Let's talk Re@l: Directed Acyclic Graphs (DAGs)

An introduction to this tool that helps study causal relationships between exposure and outcome by Barbara Torlinska and Rachel Tham

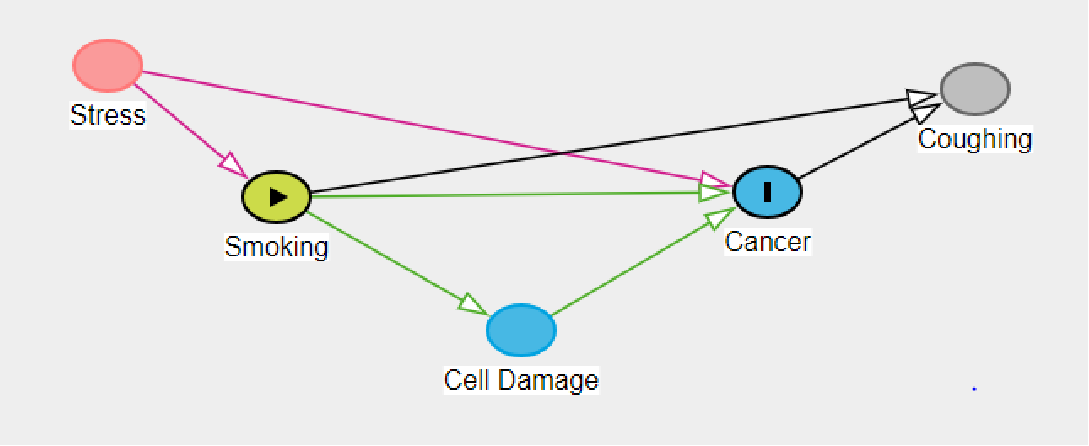
Whether assessing treatment effectiveness, evaluating surrogacy, generating a prediction model, or devising nearly any other task leading to the improvement of health outcomes, the researcher is interested in the causal relationship between the exposure and the outcome. However, in real life, there are multiple interplaying factors, making the exposure-outcome relationship complex and/or indirect. These complexities may influence the estimated effect or lead to the discovery of association without causation.

In this blog, we would like to introduce the basic features of a graphical tool depicting causality called directed acyclic graph (DAG). Familiarisation with DAGs and their underlying concepts can benefit researchers and be a useful tool to discuss challenging study elements with physicians, patients, and payers.

What is a DAG?

A DAG is a pictorial description of directional causal relationships between multiple variables, see figure 1. Each variable is a node, and arrows between the nodes go from cause to effect. As cause precedes effect a temporal rule that time flows from left to right applies.

Figure 1. Example DAG showing the causal relationship between smoking and cancer



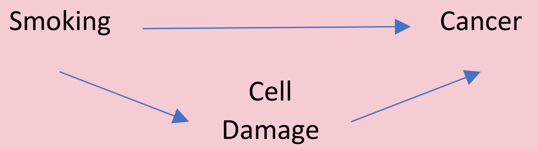
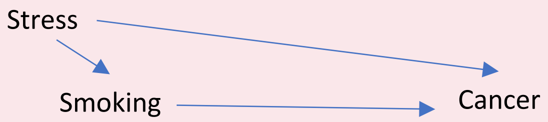
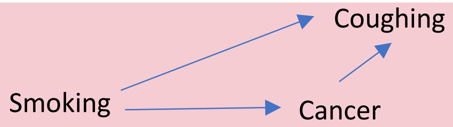
Fundamentally, a DAG follows 2 simple rules:

- A DAG must be acyclic (as the name implies), where you can't follow the arrows and end up back at the variable you started at. This means there can be no "feedback" loops in the DAG.
- For a DAG to be complete, the shared cause of any two variables must be included. This means that the DAG should include both measured and unmeasured variables along with all causal arrows. Therefore, if there is no arrow between two variables, then there is the assumption that there is no causal relationship that can be made between the two variables.

A DAG typically requires subject matter knowledge. Thus, each DAG is a valuable tool for causal inference researchers that explain how associations are generated from causal pathways.

What to look for in a DAG?

Causal pathways are one of the most important elements to look for in a DAG. A simple pathway connects three variables and can result in identifying the role of a variable, which is neither the exposure nor the outcome. The main three roles are a confounder, a collider and a mediator. The corresponding types of pathways are depicted below using the relationship between smoking (the exposure) and cancer (the outcome) from the full example DAG shown in figure 1.

Term	Description	Term	Example of acausal pathway from smoking to cancer
Mediator between two variables	All causal arrows point from the exposure to the mediator and from the mediator to the outcome (in same direction)	Directed path (open)	
Confounder between two variables	Two variables share the same cause (confounder). Therefore, association between the two variables may be due to confounding rather than be causal	Backdoor path	
Collider of two variables	Two variables have the same effect (collider) despite no association	Closed path (blocked)	

The reason why pathways are important is that researchers can change the status of some pathways from open to closed (or vice versa) through study design or statistical approaches. This is as achieved by conditioning on a variable and is depicted by adding a box around the variable. The examples of conditioning include adjusting for covariates in a model, including or restricting a subpopulation in the inclusion/exclusion criteria or study design, stratifying or looking at different subgroups in an analysis, assessing the per-protocol analysis set, or matching to balance groups (just to name a few!).

Conditioning

Understanding which variables require conditioning in order to open or close pathways is guided by a set of criteria called d-separation rules (whether d stands for directional or dependence seems to be unclear). There are four d-separation rules:

1. If there are no variables being conditioned on, a path is blocked if and only if two arrowheads on a path collide at some variable
2. Any path that contains a non-collider that has been conditioned on is blocked
3. A collider that has been conditioned on does not block a path
4. A collider that has a descendant that has been condition on does not block a path

In other words, if a confounder forms part of a DAG it needs to be conditioned. Technically speaking the backdoor path between exposure and outcome needs to be blocked in order to remove bias introduced by the confounder. In our Figure 1 example, failure to condition on stress will distort the causal pathway between smoking and cancer and the estimates will be biased.

Variables never to be conditioned on are mediators and colliders. Conditioning on a mediator, such as cell damage (Figure 1) on smoking – cancer pathway, will close an open pathway distorting the final results. Collider, such as cough, which can be caused by either smoking or cancer (Figure 1), needs no conditioning. Any adjustment for coughing will result in bias.

Therefore, a clear understanding of the role of each variable available for analysis is needed. Without it, in a mechanistic approach for data analysis, such as in stepwise elimination regression, the researcher can conclude on associations but not on the causal relationship between exposure and outcome.

Selection bias

In the design of any study, whether observational or randomised clinical trial, the researcher needs to be aware of the selection bias. Depending on the setting, this type of bias may carry different names: Berkson's bias, loss to follow-up, non-response bias, volunteer bias, missing data bias etc. All these biases have the same mechanism and are the result of conditioning on a common effect of treatment and outcome, i.e. on the collider or its descendant. In the simplified scenario presented in the table, where coughing is a common cause of smoking and cancer, data analysis of patients from respiratory clinic (coughing patients) will be the source of selection bias. Similarly, in an RCT in which loss to follow-up or non-response will be associated with coughing, the selection bias will creep in.

Conclusions

DAGs are a powerful tool when used wisely.

DAGs are about causal relationships (not just relationships). All causal relations (measured and unmeasured) need to be added as long as they are causal - they don't need to be statistically significant.

- DAGs are helpful in interpreting research that deals with proposed causal relationships
- When conducting research, it is crucial to identify and appropriately adjust for confounders or other sources of bias (while not adjusting for mediators or colliders!)
- DAGs can illustrate threats to validity (such as confounding, selection, collider-stratification, and overadjustment)
- DAGs remind those performing analytical studies to collect sufficient data to condition on possible confounders, and appropriately adjust for these, whilst refraining from inappropriate adjustments

They are a great collaborative tool between clinical expertise and researcher/statistician/epidemiologist. Following Hernan's advice: Draw Your Assumptions Before Your Conclusions.

Want to chat about DAGs more? Reach out to Barbara (https://www.linkedin.com/in/barbara-torlinska/) or Rachel (https://www.linkedin.com/in/racheltham1/)

References

- Hernán MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC. Available at: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book

- HarvardX: Causal Diagrams: Draw Your Assumptions Before Your Conclusions. EdX online course. https://www.edx.org/learn/data-analysis/harvard-university-causal-diagrams-draw-your-assumptions-before-your-conclusions

- Williams, T.C., Bach, C.C., Matthiesen, N.B. et al. Directed acyclic graphs: a tool for causal studies in paediatrics. Pediatr Res 84, 487–493 (2018). https://doi.org/10.1038/s41390-018-0071-3