

Continued Medical Education



DAG: method for causal modelling in epidemiological research

Anuji Upekshika Gamage^{1*} & Chrishantha Abeysena²

¹Management, Development and Planning Unit, Ministry of Health, Sri Lanka; ²Department of Public Health, Faculty of Medicine, University of Kelaniya, Sri Lanka

*Correspondence: anujigamage@gmail.com

DOI: : <https://doi.org/10.4038/jccpsl.v23i4.8129>

Received on: 11 November 2017

Accepted on: 13 December 2017

“To know the causes of a disease and to understand the use of the various methods by which the disease may be prevented amounts to the same thing as being able to cure the disease.”

– Hippocrates

What is causal modelling?

Causes necessarily precede effects (1). In epidemiology, the causal links of exposures and outcomes are studied to assist in deciding on appropriate statistical analysis, thus as close as possible to answer the causal question at hand. Causal modelling is a term applied to a wide variety of formal methods for representing and facilitating inferences about causal relationships (2). Causal graphs are used as an approach to visualize the causal links between exposures and outcomes, and are considered as tools for understanding the network of structures and relationships between variables. In epidemiology, causal graphs, causal diagrams and directed acyclic graphs (DAG) are synonymously used (1). DAG approach is likely to reduce the degree of bias for effect estimate in the chosen causal relationship, as it could detect and thereby assist in the control of confounding and selection bias (3).

DAG

DAG is useful for depicting a causal structure in epidemiological settings (1). It is a flow chart that visualizes the whole causal aetiological network and is of use for embedding causality in a formal causal framework.

As the name indicates DAG stands for;

- **Directed:** All arcs in a DAG are arrows indicating causal effects
- **Acyclic:** No arrows from descendants to ancestors
- **Graph:** A picture of nodes (variables C, X, Y) and ‘arcs’ or ‘edges’ (arrows) encoding relationships between variables

DAG allows for better insights into the assumed causal mechanisms, and can assist in the selection of factors to adjust for, in order to remove their confounding effect. Therefore, DAG is useful for selecting variables for a multivariate model such as logistic regression and cox regression models. In some situations, controlling a single variable is sufficient for removing the confounding effect of two or more factors. Therefore, DAG is useful to identify a minimum but a sufficient set of factors, to adjust for their confounding effects. This structured approach serves as a visual aid in a scientific discussion by presenting the underlying relations explicitly. It can depict and be used to explain all types of bias including information bias. Owing to these advantages, DAG may be more preferable over traditional methods for identifying sources of confounding, especially in complex research questions (4).

DAG provides a way to organize the covariates based on the best possible assumptions. Before starting to draw the DAG, an investigator needs to identify the variables that might be important in answering the

research question at hand. The potential variables are identified based on subjective priori assumptions and knowledge, reported literature and theoretical considerations and expert opinion (1, 5). It is important to note that DAG should not be limited to measured variables from available data or measured data; they must be constructed independent of the available data and background knowledge of the causal network, linking effect to the outcome (1, 3, 5).

Creating a DAG

In the causal DAG approach, an arrow connecting two variables indicates causation; variables with no direct causal association are left unconnected (1, 6). The most important aspect when constructing a causal DAG is to include in the diagram any common cause of variables on the DAG. Exogenous variables may be included or omitted from the DAG. However, it is important to include common causes on the DAG to be considered causal (cause and effect). The association between exposure (X) and outcome (Y) is spurious due to confounding. Confounding is a common cause for a perceived disease exposure association (7). Causal diagrams can help to identify the set of confounders that could be controlled for during analysis and other confounders could be discarded based on DAG (6). One DAG may not be sufficient to answer a complex clinical question. Multiple DAGs may be constructed and statistical associations observed from available data may be used to evaluate the consistency of observed probability distributions with the proposed DAGs. Statistical analyses may be undertaken as informed by different DAGs, and the results can be compared (3).

Terminology used in DAG

DAG contains directed *edges/arcs* (arrows), linking *nodes* (variables), and their *paths*.

A *path* is a sequence of unbroken arrows (edges) connecting two variables (nodes), irrespective of the number or direction of arrows (3, 8).

Example: $X \longrightarrow B \longrightarrow Y$

In this example, the directed path is from X to Y, then X is an *ancestor* (cause) of Y and Y is a *descendent* (effect) of X. The node B lies in the causal pathway between X and Y and is considered an *intermediate* or *mediator* variable on the directed path (3, 8). Controlling for mediators leads to a biased estimate of the exposure-disease relationship.

However, it should be remembered that DAGs are acyclic and no node can have an arrow pointing to itself, and all edges must be directed (contain arrows) as causes must precede the effect, as shown in the below example.

Example: $X \longleftarrow C \longrightarrow Y$

A *collider* is a variable, which has two or more arrows pointing into it meaning “arrows collide”. Bias can be introduced to a study due to conditioning (restricting or analysis) on a collider (3, 8-9).

Example: $X \longrightarrow B \longrightarrow U \longleftarrow Y$
U is a collider on the path.

A path is blocked if it contains a collider (8) i.e. associations are not transmitted across colliders.

Example: $X \longrightarrow B \longrightarrow U \longleftarrow Y$
X is not associated with Y.

Controlling for mediators and modifiers leads to a biased estimate of the causal relationship. Controlling for collider may lead to the induction of a causal pathway between two other factors, which may subsequently introduce confounding bias. DAG allows the researcher to oversee all information needed and to judge whether controlling for a certain factor might introduce confounding and/or colliding bias (4).

An unblocked path is called open.

Example: $X \longrightarrow B \longrightarrow U$ or $X \longleftarrow C \longrightarrow Y$
It transmits associations along it.

A path between X and Y is *blocked* or *closed*, if it contains a non-collider that has been conditioned (controlled for) or a collider that has not been conditioned on and no descendants of the collider have been conditioned on; otherwise it is unblocked or open (10).

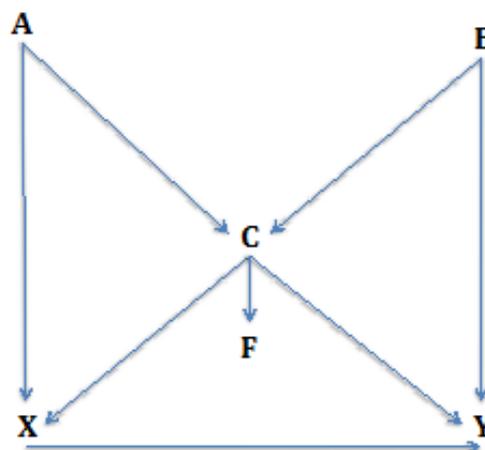


Figure 1. Hypothetical DAG used to illustrate the open backdoor path rule

As shown in Figure 1, a *front door* path from X to Y represents a causal effect of X on Y. i.e. $X \longrightarrow Y$. A *back door* path from X to Y is a path going out the “back door” of X via the head of an arrow. i.e. $X \longleftarrow C \longrightarrow Y$ (8).

Back door paths may correspond to potential confounding bias. For the purpose of controlling for confounders, the investigator should identify all backdoor paths. Confounding is present if the exposure and outcome are still connected other than the direct effect (1, 8). Also, it is important to note that the backdoor paths already blocked by a collider should be disregarded, hence could be deleted (1).

Example: $X \longleftarrow A \longrightarrow C \longleftarrow B \longrightarrow Y$

Controlling for collider may lead to induction of a causal pathway between two other factors, which may subsequently introduce confounding bias.

Example: 

The path between L and A is open after conditioning on D (a collider).

DAG allows the researcher to oversee all information needed, to judge whether controlling for a certain factor/s might introduce confounding and/or colliding bias. Restriction of the study population in relation to a collider may introduce bias to the exposure-disease association. The same will happen in a matched design where the matching factor is a collider.

After depicting the DAG diagrammatically for the purpose of analysis, the researcher needs to list out all possible backdoor pathways and unblocked pathways. Afterwards, he should identify a set of covariates that would block all still existing open backdoor paths by adjusting.

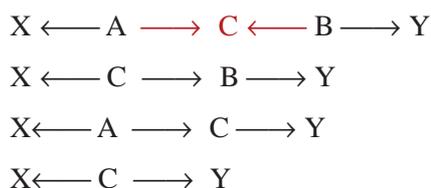


Figure 2. Finding the backdoor paths after eliminating the exposure outcome relationship and identifying the colliders

This is referred to as the *minimally sufficient adjustment set (MSAS)*. As shown in Figures 1 and 2, the MSAS is C. Adjusting for C would block all open backdoor paths (1). Generally, the more steps in the path between nodes, the weaker the effects.

Uses of DAG

- Encodes assumptions about the research question and the factors under study
- Makes those assumptions explicit and open for debate
- Provides a unifying link between DAG causal model and the statistical model used to make causal inference
- Makes us explicitly examine potential bias paths such as confounding bias, selection bias and information bias

In addition to the causal DAG that is described above, there is another called probabilistic DAG. In a probabilistic DAG, cause and effect are non-symmetrical while the strength or direction of association is not shown (1).

Figures 3 and 4 give two examples for a DAG.

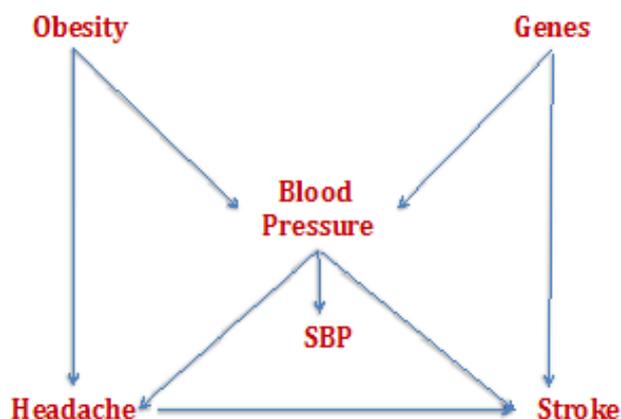


Figure 3. DAG for the association between headache and stroke

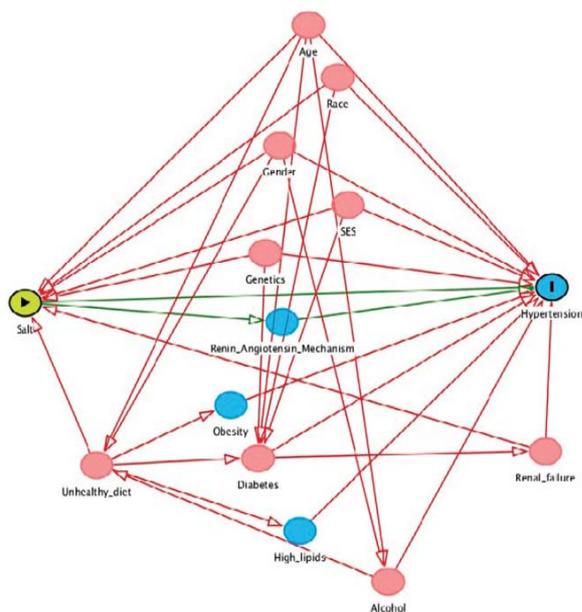


Figure 4. DAG using Dagitty to illustrate the association between salt and hypertension

Based on the above example (Figure 4) to illustrate the association between salt intake and hypertension, the minimal sufficient adjustment sets (MSAS) for estimating the total effect of salt on hypertension:

- age, alcohol, diabetes, gender, genetics, high lipids, obesity, ethnicity, renal failure, socio-economic status or
- age, gender, genetics, race, renal failure, socio-economic status, unhealthy diet

Hence, either one of the above adjustment sets is sufficient for estimating the effect of salt on hypertension.

References

1. Foraita RSJ & Zeeb H. Directed Acyclic Graphs. In Ahrens W & Pigeot I (Eds.), *Handbook of Epidemiology*. New York: Springer, 2014.
2. Hitchcock C. Causal Modelling. In Beebee H, Hitchcock C, Menzies P (Eds.), *The Oxford Handbook of Causation*. Oxford: Oxford University Press, 2009.
3. Shrier I & Platt RW. Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology* 2008; 8: 70.
4. Abeysena C. *Directed acyclic graphs in epidemiology*. Community Physicians Precis, 2016.
5. Petersen ML & Vander Laan MJ. Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology* 2014; 25(3): 418-426.
6. Sauer B & VanderWeele TJ. Use of directed acyclic graphs. In Velentgas P, Dreyer NA, Nourjah P, et al. (Eds.), *Developing a protocol for observational comparative effectiveness research: a user's guide*. Rockville (MD): Agency for Healthcare Research and Quality, Supplement 2, 2013.
7. VanderWeele TJ & Shpitser I. On the definition of a confounder. *The Annals of Statistics* 2013; 41(1): 196-220.
8. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999; 10(1): 37-48.
9. Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, Poole C. Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology* 2010; 39(2): 417-420.
10. Blackwell M. PSC 504 – *Observational Studies and Confounding*. 2013.