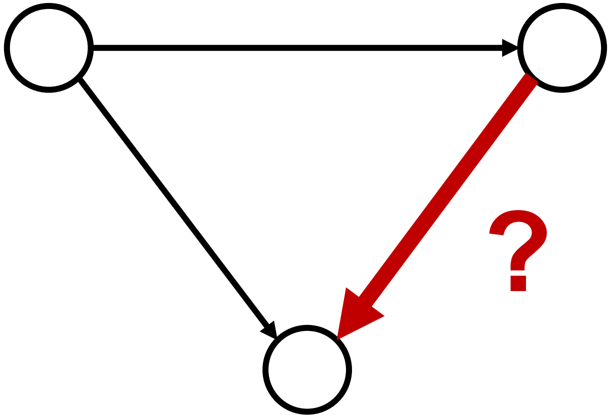
Machine Learning for Healthcare
6.871, HST.956

Lecture 13: Causality (Part 2)

Prof. Manolis Kellis

Slides Credit:
Prof. David Sontag

Reminder: Causal inference



Patient, X

(including all confounding factors)

Intervention, T

(e.g. medication, procedure)

?

Outcome, Y

High dimensional

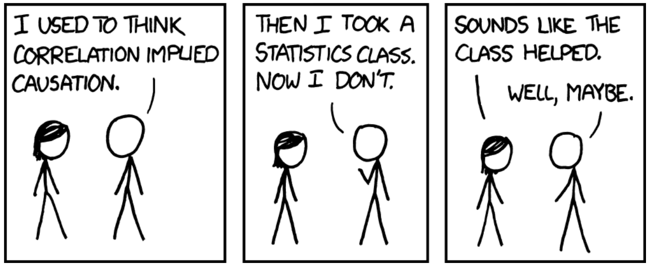
Observational data

Reminder: Potential Outcomes

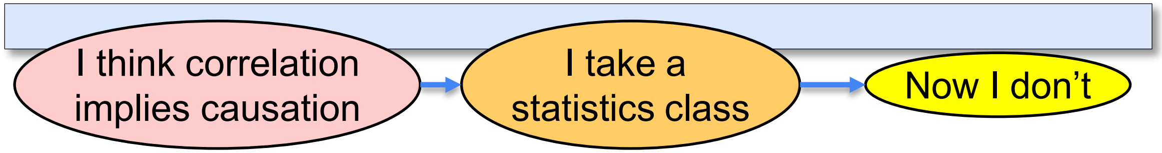
- Each unit (individual) x_i has two potential outcomes:
  - $Y_0(x_i)$ is the potential outcome had the unit not been treated: "***control outcome***"
  - $Y_1(x_i)$ is the potential outcome had the unit been treated: "***treated outcome***"

- Conditional average treatment effect for unit i:
  $CATE(x_i) = \mathbb{E}_{Y_1 \sim p(Y_1|x_i)}[Y_1|x_i] - \mathbb{E}_{Y_0 \sim p(Y_0|x_i)}[Y_0|x_i]$
- Average Treatment Effect:

$$ATE = \mathbb{E}_{x \sim p(x)}[CATE(x)]$$


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

XKCD

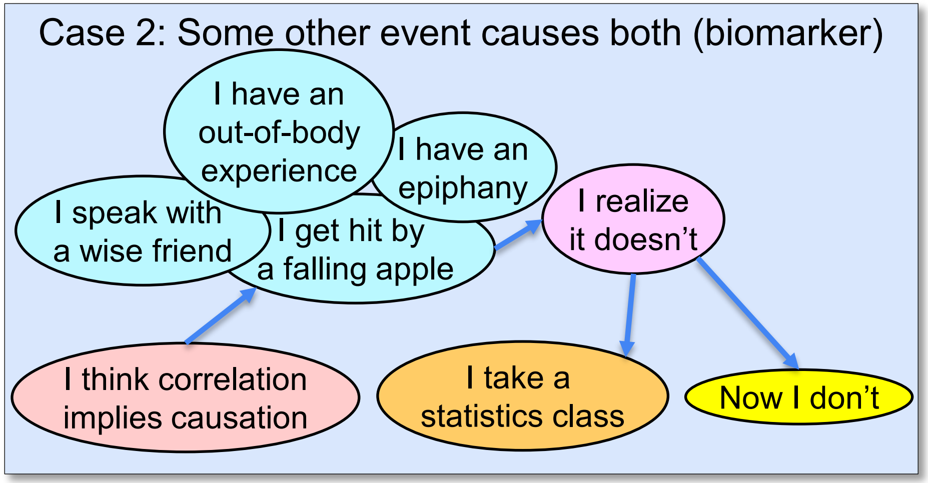


I think correlation implies causation

I take a statistics class

Now I don't

Case 1: Causation



Case 2: Some other event causes both (biomarker)

I have an out-of-body experience

I have an epiphany

I speak with a wise friend

I get hit by a falling apple

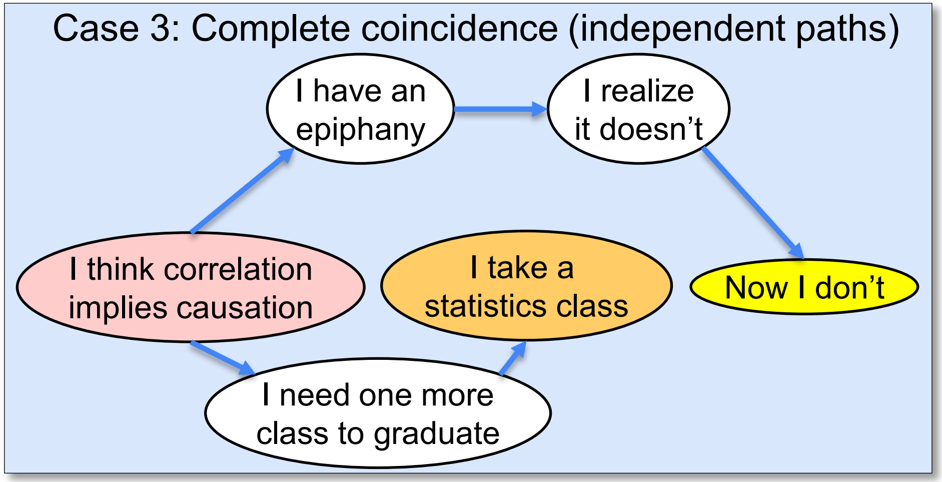
I realize it doesn't

I think correlation implies causation

I take a statistics class

Now I don't

Intuition: not everyone takes a statistics class
Perhaps something pushed me to take one.
Perhaps that same something led to the outcome



Case 3: Complete coincidence (independent paths)

I have an epiphany

I realize it doesn't

I think correlation implies causation

I take a statistics class

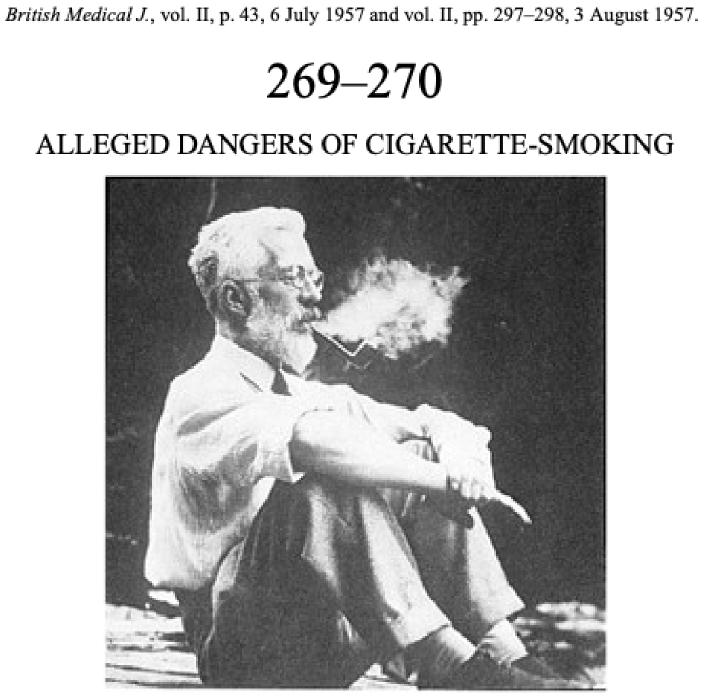
Now I don't

I need one more class to graduate

Intuition: Sometimes even the correlation is fortuitous
(solution: increase sample size ➔ correlation goes away)

Does smoking cause lung cancer?

- Think about confounding factors that we would need to collect as part of the dataset
- RA Fisher - famous statistician, rejected smoking ➔ cancer causality
- Claim: Only associational studies have been run so far.
- Monozygotic twins have more similar smoking patterns than dizygotic twins, so maybe a genetic propensity to smoke instead of a causal link?
- How many cancers were caused by this wrong interpretation?



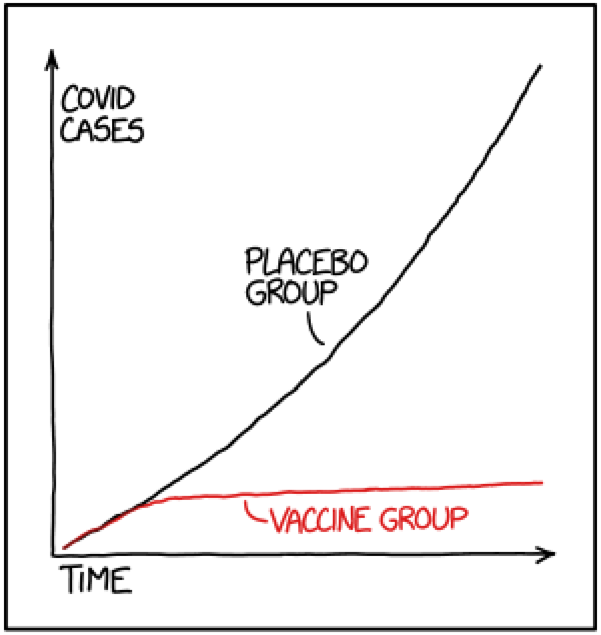
British Medical J., vol. II, p. 43, 6 July 1957 and vol. II, pp. 297–298, 3 August 1957.

269–270

ALLEGED DANGERS OF CIGARETTE-SMOKING

"Alleged benefits of covid vaccination"

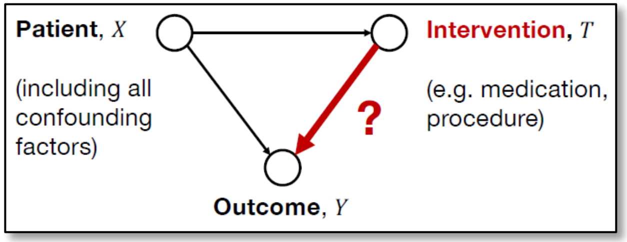
Statistics



STATISTICS TIP: ALWAYS TRY TO GET DATA THAT'S GOOD ENOUGH THAT YOU DON'T NEED TO DO STATISTICS ON IT

We reject the null hypothesis based on the 'hot damn, check out this chart' test

xkcd



Patient, X

(including all confounding factors)

Intervention, T

(e.g. medication, procedure)

Outcome, Y

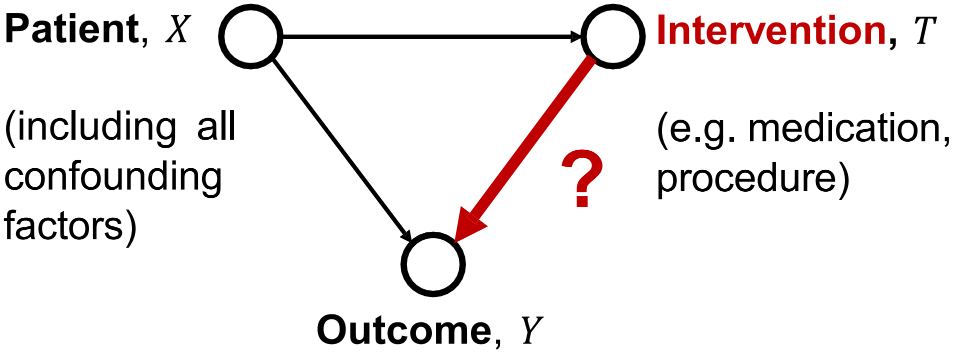
Alleged benefits of mask-wearing to protect against covid spread:

- Yes, there is plausibility
- Yes, there is correlation
- Yes, there are interventional studies

But many confounders:

- Counties who choose to mask also choose other measures
- Individuals who choose to mask also take other precautions
- Can we untangle these effects?

To properly answer, need to formulate as *causal* questions:



Patient, X

(including all confounding factors)

Intervention, T

(e.g. medication, procedure)

?

Outcome, Y

High dimensional

Observational data

ATE = Average Treatment Effect
CATE = Conditional Average Treatment Effect

Each unit (individual) x_i has two potential outcomes*:

- $Y_0(x_i)$ is the potential outcome had the unit not been treated: "***control outcome***"
- $Y_1(x_i)$ is the potential outcome had the unit been treated: "***treated outcome***"

Conditional average treatment effect for unit i:

$$CATE(x_i) = \mathbb{E}_{Y_1 \sim p(Y_1|x_i)}[Y_1|x_i] - \mathbb{E}_{Y_0 \sim p(Y_0|x_i)}[Y_0|x_i]$$

Average Treatment Effect:

$$ATE := \mathbb{E}[Y_1 - Y_0] = \mathbb{E}_{x \sim p(x)}[CATE(x)]$$

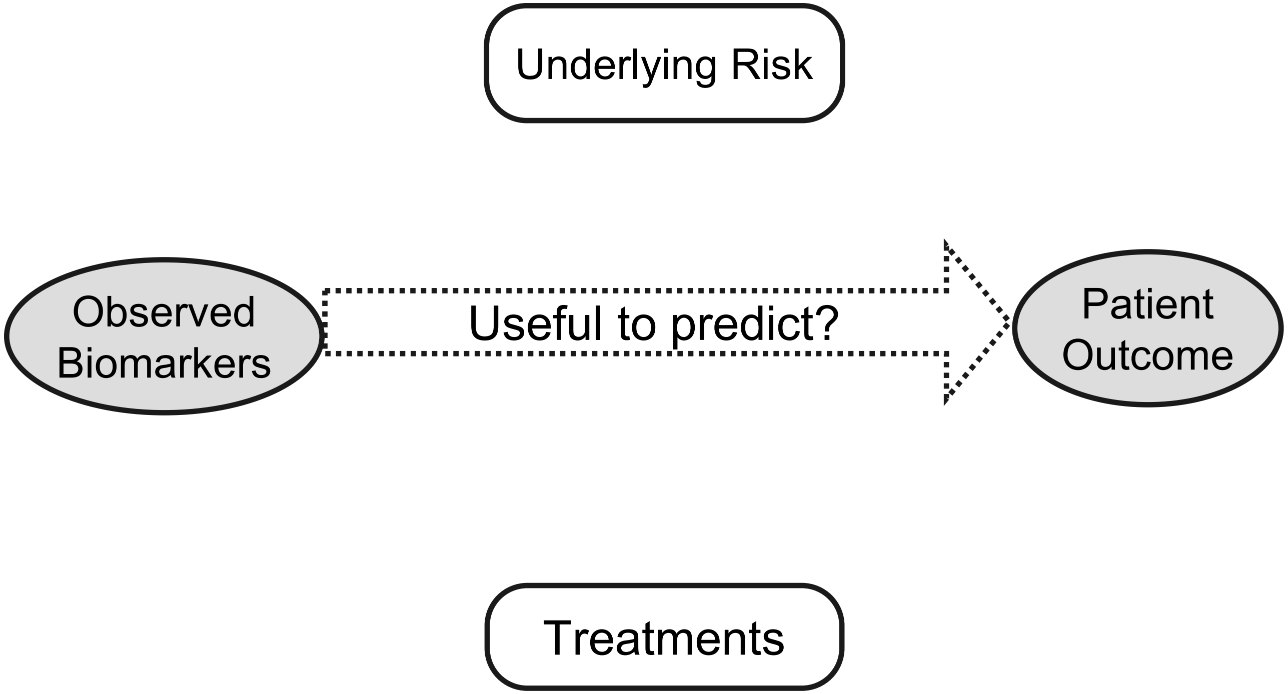
Observed factual outcome:

$$y_i = t_i Y_1(x_i) + (1 - t_i) Y_0(x_i)$$

Unobserved counterfactual outcome:

$$y_i^{CF} = (1 - t_i) Y_1(x_i) + t_i Y_0(x_i)$$

Real-world evidence comes from complex human behaviors



Underlying Risk

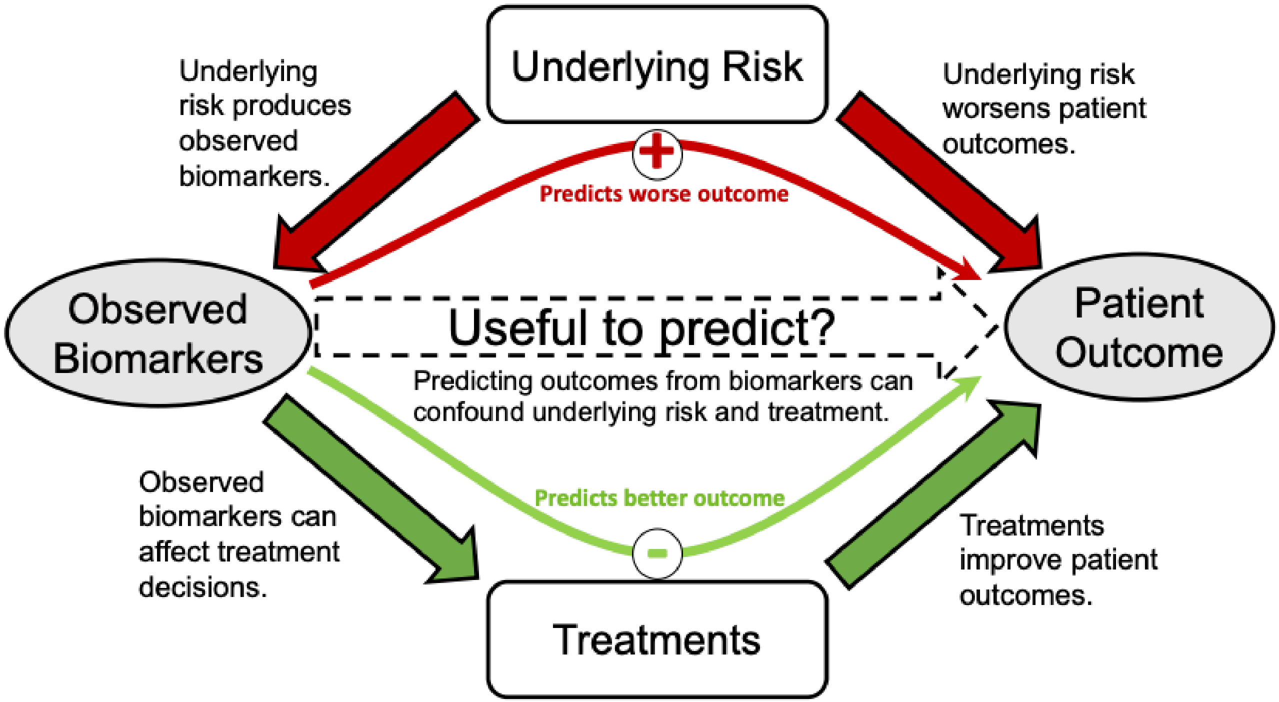
Observed Biomarkers

Useful to predict?

Patient Outcome

Treatments

Real-world evidence comes from complex human behaviors


Underlying Risk
Underlying risk produces observed biomarkers.
Underlying risk worsens patient outcomes.
+
Predicts worse outcome
Observed Biomarkers
Useful to predict?
Patient Outcome
Predicting outcomes from biomarkers can confound underlying risk and treatment.
Observed biomarkers can affect treatment decisions.
Predicts better outcome
-
Treatments improve patient outcomes.
Treatments

Two approaches for causality inference using counterfactual analysis

Covariate adjustment and matching

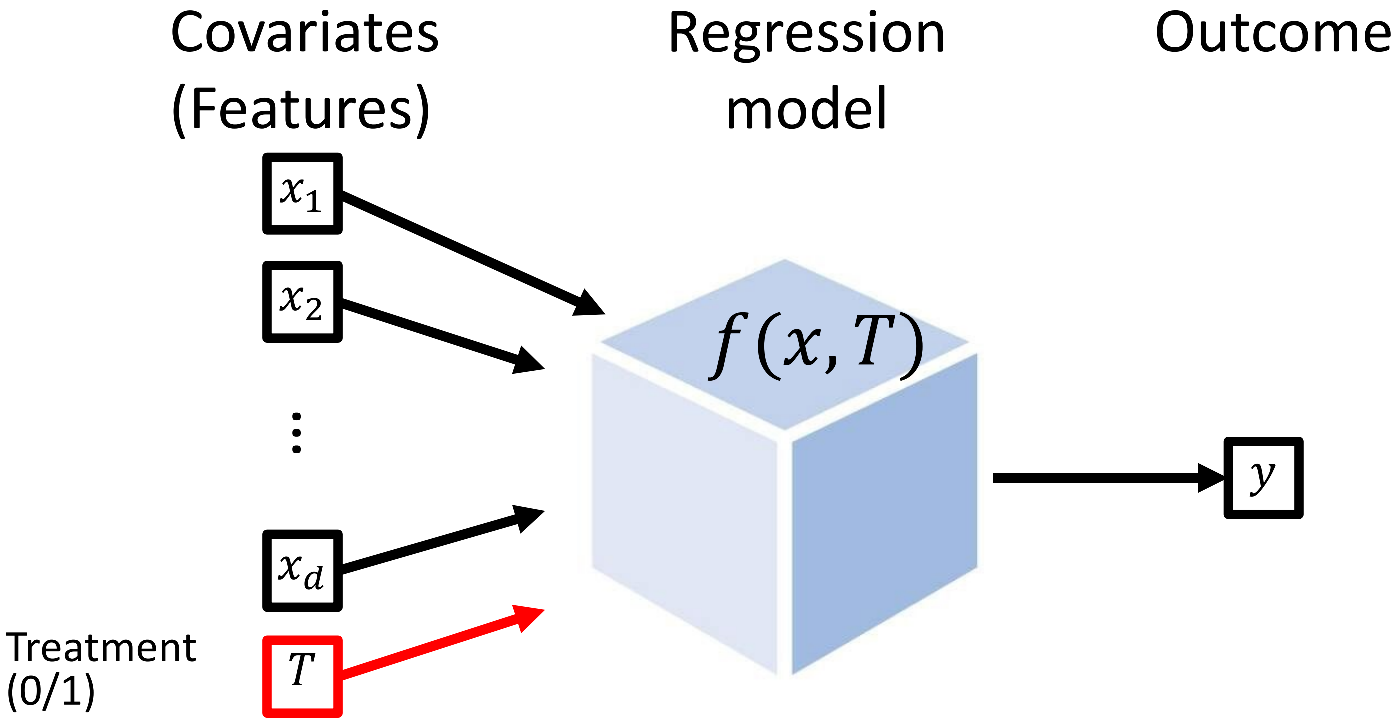
*Predict outcome given features and treatment,
then use resulting model to impute counterfactuals*

Propensity score re-weighing

*Predict treatment using features (propensity score),
then use to reweight outcome or stratify the data*

Covariate adjustment (reminder)

Explicitly model the relationship between treatment, confounders, and outcome:



Covariates (Features)

Regression model

Outcome

x_1

x_2

$\vdots$

x_d

Treatment (0/1)

T

$f(x, T)$

y

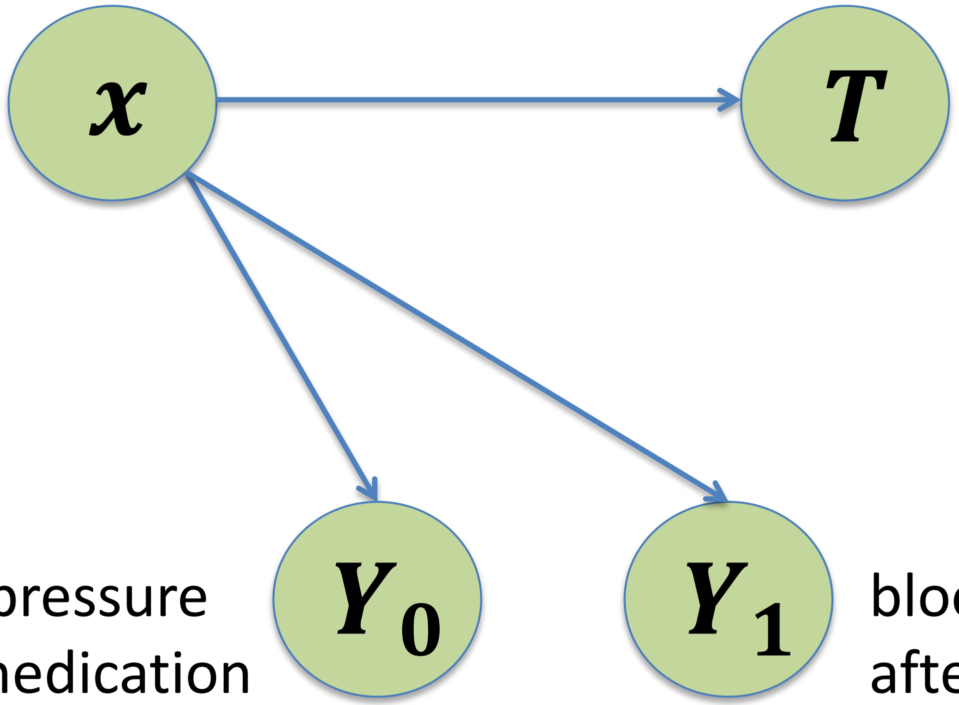
Covariate adjustment (reminder)

- Under ignorability, can use the adjustment formula:

  $ATE(x)$ =
  $\mathbb{E}_{x\sim p(x)}\big[\ \mathbb{E}[Y_1|T=1,x] - \mathbb{E}[Y_0|T=0,x]\big]$

- Fit a model $f(x,t) \approx \mathbb{E}[Y_t|T=t,x]$, then:
  $\widehat{CATE}(x) = f(x,1) - f(x,0)$.

Ignorability (no hidden confounding)



anti-hypertensive medication

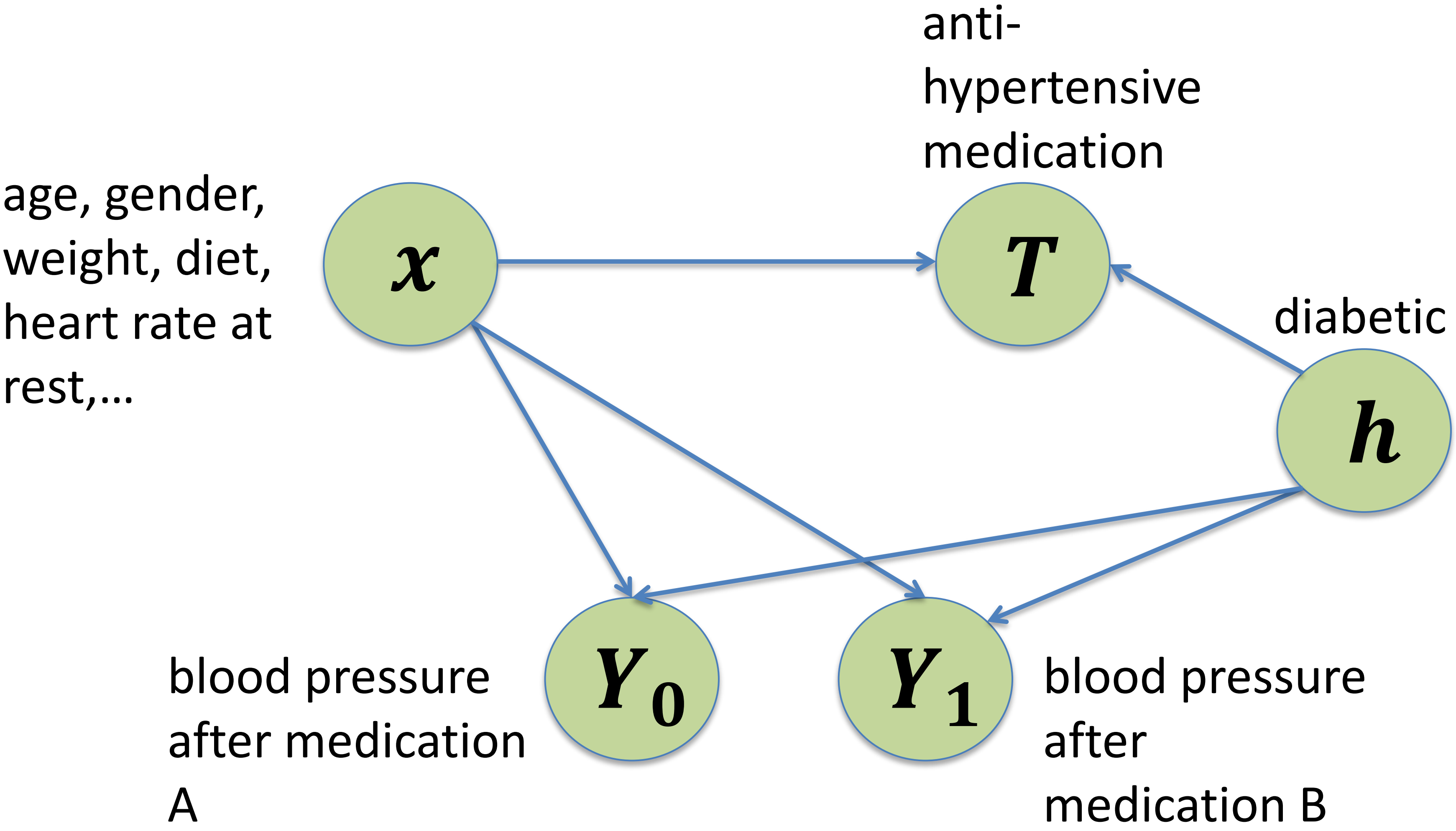
age, gender, weight, diet, heart rate at rest,…

blood pressure after medication A

blood pressure after medication B

$$(Y_0, Y_1) \perp\!\!\!\perp T \mid x$$

No Ignorability



anti-hypertensive medication

age, gender, weight, diet, heart rate at rest,…

diabetic

blood pressure after medication A

blood pressure after medication B

$$(Y_0, Y_1) \not\!\perp\!\!\!\perp T \mid x$$

Covariate adjustment with linear models

- Assume that:

Blood pressure **age** **medication**

$$Y_t(x) = \beta x + \gamma \cdot t + \epsilon_t$$
$$\mathbb{E}[\epsilon_t] = 0$$

- Then:

$$CATE(x) := \mathbb{E}[Y_1(x) - Y_0(x)] =$$

Covariate adjustment with linear models

- Assume that:

Blood pressure **age** **medication**

$$Y_t(x) = \beta x + \gamma \cdot t + \epsilon_t$$

$$\mathbb{E}[\epsilon_t] = 0$$

- Then:

$$CATE(x) := \mathbb{E}[Y_1(x) - Y_0(x)] =$$

$$\mathbb{E}[(\cancel{\beta x} + \gamma + \epsilon_1) - (\cancel{\beta x} + \epsilon_0)] = \gamma$$

$$\boxed{ATE := \mathbb{E}_{p(x)}[CATE(x)] = \gamma}$$

Covariate adjustment with linear models

- Assume that:

$$\overset{\text{Blood pressure}}{Y_t(x)} = \beta \overset{\text{age}}{x} + \gamma \cdot \overset{\text{medication}}{t} + \epsilon_t$$

$$\mathbb{E}[\epsilon_t] = 0$$

$$\boxed{ATE := \mathbb{E}_{p(x)}[CATE(x)] = \gamma}$$

- For causal inference, need to estimate γ well, not $Y_t(x)$ - **Identification, not prediction**
- *Major difference between ML and statistics*

What happens when there is misspecification?

- True data generating process, $x \in \mathbb{R}$:

$$Y_t(x) = \beta x + \gamma \cdot t + \delta \cdot x^2$$

$$ATE = \mathbb{E}[Y_1 - Y_0] = \gamma$$

- Hypothesized model:

$$\widehat{Y}_t(x) = \hat{\beta} x + \hat{\gamma} \cdot t$$

$$\hat{\gamma} = \gamma + \delta \frac{\mathbb{E}[xt]\mathbb{E}[x^2] - \mathbb{E}[t^2]\mathbb{E}[x^2 t]}{\mathbb{E}[xt]^2 - \mathbb{E}[x^2]\mathbb{E}[t^2]}$$

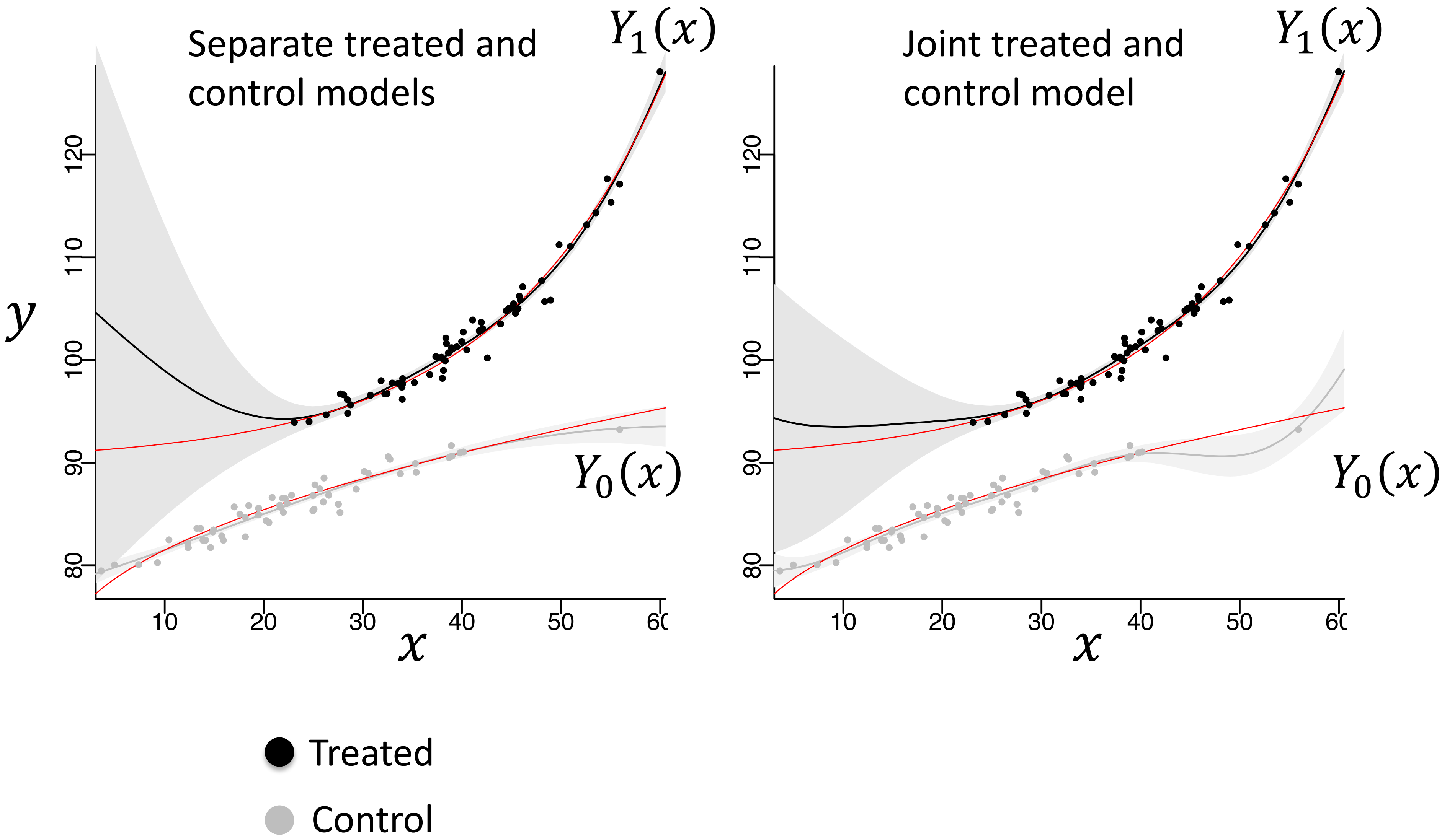
Depending on δ, can be made to be arbitrarily large or small!

Covariate adjustment with non-linear models

- Random forests and Bayesian trees
  Hill (2011), Athey & Imbens (2015), Wager & Athey (2015)
- Gaussian processes
  Hoyer et al. (2009), Zigler et al. (2012), Alaa & van der Schaar (2017)
- Neural networks
  Beck et al. (2000), Johansson et al. (2016), Shalit et al. (2016), Lopez-Paz et al. (2016)

Called *nonparametric* estimators, since they do not make assumptions about form of $\mathbb{E}[Y|X,T]$ and, given enough data, could fit any function

Example: Gaussian processes

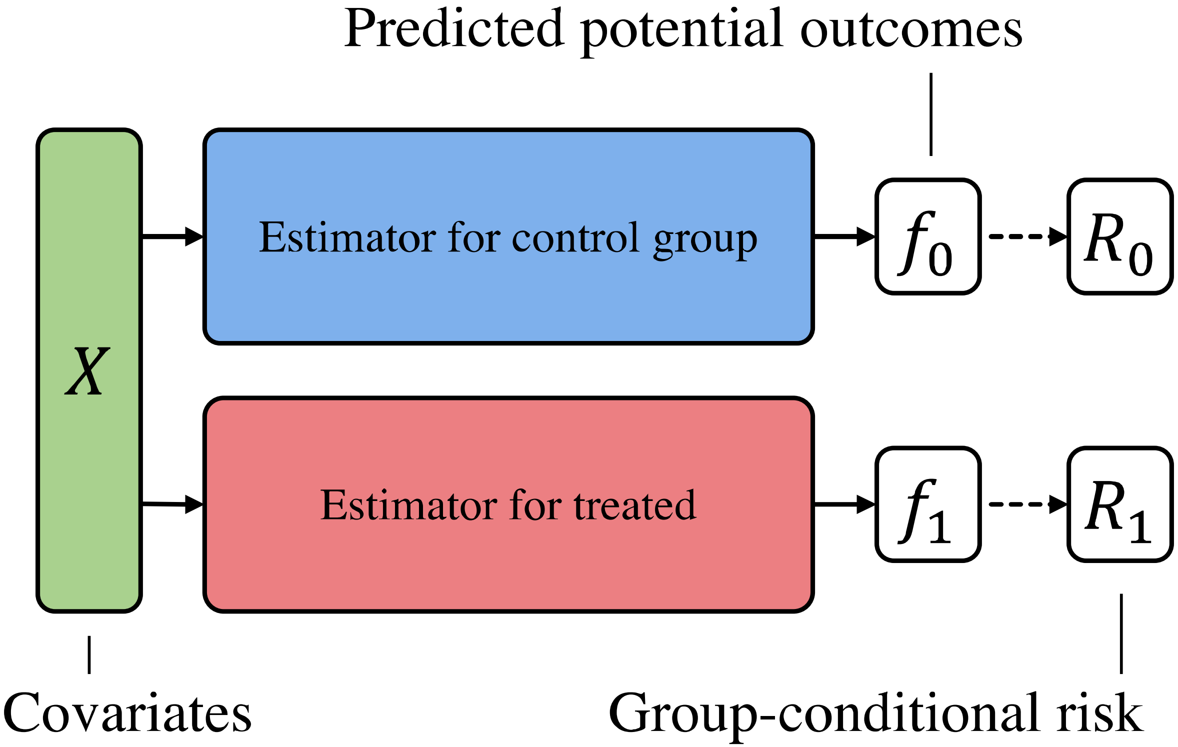


Separate treated and control models

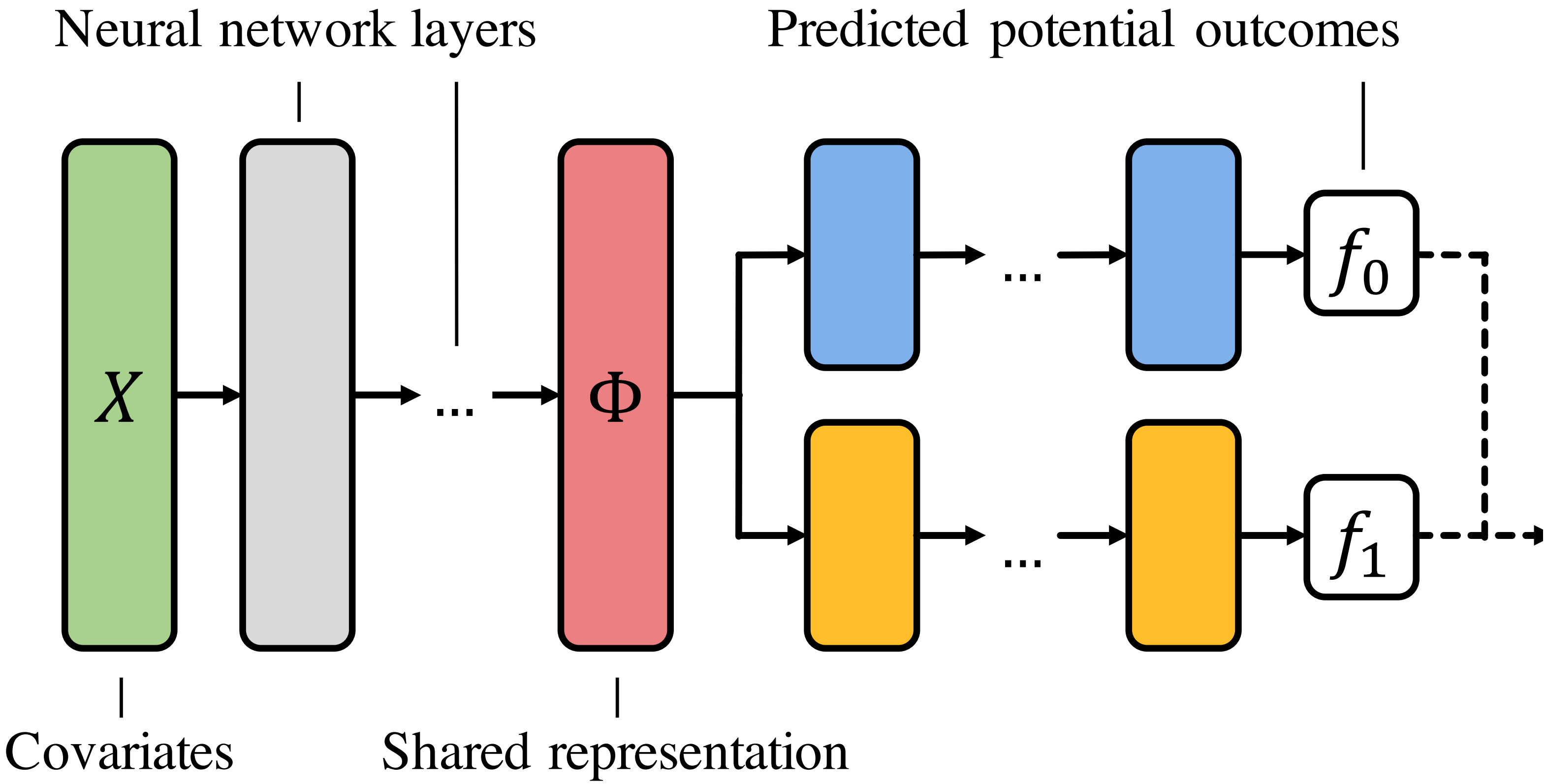
Joint treated and control model

Figures: Vincent Dorie & Jennifer Hill

Example: Neural networks



Example: Neural networks



Shalit, Johansson, Sontag. *Estimating Individual Treatment Effect: Generalization Bounds and Algorithms*. ICML, 2017

Necessary assumption for nonparametric estimation – common support

Y_0, Y_1: potential outcomes for control and treated

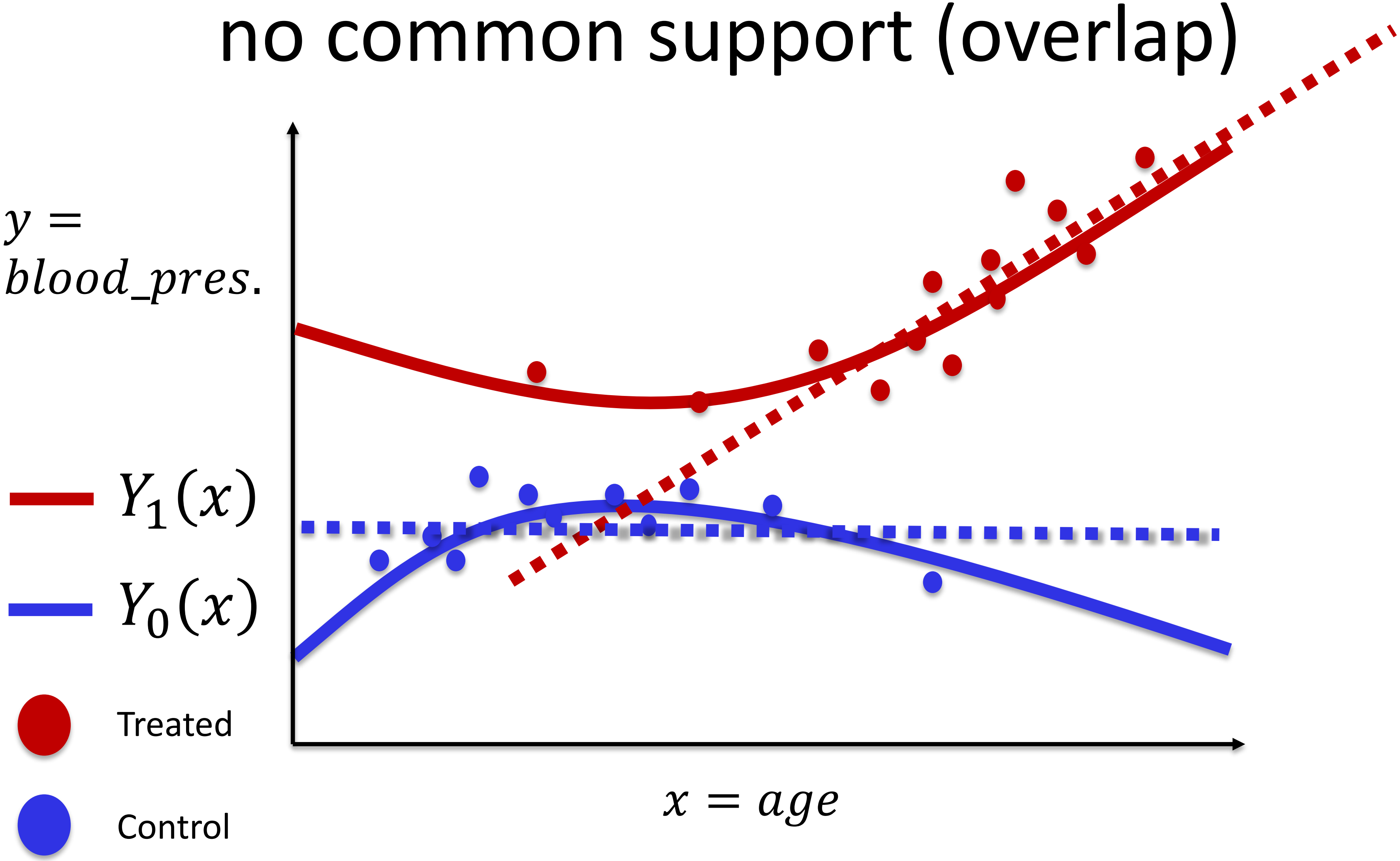
x: unit covariates (features)

T: treatment assignment

We assume:

$$p(T = t|X = x) > 0 \ \forall t, x$$

Example of how (nonparametric) covariate adjustment fails when there is no common support (overlap)



Matching

- Find each unit’s long-lost counterfactual identical twin, check up on his outcome

Matching

- Find each person's long-lost counterfactual identical twin, check up on his outcome

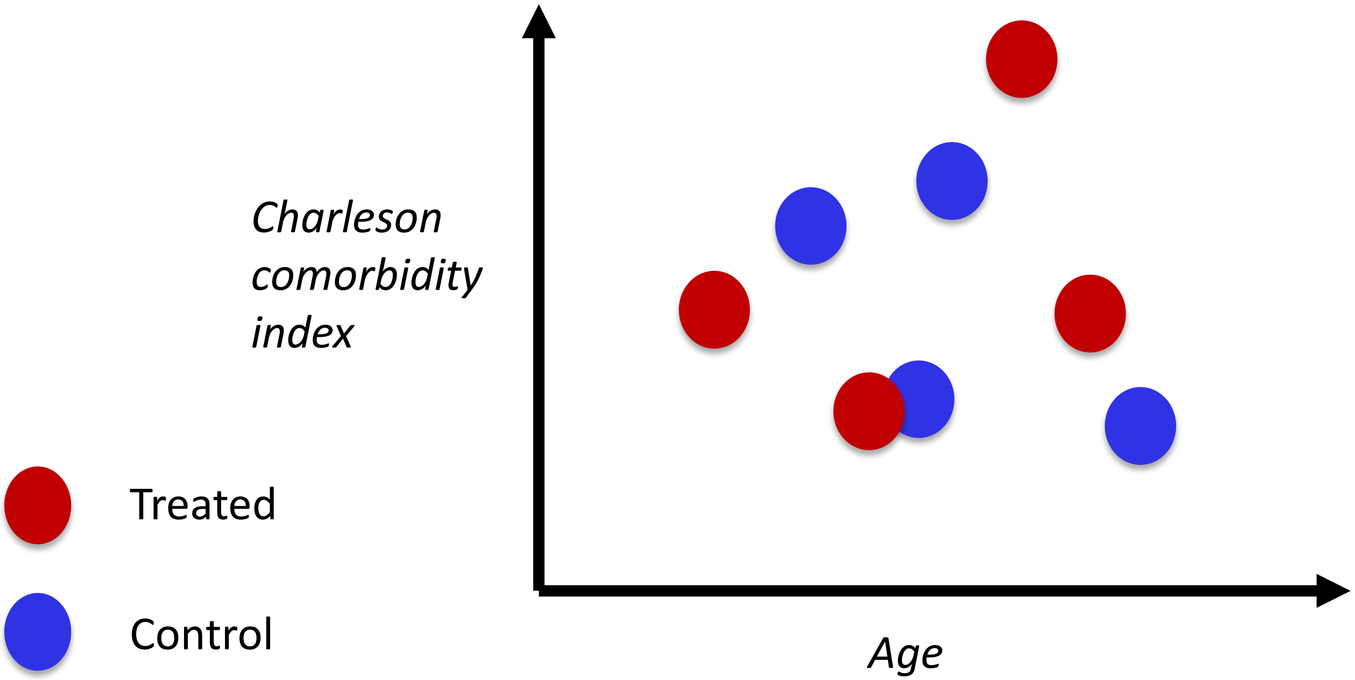
Obama, had he gone to law school

Obama, had he gone to business school

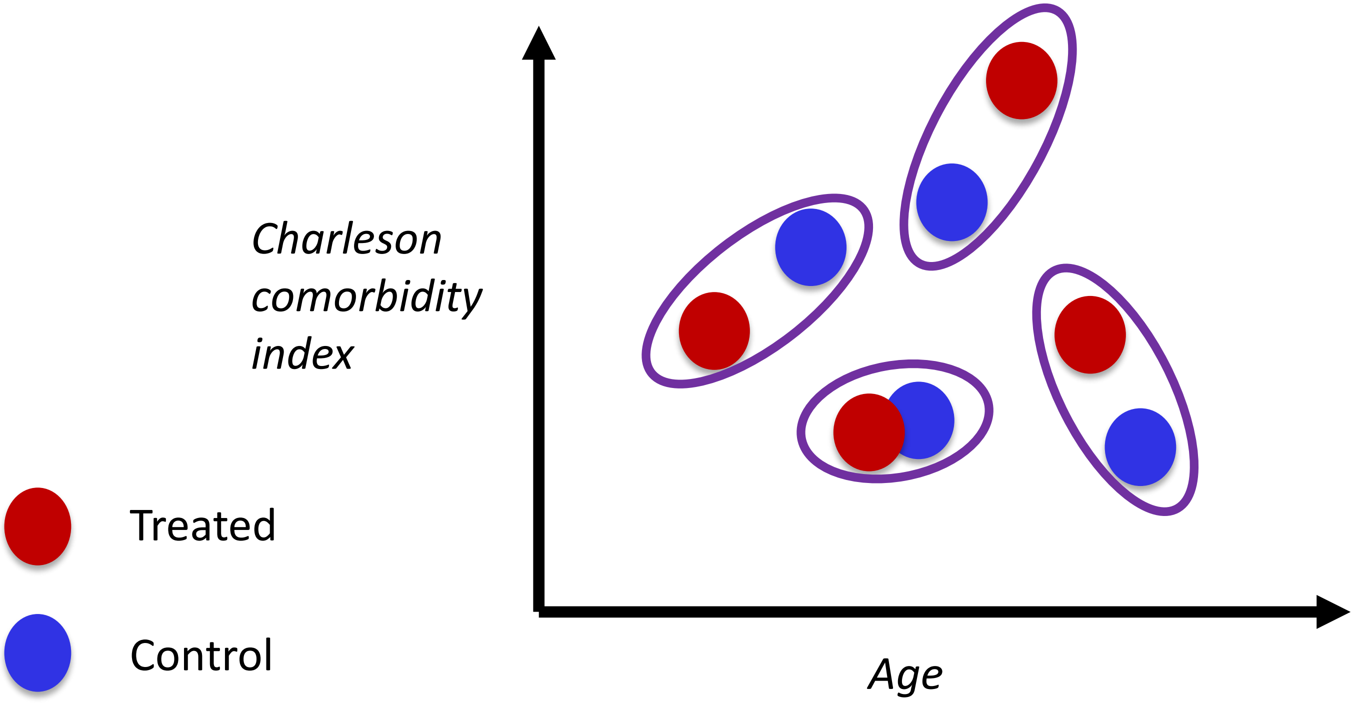
Matching

- Find each person’s long-lost counterfactual identical twin, check up on his outcome
- Used for estimating both ATE and CATE

Match to nearest neighbor from opposite group



Match to nearest neighbor from opposite group



1-NN Matching

- Let $d(\cdot,\cdot)$ be a metric between x's
- For each i, define $j(i) = \underset{j\ s.t.\ t_j \neq t_i}{\operatorname{argmin}}\ d(x_j, x_i)$

  $j(i)$ is the nearest counterfactual neighbor of i
- $t_i = 1$, unit i is treated:

$$\widehat{CATE}(x_i) = y_i - y_{j(i)}$$

- $t_i = 0$, unit i is control:

$$\widehat{CATE}(x_i) = y_{j(i)} - y_i$$

1-NN Matching

- Let $d(\cdot,\cdot)$ be a metric between x's
- For each i, define $j(i) = \underset{j\ s.t.\ t_j \neq t_i}{\operatorname{argmin}}\ d(x_j, x_i)$

  $j(i)$ is the nearest counterfactual neighbor of i

- $\widehat{CATE}(x_i) = (2t_i - 1)(y_i - y_{j(i)})$
- $\widehat{ATE} = \frac{1}{n} \sum_{i=1}^{n} \widehat{CATE}(x_i)$

Matching

- Interpretable, especially in small-sample regime
- Nonparametric
- Heavily reliant on the underlying metric
- Could be misled by features which don’t affect the outcome

Covariate adjustment and matching

- Matching is equivalent to covariate adjustment with two 1-nearest neighbor classifiers:
  $\hat{Y}_1(x) = y_{NN_1(x)}$, $\hat{Y}_0(x) = y_{NN_0(x)}$
  where $y_{NN_t(x)}$ is the nearest-neighbor of x among units with treatment assignment
  $$t = 0,1$$

- 1-NN matching is in general inconsistent, though only with small bias (Imbens 2004)

Two approaches for causality inference using counterfactual analysis

Covariate adjustment and matching

Predict outcome given features and treatment, then use resulting model to impute counterfactuals

Propensity score re-weighing

Predict treatment using features (propensity score), then use to reweight outcome or stratify the data

Propensity scores

- Tool for estimating ATE
- Imagine that we had data from a randomized control trial (RCT). Then we could simply estimate the ATE using:

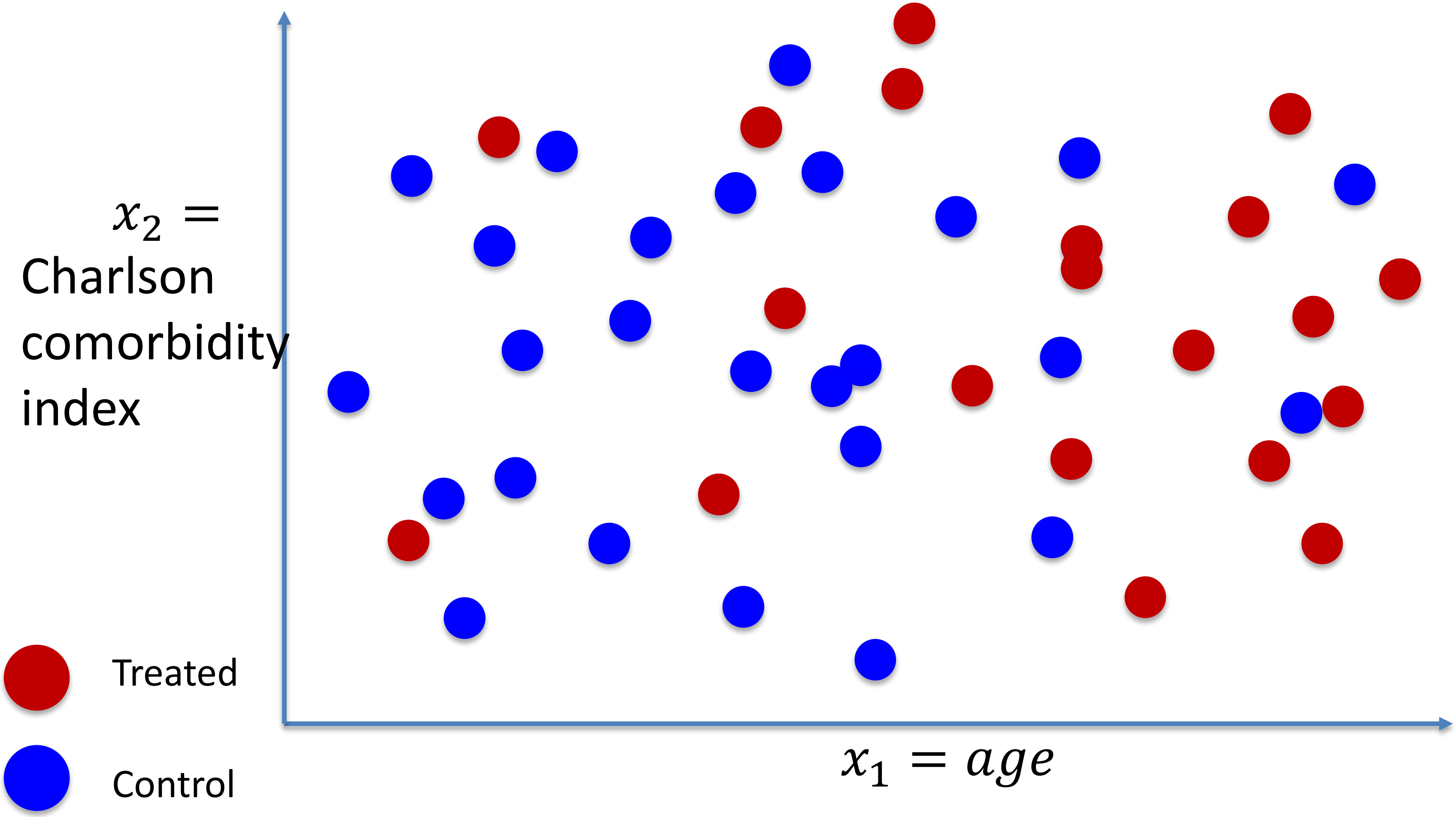
$$\frac{1}{n_1}\sum_{i\ s.t. T_i=1} Y_i - \frac{1}{n_0}\sum_{i\ s.t. T_i=0} Y_i$$

- Basic idea: turn observational study into a pseudo-randomized trial by re-weighting samples

Inverse propensity score re-weighting

$$p(x|t=0) \neq p(x|t=1)$$

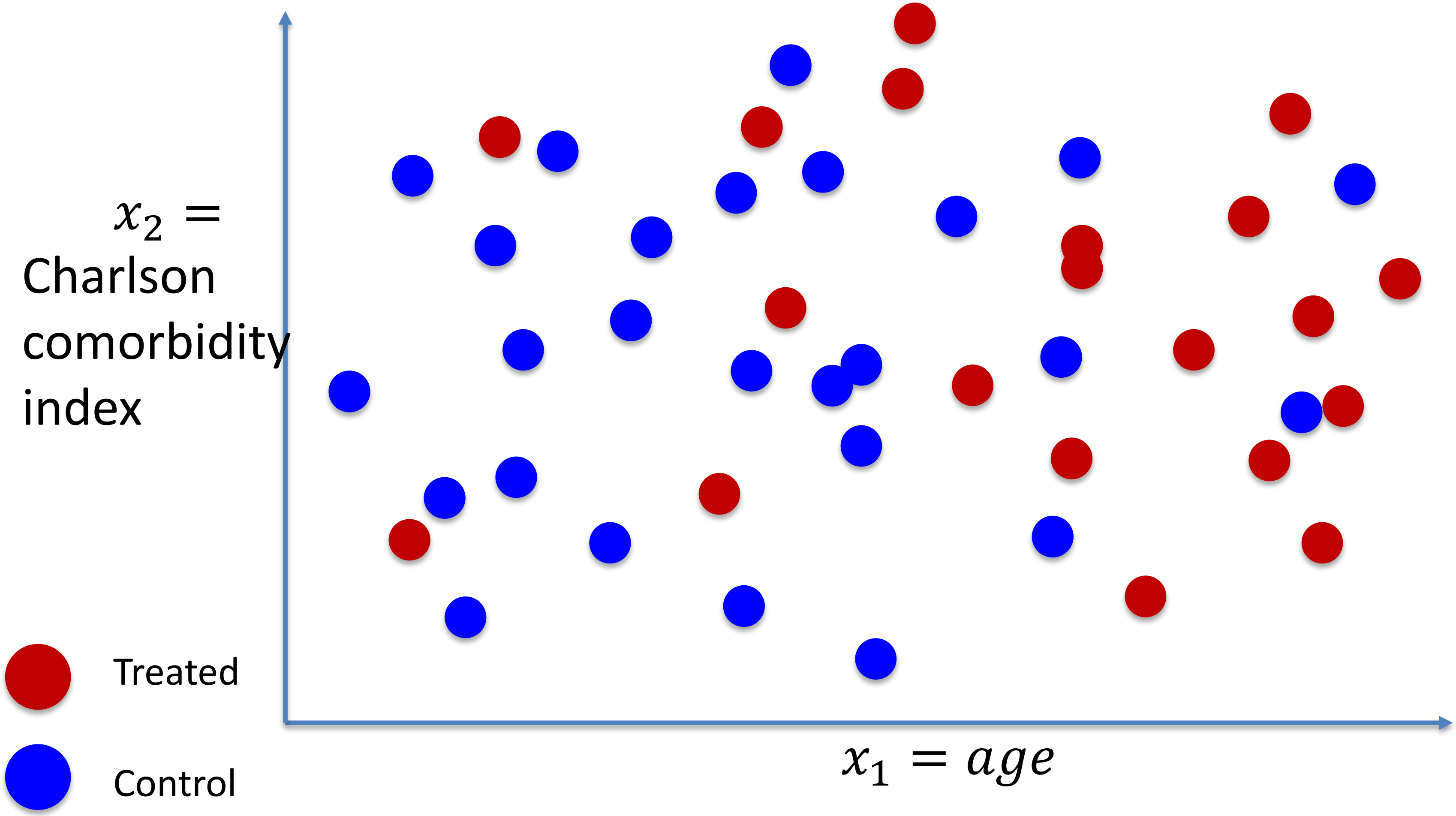
control *treated*



Inverse propensity score re-weighting

$$p(x|t=0)\cdot w_0(x) \approx p(x|t=1)\cdot w_1(x)$$

reweighted control *reweighted treated*

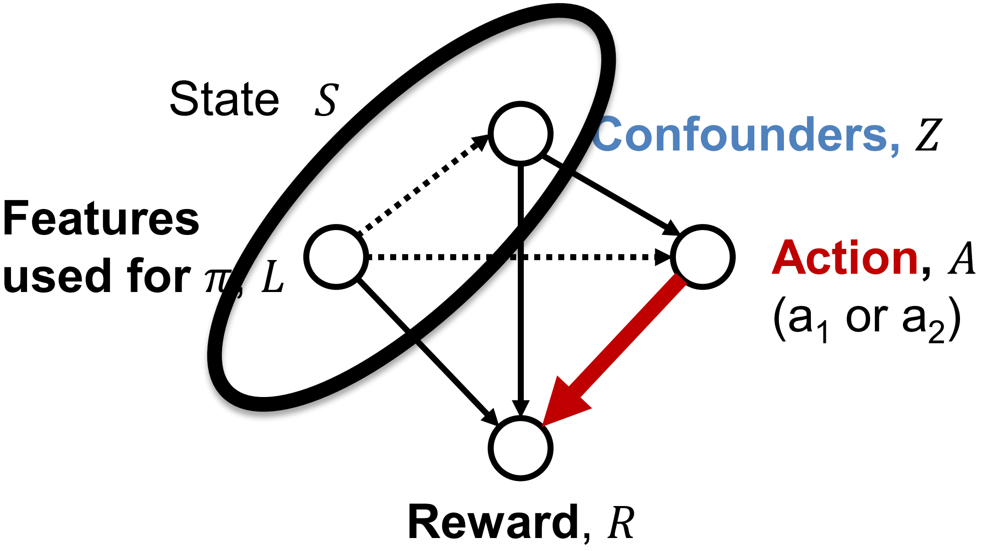


Propensity score

- Propensity score: $p(T = 1|x)$,
  using machine learning tools, e.g. logistic regression
- Samples re-weighted by the inverse propensity score of the treatment they received

Same ideas can be used for off-policy evaluation

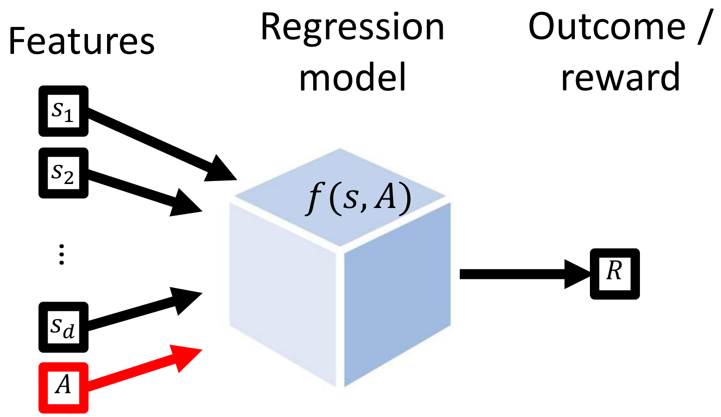
- Suppose someone gave us a policy $\pi(l)$ that outputs a_1 vs a_2
- How do we evaluate it?
- We give two approaches, one based on potential outcomes and the other based on propensity scores
- In both cases, we have to first consider the causal graph that underlies the *observational data*



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, SUTVA)

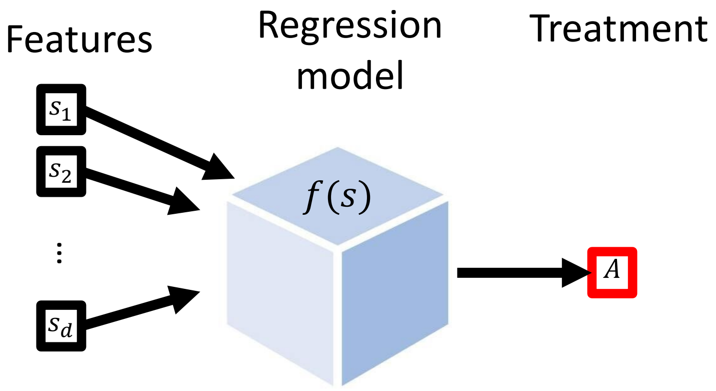


- Then, use this model to impute policy outcomes:

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

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

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

$$o_i = \arg\max_A f(l_i, z_i, A)$$

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)

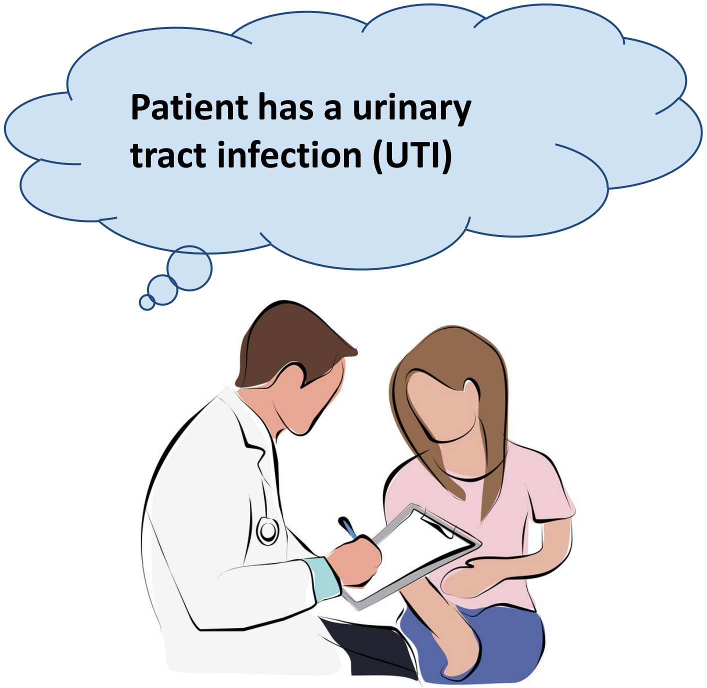
Reinforcement Learning for policy evaluation

Using observational data

Evaluate *policies* using observational data with Reinforcement Learning

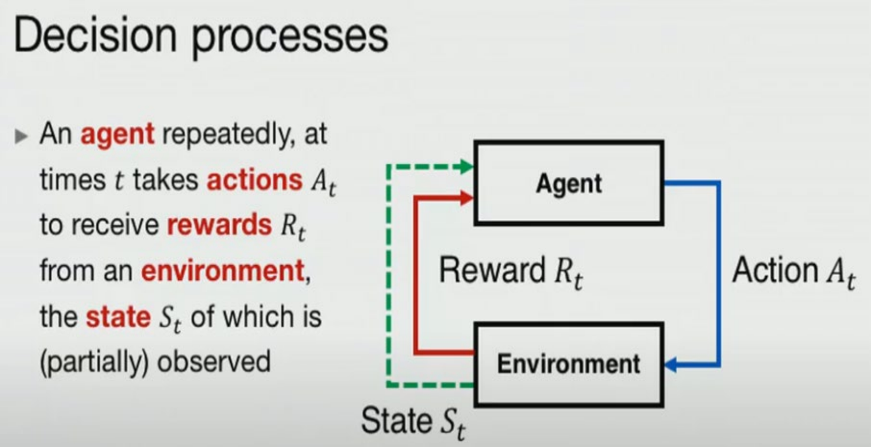
- Suppose someone gave us a policy $\pi(l)$ that outputs a_1 vs a_2

Example: which antibiotic to prescribe?



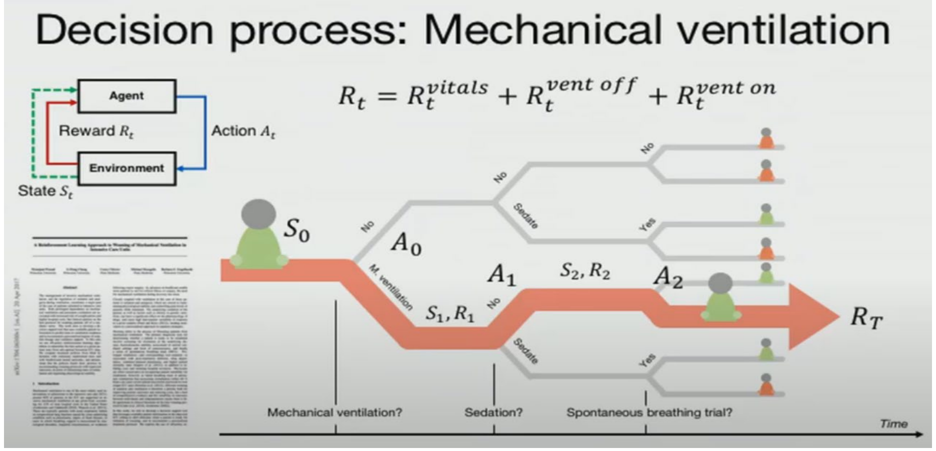
Affects 1 in 2 women during lifetime; 3rd most common cause for antibiotic treatment

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



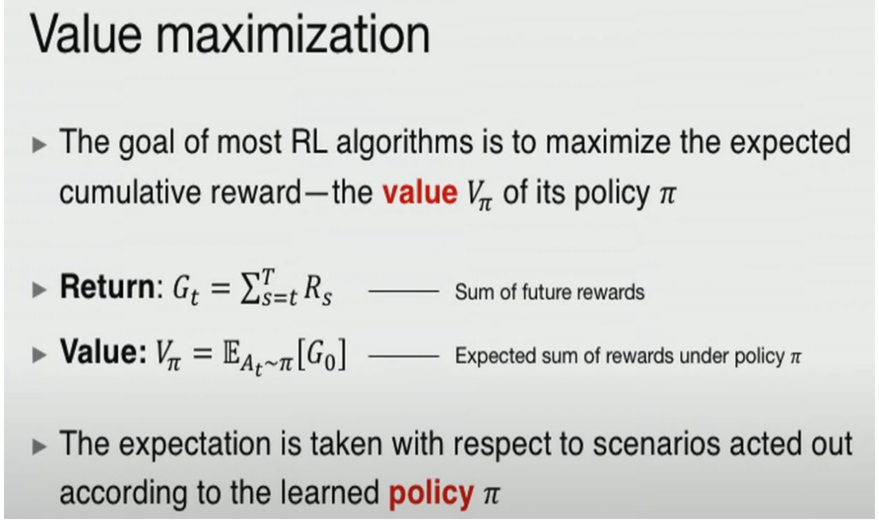
Decision processes

- An **agent** repeatedly, at times t takes **actions** A_t to receive **rewards** R_t from an **environment**, the **state** S_t of which is (partially) observed



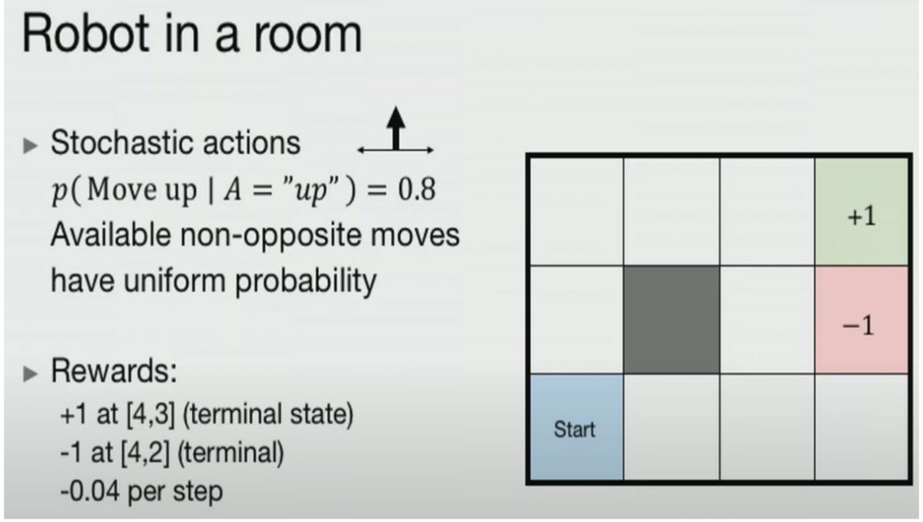
Decision process: Mechanical ventilation

$$R_t = R_t^{vitals} + R_t^{vent\ off} + R_t^{vent\ on}$$



Value maximization

- The goal of most RL algorithms is to maximize the expected cumulative reward—the **value** V_π of its policy π
- **Return:** $G_t = \sum_{s=t}^{T} R_s$ —— Sum of future rewards
- **Value:** $V_\pi = \mathbb{E}_{A_t \sim \pi}[G_0]$ —— Expected sum of rewards under policy π
- The expectation is taken with respect to scenarios acted out according to the learned **policy** π



Robot in a room

- Stochastic actions

  $p(\text{Move up} \mid A = \text{"}up\text{"}) = 0.8$

  Available non-opposite moves have uniform probability

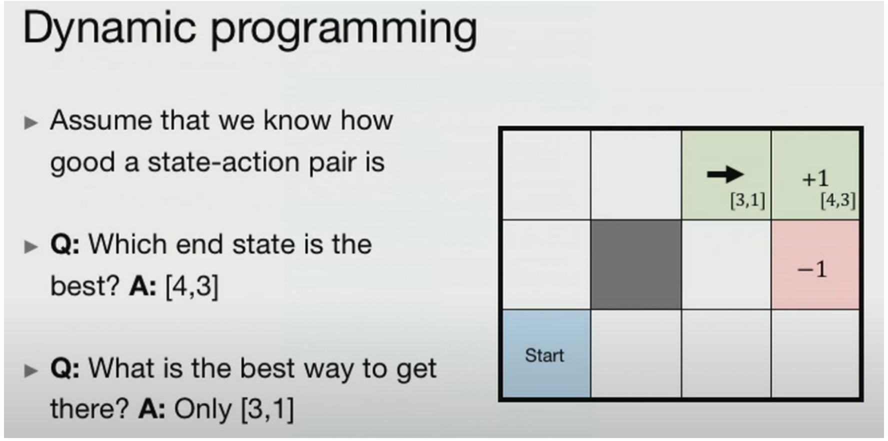
- Rewards:

  +1 at [4,3] (terminal state)

  -1 at [4,2] (terminal)

  -0.04 per step

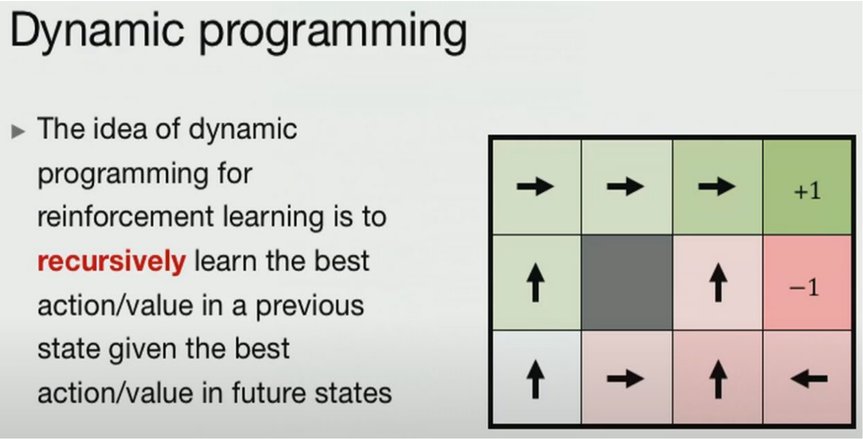
			+1
	▨		−1
Start			



Dynamic programming

- Assume that we know how good a state-action pair is
- **Q:** Which end state is the best? **A:** [4,3]
- **Q:** What is the best way to get there? **A:** Only [3,1]

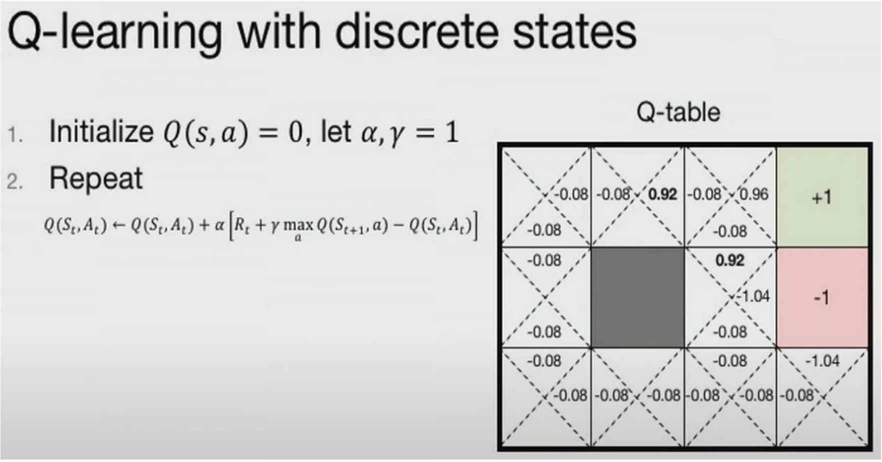
		→ [3,1]	+1 [4,3]
			−1
Start			



Dynamic programming

- The idea of dynamic programming for reinforcement learning is to **recursively** learn the best action/value in a previous state given the best action/value in future states

→	→	→	+1
↑		↑	−1
↑	→	↑	←



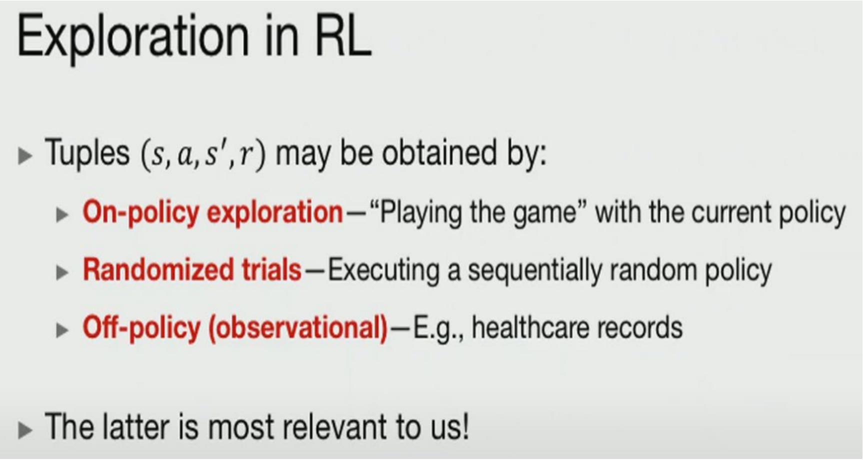
Q-learning with discrete states

1. Initialize $Q(s,a) = 0$, let $\alpha, \gamma = 1$
2. Repeat

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[R_t + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right]$$

Q-table

-0.08 -0.08	-0.08 **0.92**	-0.08 0.96 -0.08	+1
-0.08 -0.08		**0.92** -1.04 -0.08	-1
-0.08 -0.08	-0.08 -0.08	-0.08 -0.08 -0.08	-1.04 -0.08



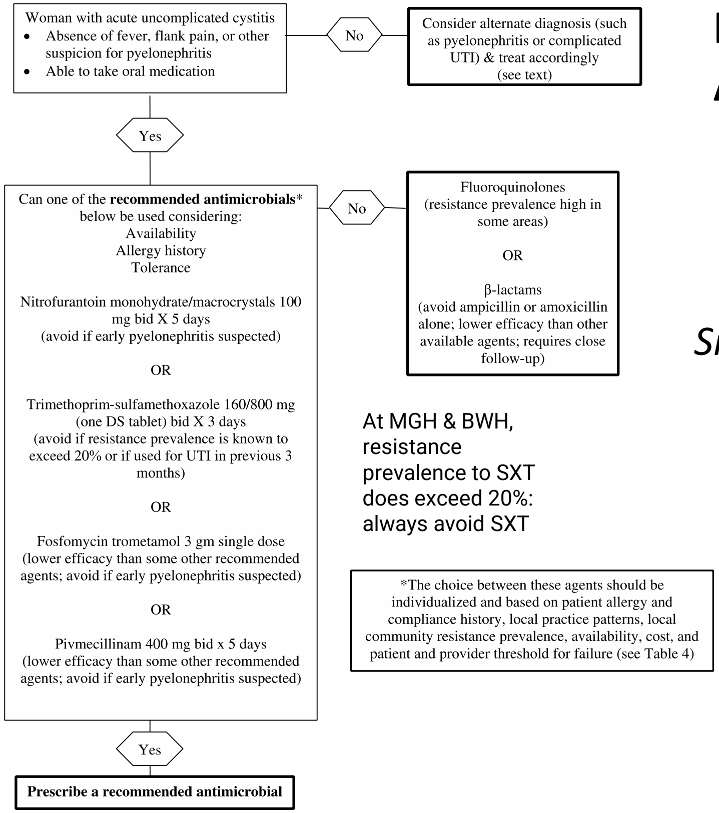
Exploration in RL

- Tuples (s, a, s', r) may be obtained by:
  - **On-policy exploration**—"Playing the game" with the current policy
  - **Randomized trials**—Executing a sequentially random policy
  - **Off-policy (observational)**—E.g., healthcare records
- The latter is most relevant to us!

Evaluate *policies* using observational data with Reinforcement Learning

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

Example: which antibiotic to prescribe?


Woman with acute uncomplicated cystitis
- Absence of fever, flank pain, or other suspicion for pyelonephritis
- Able to take oral medication

No

Consider alternate diagnosis (such as pyelonephritis or complicated UTI) & treat accordingly (see text)

Yes

Can one of the **recommended antimicrobials*** below be used considering:
Availability
Allergy history
Tolerance

Nitrofurantoin monohydrate/macrocrystals 100 mg bid X 5 days (avoid if early pyelonephritis suspected)

OR

Trimethoprim-sulfamethoxazole 160/800 mg (one DS tablet) bid X 3 days (avoid if resistance prevalence is known to exceed 20% or if used for UTI in previous 3 months)

OR

Fosfomycin trometamol 3 gm single dose (lower efficacy than some other recommended agents; avoid if early pyelonephritis suspected)

OR

Pivmecillinam 400 mg bid x 5 days (lower efficacy than some other recommended agents; avoid if early pyelonephritis suspected)

No

Fluoroquinolones (resistance prevalence high in some areas)

OR

β-lactams (avoid ampicillin or amoxicillin alone; lower efficacy than other available agents; requires close follow-up)

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

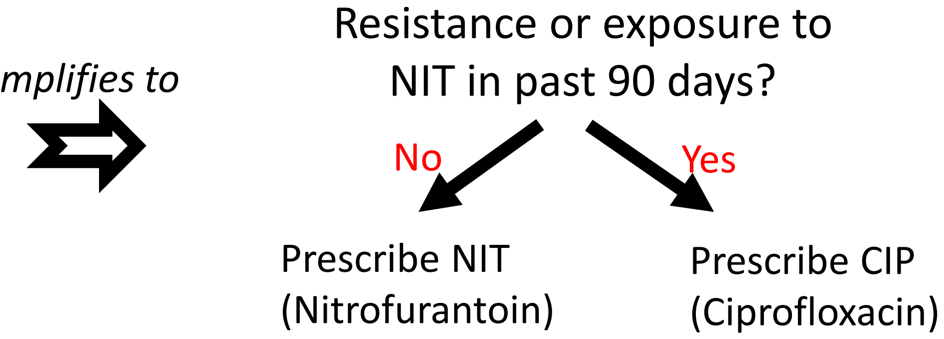
*The choice between these agents should be individualized and based on patient allergy and compliance history, local practice patterns, local community resistance prevalence, availability, cost, and patient and provider threshold for failure (see Table 4)

Yes

Prescribe a recommended antimicrobial

Infectious Disease Society of America (IDSA) guidelines

Simplifies to


Resistance or exposure to NIT in past 90 days?

No

Yes

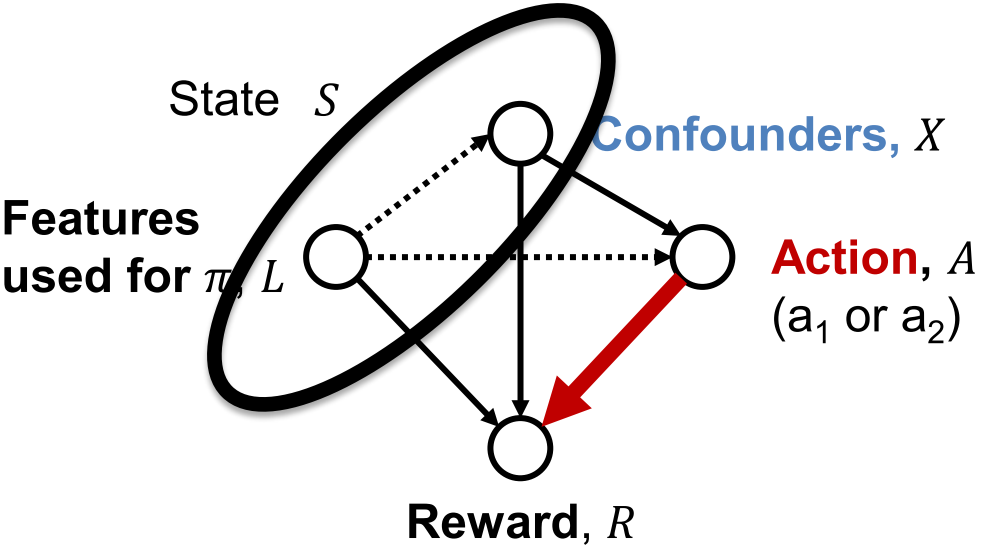
Prescribe NIT (Nitrofurantoin)

Prescribe CIP (Ciprofloxacin)

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

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

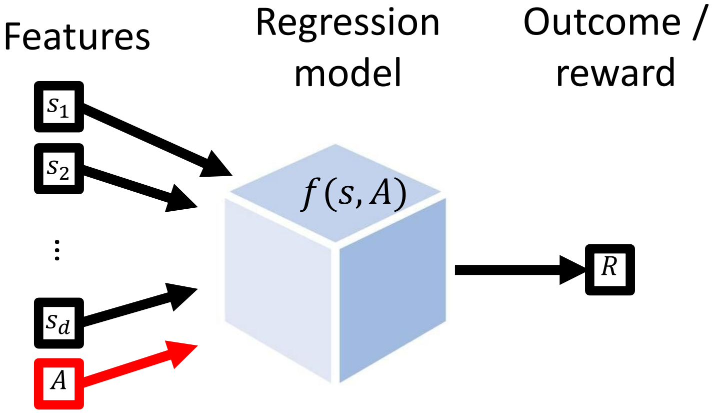
- Suppose someone gave us a policy $\pi(l)$ that outputs a_1 vs a_2
- **How do we evaluate it?**
- We give two approaches, one based on potential outcomes and the other based on propensity scores
- In both cases, we have to first consider the **causal graph** that underlies the *observational data*



Switched notation to what's more typically used in Reinforcement Learning
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)

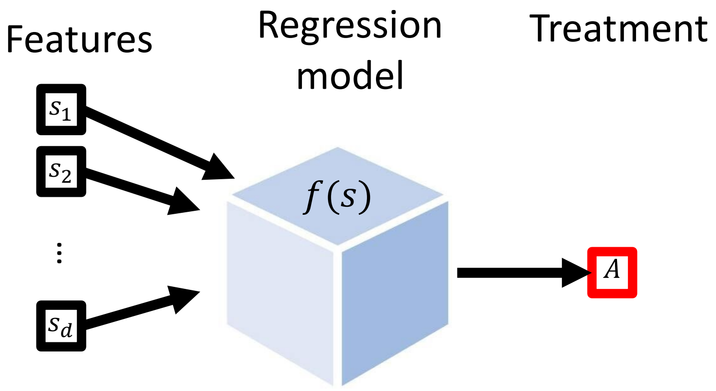


- Then, use this model to impute policy outcomes:

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

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

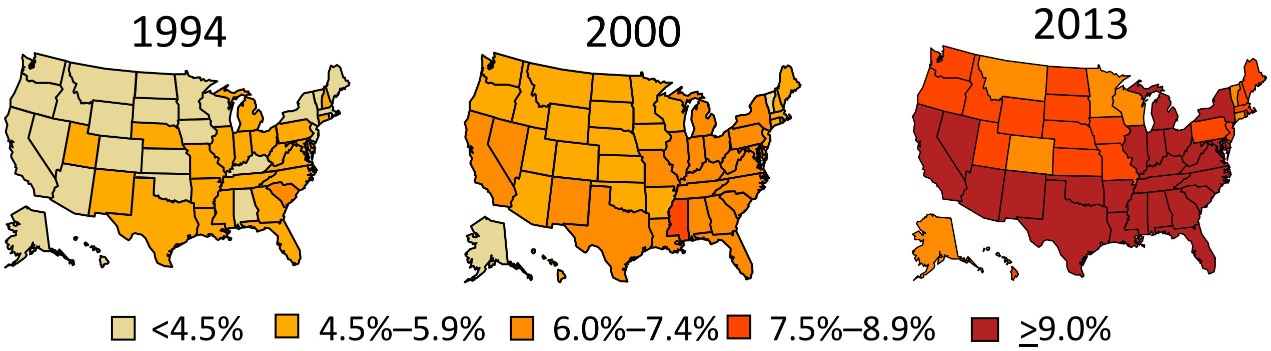
$$o_i = \arg\max_A f(l_i, x_i, A)$$

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*!

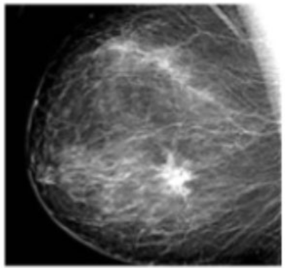
(Makar, Swaminathan, Kiciman. A distillation approach to data efficient individual treatment effect estimation. AAAI, 2019)

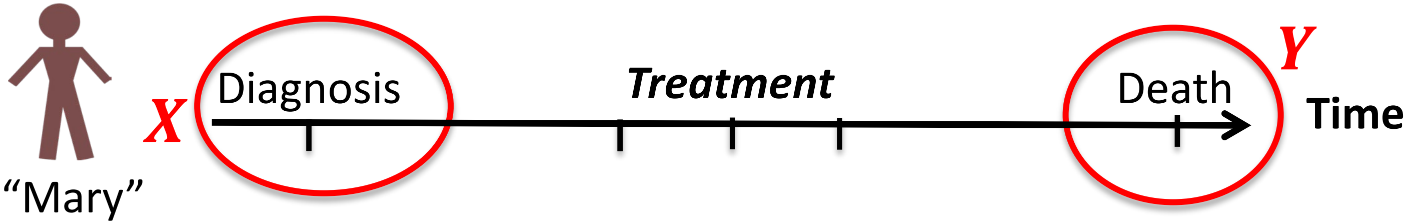
Does gastric bypass surgery prevent onset of diabetes?



- 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.

Reinforcement Learning for policy evaluation

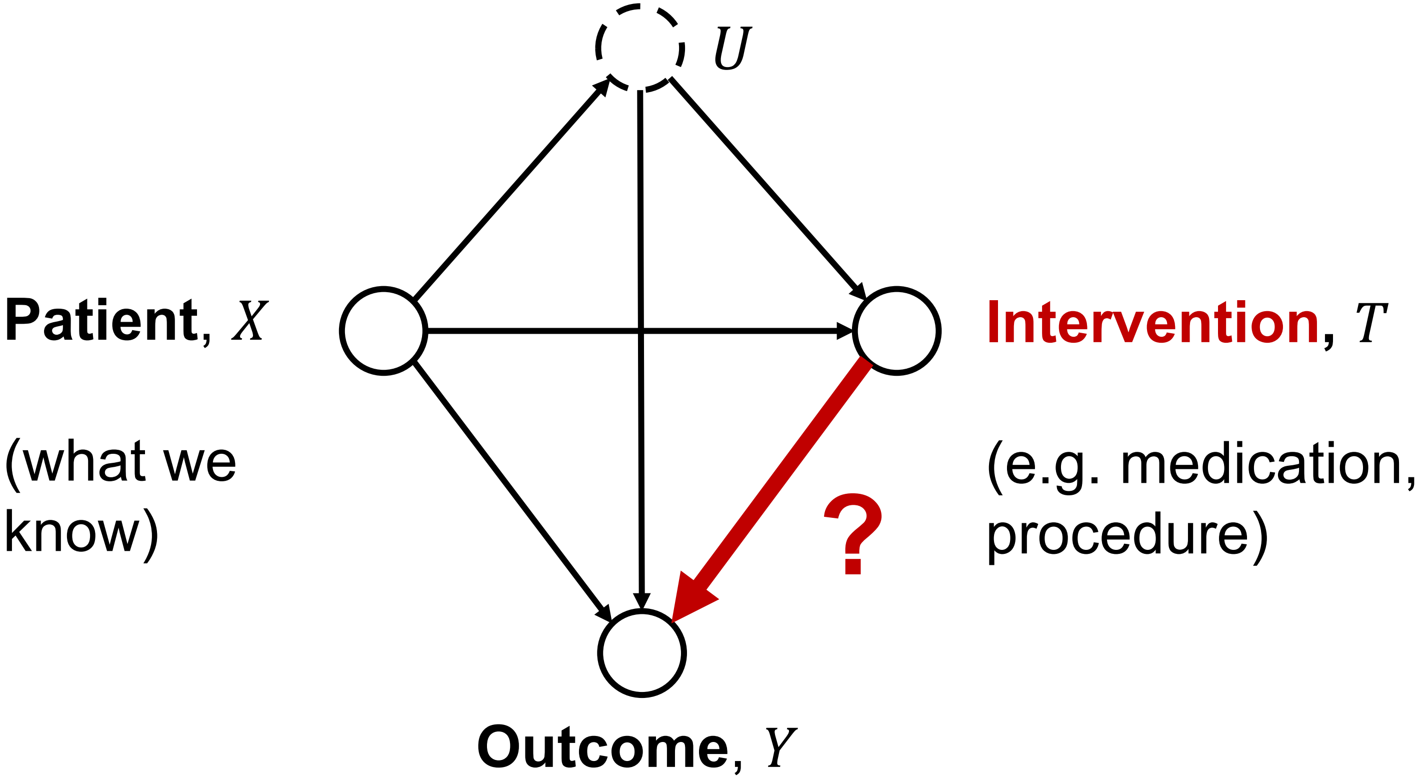
Using observational data

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*



U

Patient, X

(what we know)

Intervention, T

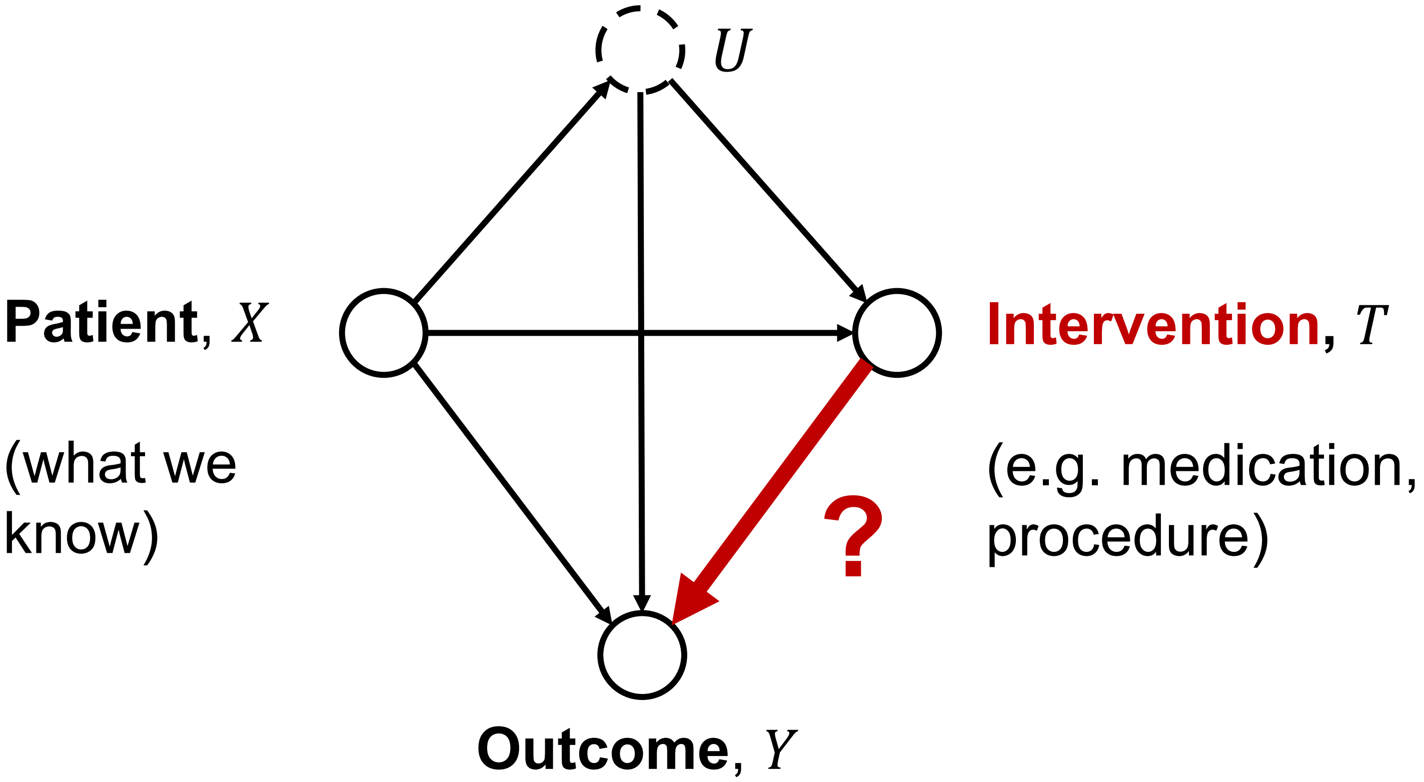
(e.g. medication, procedure)

?

Outcome, Y

Estimation using an instrumental variable

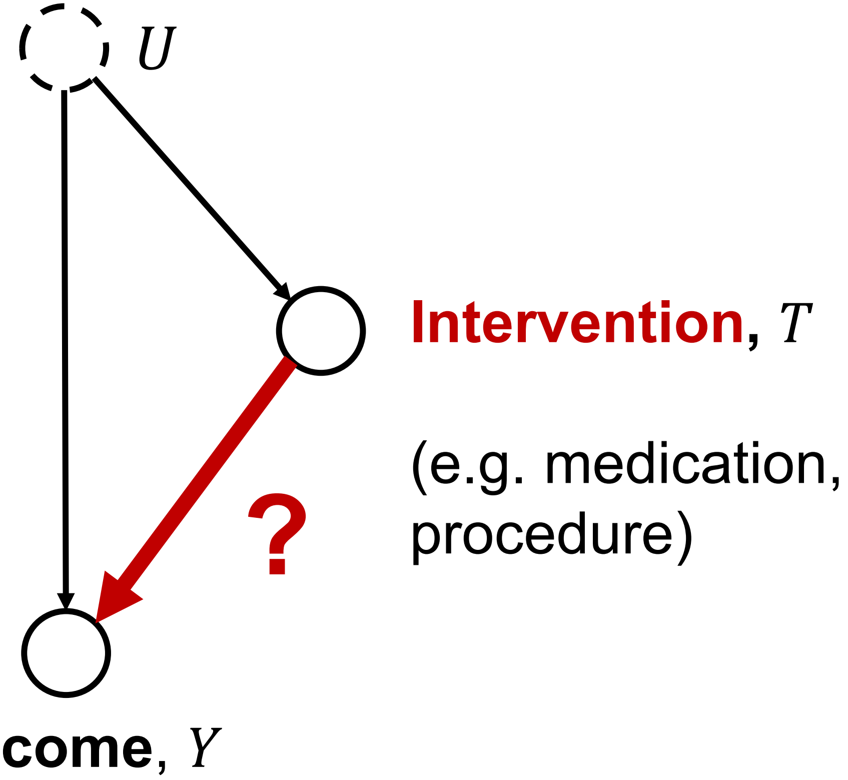
First, assume no patient covariates (with this, we will only be able to estimate ATE not CATE)



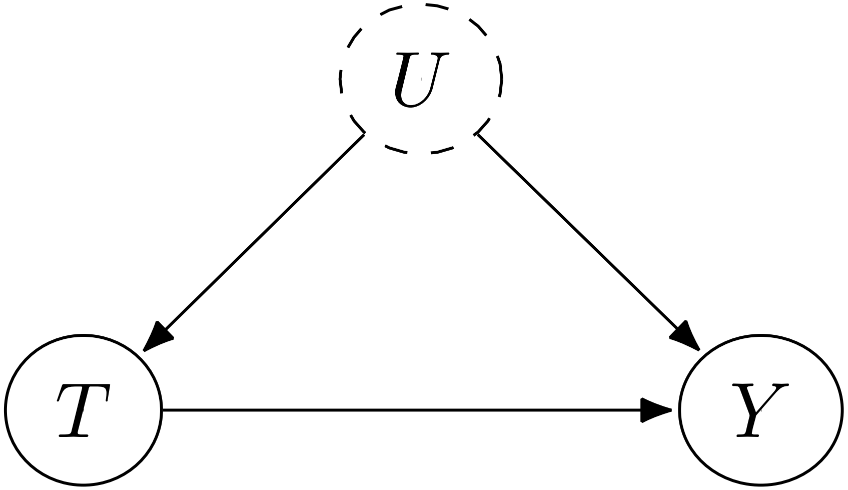
Estimation using an instrumental variable

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)



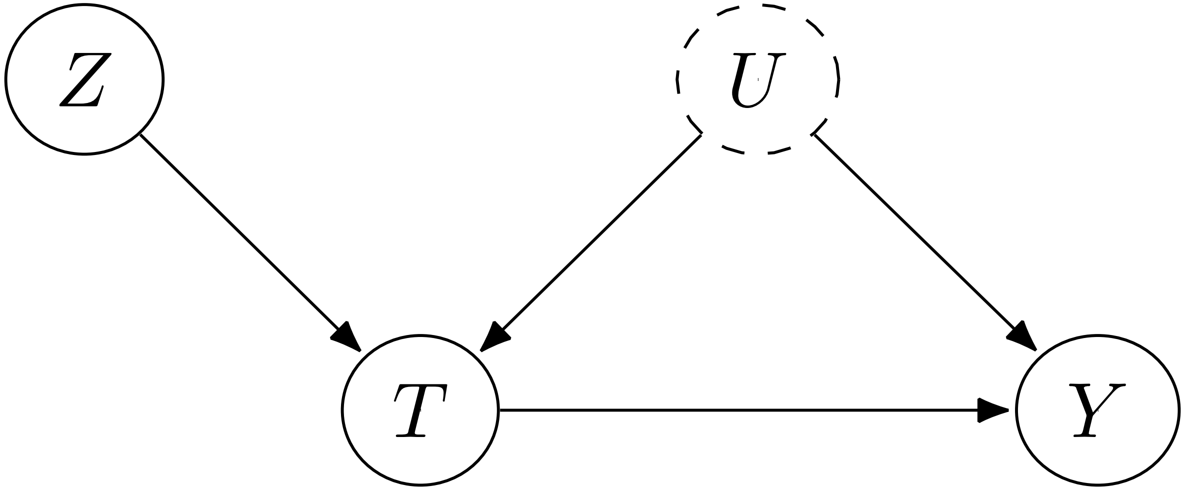
Estimation using an instrumental variable



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

Estimation using an instrumental variable

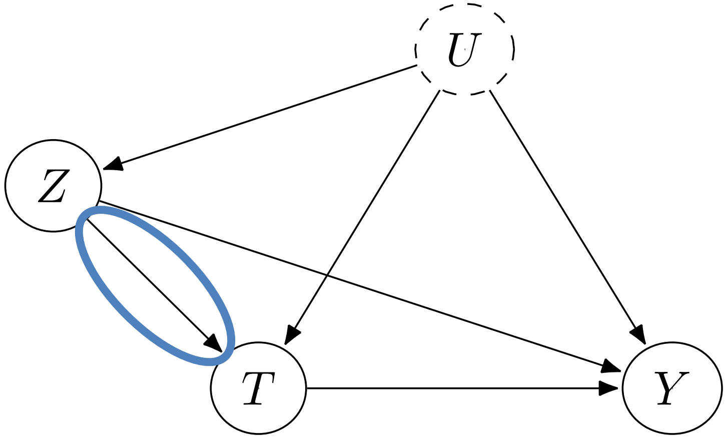
Instrument (e.g., voucher)



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

Assumption 1: Relevance

Z has a causal effect on T

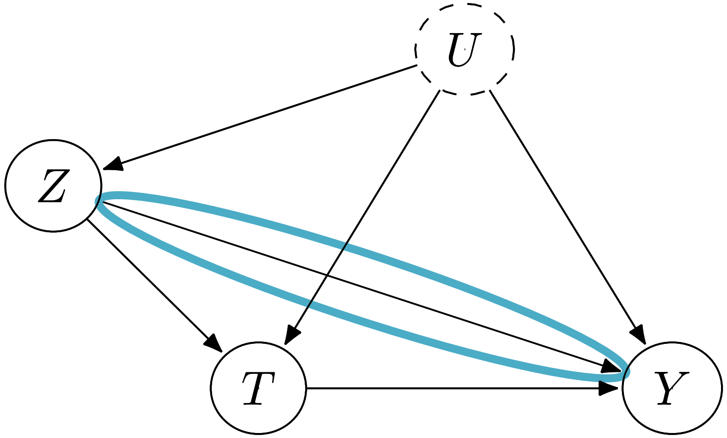


What is an Instrument?

(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

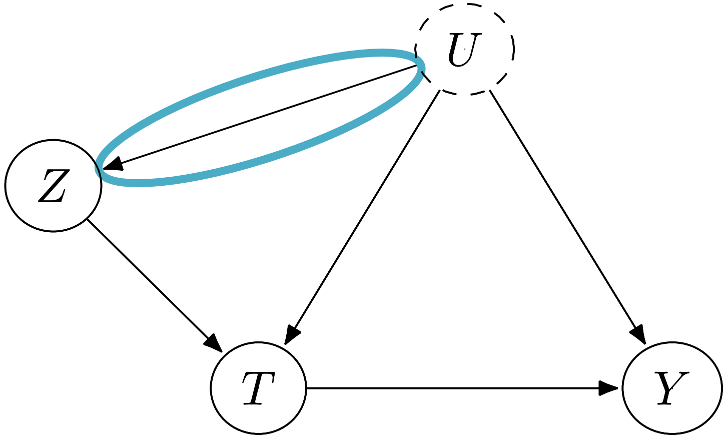


What is an Instrument?

(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?

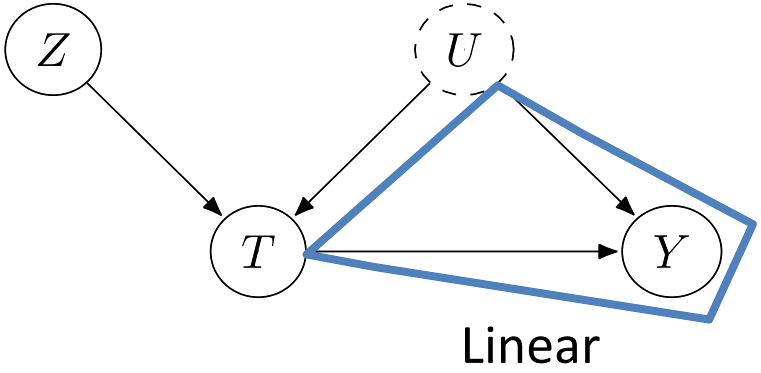
(Slides adapted from Brady Neal's Introduction to Causal Inference class)

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

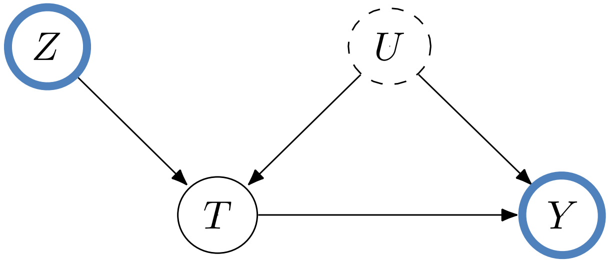


(Slides adapted from Brady Neal's Introduction to Causal Inference class)

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

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

$$\delta = \frac{\mathbb{E}[Y \mid Z=1] - \mathbb{E}[Y \mid Z=0]}{\underbrace{\mathbb{E}[T \mid Z=1] - \mathbb{E}[T \mid Z=0]}_{\text{(non-zero due to relevance assumption)}}}$$



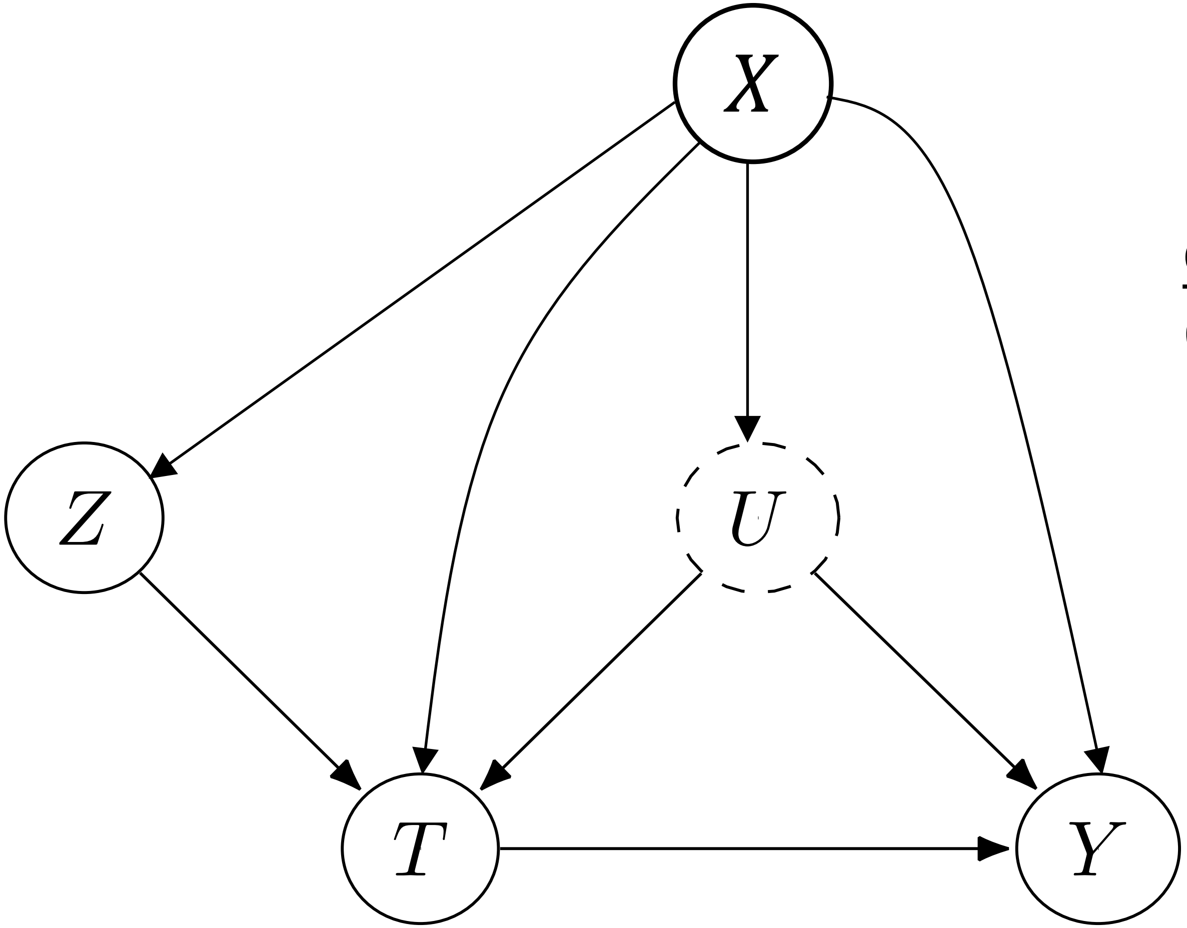
$$Y_t(U) = \alpha_u U + \delta \cdot t + \epsilon_t$$

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

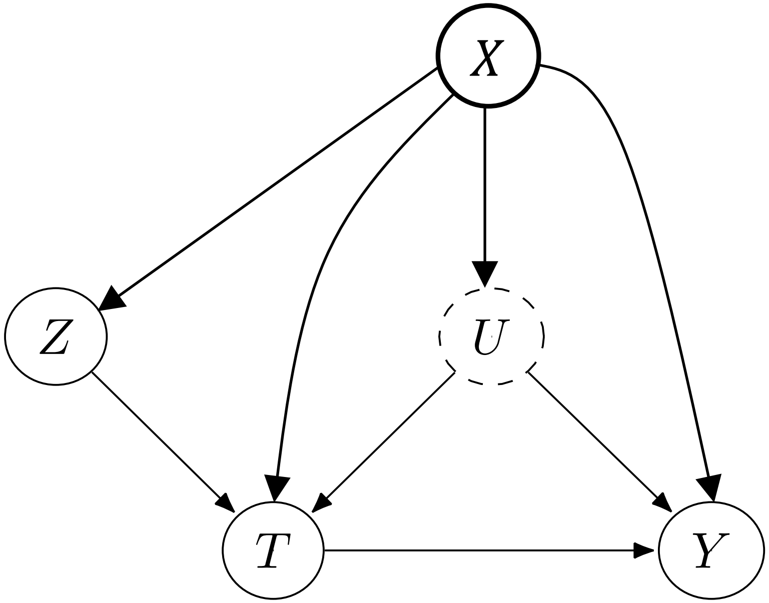
Estimation using (conditional) instruments

Assume potential outcomes given by:

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

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

(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.

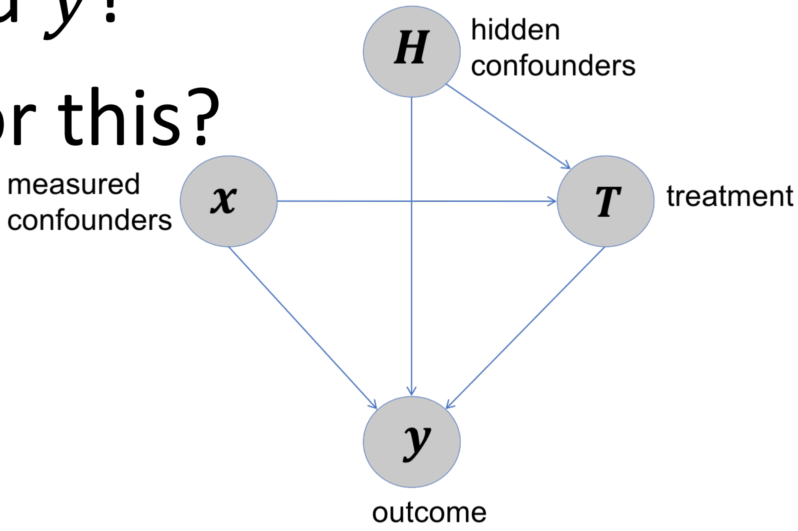
There is believed to be some heritability for both addiction and lung cancer.

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)

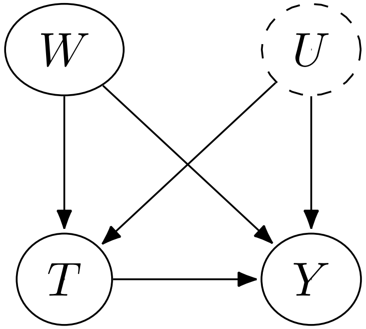
Sensitivity analysis and hidden confounding

- 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

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

Linear Y

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

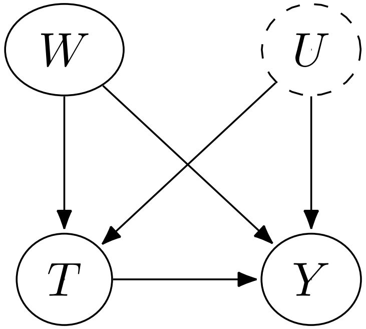
Goal: recover δ

(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

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

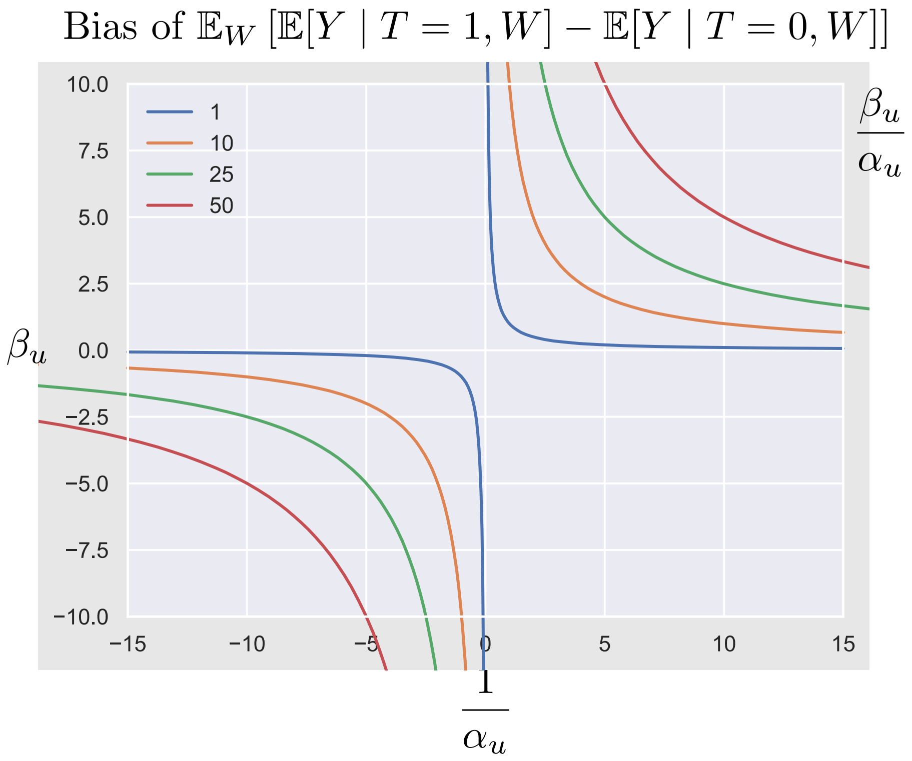
$$\mathbb{E}_W\left[\mathbb{E}[Y \mid T=1, W] - \mathbb{E}[Y \mid T=0, W]\right] \overset{?}{=} \delta + \frac{\beta_u}{\alpha_u}$$

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

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:

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

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

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

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

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

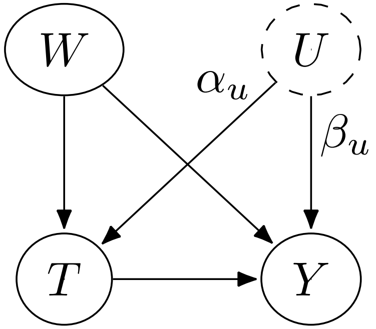
$$\begin{aligned}\mathbb{E}_W\left[\mathbb{E}[Y \mid T=1, W] - \mathbb{E}[Y \mid T=0, W]\right] &= \left(\delta + \frac{\beta_u}{\alpha_u}\right)(1) + \left(\beta_w - \frac{\beta_u \alpha_w}{\alpha_u}\right) \mathbb{E}[W] \\ &\quad - \left[\left(\delta + \frac{\beta_u}{\alpha_u}\right)(0) + \left(\beta_w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

Sensitivity Analysis: Linear Single Confounder

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

Bias in Simple Linear Setting Proof: Step 3

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



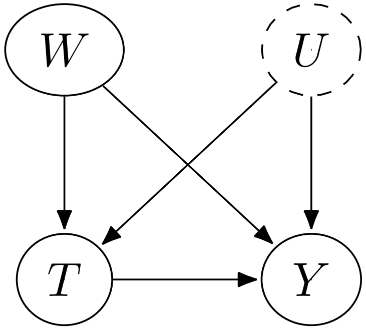
$$T := \alpha_w W + \alpha_u U$$
$$Y := \beta_w W + \beta_u U + \delta T$$

Sensitivity Analysis: Linear Single Confounder

(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)

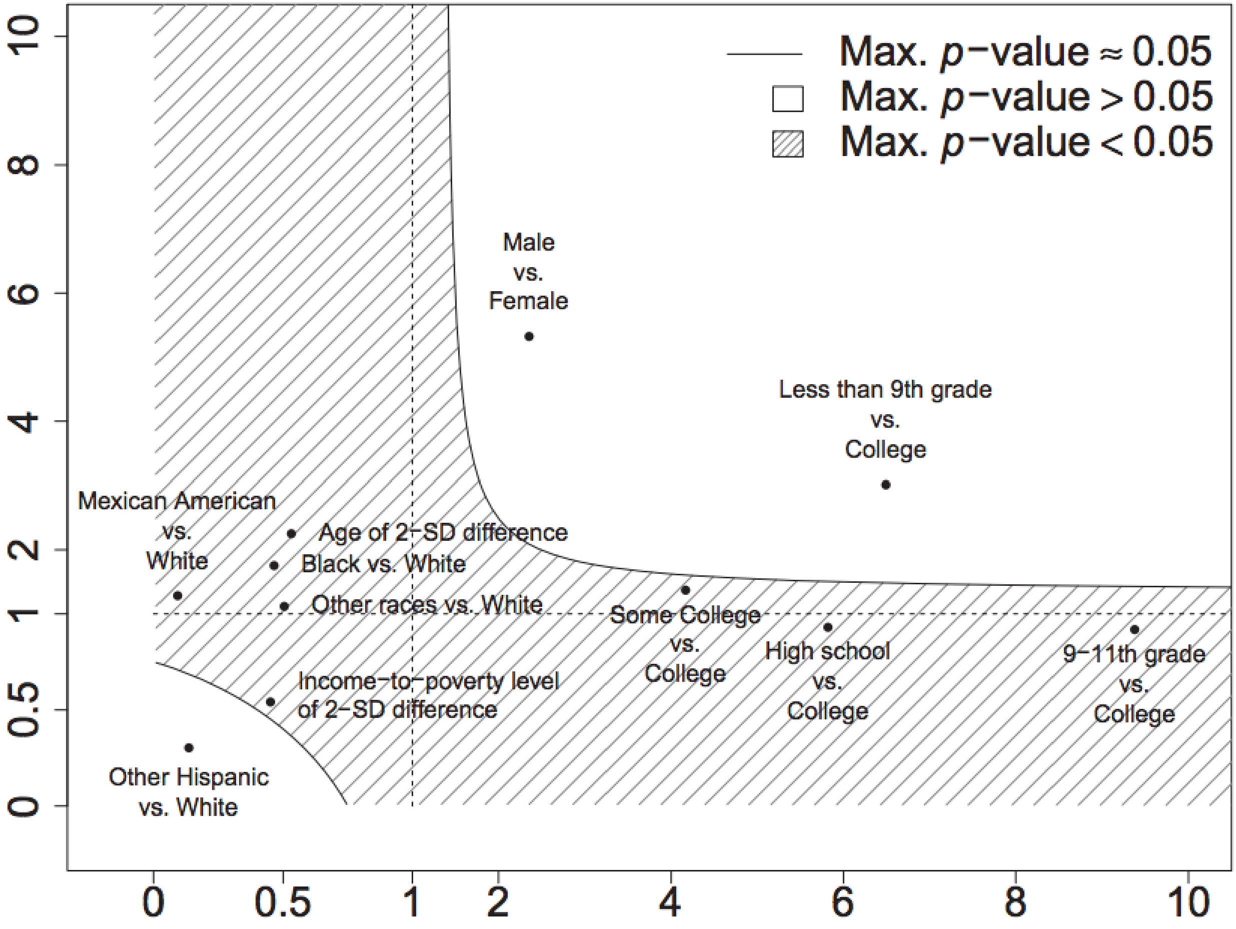
Sensitivity analysis with binary treatment

- 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 $U \to Y$



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

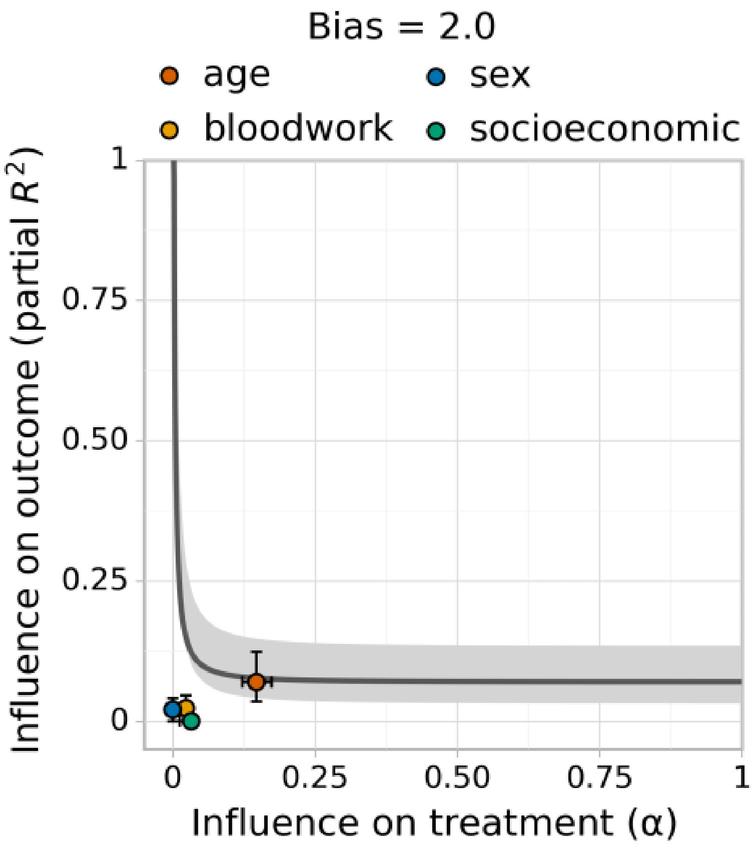
Hsu & Small, 2013

(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.