史（1：有，0：无），D 代表是否患有心肌梗死（1：患病，0：不患病），C 代表年龄（1：≥40 岁，0：<40 岁），Z 代表是否纳入研究（1：纳入，0：不纳入）。D→Z=1 表示患者比正常人更容易被纳入研究，C→Z=1 表示考虑到年龄对是否患有心肌梗死存在影响，故在对照组中年龄较大者更加容易被纳入研究。C 与 D 之间存在 3 条开放路径 C→D、C→Z=1←D、C→E→D，对变量 C 进行匹配，C 与 D 相互独立，C→D、C→Z=1←D、C→E→D 这 3 条路径相互抵消，意味着 C→D、C→Z=1←D 之间的净关联不为零，也意味着

E 与 D 之间的 3 条开放路径 E→D、E←C→D、E←C→Z=1←D 中的 2 条混杂路径净关联不为零，因此匹配 C 本身并不能打断混杂路径 E←C→D，研究者在数据分析时需对 C 进行调整才能获得 E→D 的真实因果关联[6-7]。如图 2B 所示，当 E→D 无因果关系时，C 与 D 之间存在两条开放路径 C→D、C→Z=1←D，对变量 C 进行匹配，C 与 D 相互独立，C→D、C→Z=1←D 这两条路径相互抵消，意味着 E 与 D 之间的两条开放混杂路径 E←C→D、E←C→Z=1←D 相互抵消，因此匹配 C 打断混杂路径 E←C→D，此时研究者在数据分析时不控制匹配因素也能准确获得 E 与 D 之间的零关联。



注：匹配因素为混杂因素

图 2　匹配性病例对照研究

2. 匹配性病例对照研究中匹配变量的选择：

在匹配性病例对照研究中，研究者在选择匹配变量时，并不能确保某欲匹配变量一定为混杂变量或非混杂变量，前者若不匹配则可能导致混杂偏倚，若误将后者匹配则可能造成过度匹配，故本文根据欲匹配变量的不同情况进行如下讨论，为研究者制定匹配变量标准提供一些建议。

（1）如果某一变量表现为对暴露存在明确效应，但是其与结局的关系尚不明确时，其基本 DAGs 如图 3A 所示，其中---加？代表两变量之间的因果关系不确定。若变量之间的真实情况如图 3B 所示，C 对结局 D 不存在直接效应，类似于暴露修饰变量[13]，此时若误将 C 进行匹配，C 在 D 中各组间分布一致，C 与 D 条件独立，C 与 D 之间的开放路径 C→Z=1←D 与 C→E→D 相互抵消，E 与 D 之间产生无法抵消的偏倚路径 E←C→Z=1←D，该偏倚路径会削弱 E→D 的关联强度，从而使得 E 与 D 之间的关联趋向于零关联，这是因为当 C 与 E 存在关系，对 C 在 D 的各组间进行匹配，会导致 E 在 D 的各组间分布也趋于相同，此时对 C 进行调整虽也能获得拟研究因果关系 E→D 的正确点估计，但是会导致研究的统计效率降低[5]。若变量之间的真实情况如图 3C 所示，C 与 D 之间存在直接关系，即 C 为混杂因素，对其匹配则是正确的，分析过程与本文前述一致。若变量之间的真实情况如图 3D 所示，结局 D 通过 C 对暴露 E 产生因果效应，从而产生循环路径，C 为反向效应的中间变量[13]，根据 DAGs 的基本规则以及变量之间的时序关系，此时暴露 E 处于两个不同的时间点，应分为 E_0 和 E_1 两个变量，而不是简单的一个变量 E，变量之间的因果路径应为 E_0→D→C→E_1（如现患病例偏倚），故本文对此不进行讨论。

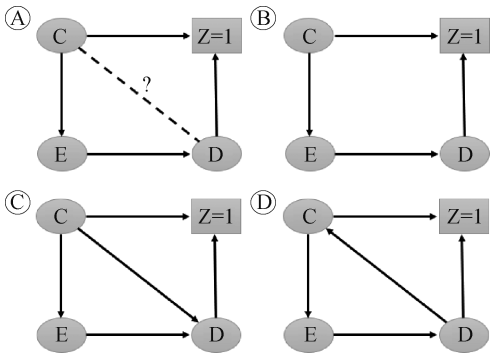


图 3　候选匹配因素对暴露存在明确效应

（2）如果某一变量表现为对结局存在明确效应，但是其与暴露的关系尚不明确时，其基本 DAGs 如图 4A 所示。若变量之间的真实情况如图 4B 所示，C 与 E 之间不存在关联，C 类似于效应修饰变量[13]，此时误将 C 进行匹配，因路径 C→D 与路径 C→Z=1←D 相互抵消，虽然匹配 C 不会对研究暴露结局关联产生偏倚，因为匹配 C 并不会对结局各组间的暴露分布产生影响，但匹配 C 会增加研究的工作量、某些病例因不存在对照而丢弃导致的统计效率下降，从而形成过度匹配。若变量之间的真实情况如图 4C 所示，C 对暴露 E 存在直接效应，即 C 为混杂因素，那么匹配则是正确的，分析过程与本文前述一致。若变量之间的真实情况如图 4D 所示，暴露 E 对 C 存在直接效应，则 C 为中间变量，此情形在稍后进行详述。

（3）如果某一变量与暴露及结局都存在关系，

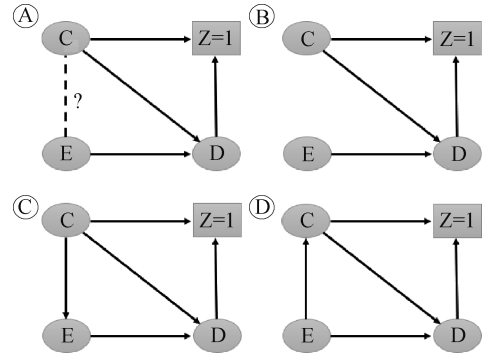


图 4　候选匹配因素对结局存在明确效应

但是变量之间关联的方向并不确定时，其基本DAG如图5A所示，其中↔加？代表两变量之间的因果关系的方向不确定。若变量之间的真实情况如图5B所示，C既是暴露的效应变量，又是结局的效应变量，此时C为碰撞变量[13]，碰撞变量较为特殊，正常情况下碰撞变量的存在并不会对拟研究的暴露和结局之间的因果效应产生影响，但当对碰撞变量进行调整、控制时，则会引入原本不存在的碰撞偏倚[14]，此时误将C进行匹配，E与D之间存在E→D、E→C→Z=1←D、E→C←D 3条开放路径，即使后续数据分析对C进行控制，碰撞路径E→C←D也不会关闭，从而造成研究结果存在偏倚[7]。若变量之间的真实情况如图5C所示，即暴露可通过C影响结局[13]，此时误将C进行匹配，E与D之间直接存在3条开放路径E→D、E→C→D、E→C→Z=1←D，匹配会打断因果路径E→C→D，研究者只能获得E→D的直接效应，造成研究结果有偏。若变量之间的真实情况如图5D所示，C对暴露E存在直接效应，即C为混杂因素，那么匹配则是正确的，分析过程与本文前述一致。

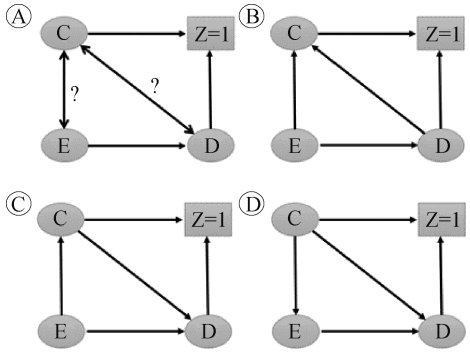


图5　候选匹配因素与暴露和结局都存在关联

3. 总结：目前对于匹配在观察性研究设计中的作用存在两个误解：匹配本身消除了匹配混杂因素造成的偏倚；如果进行了匹配研究设计，则必须进行“匹配分析”[15]。本文主要针对第一个误解进行了综合分析。研究结果表明，对于第一个误解，研究人员需注意即使匹配因素为混杂因素，匹配本身也并不能保证结果无混杂偏倚，这取决于研究设计类型。在匹配队列研究中，由于匹配所引入的选择偏倚，抵消了混杂因素固有的混杂偏倚[16]，因此产生一种类似混杂路径打断的效果，使得研究结果无混杂偏倚[12]；在病例对照研究中，由于匹配所引入的研究对象选择性偏倚不足以完全抵消混杂因素固有的混杂偏倚，因此需额外地对匹配因素进行调整，才能获得真实的因果关联[16]，虽然当暴露与结局不存在因果关联时，由于匹配混杂因素引入的研究对象选择性偏倚可完全抵消混杂因素固有的混杂因素偏倚，因此不需额外地对C进行调整，即可获得E与D真实的零关联，但是客观世界研究中，研究者事先无法得知E与D之间是否存在因果关联，因此在匹配性病例对照研究的数据分析中，控制匹配因素值得推荐。

在匹配性病例对照研究中，研究者不能确定某一候选的欲匹配变量一定为混杂变量或非混杂变量，若对其进行匹配，则可能因匹配非混杂因素而产生类似过度调整偏倚[17]，若不进行匹配，则可能因遗漏混杂因素导致混杂因素控制不完全，从而导致混杂偏倚。本文根据不同的变量情形进行了详细分析，认为研究者可根据具体情形来提出候选匹配变量的选择标准。如果某一变量C表现为对暴露存在明确效应，但是其与结局的关系尚不明确时，虽然C可能为暴露修饰变量，从而形成过度匹配，导致统计效率下降，但若C为混杂因素，不匹配会导致混杂偏倚，考虑到匹配C的损失和收益，在此情形下研究者似乎可以对C进行匹配；如果某一变量C表现为对结局存在明确效应，但是其与暴露的关系尚不明确时，虽然C可能为效应修饰变量，但对其进行匹配并不会引入额外的偏倚，而是某种程度上增加研究工作量以及因不存在对照而丢弃某些病例导致的统计效率下降，但若C为混杂因素，不匹配会导致混杂偏倚，因此研究者似乎也可将其纳入匹配的考虑范围；此外即使某变量C与暴露和结局都有关联，匹配也应谨慎考虑，因为混杂变量与碰撞变量、中间变量仅仅通过关联是很难区分的，研究者需要对变量之间的因果方向有明确的认识，且对碰撞变量、中间变量进行误匹配的后果往往比误匹配暴露修饰变量、效应修饰变量严重，会对研究结果产生不可逆的偏倚[18]，而不仅仅是损失统计效率和增加工作量的问题。综上所述，研究者在选择候选匹配变量时，应明确变量之间的因果关系，并根据不同的情形进行综合的考量。

本文利用DAG的优点，图形化、直观化地将匹配在各种研究设计、各种匹配因素情形下的变量之间因果关系图进行展示，对于厘清匹配在流行病学研究设计中的作用具有一定的积极意义，提示研究人员在研究设计以及数据分析时，尽可能的利用DAG将研究设计、各变量之间关系进行图形可视化，有利于指导数据分析与研究结果的评价。本研

究存在不足，首先，研究为了简明易懂，未考虑数据缺失、存在测量误差以及各个变量之间的效应存在交互作用等情况下的各个变量之间的因果关系图，而这些情形下的因果关系图将更为复杂，但更接近实际情况；其次，本次研究仅利用DAG对匹配在各种情况下的因果关系进行了定性分析，具体的匹配对因果关系估计的影响大小，尤其对因果关系的区间估计的影响仍待进一步研究。

利益冲突　所有作者均声明不存在利益冲突

参 考 文 献

[1] Grimes DA, Schulz KF. Bias and causal associations in observational research[J]. Lancet, 2002, 359(9302): 248-252. DOI:10.1016/S0140-6736(02)07451-2.

[2] 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.

[3] 詹思延，叶冬青，谭红专. 流行病学[M]. 8版. 北京：人民卫生出版社，2017.
Zhan SY, Ye DQ, Tan HZ. Epidemiology[M]. 8th ed. Beijing: People's Medical Publishing House, 2017.

[4] Deb S, Austin PC, Tu JV, et al. A review of propensity-score methods and their use in cardiovascular research[J]. Can J Cardiol, 2016, 32(2): 259-265. DOI: 10.1016/j. cjca. 2015. 05.015.

[5] Rothman KJ, Lash TL, Greenland S. Modern epidemiology [M]. 3rd ed. Philadelphia: Lippincott Williams and Wilkins, 2013.

[6] Ali Mansournia M, Jewell NP, Greenland S. Case-control matching: effects, misconceptions, and recommendations [J]. Eur J Epidemiol, 2018, 33(1): 5-14. DOI: 10.1007/ s10654-017-0325-0.

[7] Mansournia MA, Hernán MA, Greenland S. Matched designs and causal diagrams[J]. Int J Epidemio, 2013, 42(3):860-869. DOI:10.1093/ije/dyt083.

[8] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research[J]. Epidemiology, 1999, 10(1): 37-48. DOI:10.1097/00001648-199901000-00008.

[9] 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.

[10] 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.

[11] 郑英杰，赵耐青，何一宁. 客观世界的因果关系：基于有向无环图的结构解析[J]. 中华流行病学杂志，2018, 39(1): 90-93. DOI:10.3760/cma.j.issn.0254-6450.2018.01.019.
Zheng YJ, Zhao NQ, He YN. Causality in objective world: directed acyclic graphs-based structural parsing[J]. Chin J Epidemiol, 2018, 39(1): 90-93. DOI: 10.3760/cma. j. issn. 0254-6450.2018.01.019.

[12] 何一宁，刘丽丽，蔡倩莹，等. 研究设计时混杂控制策略的结构分类[J]. 中华流行病学杂志，2018, 39(7):999-1002. DOI:10.3760/cma.j.issn.0254-6450.2018.07.025.
He YN, Liu LL, Cai QY, et al. A structural classification of strategies for confounding control in research design[J]. Chin J Epidemiol, 2018, 39(7): 999-1002. DOI: 10.3760/ cma.j.issn.0254-6450.2018.07.025.

[13] 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.

[14] Liu W, Brookhart MA, Schneeweiss S, et al. Implications of M bias in epidemiologic studies:a simulation study[J]. Am J Epidemiol, 2012, 176(10): 938-948. DOI: 10.1093/aje/ kws165.

[15] Pearce N. Analysis of matched case-control studies[J]. BMJ, 2016, 352:i969. DOI:10.1136/bmj.i969.

[16] Sjölander A, Greenland S. Ignoring the matching variables in cohort studies——when is it valid and why? [J]. Stat Med, 2013, 32(27):4696-4708. DOI:10.1002/sim.5879.

[17] Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies[J]. Epidemiology, 2009, 20(4): 488-495. DOI: 10.1097/EDE. 0b013e3181a819a1.

[18] Greenland S. Quantifying biases in causal models: classical confounding *vs* collider-stratification bias[J]. Epidemiology, 2003, 14(3): 300-306. DOI: 10.1097/01.EDE.0000042804. 12056.6C.