Directed Acyclic Graphs (DAGs)

David Evans[1,2,3]

1 Institute for Molecular Bioscience, University of Queensland

2 University of Queensland Diamantina Institute

3 MRC Integrative Epidemiology Unit, University of Bristol

Why Study DAGs?

- To understand whether to condition on a variable or not
- To understand selection bias and loss to follow up bias
- To understand the impact of missing data

To Condition or Not to Condition?

	Drug	No Drug
Men	81 out of 87 recovered (97%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

- In a study, a group of sick patients is given the option to try a new drug
- The drug appears to hurt men and women separately, but be beneficial for the population
- So if we know the patient’s sex we shouldn’t prescribe the drug?
- But that is ridiculous…
- Conditional or Unconditional? Let’s vote!

To Condition or Not to Condition?

- Now imagine that we recorded individuals’ blood pressure at the end of the study
- Imagine that we know that the drug affects recovery by lowering the blood pressure of those who take it
- But it also involves a toxic side effect…
- Would you recommend the drug?
- Conditional or Unconditional? Let’s Vote!

	No Drug	Drug
Low BP	81 out of 87 recovered (97%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

To Condition or Not to Condition?

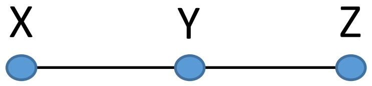
	Drug	No Drug
Men	81 out of 87 recovered (97%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

	No Drug	Drug
Low BP	81 out of 87 recovered (97%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

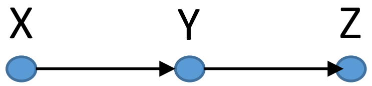
- Notice that the numbers are *exactly* the same in the two tables
- The decision to condition or not is driven by knowledge of the data generating mechanism, not by the data itself!

Graphs

- A graph is a series of nodes/vertices (variables) and edges



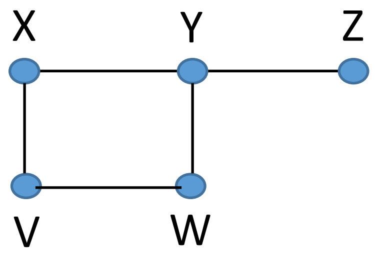
- A graph can be directed or undirected



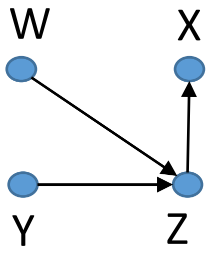
- A path between two nodes X and Z is a sequence of nodes beginning with X and ending with Z in which each node is connected to the next by an edge. A "directed path" follows the arrow heads.
- A Directed Acyclic Graph (DAG) is a graph that is Directed (has arrows) and Acyclic (no feedback loops).

DAG or not DAG?

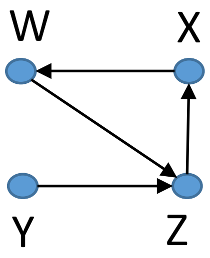
(A)



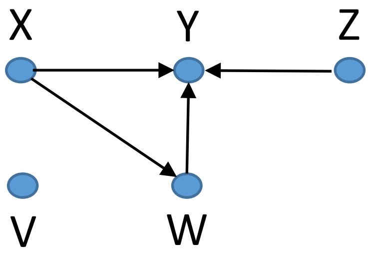
(B)



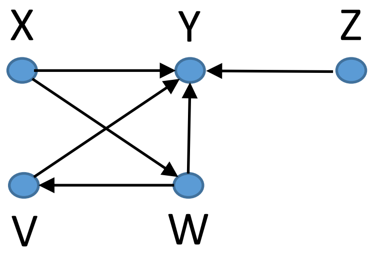
(C)



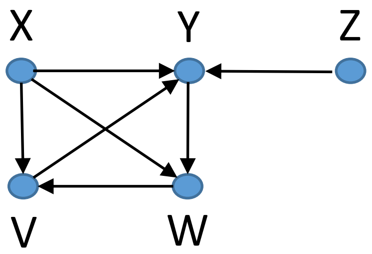
(D)



(E)



(F)



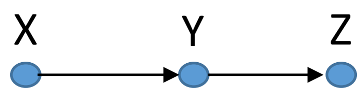
DAGs vs SEMs / Path Models

- DAGs and path models are related but not the same!

DAGs	Path Models
Distribution free	Assumes linearity and normality
Implies probabilistic dependencies in model	Implies (linear) covariances and variances in model
One headed arrows only	One headed and two headed arrows
Acylic	Feedback loops allowed
Boxes indicate conditioning	Boxes indicate observed variables

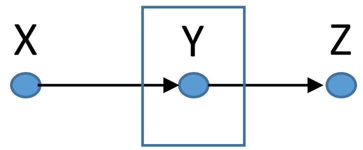
Structure #1 “Chains”

- Y and X are dependent
- Z and Y are dependent
- Z and X are (likely) dependent
- Z and X are independent conditional on Y (one way of thinking about conditioning on Y is like holding Y constant)



Structure #1 "Chains"

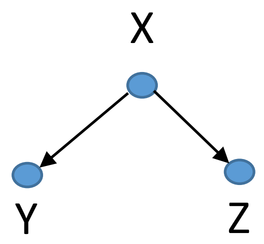
- Y and X are dependent
- Z and Y are dependent
- Z and X are (likely) dependent
- Z and X are independent conditional on Y (one way of thinking about conditioning on Y is like holding Y constant)



$$X \longrightarrow \boxed{Y} \longrightarrow Z$$

Structure #2 "Forks" (Confounders)

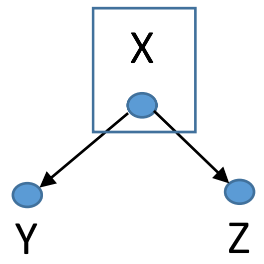
- Y = # of ice-cream cones eaten in a day
- Z = # of drownings in a day
- X = Temperature of day



- X and Y are dependent
- X and Z are dependent
- Z and Y are (likely) dependent
- Y and Z are independent conditional on X
- In epidemiology, X is a "confounder" that we will often want to control for

Structure #2 “Forks” (Confounders)

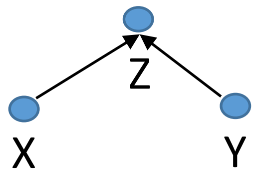
- Y = # of ice-cream cones eaten in a day
- Z = # of drownings in a day
- X = Temperature of day



- X and Y are dependent
- X and Z are dependent
- Z and Y are (likely) dependent
- Y and Z are independent conditional on X
- In epidemiology, X is a “confounder” that we will often want to control for

Structure #3 “Colliders”

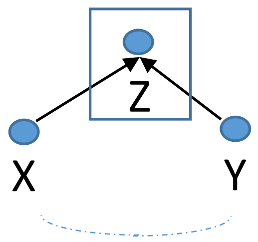
- Let X be musical ability
- Let Y be academic ability
- Let Z represent admittance to an exclusive school



- X and Z are dependent
- Y and Z are dependent
- X and Y are independent
- X and Y are dependent conditional on Z
- In epidemiology, Z is a “collider” and we often <u>do not</u> want to control for it.
- Z may represent missingness, selection into a study, loss to follow up etc

Structure #3 “Colliders”

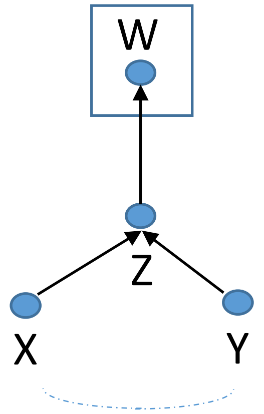
- Let X be musical ability
- Let Y be academic ability
- Let Z represent admittance to an exclusive school



- X and Z are dependent
- Y and Z are dependent
- X and Y are independent
- X and Y are dependent conditional on Z
- In epidemiology, Z is a “collider” and we often <u>do not</u> want to control for it.
- Z may represent missingness, selection into a study, loss to follow up etc

Structure #3 "Colliders"

- Also an issue if the variable is a descendent of a collider



- E.g. Let W represent wearing a posh school uniform

Dependent or Independent?

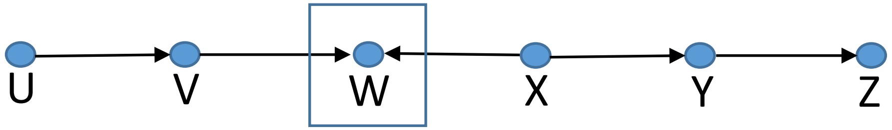
- Graphs allow us to determine whether two variables are independent or (likely) dependent
- Two variables are independent if every path between them is blocked
- If even one path between X and Y is unblocked, then X and Y are (likely) dependent

- Colliders block paths between variables
- The act of conditioning on a variable can block a path
- However, conditioning on a collider opens paths…

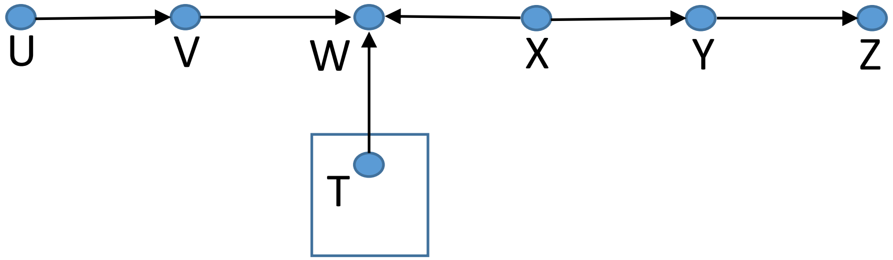
Exercise: Dependent or Independent?

Are U and Z independent?

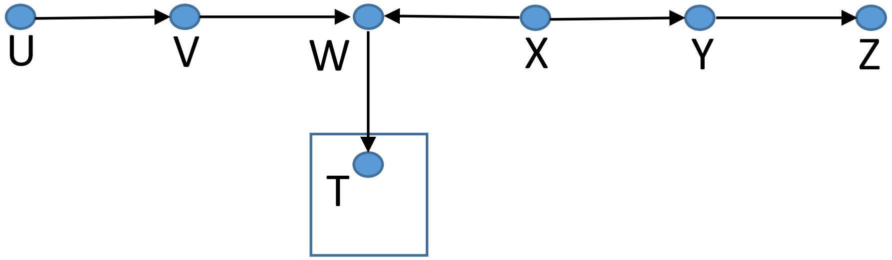
Are U and Z independent?



Are U and Z independent?

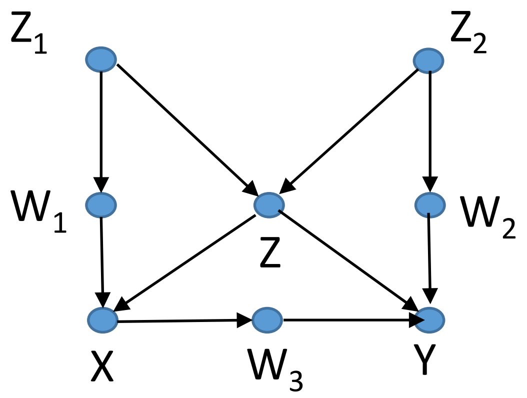


Are U and Z independent?



Exercises Continued

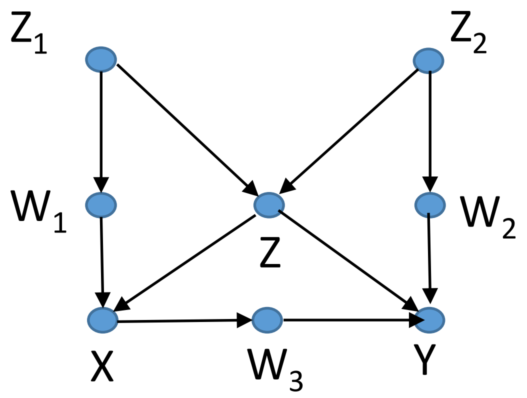
What's the minimum number of variables to condition on to make Z1 and Y conditionally independent? Which variables?



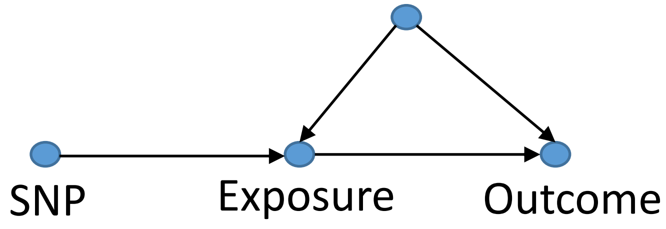
In a Mendelian randomization analysis, why do we not condition on the exposure variable (i.e. check if it blocks the path from the SNP to the outcome)?

Exercises Continued

What's the minimum number of variables to condition on to make Z1 and Y conditionally independent? Which variables?



In a Mendelian randomization analysis, why do we not condition on the exposure variable (i.e. check if it blocks the path from the SNP to the outcome)?



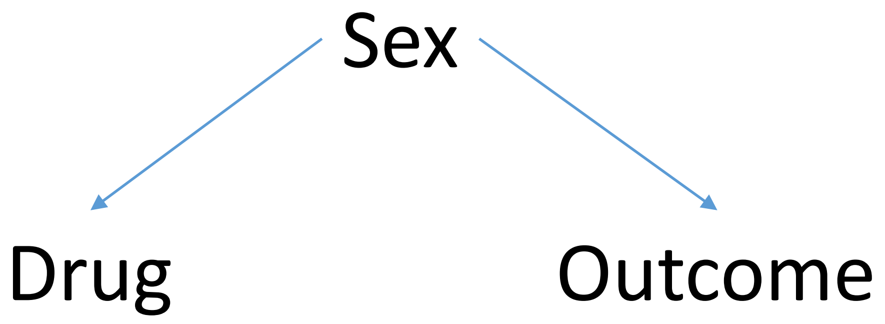
To Condition or Not to Condition?

	Drug	No Drug
Men	81 out of 87 recovered (97%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

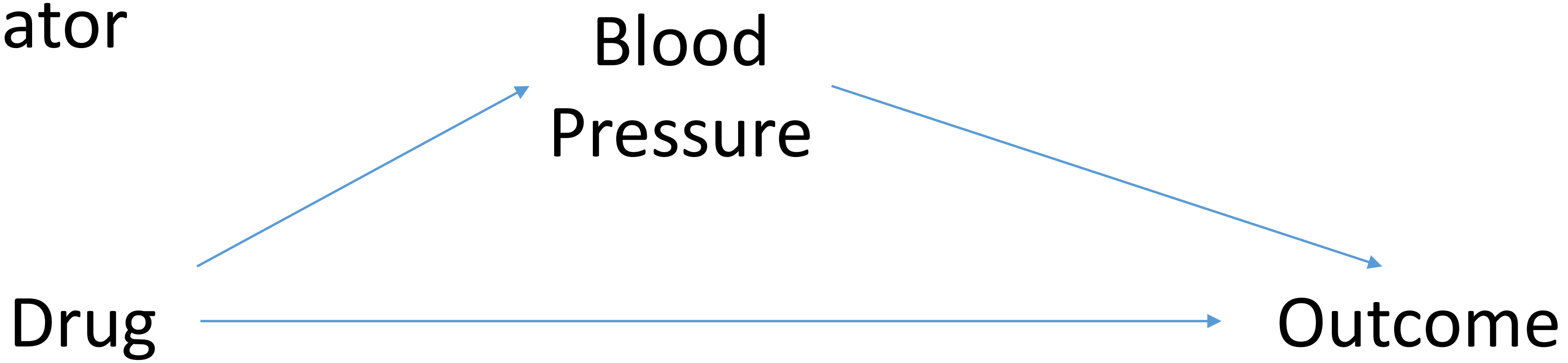
	No Drug	Drug
Low BP	81 out of 87 recovered (97%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

Mediator or Confounder?

(1) Confounder



(2) Mediator



DAGs To Understand Ascertainment Bias

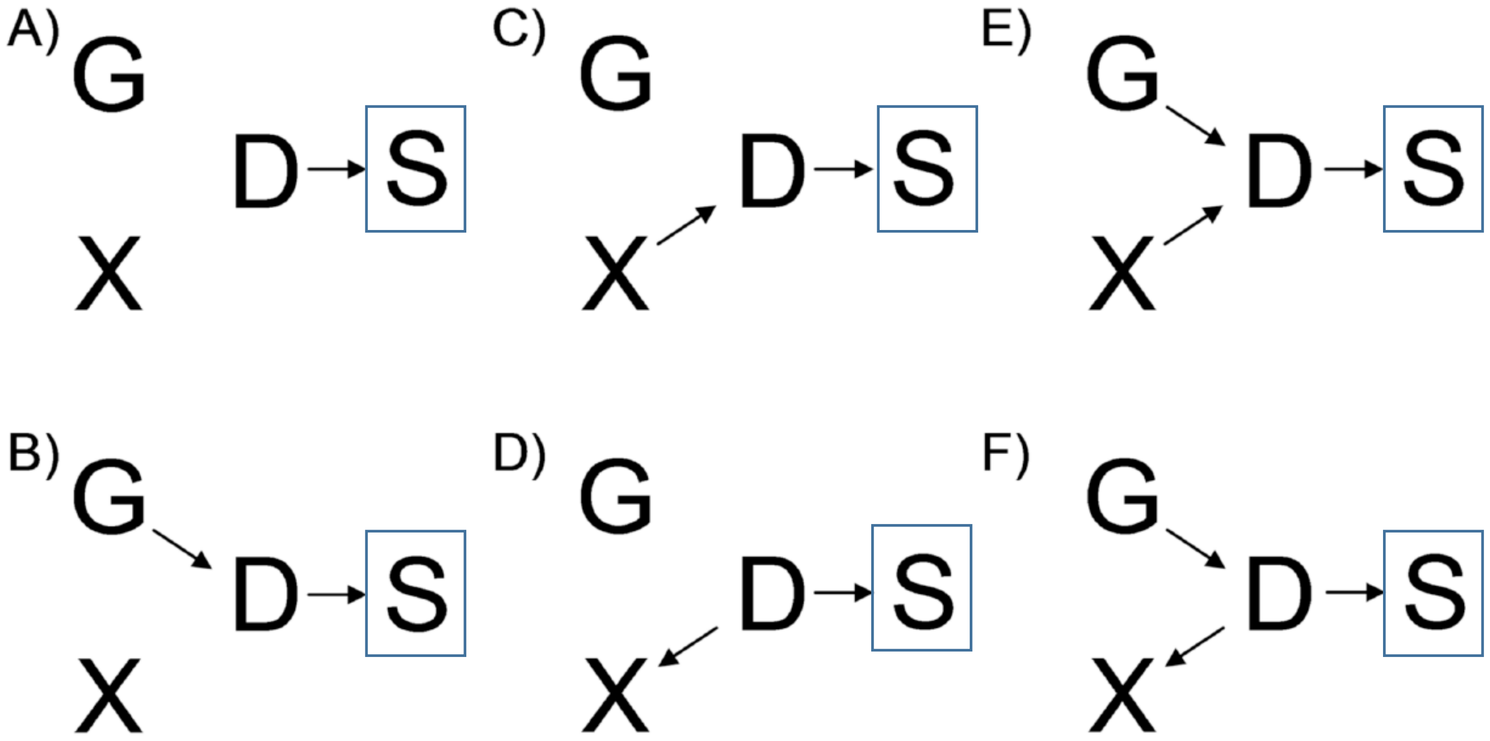


Figure 1.
Directed Acyclic Graphs (DAGs) describing the joint probabilities and conditional independence structure for genotype (G), disease status (D), secondary trait (X), and sampling indicator (S) for the six scenarios described in the text.

Monsees et al. (2009) Genet Epi

Further Reading

- Aschard H et al. (2015). Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *Am J Hum Genet*, 96(2), 329-39.
- Cole SR et al. (2010). Illustrating bias due to conditioning on a collider. *Int J Epidemiol*, 39(2), 417-20.
- Daniel RM et al. (2011). Using causal diagrams to guide analysis in missing data problems. *Stat Meth Med Res*, 21(3), 243-256.
- Mackenzie D. & Pearl J. (2018). The book of why.
- Monsees GM. et al (2009). Genome-wide association scans for secondary traits using case-control samples. *Genet Epidemiol*, 33(8), 717-728.
- Munafo MR. et al (2018). Collider scope: when selection bias can substantially influence observed associations. *Int J Epidemiol*, 47(1), 226-235.
- Pearl J. (2000). Causality.
- Pearl J., Glymour M., Jewell N.P (2016). Causal inference in statistics: A Primer.