Directed Acyclic Graphs (DAGs): Introduction and applications

Hsin-Yi Weng, BVM, MPH, PhD
Associate Professor
Department of Comparative Pathobiology
Purdue University – College of Veterinary Medicine

PURDUE UNIVERSITY | College of Veterinary Medicine

Outline

- Introduction to Directed Acyclic Graphs (DAGs)
- Confounder identification using DAGs
- DAGitty: An online tool for building and analyzing DAGs

Epidemiology is a measurement science: the goal is to quantify an **unbiased** relationship between an exposure and an outcome

Making inference from observational studies

- Always needs to consider confounding
- Confounder identification requires causal assumptions

Traditional definition of a confounder

1. Associated with study exposure
2. Associated with study outcome
3. Not affected by study exposure or outcome

Most common way to select confounders for adjustment

Select among all available variables based on statistical significance (e.g., $P < 0.05$).

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

Flaws in the traditional approaches for confounder identification

- Traditional definition is not testable using the data collected
- Selecting confounders for adjustment based on statistical significance may miss important confounders or result in harmful adjustment

Directed Acyclic Graphs (DAGs): a type of causal diagrams

PURDUE UNIVERSITY | College of Veterinary Medicine

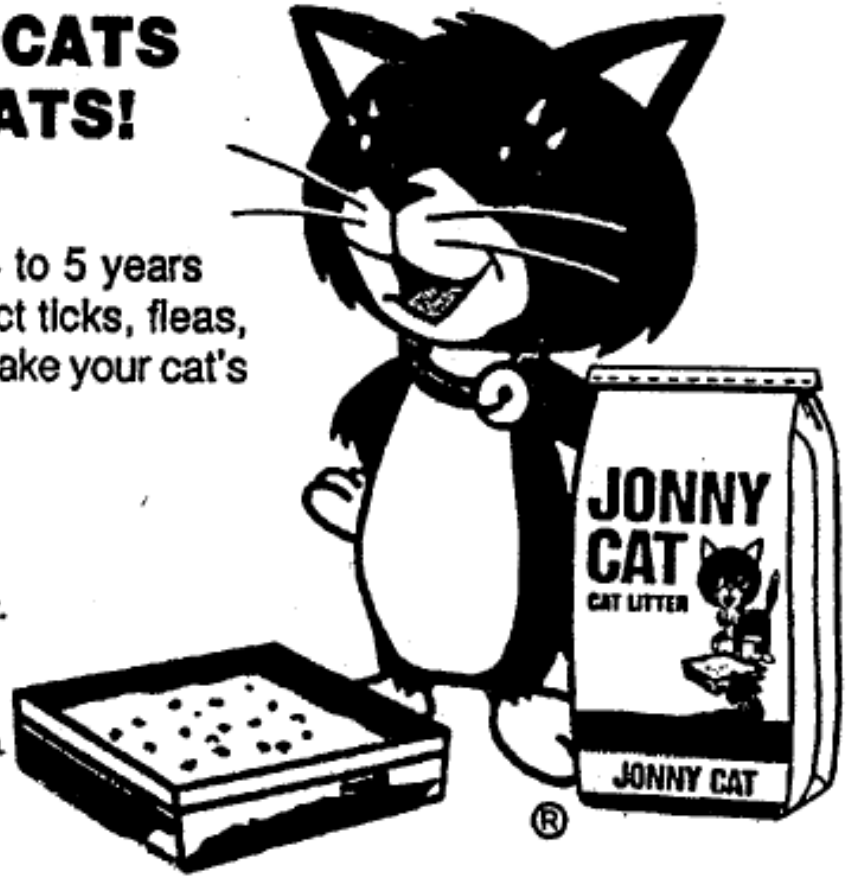
Example: Litterbox and longevity in cats

THE HAPPIEST CATS ARE JONNY CATS!

Good News! Research shows:
cats that use a litter box can live 4 to 5 years longer and are less likely to contract ticks, fleas, and diseases. A litter box helps to make your cat's life a safe and healthy one.

HOW TO USE A CAT LITTER BOX

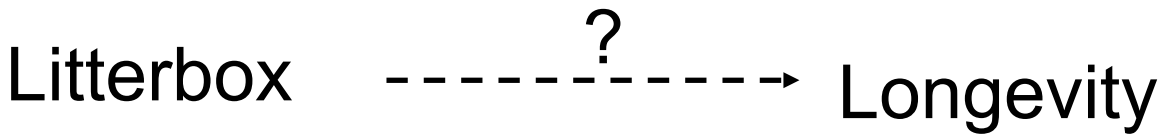
1. Spread three inches of litter in bottom of clean cat box.
2. Occasionally remove solid wastes.
3. Dispose of used litter in the trash only.
4. Frequently wash litter box with hot water and ammonia.
5. Refill with Jonny Cat!



Courtesy of Dr. Philip Kass, UC Davis

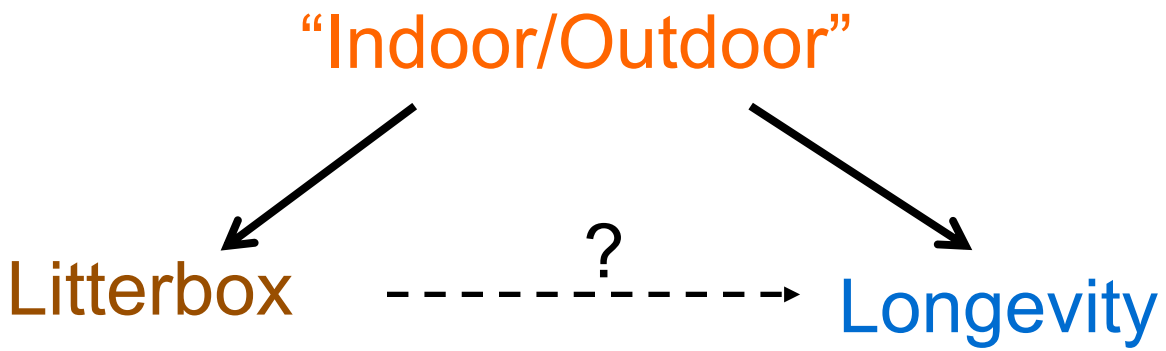
Is “indoor/outdoor” status a confounder?

To assess the association between litterbox use and longevity in cats, should I adjust for indoor/outdoor status as a confounder?

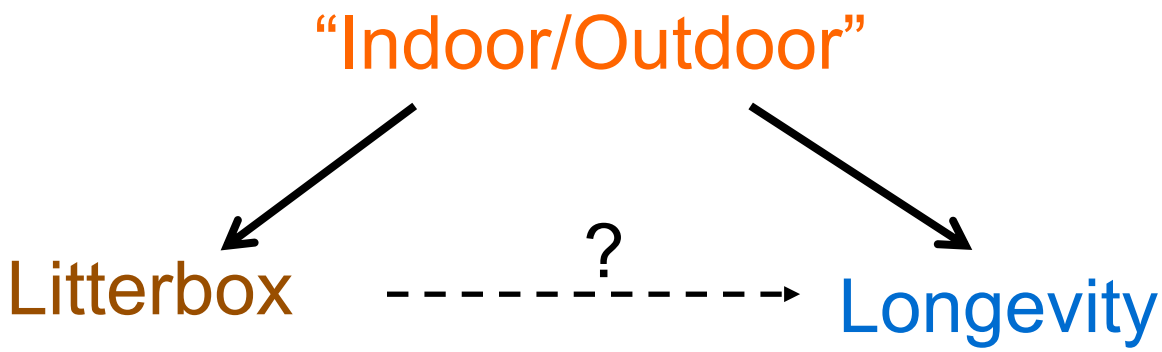


The role of study variables

- Exposure: Litterbox
- Outcome: Longevity
- Covariate: Indoor/outdoor
- **Association of interest**: between litterbox use and longevity of cat



Elements and terminology

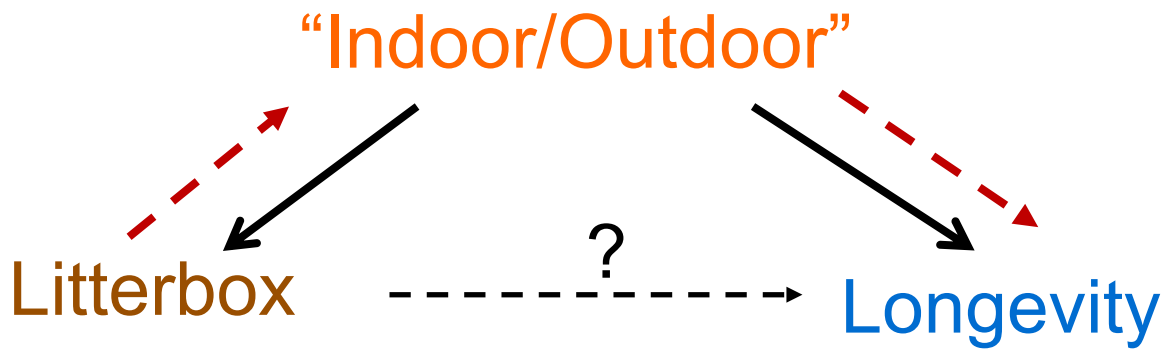


In this diagram, Indoor/Outdoor is the parent of both Litterbox and Longevity. The relationships presented are directed (indicated by arrows) and acyclic (the arrows never point from a given variable back to any other variable in its past). This type of causal diagram is called **D**irected **A**cyclic **G**raph (a.k.a. DAG)

Marginal association

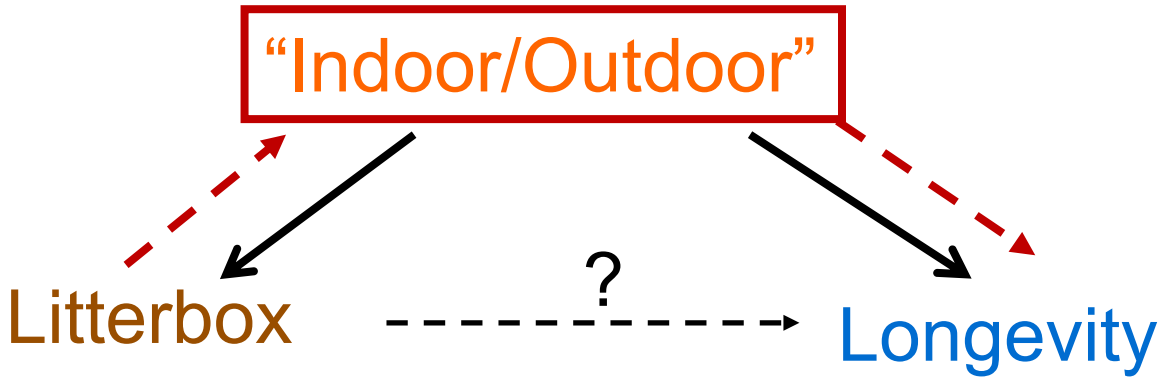
- A marginal association between two variables in a DAG requires that there be an unblocked path between them.
- Two kinds of unblocked paths:
  - Directed path, implying a causal association
  - Backdoor path, implying a non-causal association
- Thus, to correctly estimate the causal association, all backdoor paths should be *blocked*.

Backdoor path and confounding



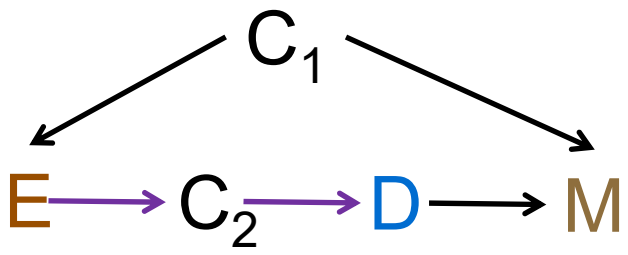
- There is a **backdoor path** from Litterbox to Longevity through Indoor/Outdoor.
- This suggests that Indoor/Outdoor is a confounder, based on this set of causal assumptions.

Backdoor path and confounding



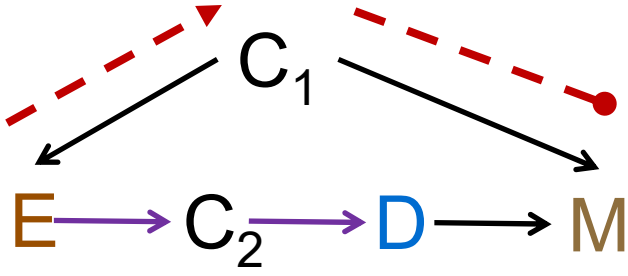
- To correctly estimate the (causal) association (i.e., via a directed path) between litterbox and longevity in cats, we need to block the backdoor path.
- This can be done by statistical adjustment.

Another DAG



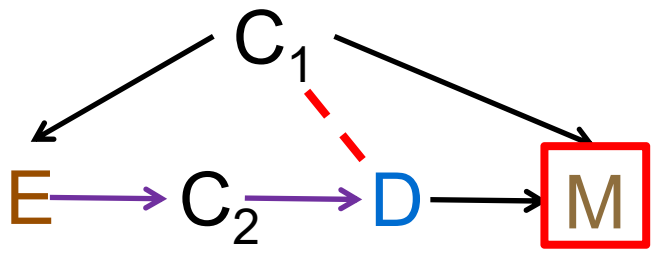
- In this DAG, E is associated with D through an intermediate C_2. This is a directed path.
- The path through C_1 and M in figure 2 is blocked at M because M is a collider.

More on colliders



- A collider is a variable that has two or more arrowheads pointing to it (e.g., M).
- A path is blocked if it enters a variable through an arrowhead and only can exit through another arrowhead (e.g., $C_1 \rightarrow M \leftarrow$).

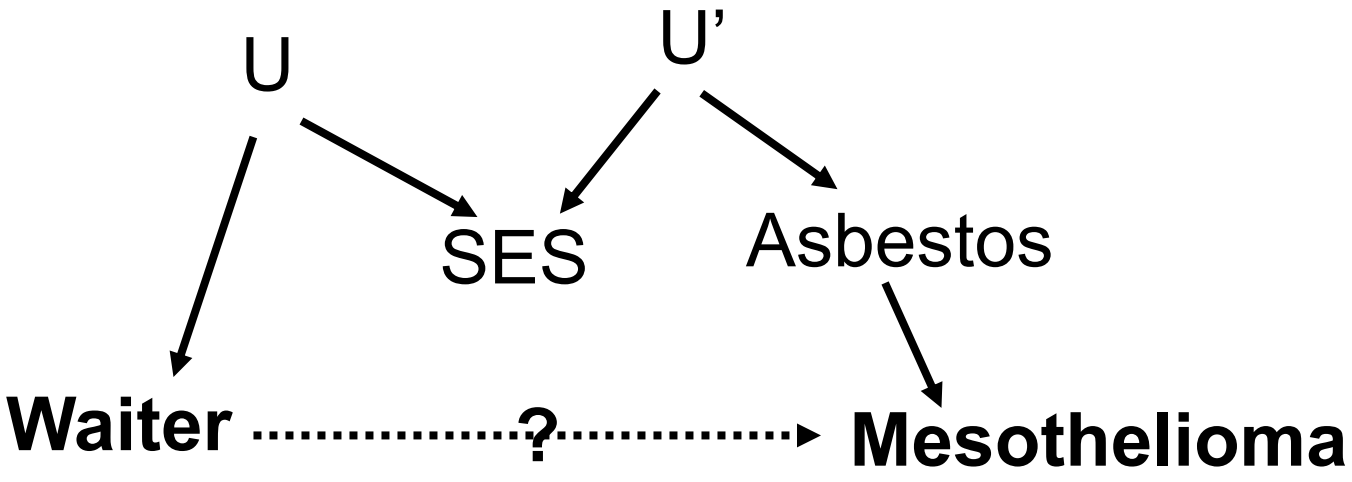
Adjust for a collider will "marry" its parents



Statistically adjust for M, a collider, will create a marginal association between C1 and D, which results in a backdoor path: E ← C1 --- D.

Exercise:
Using DAGs to identify confounders for control (Block all backdoor paths!)

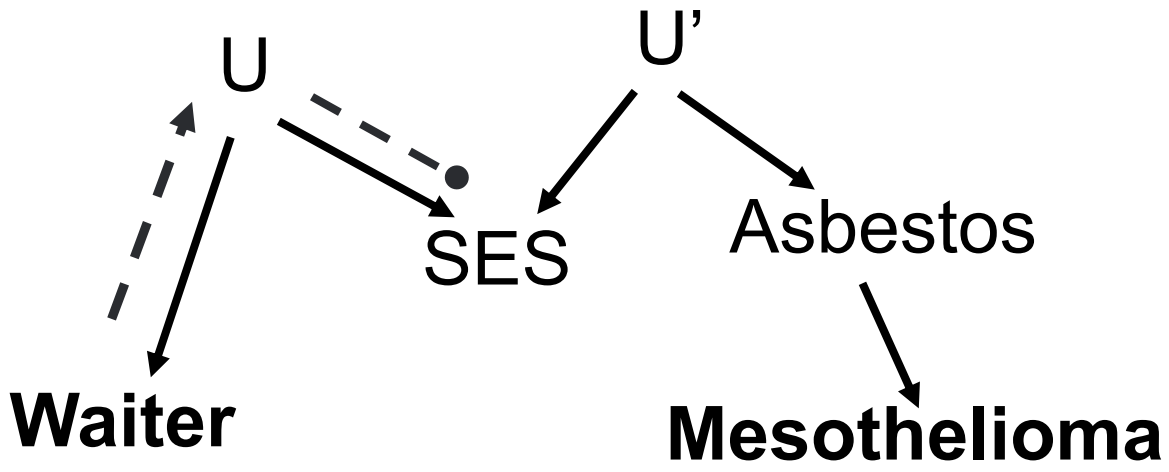
Waiter and Mesothelioma: Should we adjust for SES?



Having worked as a waiter (Waiter) and socioeconomic status (SES) are associated with each other, as they share common unmeasured causes (U). Similarly, there are unmeasured factors (U') that determine an association between SES and working in occupations entailing exposure to asbestos (Asbestos), which, in turn, causes mesothelioma.

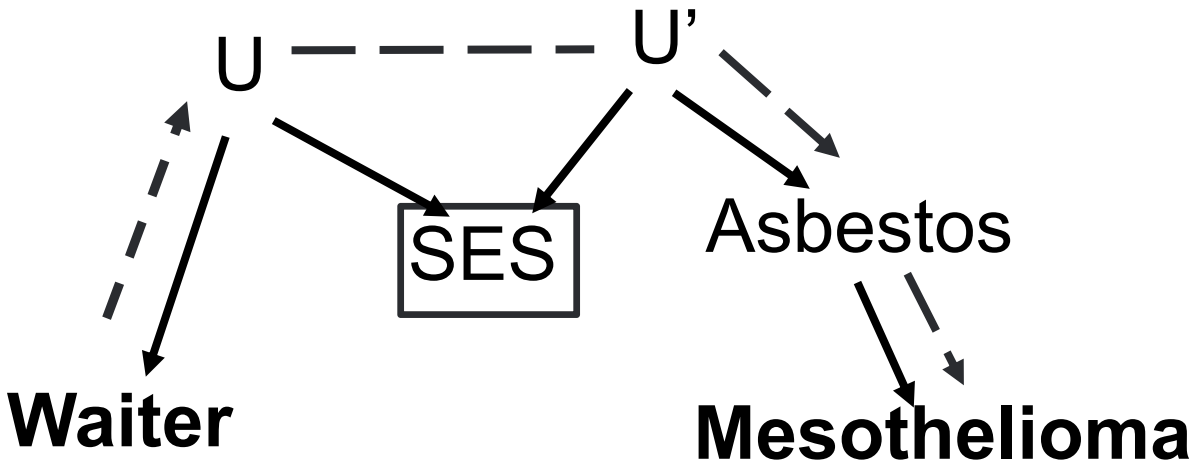
Richiardi et al. 2008

Unnecessary or harmful adjustment



In this example, there is no backdoor path connecting Waiter to Mesothelioma, and hence no need to adjust for any of the covariates.

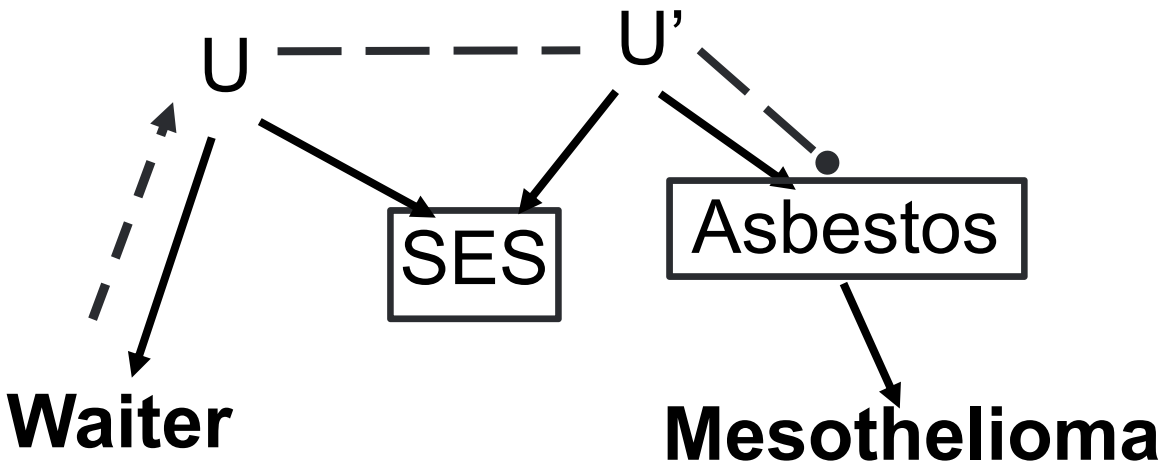
Unnecessary or harmful adjustment



SES is a collider. Adjusting for SES will create a marginal association between U and U', thus a backdoor path:

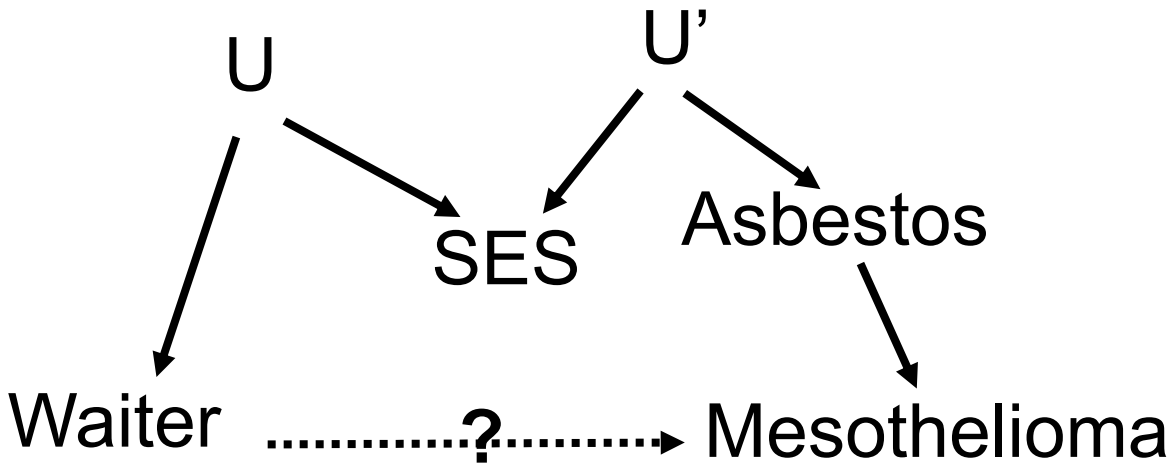
Waiter ← U---U' → Asbestos → Mesothelioma.

Unnecessary or harmful adjustment



Adjusting for both SES and Asbestos, using certain regression models, may cause loss of precision (i.e., increase SE estimates).

Conflicting results when compared to traditional approaches



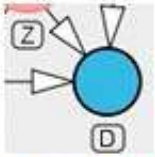
- SES is associated with both Waiter and Mesothelioma and is not affected by either **Suggests to control for SES**
- SES and Asbestos are associated with Mesothelioma, thus may be selected based on statistical significance

$$g(Mesothelioma) = \beta_0 + \beta_1 SES + \beta_1 Asbestos$$

Confounder identification requires causal assumptions

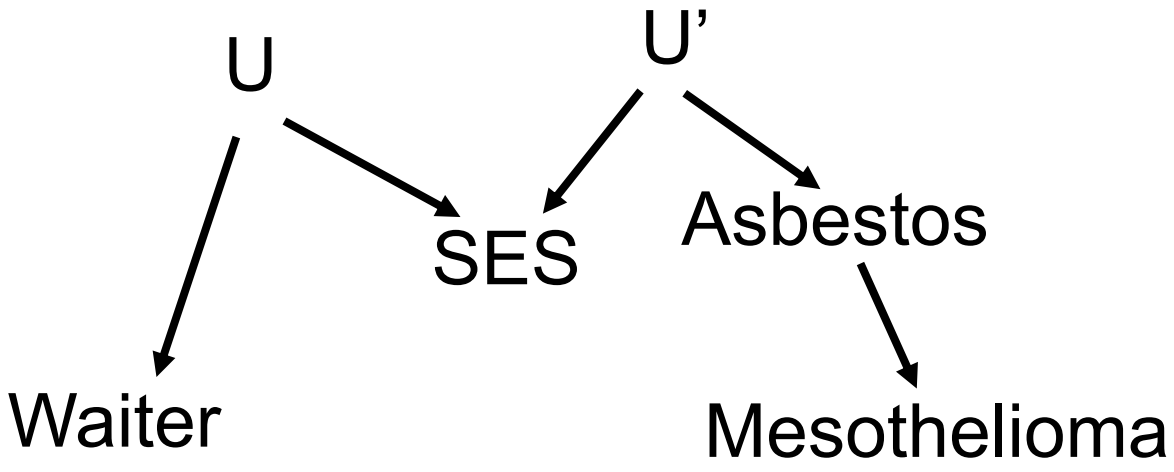
- **Confounder identification always requires causal assumptions**
- Traditional definition is not testable using the data collected
- Selecting confounders based on statistical significance may miss important confounders and/or result in harmful adjustment

Using DAGitty to build DAGs and identify confounders for control

Launch	Download	Learn	Code
 Launch DAGitty online in your browser.	 Download DAGitty's source for offline use.	 Learn more about DAGs and DAGitty.	 The R package "dagitty" is available on CRAN or github.

http://www.dagitty.net/

Exercise



Construct the above DAG in DAGitty

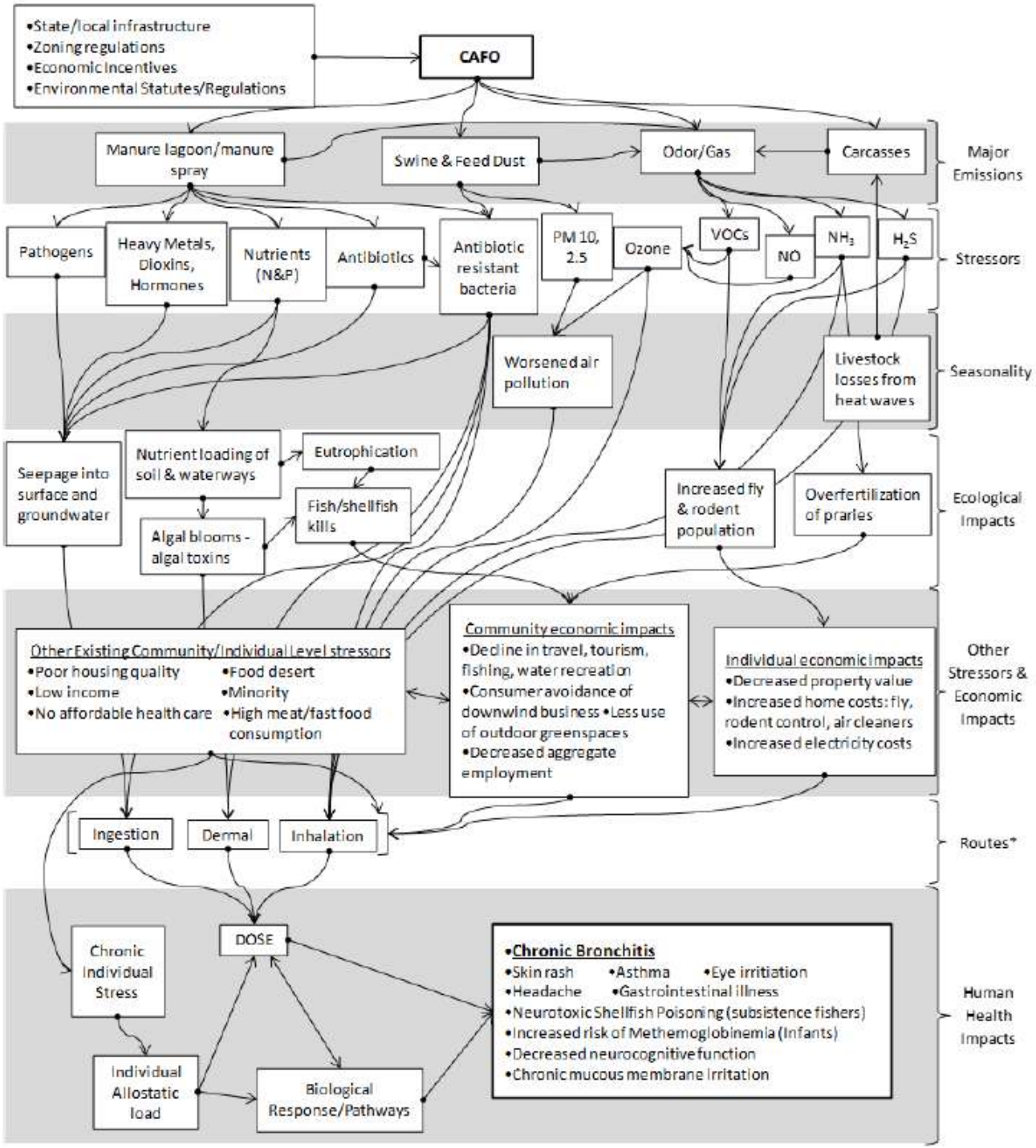
http://www.dagitty.net/

Example 2

Causal inference in cumulative risk assessment - A hypothetical community located near a concentrated animal feeding operation

BL Brewer et al. Causal inference in cumulative risk assessment: The roles of directed acyclic graphs. Env Int, 2017;102:30-41.

PURDUE UNIVERSITY | College of Veterinary Medicine



Conceptual model for a cumulative risk assessment in a hypothetical community near a concentrated animal feeding operation. *Brewer et al. (2017)*

Multiple exposures and the code

- Users can define multiple exposures in DAGitty.
- DAGitty has built-in algorithm to analyze complicated DAGs.
- DAGitty will automatically generate the code while building a DAG. The code can be saved and used to reproduce the DAG.
- Users also can export R code generated by DAGitty.

```
dag {
"AR Bact" [pos="-1.629,-0.115"]
"Air pollution" [pos="-1.540,0.280"]
"Algal blooms/toxins" [pos="-1.721,1.105"]
"Allostatic load" [pos="-0.492,1.529"]
"Animal&Feed" [pos="-1.474,-1.358"]
"Chronic Bronchitis" [outcome,pos="-1.121,1.738"]
"Chronic stress" [pos="-0.346,1.099"]
"Community economic impacts" [pos="-0.352,0.692"]
"Fish/shellfish kills" [pos="-1.170,0.453"]
"Fly&Rodent" [pos="-0.326,0.214"]
"HM,Dioxins,Hormones" [pos="-2.037,-0.073"]
"Individual economic impacts" [pos="-1.440,1.141"]
"Manure lagoon" [pos="-1.919,-1.029"]
"N&P" [pos="-1.902,-0.431"]
"Nutrient loading of soil & waterways" [pos="-1.825,0.872"]
"Odor/Gas" [pos="-1.202,-0.868"]
"Other Community/Individual stressors" [pos="-0.665,1.260"]
"PM 10,2.5" [exposure,pos="-1.336,-0.473"]
"Surface&GW" [pos="-2.169,1.278"]
"Swine & Feed Dust" [pos="-1.494,-0.880"]
Antibiotics [pos="-1.744,-0.354"]
Carcasses [pos="-0.731,-1.185"]
Eutrophication [pos="-1.569,0.597"]
H2S [exposure,pos="-0.449,-1.089"]
NH3 [exposure,pos="-0.498,-0.491"]
NO [exposure,pos="-0.777,-0.491"]
Overfertilization [pos="-1.153,0.053"]
Ozone [exposure,pos="-1.009,-0.551"]
Pathogens [pos="-2.149,-0.509"]
VOCs [exposure,pos="-0.702,-0.802"]
"AR Bact" -> "Chronic Bronchitis"
"AR Bact" -> "Surface&GW"
"Air pollution" -> "Chronic Bronchitis"
"Algal blooms/toxins" -> "Chronic Bronchitis"
"Algal blooms/toxins" -> "Fish/shellfish kills"
"Allostatic load" -> "Chronic Bronchitis"
"Animal&Feed" -> "Manure lagoon"
"Animal&Feed" -> "Odor/Gas"
"Animal&Feed" -> "Swine & Feed Dust"
"Animal&Feed" -> Carcasses
"Chronic stress" -> "Allostatic load"
"Community economic impacts" -> "Chronic Bronchitis"
"Fish/shellfish kills" -> "Community economic impacts"
"Fly&Rodent" -> "Chronic Bronchitis"
"Fly&Rodent" -> "Individual economic impacts"
"HM,Dioxins,Hormones" -> "Surface&GW"
"Individual economic impacts" -> "Algal blooms/toxins"
"Individual economic impacts" -> "Chronic Bronchitis"
"Manure lagoon" -> "AR Bact"
"Manure lagoon" -> "HM,Dioxins,Hormones"
"Manure lagoon" -> "N&P"
"Manure lagoon" -> Antibiotics
"Manure lagoon" -> Carcasses
"Manure lagoon" -> Pathogens
"N&P" -> "Nutrient loading of soil & waterways"
"N&P" -> "Surface&GW"
"Nutrient loading of soil & waterways" -> "Algal blooms/toxins"
"Nutrient loading of soil & waterways" -> Eutrophication
"Odor/Gas" -> "Chronic Bronchitis"
"Odor/Gas" -> H2S
"Odor/Gas" -> NH3
"Odor/Gas" -> NO
"Odor/Gas" -> VOCs
"Other Community/Individual stressors" -> "Chronic Bronchitis"
"Other Community/Individual stressors" -> "Chronic stress"
"Other Community/Individual stressors" -> "Community economic impacts"
"Other Community/Individual stressors" -> "Individual economic impacts"
"PM 10,2.5" -> "Air pollution"
"Surface&GW" -> "Chronic Bronchitis"
"Swine & Feed Dust" -> "AR Bact"
"Swine & Feed Dust" -> "Odor/Gas"
"Swine & Feed Dust" -> "PM 10,2.5"
Antibiotics -> "AR Bact"
Antibiotics -> "Surface&GW"
Carcasses -> "Odor/Gas"
Eutrophication -> "Fish/shellfish kills"
H2S -> "Chronic Bronchitis"
H2S -> "Fly&Rodent"
NH3 -> "Chronic Bronchitis"
NH3 -> "Fly&Rodent"
NH3 -> Overfertilization
NO -> Ozone
Overfertilization -> "Community economic impacts"
Ozone -> "Air pollution"
Ozone -> "Chronic Bronchitis"
Pathogens -> "PM 10,2.5"
Pathogens -> "Surface&GW"
VOCs -> "Fly&Rodent"
VOCs -> Ozone
}
```

PURDUE UNIVERSITY | College of Veterinary Medicine

Take-home message

- Confounder identification always requires causal assumptions
- Construct a DAG **before** data collection to ensure that data on important confounders are obtained
- DAGs guide data analysis to avoid unnecessary and harmful adjustment
- There are online applications available for constructing and analyzing DAGs

Acknowledgments

I thank Dr. David Miller for inviting me to give this talk and also for providing me feedback during the practice run.

PURDUE UNIVERSITY | College of Veterinary Medicine

References & Resources

1. Greenland et al. Causal diagrams for epidemiologic research. Epidemiol, 1999;10:37-48.
2. Hernán et al. Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. Am J Epidemiol, 2002;155:176-184.
3. Robins. Data, design, and background knowledge in etiologic inference. Epidemiol, 2001;12:313-320.
4. Textor et al. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. Int J Epidemiol, 2016;45:1887-1894. (http://www.dagitty.net/)

References & Resources

5. Brewer et al. Causal inference in cumulative risk assessment: The roles of directed acyclic graphs. Env Int, 2017;102:30-41.
6. Richiardi et al. Using directed acyclic graphs to consider adjustment for socioeconomic status in occupational cancer studies. J Epidemiol Community Health, 2008;62:e14.
7. Naimi et al. Assessing the component associations of the healthy worker survivor bias: occupational asbestos exposure and lung cancer mortality. Ann Epidemiol, 2013;23:334-341.