# 混杂因素的定义与识别

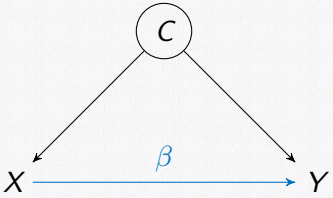
战义强

公共卫生学院（深圳）
中山大学

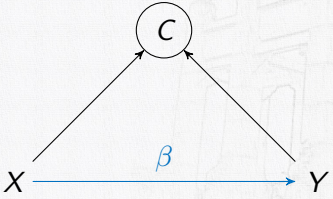
2025 年春

# 混杂 & 混杂因素



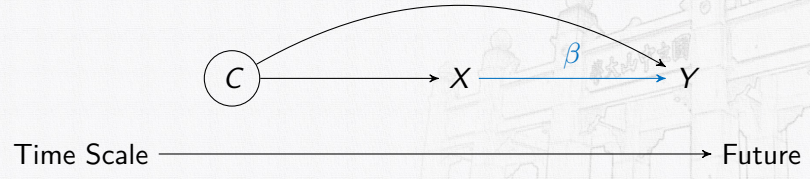



- $C$ 与 $X$ 有关
- $C$ 与 $Y$ 有关
- $C$ 不是 $X$ 和 $Y$ 的中介因子

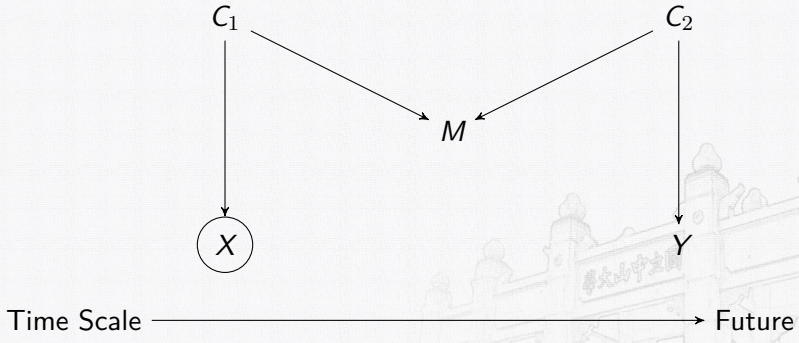


# 混杂 & 混杂因素

- 混杂是一种系统偏倚 bias，其歪曲了暴露和结局之间关系的真实联系
- 混杂又被称为共同的病因 common (shared) cause.
- 导致混杂的变量称为混杂因素 confounders.
- 在因果推断领域，观察到到 $X$ 和 $Y$ 之间到关系，通常分为两部分：真正的因果效应 $\beta$ ($X \rightarrow Y$) 和非因果效应 ($X \leftarrow C \rightarrow Y$)

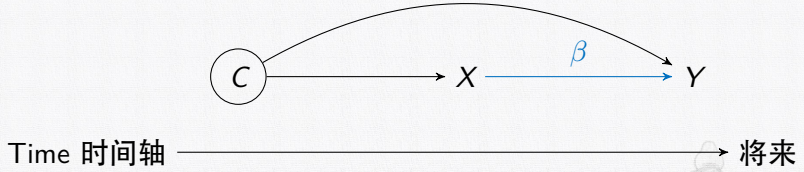


# 混杂 & 混杂因素



# Directed Acyclic Graph (单向无环图/图模型)

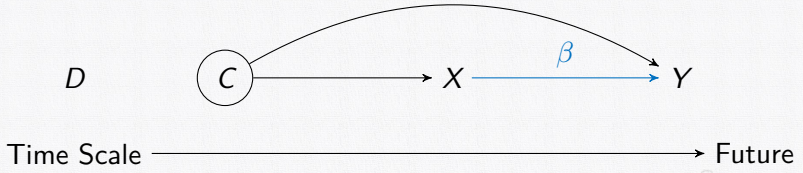





- DAG 单向无环图可以用来辅助确定哪些因素是混杂因素（哪些因素需要在模型中进行调整）
- 它以图形的形式展示了潜在的因果结构（因果链）
- 在变量选择方面，DAG 已经植入了我们的先验认知（既往知识）

# DAG 相关术语

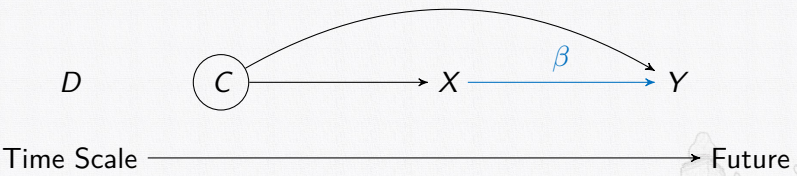





- 箭头 arrow：每个箭头代表一个因果效应；如果没有画箭头，就代表没有因果效应
- 单向 Directed：两个变量（节点 node）之间连线的箭头，代表因果关系的方向
- 无环 Acyclic：从某一个变量（节点 node）沿着箭头方向往前走，永远不会再返回到这个变量（节点 node）；过去影响将来，但是将来不影响过去（时间轴往前走）

# DAG: 单向



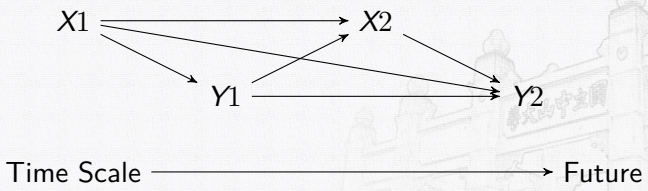



- $X$ 到 $Y$ 的箭头意味着 $X$ 可能影响 $Y$, 而不是 $Y$ 影响 $X$
- $X$ 和 $Y$ 之间有箭头意味着 $X$ 可能影响或者不影响 $Y$
- $D$ 和 $C, X, Y$ 之间没有箭头意味着 $D$ 不可能影响 $C, X, Y$ 中的任何一个

# DAG: 无环

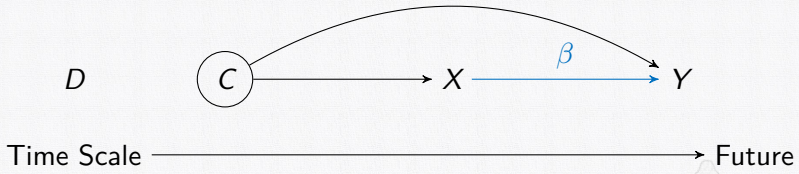


- 无环意味着一个变量不可能影响它自身，例如：我今天的体重不能影响我今天的体重
- 但是，我今天的体重肯定会影响我明天的体重
- 这个随时间变化的过程可以使用如下图示来展示：

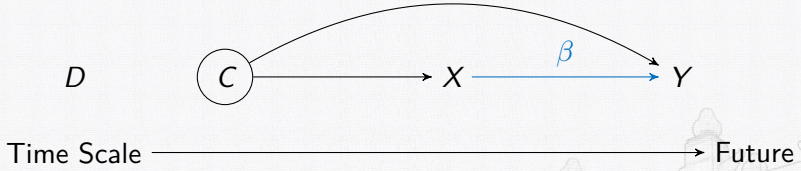


# DAG: 共同的病因





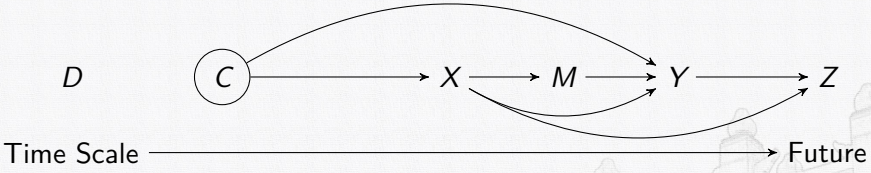

- $C$ 存在，意味着 $X$ 和 $Y$ 有共同的病因- $C$.
- $C$ 和 $X$ 没有共同的病因.
- $C$ 和 $Y$ 没有共同的病因.
- 共同的病因 Common causes 又被称为 shared causes (遗传学)；也被称为混杂因素.




- 一个变量 $V$ 的祖先是所有直接或者间接影响 $V$ 的变量
- 一个变量 $V$ 的后代是所有直接或者间接被 $V$ 影响的变量

# DAG: Paths 路径

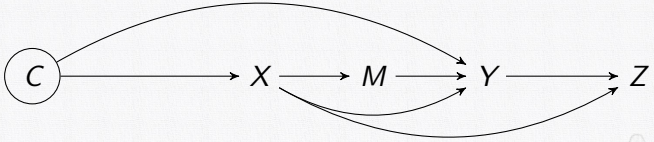





- 一条路径是两个变量之间的路线；不一定必须是沿着箭头的方向
- $X$ 和 $Y$ 之间的路径是什么?

# DAG: 路径

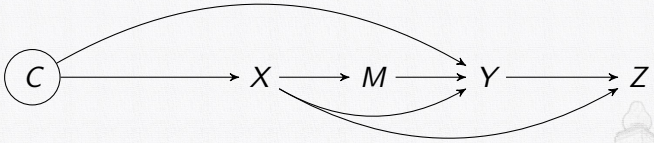





- $X \to Y$
- $X \to M \to Y$
- $X \leftarrow C \to Y$
- $X \to Z \leftarrow Y$

# DAG: 因果路径 Causal Paths

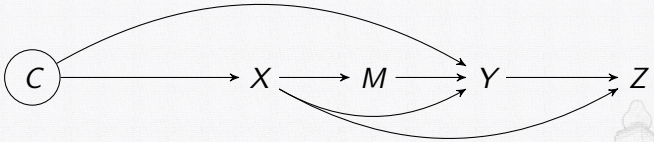





- 因果路径 causal path 两个变量之间沿着箭头方向的路线
- $X$ 和 $Y$ 之间的因果路径是什么?

# DAG: 因果路径 Causal Paths

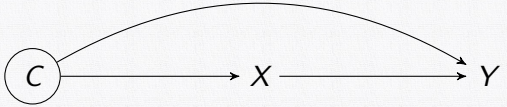





- $X \to Y$
- $X \to M \to Y$

# DAG: Blocking of Paths 路径的阻断






- 所有的路径（因果路径和非因果路径）可以根据两条判断规则来决定他们是开放路径 open paths 还是阻断路径 blocked paths

# DAG: 阻断规则 Blocking Rules



- 如果有一个变量 $V$ 位于一个链条 chain 上 $\to V \to$ 或者叉子 fork 上 $\leftarrow V \to$，并且控制 control 了 $V$, 那么这条路径就被阻断了.
- 如果有一个变量 $V$ 位于倒叉子"inverted fork" $\to V \leftarrow$ 上, 并且没有控制 $V$ 和它的任何后代，那么这条路径就被阻断了.

# 举例




- 控制 $A$ 阻断了 $X$ 到 $Y$ 之间的路径 (规则 1)
- 控制 $B$ 打开了 $X$ 到 $Y$ 之间的路径 (规则 2)
- 同时控制 $A$ 和 $B$ 阻断了 $X$ 到 $Y$ 之间的路径
- 不控制 $A$ 或者 $B$ 阻断了 $X$ 到 $Y$ (一处被阻断，整条路径都被阻断)

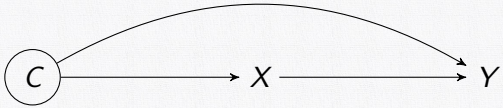
# 注意：



- 如果 $X$ 到 $Y$ 之间到所有通路都被阻断了，那么 $X$ 和 $Y$ 是独立的、不相关的
- 如果 $X$ 到 $Y$ 之间至少有一条通路是开放的，那么 $X$ 和 $Y$ 一般是有关系的、不独立的

# 举例

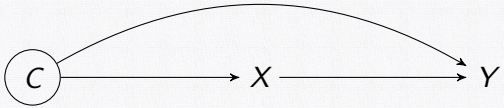





- 假设以上 DAG 描述了真实的因果关系网络
- 拟检验 $X$ 是否对 $Y$ 有因果关系
- 是否应该控制 $C$, 为什么?

# 举例

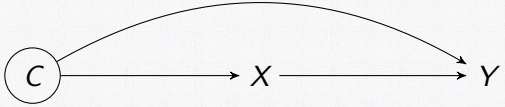





- 假设不控制 $C$，我们发现 $X$ 和 $Y$ 之间存在关系.
- 一个解释: 因果路径 $X \to Y$
- 另外的解释: 开放的非因果路径 $X \leftarrow C \to Y$
- 因此，如果 $X$ 和 $Y$ 之间存在关系, 在不控制 $C$ 的情况下, 不能证明因果路径 $X \to Y$ 是存在的.

# 举例

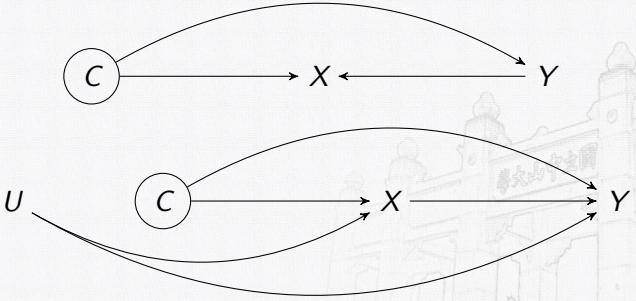





- 假设控制了 $C$, 我们发现 $X$ 和 $Y$ 之间存在关系.
- 唯一的解释: 因果路径 $X \to Y$
- 因为这个非因果路径 $X \leftarrow C \to Y$ 在控制了 $C$ 的情况下是被阻断的
- 因此，在控制了 $C$ 的情况下，$X$ 和 $Y$ 存在关系, 证明了 $X \to Y$ 因果路径存在.

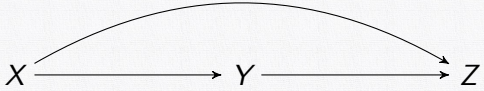
# 然而

- 如果 DAG 不正确，控制 $C$ 也不能证明因果关系
- 反向因果关系
- 未被测量的混杂



# 举例

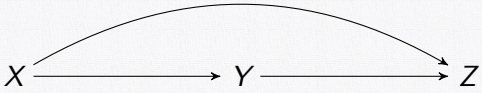





- 假设以上 DAG 是正确的
- 拟检验 $X$ 对 $Y$ 的因果关系
- 是否应该控制 $Z$, 为什么?

# 举例

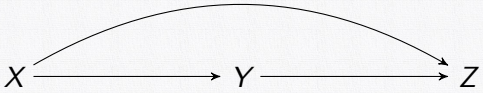





- 假设控制了 $Z$, 我们发现 $X$ 和 $Y$ 之间存在关系
- 一个解释: 因果路径 $X \to Y$
- 另外一个解释: 非因果路径 $X \to Z \leftarrow Y$ 被打开了 (规则 2)
- 因此, 当控制 $Z$ 时，发现 $X$ 和 $Y$ 之间有关系, 不能证明 $X \to Y$ 路径是存在的.

# 举例



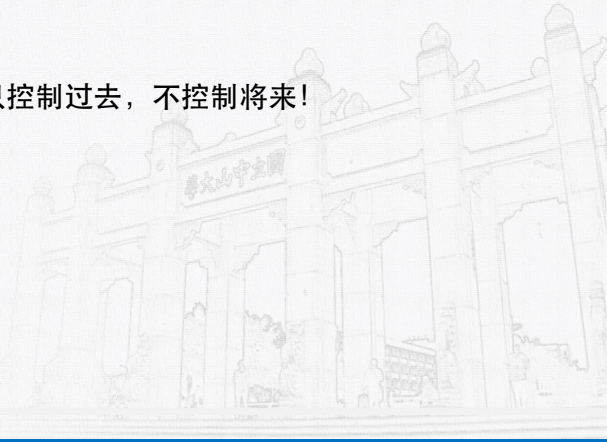



- 假设不控制 $Z$, 发现 $X$ 和 $Y$ 之间有关系.
- 唯一的解释: 因果路径 $X \to Y$
- 非因果路径 $X \to Z \leftarrow Y$ 被天然的阻断了 (规则 2)
- 因此，当不控制 $Z$ 时，$X$ 和 $Y$ 之间存在关系，可以证明因果路径 $X \to Y$ 存在.

# 一般策略



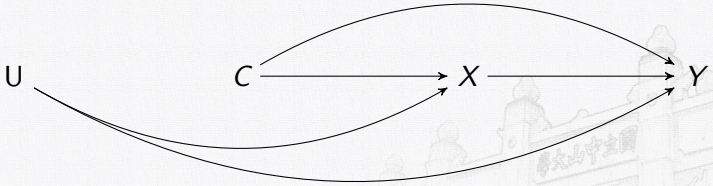
- 只控制可以阻断非因果通路的变量
- 不控制可以打开非因果通路的变量
- 请回忆 DAG“时间轴”这个概念，只控制过去，不控制将来！



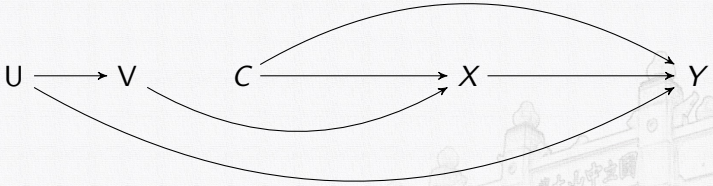
# 可能的问题



- 未被测量的混杂
- 尽可能的减少混杂



- 然而，如果 $V$ 是 $U$ 到 $X$ 之间的中介变量，我们仍然可以控制未被测量的混杂



# 更大的 DAG



- 更多的变量意味着更复杂的模型，可能在模型设置的时候会出错
- 某些变量可能更容易出现测量误差（信息偏倚），这将导致更大的偏倚 bias
- 某些变量可能会降低统计把握度
- 当出现这些情况的时候，某些变量的混杂效应较小时，可以不控制他们

# 小结



- 传统的控制变量（选择混杂因素）的方法可能比较主观、难以把握
- DAG 可以用来选择变量，但是需要依靠寄往的知识
- DAG 更重要的是用来方便交流（你与临床医生、与审稿人等）
- www.dagitty.net