An Introduction to the Use of Directed Acyclic Graphs (DAGs) in Epidemiologic Research

OSCTR BERD SEMINAR
August 19th, 2016

Tabitha Garwe, PhD
Amanda Janitz, PhD
Sydney Martinez, PhD

Seminar Outline

• Theoretical Background – Tabitha Garwe
  • Confounding and Directed Acyclic Graphs (DAGs) – Background and Importance
  • DAG terminology
  • Assessing confounding using DAGs
  • Limitations of DAGs

• Applied Example – Amanda Janitz

• Daggity® Software Demonstration – Sydney Martinez
Interpreting associations

Assuming no systemic or random error, where do crude associations in our data come from?

1) Exposure causes disease

Smoking → Tar → Mutations → Tumor

Source: Petersen, M. Presentation 11/3/04

2) Exposure and disease share common cause

Source: Petersen, M. Presentation 11/3/04
Three Different Ways of Thinking About Confounding

1. Classical approach
2. Collapsibility approach
3. Counterfactual approach
Collapsibility Approach

• According to this view, a factor is a confounding variable if
  • the effect measure is homogeneous across the strata defined by the confounder and there is “lack of collapsibility”
  • *Collapsibility is equality of stratum-specific measures of effect with the crude (collapsed), unstratified measure* - Porta, 2008

Counterfactual Model View (Causality)

“*Confounding* is present if the substitute population imperfectly represents what the target would have been like under the counterfactual condition”

Maldonado & Greenland, Int J Epi 2002;31:422-29
Practical Implications of the Different Views

- **Counterfactual** – identifies specific conditions that must be met in order for observed associations to reflect accurately, a causal association
  - Limited value in practice – unobservable quantities
- **Classical** and **Collapsibility** approaches are more empirical in orientation
  - Ultimately the collapsibility view leads to what is arguably the most practical and efficient approach

• Why do statisticians and epidemiologists adjust for potentially confounding variables?
  - Because they can.
  - **But should they?**
  - Strategies for adjustment should account for “causal knowledge” (Hernan et al. AJE 2002)
Common Approaches to Evaluating Confounding

• Apply automatic variable selection procedures
• Compare adjusted and unadjusted effect estimates.
• Check whether the necessary criteria for confounding are met (classical approach).
• Approaches may introduce *conditional associations* and *create bias* where none existed

Hernan et al. AJE 2002
Control of Confounding: Analysis Stage

- Randomization assumption
- Conventional approaches:
  - Stratification
  - Multivariable Analysis
    - *Counterfactual model* provides a firm basis to discuss causation and confounding
    - But a large number of variables leads to a complicated scenario

Directed Acyclic Graphs: Uses *(AKA Causal Graphs)*

- Effectively minimize the number of confounding variables to measure or consider in the analysis
- Explicitly express assumptions about the causal structure (web of causation)
- Refine thinking about conditions on the directions of associations that are necessary for confounding

Under my *prior* assumptions, would the statistical analysis proposed here provide a *valid* test of a *causal* hypothesis?
Directed Acyclic Graphs: Other Uses

- Selection bias – Hernan, 2004
- Information bias – not as widely used for this yet
- DAG theory in the context of interaction/effect modification is still evolving

Issue of Interest

What is the effect of maternal multivitamin use on birth defects?

A priori knowledge allows us to make the following assumptions:

1) Prenatal care leads to an increase in vitamin use
2) Prenatal care protects against birth defects through pathways other than vitamin use
3) Difficulty conceiving may cause a woman to seek PNC once she becomes pregnant
4) Maternal genetics that lead to conception difficulty may also lead to birth defects
5) Socioeconomic characteristics directly effect both access to PNC and use of multivitamins

Sources: Hernan et al. 2002
Peterson, M. 11/3/04
Your Mission*

Draw a diagram to represent these causal relationships

*(should you choose to accept it...)

The Causal Diagram

- Maternal Genetics
- Conception Difficulty
- Prenatal Care
- SES
- Vitamin Use
- Birth Defects
• Under my prior assumptions, would the statistical analysis proposed here provide a valid test of a causal hypothesis?

What do DAGs include?

• Exposure and outcome for research question
• Suspected confounders
• Additional variables
• Both measured and unmeasured variables
  • This represents relationships between variables in a source population
• What about unknown relations?
  • Ideally based on subject matter expertise
  • When in doubt, draw multiple DAGs to see if meaningfully different

Source: Penny Howards, MCH EnRICH Webinar, 2013
TERMINOLOGY

DAG Notation

Node
Vertex

Y

Arc

Link
Edge

Z

Node
Vertex
DAG Notation

Y → Z

Parent  Child

DAG Notation: Paths

• Any way to connect two variables through a series of edges
  • Arrows can point in any direction

E ← X → M → O
E ← Z → A ← O
E ← L ← G → O
E → F ← H → O

Source: Penny Howards, MCH EnRICH Webinar, 2013
A directed path between two nodes is a path connecting the nodes where each edge of the path is an arrow that always follows the direction of the path – such a path aka causal path.

**Directed Paths:**
- X-Y-Z

**Not Directed Paths:**
- X-Z-Y
- W-Z-X

Directed paths – every edge has a single directed arrow: No variable can be a cause and effect of another variable at the same time.
A directed graph is called **acyclic** if no direct path forms a closed loop.

**DAG Notation**

- **Y** → **Z**
- **W** → **Z**
- **X** → **Z**

**Cyclic Graph: Smoking Status**

- Mother’s smoking status may be **both** a *cause* and a *consequence* of child’s respiratory condition.

Jewell, Ch 8
Acyclic Graph: Smoking status

Mother’s smoking status at time $t = 0$ → Mother’s smoking status at time $t = 1$

Child’s respiratory condition at time $t = 0$ → Child’s respiratory condition at time $t = 1$

DAG Notation

Mediators/interceptors can be considered on the causal pathway.
Other DAG Notation

Front Door vs. Back Door Paths

• Door is defined relative to your exposure
  • Front door paths – arrow leaving your exposure
    \[ E \rightarrow \]
  • Backdoor paths – arrow sneaking into your exposure
    \[ E \leftarrow \]

Source: Penny Howards, MCH EnRICH Webinar, 2013
Sample Backdoor Paths from E to O

- Key – arrow going into E
- OK for other arrows to point either way

E ➔ X ➔ M ➔ O
E ➔ X ← M ➔ O

Non-causal Paths (from E to O)

- Non-causal path – any path that is not a causal path from your exposure to your outcome
  - Classic example = backdoor path from E to O (confounding)
    - No causal path in this example
    - E and O associated solely because of confounding

E ← X ➔ M ➔ O

Source: Penny Howards, MCH EnRICH Webinar, 2013
More DAG Notation: Colliders

Closed/Blocked Path
Collider

W → X ← Y → Z

Open/Unblocked Path
Non-Collider

W → X → Y → Z

Open or Unblocked Paths

• **Association observed** in data - Could be causal or non-causal
  • Path with no colliders - **OPEN**
    • No variables conditioned on
  • Conditioning on a collider **OPENS** a path
    • (If there are no other colliders on the path)

E → X ← O

• Adjusting for X = spurious association between E and O

Modified from: Penny Howards, MCH EnRICH Webinar, 2013
Closed or Blocked Paths

- **No association observed** in data from that path

- Path includes a collider → **CLOSED**

![Diagram showing a collider and blocked path]

- Conditioning on a non-collider closes or blocks a path
  - Box indicates conditioning on a variable

Paths

Causal Question: What is the relation between childhood vaccination and risk of a subsequent health condition?

![Diagram showing various paths and variables]

Direct Path: (Vaccination → Health Outcome)
Backdoor Path(s)?
Blocked Path(s)?
Assumption regarding the relationship between SES and Family Hx?

Source: Jewell, Chap. 8
In any DAG, the only pathways between two distinct variables are either (1) a **directed path** or (2) **backdoor path** through a common ancestor.

Figure 8.4 *A directed acyclic causal graph that includes unmeasured variables U.*

Source: Jewell, Chap. 8
Assessing Confounding

1. 
\[ \text{C} \rightarrow \text{E} \rightarrow \text{D} \]

2. 
\[ \text{C} \rightarrow \text{E} \rightarrow \text{D} \]

Step 1: Delete all arrows from E that point to any other node
Step 2: Any unblocked backdoor paths from E to D?

Source: Jewell, Chap. 8

DAG Confounding

1. 
\[ \text{C} \rightarrow \text{F} \rightarrow \text{E} \rightarrow \text{D} \]

2. 
\[ \text{C} \rightarrow \text{F} \rightarrow \text{E} \rightarrow \text{D} \]

3. 
\[ \text{C} \leftarrow \text{F} \leftarrow \text{E} \rightarrow \text{D} \]

4. 
\[ \text{C} \leftarrow \text{F} \leftarrow \text{E} \rightarrow \text{D} \]

Source: Jewell, Chap. 8
Issue of Interest

What is the effect of maternal multivitamin use on birth defects?

A priori knowledge allows us to make the following assumptions:

1) Prenatal care leads to an increase in vitamin use
2) Prenatal care protects against birth defects through pathways other than vitamin use
3) Difficulty conceiving may cause a woman to seek PNC once she becomes pregnant
4) Maternal genetics that lead to conception difficulty may also lead to birth defects
5) Socioeconomic characteristics directly effect both access to PNC and use of multivitamins

DAGs and confounding

- Step 1: No variables in C should be descendants of E
- Step 2: Delete all non-ancestors of [E, D, C]
- Step 3: Delete all arrows emanating at E
- Step 4: Connect any two parents with a common child
- Step 5: Strip arrowheads from all edges
- Step 6: Delete C

Test: If E is disconnected from D in the remaining graph, then adjustment for C is sufficient to remove confounding.
If E and D are still connected, additional adjustment is required.
Confounding?

Remove all direct effects of E

- Conception Difficulty
- PreNatal Care
- SES
- Vitamin Use

Q: Do E and D share common cause?

- Conception Difficulty
- PreNatal Care
- SES
- Vitamin Use

Birth Defects

Birth Defects
Adjustment

Step 1: No variables in C should be descendants of E
Prenatal care caused by vitamin use?

Maternal Genetics

Conception Difficulty

PreNatal Care

SES

Vitamin Use

Birth Defects

Adjustment

Step 2: Delete all non-ancestors of vitamin use (E), birth defects (D), and prenatal care (C)

Maternal Genetics

Conception Difficulty

PreNatal Care

SES

Vitamin Use

Birth Defects
Adjustment

Step 3: Delete all direct effects of vitamin use (all edges emanating from vitamin use)

- Maternal Genetics
- Conception Difficulty
- PreNatal Care
- SES
- Vitamin Use
- Birth Defects

Adjustment

Step 4: Connect any two causes sharing a common effect

- Maternal Genetics
- Conception Difficulty
- PreNatal Care
- SES
- Vitamin Use
- Birth Defects
Adjustment

Step 5: Strip arrow heads from all edges

Maternal Genetics

Conception Difficulty

PreNatal Care

SES

Vitamin Use

Birth Defects

Adjustment

Step 6: Delete prenatal care (and all associated edges)

Maternal Genetics

Conception Difficulty

PreNatal Care

SES

Vitamin Use

Birth Defects
Adjustment

Test: Are vitamins and birth defects still connected?

YES: How else can we control for confounding?
Review: DAGs and confounding

- Step 1: No variables in C should be descendants of E
- Step 2: Delete all non-ancestors of [E, D, C]
- Step 3: Delete all arrows emanating at E
- Step 4: Connect any two parents with a common child
- Step 5: Strip arrowheads from all edges
- Step 6: Delete C

**Test:** If E is disconnected from D in the remaining graph, then adjustment for C is sufficient to remove confounding.
If E and D are still connected, additional adjustment is required.

Sources: Petersen, M. 11/3/04, Pearl, J. Causality

Assessing confounding

Remove direct effect of E on D

```
SES → Health Care Access
Vaccination → Health Outcome
Family Hx
```

Source: Jewell, Chap. 8
Assessing confounding: Another Example

Step 1: Delete direct effects of exposure of interest

 SES  →  Health Care Access  ←  Family Hx
     |       |                  |
  Vaccination  |                 |
     |       |                  |
   →              |
   Health Outcome

Source: Jewell, Chap. 8

Assessing confounding

Step 2: Delete all non-ancestors of E, D, C

 SES  →  Health Care Access  ←  Family Hx
     |       |                  |
  Vaccination  |                 |
     |       |                  |
   →              |
   Health Outcome

Source: Jewell, Chap. 8
Assessing confounding

Step 3: Delete all direct effects of E

SES → Vaccination → Health Care Access → Health Outcome

Family Hx

Source: Jewell, Chap. 8

Assessing confounding

Step 4: Connect any two causes sharing a common effect

SES → Vaccination

Health Care Access → Health Outcome

Family Hx

Source: Jewell, Chap. 8
Assessing confounding

Step 5: Delete arrow heads from all edges

Assessing confounding

Step 6: Delete C and all associated edges

Source: Jewell, Chap. 8
Assessing confounding

Step 6: Delete C and all associated edges

SES

Family Hx

Vaccination

Health Outcome

Are E and D still connected?

Source: Jewell, Chap. 8

Overadjustment Bias and Unnecessary Adjustment in Epidemiologic Studies

*Enrique F. Schisterman, Stephen R. Cole,* and Robert W. Platt*

Abstract: Overadjustment is defined inconsistently. This term is meant to describe control (e.g., by regression adjustment, stratification, or restriction) for a variable that either increases net bias or decreases precision without affecting bias. We define overadjustment bias as control for an intermediate variable (or a confounding proxy for an intermediate variable) on a causal path from exposure to outcome. We define unnecessary adjustment as control for a variable that does not affect bias of the causal relation between exposure and outcome but may affect its precision. We use causal diagrams and an empirical example (the effect of maternal smoking on neonatal mortality) to illustrate and clarify the definition of overadjustment bias, and to distinguish overadjustment bias from unnecessary adjustment. Using simulations, we quantify the amount of bias associated with overadjustment. Moreover, we show that this bias is based on a different causal structure from confounding or selection biases. Overadjustment bias is not a finite sample bias, while inefficiencies due to control for unnecessary variables are a function of sample size.

(Epidemiology 2009;20: 488–495)
Unnecessary and Harmful Adjustment

A few DAG limitations

• Not built to handle effect modification
• Assumption in model is that there is no information bias or selection bias
• If time-dependent confounding is present, simple confounder adjustment as described here not sufficient to control for confounding
• Subject matter knowledge is crucial!
The ultimate complex causal graph!

A PowerPoint diagram meant to portray the complexity of American strategy in Afghanistan!

Jewell, N. *Statistics for Epidemiology*, Chapter 8

Shrier and Platt (2008). Reducing bias through directed acyclic graphs. BMC Medical Research Methodology. 8:70

Glymour, M. “Using causal diagrams to understand common problems in social epidemiology,” in *Methods in Social Epidemiology*.

Petersen, M. “Causal diagrams: Directed acyclic graphs to understand, identify, and control for confounding.” *Presentation to Epidemiologic Methods II, UC Berkeley November 3, 2004*.

Magzamen S. BSE Seminar, OUHSC College of Public Health, 2011


Applied Example

Example: Congenital anomalies and childhood cancer

Minimally sufficient set for the total effect of childhood of congenital anomalies on childhood cancer:

• Gender, family history of anomalies and cancer, maternal age, plurality, prenatal vitamin use, and SES.

Janitz et al., 2016
Example: Benzene and childhood leukemia

Minimally sufficient adjustment sets for estimating the total effect of benzene on acute leukemia:

- Degree of Urban Development, Parental Smoking, Socioeconomic Status (maternal education)
- EMFs, Parental Smoking, Socioeconomic Status (maternal education)
Example: Benzene and childhood leukemia EXCLUDING urbanization

Minimally sufficient adjustment sets for estimating the total effect of benzene on acute leukemia:

- Parental Smoking, Socioeconomic Status (maternal education)

Janitz et al., 2016

How I used DAGs...

- Conduct a thorough literature review
  - Risk factors for exposure and outcome
  - Common confounders evaluated
- Draw DAG (may take many iterations)
  - Understand relationships between all variables included in the DAG
- Identify minimally sufficient set(s)
- Conduct statistical analysis
  - Only including minimally sufficient set(s)
  - Including other potential confounders identified in the literature
Using Dagitty®

• http://dagitty.net/

Welcome to DAGitty!

What is this?

DAGitty is a browser-based environment for creating, editing, and analyzing causal models (also known as directed acyclic graphs or causal Bayesian networks). The focus is on the use of causal diagrams for minimizing bias in empirical studies in epidemiology and other disciplines. For background information, see the “Learn” page.

DAGitty is developed and maintained by Johannes Textor (Theoretical Biology & Bioinformatics group, University of Utrecht).

Versions

The following versions of DAGitty are available:

- Development version
  This is the current development snapshot. May contain new features, but could also contain new bugs.
- 2.3: Released 2015-09-19
- 2.2: Released 2014-10-30
- 2.1: Released 2014-02-06
- 2.0: Released 2013-02-12
- 1.1: Released 2011-11-28
- 1.0: Released 2011-02-24
- 0.9b: Released 2010-11-24
- 0.8a: Released 2010-09-21

News on Twitter

#dagitty