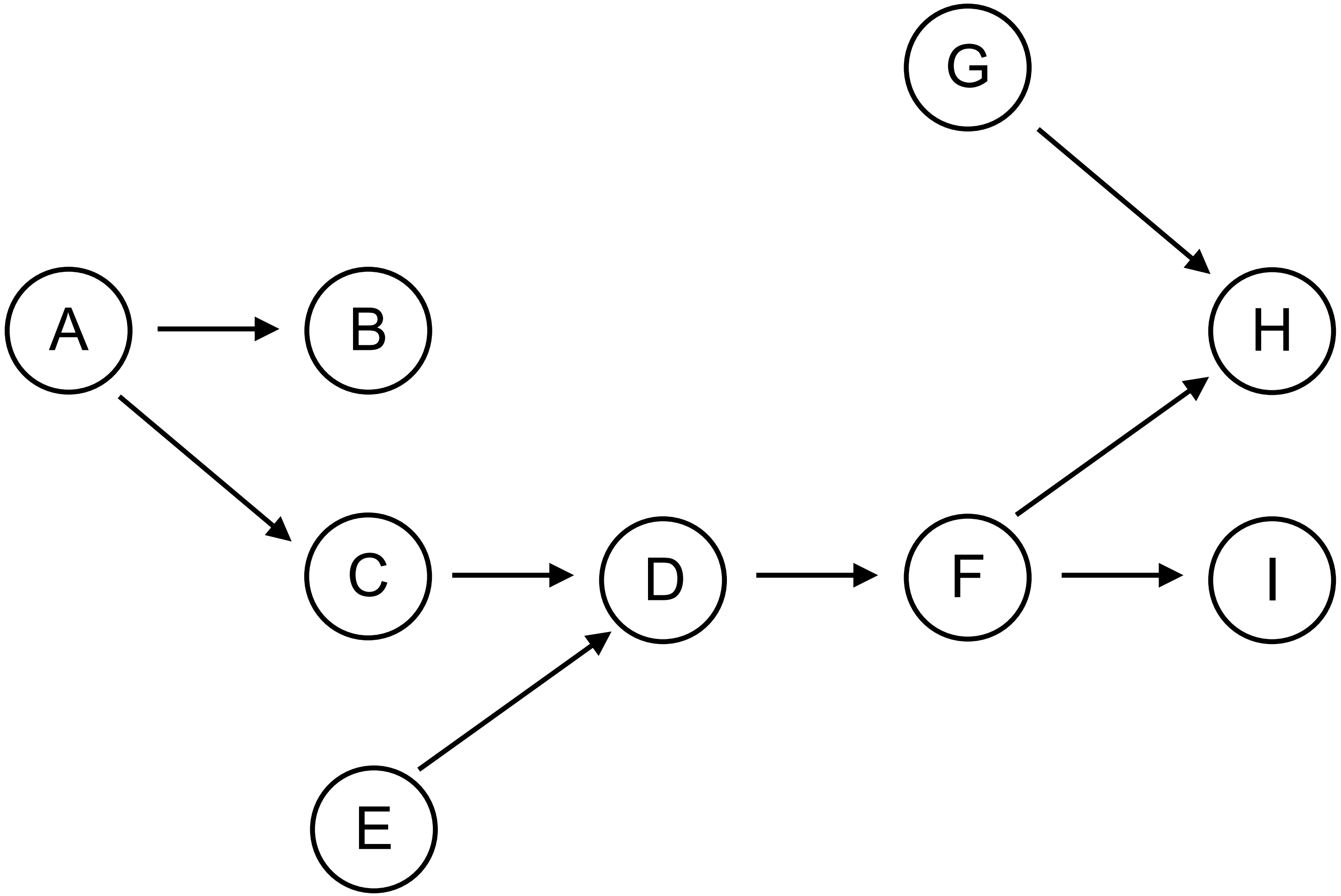


Introduction to causal inference

Roni Kobrosly, PhD

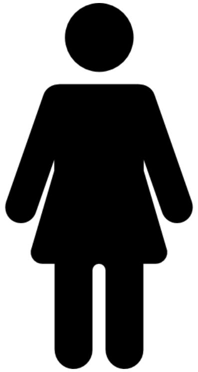
By the end of this tutorial, you should be able to

- Understand the pitfalls of observational data analysis
- Know the various types of causal relationships to look out for
- Describe the hierarchy of statistical analyses, causal inference, and experiments
- Start conducting preliminary causal analyses on your own data
- Confidently explore the topic on your own (now that you have a solid foundational understanding of causal thinking)

Does Vitamin D supplementation prevent severe covid symptoms?

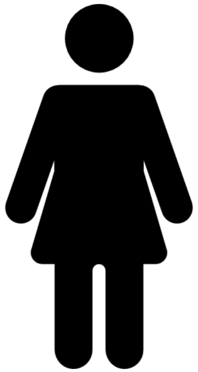
The alternative universe example

Our
Observed
reality



The alternative universe example

Our Observed reality

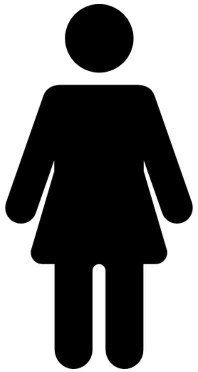


Early 2020

Starts Taking Vitamin D

The alternative universe example

Our Observed reality

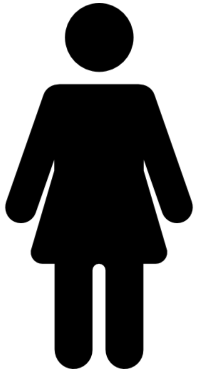


Early 2020

Starts Taking Vitamin D → Lives life

The alternative universe example

Our
Observed
reality



Early 2020

Starts Taking
Vitamin D

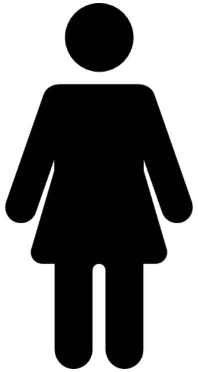
Lives
life

Mid 2020

Gets covid with
Mild symptoms

The alternative universe example

Our Observed reality

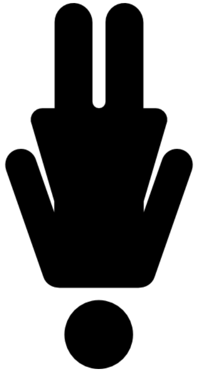


Early 2020

Starts Taking Vitamin D → Lives life → Mid 2020

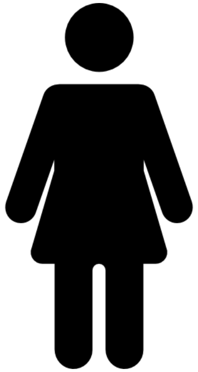
Gets covid with Mild symptoms

An alternative reality



The alternative universe example

Our Observed reality

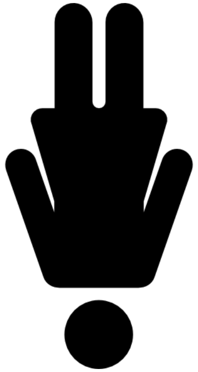


Early 2020

Starts Taking Vitamin D → Lives life → Mid 2020

Gets covid with Mild symptoms

An alternative reality

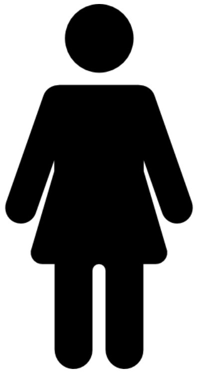


Early 2020

Never takes vitamin D

The alternative universe example

Our Observed reality



Early 2020

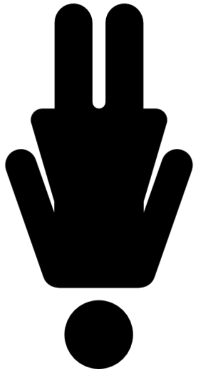
Starts Taking Vitamin D

Lives life

Mid 2020

Gets covid with Mild symptoms

An alternative reality



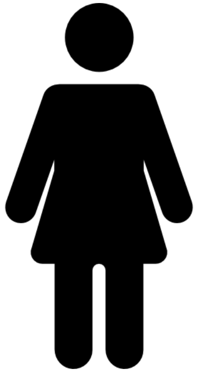
Early 2020

Never takes vitamin D

Lives life

The alternative universe example

Our Observed reality

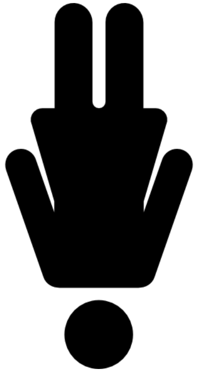


Early 2020

Starts Taking Vitamin D → Lives life → Mid 2020

Gets covid with Mild symptoms

An alternative reality



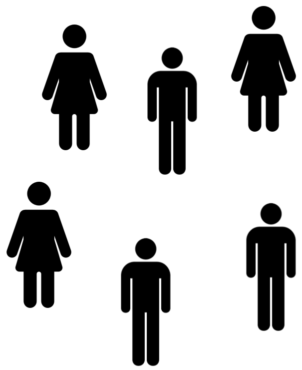
Early 2020

Never takes vitamin D → Lives life → Mid 2020

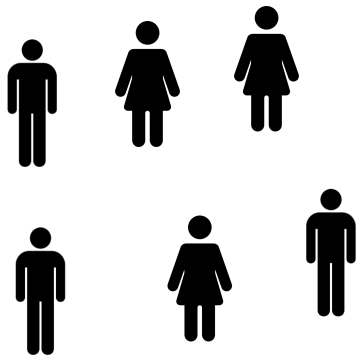
???

Experiments (AKA A/B Tests, AKA Randomized Controlled Trials)

Treatment group

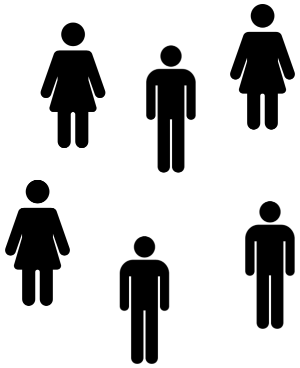


Control group



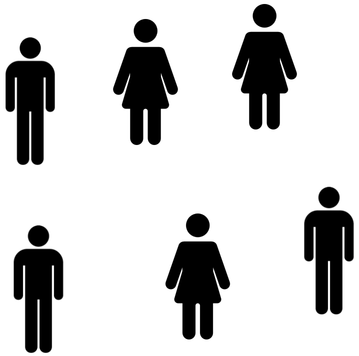
Experiments (AKA A/B Tests, AKA Randomized Controlled Trials)

Treatment group



given blank vitamin D pills

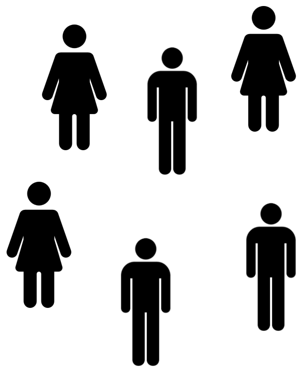
Control group



given blank sugar pills

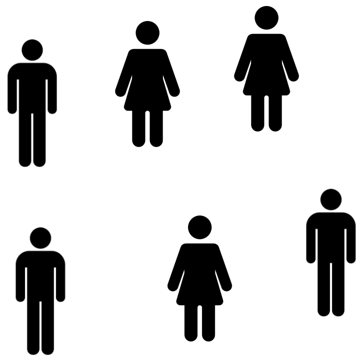
Experiments (AKA A/B Tests, AKA Randomized Controlled Trials)

Treatment group



→ given blank vitamin D pills → See how many people get severe covid symptoms over next 2 months

Control group



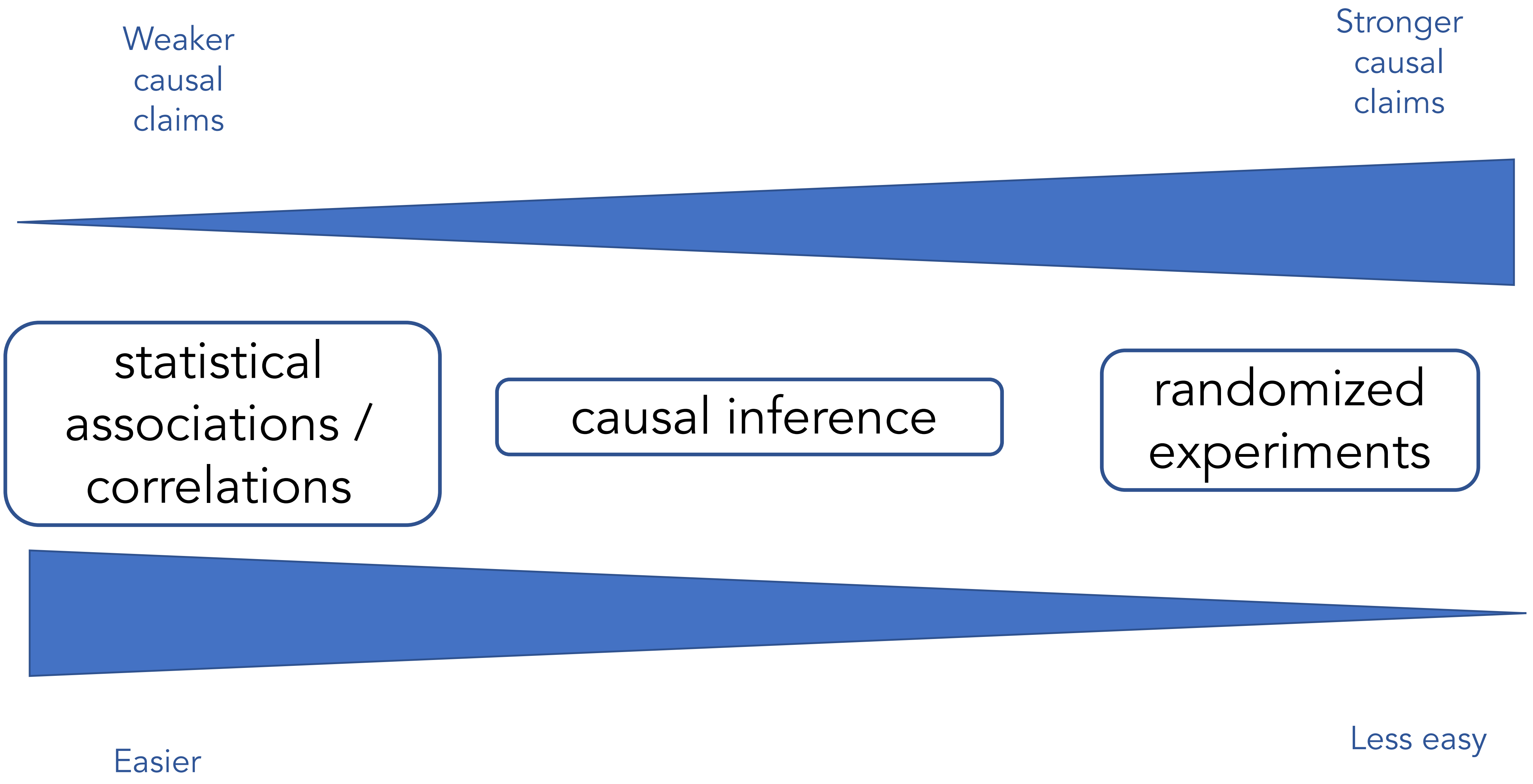
→ given blank sugar pills → See how many people get severe covid symptoms over next 2 months

Experiments won’t always save us

NOT ETHICAL: randomly assign some people to be exposed to lead paint while others are not, then see which group is more likely to develop neurological disorders.

NOT FEASIBLE: modify household incomes in neighborhoods, to see if reducing a neighborhood’s income inequality reduces the local crime rate.

A simple hierarchy…



Causal Inference vs Typical ML Project Questions

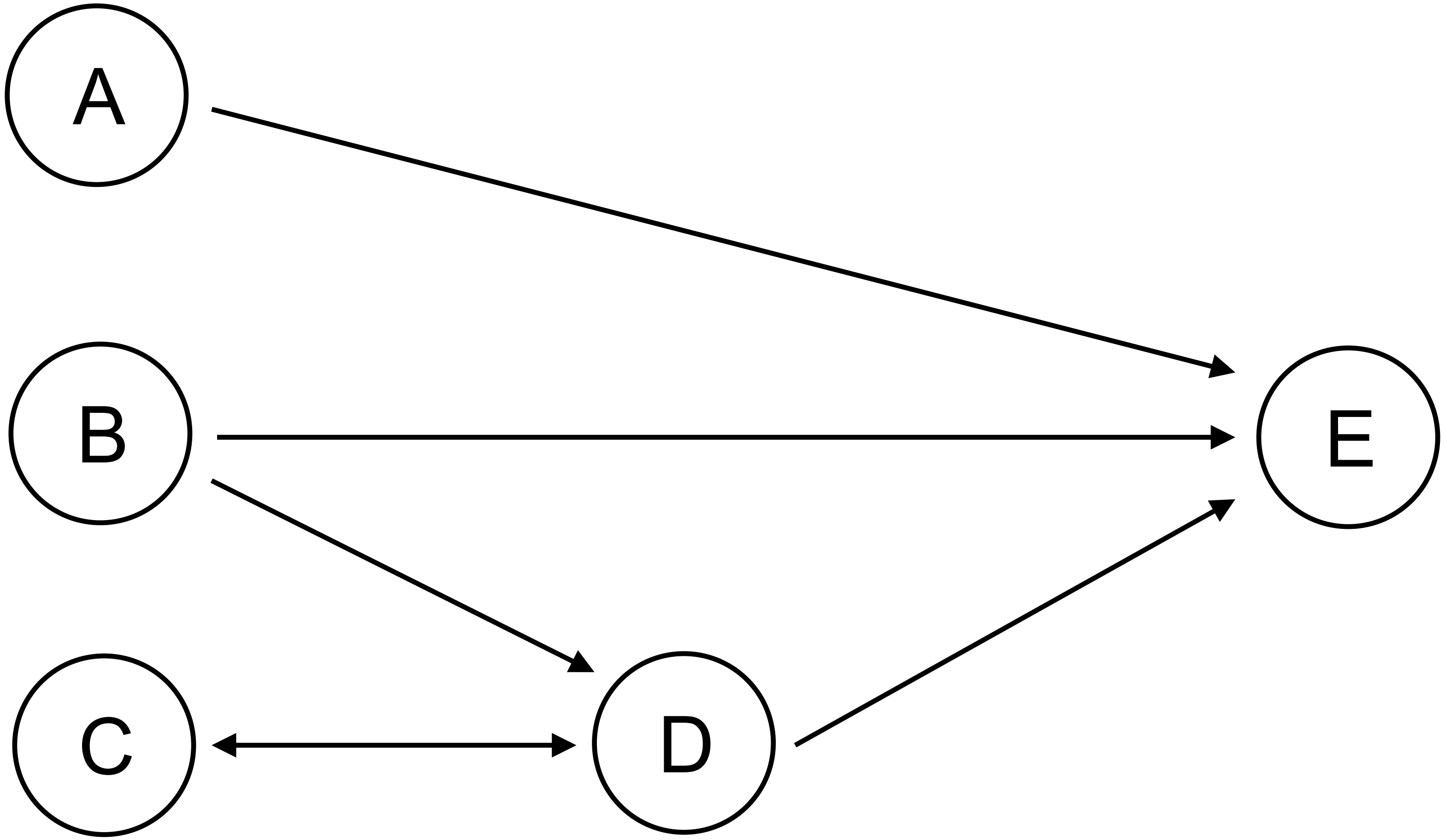
Causal Inference:

- How does improving neighborhood income inequality reduce neighborhood crime rate?
- How does increasing or decreasing the price of a product would impact demand?
- What would be the impact on the number of people with diabetes if we enacted a policy to reduce the average amount of sugar consumed per day by X grams.

Typical ML:

- Can I cluster neighborhoods by their characteristics and tell a story about these different segments and how it relates to crime rates?
- Can I predict whether someone will convert from a lead to a customer?
- How well can I predict whether a patient will be diagnosed with diabetes later in life?

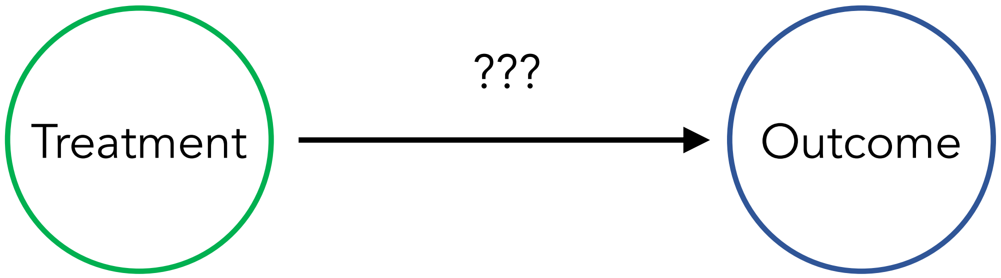
A causal graph



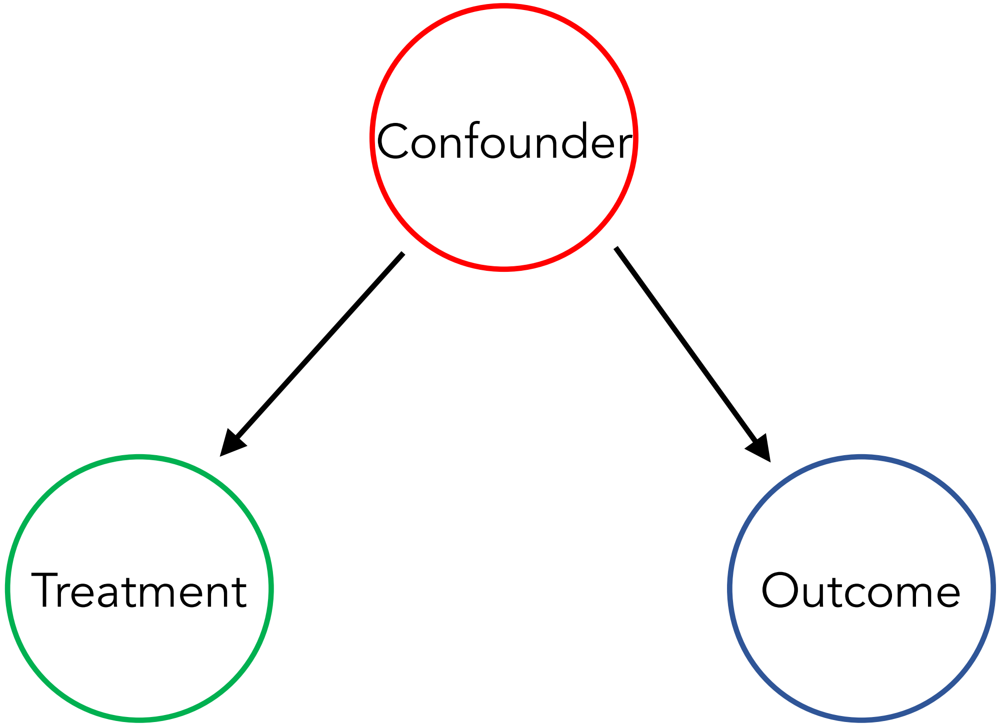
Four types of causal relationships…

1) Confounders

Confounders

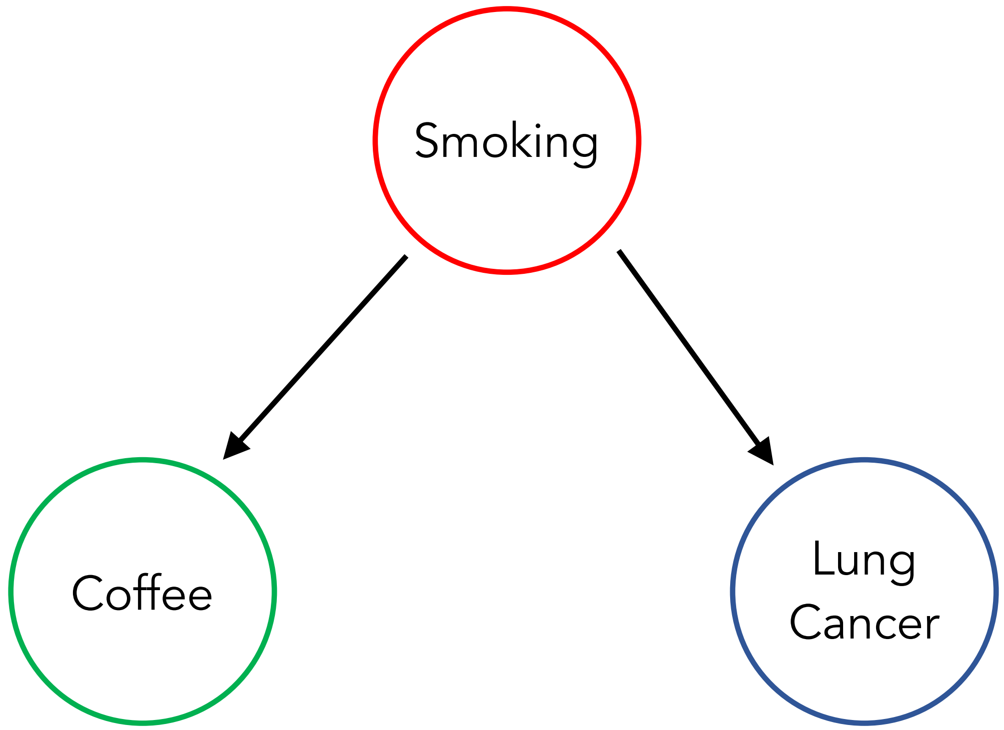


Confounders



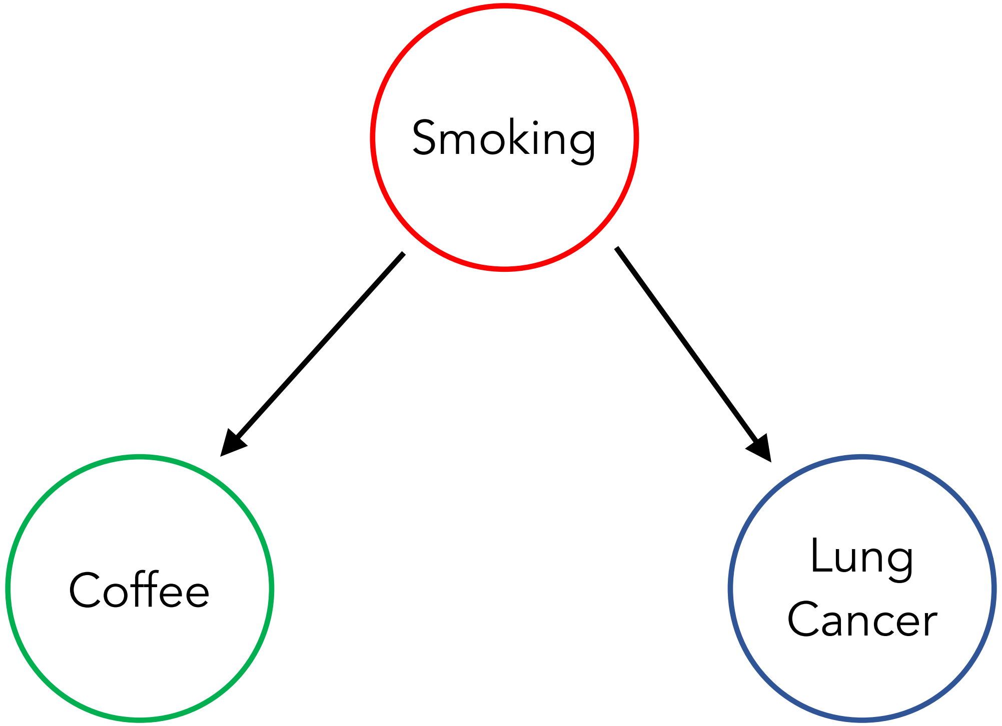
Confounders

- Always want to control for / condition on confounders in inferential modeling
- Confounding changes the effect size and possibly statistical significance of your association of interest
- Confounders can also flip the direction of your association of interest
- A model will ideally control for confounding, but leftover confounding in a model is named "residual confounding"

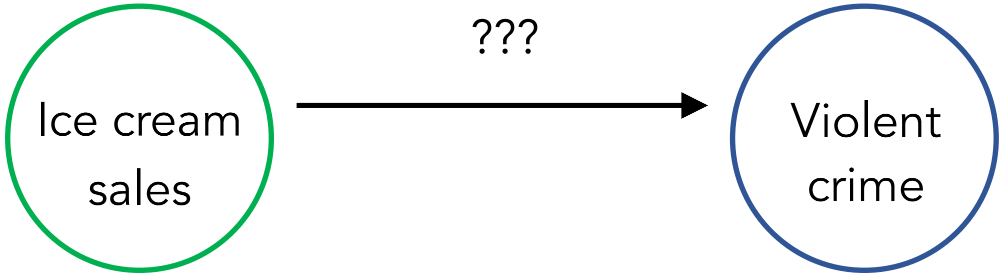


Confounders

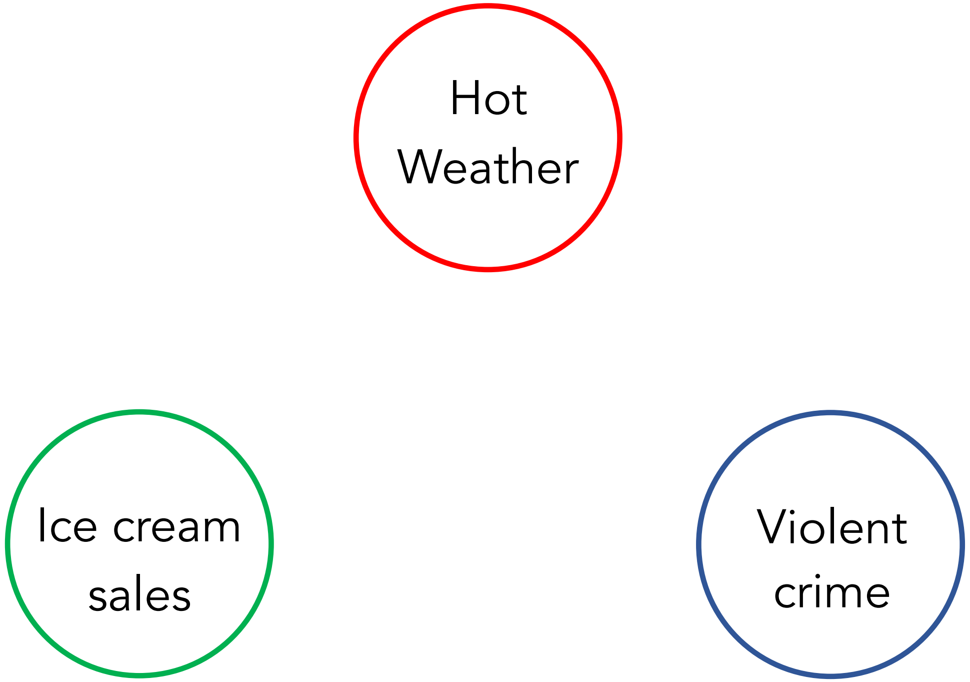
- Positive confounding: confounder introduces a bias that pushes association of interest away from the "null"
- Negative confounding: confounder biases association towards the "null"



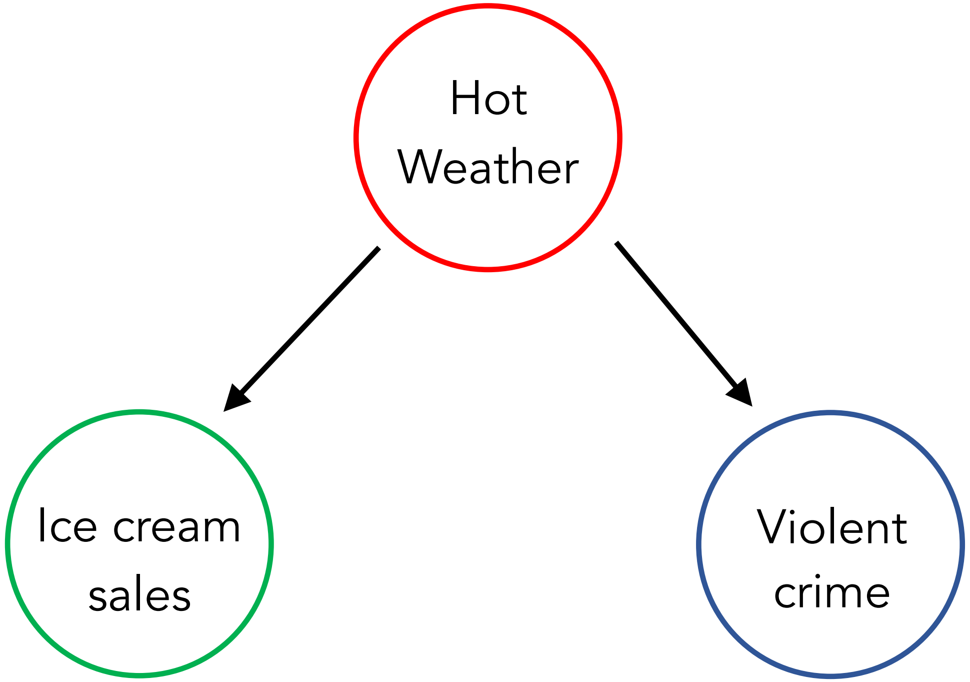
Violent crime in your city!



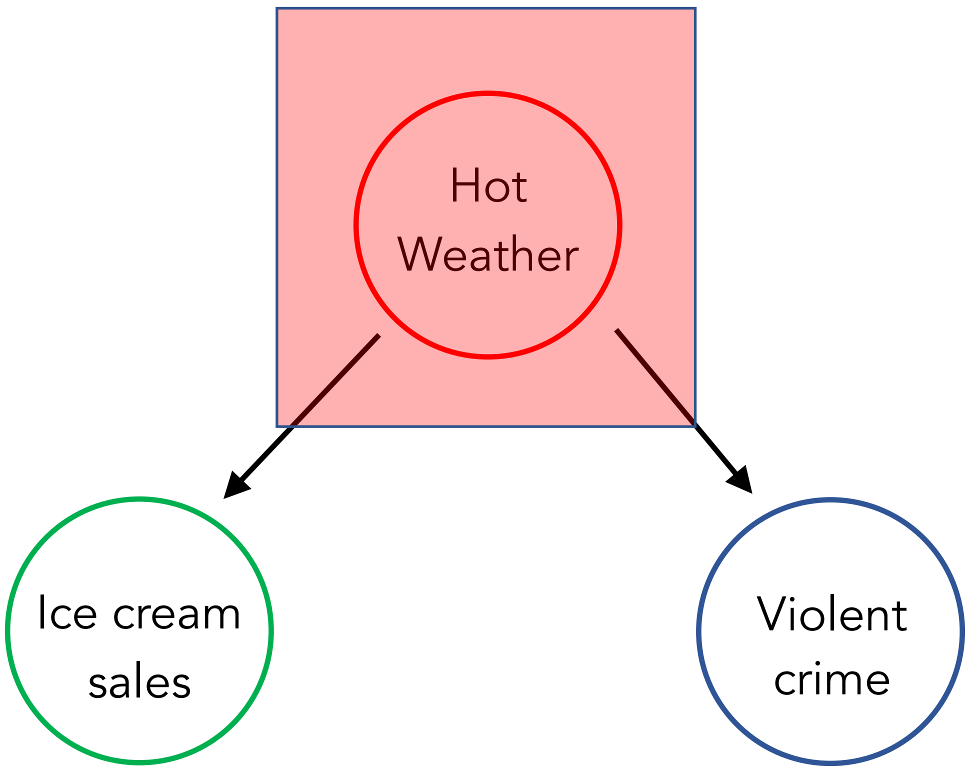
Summer weather induces a false association between ice cream sales and violent crime



Summer weather induces a false association between ice cream sales and violent crime

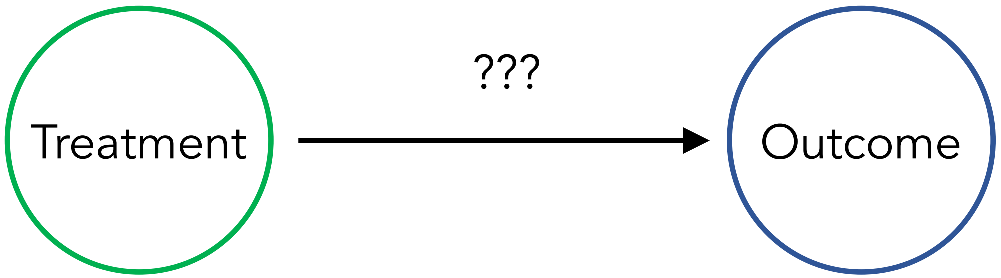


If you control for the season, any ice cream-violent crime association in your dataset will disappear

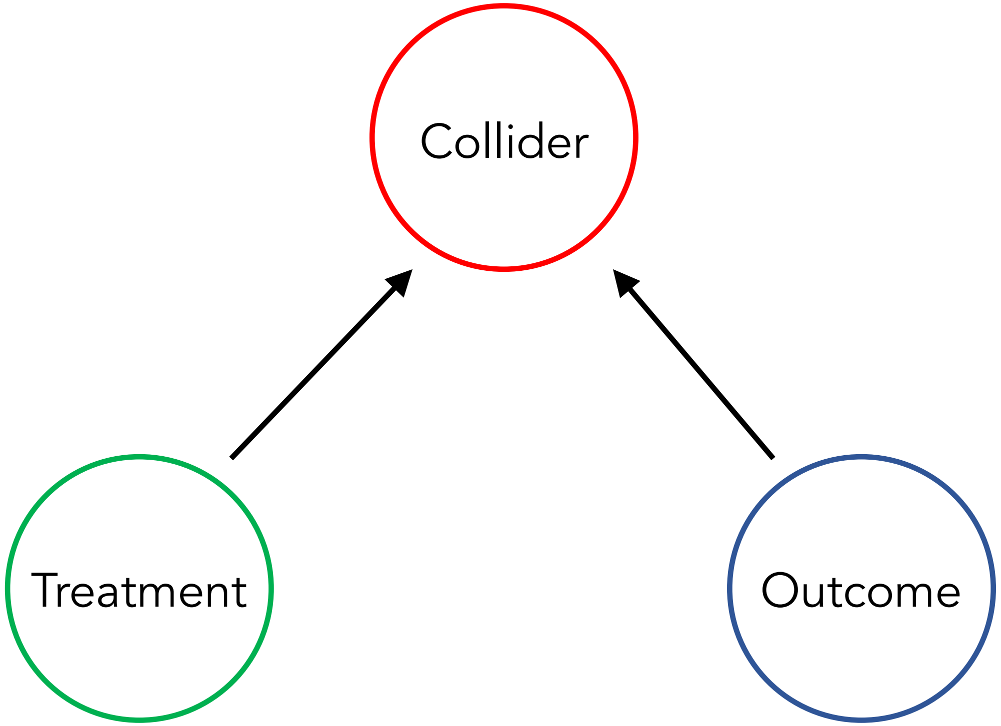


2) Colliders

Colliders

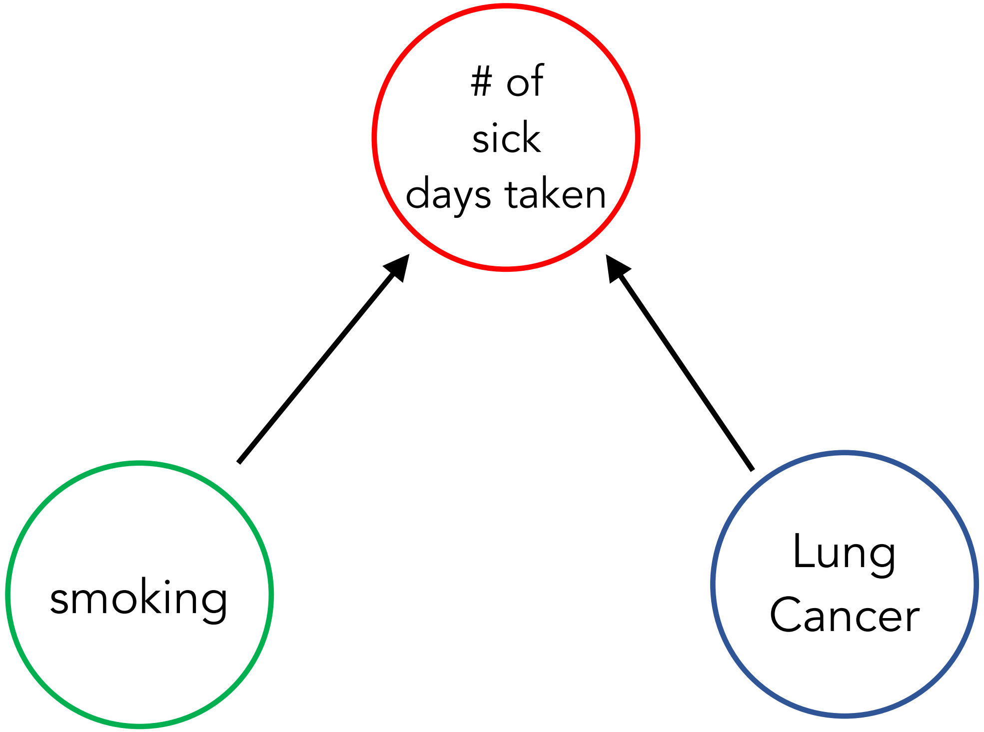


Colliders



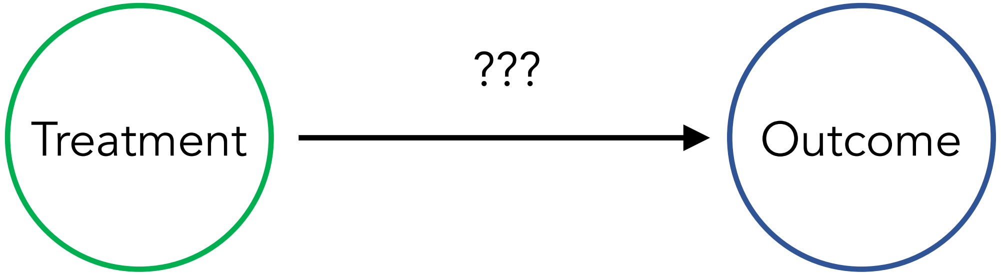
Colliders

- Never want to control for / condition on colliders
- Conditioning on a common effect causes **collider bias**, which can be in positive or negative direction

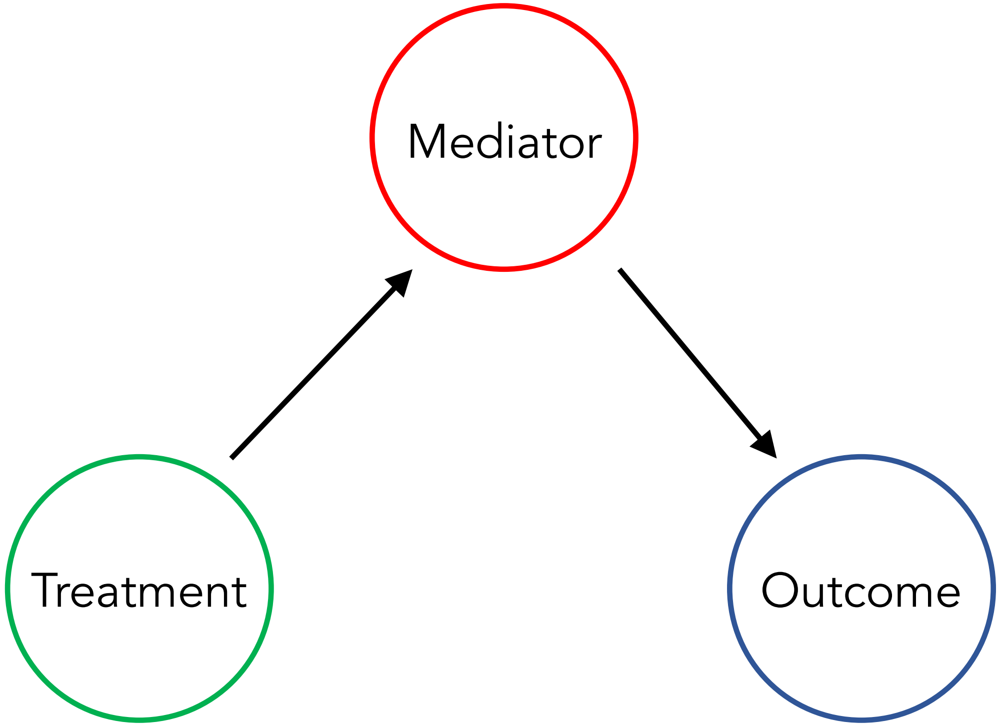


3) Mediators

Mediators

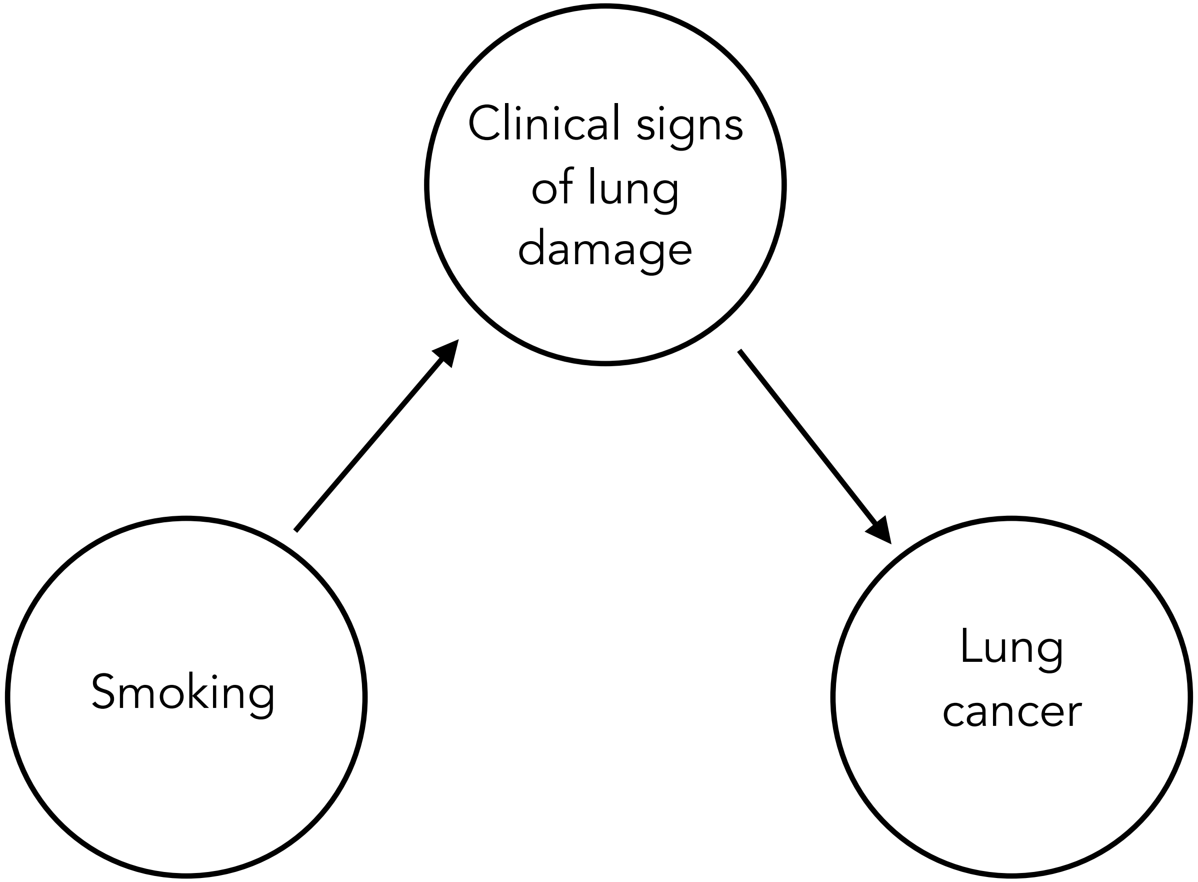


Mediators



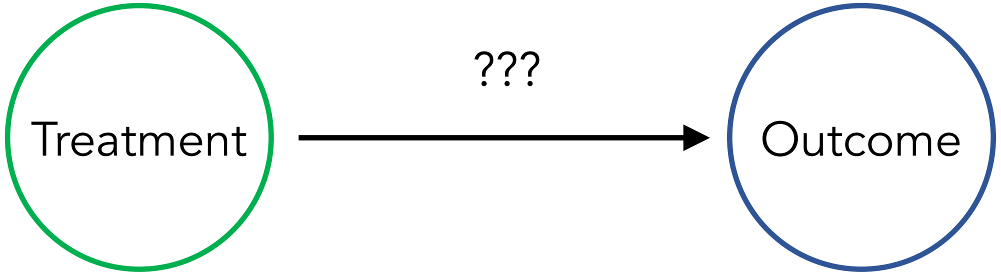
Mediators

- Controlling for a mediator will nullify associations of interest
- There are statistical tests of mediation you can use to help determine causal relationships in observational data

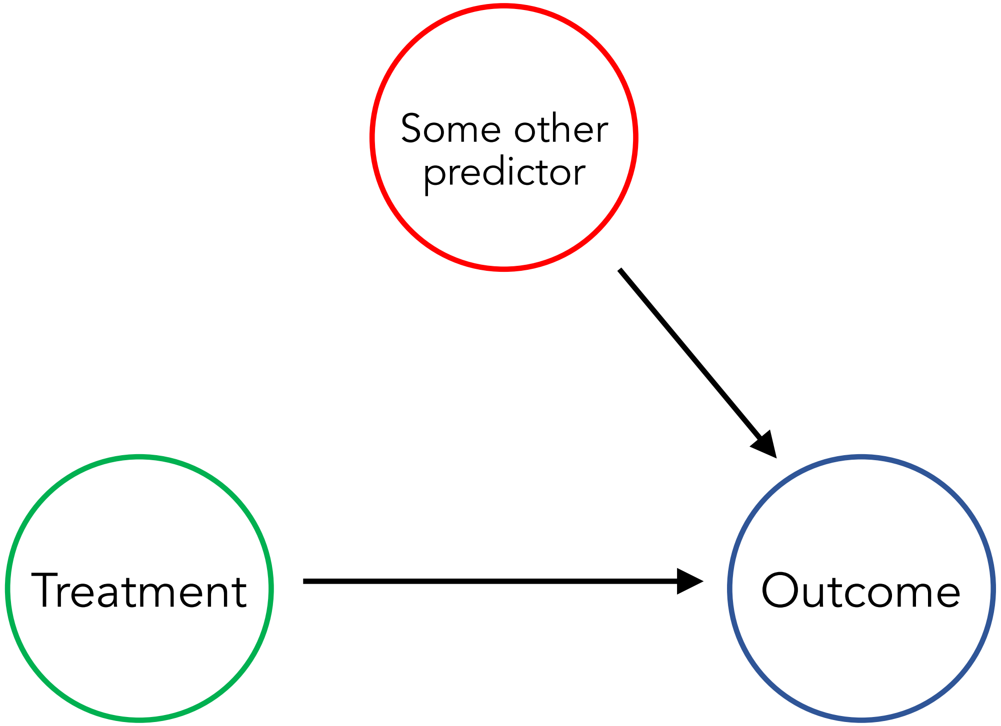


4) Unrelated Predictors

Unrelated predictors

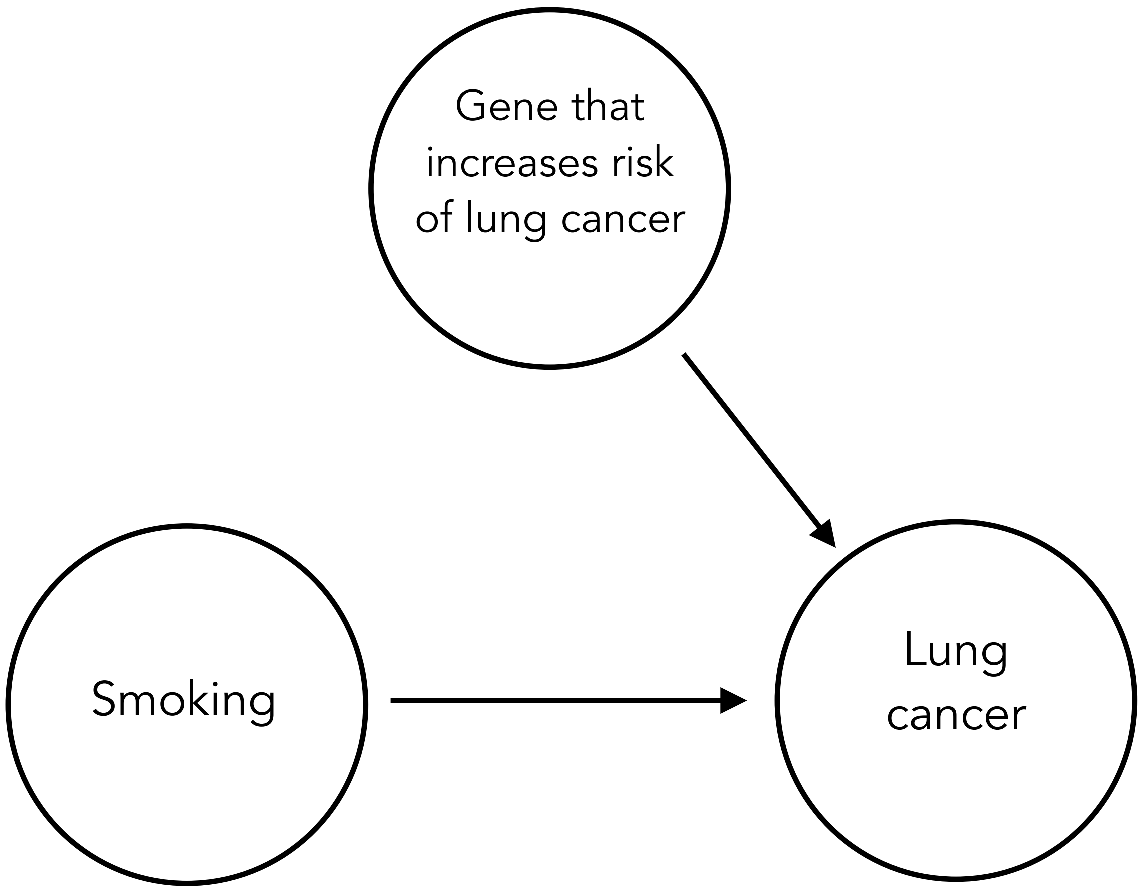


Unrelated predictors



Unrelated predictors

- If they are unrelated to your independent variable / treatment / exposure of interest, there is no harm in controlling for them.
- In fact, leaving them in could improve model performance.



G-computation

1) Start with a set of participants for whom we have complete treatment, outcome, and covariate data

ID#	Covar 1	Covar 2	treat	outcome
1	…	…	1	20
2	…	…	1	15
3	…	…	0	10
4	…	…	0	10
5	…	…	1	20

2) Train a model that predicts the outcome from all covariates and treatment variable. Aim for high recall and precision.

ID#	Covar 1	Covar 2	treat	outcome
1	...	...	1	20
2	...	...	1	15
3	...	...	0	10
4	...	...	0	10
5	...	...	1	20

3) "Force" every observation in the dataset to receive the treatment

ID#	Covar 1	Covar 2	treat	outcome
1	…	…	1	20
2	…	…	1	15
3	…	…	1	10
4	…	…	1	10
5	…	…	1	20

4) Predict outcome values with these covariate and treatment values

ID#	Covar 1	Covar 2	treat	outcome	$\hat{O}_{treat}$
1	…	…	1	20	22.5
2	…	…	1	15	16.0
3	…	…	1	10	14.0
4	…	…	1	10	17.0
5	…	…	1	20	22.5

5) Now “force” every observation to not receive treatment, And make outcome predictions again

ID#	Covar 1	Covar 2	treat	outcome	$\hat{O}_{treat}$	$\hat{O}_{untreat}$
1	…	…	0	20	22.5	18.5
2	…	…	0	15	16.0	14.0
3	…	…	0	10	14.0	11.5
4	…	…	0	10	17.0	13.0
5	…	…	0	20	22.5	19.5

6) Calculate the average difference between treated and untreated outcome estimates

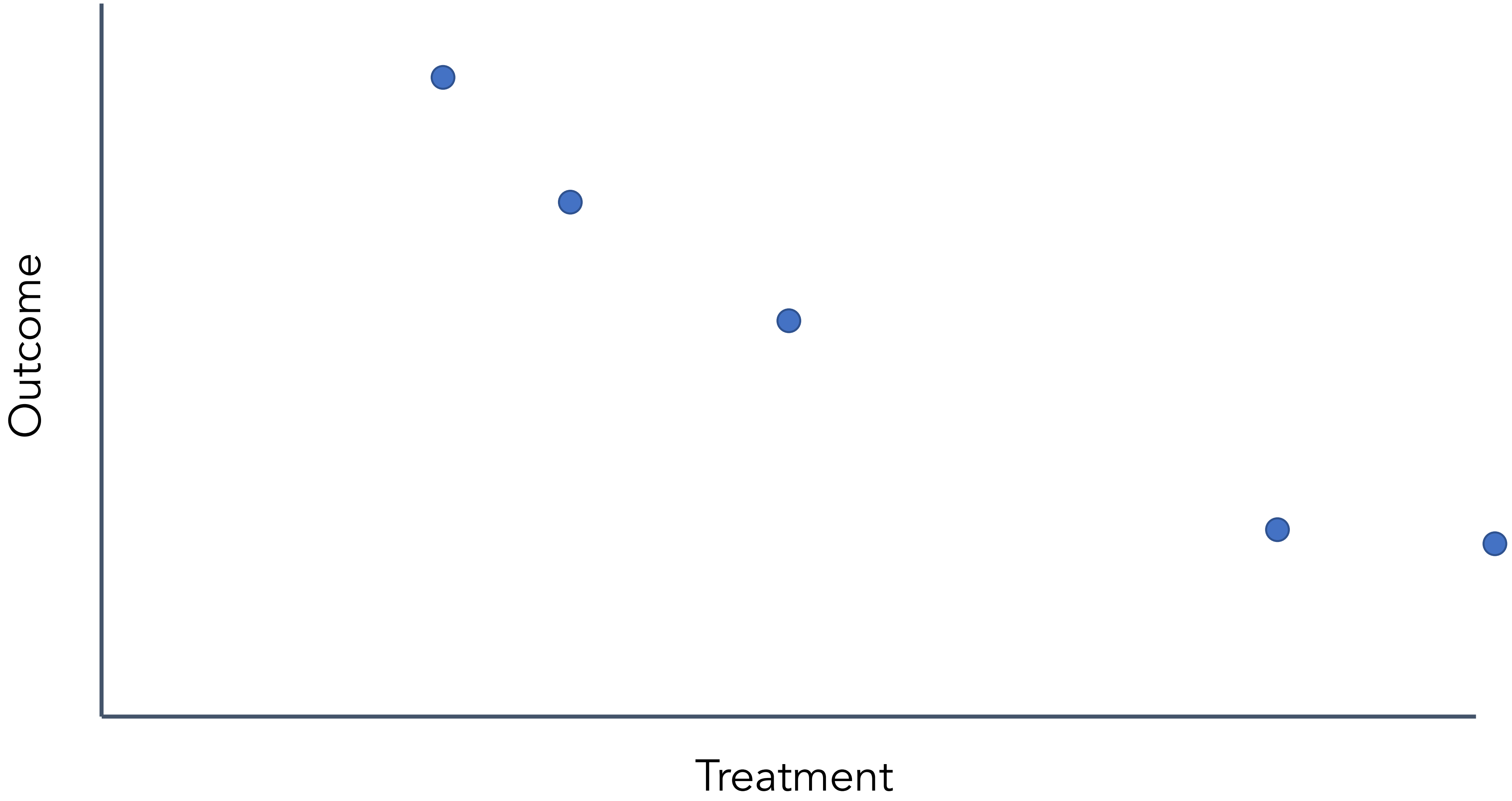
ID#	$\hat{O}_{treat}$	$\hat{O}_{untreat}$	Δ
1	22.5	18.5	4.0
2	16.0	14.0	2.0
3	14.0	11.5	2.5
4	17.0	13.0	4.0
5	22.5	19.5	3.0

$$\mu_{\Delta} = 3.1$$

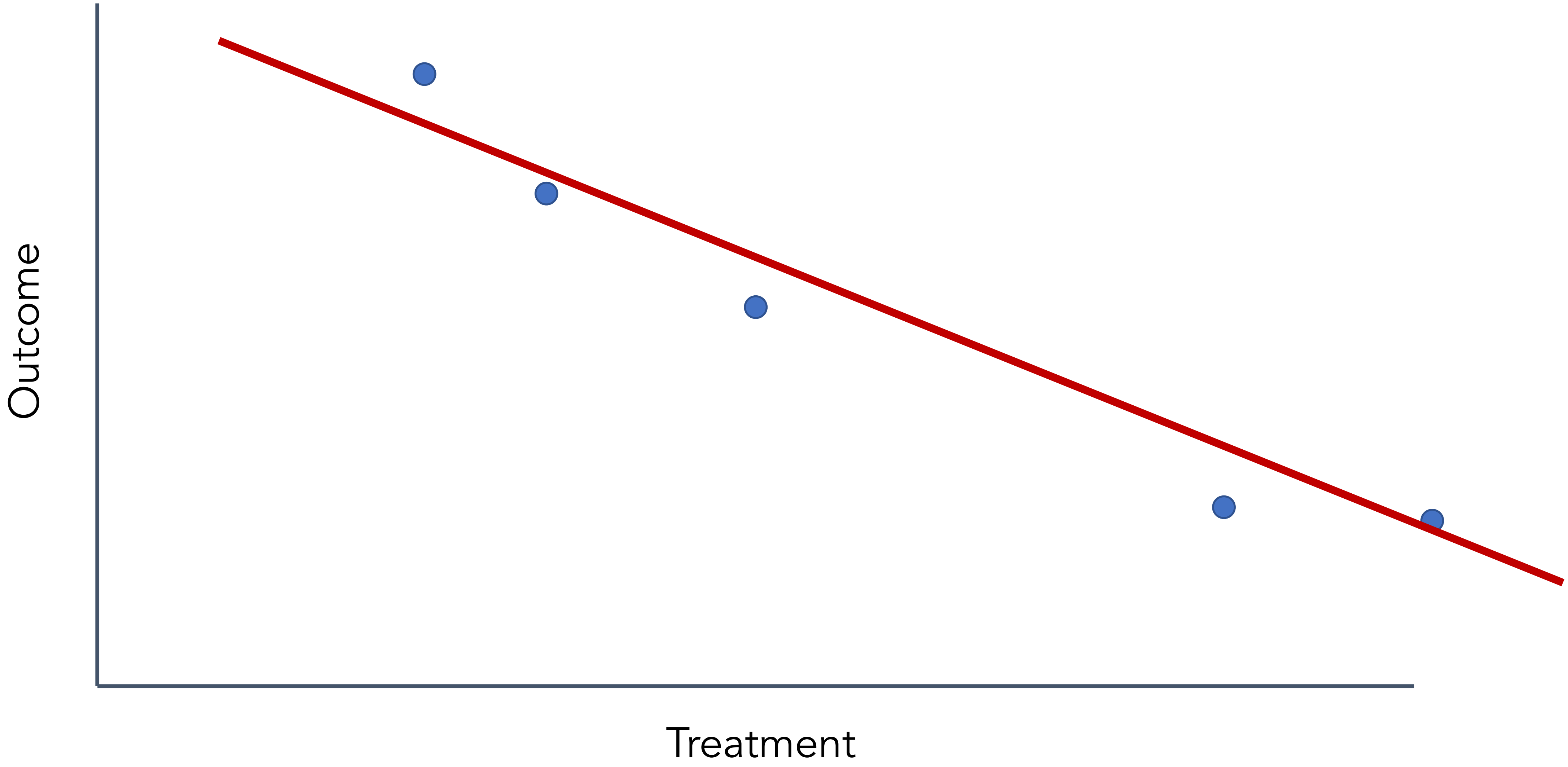
Notebook exercise #1

Causal dose-response curve estimation (AKA estimating the causal curve)

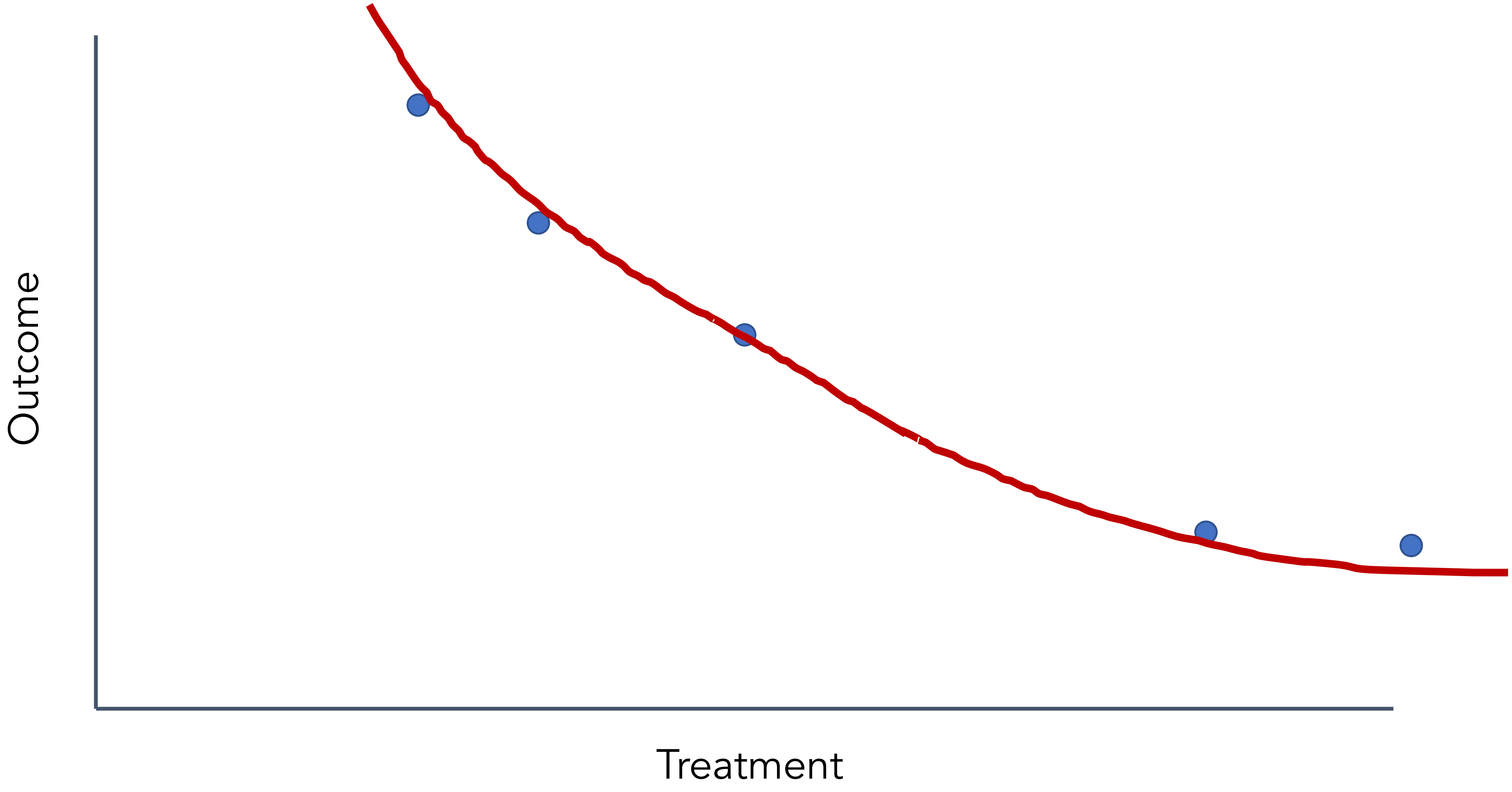
Counterfactuals (with a continuous treatment)



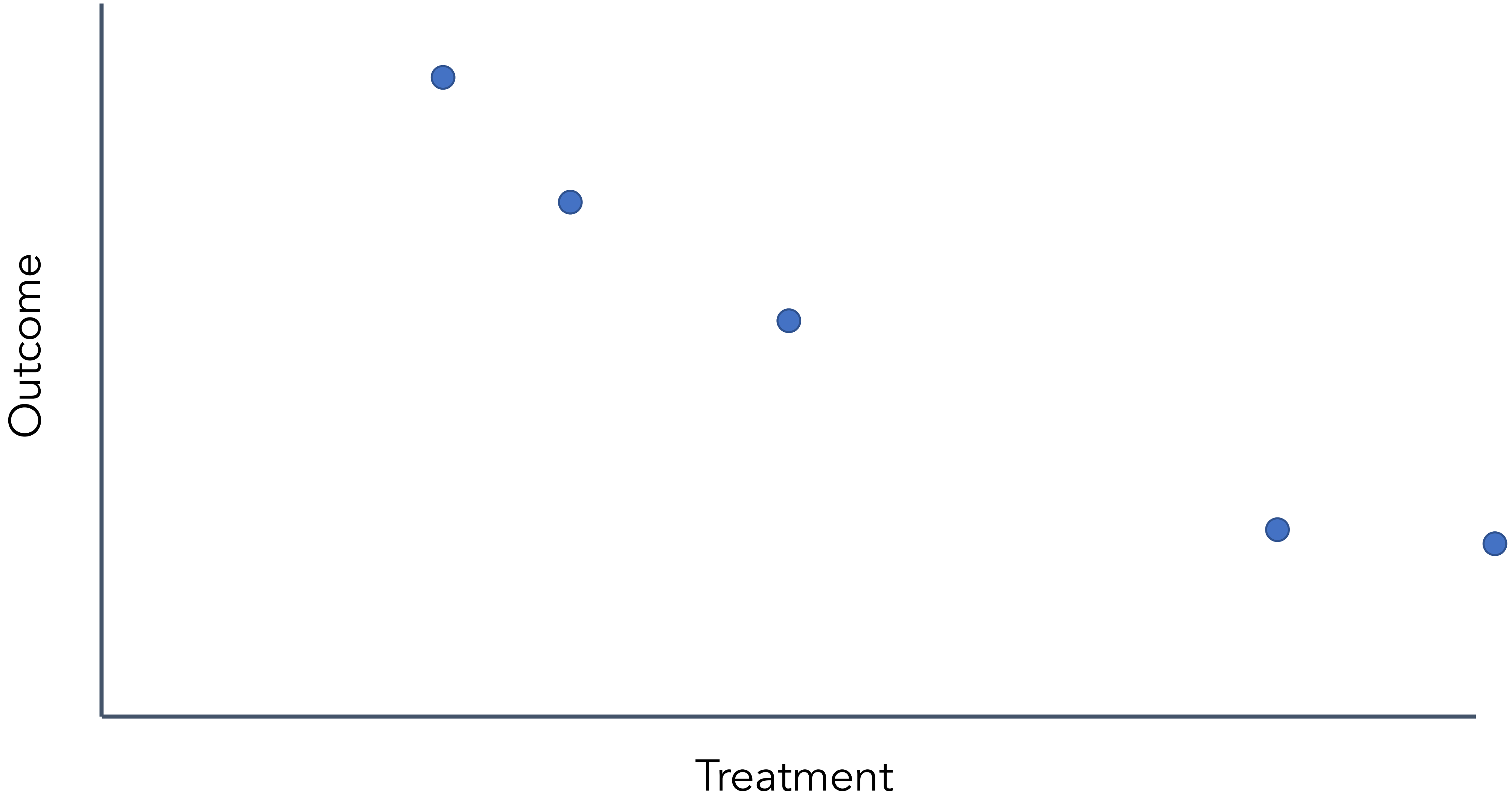
Counterfactuals (with a continuous treatment)



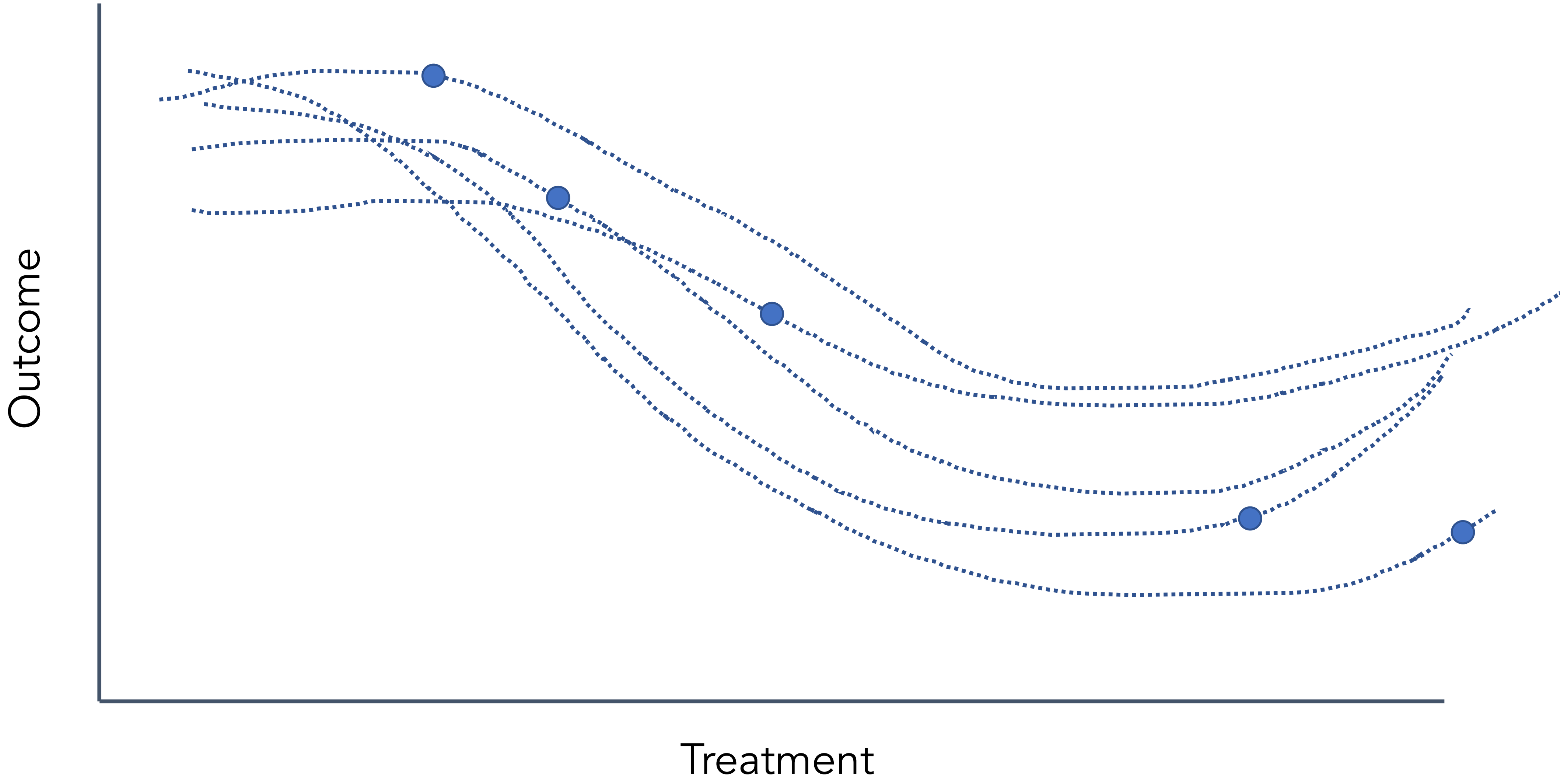
Counterfactuals (with a continuous treatment)



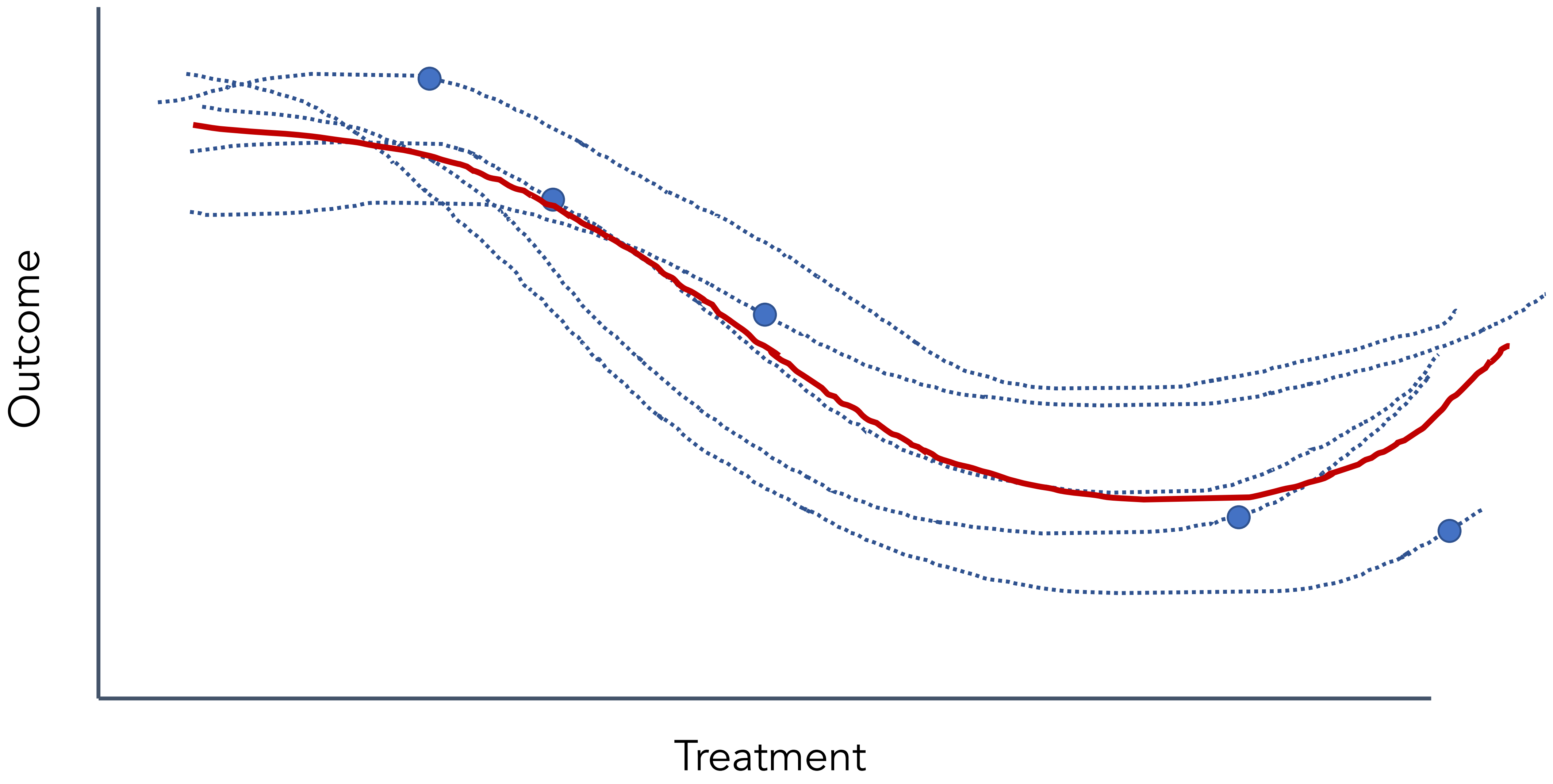
Counterfactuals (with a continuous treatment)



Counterfactuals (with a continuous treatment)



Counterfactuals (with a continuous treatment)



Estimating the "causal curve"

GPS is an extension of the standard propensity score method. It is the treatment assignment density calculated at a particular treatment value

1) Calculate the GPS associated with each treatment value observation
2) Fit a curve of treatment values predicting outcome values, adjusted for the GPS
3) The resulting treatment against outcome curve is your causal dose response curve (AKA your causal curve)

Notebook exercise #2

Closing thoughts: troubleshooting

- Having domain knowledge and understanding the data-generating process is often way more productive than just throwing an algo at the problem
- There is value in trying multiple techniques to understand their range of estimates (but use p-value correction if you're running lots of analyses)
- You'll never be able to capture all confounders, but do aim to capture the major ones
- If your results don't make sense and your code isn't buggy, you're probably missing a big source of bias
- Causal inference and modeling is powerful but still not as trustworthy as running a proper experiment. Approach all results with healthy skepticism.

Closing thoughts: the perils of multiple testing…

Statistics

Priya Ranganathan, C. S. Pramesh[1], Marc Buyse[2,3]

Department of Anaesthesiology, Tata Memorial Centre, [1]Department of Surgical Oncology, Division of Thoracic Surgery, Tata Memorial Centre, Mumbai, Maharashtra, India, [2]International Drug Development Institute, San Francisco, California, USA, [3]Department of Biostatistics, Hasselt University, Hasselt, Belgium

Common pitfalls in statistical analysis: The perils of multiple testing

Closing thoughts: be humble, it’s likely your research or business idea doesn’t work!

O’REILLY®

TEAMS INDIVIDUALS FEATURES BLOG CONTENT SPONSORSHIP

Radar / Business

The Sobering Truth About the Impact of Your Business Ideas

By Eric Colson, Daragh Sibley and Dave Spiegel

October 26, 2021