控制谁?

Introduction to causal diagrams for confounder selection

北京协和医学院流统协会
真实世界研究系列讲座

孙殿钦

中国医学科学院肿瘤医院

2019-11-09

内容大纲

- Motivation
  - Confounder
  - Practical guide
  - Limitation
- Something new
  - Directed Acyclic Graph (DAGs)
  - Backdoor criteria
  - Principles of confounder selection based on DAGs
  - Tools for creating DAGs

什么是混杂因素?

举个栗子?

什么是混杂因素?

混杂因素亦称外来因素、混杂因子或混杂变量，是指与研究因素和研究疾病均有关，若在比较的人群组中分布不均衡，可以歪曲（缩小或扩大）研究因素与疾病之间真实联系的因素。

混杂因素的基本特点是：

1是所研究疾病的危险因素

2与所研究的因素有关

3不是研究因素与研究疾病因果链上的中间变量

——《流行病学·第7版》

什么是混杂因素?

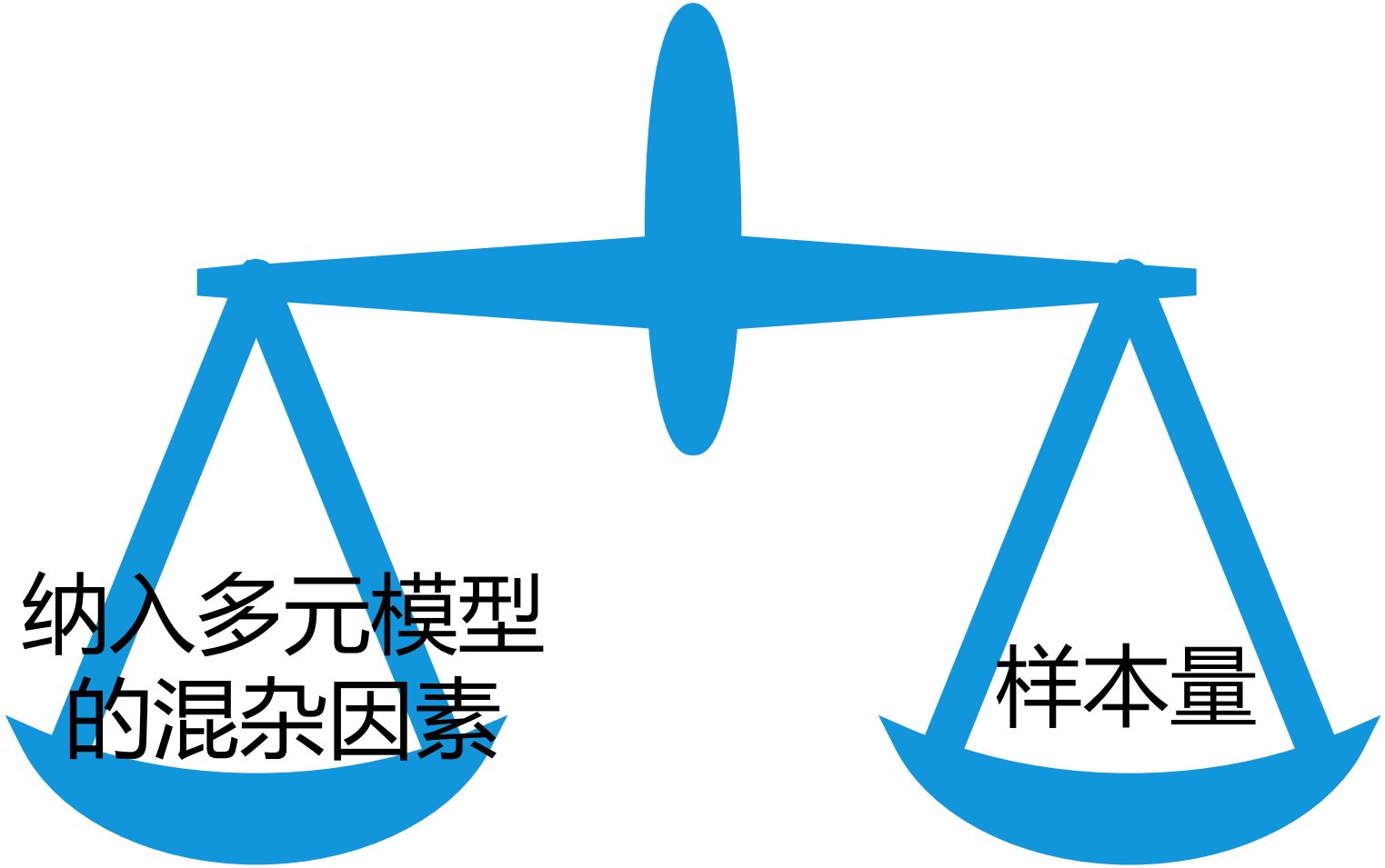
We propose that a “confounder” be defined as a pre-exposure covariate C for which there exists a set of other covariates X such that effect of the exposure on the outcome is unconfounded conditional on (X, C) but such that for no proper subset of (X, C) is the effect of the exposure on the outcome unconfounded given the subset.

VanderWeele TJ, Shpitser I. On the definition of a confounder. *Ann Stat*. 2013;41(1):196–220.

什么是混杂因素?

混杂因素的定义并不重要，重要的是我们要控制谁?

如何筛选混杂因素?



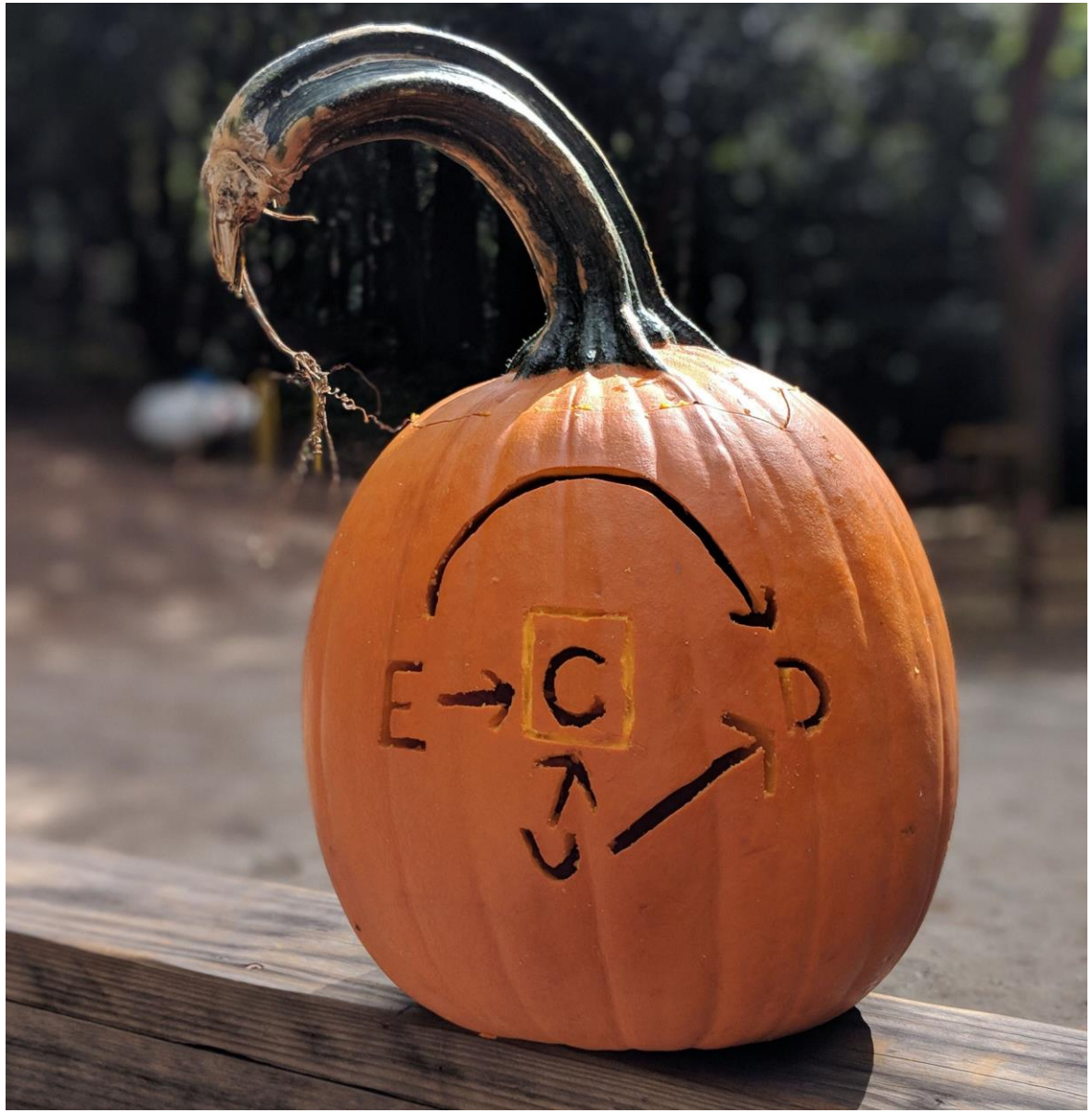
如何筛选混杂因素?

- 临床相关性
- 单因素分析
- 是否会引起效应值较大变化
- 向前/后选择（forward/backward selection）

实际操作中的局限性

- 无法直观考虑变量间错综复杂的关系
- 可能会错误调整中介变量
- 并不是所有的混杂都需要调整
- 如何让审稿人信服？怎么证明你有理有据，没有P-hacking?

有向无环图Directed Acyclic Graph



Current situation

Table 1 Variable selection methods used in explicative studies published in four major epidemiological journals in 2015

	American Journal of Epidemiology	Epidemiology	European Journal of Epidemiology	International Journal of Epidemiology	Total
Prior knowledge or causal graphs	55 (47%)	33 (59%)	27 (46%)	31 (52%)	146 (50%)
Prior knowledge or causal graphs only	40 (34%)	29 (52%)	19 (32%)	28 (47%)	*116 (40%)*
Change in estimate	20 (17%)	5 (9%)	5 (8%)	4 (7%)	34 (12%)
Stepwise	5 (4%)	3 (5%)	7 (12%)	1 (2%)	16 (5%)
Univariate analyses	16 (14%)	4 (7%)	5 (8%)	1 (2%)	26 (9%)
Other	3 (3%)	1 (2%)	1 (2%)	0 (0%)	5 (2%)
Insufficiently detailed	42 (36%)	16 (29%)	24 (41%)	25 (42%)	107 (37%)
Total	118	56	59	59	292

Results are reported as frequency (%). More than one method could be used in each study; as such, percentages do not add up to 100%

Talbot, D. & Massamba, V.K. A descriptive review of variable selection methods in four epidemiologic journals: there is still room for improvement. Eur J Epidemiol (2019) 34: 725. https://doi.org/10.1007/s10654-019-00529-y

举个栗子

- 上世纪70年代，流行病学家发现服用雌激素的女性宫颈癌确诊率较高。
- 对此现象，有两个可能的解释：
    I. 雌激素导致了宫颈癌的发生
    II. 雌激素增加了子宫流血的风险，而宫颈癌也会导致子宫流血，促使了宫颈癌的发现和诊断

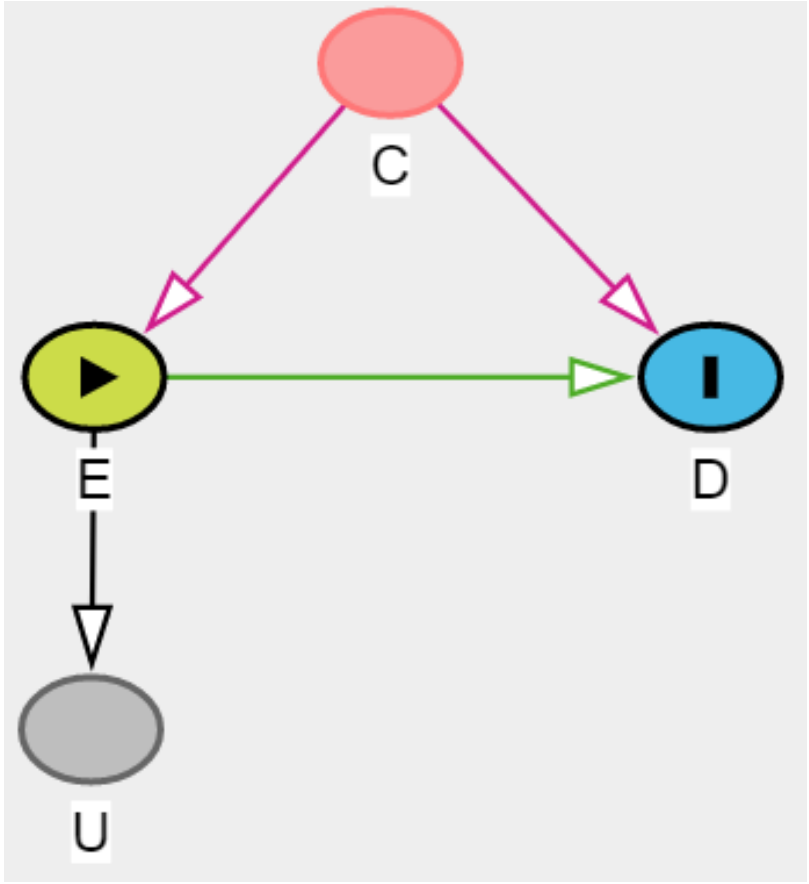
举个栗子

两位来自Yale的研究者提出根据是否出现子宫流血进行分层分析。

举个栗子

来自波士顿和哈佛的研究者表达了反对意见，没有必要控制子宫是否出现流血这一变量。

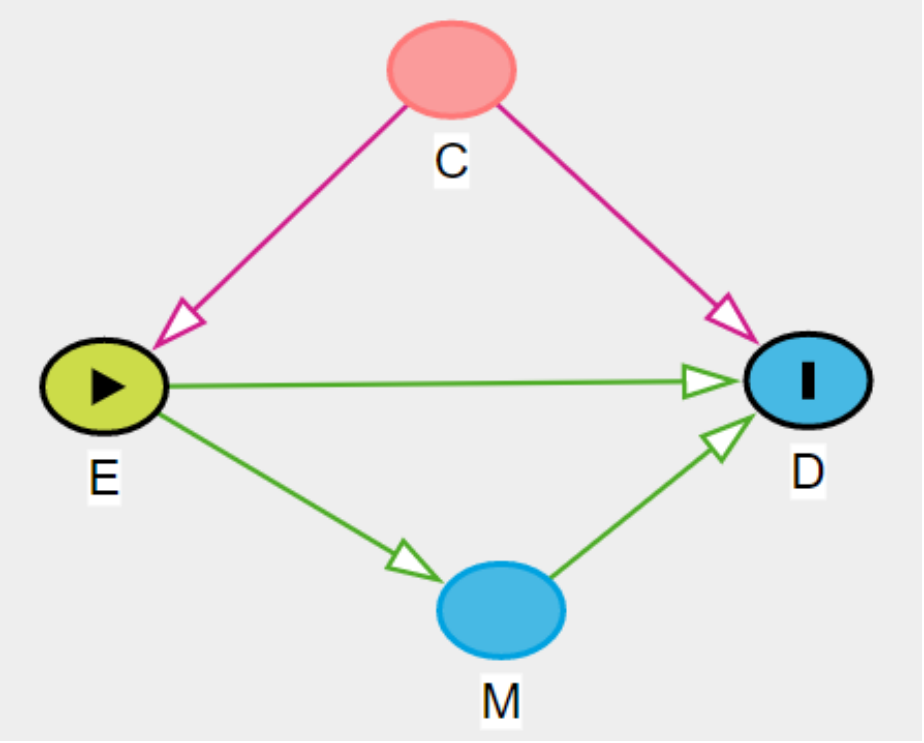
有向无环图Directed Acyclic Graph



- Balls represent variables (nodes)
- Links represent associations
- Arrows represent causal relationships
- Path

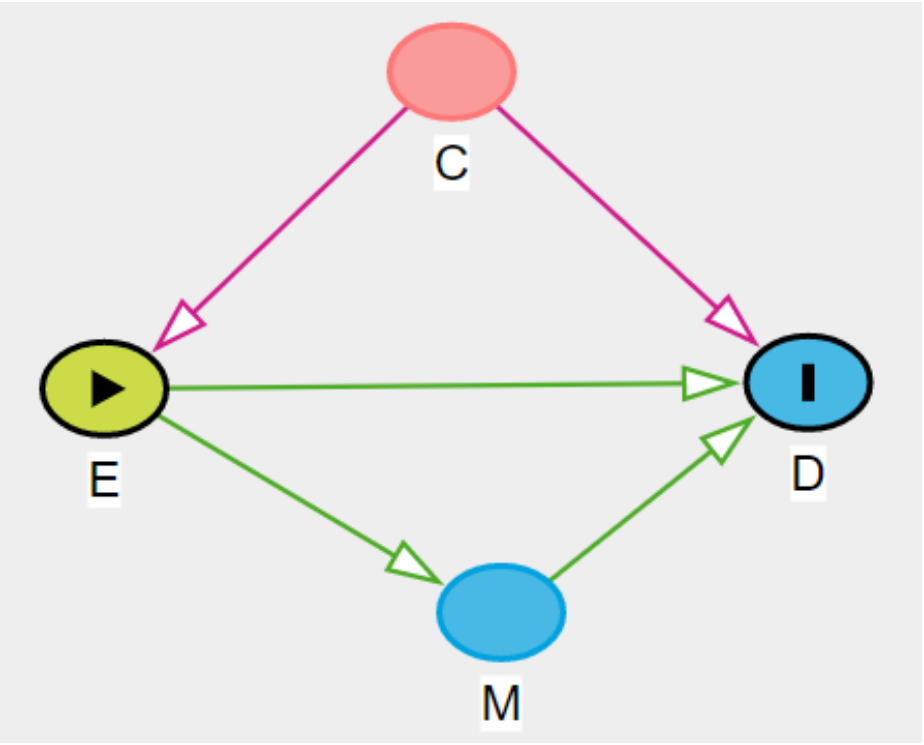
- Directed paths are denoted by arrowed links
- Acylic: No loops, the future does not cause the past

有向无环图Directed Acyclic Graph



- Encoded assumptions
  - Absence of variables: all common (observed and unobserved) causes of any pair of variables
  - Absence of arrows: zero causal effect

有向无环图Directed Acyclic Graph

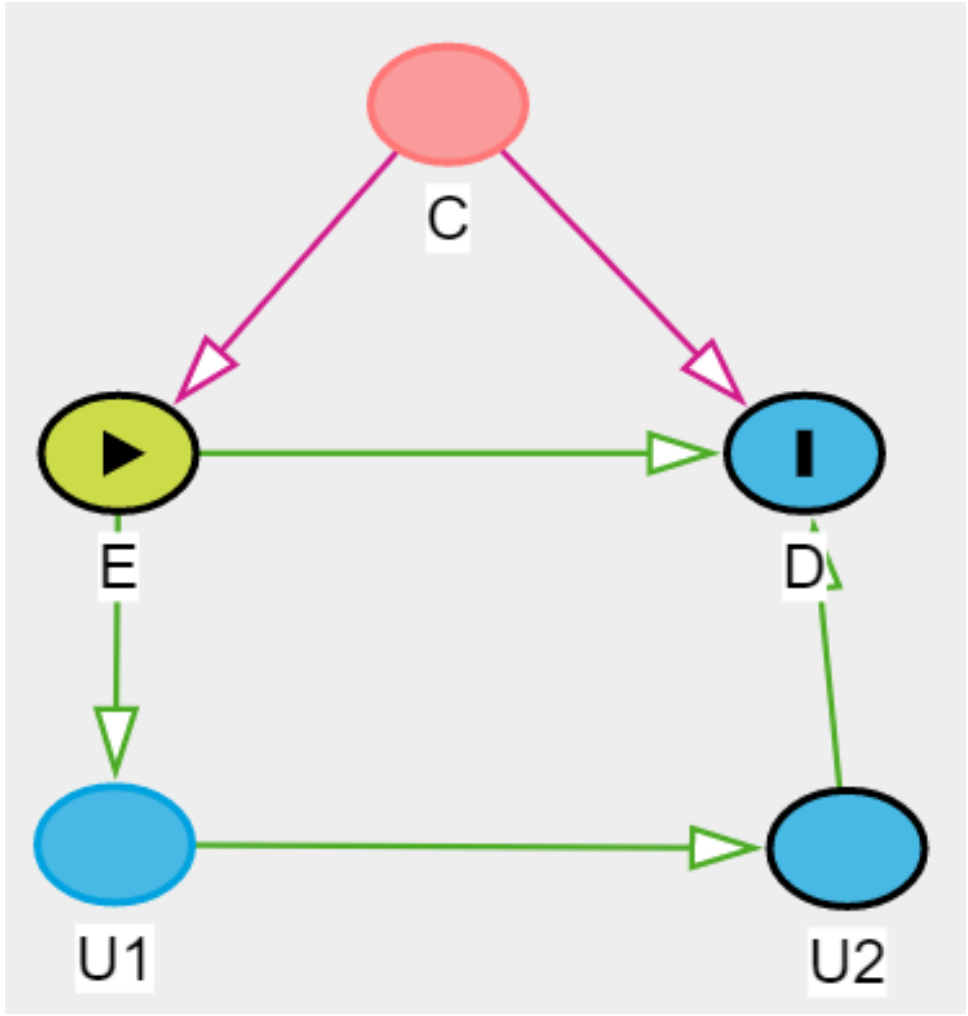


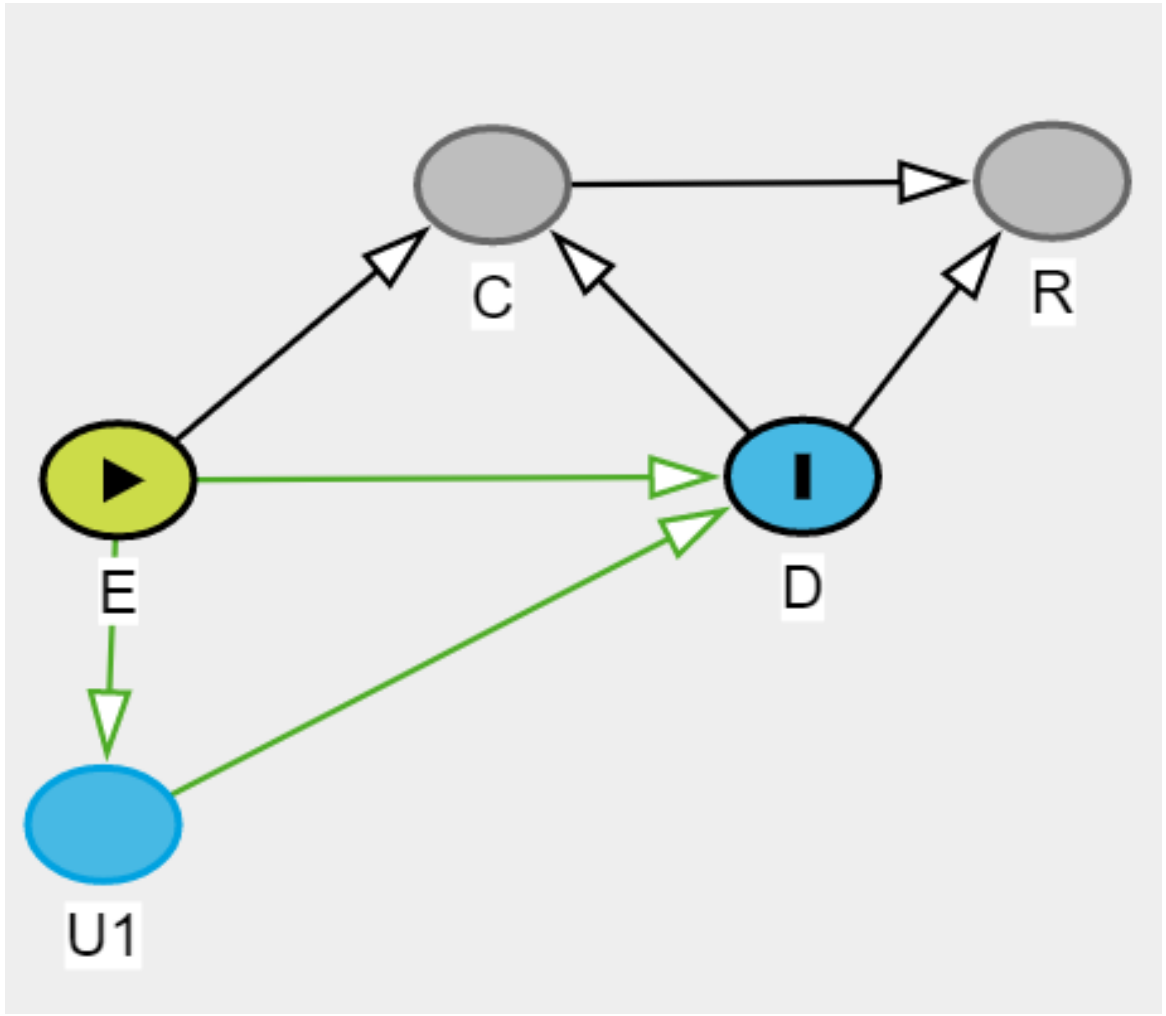
- Frontdoor paths: A path in which all arrows point away from the exposure, E, to the outcome, D
- 前门路径：从暴露出发到结局，路径方向完全顺着该路径上变量间的箭头方向

- Backdoor paths: A path connecting E and D in which at least one arrow points against the flow of time
- 后门路径：从暴露出发到结局，路径方向与路径上变量间箭头方向不完全一致

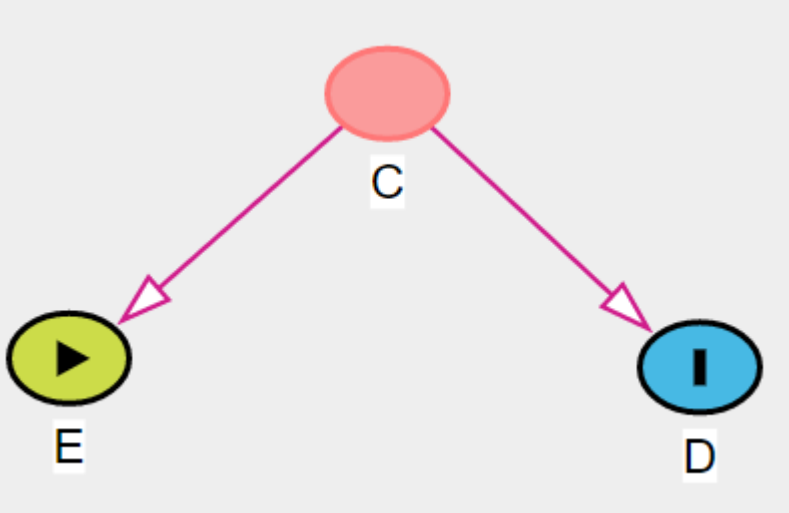
Frontdoor/Backdoor paths ·Practice

Frontdoor paths & Backdoor paths

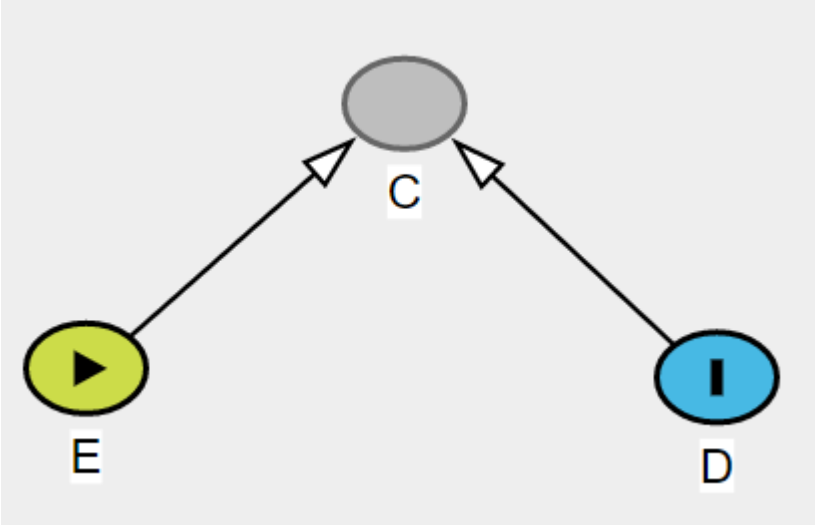




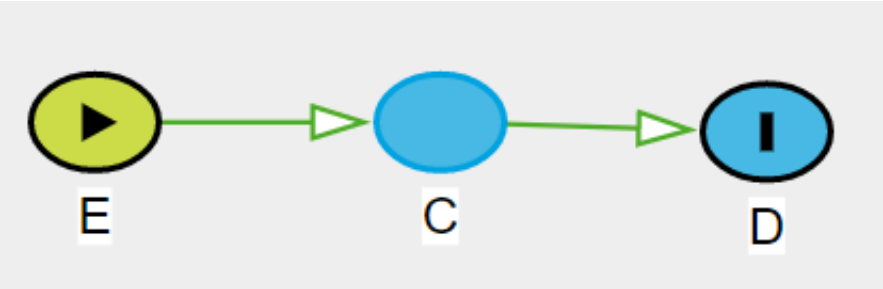
DAG·Elementary DAG



Common cause (confounder)

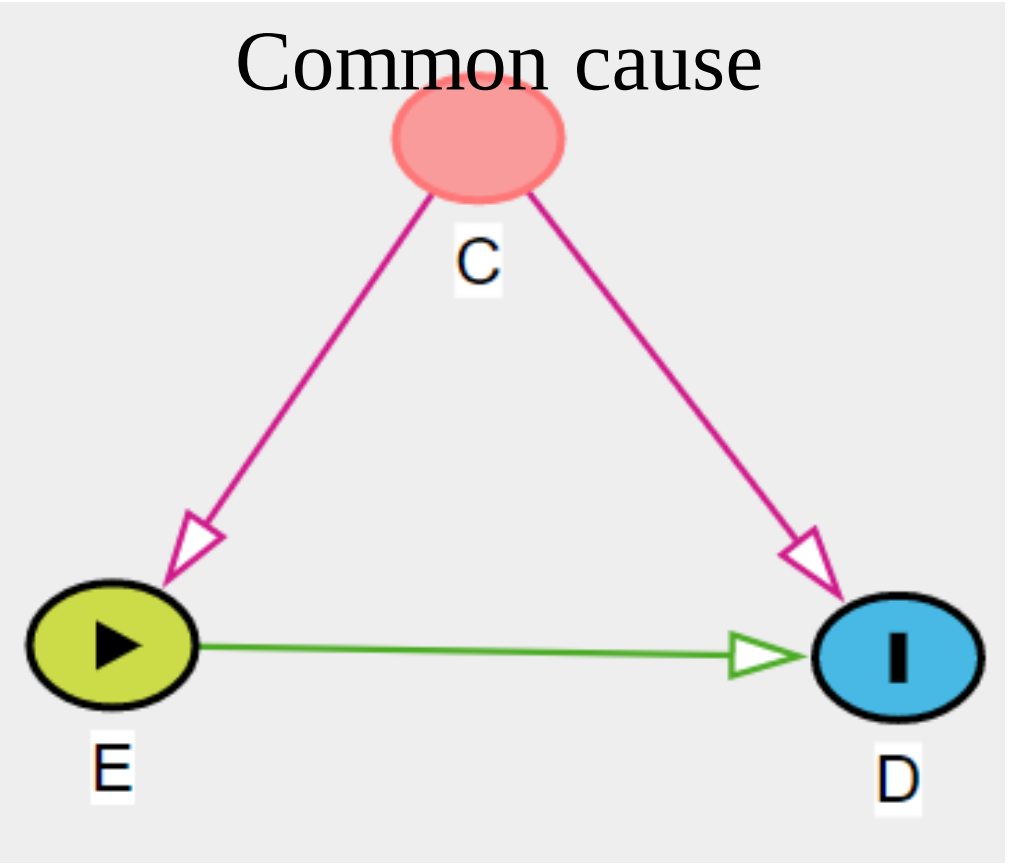


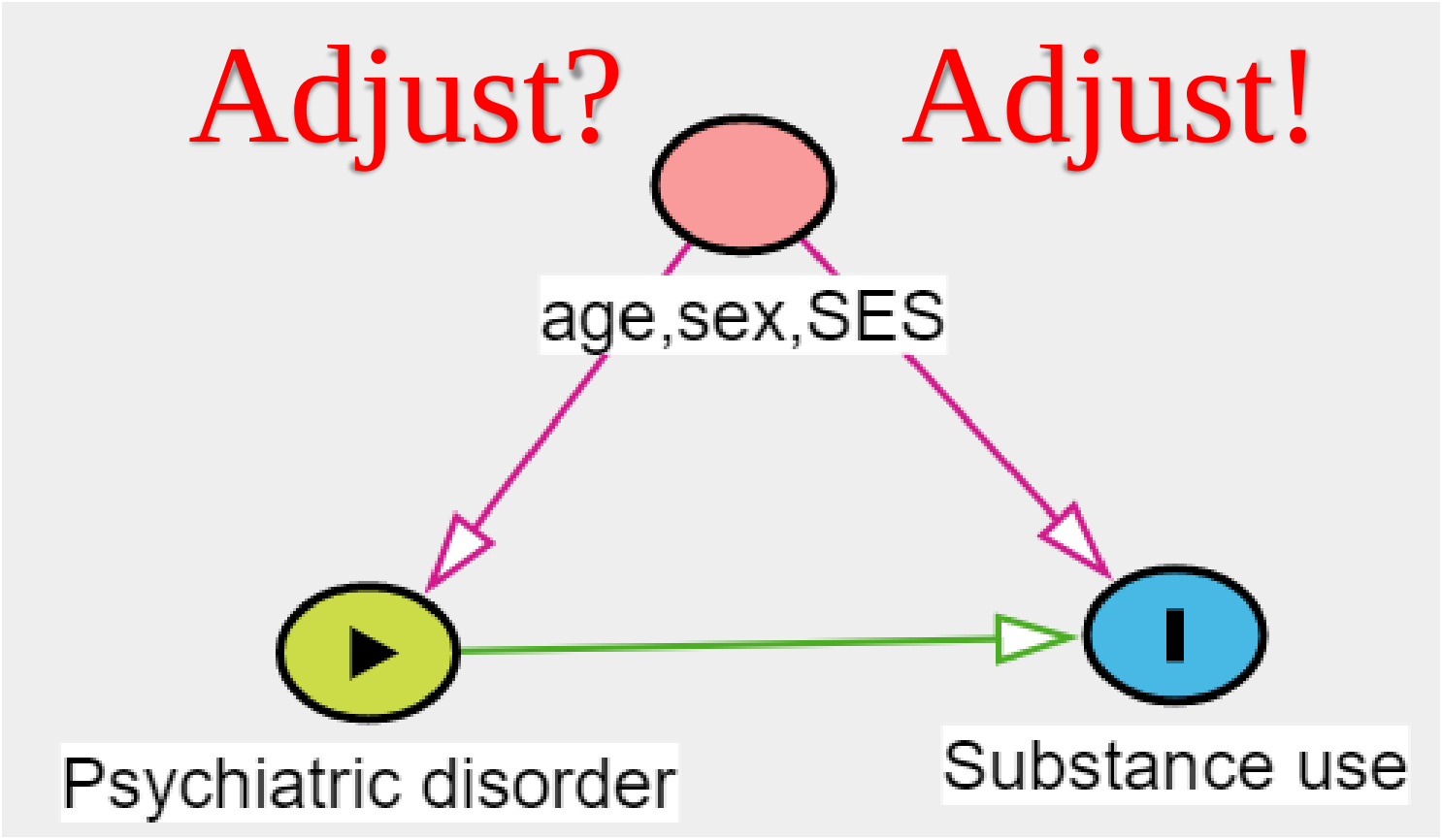
Common effect (collider)



Causal chain (mediator)

DAG·Confounder

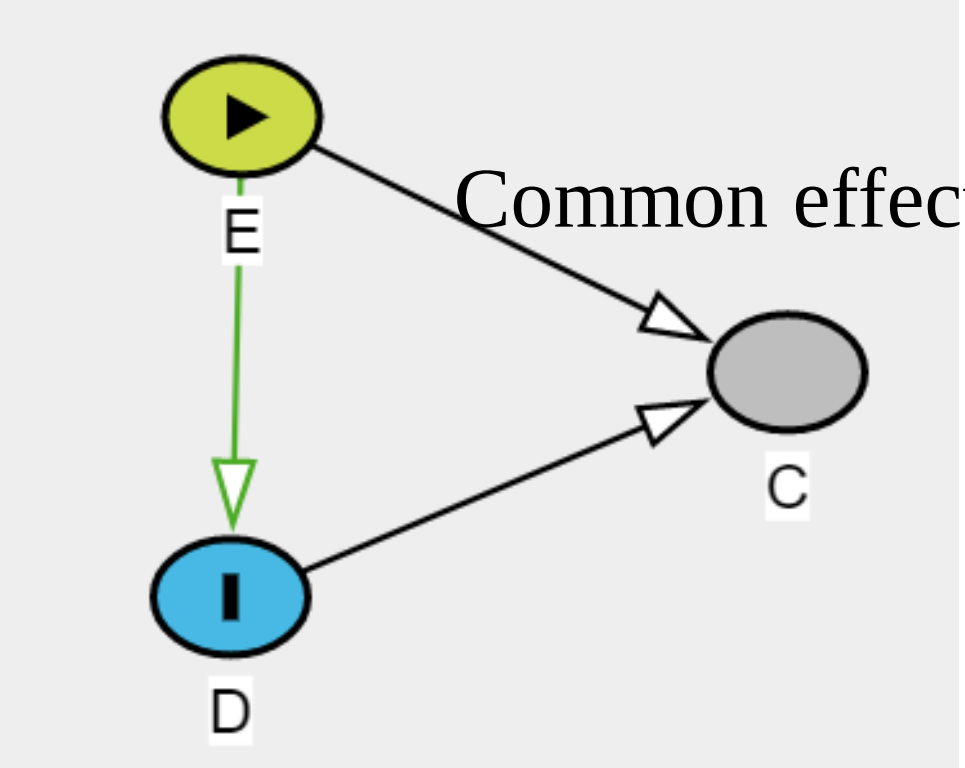


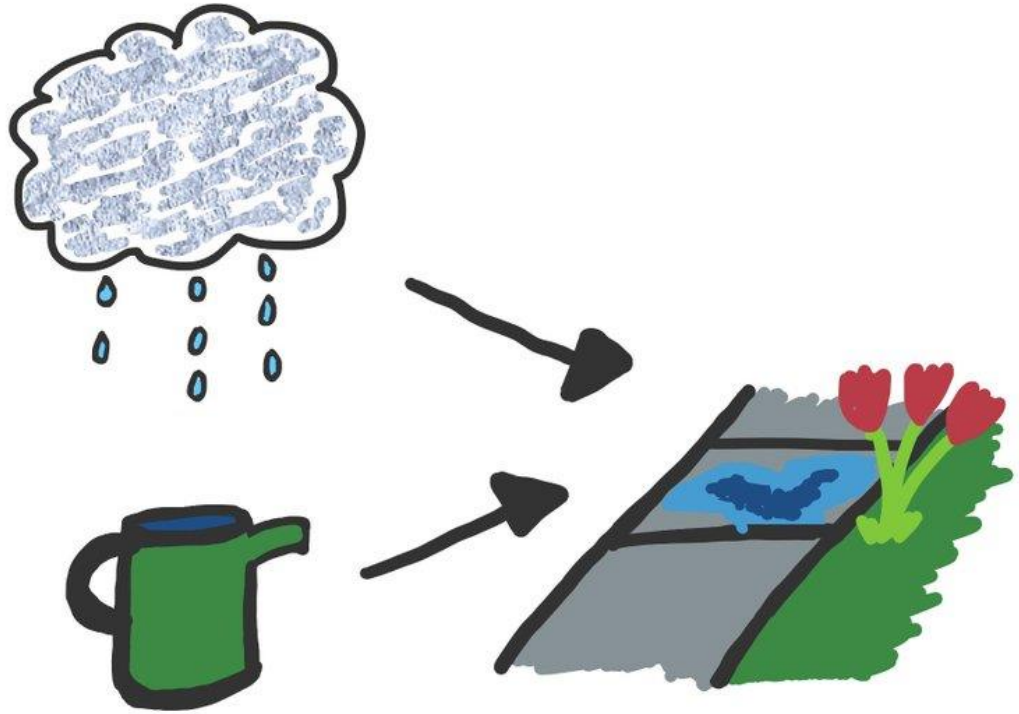


Mangerud, W.L., Bjerkeset, O., Holmen, T.L., Lydersen, S., & Indredavik, M.S. 2014. Smoking, alcohol consumption, and drug use among adolescents with psychiatric disorders compared with a population based sample. Journal of Adolescence, 37 (7) 1189-1199

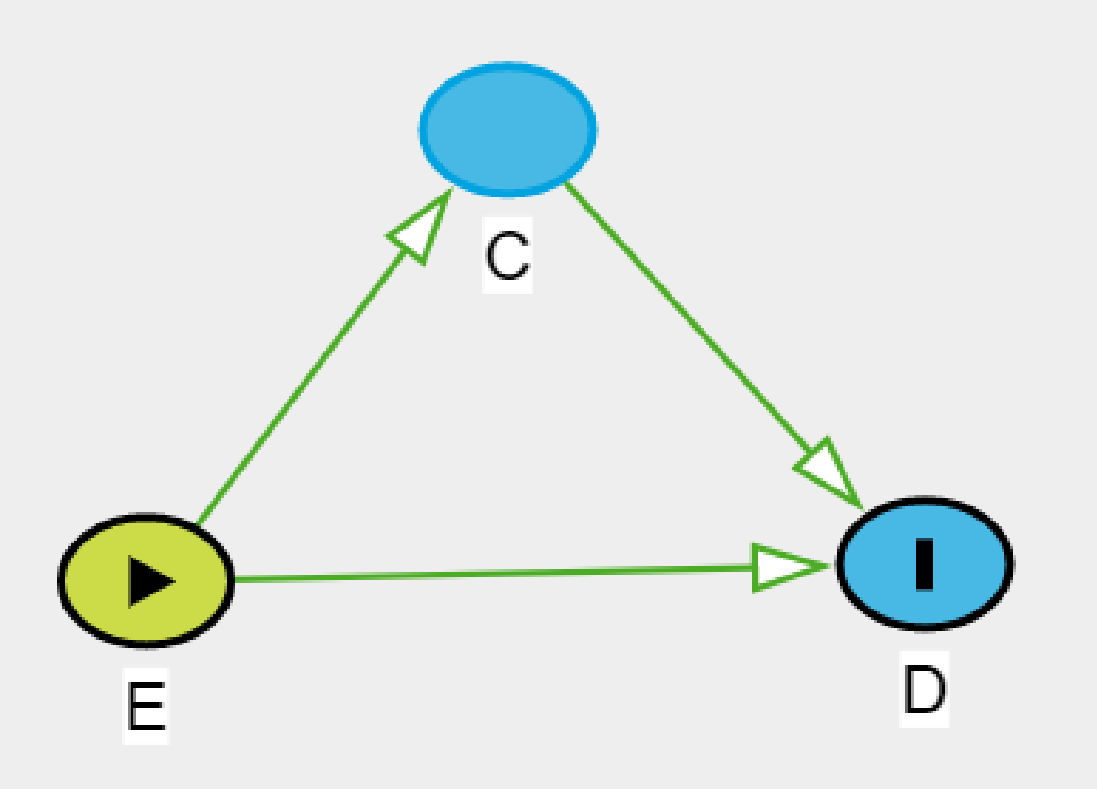
DAG·Collider

Adjust? Not adjust!



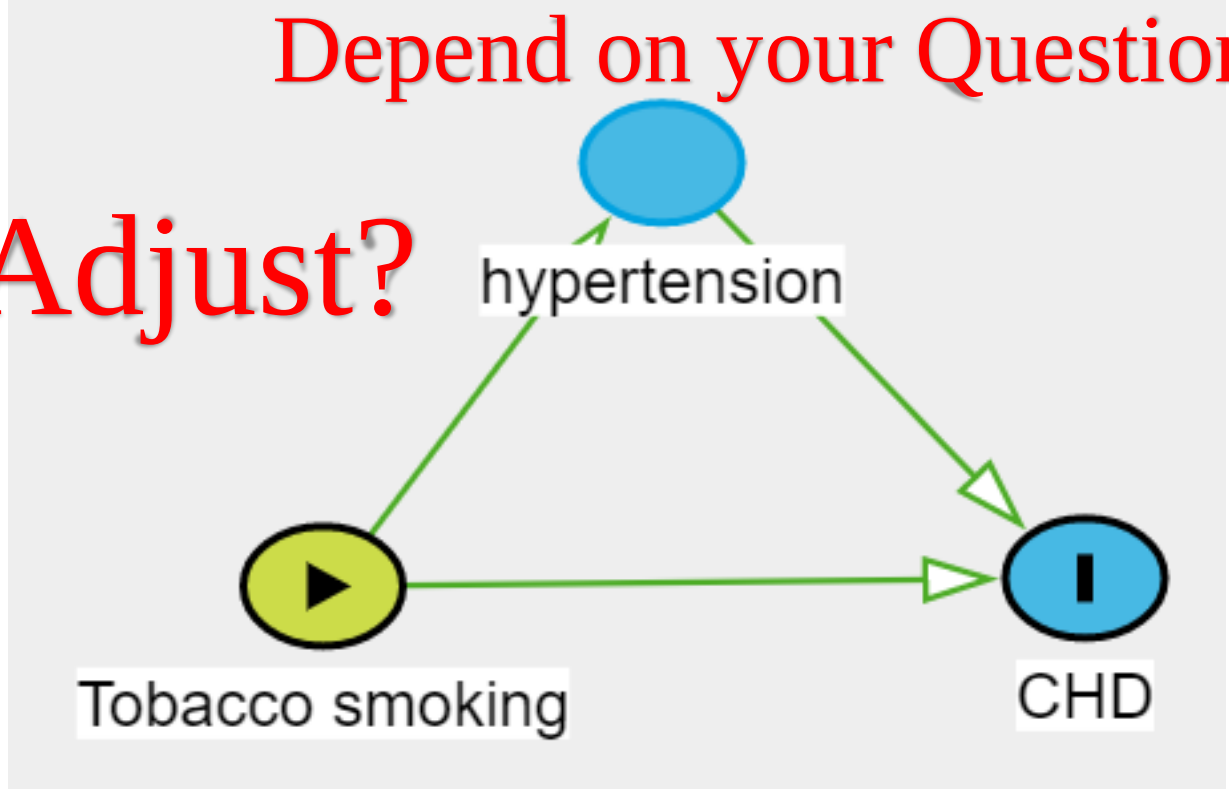


DAG·Mediator



Depend on your Question!

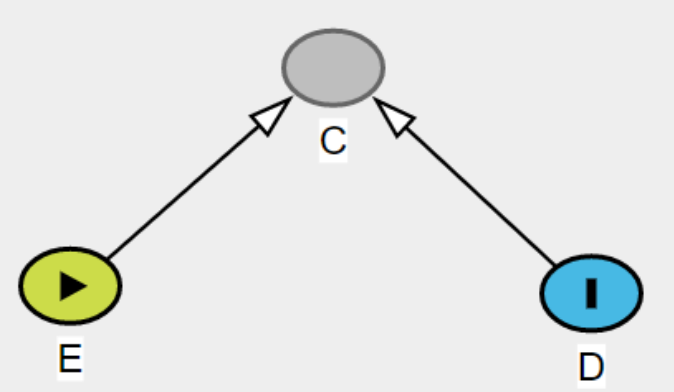
Adjust?

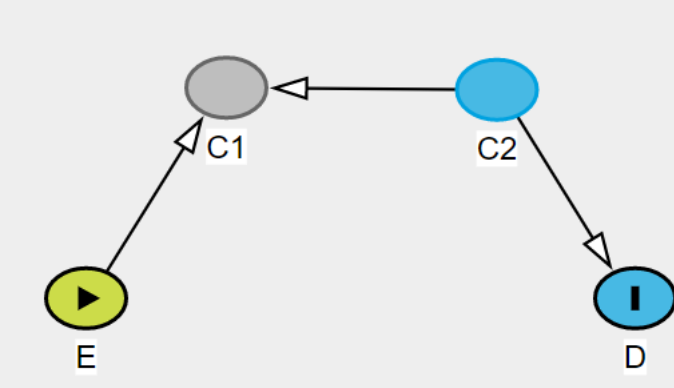


Graphical identification criteria

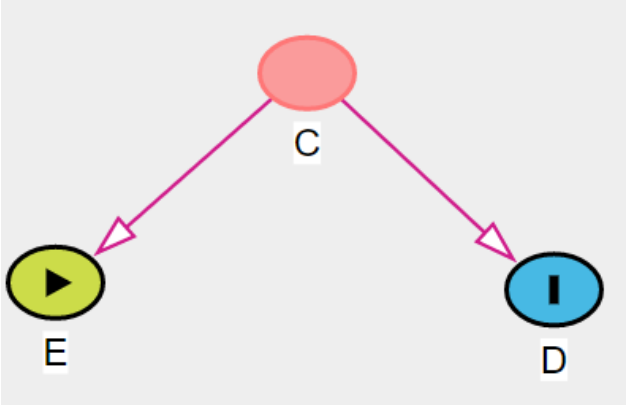
"Unblocked" paths & "Blocked" paths

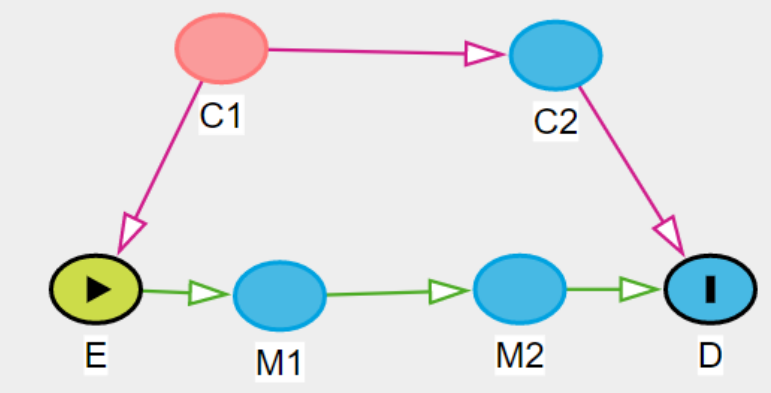
关闭路径（路径上存在碰撞变量）：碰撞变量两侧的变量之间无条件独立。





开放路径（路径上不存在碰撞变量）：该路径上的变量之间存在关联。



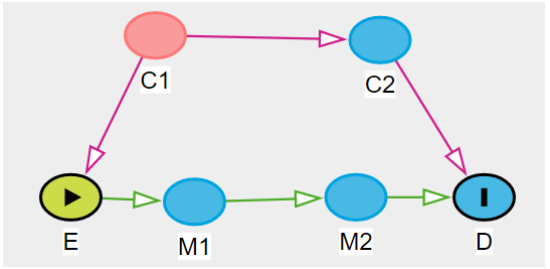


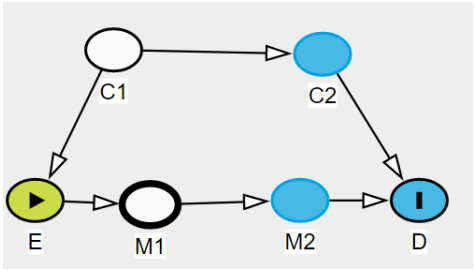
Graphical identification criteria

"Unblocked" paths & "Blocked" paths

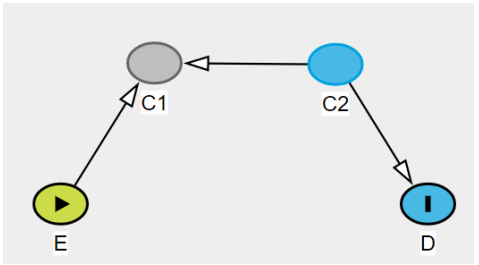
Three blocking criteria (key!!)

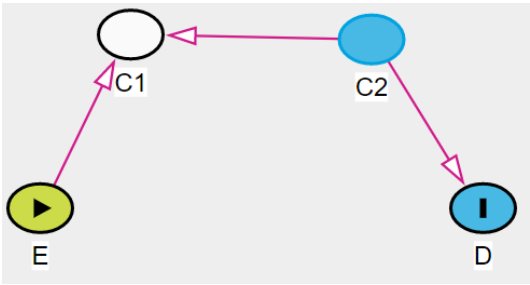
1. Conditioning on a non-collider blocks a path 控制非碰撞变量会关闭路径





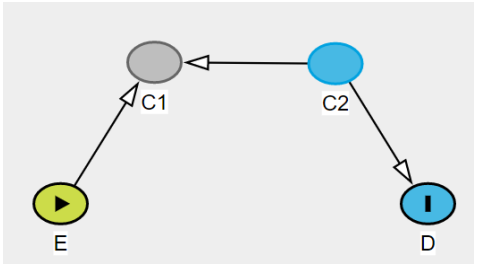
2. Conditioning on a collider, or a descendent of a collider, unblocks a path控制碰撞变量或者碰撞变量子变量，会打开路径



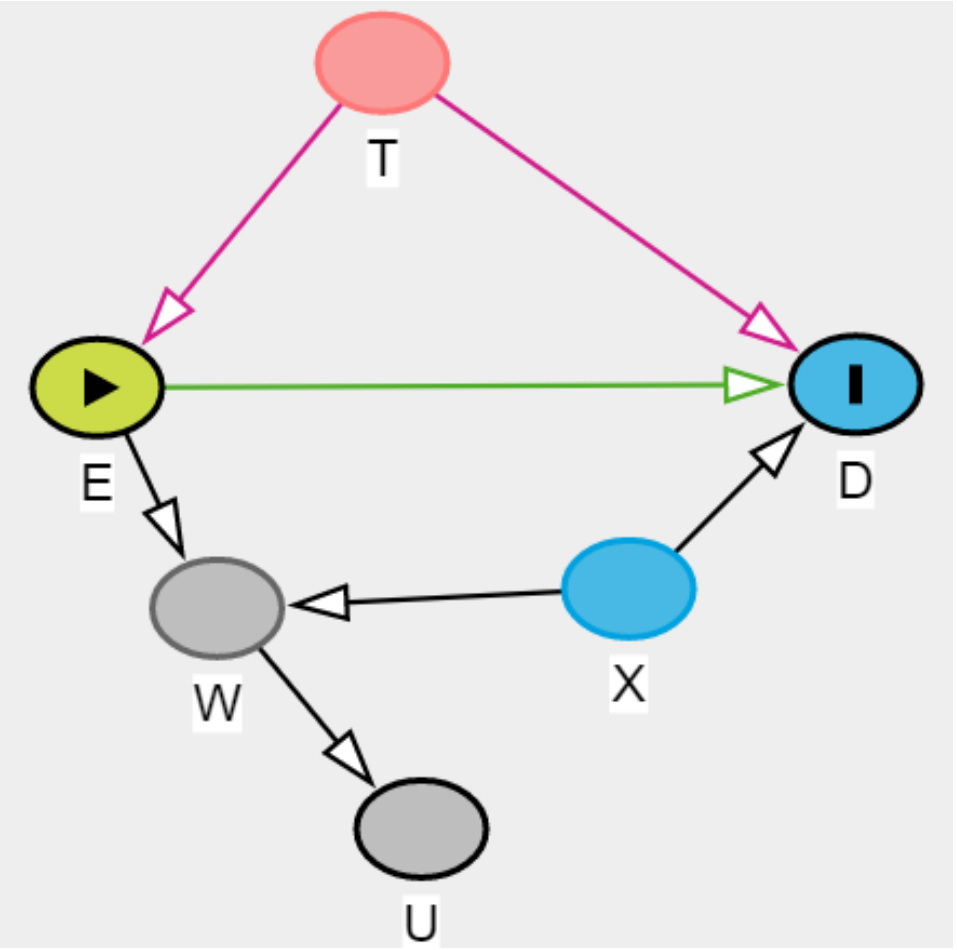


3. Not conditioning on a collider leaves a path "naturally" blocked.

不控制路径中的碰撞变量，该路径就是自然关闭的



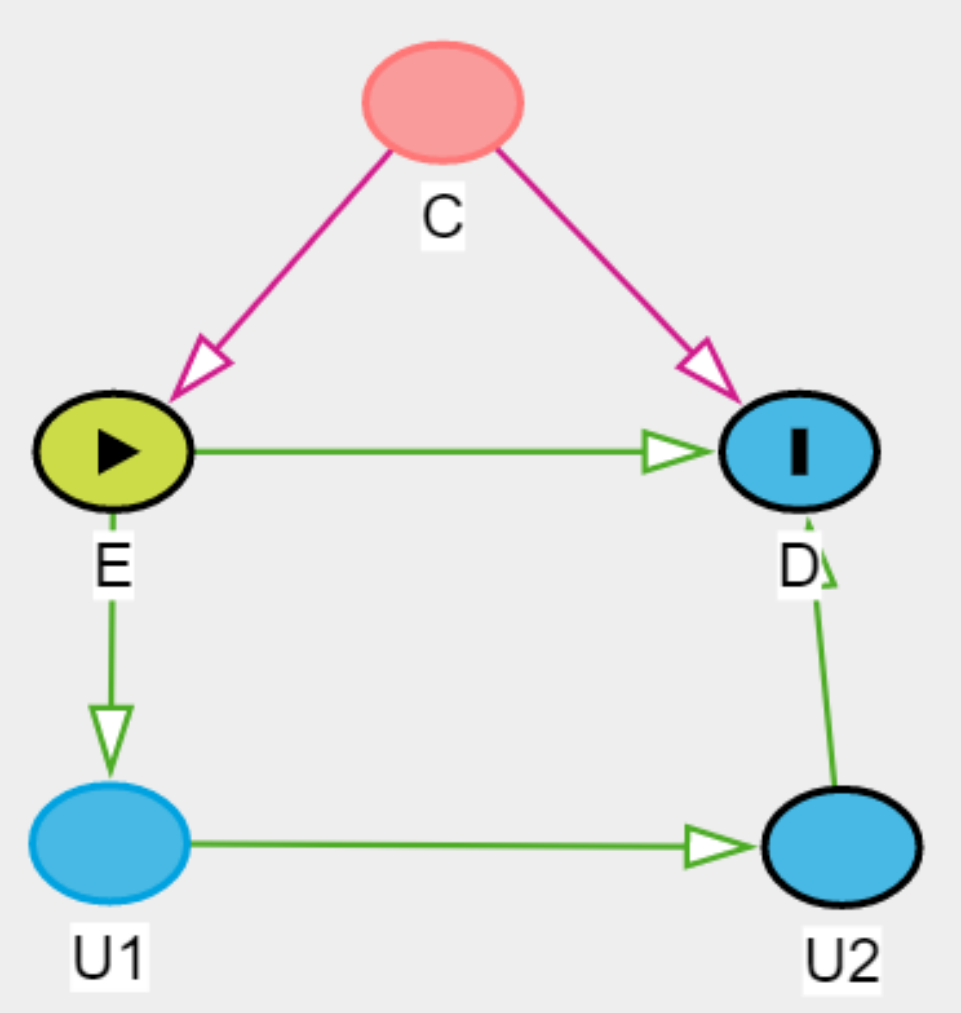
Graphical identification criteria
Backdoor criterion

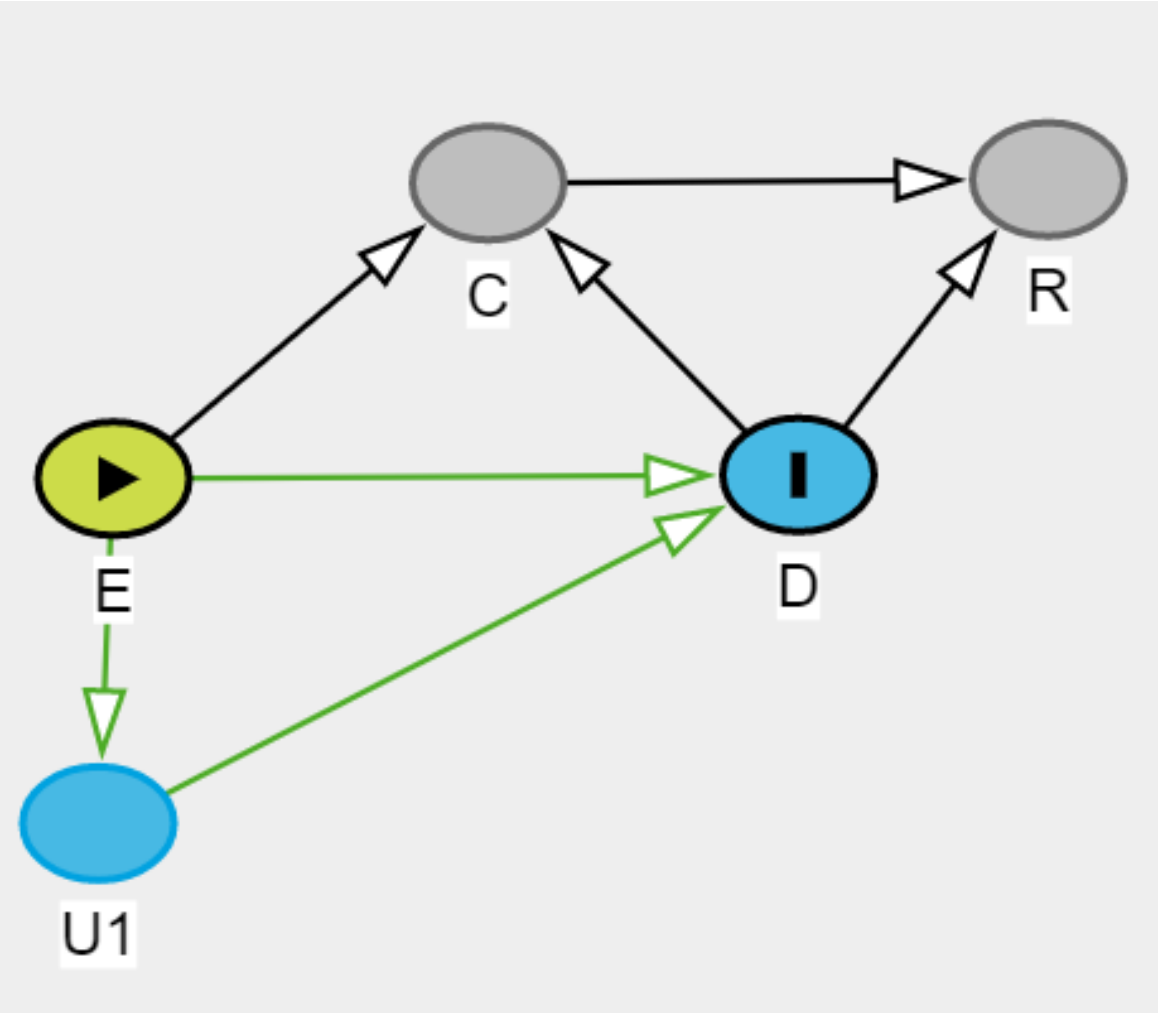


- Backdoor criteria: We need to block all backdoor paths between E and D.
- （后门准则：关闭所有开放的后门路径即可充分控制混杂）

- T, (X,T),(W,X,T),(U,X,T),(W,U,X,T)

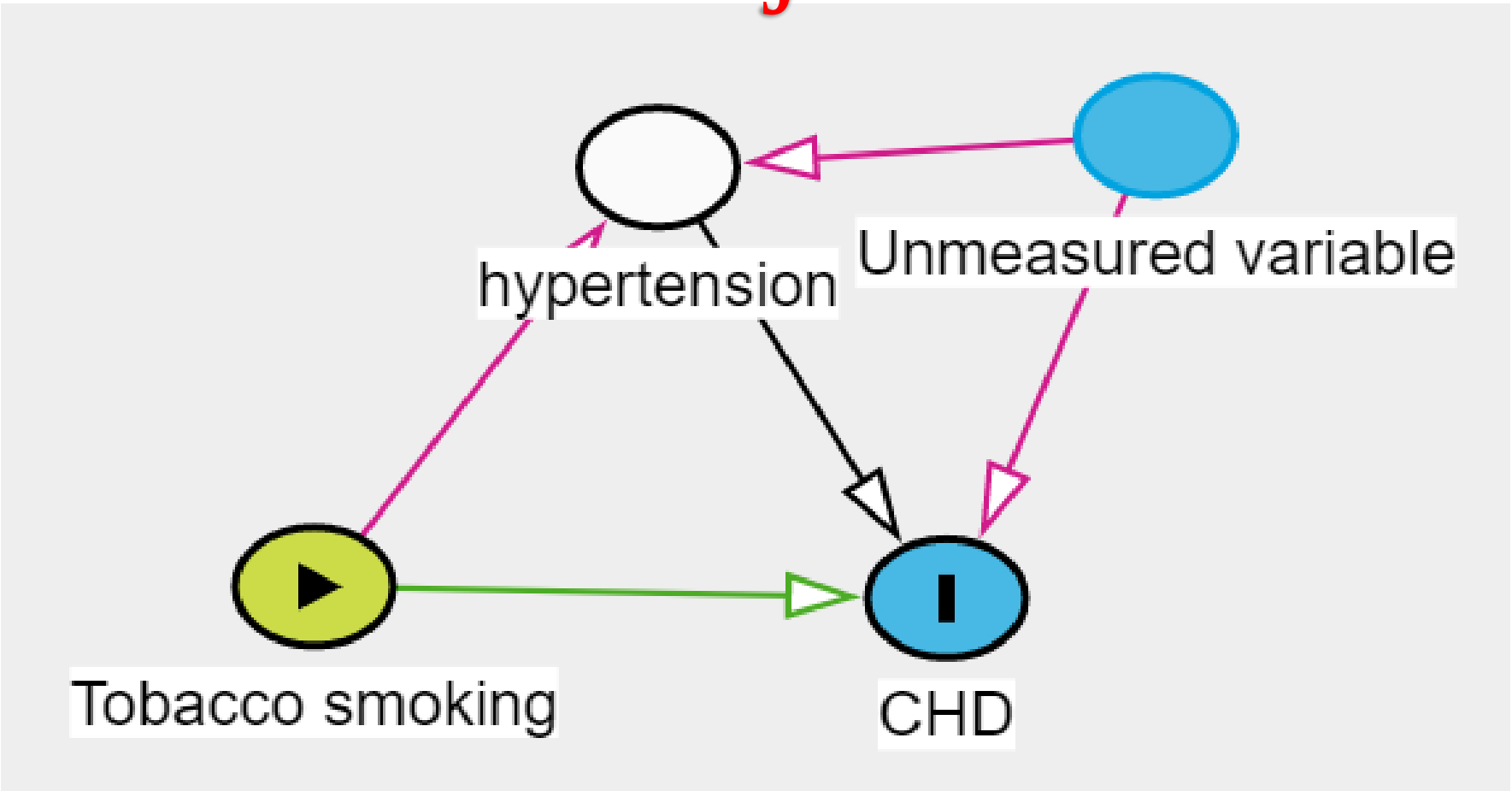
Backdoor criterion·Practice





DAG·Mediator

Adjust?



Ananth C V , Schisterman E F . Confounding, Causality and Confusion: The Role of Intermediate Variables in Interpreting Observational Studies in Obstetrics[J]. American Journal of Obstetrics and Gynecology, 2017:S0002937817305173.

DAG·Mediator

子痫前期→早产/低出生体重→新生儿脑性麻痹

thebmj

CrossMark

BMJ 2013;347:f4089 doi: 10.1136/bmj.f4089 (Published 9 July 2013)

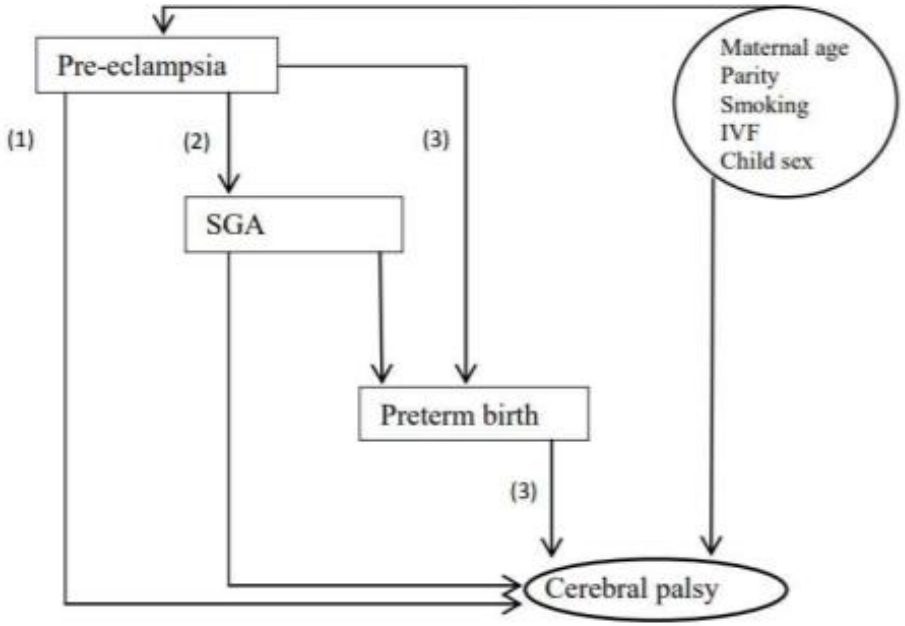
RESEARCH

Mediators of the association between pre-eclampsia and cerebral palsy: population based cohort study

Kristin Melheim Strand *medical student*[1], Runa Heimstad *senior consultant in obstetrics and gynaecology*[2], Ann-Charlotte Iversen *senior researcher*[1], Rigmor Austgulen *professor of paediatrics*[1], Stian Lydersen *professor of medical statistics*[3], Guro L Andersen *senior consultant in paediatrics*[4], Lorentz M Irgens *professor of preventive medicine*[5], Torstein Vik *professor of paediatrics*[6]

Supplementary figure 2. Proposed conseptual hierarchical framework for the relationship between pre-eclampsia and cerebral palsy (CP). Hypothesized causal pathways added in models 1-3, as well as potential confounders, are shown in the figure. Covariates in the models were:

Model 1: Pre-eclampsia
Model 2: Pre-eclampsia + small for gestational age (SGA)
Model 3: Pre-eclampsia + SGA + gestational age (GA)
Model 4: Pre-eclampsia + SGA + GA + Pre-eclampsia*GA



Strand K M , Heimstad R , Iversen A C , et al. Mediators of the association between pre-eclampsia and cerebral palsy: population based cohort study[J]. BMJ, 2013, 347(jul09 2):f4089-f4089.

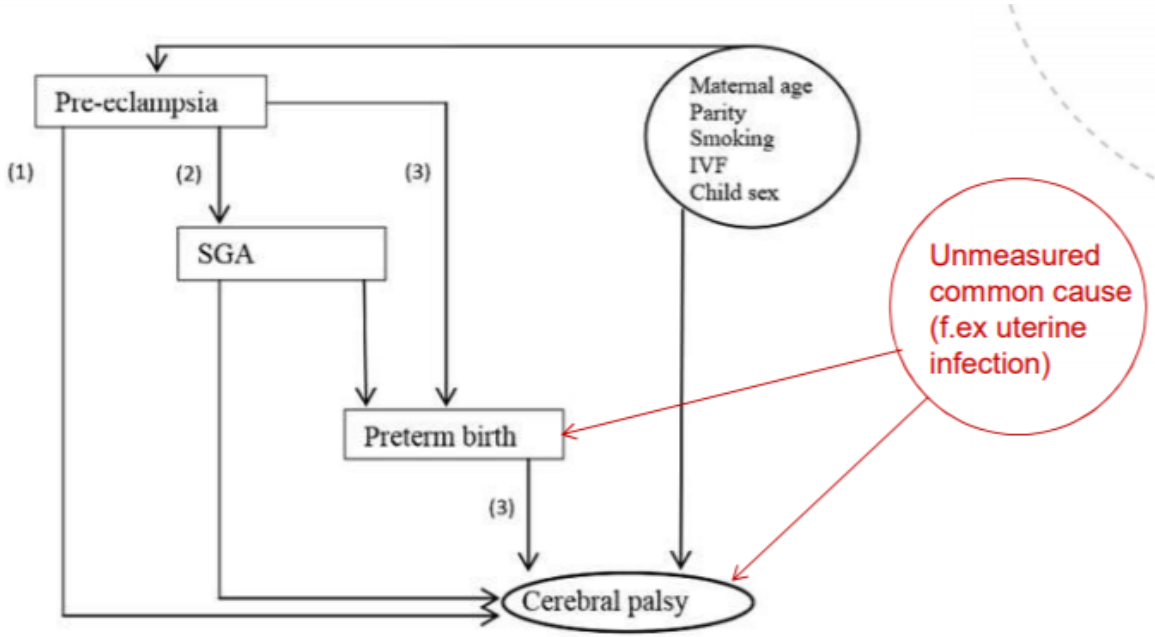
DAG·Mediator

子痫前期→早产/低出生体重→新生儿脑性麻痹

Table 2| **Unadjusted (model 1) and adjusted odds ratios for cerebral palsy after exposure to pre-eclampsia**

Potential mediators	Odds ratios (95% CI)		
	Model 1*	Model 2†	Model 3‡
Pre-eclampsia	2.52 (1.98 to 3.19)	2.14 (1.67 to 2.74)	0.73 (0.56 to 0.96)
Small for gestational age	—	2.30 (1.91 to 2.76)	1.90 (1.58 to 2.30)
Duration of pregnancy:			
37-40 weeks	—	—	1.00 (reference)
32-36 weeks	—	—	5.10 (4.18 to 6.20)
<32 weeks	—	—	40.71 (33.70 to 49.17)

*Unadjusted odds ratio for association between pre-eclampsia and cerebral palsy.
†Adjusted for small for gestational age.
‡Adjusted for small for gestational age and duration of pregnancy.

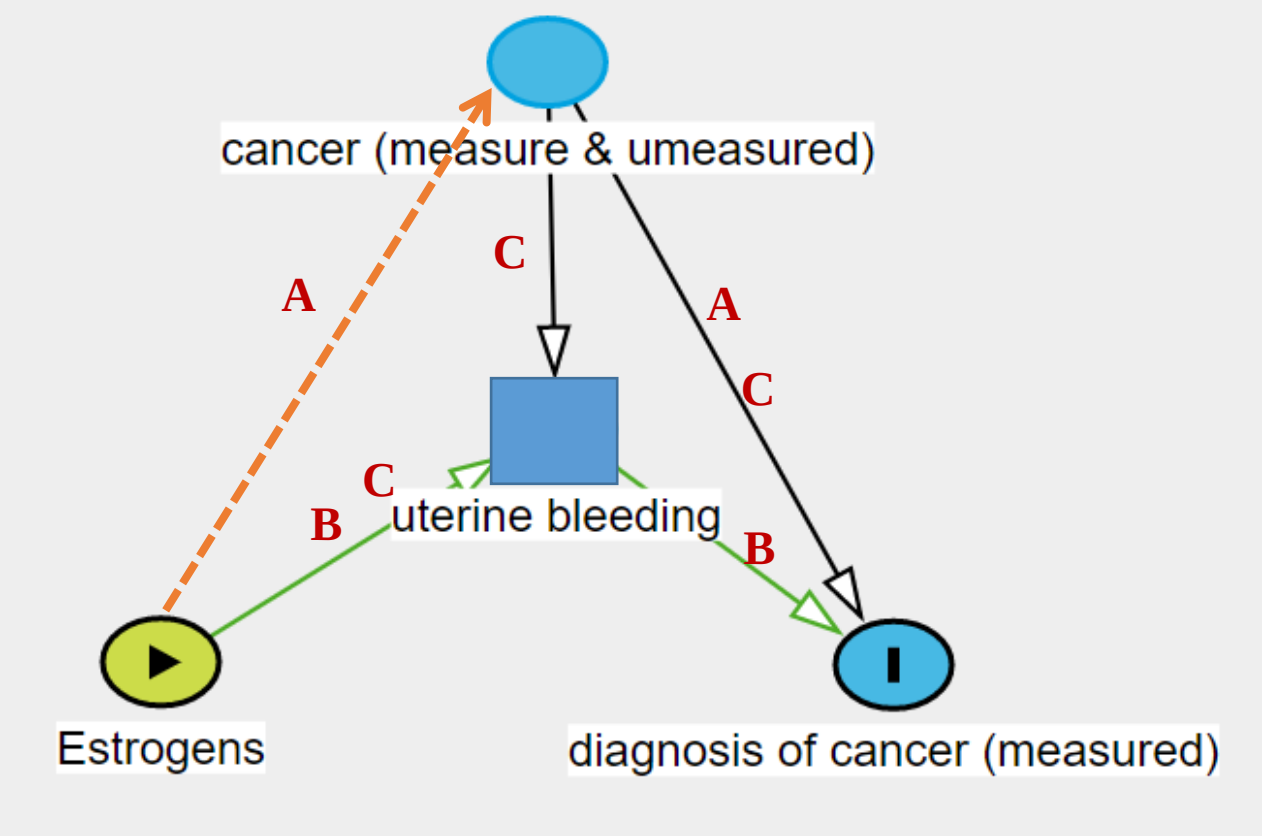


Strand K M , Heimstad R , Iversen A C , et al. Mediators of the association between pre-eclampsia and cerebral palsy: population based cohort study[J]. BMJ, 2013, 347(jul09 2):f4089-f4089.

举个栗子

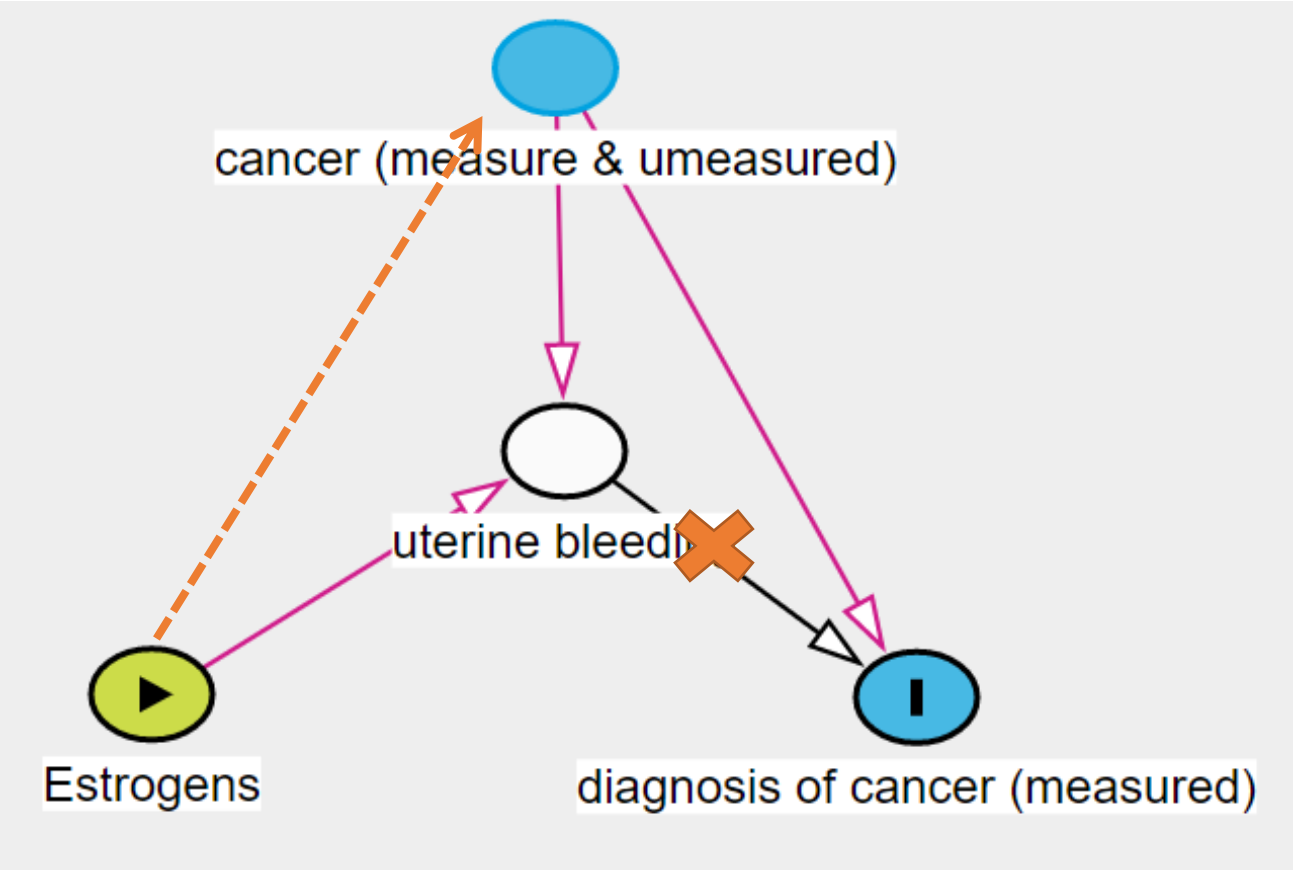
- 上世纪70年代，流行病学家发现服用雌激素的女性宫颈癌确诊率较高。
- 对此现象，有两个可能的解释：
    I. 雌激素导致了宫颈癌的发生
    II. 雌激素增加了子宫流血的风险，而宫颈癌也会导致子宫流血，促使了宫颈癌的发现和诊断

DAG·Collider bias



- 观察到的使用雌激素与被诊断子宫颈癌有关，有以下两种路径
  - A雌激素导致宫颈癌发生（真实因果）
  - B 雌激素导致子宫出血，子宫出血促使了宫颈癌的诊断（检出症候偏倚）
- 如果控制了子宫是否出血这一变量，可以阻断路径B，但又开放了路径C

DAG·Collider bias

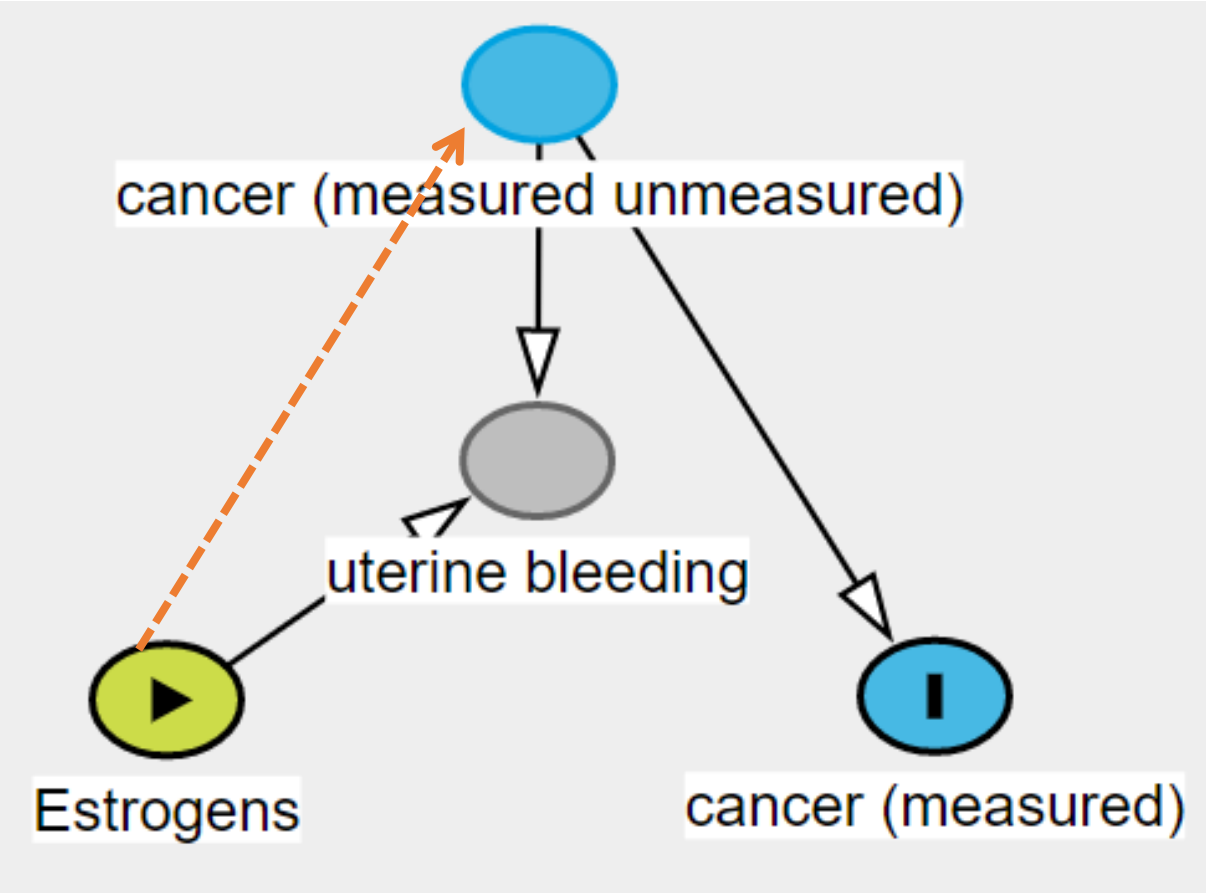


- 如何证实使用雌激素导致被诊断为子宫颈癌的机会增高

是因为

使用雌激素导致了子宫颈癌发病呢？ （如何避免检出症候偏倚？）

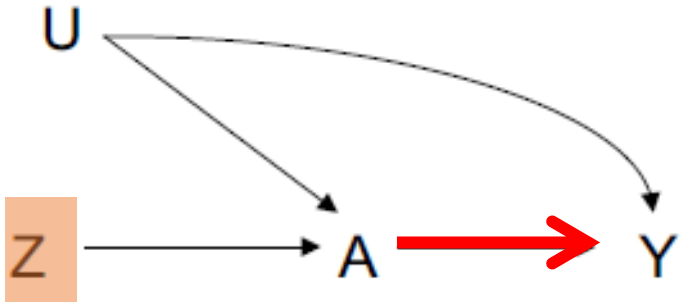
DAG·Collider bias

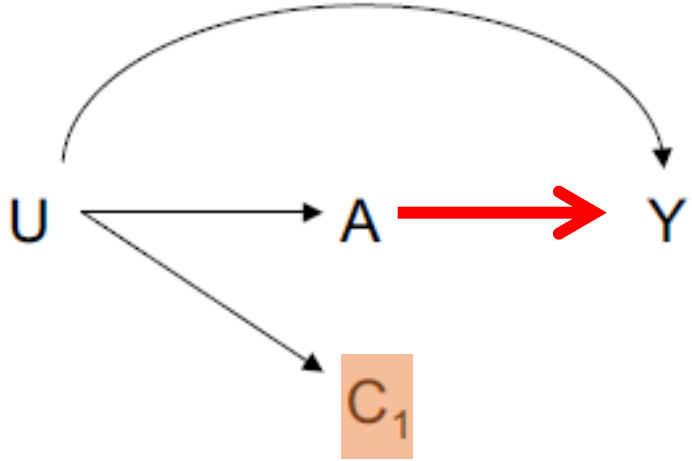


If the DAG is Not Fully Known

Principles of confounder selection

- Control for each covariate that is a cause of the exposure, or of the outcome, or of both;控制暴露或者结局或者二者共同的cause
- Exclude from this set any variable known to be an instrumental variable; 不要控制工具变量
- Include as a covariate any proxy for an unmeasured variable that is a common cause of both the exposure and the outcome.有未观测的混杂时，可以控制它的子变量

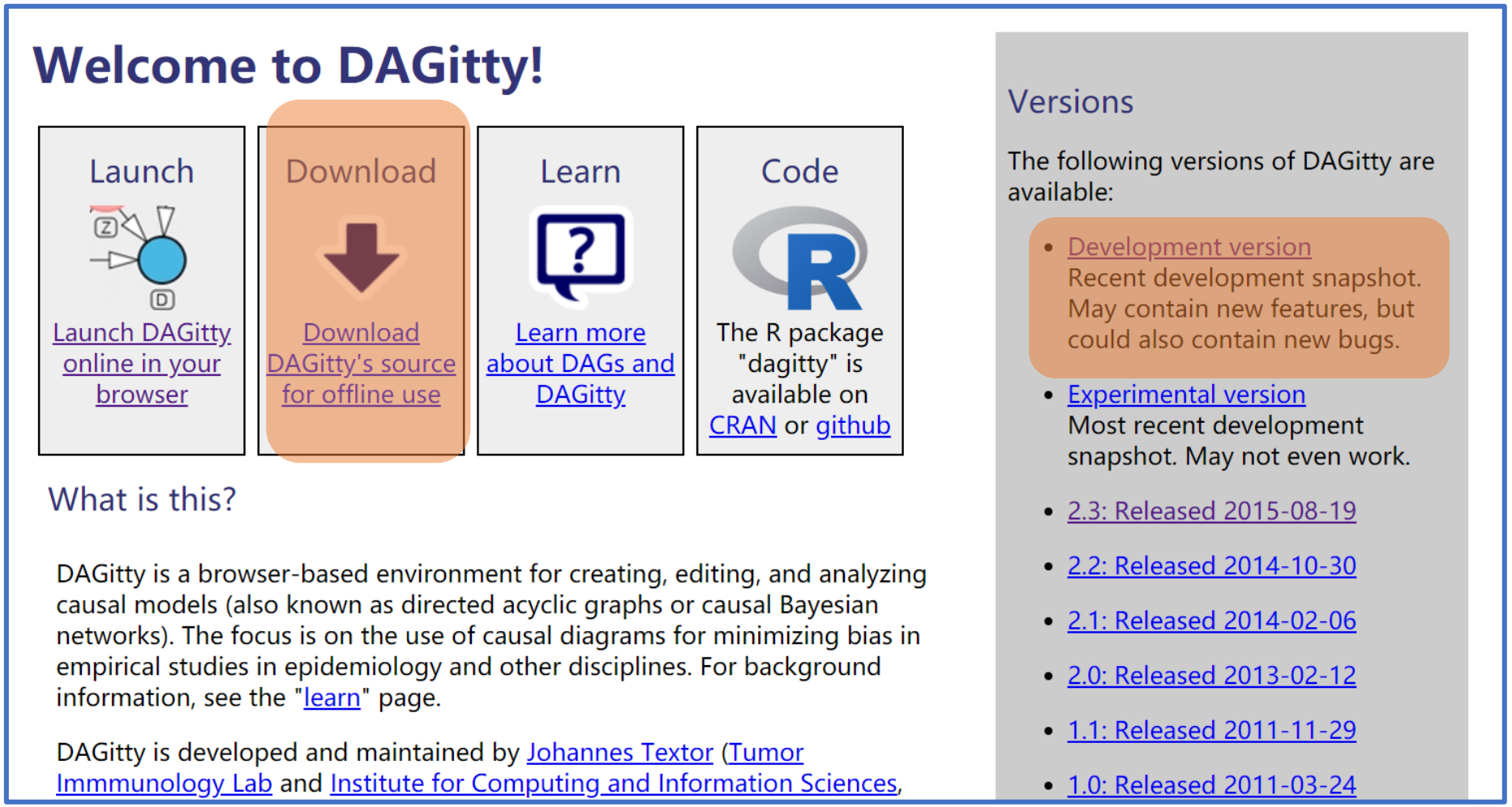




VanderWeele, T.J. Principles of confounder selection.Eur J Epidemiol (2019) 34: 211. https://doi.org/10.1007/s10654-019-00494-6

Create DAGs·Tools

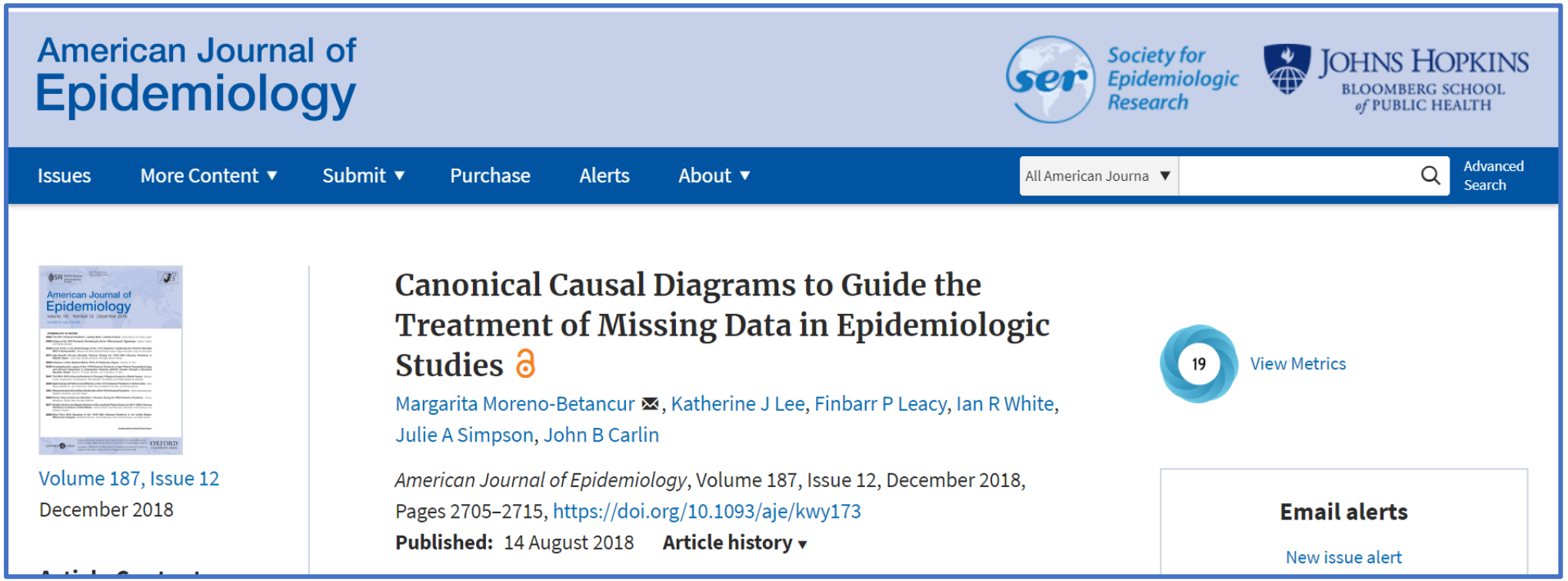
http://dagitty.net/



Textor, J., Hardt, J., Knuppel, S., 2011. DAGitty: a graphical tool for analyzing causal diagrams. Epidemiology 22, 745. https://doi.org/10.1097/EDE.0b013e318225c2be.

DAGs的其他应用

指导缺失数据的处理

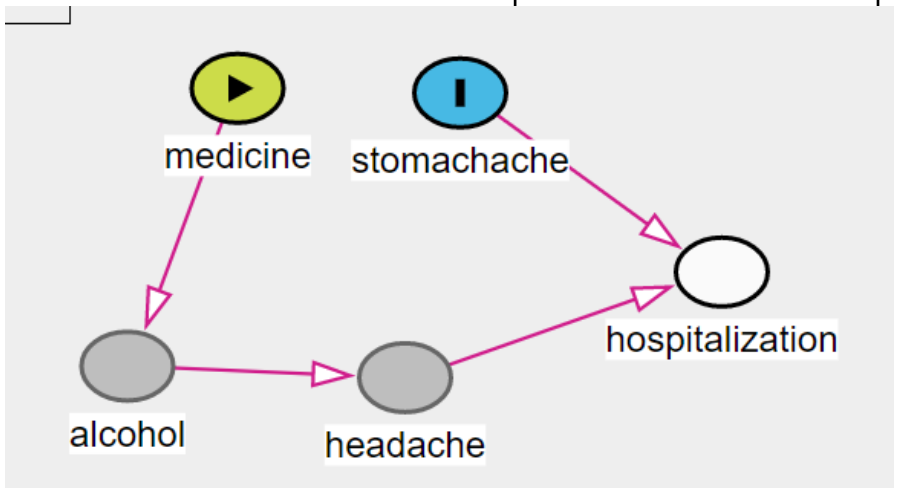


Margarita Moreno-Betancur, Katherine J Lee, Finbarr P Leacy, Ian R White, Julie A Simpson, John B Carlin, Canonical Causal Diagrams to Guide the Treatment of Missing Data in Epidemiologic Studies, American Journal of Epidemiology, Volume 187, Issue 12, December 2018, Pages 2705–2715, https://doi.org/10.1093/aje/kwy173

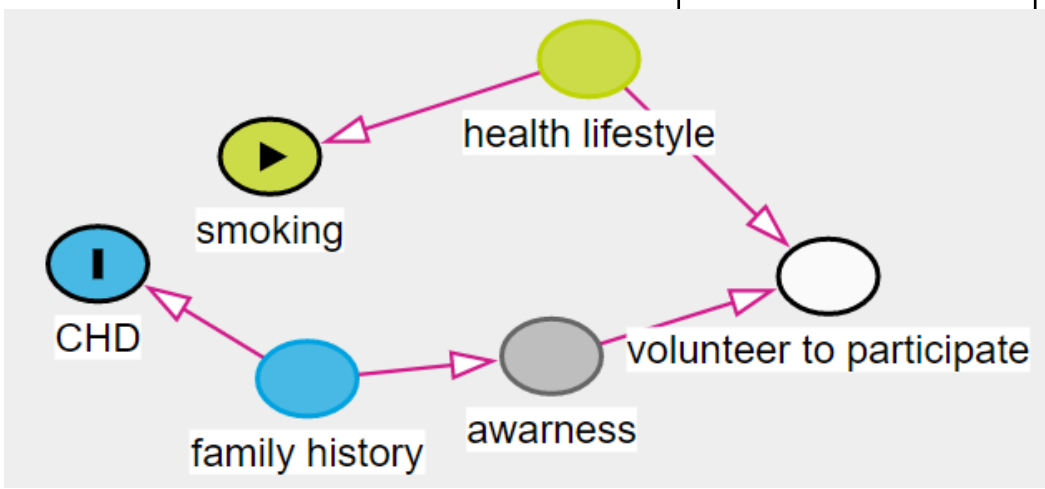
DAGs的其他应用

辅助教学：偏倚的理解

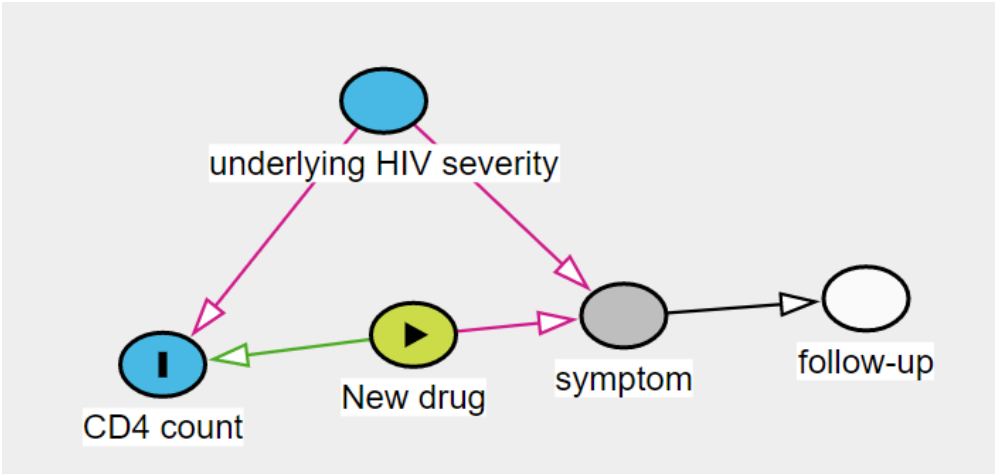
Berkson's Bias



Volunteer Bias



Differential Loss to Follow Up



Hernán, Miguel A, Hernández-Díaz, Sonia, Robins J M . A Structural Approach to Selection Bias[J]. Epidemiology, 2004, 15(5):615-625.

参考文献

- https://www.edx.org/course/causal-diagrams-draw-your-assumptions-before-your-conclusions-2
- https://www.coursera.org/learn/crash-course-in-causality#syllabus
- 向韧, 戴文杰, 熊元, et al. 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, 6(7):107-108.
- 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(08):1140-1144.
- 刘子言, 吴小丽, 解美秋, et al. 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(3).
- Strand K M , Heimstad R , Iversen A C , et al. Mediators of the association between pre-eclampsia and cerebral palsy: population based cohort study[J]. BMJ, 2013, 347(jul09 2):f4089-f4089.
- Cole SR, Platt RW, Schisterman EF, et al. Illustrating bias due to conditioning on a collider. Int J Epidemiol. 2010;39(2):417–420. doi:10.1093/ije/dyp334
- VanderWeele, T.J. Eur J Epidemiol (2019) 34: 211. https://doi.org/10.1007/s10654-019-00494-6

参考文献

- Strand K M , Heimstad R , Iversen A C , et al. Mediators of the association between pre-eclampsia and cerebral palsy: population based cohort study[J]. BMJ, 2013, 347(jul09 2):f4089-f4089.

- Karl DF, Mark MC, Srinivasa VK, et al. Evidence synthesis for constructing directed acyclic graphs (ESC-DAGs): a novel and systematic method for building directed acyclic graphs, International Journal of Epidemiology, 2019, dyz150, https://doi.org/10.1093/ije/dyz150

- Textor, J., Hardt, J., Knuppel, S., 2011. DAGitty: a graphical tool for analyzing causal diagrams. Epidemiology 22, 745. https://doi.org/10.1097/EDE.0b013e318225c2be

- Talbot, D. & Massamba, V.K. A descriptive review of variable selection methods in four epidemiologic journals: there is still room for improvement. Eur J Epidemiol (2019) 34: 725. https://doi.org/10.1007/s10654-019-00529-y

- Margarita Moreno-Betancur, Katherine J Lee, Finbarr P Leacy, Ian R White, Julie A Simpson, John B Carlin, Canonical Causal Diagrams to Guide the Treatment of Missing Data in Epidemiologic Studies, American Journal of Epidemiology, Volume 187, Issue 12, December 2018, Pages 2705–2715, https://doi.org/10.1093/aje/kwy173

- Hernán, Miguel A, Hernández-Díaz, Sonia, Robins J M . A Structural Approach to Selection Bias[J]. Epidemiology, 2004, 15(5):615-625.

Thanks!

Directed Acyclic Graph·Key player

Computer Science: Judea Pearl, Jin Tian, Thomas Verma

Philosophy: Peter Spirtes, Clark Glymour, Richard Scheines

Biostatistics: Jamie Robins, Sander Greenland, Tyler VanderWeele, Miguel Hernan

Create DAGs·Method

Mapping Translation Integration 1:synthesis Integration 2: recombination

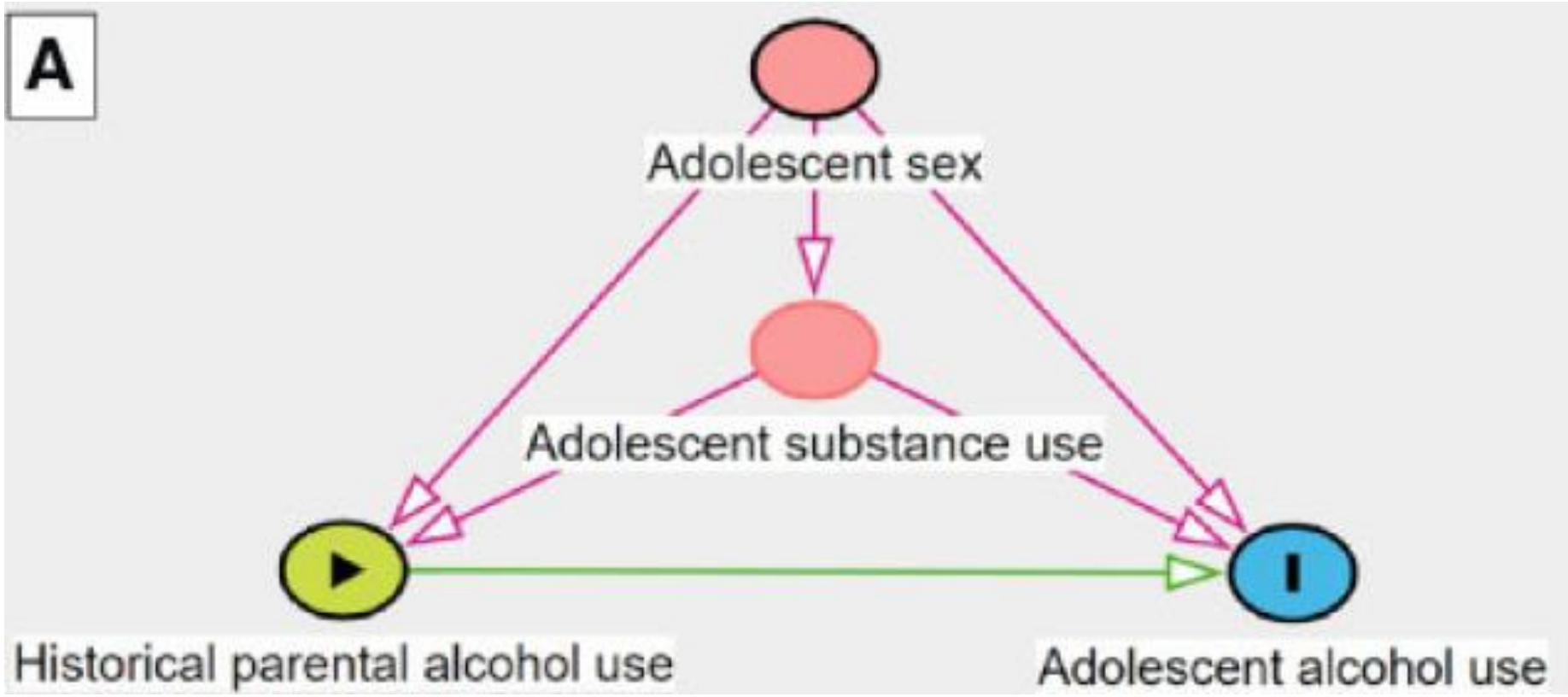
To produce a DAG that corresponds to the conclusions of the study under review (accurate or otherwise) and is ‘saturated’ with an edge between all pairs of nodes.

1. Outcome variable of interest is set as DAG outcome(s).
2. Exposure variable(s) of interest is set as DAG exposure(s).
3. A directed edge is drawn originating from the exposure(s), terminating at the outcome(s).
4. All control variables are entered as unassigned variables.
5. A directed edge is drawn originating from each control to the exposure(s) and outcome(s).
6. Mediators, instrumental variables etc. are mapped as per the study’s conclusions.
7. The ‘implied graph’ is saturated by drawing directed or undirected edges between all confounders (direction does not matter until the translation stage). The recombination process can be performed at this stage to help simplify an overly complex IG.

Karl DF, Mark MC, Srinivasa VK, et al. Evidence synthesis for constructing directed acyclic graphs (ESC-DAGs): a novel and systematic method for building directed acyclic graphs, *International Journal of Epidemiology*, 2019, dyz150, https://doi.org/10.1093/ije/dyz150

Create DAGs·Method

Mapping Translation Integration 1:synthesis Integration 2: recombination



Implied graph for hypothetical study

Karl DF, Mark MC, Srinivasa VK, et al. Evidence synthesis for constructing directed acyclic graphs (ESC-DAGs): a novel and systematic method for building directed acyclic graphs, *International Journal of Epidemiology*, 2019, dyz150, https://doi.org/10.1093/ije/dyz150

Create DAGs·Method

Mapping **Translation** Integration 1:synthesis Integration 2: recombination

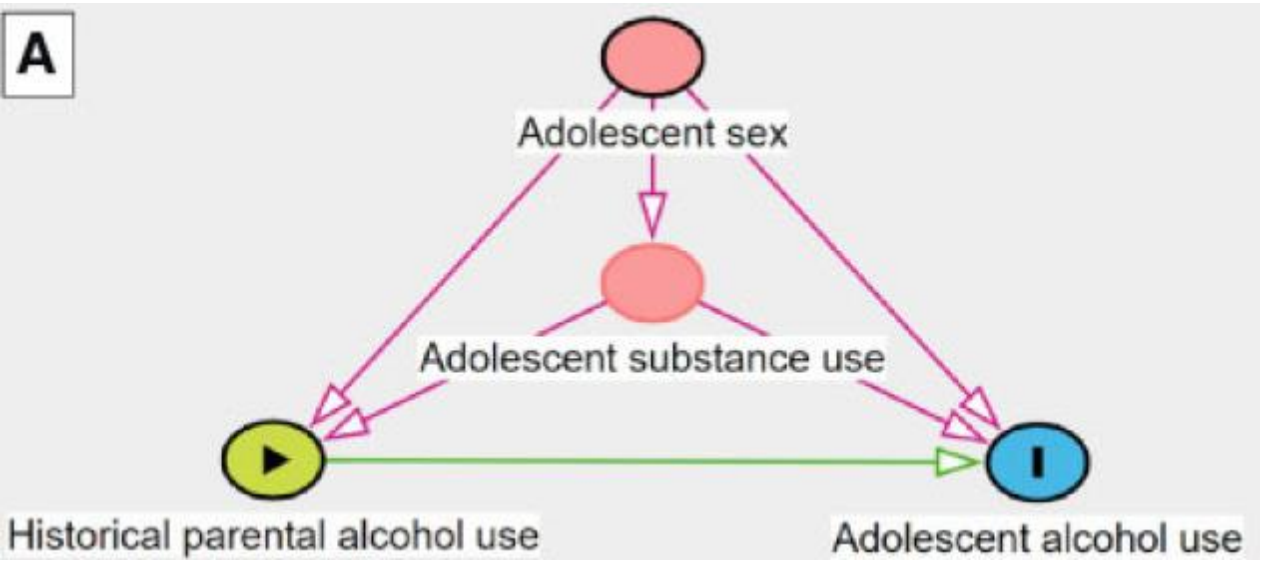
Each relationship in the IG is assessed under sequential causal criteria and a counterfactual thought experiment

1.Temporality—does the posited cause precede effect? (If 'yes', proceed to next criterion. If not, assess reverse relationship.)
2.Face-validity—is the posited relationship plausible? (If 'yes', proceed to next criterion. If not, assess reverse relationship.)
3.Recourse to theory—is the posited relationship supported by theory? (Always proceed to the counterfactual thought experiment.)
4.Counterfactual thought experiment—is the posited relationship supported by a systematic thought experiment informed by the POF? (Once completed, always assess the reverse relationship unless already assessed.)

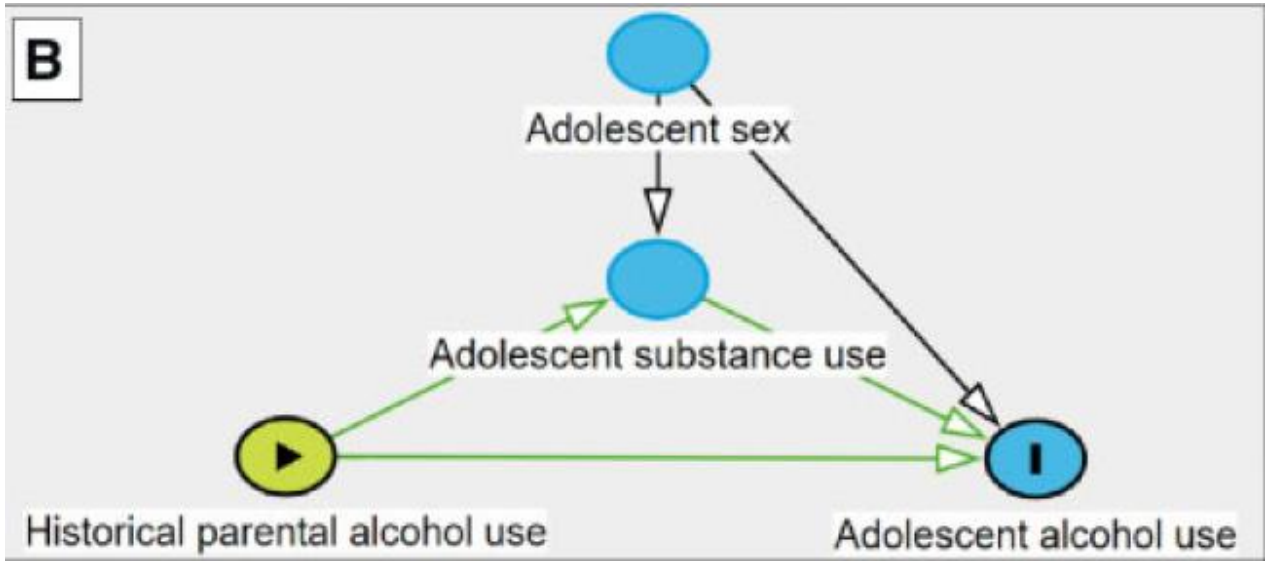
Karl DF, Mark MC, Srinivasa VK, et al. Evidence synthesis for constructing directed acyclic graphs (ESC-DAGs): a novel and systematic method for building directed acyclic graphs, *International Journal of Epidemiology*, 2019, dyz150, https://doi.org/10.1093/ije/dyz150

Create DAGs·Method

Mapping **Translation** Integration 1:synthesis Integration 2: recombination



Implied graph for hypothetical study



DAG for hypothetical study

Karl DF, Mark MC, Srinivasa VK, et al. Evidence synthesis for constructing directed acyclic graphs (ESC-DAGs): a novel and systematic method for building directed acyclic graphs, *International Journal of Epidemiology*, 2019, dyz150, https://doi.org/10.1093/ije/dyz150

Create DAGs·Method

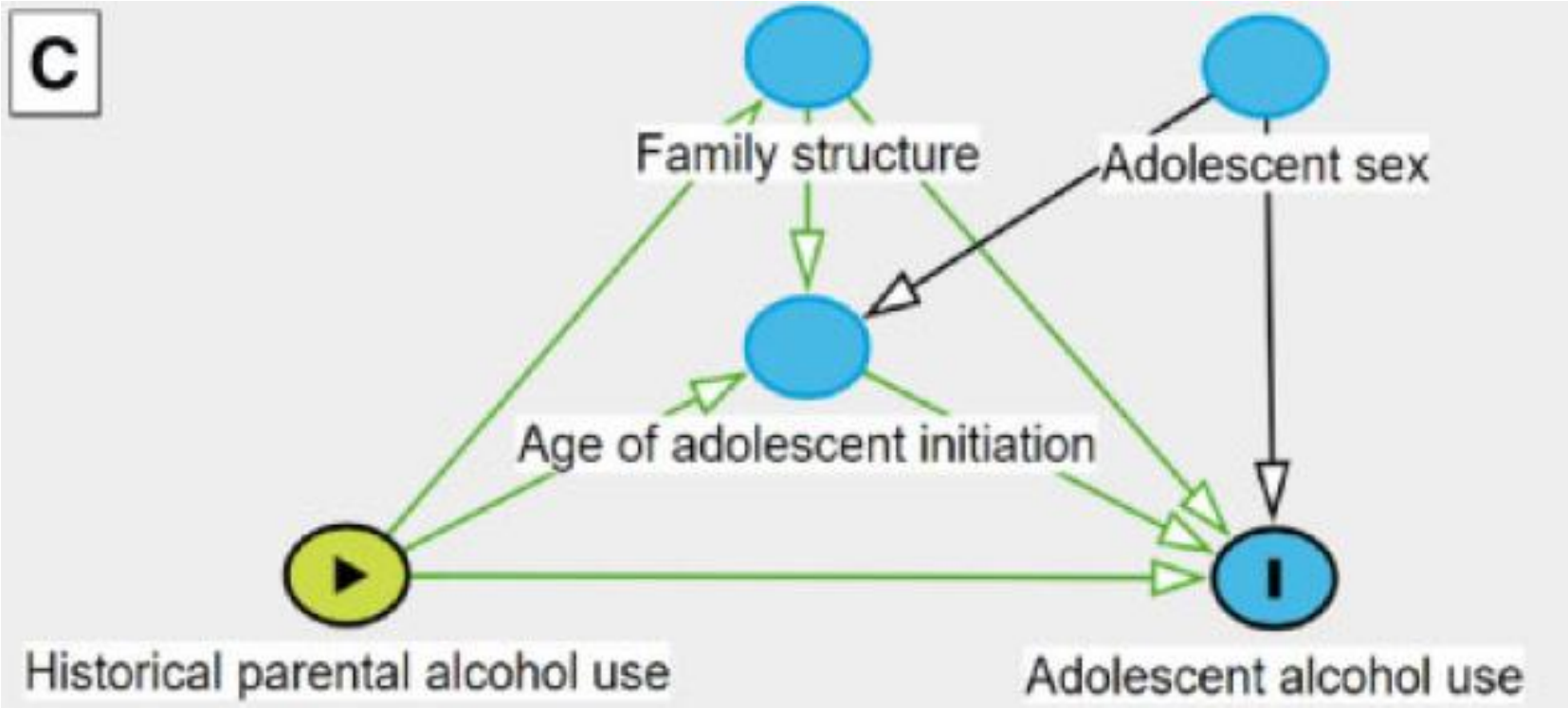
Mapping Translation **Integration 1:synthesis** Integration 2: recombination

1.A new DAG is created to serve as the integrated DAG (I-DAG).
2.The focal relationship is added to the I-DAG (as per mapping steps 1–3).
3.Each indexed directed edge pertaining to the focal relationship (including its corresponding node) is added to the diagram.
4.Each indexed directed edge pertaining to other nodes is added (e.g. between confounders).
5.Conceptually similar nodes should be grouped together in virtual space to aid the recombination process.

Karl DF, Mark MC, Srinivasa VK, et al. Evidence synthesis for constructing directed acyclic graphs (ESC-DAGs): a novel and systematic method for building directed acyclic graphs, *International Journal of Epidemiology*, 2019, dyz150, https://doi.org/10.1093/ije/dyz150

Create DAGs·Method

Mapping Translation **Integration 1:synthesis** Integration 2: recombination



DAG for the study by Seljamo et al

Karl DF, Mark MC, Srinivasa VK, et al. Evidence synthesis for constructing directed acyclic graphs (ESC-DAGs): a novel and systematic method for building directed acyclic graphs, *International Journal of Epidemiology*, 2019, dyz150, https://doi.org/10.1093/ije/dyz150

Create DAGs·Method

Mapping Translation Integration 1:synthesis **Integration 2: recombination**

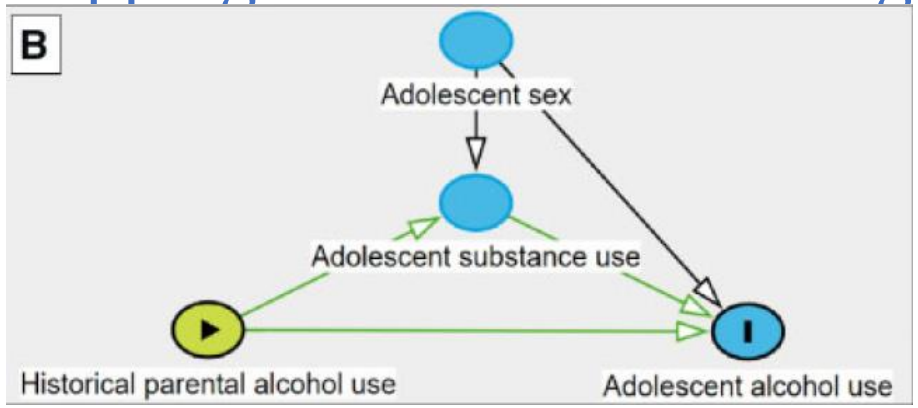
To combine nodes for either practical reasons (i.e. to reduce complexity) or substantive reasons (i.e. to establish consistency).

1.Is there theoretical support for combining two variables/nodes?
2.Do the conceptually related nodes have similar inputs and outputs (i.e. do they 'send to' and 'receive from' the same nodes)?

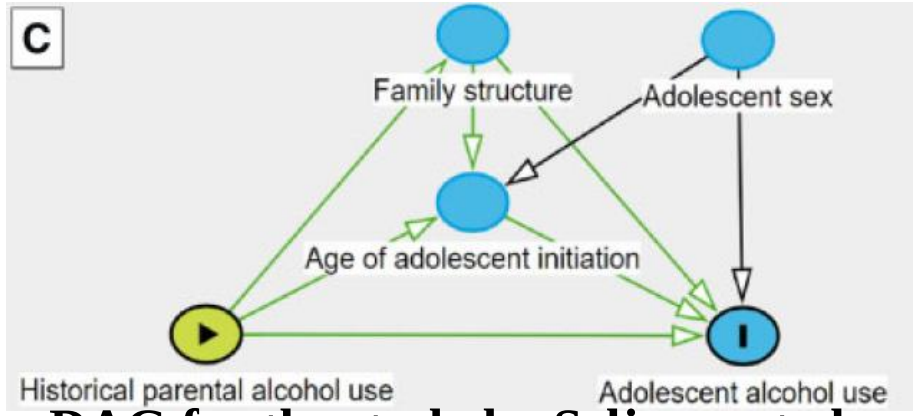
Karl DF, Mark MC, Srinivasa VK, et al. Evidence synthesis for constructing directed acyclic graphs (ESC-DAGs): a novel and systematic method for building directed acyclic graphs, *International Journal of Epidemiology*, 2019, dyz150, https://doi.org/10.1093/ije/dyz150

Create DAGs·Method

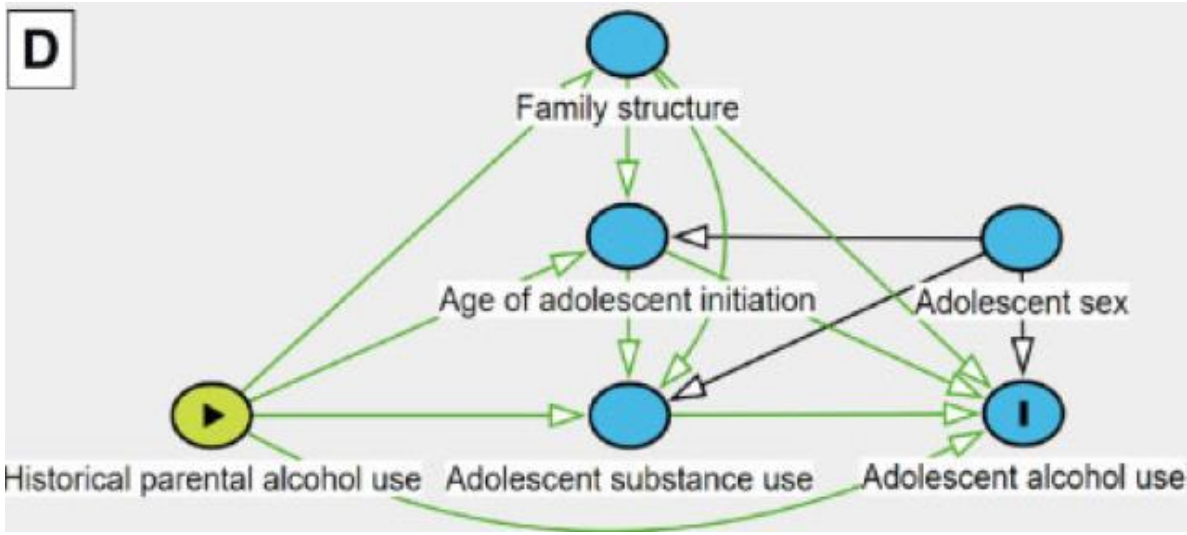
Mapping Translation Integration 1:synthesis **Integration 2: recombination**



DAG for hypothetical study



DAG for the study by Seljamo et al



Integratd DAG for hypothetical study and the study by Seljamo et al

Karl DF, Mark MC, Srinivasa VK, et al. Evidence synthesis for constructing directed acyclic graphs (ESC-DAGs): a novel and systematic method for building directed acyclic graphs, *International Journal of Epidemiology*, 2019, dyz150, https://doi.org/10.1093/ije/dyz150

Effect modification or Interaction

- VanderWeele, Tyler J . On the Distinction Between Interaction and Effect Modification[J]. Epidemiology, 2009, 20(6):863-871.

- Weinberg, Clarice R . Can DAGs Clarify Effect Modification?[J]. Epidemiology, 2007, 18(5):569-572.

- Tyler VanderWeele;James Robins. Four Types of Effect Modification: A Classification Based on Directed Acyclic Graphs. Epidemiology, 200718(5):561-568.