Machine Learning for Healthcare

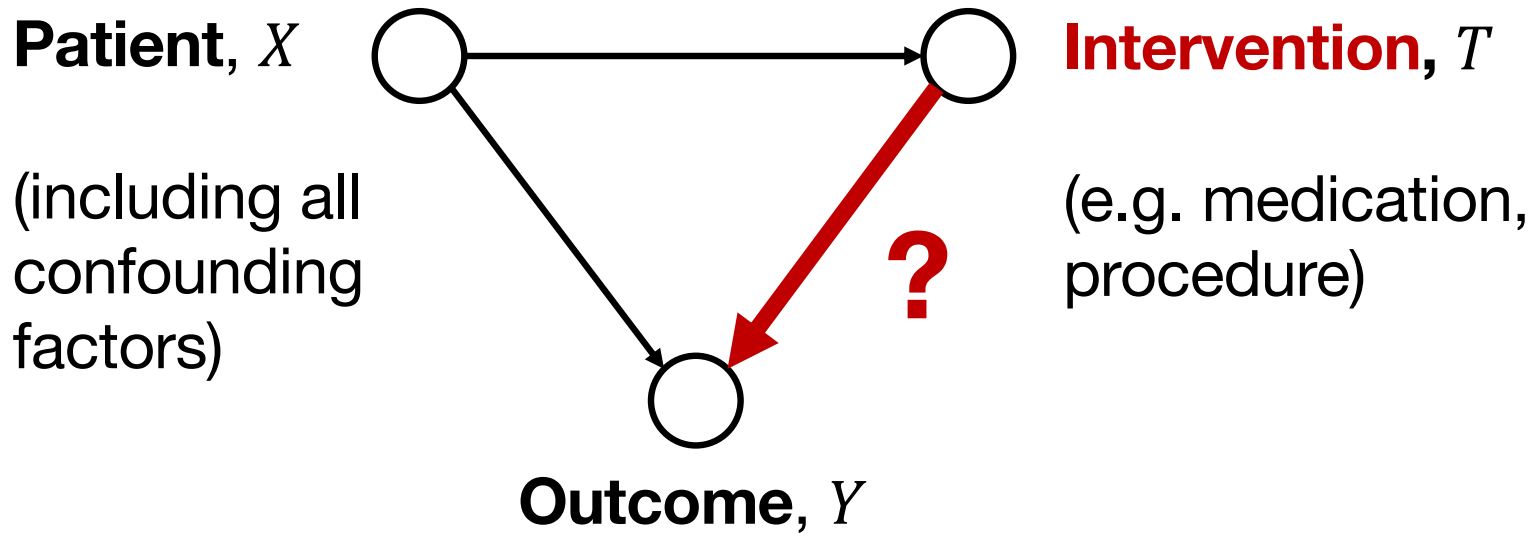
6.871, HST.956

Lecture 12: Causal Inference Part 3

David Sontag

HEALTH SCIENCES
& TECHNOLOGY

Reminder: Causal inference



High dimensional

Observational data

Reminder: Causal inference

- Two approaches to use machine learning for causal inference
  - Predict outcome given features and treatment – i.e., E[Y | X, T] – then use to impute counterfactuals (*covariate adjustment*)
  - Predict treatment using features (*propensity score*) – i.e., Pr(T|X) – then use to reweight outcomes

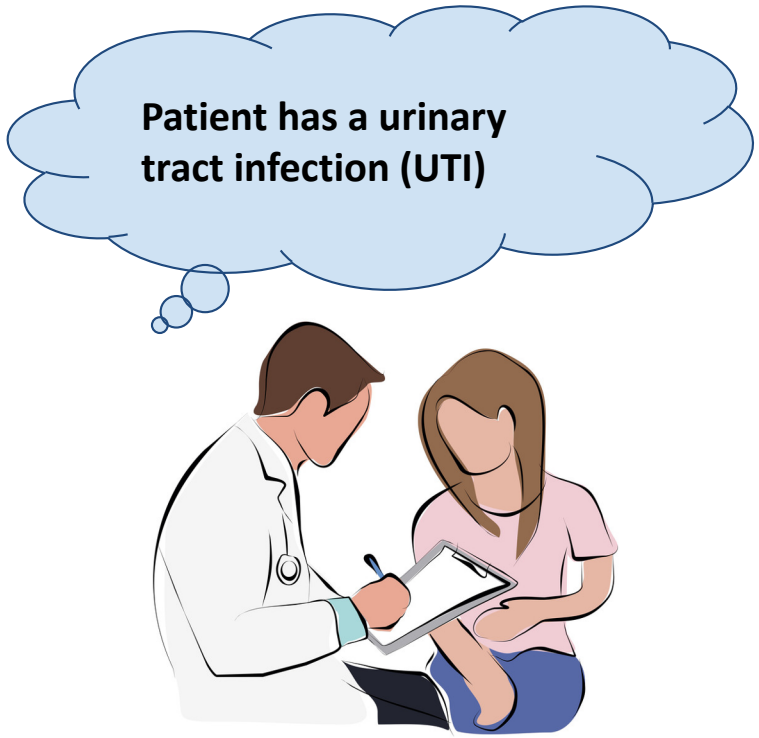
Consistency of estimates depend on:

- Causal graph being correct (i.e., no unobserved confounding)
- Identifiability of causal effect (i.e., overlap or correctly specified model)

Same ideas can be used to evaluate *policies* using observational data

- Suppose someone gave us a policy $\pi(l)$ that outputs a_1 vs a_2

  Example: which antibiotic to prescribe?



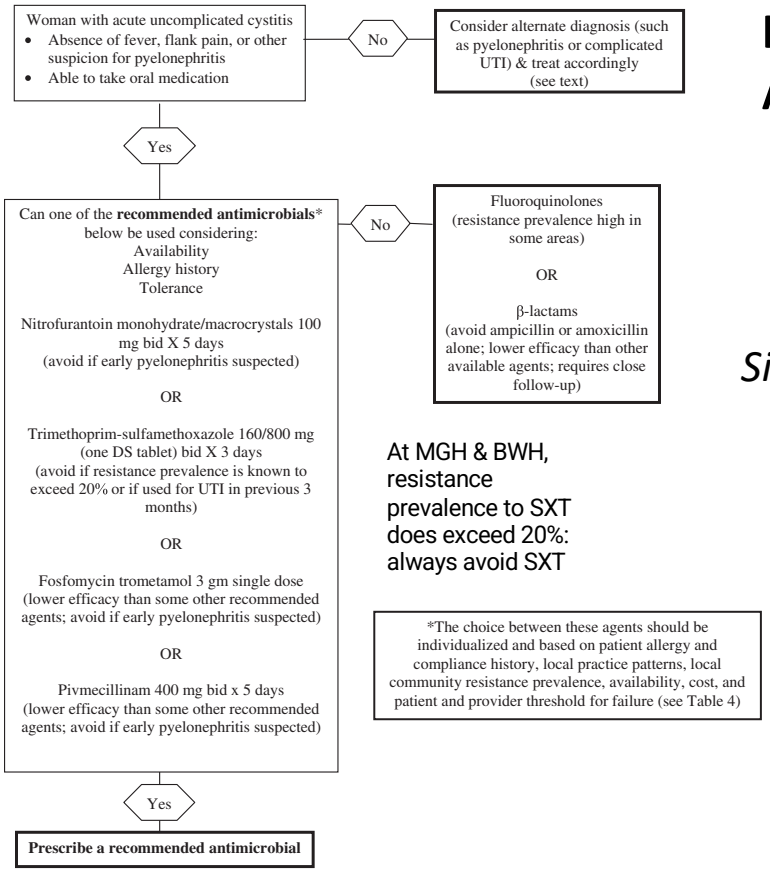
Affects 1 in 2 women during lifetime; 3rd most common cause for antibiotic treatment

[Kanjilal et al., A decision algorithm to promote outpatient antimicrobial stewardship for uncomplicated urinary tract infection. *Science Translational Medicine*, 2020.]

Same ideas can be used to evaluate *policies* using observational data

- Suppose someone gave us a policy $\pi(l)$ that outputs a_1 vs a_2

Example: which antibiotic to prescribe?



Infectious Disease Society of America (IDSA) guidelines

At MGH & BWH, resistance prevalence to SXT does exceed 20%: always avoid SXT

Simplifies to ⇒

Resistance or exposure to NIT in past 90 days?

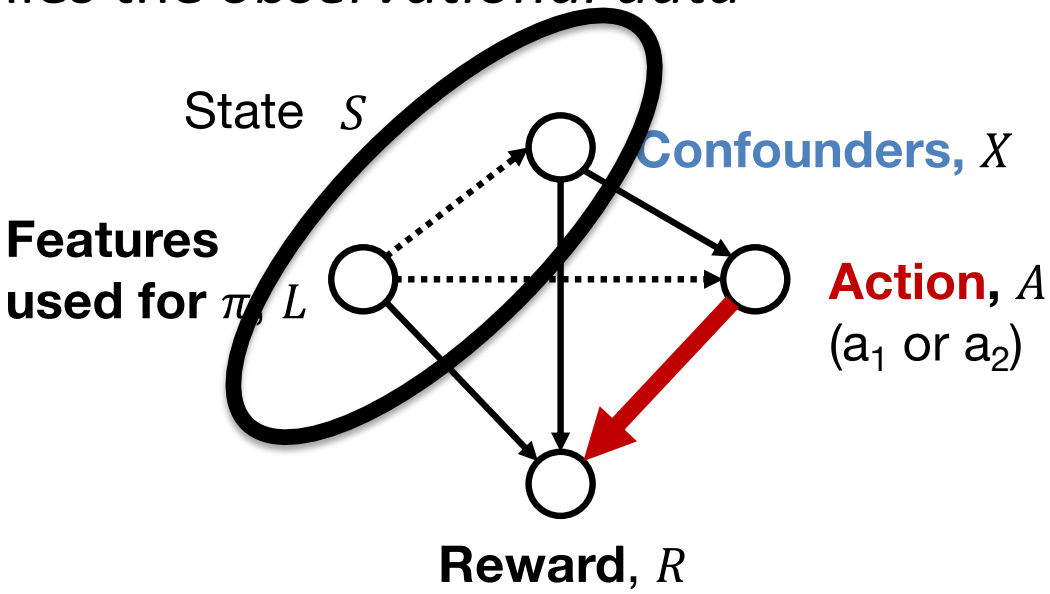
No → Prescribe NIT (Nitrofurantoin)

Yes → Prescribe CIP (Ciprofloxacin)

[Gupta et al., *Clinical Infectious Diseases*, 2011.]

Same ideas can be used to evaluate *policies* using observational data

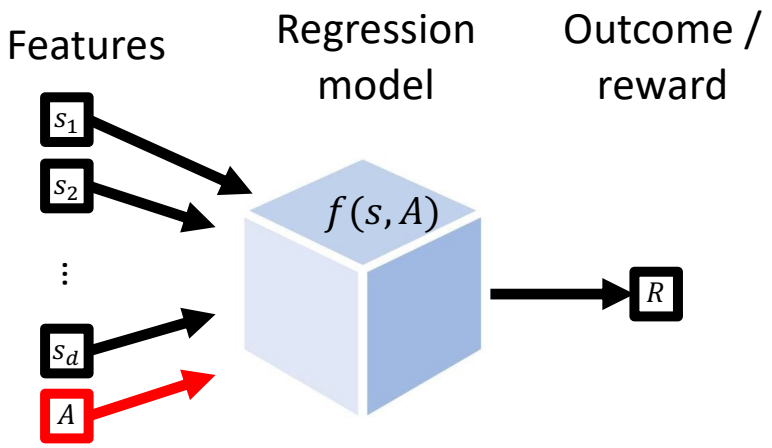
- Suppose someone gave us a policy $\pi(l)$ that outputs a_1 vs a_2
- How do we evaluate it?
- We give two approaches, one based on potential outcomes and the other based on propensity scores
- In both cases, we have to first consider the causal graph that underlies the *observational data*



Switched notation to what's more typically used in RL
action *A*: Treatment *T*
reward *R*: Outcome *Y*

Evaluating policies using potential outcomes

- First, use machine learning to obtain a model that can predict potential outcomes (we need ignorability, overlap)

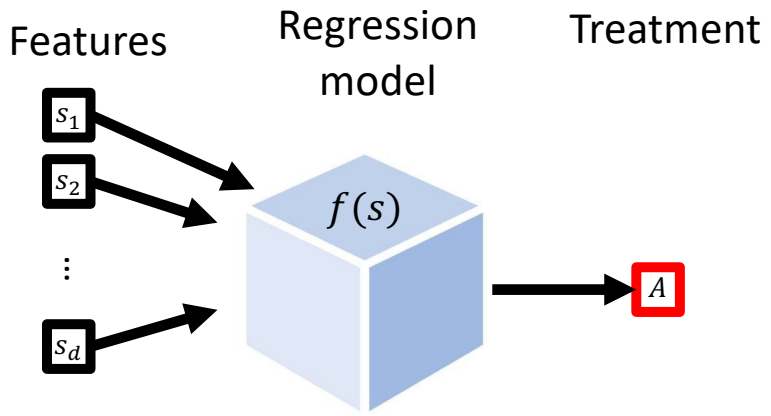


- Then, use this model to impute policy outcomes:

$$\hat{Q}(\pi) = \frac{1}{n} \sum_{i=1}^{n} f(l_i, x_i, \pi(l_i))$$

Evaluating policies using inverse propensity scores

- First, use machine learning to obtain $\hat{p}(A|s) = f(s)$, estimated propensity scores



- Then, use this model to reweight the outcomes:

$$\hat{Q}^{IPW}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\frac{1[a_i = \pi(l_i)]}{\hat{p}(a_i \mid s_i)}R_i$$

Aside: is this the right goal? What if we wanted to control worst-case reward instead of average?

Learning policies from observational data

- Consider our first estimator: $\hat{Q}(\pi) = \frac{1}{n}\sum_{i=1}^{n} f(l_i, x_i, \pi(l_i))$

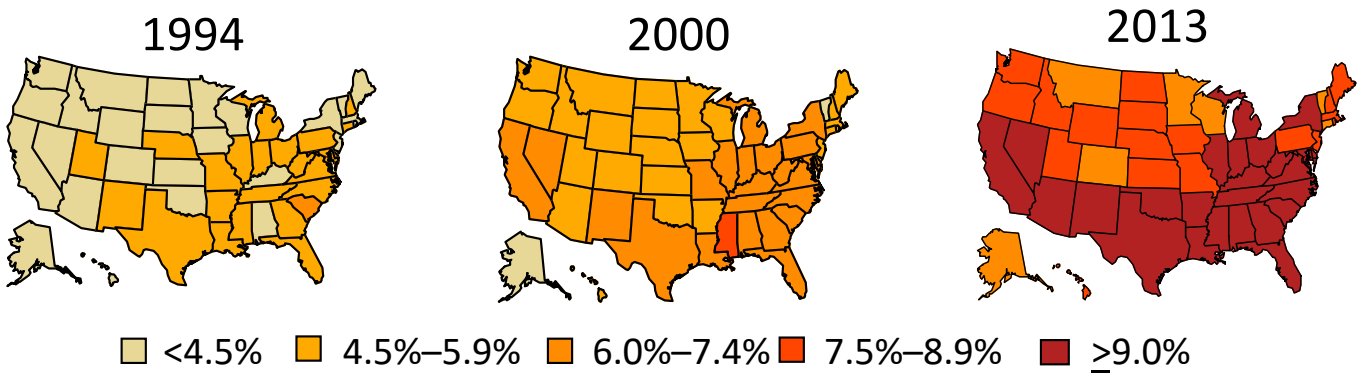
- Create data set $\{(l_i, o_i)\}$ where

$$o_i = \arg\max_A f(l_i, x_i, A)$$ Notice relationship to CATE

- Use an (interpretable) ML algorithm to fit this new dataset
- The resulting policy may be a much simpler function than *f*!

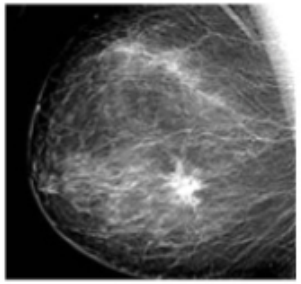
(Makar, Swaminathan, Kiciman. A distillation approach to data efficient individual treatment effect estimation. AAAI, 2019)

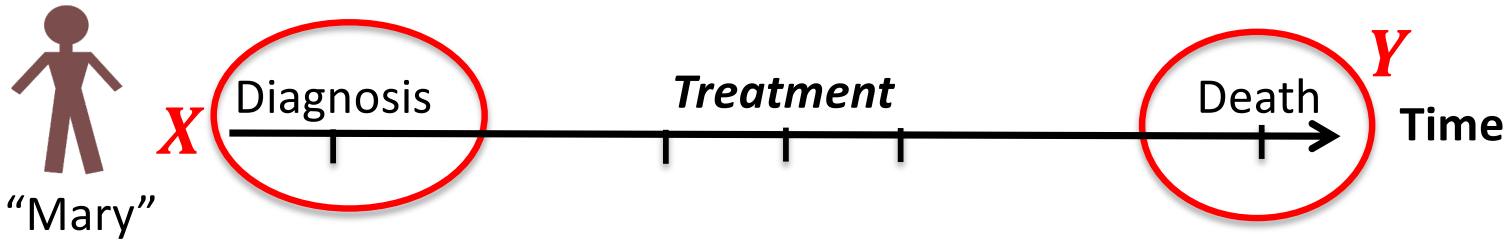
Does gastric bypass surgery prevent onset of diabetes?



- Gastric bypass surgery is the highest negative weight (9th most predictive feature)
  - Does this mean it would be a good intervention?
- Yes, *if*….
  - Interpret ‘gastric bypass surgery’ feature as T
  - Interpret all the other features as X; assume they all include all relevant confounders and do not include anything post-treatment
  - True potential outcome function is linear

What is the likelihood this patient, with breast cancer, will survive 5 years?





A long survival time may be because of treatment!

- Group into K categories of treatment strategies T (one of which might be "no treatment")
- Gather data on confounding factors C that might influence both treatment decision and outcome
- Learn f(X,C,T) to predict Y (survival time)
- Assess overlap* by looking at p(X,C|T) or p(T|X,C)
- Predict survival under a specific treatment regime *k* using f(X,C,*k*)
- Will survive 5 years when treated *optimally* if $\max_k$ f(X,C, *k*) > 5

* See, e.g., Oberst, Johansson, Wei, Gao, Brat, Sontag, Varshney. Characterization of Overlap in Observational Studies, Conference on Artificial Intelligence and Statistics (AI-STATS), 2020.

Many more ideas and methods

- Doubly robust estimators that combine both regression and IPW
- Natural experiments & regression discontinuity
- Instrumental variables
- Sensitivity analyses

Many more ideas and methods – Natural experiments

- Does stress during pregnancy affect later child development?
- Confounding: genetic, mother personality, economic factors…
- Natural experiment: the Cuban missile crisis of October 1962. Many people were afraid a nuclear war is about to break out.
- Compare children who were in utero during the crisis with children from immediately before and after

Many more ideas and methods – Instrumental variables

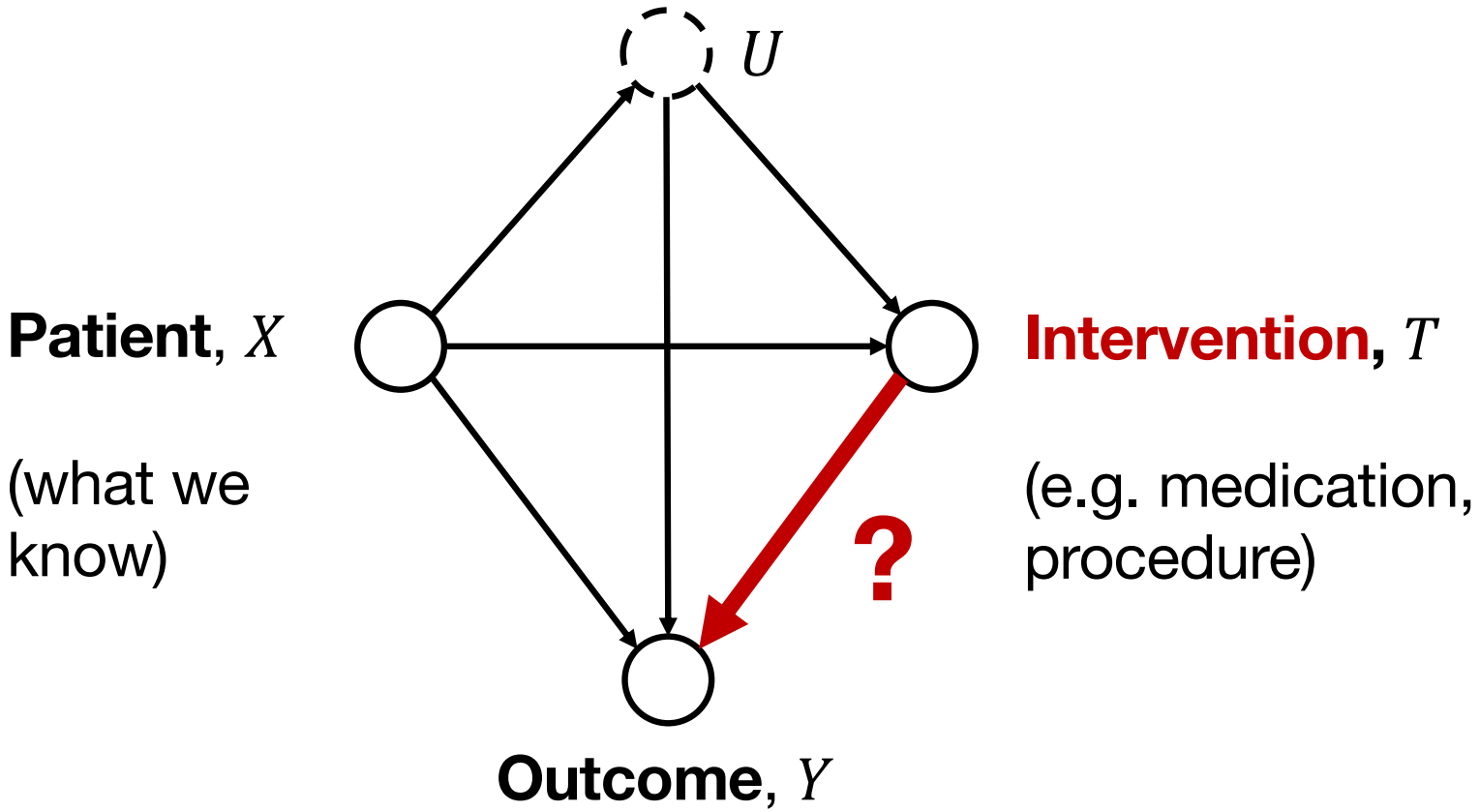
- Informally: a variable which affects treatment assignment but not the outcome
- Example: are private schools better than public schools? Which students would benefit the most?
- Confounding: different student population, different teacher population
- Can’t force people which school to go to

Many more ideas and methods – Instrumental variables

- Informally: a variable which affects treatment assignment but not the outcome
- Example: are private schools better than public schools? Which students would benefit the most?
- Can't force people which school to go to
- Can *randomly* give out vouchers to some children, giving them an opportunity to attend private schools
- The voucher assignment is the instrumental variable

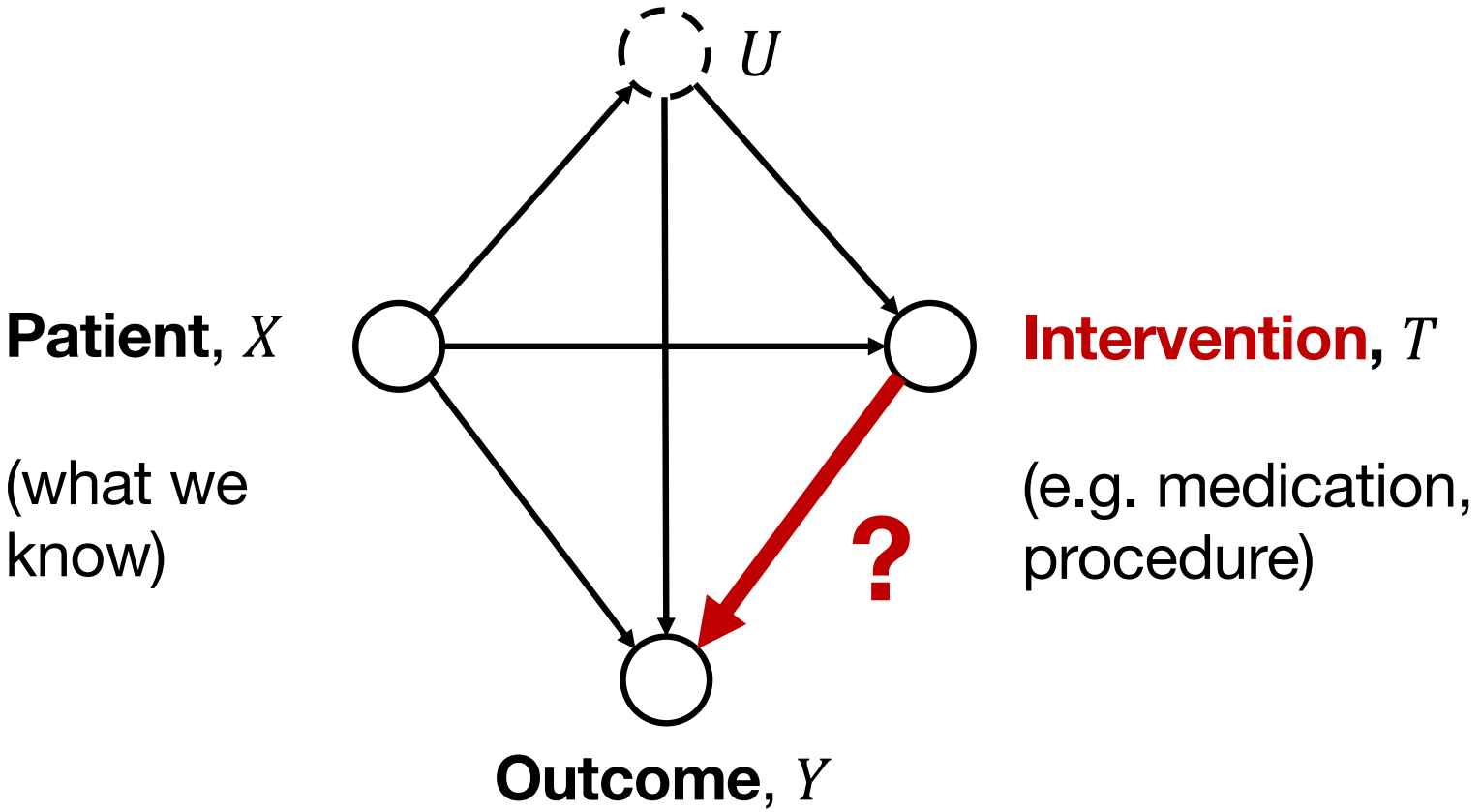
Estimation using an instrumental variable

Goal: estimation in setting where there are unobserved confounders, *U*, not captured in *X*



Estimation using an instrumental variable

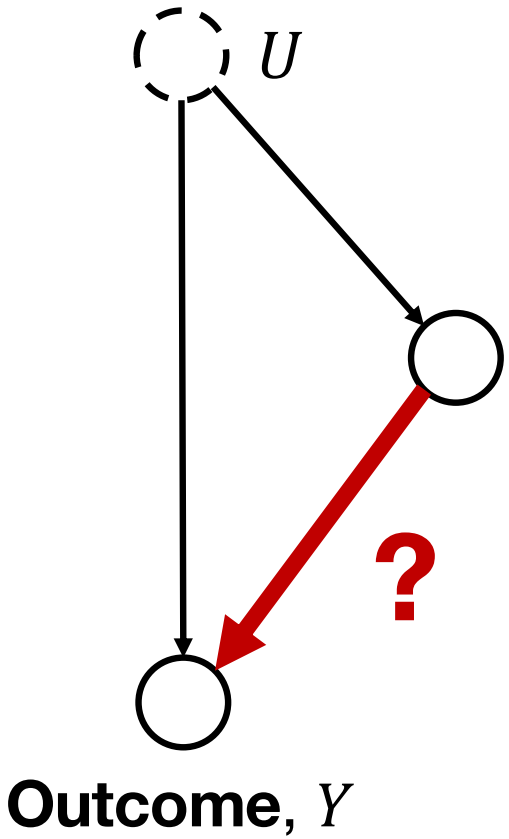
First, assume no patient covariates (with this, we will only be able to estimate ATE not CATE)



Estimation using an instrumental variable

First, assume no patient covariates (with this, we will only be able to estimate ATE not CATE)

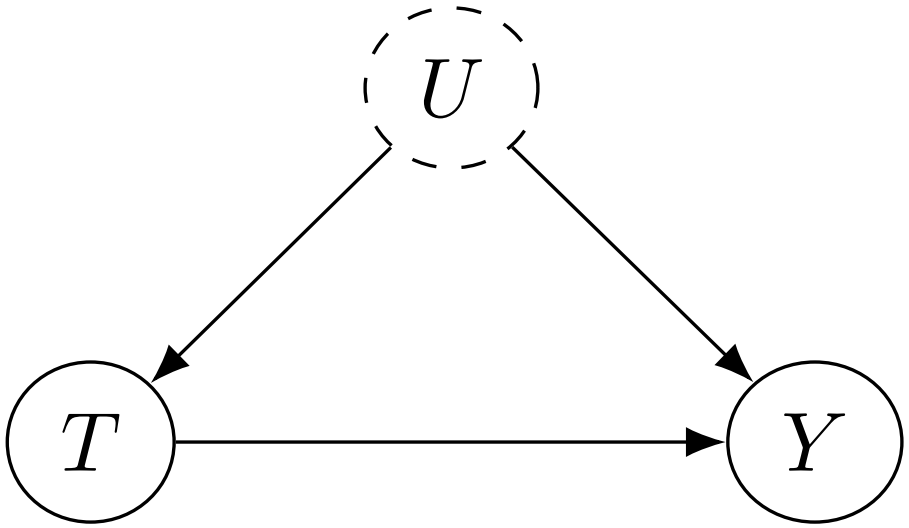
Note: this is without loss of generality (since U could include all of X)



Intervention, T

(e.g. medication, procedure)

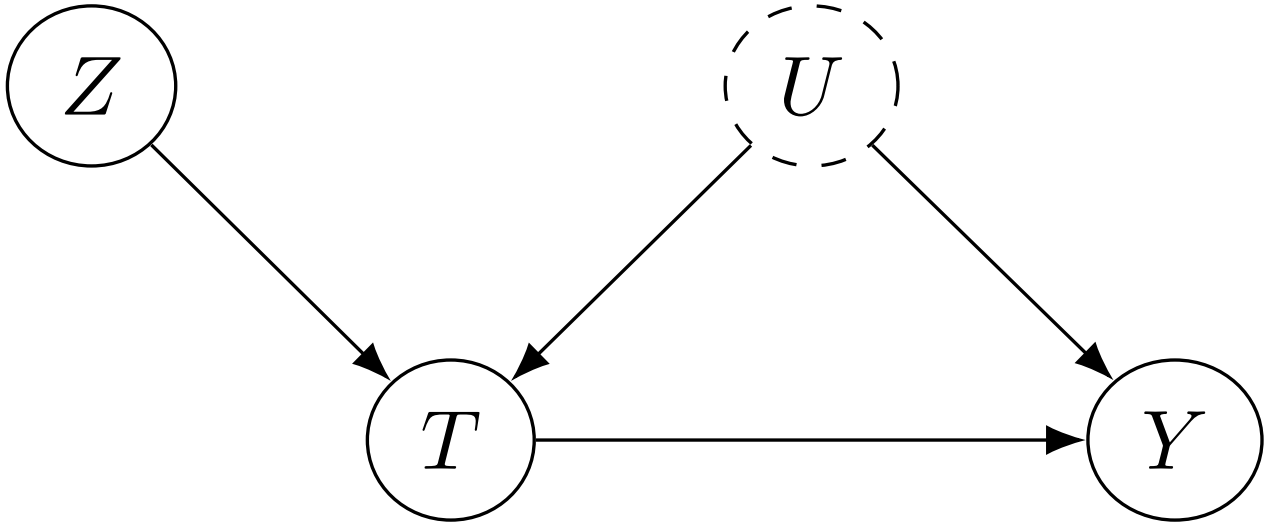
Estimation using an instrumental variable



(Slides adapted from Brady Neal's Introduction to Causal Inference class)

Estimation using an instrumental variable

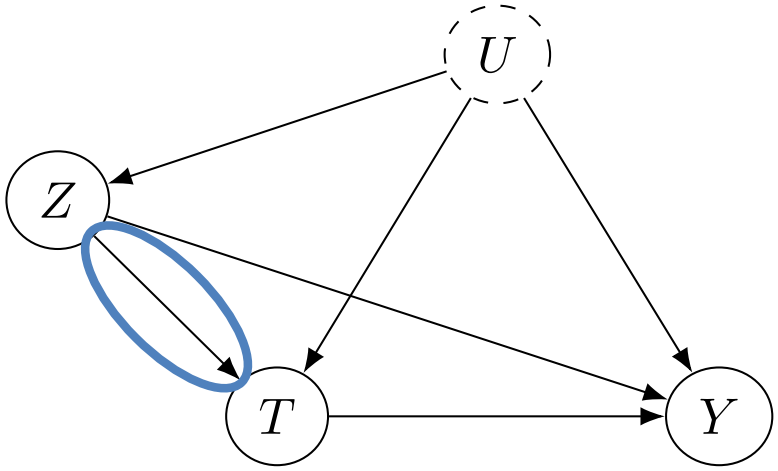
Instrument (e.g., voucher)



(Slides adapted from Brady Neal's Introduction to Causal Inference class)

Assumption 1: Relevance

Z has a causal effect on T

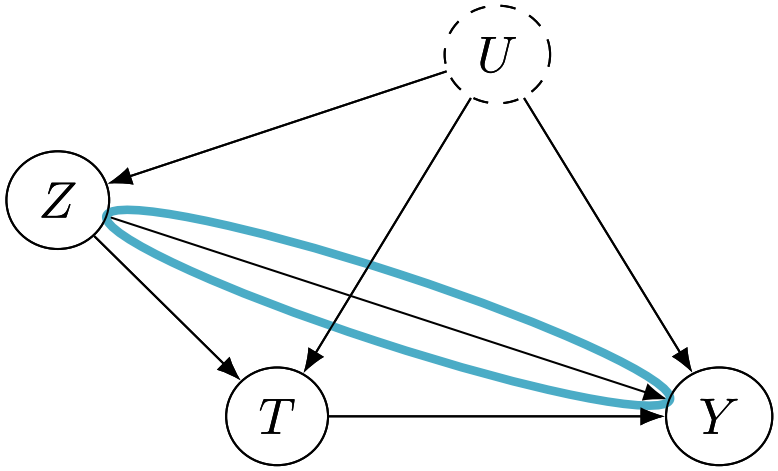


What is an Instrument?

(Slides adapted from Brady Neal's Introduction to Causal Inference class)

Assumption 2: Exclusion Restriction

The causal effect of Z on Y is fully mediated by T

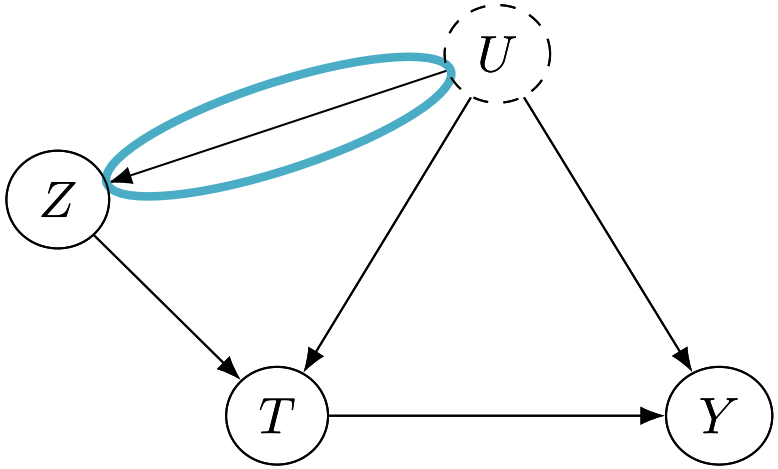


What is an Instrument?

(Slides adapted from Brady Neal's Introduction to Causal Inference class)

Assumption 3: Instrumental Unconfoundedness

Z is unconfounded (in the setting of no *X*, this simply means *U* and *Z* are independent)



What is an Instrument?

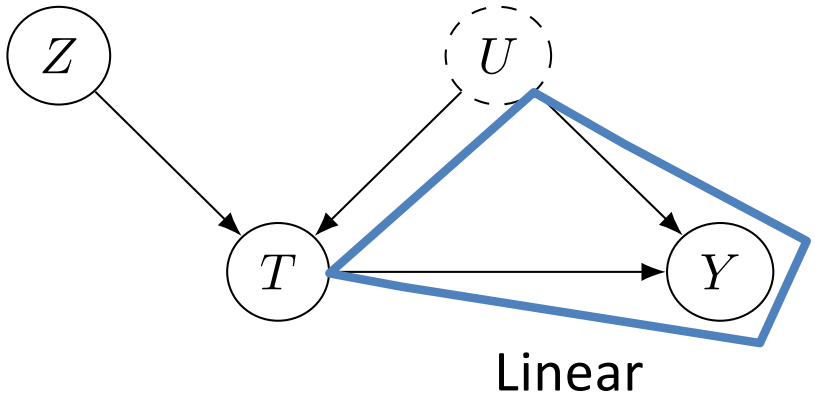
(Slides adapted from Brady Neal's Introduction to Causal Inference class)

Warm-up: linear potential outcome, no *X*

Assume potential outcomes given by the linear model,

$$Y_t(U) = \alpha_u U + \delta \cdot t + \epsilon_t, \quad \mathbb{E}[\epsilon_t] = 0$$

Z doesn't appear because of the exclusion restriction assumption



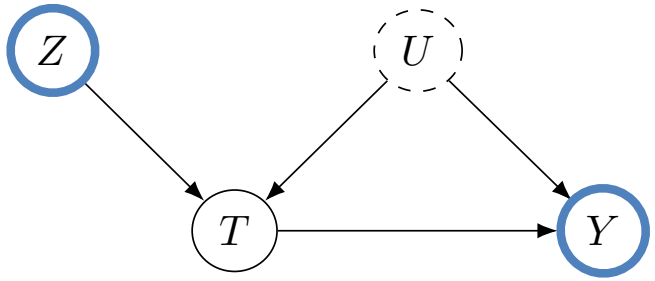
(Slides adapted from Brady Neal's Introduction to Causal Inference class)

Warm-up: linear potential outcome, no *X*

$$
\begin{aligned}
&\mathbb{E}[Y \mid Z=1] - \mathbb{E}[Y \mid Z=0] \\
&\quad = \mathbb{E}[\delta T + \alpha_u U \mid Z=1] - \mathbb{E}[\delta T + \alpha_u U \mid Z=0] && \text{(exclusion restriction and linear outcome assumptions)} \\
&\quad = \delta\left(\mathbb{E}[T \mid Z=1] - \mathbb{E}[T \mid Z=0]\right) + \alpha_u\left(\mathbb{E}[U \mid Z=1] - \mathbb{E}[U \mid Z=0]\right) \\
&\quad = \delta\left(\mathbb{E}[T \mid Z=1] - \mathbb{E}[T \mid Z=0]\right) + \alpha_u\left(\mathbb{E}[U] - \mathbb{E}[U]\right) && \text{(instrumental unconfoundedness assumption)} \\
&\quad = \delta\left(\mathbb{E}[T \mid Z=1] - \mathbb{E}[T \mid Z=0]\right)
\end{aligned}
$$

$$
\delta = \frac{\mathbb{E}[Y \mid Z=1] - \mathbb{E}[Y \mid Z=0]}{\mathbb{E}[T \mid Z=1] - \mathbb{E}[T \mid Z=0]}
$$

(non-zero due to relevance assumption)



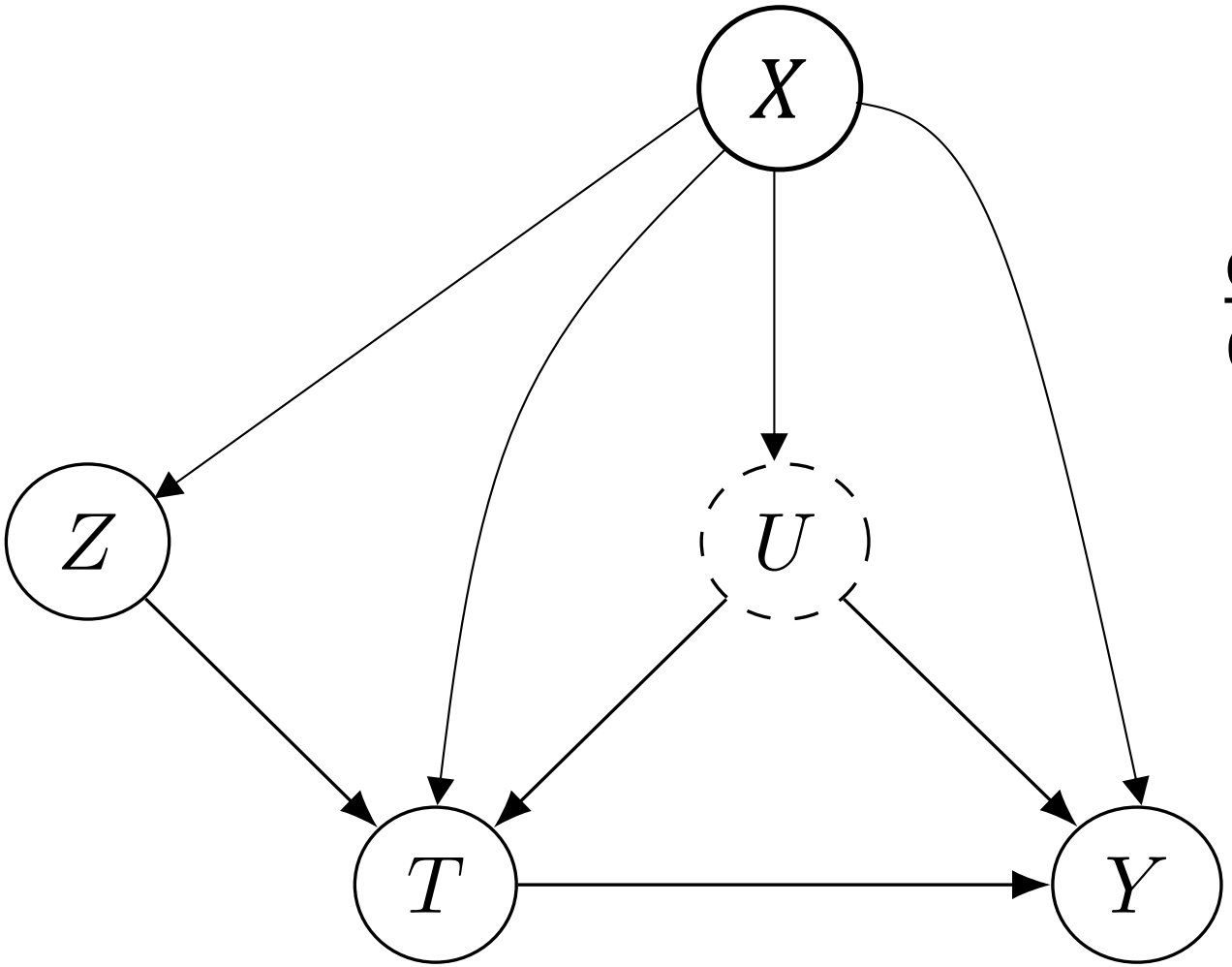
$$
Y_t(U) = \alpha_u U + \delta \cdot t + \epsilon_t
$$

(Slides adapted from Brady Neal's Introduction to Causal Inference class)

Estimation using (conditional) instruments

Assume potential outcomes given by:

$$Y_T(x, U) = \delta(x)T + g(x, U) + \epsilon_T$$



Goal: estimate

$$\text{CATE}(x) = \delta(x)$$

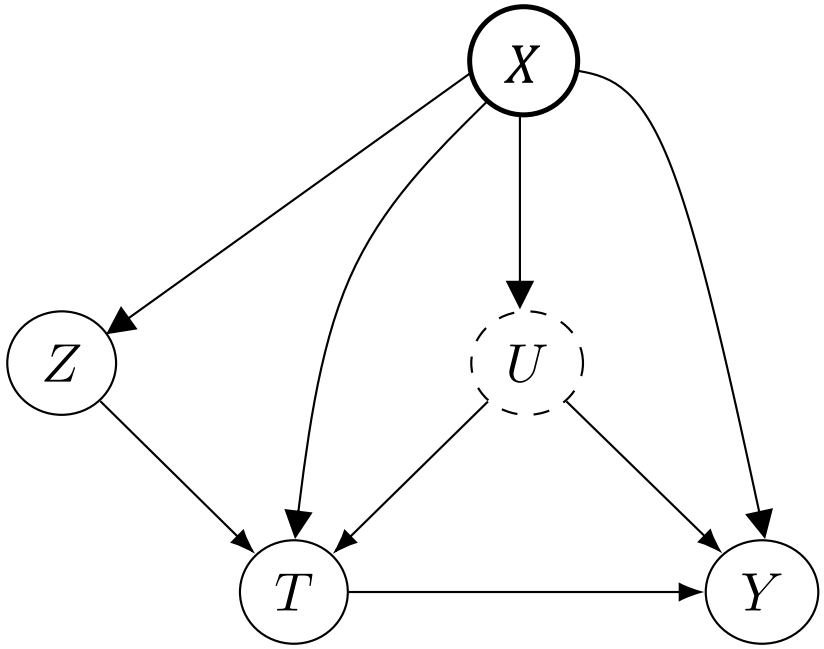
Estimation using (conditional) instruments

Assume potential outcomes given by:

$$Y_T(x, U) = \delta(x)T + g(x, U) + \epsilon_T(x)$$

Theorem: $\text{CATE}(x) = \delta(x) = \dfrac{\mathbb{E}[Y|Z=1,x] - \mathbb{E}[Y|Z=0,x]}{p(T=1 \mid Z=1,x) - p(T=1 \mid Z=0,x)}$

(proof shown on board)



Assume

$\mathbb{E}[\epsilon_0 | x] = 0$

$\mathbb{E}[\epsilon_1 | x] = 0$

What if you have unobserved confounding but no instrument?

Sensitivity analysis will help us build intuition on how biased our estimates might be

Sensitivity analysis and hidden confounding

- Major challenge: how to define the amount of hidden confounding?
- This is not a purely mathematical problem! We need to frame it in terms that enable us to make judgement calls about plausible and implausible levels of hidden confounding

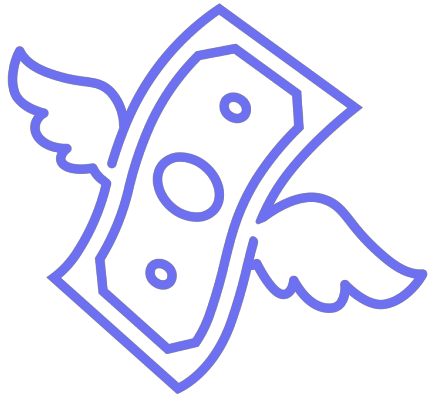
(Slides adapted from Uri Shalit's causal inference class)

Scenario #1

Patients treated with blood pressure drug A live longer than patients without on average.

However, drug A is very expensive, so mostly wealthy patients get drug A.

If income is not in our dataset, it could be very likely that it explains much or all of the ATE due to general lifestyle factors

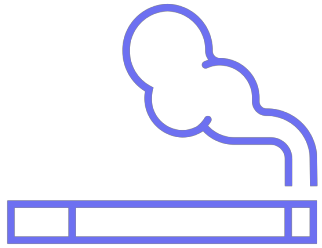


(Example from Monica Agrawal)

Scenario #2

Patients who smoke are likelier to develop lung cancer than patients who don't.

There is believed to be some heritability for both addiction and lung cancer.

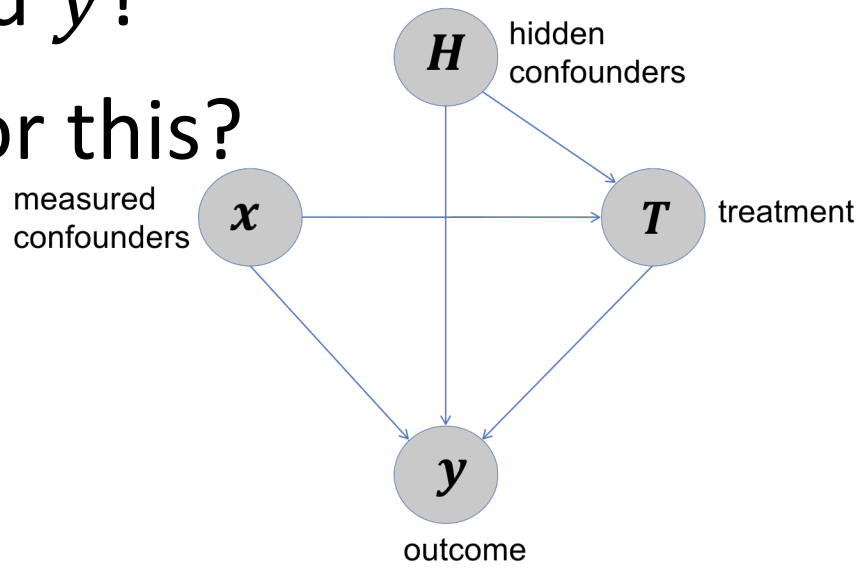


Even if patients' mutations are not in the dataset, it is unlikely that the genetic factors are sufficient to overpower the overwhelming ATE.

(Example from Monica Agrawal)

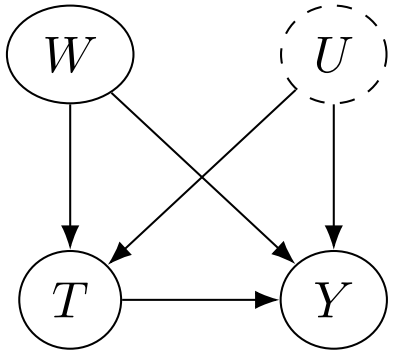
Sensitivity analysis and hidden confounding

- How to define the amount of hidden confounding?
- How much H affects T and y?
- What "units" do we use for this? How to ground it?



(Slides adapted from Uri Shalit's causal inference class)

Special case to build intuition



Continuous T

Notation change (!)
these slides use *W*
instead of *X*

Linear T and no randomness

Linear Y

$$T := \alpha_w W + \alpha_u U$$

$$Y := \beta_w W + \beta_u U + \delta T$$

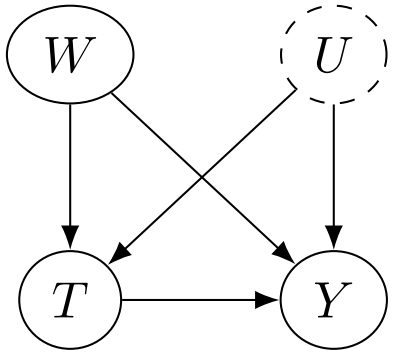
Goal: recover δ

Sensitivity Analysis: Linear Single Confounder

(Slides adapted from Brady Neal's Introduction to Causal Inference class)

Bias in Simple Linear Setting

$$T := \alpha_w W + \alpha_u U$$
$$Y := \beta_w W + \beta_u U + \delta T$$



Proof coming after next slide

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_{W,U}\left[\mathbb{E}[Y \mid T = 1, W, U] - \mathbb{E}[Y \mid T = 0, W, U]\right] = \delta$$

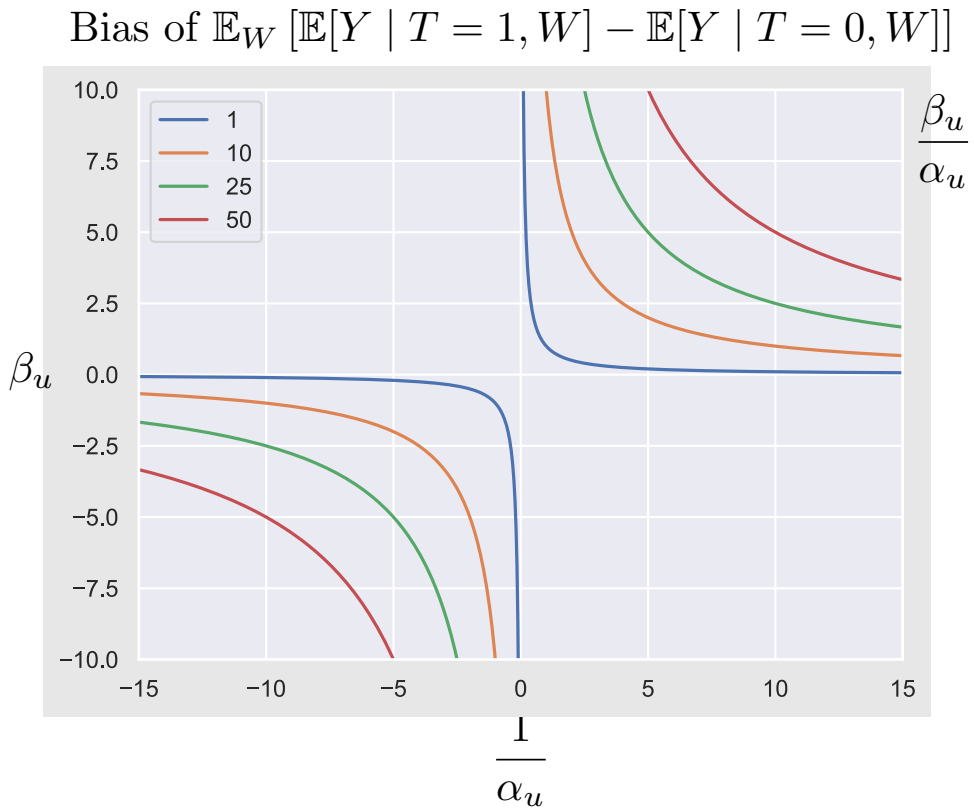
$$\mathbb{E}_W\left[\mathbb{E}[Y \mid T = 1, W] - \mathbb{E}[Y \mid T = 0, W]\right] \stackrel{?}{=} \delta + \frac{\beta_u}{\alpha_u}$$

$$\text{Bias of } \mathbb{E}_W\left[\mathbb{E}[Y \mid T = 1, W] - \mathbb{E}[Y \mid T = 0, W]\right] = \delta + \frac{\beta_u}{\alpha_u} - \delta = \frac{\beta_u}{\alpha_u}$$

Sensitivity Analysis: Linear Single Confounder

(Slides adapted from Brady Neal's Introduction to Causal Inference class)

Contour Plots for Sensitivity to Confounding



Sensitivity Analysis: Linear Single Confounder

(Slides adapted from Brady Neal's Introduction to Causal Inference class)

Bias in Simple Linear Setting Proof: Step 1

Assumed SCM:

$$T := \alpha_w W + \alpha_u U$$
$$Y := \beta_w W + \beta_u U + \delta T$$

$$U = \frac{T - \alpha_w W}{\alpha_u}$$

Get a closed-form expression for $\mathbb{E}[Y \mid T = t, W]$ in terms of α_w, α_u, β_w, and β_u.

$$\begin{aligned}
\mathbb{E}[Y \mid T = t, W] &= \mathbb{E}_U [\beta_w W + \beta_u U + \delta T \mid T = t, W] \\
&= \mathbb{E}_W [\beta_w W + \beta_u \mathbb{E}[U \mid T = t, W] + \delta t] \\
&= \mathbb{E}_W \left[\beta_w W + \beta_u \left(\frac{t - \alpha_w W}{\alpha_u}\right) + \delta t\right] \\
&= \mathbb{E}_W \left[\beta_w W + \frac{\beta_u}{\alpha_u} t - \frac{\beta_u \alpha_w}{\alpha_u} W + \delta t\right] \\
&= \beta_w \mathbb{E}[W] + \frac{\beta_u}{\alpha_u} t - \frac{\beta_u \alpha_w}{\alpha_u} \mathbb{E}[W] + \delta t \\
&= \left(\delta + \frac{\beta_u}{\alpha_u}\right) t + \left(\beta_w - \frac{\beta_u \alpha_w}{\alpha_u}\right) \mathbb{E}[W]
\end{aligned}$$

Sensitivity Analysis: Linear Single Confounder

(Slides adapted from Brady Neal's Introduction to Causal Inference class)

Bias in Simple Linear Setting Proof: Step 2

Step 1: $$\mathbb{E}_W\left[\mathbb{E}[Y \mid T=t, W]\right] = \left(\delta + \frac{\beta_u}{\alpha_u}\right) t + \left(\beta_w - \frac{\beta_u \alpha_w}{\alpha_u}\right) \mathbb{E}[W]$$

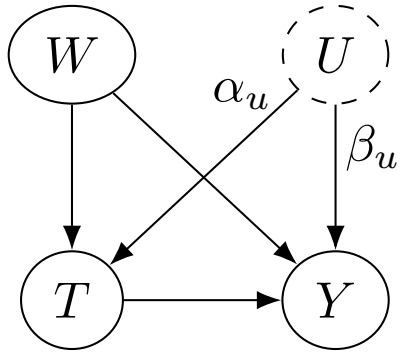
$$\begin{aligned}\mathbb{E}_W\left[\mathbb{E}[Y \mid T=1, W] - \mathbb{E}[Y \mid T=0, W]\right] &= \left(\delta + \frac{\beta_u}{\alpha_u}\right)(1) + \left(\beta_w - \frac{\beta_u \alpha_w}{\alpha_u}\right) \mathbb{E}[W] \\ &\quad - \left[\left(\delta + \frac{\beta_u}{\alpha_u}\right)(0) + \left(\beta_w - \frac{\beta_u \alpha_w}{\alpha_u}\right) \mathbb{E}[W]\right] \\ &= \delta + \frac{\beta_u}{\alpha_u}\end{aligned}$$

Sensitivity Analysis: Linear Single Confounder

(Slides adapted from Brady Neal's Introduction to Causal Inference class)

Bias in Simple Linear Setting Proof: Step 3

$$
\begin{aligned}
\text{Bias} &= \mathbb{E}_W \left[\mathbb{E}[Y \mid T=1, W] - \mathbb{E}[Y \mid T=0, W]\right] \\
&\quad - \mathbb{E}_{W,U} \left[\mathbb{E}[Y \mid T=1, W, U] - \mathbb{E}[Y \mid T=0, W, U]\right] \\
&= \delta + \frac{\beta_u}{\alpha_u} - \delta \\
&= \frac{\beta_u}{\alpha_u}
\end{aligned}
$$

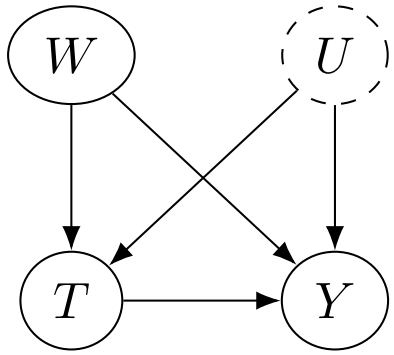


$$T := \alpha_w W + \alpha_u U$$
$$Y := \beta_w W + \beta_u U + \delta T$$

(Slides adapted from Brady Neal's Introduction to Causal Inference class)

Sensitivity analysis with binary treatment

$$T := \alpha_w W + \alpha_u U$$
$$Y := \beta_w W + \beta_u U + \delta T$$



$$P(T = 1 \mid W, U) := \text{sigmoid}\left(\alpha_w W + \alpha_u U\right)$$
$$Y := \beta_w W + \beta_u U + \delta T + N$$
where $\text{sigmoid}(x) = \frac{1}{1 + e^{-x}}$

Rosenbaum & Rubin (1983) and Imbens (2003)

- Simple parametric form for T
- Simple parametric form for Y
- U is binary
- U is a scalar (only one unobserved confounder)

(Slides adapted from Brady Neal's Introduction to Causal Inference class)

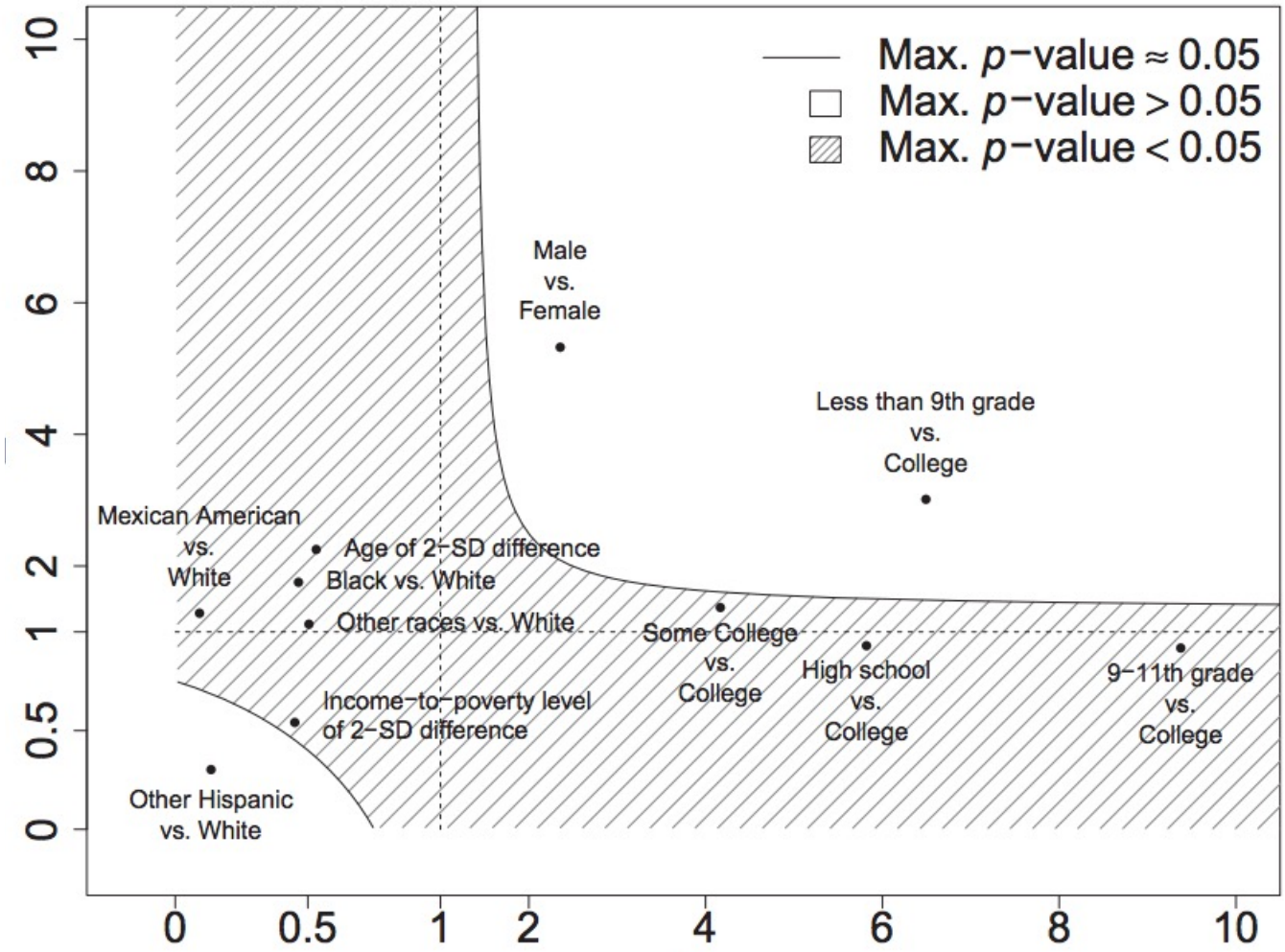
Sensitivity analysis with binary treatment

- How much unmeasured confounding to flip our conclusions?

(Slides adapted from Uri Shalit's causal inference class)

Does cigarette smoking increase blood lead?

Unmeasured confounding $\exp(\delta)$ in outcome model $U \rightarrow Y$



Unmeasured confounding $\exp(\gamma)$ in treatment assignment $U \rightarrow T$

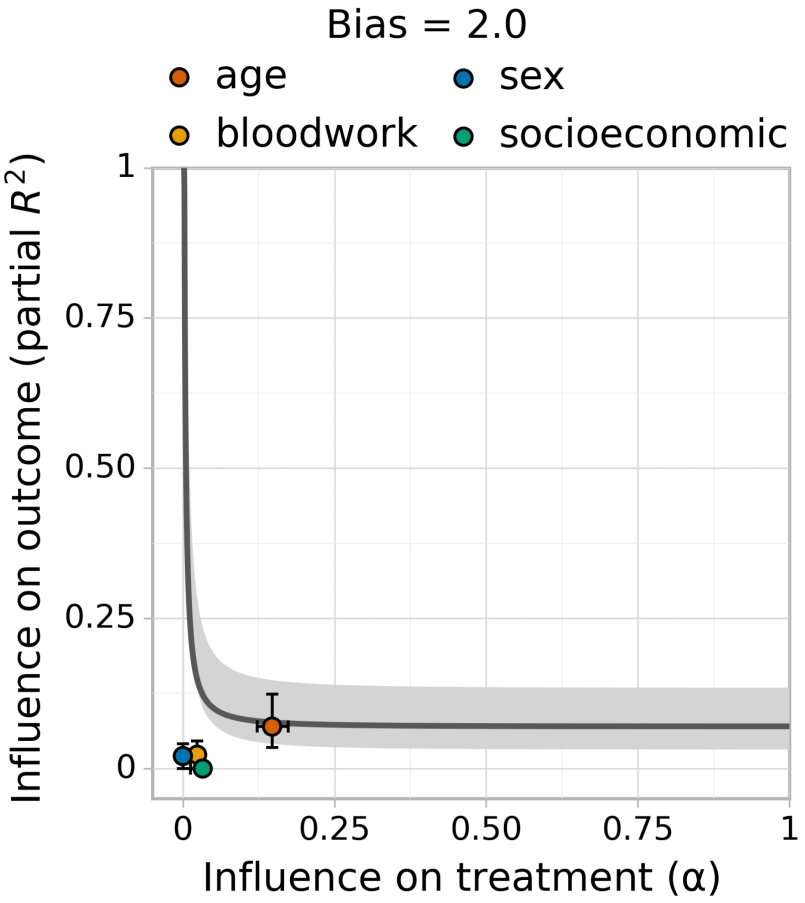
Hsu & Small, 2013

(Slides adapted from Uri Shalit's causal inference class)

Generalization: Austen plots

- Here, both treatment mechanism and the outcome mechanism can be modeled with **arbitrary machine learning models**

- Assumptions on how hidden confounders modify treatment & outcome models



(Veitch & Zaveri, Sense and Sensitivity Analysis: Simple Post-Hoc Analysis of Bias Due to Unobserved Confounding. NeurIPS 2020)

Summary

- Close connection between causal inference and off-policy evaluation
  - Will return to this later when we talk about off-policy *reinforcement learning*
- Instrumental variables can be used to estimate ATE and CATE when there is unobserved confounding
- Sensitivity analysis can help build intuition for how unobserved confounding affects bias

References

- Introduction to causal inference from a machine learning perspective by Brady Neal, 2020.
  - Section 8.2: Sensitivity Analysis
  - Chapter 9: Instrumental Variables

  (See also the many references within for both recent literature and where these methods were originally introduced.)
- Syrgkanis et al., Machine Learning Estimation of Heterogeneous Treatment Effects with Instruments, NeurIPS 2019.
- Boominathan et al., Treatment Policy Learning in Multiobjective Settings with Fully Observed Outcomes, KDD 2020.