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

Oklahoma Shared Clinical and Translational Resources
http://osctr.ouhsc.edu
NIGMS award U54GM104938

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

SCTR Oklahoma Shared Clinical & Translational Resources

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

3

Interpreting associations

2) Exposure and disease share common cause

SCTR Oklahoma Shared Clinical & Translational Resources

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

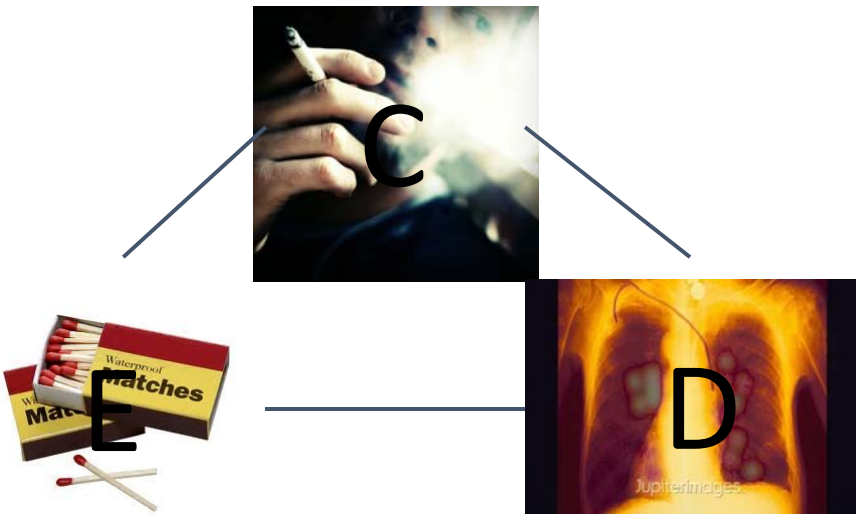
4

Three Different Ways of Thinking About Confounding

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

OSCTR Oklahoma Shared Clinical & Translational Resources

Confounding: Classical View



OSCTR Oklahoma Shared Clinical & Translational Resources

6

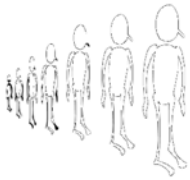
Collapsibility Approach

- 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*

OSCTR Oklahoma Shared Clinical & Translational Resources

Counterfactual Model View (Causality)

Counterfactual, unexposed cohort



Substitute, unexposed cohort

"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

OSCTR Oklahoma Shared Clinical & Translational Resources

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

OSCTR Oklahoma Shared Clinical & Translational Resources

- 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)**

OSCTR Oklahoma Shared Clinical & Translational Resources

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

OSCTR Oklahoma Shared Clinical & Translational Resources

Hernan et al. AJE 2002

American Journal of Epidemiology
Copyright © 2002 by the Johns Hopkins Bloomberg School of Public Health
All rights reserved

Vol. 155, No. 2
Printed in U.S.A.

Causal Knowledge as a Prerequisite for Confounding Evaluation: An Application to Birth Defects Epidemiology

Miguel A. Hernán,[1] Sonia Hernández-Díaz,[2] Martha M. Werler,[2] and Allen A. Mitchell[2]

Common strategies to decide whether a variable is a confounder that should be adjusted for in the analysis rely mostly on statistical criteria. The authors present findings from the Slone Epidemiology Unit Birth Defects Study, 1992–1997, a case-control study on folic acid supplementation and risk of neural tube defects. When statistical strategies for confounding evaluation are used, the adjusted odds ratio is 0.80 (95% confidence interval: 0.62, 1.21). However, the consideration of a priori causal knowledge suggests that the crude odds ratio of 0.65 (95% confidence interval: 0.46, 0.94) should be used because the adjusted odds ratio is invalid. Causal diagrams are used to encode qualitative a priori subject matter knowledge. *Am J Epidemiol* 2002;155:176–84.

abnormalities; causality; confounding factors (epidemiology); inference; selection bias

OSCTR Oklahoma Shared Clinical & Translational Resources

Control of Confounding: Analysis Stage

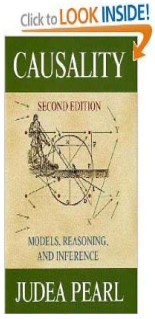
- Randomization assumption
- Conventional approaches:
  - Stratification
  - Multivariable Analysis
    - *Counterfactual model* provides a firm basis to discuss causation and confounding
    - But a large number of variables leads to a complicated scenario

OSCTR Oklahoma Shared Clinical & Translational Resources

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

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?

OSCTR Oklahoma Shared Clinical & Translational Resources

14

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

OSCTR Oklahoma Shared Clinical & Translational Resources

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

OSCTR Oklahoma Shared Clinical & Translational Resources

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

16

Your Mission*

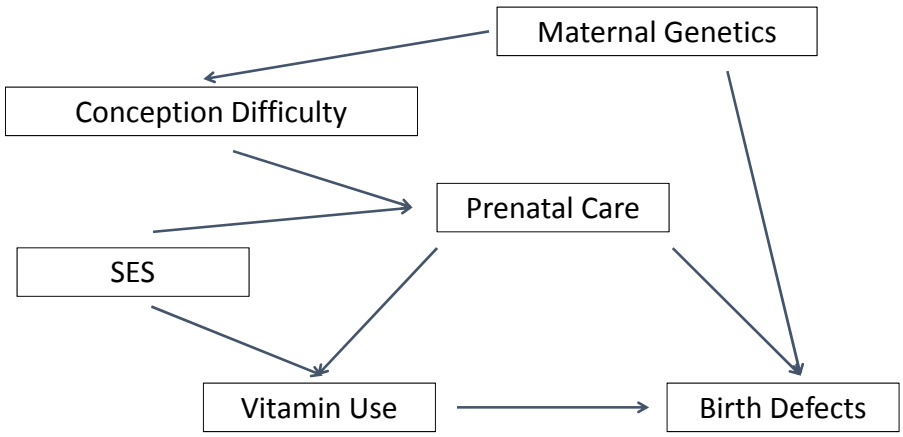
Draw a diagram to represent these **causal** relationships

*(should you choose to accept it…)

OSCTR Oklahoma Shared Clinical & Translational Resources

17

The Causal Diagram



OSCTR Oklahoma Shared Clinical & Translational Resources

18

- Under my *prior* assumptions, would the statistical analysis proposed here provide a valid test of a causal hypothesis?

OSCTR Oklahoma Shared Clinical & Translational Resources

19

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

OSCTR Oklahoma Shared Clinical & Translational Resources

Source: Penny Howards, MCH EnRICH Webinar, 2013

TERMINOLOGY

OSCTR Oklahoma Shared Clinical & Translational Resources

DAG Notation

Link

Edge

Arc

Node

Vertex

Node

Vertex

OSCTR Oklahoma Shared Clinical & Translational Resources

22

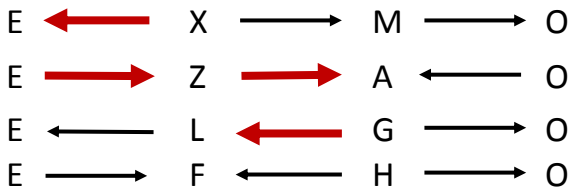
DAG Notation

OSCTR Oklahoma Shared Clinical & Translational Resources

23

DAG Notation: Paths

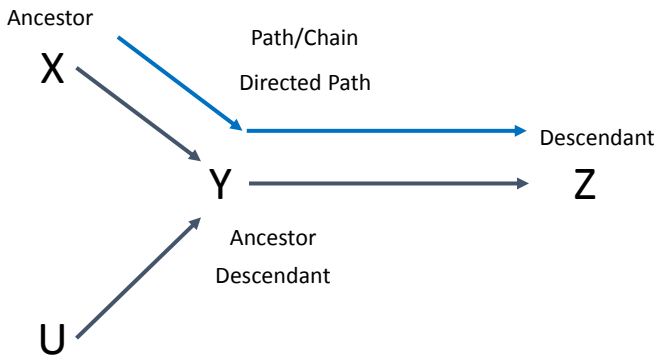
- Any way to connect two variables through a series of edges
  - Arrows can point in any direction



OSCTR Oklahoma Shared Clinical & Translational Resources

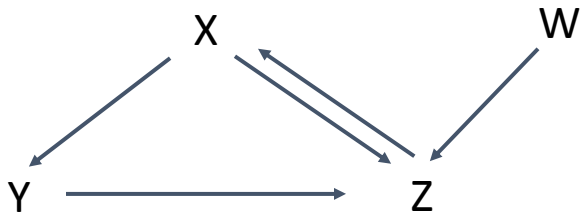
Source: Penny Howards, MCH EnRICH Webinar, 2013

DAG Notation



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

DAG Notation



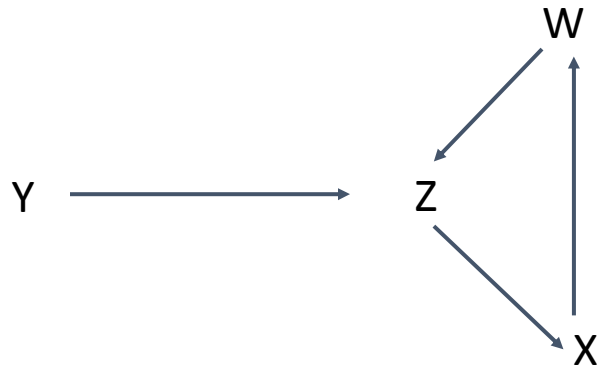
Directed Paths: X-Y-Z

<u>Not</u> 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.

DAG Notation

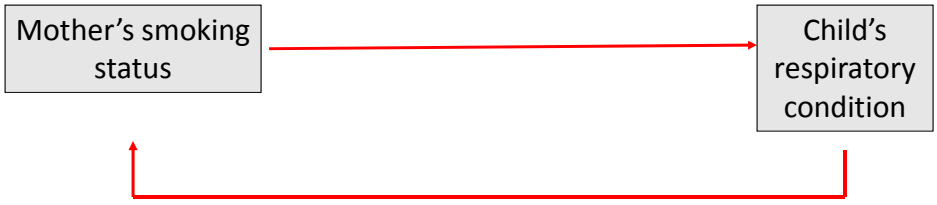


A directed graph is called ***acyclic*** if no direct path forms a closed loop

OSCTR Oklahoma Shared Clinical & Translational Resources

27

Cyclic Graph: Smoking Status



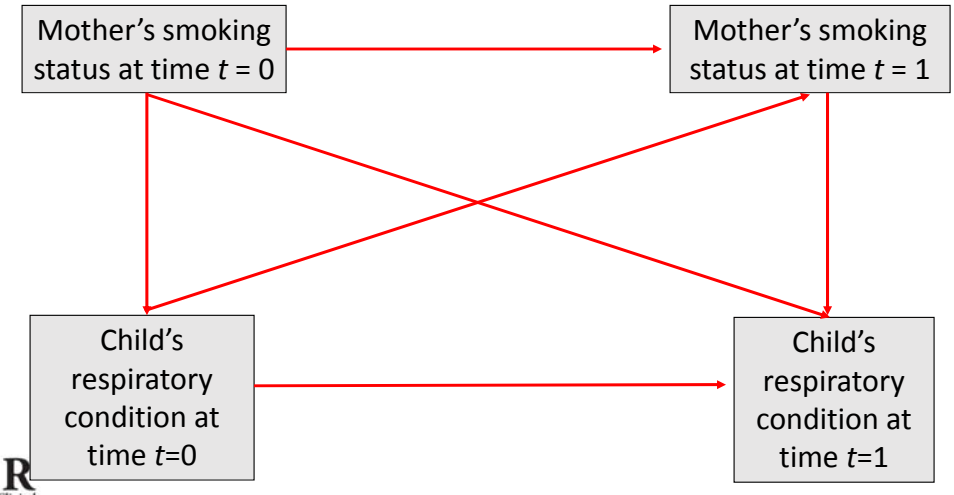
• Mother's smoking status may be **both** a **cause** and a **consequence** of child's respiratory condition

OSCTR Oklahoma Shared Clinical & Translational Resources

Jewell, Ch 8

28

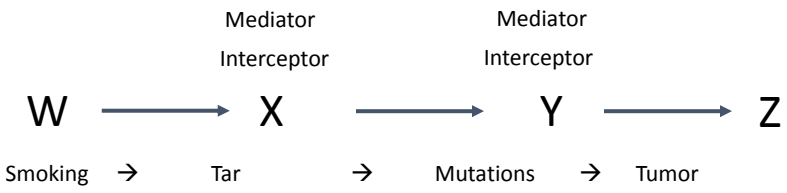
Acyclic Graph: Smoking status



OSCTR Oklahoma Shared Clinical & Translational Resources

29

DAG Notation

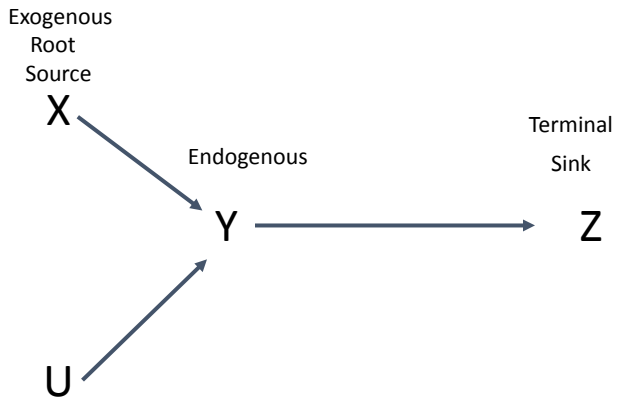


Mediators/interceptors can be considered *on the causal pathway.*

OSCTR Oklahoma Shared Clinical & Translational Resources

30

Other DAG Notation



Front Door vs. Back Door Paths

- Door is defined relative to your exposure
  - Front door paths – arrow leaving your exposure

E ⟶

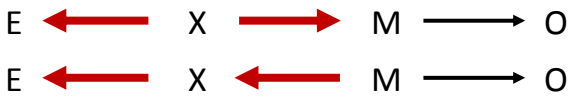
  - Backdoor paths – arrow sneaking into your exposure

E ⟵

Source: Penny Howards, MCH EnRICH Webinar, 2013

Sample Backdoor Paths from E to O

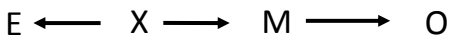
- Key – arrow going into E
- OK for other arrows to point either way



Source: Penny Howards, MCH EnRICH Webinar, 2013

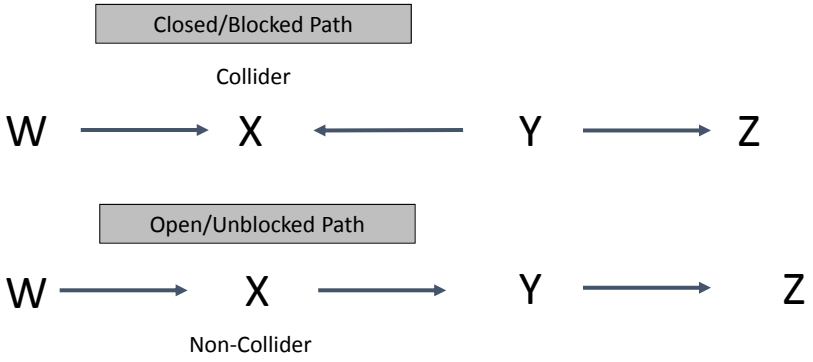
Non-causal Paths (from E to O)

- 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



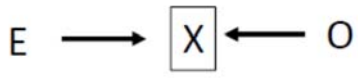
Source: Penny Howards, MCH EnRICH Webinar, 2013

More DAG Notation: Colliders



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)



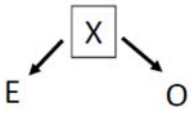
- Adjusting for X = spurious association between E and O

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**

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

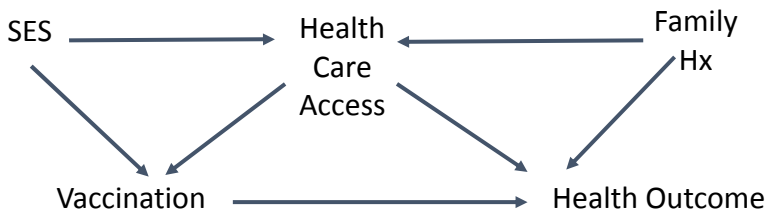


OSCTR Oklahoma Shared Clinical & Translational Resources

Modified from : Penny Howards, MCH EnRICH Webinar, 2013

Paths

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



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

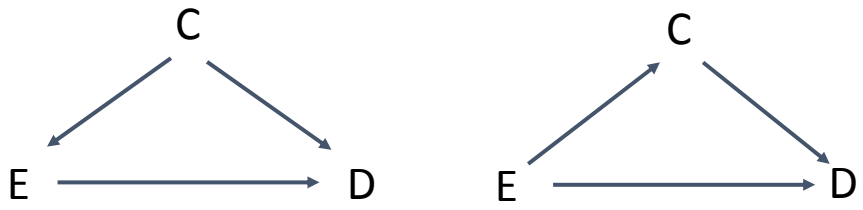
OSCTR Oklahoma Shared Clinical & Translational Resources

Source: Jewell, Chap. 8

38

Representing Confounding

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.

Representing Confounding

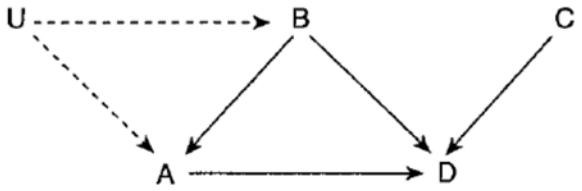
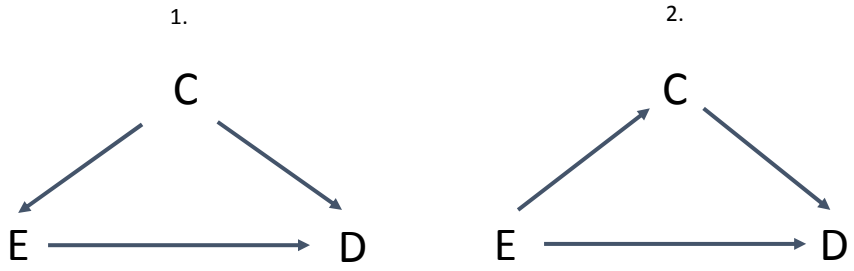


Figure 8.4 *A directed acyclic causal graph that includes unmeasured variables U.*

Source: Jewell, Chap. 8

Assessing Confounding

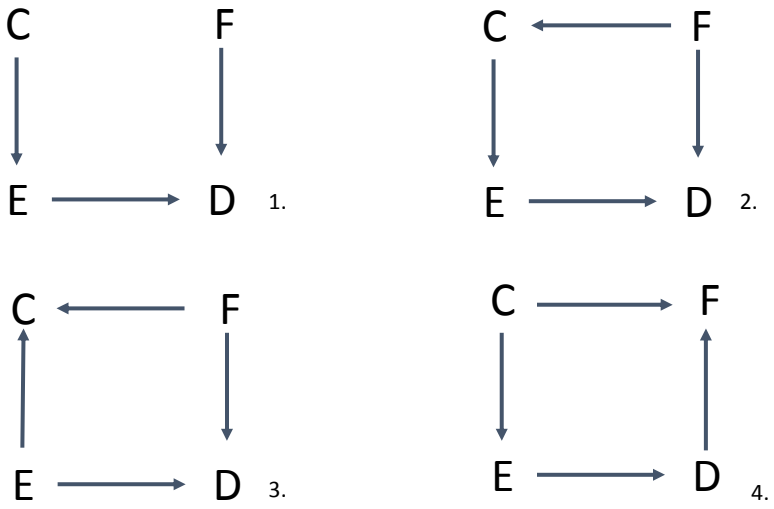


Step 1*: Delete all arrows from E that point to any other node*
Step 2: *Any* ***unblocked backdoor*** *paths from E to D?*

Source: Jewell, Chap. 8

41

DAG Confounding



Source: Jewell, Chap. 8

42

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

OSCTR Oklahoma Shared Clinical & Translational Resources

Source: Hernan *et al.* 2002

43

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.

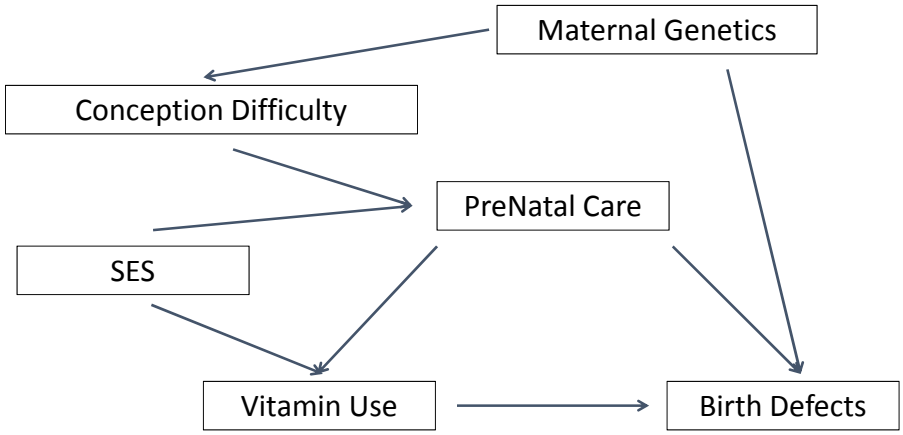
OSCTR Oklahoma Shared Clinical & Translational Resources

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

44

Confounding?

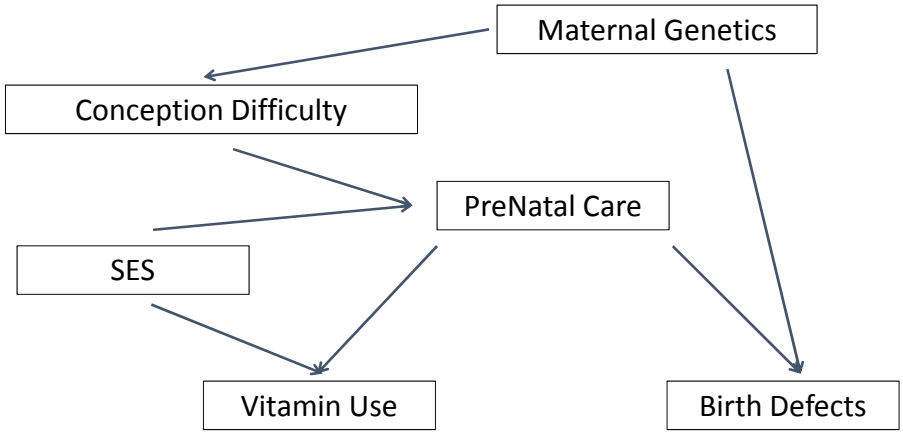
Remove all direct effects of E



OSCTR Oklahoma Shared Clinical & Translational Resources

45

Q: Do E and D share common cause?



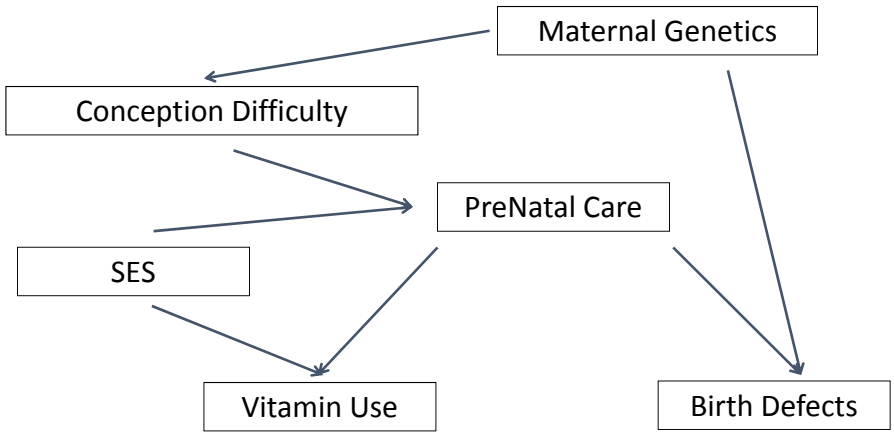
OSCTR Oklahoma Shared Clinical & Translational Resources

46

Adjustment

Step 1: No variables in C should be descendants of E

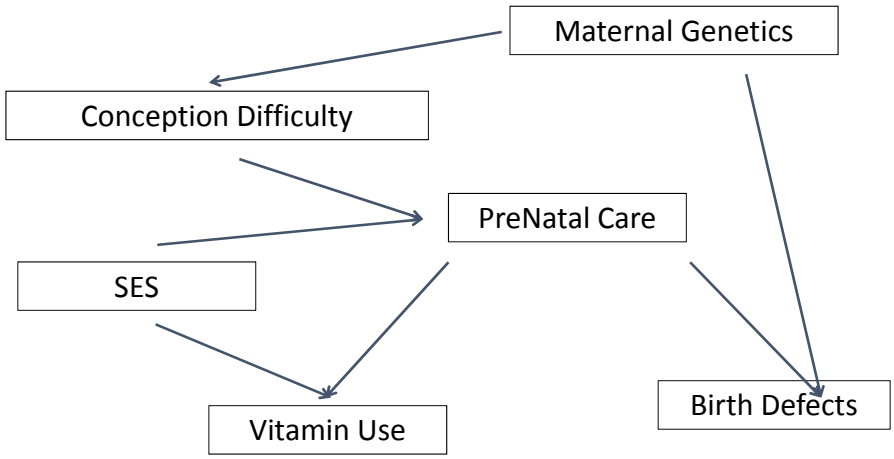
Prenatal care *caused* by vitamin use?



47

Adjustment

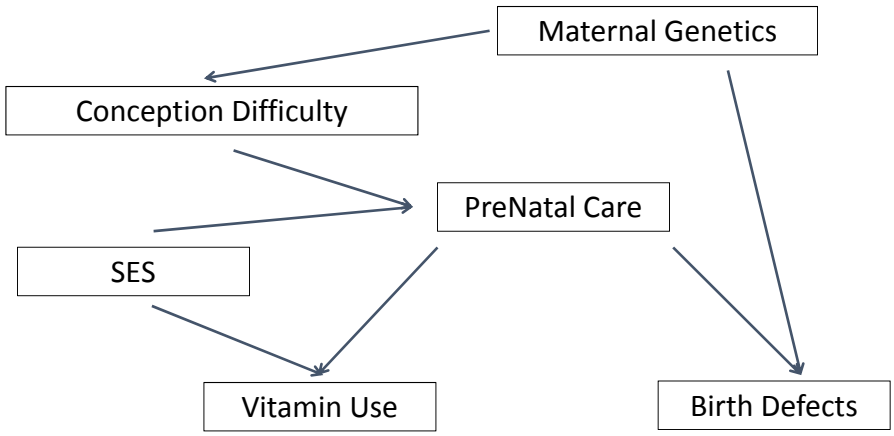
Step 2: Delete all non-ancestors of vitamin use(E), birth defects (D), and prenatal care (C)



48

Adjustment

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

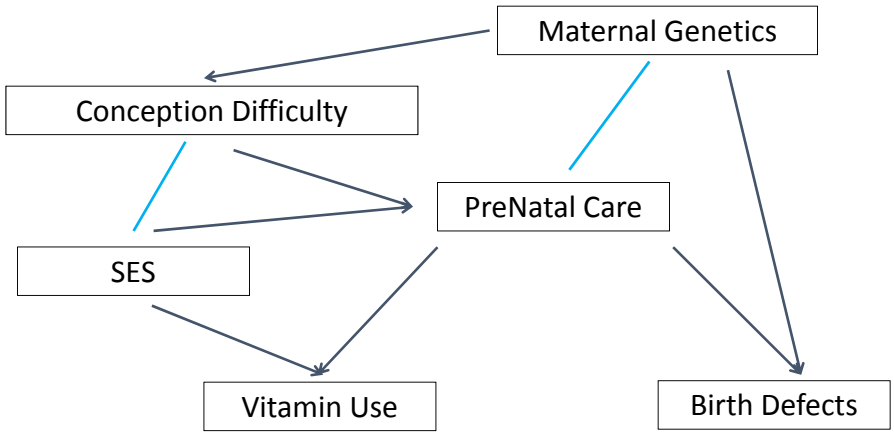


OSCTR Oklahoma Shared Clinical & Translational Resources

49

Adjustment

Step 4: Connect any two causes sharing a common effect

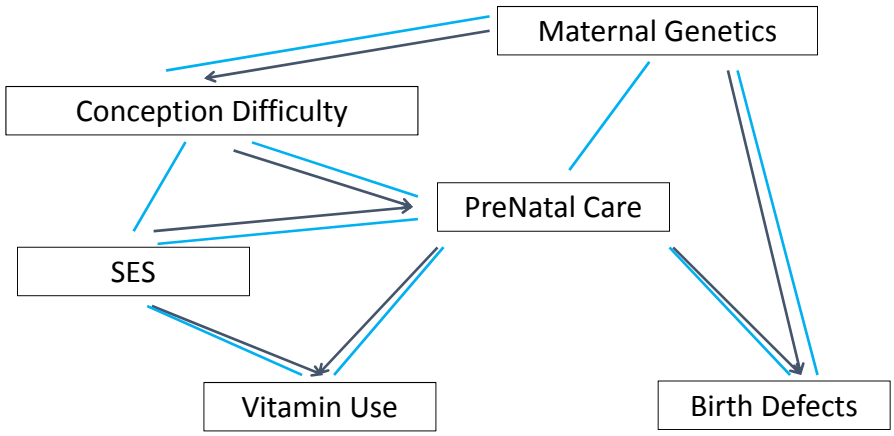


OSCTR Oklahoma Shared Clinical & Translational Resources

50

Adjustment

Step 5: Strip arrow heads from all edges

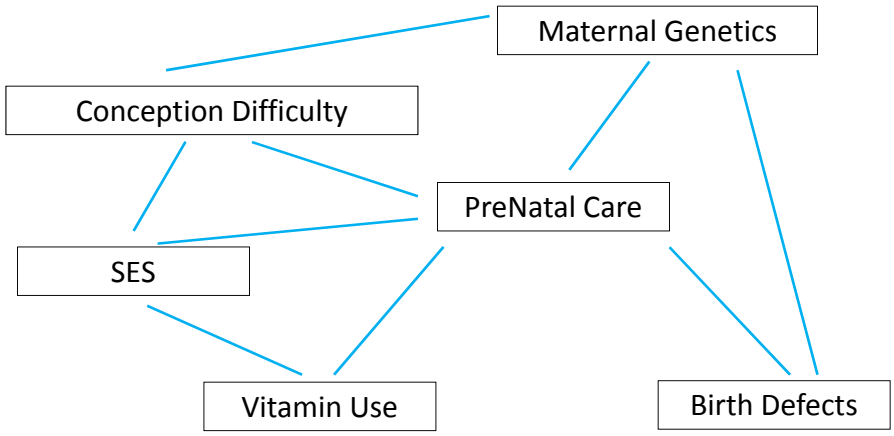


OSCTR Oklahoma Shared Clinical & Translational Resources

51

Adjustment

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

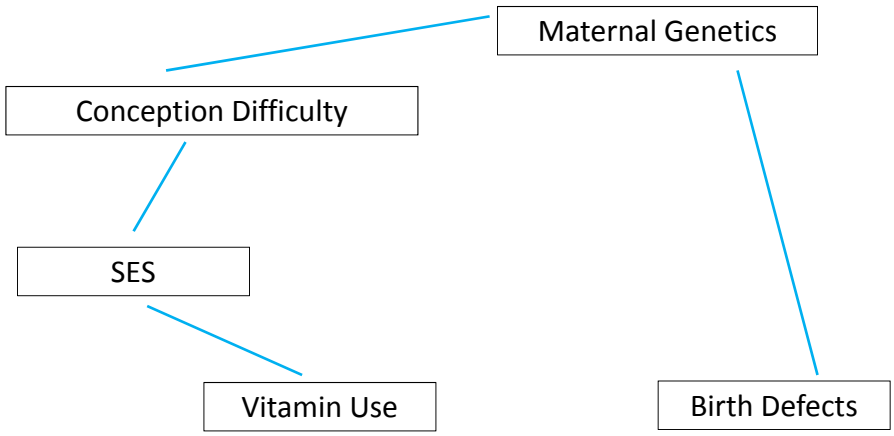


OSCTR Oklahoma Shared Clinical & Translational Resources

52

Adjustment

Test: Are vitamins and birth defects still connected?

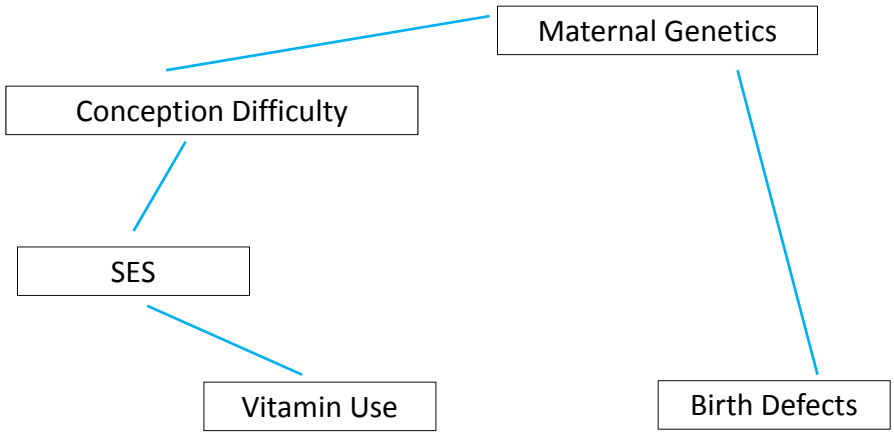


OSCTR Oklahoma Shared Clinical & Translational Resources

53

Adjustment

YES: How else can we control for confounding?



OSCTR Oklahoma Shared Clinical & Translational Resources

54

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.

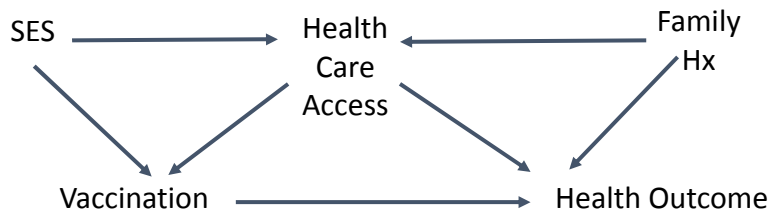
If E and D are still connected, additional adjustment is required.

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

55

Assessing confounding

Remove direct effect of E on D

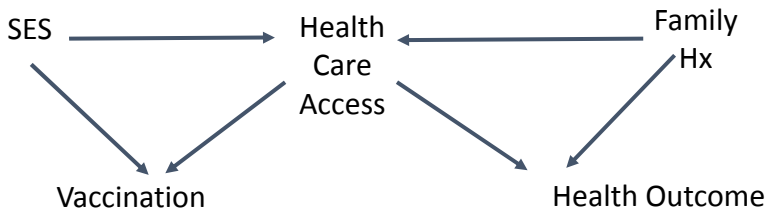


Source: Jewell, Chap. 8

56

Assessing confounding: Another Example

Step 1: Delete direct effects of exposure of interest

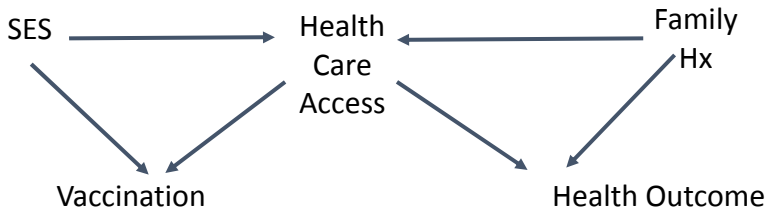


Source: Jewell, Chap. 8

57

Assessing confounding

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

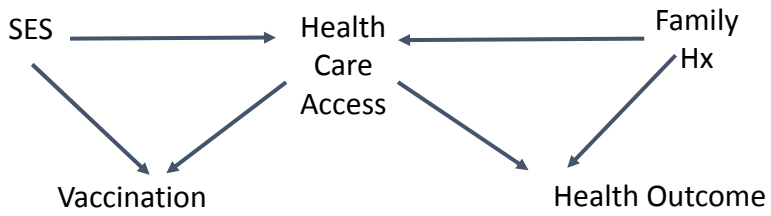


Source: Jewell, Chap. 8

58

Assessing confounding

Step 3: Delete all direct effects of E

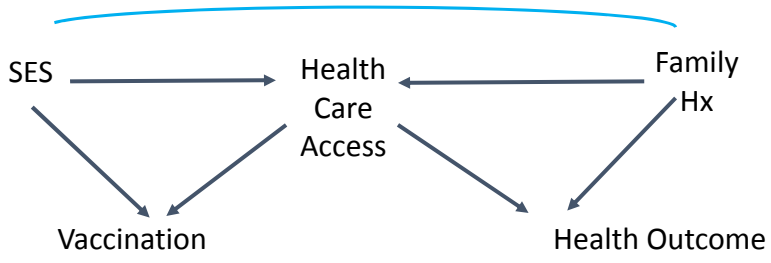


Source: Jewell, Chap. 8

59

Assessing confounding

Step 4: Connect any two causes sharing a common effect

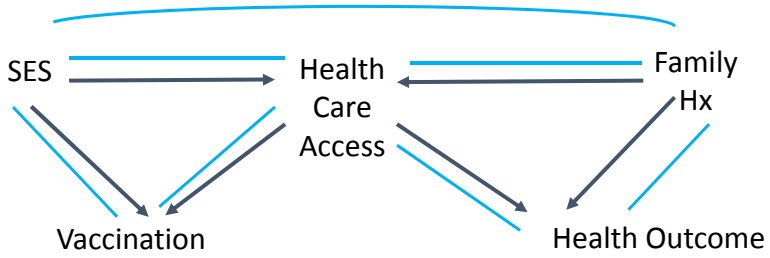


Source: Jewell, Chap. 8

60

Assessing confounding

Step 5: Delete arrow heads from all edges

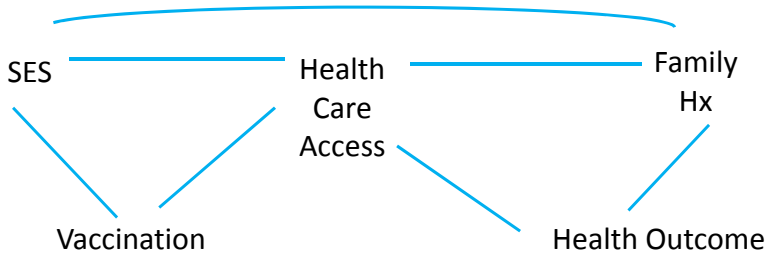


Source: Jewell, Chap. 8

61

Assessing confounding

Step 6: Delete C and all associated edges



Source: Jewell, Chap. 8

62

Assessing confounding

Step 6: Delete C and all associated edges

Are E and D still connected?

Source: Jewell, Chap. 8

63

Overadjustment Bias and Unnecessary Adjustment in Epidemiologic Studies

Enrique F. Schisterman,[a] Stephen R. Cole,[b] and Robert W. Platt[c]

Abstract: Overadjustment is defined inconsistently. This term is meant to describe control (eg, 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 descending 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.

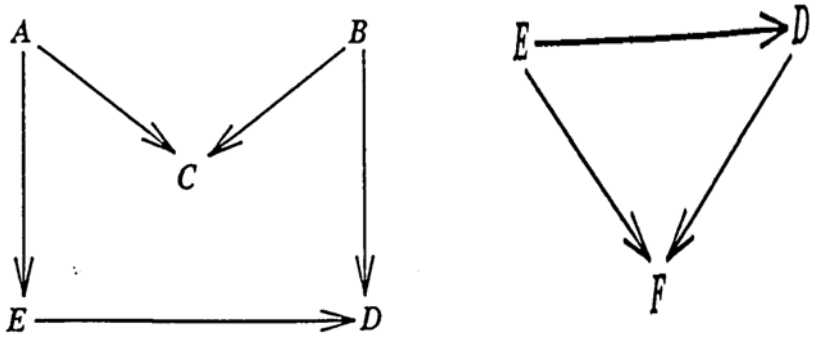
(*Epidemiology* 2009;20: 488–495)

confounding[1] and selection biases[2,3] have been discussed extensively in the epidemiologic literature, the concept of "overadjustment" has had relatively little attention. The definition of overadjustment remains vague and the causal structure of this concept has not been well described.

The Dictionary of Epidemiology[4] cites a seminal paper by Breslow[5] in broadly defining overadjustment as "Statistical adjustment by an excessive number of variables or parameters, uninformed by substantive knowledge (eg, lacking coherence with biologic, clinical, epidemiological, or social knowledge). It can obscure a true effect or create an apparent effect when none exists." Rothman and Greenland[6] discuss overadjustment in the context of intermediate variables: "Intermediate variables, if controlled in an analysis, would usually bias results towards the null. . . . . Such control of an intermediate may be viewed as a form of overadjustment." One also finds reference to the term overadjustment in settings with unnecessary control for variables.[7] In summary, overadjustment sometimes means control (eg, by regression

64

Unnecessary and Harmful Adjustment



Greenland et al., 1999

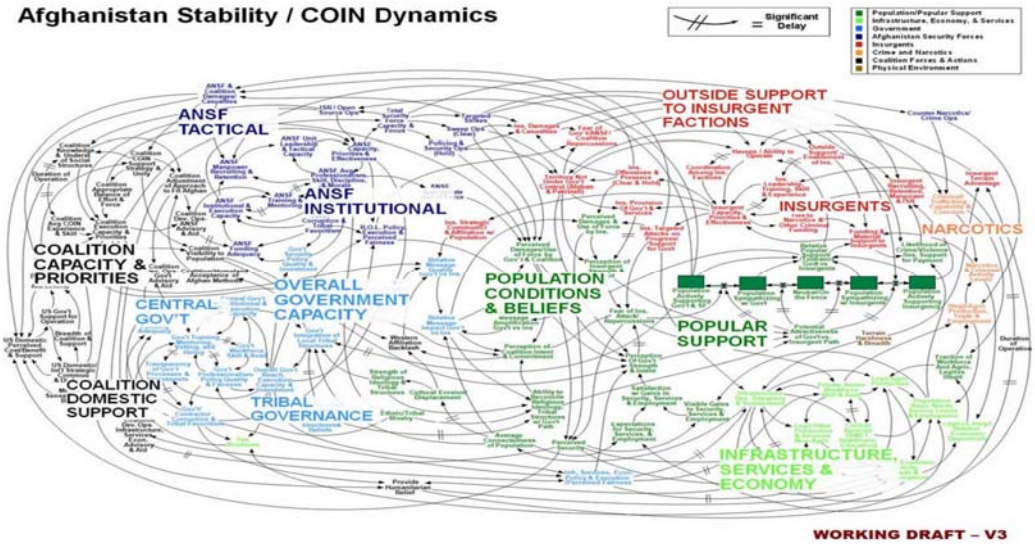
65

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!

66

The ultimate complex causal graph!



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

Madhukar Pai, McGill University

Sources

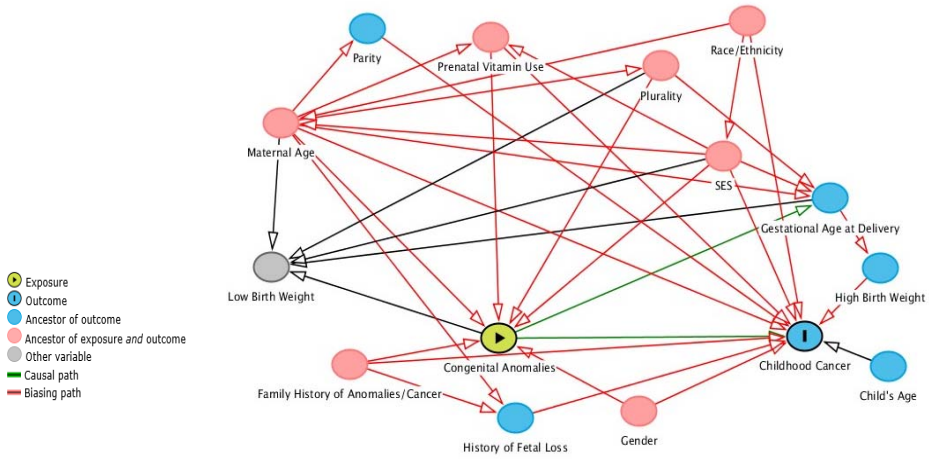
- Jewell, N. *Statistics for Epidemiology,* Chapter 8
- Shrier and Platt (2008). Reducing bias through directed acyclic graphs. BMC Medical Research Methodology. 8:70
- Glymour, M. "Using causal diagrams to understand common problems in social epidemiology," in *Methods in Social Epidemiology.*
- Petersen, 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*
- Hernan, M *et al.* Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. *Am J Epi* 2002; 155: 176 – 84.
- Greenland, S *et al.* Causal diagrams for epidemiologic research. *Epidemiology* 1999; 10: 37 – 48.

68

Applied Example

OSCTR Oklahoma Shared Clinical & Translational Resources

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

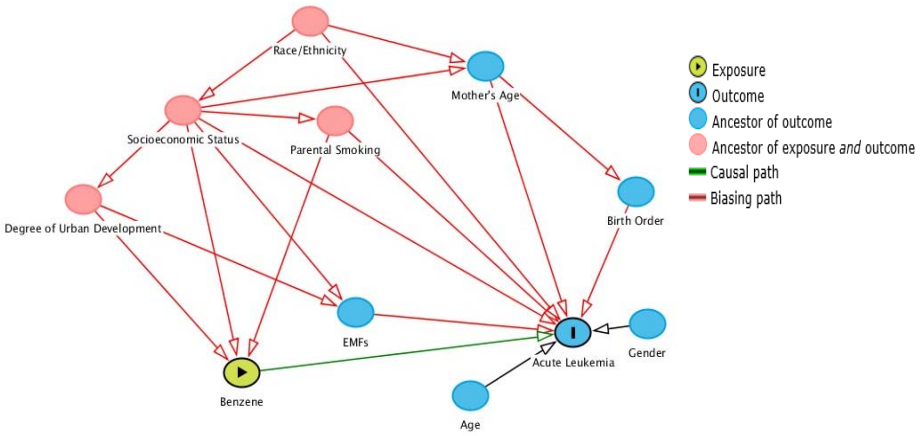
OSCTR Oklahoma Shared Clinical & Translational Resources

Table 37. Relationship between potential confounding variables, congenital anomalies and childhood cancer for DAG.

Variable	Association with Congenital Anomalies	Association with Cancer	Association with other variables	Collider	Included in Minimally Sufficient Set	Available for Analysis	Source
Plurality	Parent		Parents: Maternal age, CA Children: Low birth weight, gestational age	No	Yes	Limited to singletons	Carozza, 2012; Sunderam, 2012
Prenatal Vitamin Use	Parent	Parent	Parents: SES, maternal age Children: CA, CC	Yes	Yes	Can analyze prenatal care (yes/no), but only 1.5% did not have prenatal care	CDC, 2008; Ross, 2005; Thompson, 2001; Wen 2002; Werler, 1999
SES	Parent	Parent	Parents: Race/ethnicity, CA, CC Children: Prenatal vitamin use, maternal age, low birth weight, gestational age	Yes	Yes	Analyze maternal education (≤ high school v. > high school)	Mertens, 1998; Menegaux, 2005; Botto, 2013; Ries, 1999; Yang, 2008; Carolan, 2011; Dubay, 2001; CDC, 2008; DHHS, 2011
Maternal Age	Parent	Parent	Parents: SES, Race/ethnicity Children: Prenatal vitamin use, plurality, parity, low birth weight, CA, CC, history of fetal loss	Yes	Yes	Yes	Altmann, 1998; Botto, 2013; Fisher, 2012; Agha, 2005; Carozza, 2012; Partap, 2011; Ries, 1999; CDC, 2011; WHO, 2013; Usta 2008; Carolan, 2011; DHHS, 2011; Nybo Anderson, 2000
Race/ethnicity		Parent	Parents: Children: maternal age, SES, CC	No	No	Yes	Carozza, 2012; Botta, 2013; Mertens, 1998; Partap, 2011; Ries, 1999; ;APA, 2013, DHHS, 2011
Child's Age		Parent	Child: CC		No	Age at diagnosis	Menegaux, 2005; Zierhut, 2011; Botto, 2013; Savitz 1994; Merks, 2008; Windham, 1985

OSCTR Oklahoma Shared Clinical & Translational Resources

Example: Benzene and childhood leukemia

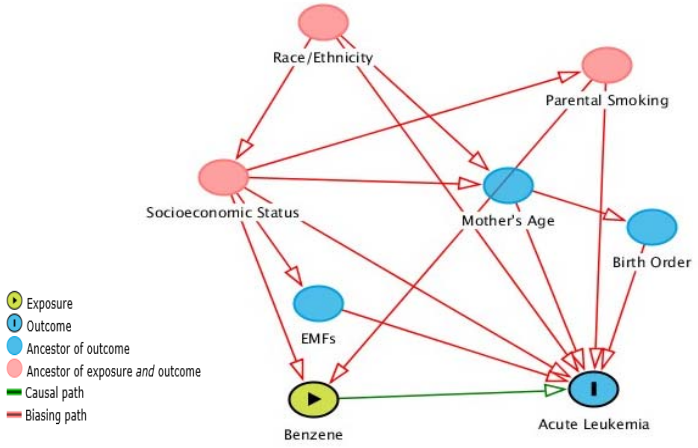


Minimally sufficient adjustment sets for estimating the total effect of benzene on acute leukemia:

- Degree of Urban Development, Parental Smoking, Socioeconomic Status (maternal education)
- EMFs, Parental Smoking, Socioeconomic Status (maternal education)

OSCTR Oklahoma Shared Clinical & Translational Resources

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)

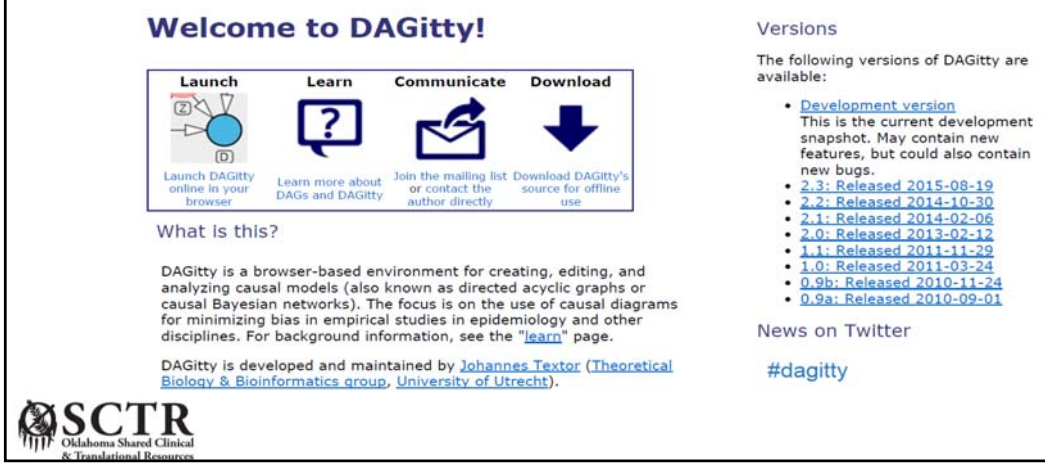
Janitz et al., 2016

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/



OSCTR Oklahoma Shared Clinical & Translational Resources