National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

NIDDK-CR Resources for Research

Data Science and Meet the Expert Webinar Series

February 26, 2026

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

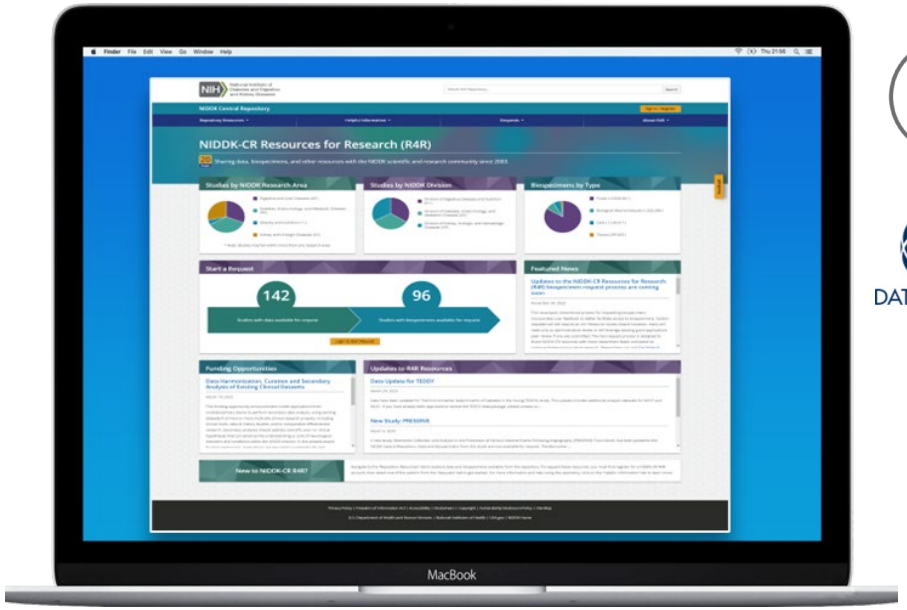
NIDDK Central Repository Overview

Mission

Established in 2003 to **facilitate sharing of data, biospecimens, and other resources** generated from studies supported by NIDDK and within NIDDK's mission by making these **resources available for request to the broader scientific and research community**.

- Supports receipt and distribution of data and biospecimens in a manner that is ethical, equitable, and efficient
- Enables investigators not involved with the original work to test new hypotheses without the need to collect new data or biospecimens
- Promotes FAIR (Findable, Accessible, Interoperable, and Reusable) and TRUST (Transparency, Responsibility, User focus, Sustainability, and Technology) principles

Recorded past tutorials, webinars, and other educational resources can be found on the NIDDK-CR website



CORE TRUST SEAL

WORLD DATA SYSTEM

Imaging Data Files
15.8 M

Clinical Datasets
>11,400
from 194 clinical studies

Biospecimens
>16 M

Registered Users
7,256

Weekly Users
>4,500

Public Releases
>900

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

NIDDK Analytics Workbench Environment

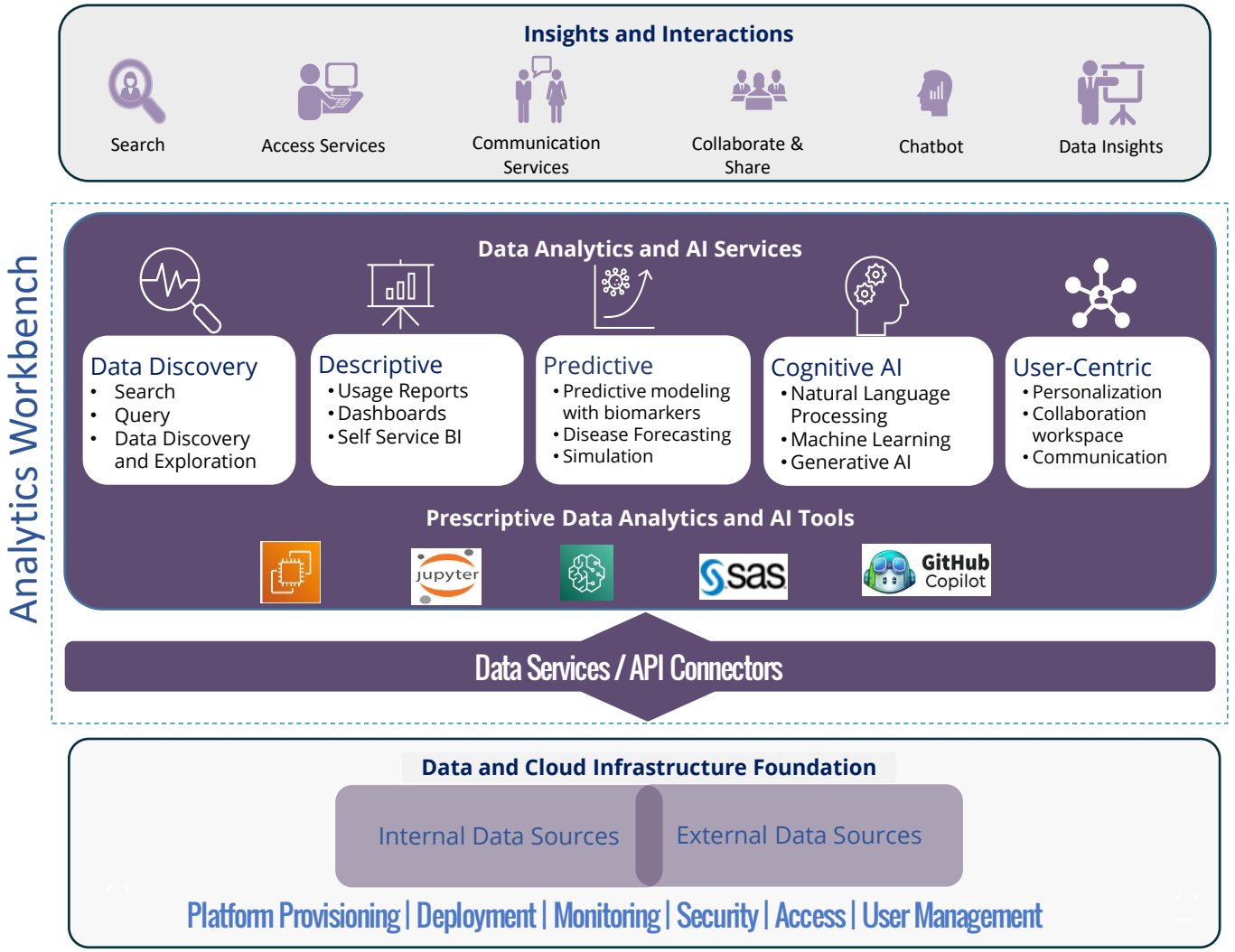
Streamlining *end-to-end data science lifecycle and discovery of data-driven biomedical insights.*

Innovation and ease of use

A cloud-based analytics environment where researchers and data scientists can access a suite of integrated analytics tools and cloud computing resources to participate in data challenges and AI innovation.

Expected Benefits of Analytics Workbench:

- **Promote Collaboration**
- **Support AI Innovation**
- **Minimize Data Movement**
- **Improve User Experience**
- **Discover Data Insights**
- **Advance NIDDK Research Mission**



National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Secondary Data Science and Meet the Expert Webinar Series

About the Series

- Aims to accelerate data science and AI-driven biomedical research by fostering collaboration between biomedical researchers and experts in the field
- Monthly webinar held on the **last Thursday of each month**

Upcoming Webinars

Neural Networks for Detecting Subtle Epileptogenic Lesions and Supporting Clinical Decision-Making

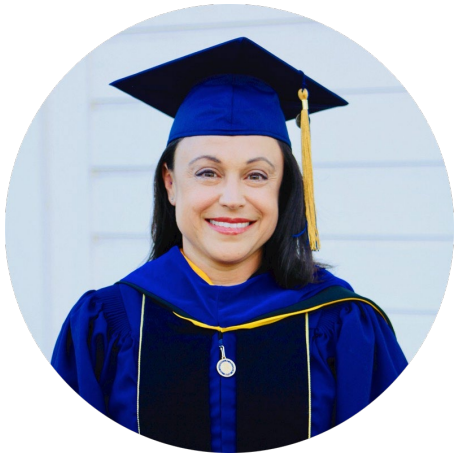
- Date: **March 26, from 2-3pm ET**
- Expert: Dr. Mathilde Ripart, Kings College London

Scan the QR code to register

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Meet the Expert



Dr. Courtney D. Shelley, PhD, is a Health Data Scientist at Booz Allen Hamilton, where she focuses on data science education and AI-readiness of health-related data. She has supported the NIH Office of Data Science Strategy to develop online data science learning resources for pre-college and collegiate audiences, and to assess data science education across US universities to promote collaborative research between biomedical researchers and AI professionals. Prior to working at Booz Allen Hamilton, Dr. Shelley worked at Los Alamos National Laboratory, where she received the Postdoctoral Distinguished Performance Award for COVID-19 response efforts at local, state, and federal levels, as well as conducted research in suicide prevention with the support of the Department of Veterans Affairs and Million Veteran Program. She completed her PhD in Epidemiology with a focus on causal inference at University of California, Davis.

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Causal Inference in Clinical Research Data

NIDDK-CR Data Science Meet the Experts Webinar Series

February 26, 2026

Presented by: Booz Allen Hamilton

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Agenda

1. Correlation vs Causation
   - Statistical relationships (independent vs. dependent variables)
   - Role of expert knowledge and DAGs
2. Three-Variable Relationships
   - Mediators, Confounders, and Colliders
3. Real-World Insights
   - Birth-Weight Paradox and Selection Bias
   - Berkson's Bias in observational studies
4. Key Biases and Paradoxes
   - Selection bias and its effects
   - Simpson's Paradox: Aggregated vs. stratified data
5. Causal Inference Analysis/Demo in NIDDK-CR Workbench
6. Takeaways and Q&A

Correlation

A B

I have two variables, A and B.

A and B are **independent**

*(i.e., there is **no** association)*

We can test this assumption with a simple correlation:

```
corr(A, B)
0.0129644
```

We can denote our findings as **A ⫫ B**.

You may also see this denoted as ⊥.

Correlation

I have two variables, A and B.

A and B are **dependent**

*(i.e., there is **an** association)*

We can test this assumption with a simple correlation:

```
corr(A, B)
0.6946628
```

We can denote our findings as **A ⫫̸ B** or **A⊤B**.

You may also see this denoted as ⫫̸ or ⊤.

Correlation

I have two variables, A and B.

A and B are **dependent**

*(i.e., there is **an** association)*

This notation (and correlation measure) only says whether or not the two variables are statistically related.

*It says nothing about **how** they are related.*

Causation

A causes B

Ex. A = high blood pressure
B = experimental blood pressure-reducing drug

B causes A

Ex. A = high blood pressure
B = experimental anti-blood clotting drug

I have two variables, A and B **that are causally related.**

*While the data will only show a correlation, I can incorporate **expert knowledge** regarding the **data generating process** and **causal relationships** into **a knowledge graph,** commonly a **DAG*.***

**A Directed Acyclic Graph*

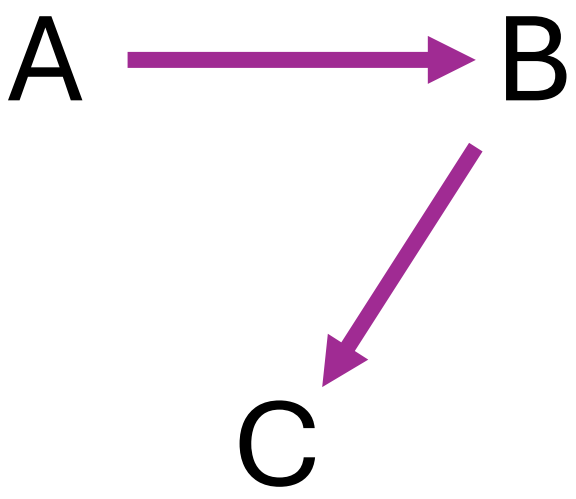
Three Variables

A B

C

If I have three variables, A and B and C, they may represent many potential causal relationships.

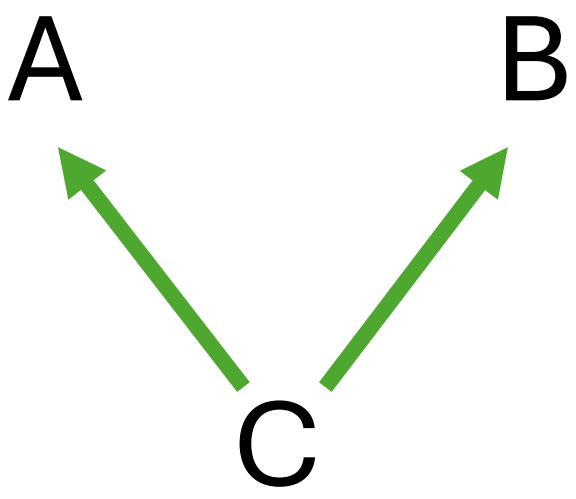
Three Variables



If I have three variables, A and B and C, they may represent many potential causal relationships.

B is a **mediator**. It is intermediate along the causal path from A to C and **describes *how* A acts on C**.

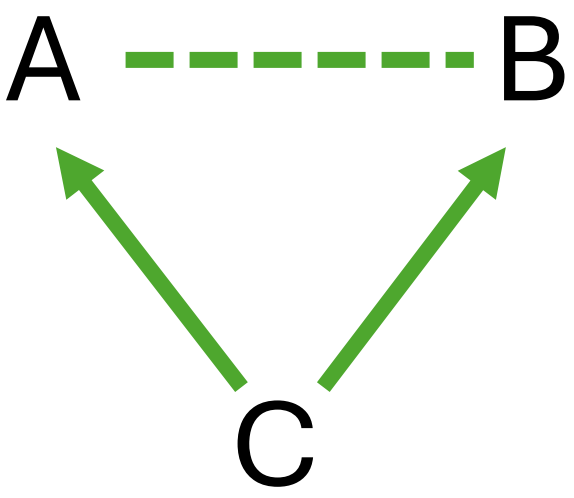
Three Variables



If I have three variables, A and B and C, they may represent many potential causal relationships.

C is a **confounder**. It causes both A and B.

Three Variables



If I have three variables, A and B and C, they may represent many potential causal relationships.

C is a **confounder**. It causes both A and B.

C ***explains*** an observed association between A and B

The Birth-Weight Paradox

Smoking

Birth Weight

Infant Mortality

- In the mid-1960s, Jacob Yerushalmy noted that a mother's smoking during pregnancy seemed to benefit the health of her newborn baby **if it was born underweight.**

The Birth-Weight Paradox

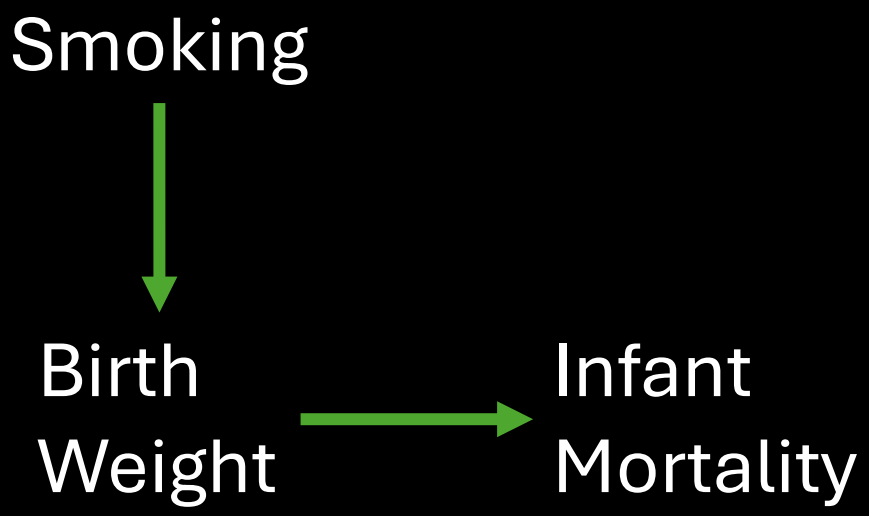
Smoking

Birth Weight

Infant Mortality

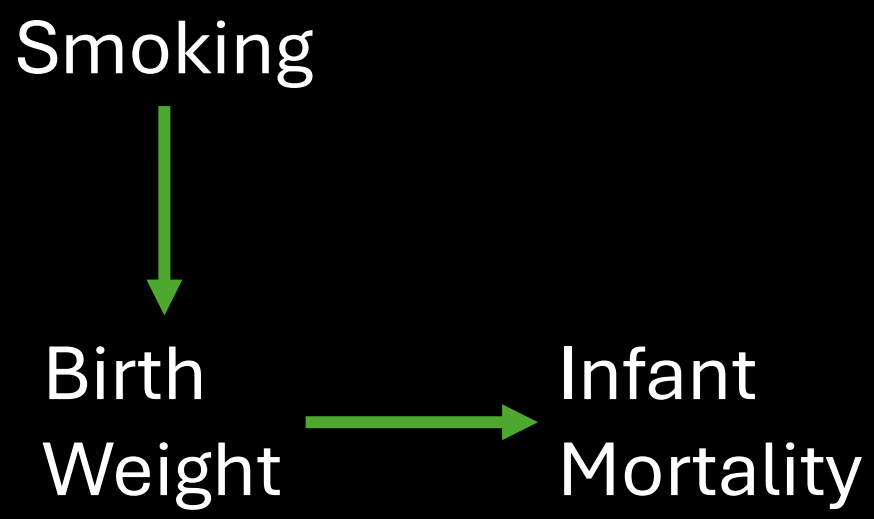
- In the mid-1960s, Jacob Yerushalmy noted that a mother's smoking during pregnancy seemed to benefit the health of her newborn baby **if it was born underweight.**
- **It was already observed that babies of mothers who smoked had lower birth weights.**

The Birth-Weight Paradox



- In the mid-1960s, Jacob Yerushalmy noted that a mother's smoking during pregnancy seemed to benefit the health of her newborn baby **if it was born underweight.**
- It was already observed that babies of mothers who smoked had lower birth weights.
- **A nationwide study also showed that low birth-weight babies had death rate more than 20 times higher than normal birth-weight babies.**

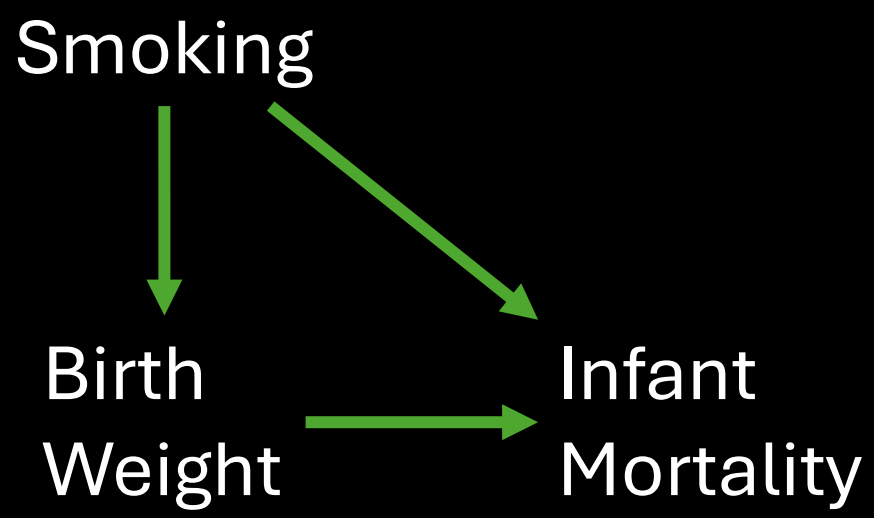
The Birth-Weight Paradox



- In the mid-1960s, Jacob Yerushalmy noted that a mother's smoking during pregnancy seemed to benefit the health of her newborn baby **if it was born underweight.**
- It was already observed that babies of mothers who smoked had lower birth weights.
- A nationwide study also showed that low birth-weight babies had death rate more than 20 times higher than normal birth-weight babies.

- Yerushalmy conducted a study on 15,000 children in the San Francisco Bay Area. He confirmed that babies of smokers had lower birth rate

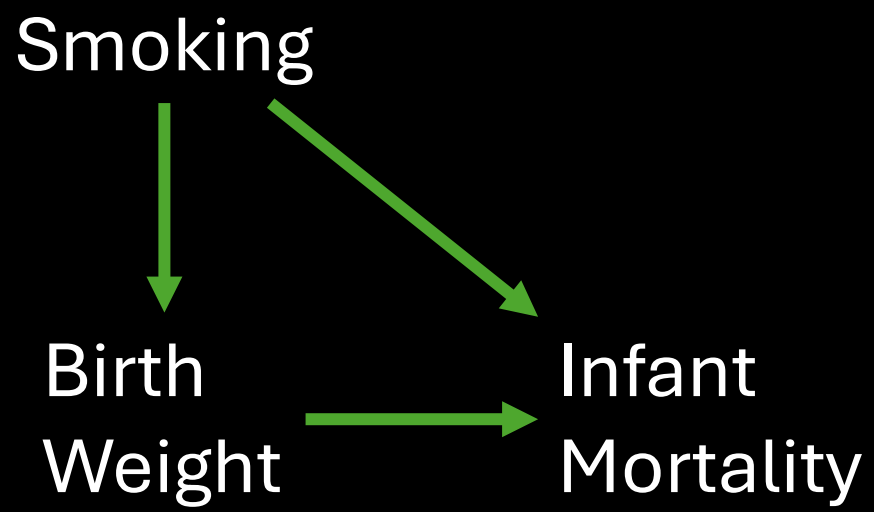
The Birth-Weight Paradox



- In the mid-1960s, Jacob Yerushalmy noted that a mother's smoking during pregnancy seemed to benefit the health of her newborn baby **if it was born underweight.**
- It was already observed that babies of mothers who smoked had lower birth weights.
- A nationwide study also showed that low birth-weight babies had death rate more than 20 times higher than normal birth-weight babies.

- Yerushalmy conducted a study on 15,000 children in the San Francisco Bay Area. He confirmed that babies of smokers had lower birth rate but those low birth-weight babies **had a better survival rate** than those born to nonsmokers.

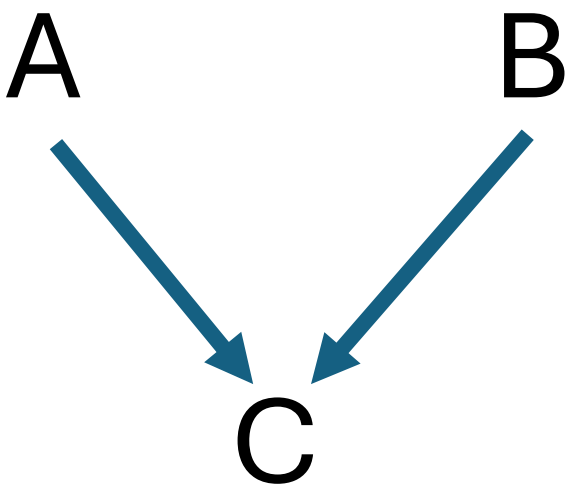
The Birth-Weight Paradox



- In the mid-1960s, Jacob Yerushalmy noted that a mother's smoking during pregnancy seemed to benefit the health of her newborn baby **if it was born underweight.**
- It was already observed that babies of mothers who smoked had lower birth weights.
- A nationwide study also showed that low birth-weight babies had death rate more than 20 times higher than normal birth-weight babies.

- Yerushalmy conducted a study on 15,000 children in the San Francisco Bay Area. He confirmed that babies of smokers had lower birth rate but those low birth-weight babies **had a better survival rate** than those born to nonsmokers.

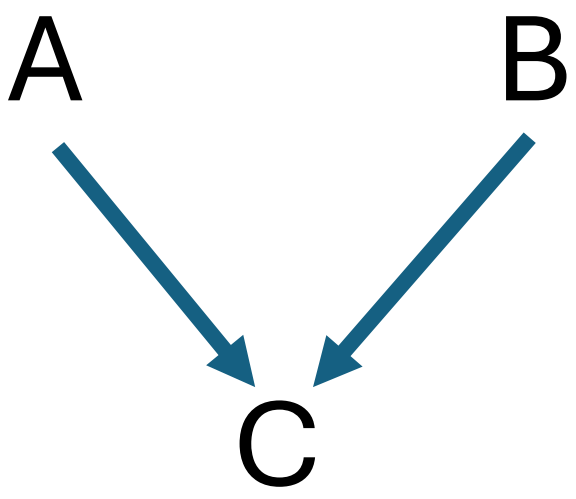
Three Variables



If I have three variables, A and B and C, they may represent many potential causal relationships.

C is a **collider**. It is caused by both A and B.

Three Variables

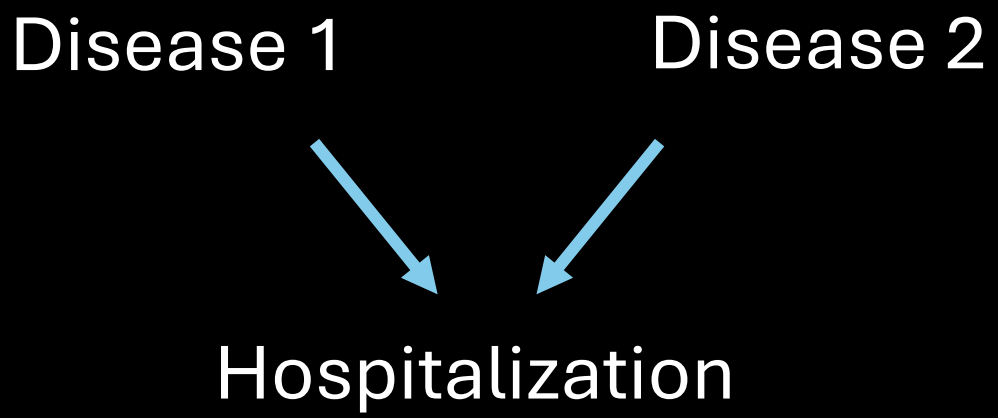


If I have three variables, A and B and C, they may represent many potential causal relationships.

C is a **collider**. It is caused by both A and B.

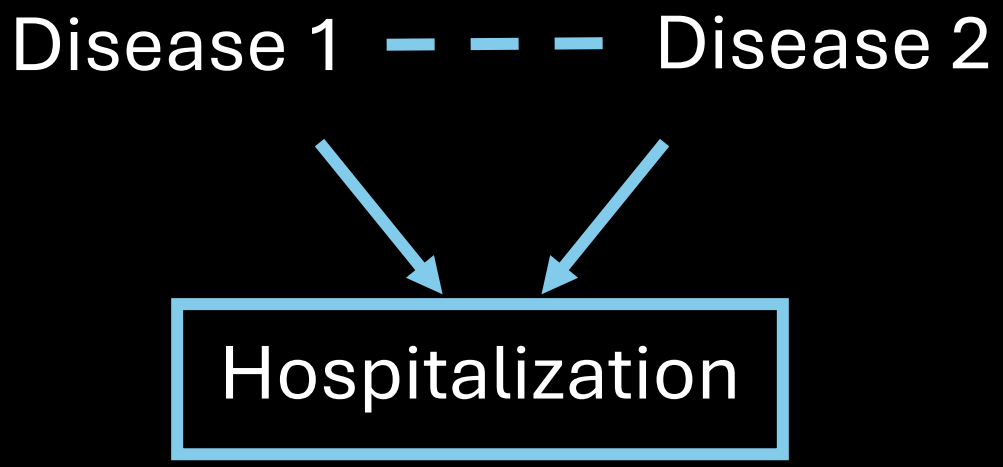
A common example of adjusting for a collider is **selection bias.**

Berkson’s Bias



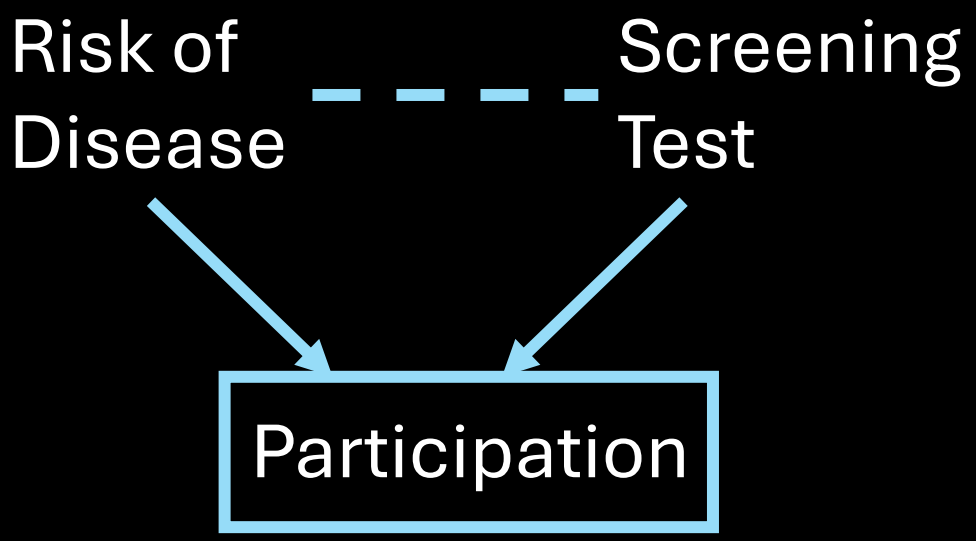
- In 1946, Joseph Berkson, a biostatistician at the Mayo Clinic, noted on oddity in observational studies conducted in hospital settings.
- Even if two diseases were unassociated in the general public, they appeared to be associated among hospitalized patients.

Berkson’s Bias



- In 1946, Joseph Berkson, a biostatistician at the Mayo Clinic, noted on oddity in observational studies conducted in hospital settings.
- Even if two diseases were unassociated in the general public, they appeared to be associated among hospitalized patients.
- By performing a study in hospitalized patients, we are **controlling for a collider**.

Selection Bias



- Selection bias can occur when we examine the effectiveness of a screening test by offering voluntary testing.
- The population being testing will often not represent the general population, perhaps because they have a family history for the disease, have signs and symptoms in line with the disease, or are more generally health-conscious.

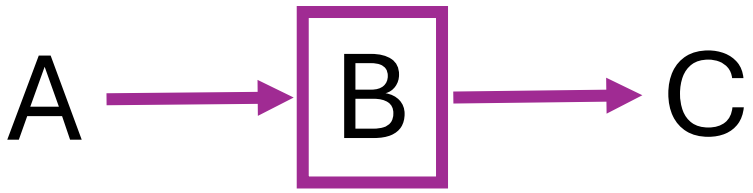
- It's unclear whether the study results will be **biased upward** (increased positive correlation) or **biased downward** (increased negative correlation)

Three Variables

Mediation

- We expect an association between A and B
- We expect an association between B and C

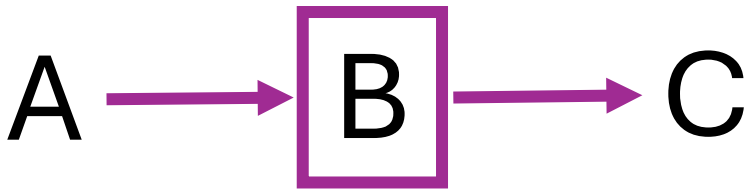
Three Variables

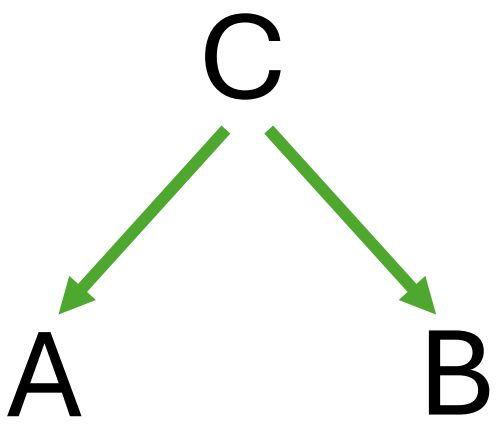


Mediation

- We expect an association between A and B
- We expect an association between B and C
- If we assume B is binary or categorical, we can calculate the **causal effect of A on C across levels of B.**

Three Variables





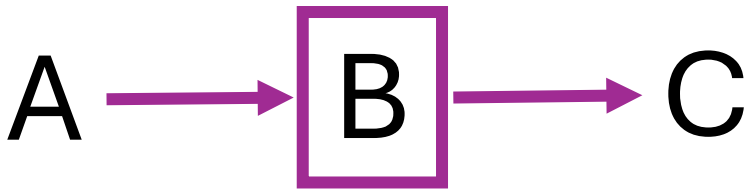
Mediation

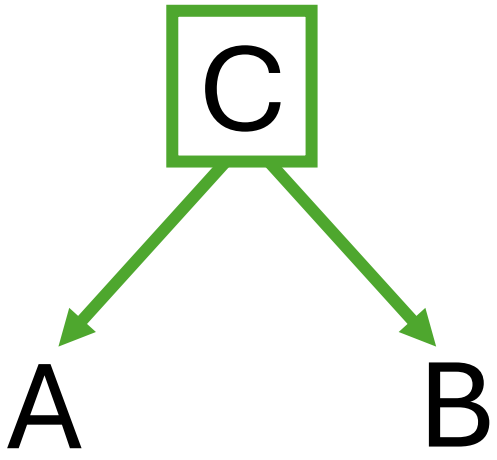
- **B is a mediator**
- We expect an association between A and B
- We expect an association between B and C
- If we assume B is binary or categorical, we can calculate the **causal effect of A on C across levels of B.**

Confounder

- **C is a confounder**. It causes A and separately causes B
- We expect an association between A and C
- We expect an association between B and C

Three Variables





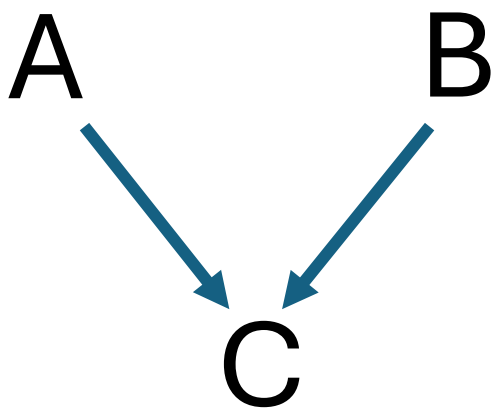
Mediation

- We expect an association between A and B
- We expect an association between B and C
- If we assume B is binary or categorical, we can calculate the **causal effect of A on C across levels of B.**

Confounder

- **C is a confounder**. It causes A and separately causes B
- We expect an association between A and C
- We expect an association between B and C
- If we assume C is binary or categorical, we can examine **whether A and B are associated "after controlling for" (i.e., across levels of) C**.
- If there is no association after controlling for C, we can write A ⫫ B | C (where | here is read as "conditional on" or "given")

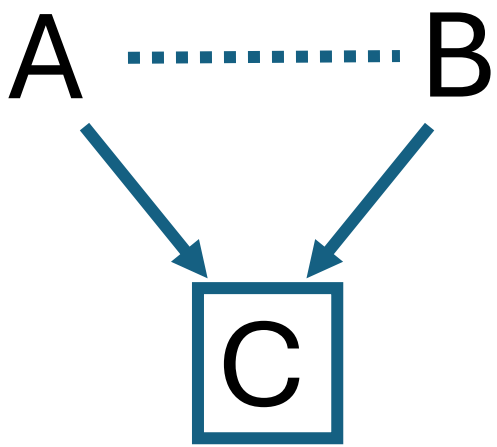
Three Variables



Collider

- If we ignore C, A and B are independent: $A \perp\!\!\!\perp B$
  - *We call this "marginalizing over" C*

Three Variables

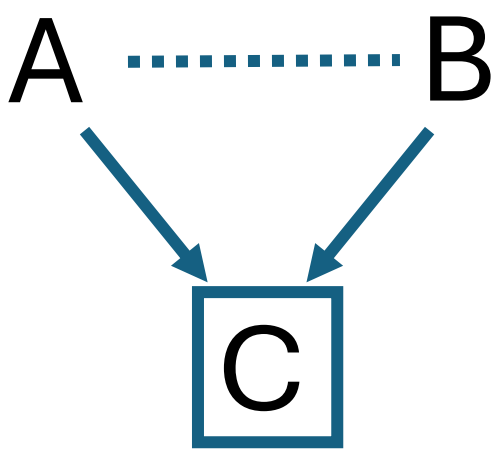


Collider

- If we ignore C, A and B are independent: $A \perp\!\!\!\perp B$
  - *We call this "marginalizing over" C*

- If we were to control for C, we would introduce an association between A and B: $A \not\perp\!\!\!\perp B \mid C$

Three Variables



Collider

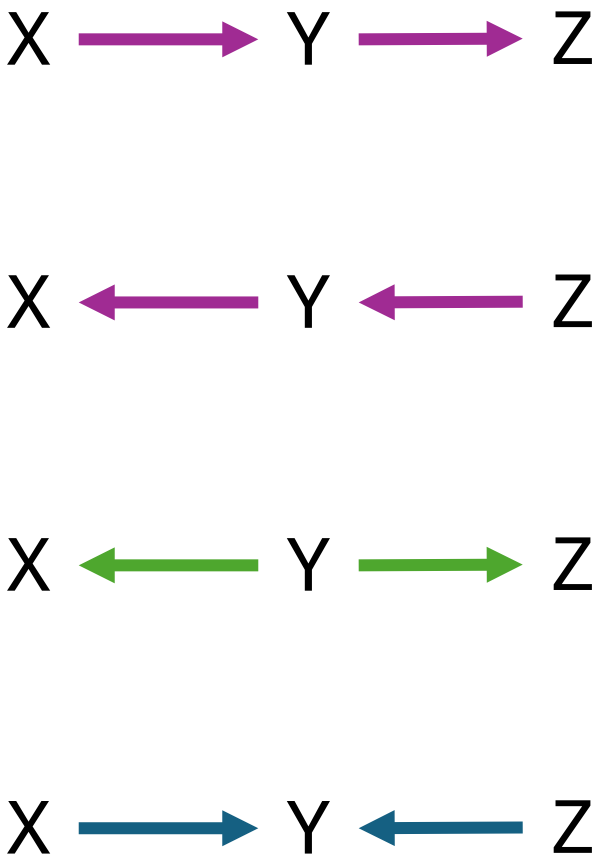
- If we ignore C, A and B are independent: $A \perp\!\!\!\perp B$
  - *We call this "marginalizing over" C*

- If we were to control for C, we would introduce an association between A and B: $A \not\perp\!\!\!\perp B \mid C$

A common example of this is **selection bias**:

A = *potential causal variable*
B = *disease status*
C = *selection criteria into study*

Three Variables



How can we tell these apart with the data?

Mediation Associations

$X \not\perp\!\!\!\perp Y$ $Y \not\perp\!\!\!\perp Z$

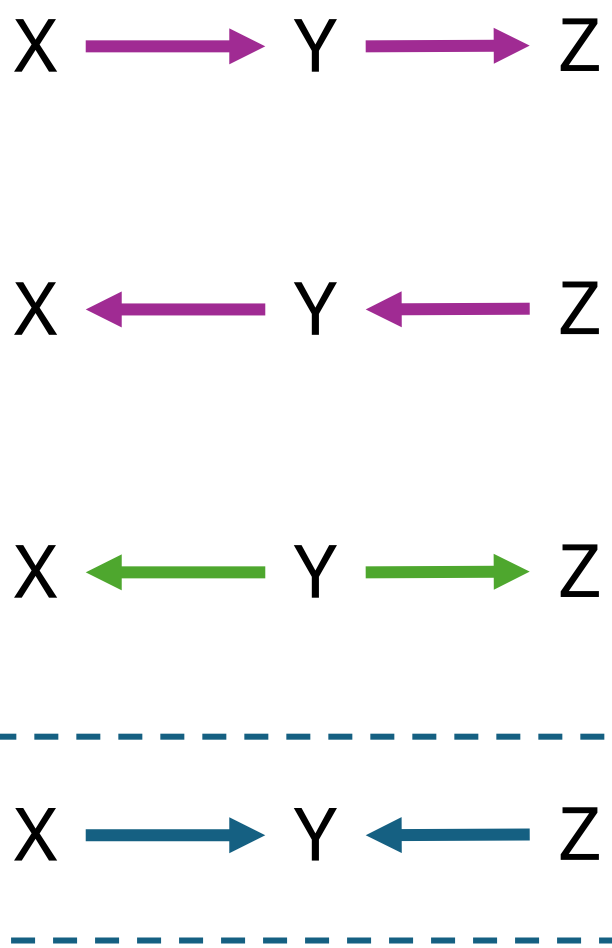
Confounder Associations

$X \not\perp\!\!\!\perp Y$ $Y \not\perp\!\!\!\perp Z$

Collider Associations

$X \not\perp\!\!\!\perp Y$ $Y \not\perp\!\!\!\perp Z$

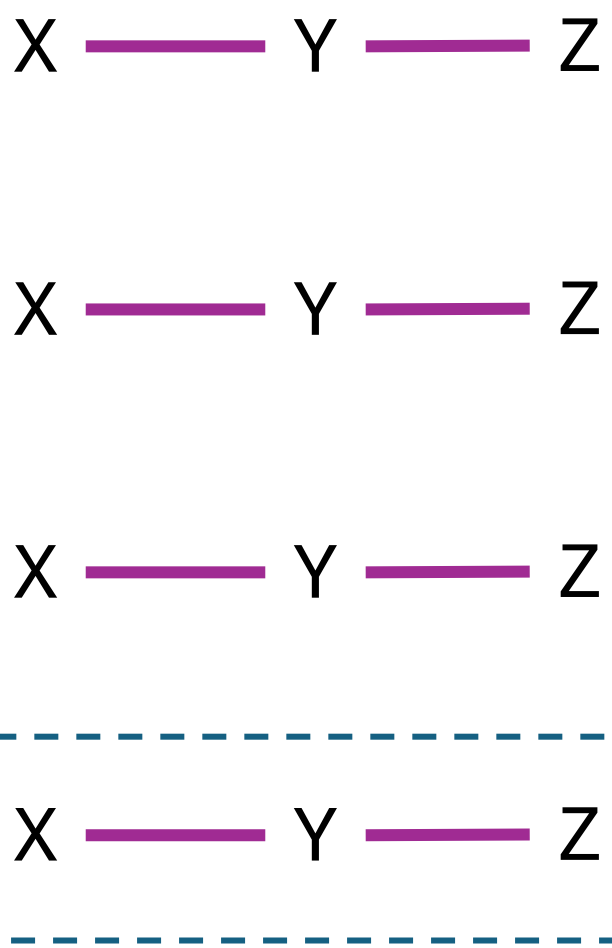
Three Variables



How can we tell these apart with the data?

1. **Define an *immorality* in a DAG as a configuration of 3 nodes - A, B, and C – such that C is a child of both A and B, and A and B are not directly connected.**

Three Variables

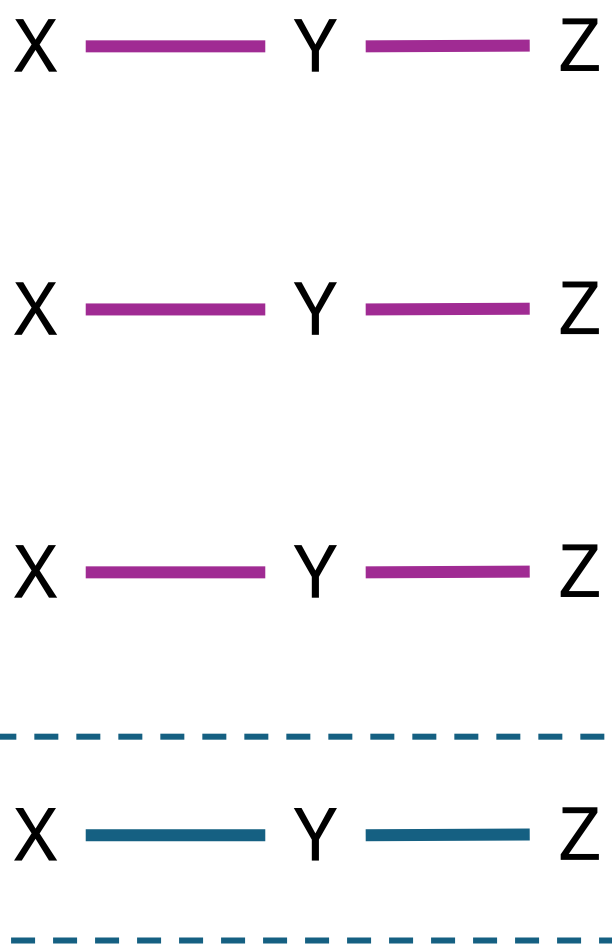


How can we tell these apart with the data?

1. **Define an *immorality* in a DAG as a configuration of 3 nodes - A, B, and C – such that C is a child of both A and B, and A and B are not directly connected.**

2. **Replace all directed edges with undirected edges.**

Three Variables



How can we tell these apart with the data?

1. **Define an *immorality* in a DAG as a configuration of 3 nodes - A, B, and C – such that C is a child of both A and B, and A and B are not directly connected.**

2. **Replace all directed edges with undirected edges.**

The first three graphs are *Markov Equivalent*, meaning you can't tell them apart by the data relationships.

Simpson's Paradox

Dr. Simpson reads an article about a promising new drug that seems to reduce the risk of heart attacks.

	Control Group (No Drug)		Treatment Group (Took Drug)	
	Heart Attack	No Heart Attack	Heart Attack	No Heart Attack
TOTAL	13	47	11	49

$$IR_C = \frac{13}{13+47} = 21.7\% \qquad IR_T = \frac{11}{11+49} = 18.3\%$$

$$IRR = \frac{21.7\%}{18.3\%} = \boxed{1.2}$$

Simpson's Paradox

Drug $\xrightarrow{(-)}$ Heart Attack

Females

$$IR_{CF} = \frac{1}{1+19} = 5\% \qquad IR_{TF} = \frac{3}{3+37} = 7.5\%$$

$$IRR_F = \frac{5\%}{7.5\%} = \boxed{0.66}$$

	Control Group (No Drug)		Treatment Group (Took Drug)	
	Heart Attack	No Heart Attack	Heart Attack	No Heart Attack
FEMALE	1	19	3	37
TOTAL	13	47	11	49

Simpson's Paradox

Drug $\xrightarrow{(-)}$ Heart Attack

Females

$$IRR_F = \frac{5\%}{7.5\%} = \boxed{0.66}$$

Drug $\xrightarrow{(-)}$ Hearth Attack

Males

$$IR_{CM} = \frac{12}{12+28} = 30\% \qquad IR_{TM} = \frac{8}{8+12} = 40\%$$

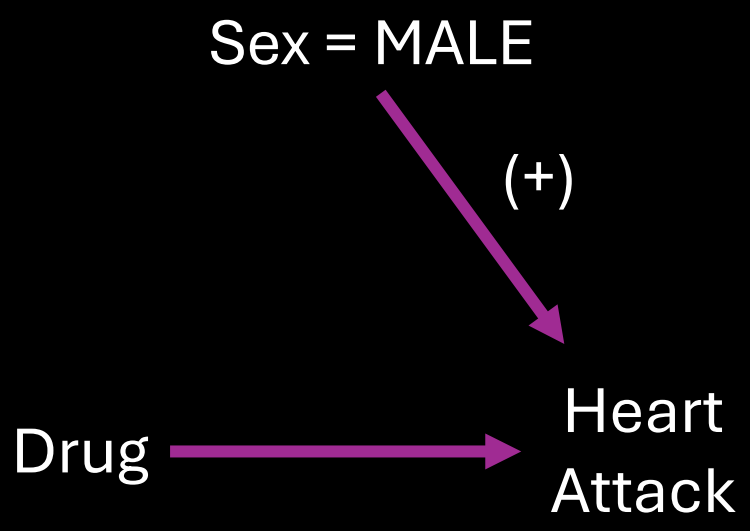
$$IRR_M = \frac{30\%}{40\%} = \boxed{0.75}$$

	Control Group (No Drug)		Treatment Group (Took Drug)	
	Heart Attack	No Heart Attack	Heart Attack	No Heart Attack
FEMALE	1	19	3	37
MALE	12	28	8	12
TOTAL	13	47	11	49

Simpson's Paradox

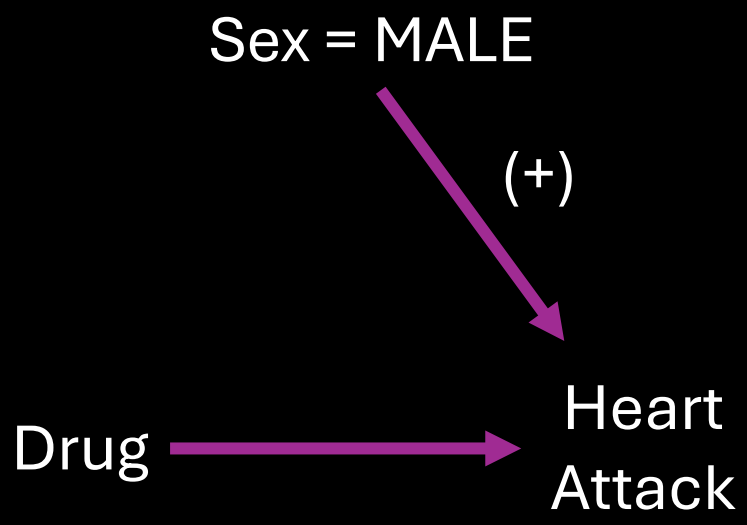
	Control Group (No Drug)		Treatment Group (Took Drug)	
	Heart Attack	No Heart Attack	Heart Attack	No Heart Attack
FEMALE	1	19	3	37
MALE	12	28	8	12
TOTAL	13	47	11	49

Simpson's Paradox



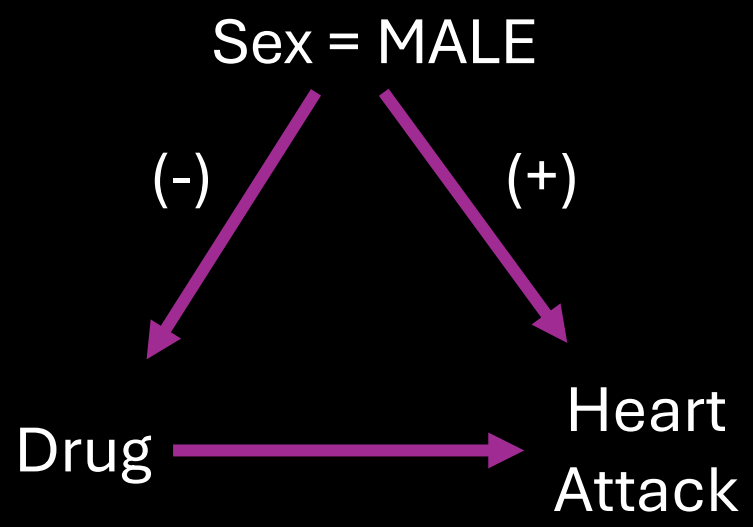
	Control Group (No Drug)		Treatment Group (Took Drug)	
	Heart Attack	No Heart Attack	Heart Attack	No Heart Attack
FEMALE	1	19	3	37
MALE	12	28	8	12
TOTAL	13	47	11	49

Simpson's Paradox



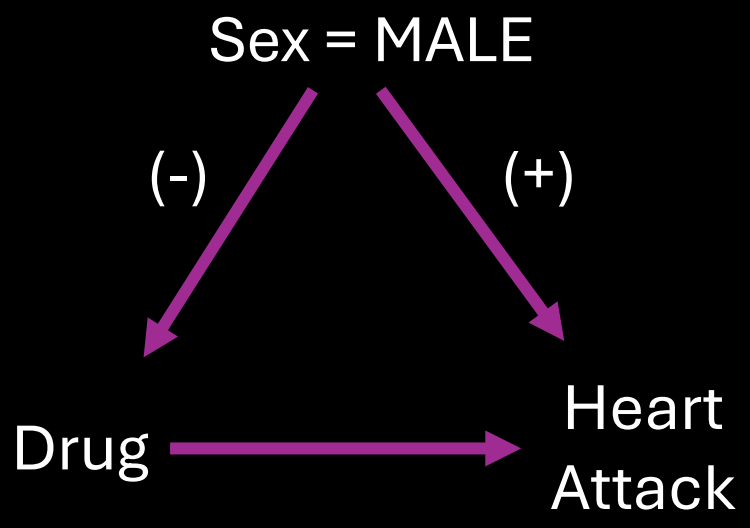
	Control Group (No Drug)		Treatment Group (Took Drug)	
	Heart Attack	No Heart Attack	Heart Attack	No Heart Attack
FEMALE	1	19	3	37
MALE	12	28	8	12
TOTAL	13	47	11	49

Simpson's Paradox



	Control Group (No Drug)		Treatment Group (Took Drug)	
	Heart Attack	No Heart Attack	Heart Attack	No Heart Attack
FEMALE	1	19	3	37
MALE	12	28	8	12
TOTAL	13	47	11	49

Simpson's Paradox



	Control Group (No Drug)		Treatment Group (Took Drug)	
	Heart Attack	No Heart Attack	Heart Attack	No Heart Attack
FEMALE	1	19	3	37
MALE	12	28	8	12
TOTAL	13	47	11	49

$$IR_C = 0.5\left(\frac{1}{20}\right) + 0.5\left(\frac{12}{40}\right) = 0.175$$

female *male*

$$IR_T = 0.5\left(\frac{3}{40}\right) + 0.5\left(\frac{8}{20}\right) = 0.2375$$

$$IRR = \frac{0.175}{0.2375}$$

$$= \boxed{0.737}$$

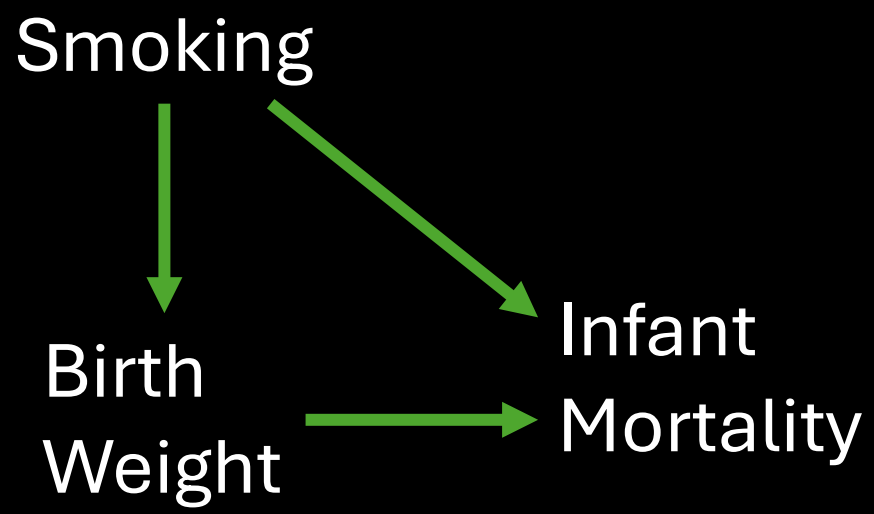
National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Demo

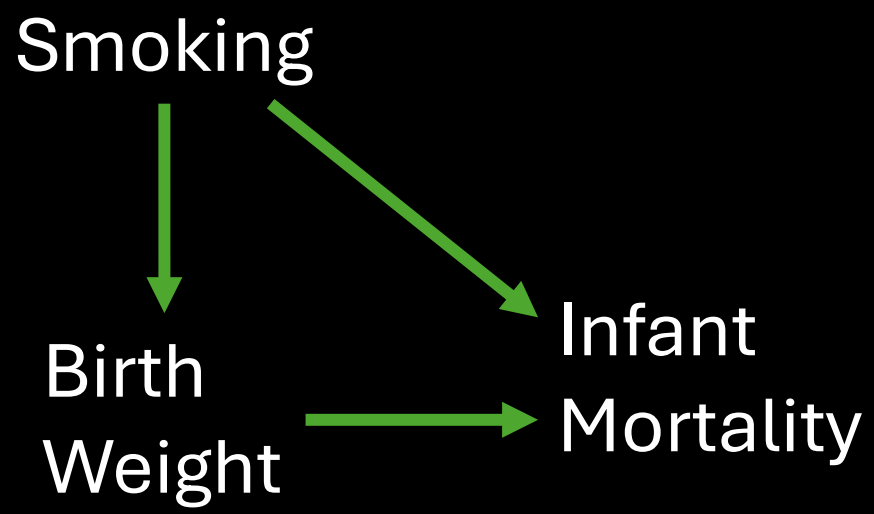
- Demo of causal inference example using ***NIDDK-CR Analytics Workbench*** – a secure, cloud-based research environment that enables:
  - ✓ Real-time collaboration
  - ✓ On-demand access to suite of tools for data science, analysis, and ML
  - ✓ Scalable computing resources
  - ✓ Integrated governance and access control
  - ✓ Efficient data processing and engineering pipelines

The Birth-Weight Paradox



- A mother's smoking during pregnancy seemed to benefit the health of her newborn baby **if it was born underweight.**

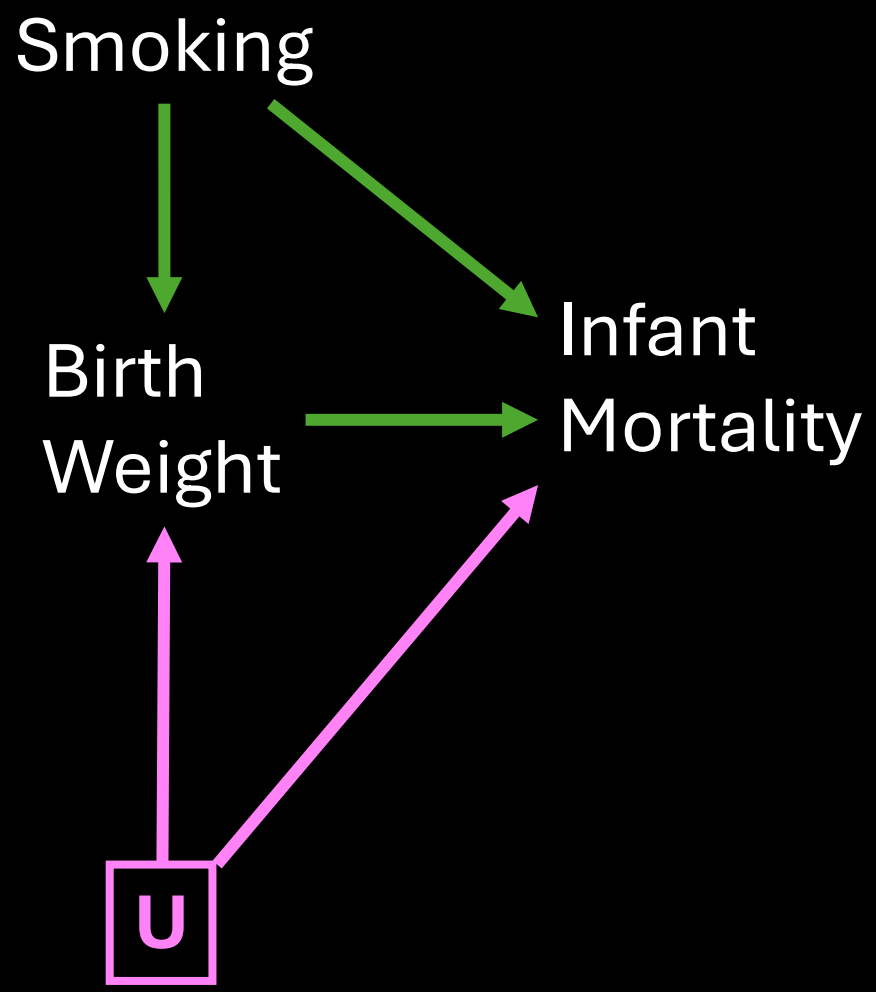
The Birth-Weight Paradox



- A mother's smoking during pregnancy seemed to benefit the health of her newborn baby **if it was born underweight.**

- **Selection bias is a recognized problem when studying newborns.** You may recall we discuss disease prevalence in newborns rather than incidence because we cannot know when disease occurs in utero, we can only observe at birth.

The Birth-Weight Paradox



- A mother's smoking during pregnancy seemed to benefit the health of her newborn baby **if it was born underweight.**

- **Selection bias is a recognized problem when studying newborns.** You may recall we discuss disease prevalence in newborns rather than incidence because we cannot know when disease occurs in utero, we can only observe at birth.

- A very plausible explanation for the birthweight paradox is that, by examining newborn infants, **we have inadvertently controlled for an unknown causal factor that effects both birthweight and spontaneous abortion.**

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Thank you!

Contact: Dr. Courtney D. Shelley, PhD - shelley_courtney@bah.com

Upcoming Webinar

Neural Networks for Detecting Subtle Epileptogenic Lesions and Supporting Clinical Decision-Making

- Date: **March 26, from 2-3pm ET**
- Experts: Dr. Mathilde Ripart, Kings College London
- *Scan the QR code register*