



Introduction to Causal Inference - a Machine Learning Perspective

Dr. Uri Shalit

Course number 097400
2020-2021

Lesson 2

“A causes B” – what does it mean?

- Long philosophical history (Greek, Hindu, Buddhist...)
- More recent:
- Hume (1793)
“...an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second

“A causes B” – what does it mean?

- Long philosophical history (Greek, Hindu, Buddhist...)
- More recent:
- Hume (1793)
“...an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second, where, if the first object had not been, the second never had existed.”

“A causes B” – what does it mean?

- Hume (1793)

“...an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second, where, if the first object had not been, the second never had existed.”

- Lewis (1973)

A causes B if:

- A and B both occur
- If A had not occurred and all else remained the same, then B would not have occurred

“A causes B” – what does it mean?

- Lewis (1973)

A causes B if:

- A and B both occur
- If A had not occurred **and all else remained the same,**
then B would not have occurred

“A causes B” – what does it mean?

- Lewis (1973)

A causes B if:

- A and B both occur
- If A had not occurred **and all else remained the same,**
then B would not have occurred
- Main challenge: define a world where A had not occurred
and all else remained the same

“Does A cause B?”: The role of experiments (manipulation, intervention)

- Main challenge: define a world where A had not occurred and all else remained the same
- Rubin’s dictum: “no causation without manipulation”
- Imagine a world where we can change A
- Examples:
 - Does medication lower blood sugar?
 - Does attending university improve my income at age 50?
 - Does being black decrease my chances of getting a job at a law firm?
 - Does subsidizing factories increase employment over 10 years?

What is the causal question

- Does medication lower blood sugar?
 - Can easily imagine manipulation: not taking medication
- Does attending university improve my income at age 50?
 - What manipulation? Maybe specific alternative: joining startup at same age?
Different causal question
- Does being black decrease my chances of getting a job at a law firm?
 - What manipulation? Maybe change name/race on CV or official forms?
Imagine changing pigments? Again, different causal question
- Does subsidizing factories increase employment over 10 years?
 - Can easily imagine not subsidizing. But, in this world, what is the saved money being used for?

Potential outcomes

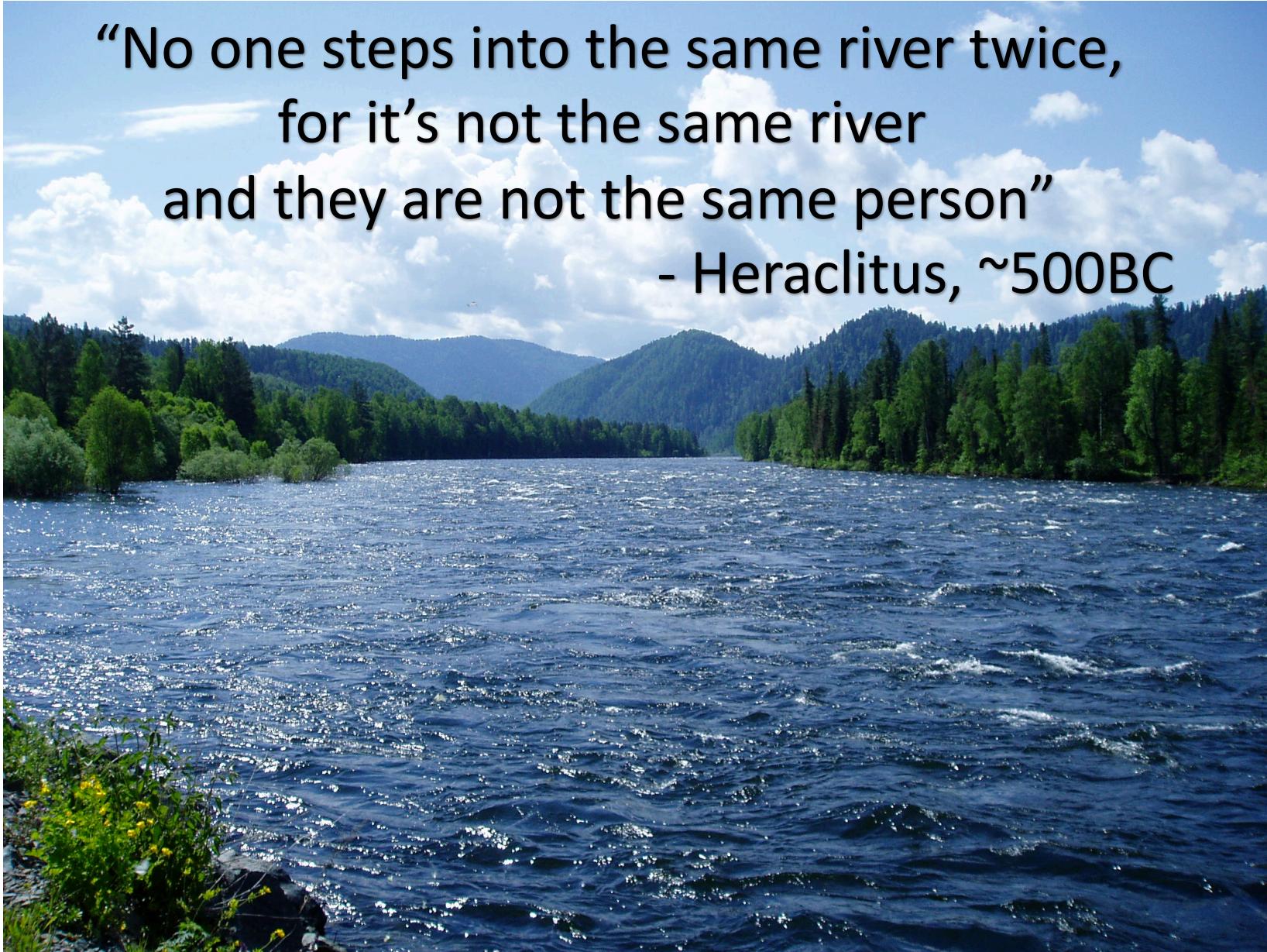
- Unit: a person, a bacteria, a company, a school, a website, a family, a piece of metal, ...
- Treatments / actions / interventions
- Potential outcomes

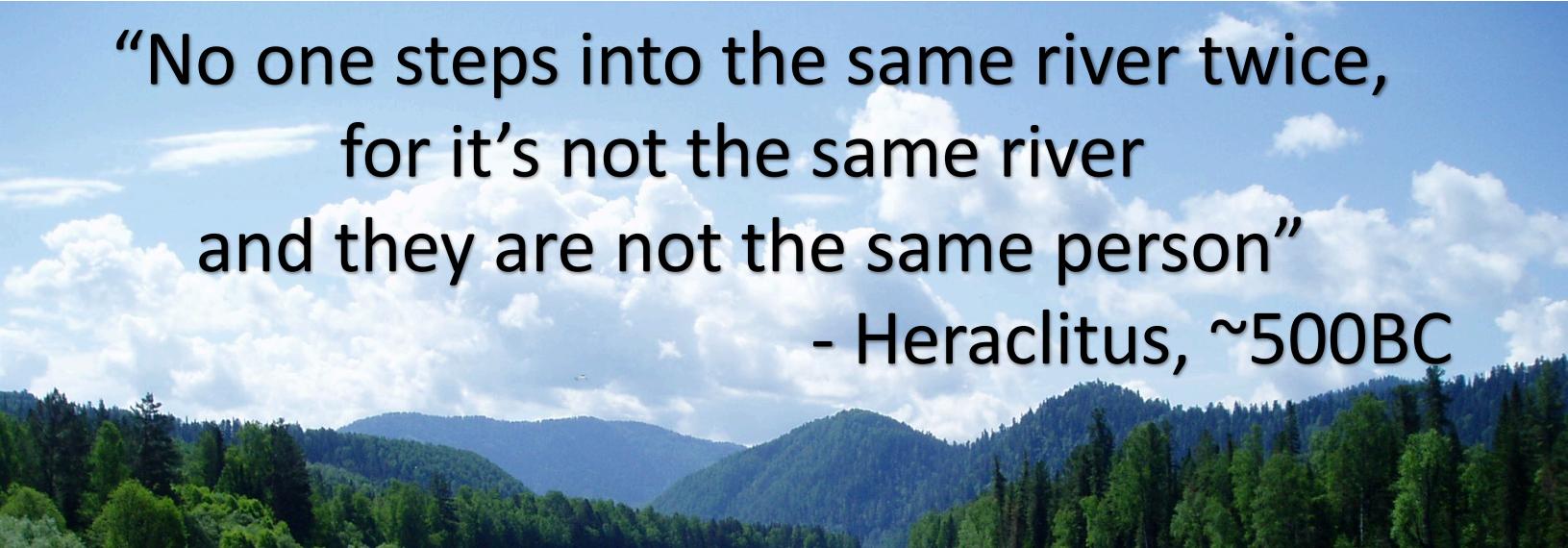
Y_1 : the unit's outcome *had they been subjected to treatment t=1*

Y_0 : the unit's outcome *had they been subjected to treatment t=0*

If number of treatments is T, we have T potential outcomes (T possibly infinite)

“No one steps into the same river twice,
for it's not the same river
and they are not the same person”
- Heraclitus, ~500BC





“No one steps into the same river twice,
for it's not the same river
and they are not the same person”

- Heraclitus, ~500BC



“You only ever see one potential outcome”

- Rubin, 1974

How do I infer the unknown?

- Multiple units
 - Light switch example
 - Self-experimentation example
 - In both examples we have **prior belief** that “everything **important** stayed the same”
- Sometimes it’s harder to believe everything important stayed the same
 - Self-experimentation with anti-diabetic medication
- We collect many units with different treatments to make an inference. This requires **identifying-assumptions**

The stages of causal inference

1. Formulate *causal assumptions* sufficient to solve the problem
 - these are mostly **untestable**
2. Under the assumptions, reduce causal problem to appropriate statistical/machine learning method
 - these methods are often specialized methods, similar but distinct from familiar methods such as regression

Identification

Estimation

Potential outcomes

- Unit: a person, a bacteria, a company, a school, a website, a family, a piece of metal, ...
- Treatments / actions / interventions
- Potential outcomes

Y_1 : the unit's outcome *had they been subjected to treatment t=1*

Y_0 : the unit's outcome *had they been subjected to treatment t=0*

If number of treatments is K, we have K potential outcomes (K possibly infinite)

Potential outcomes

- Y_0, Y_1 : potential outcomes
- T : binary treatment
- X : observed covariates
- Y : observed outcome

Consistency assumption

$$Y = TY_1 + (1 - T)Y_0:$$

Potential outcomes

- Y_0, Y_1 : potential outcomes
- T : binary treatment
- X : observed covariates
- Y : observed outcome

Consistency assumption

$$Y = TY_1 + (1 - T)Y_0:$$

Y is different from Y_0, Y_1

Potential Outcomes

- Each unit i has two potential outcomes:
 - Y_0 is the potential outcome had the unit received treatment 0
 - Y_1 is the potential outcome had the unit received treatment 1
- Average Treatment Effect:
$$ATE := \mathbb{E}[Y_1 - Y_0]$$
- The treatment assignment determines which of Y_0 and Y_1 we get to see

Potential Outcomes

- Each unit i has two potential outcomes:
 - Y_0 is the potential outcome had the unit received treatment 0
 - Y_1 is the potential outcome had the unit received treatment 1
- **Average Treatment Effect:**
$$ATE := \mathbb{E}[Y_1 - Y_0]$$
- The treatment assignment determines which of Y_0 and Y_1 we get to see

Potential Outcomes

- Y_1 can be very different from $Y|T = 1$
- Say T is job training,
- $Y|T = 1$: the income we expect from people who actually went to job training
- Y_1 : the income we expect if we *forced everyone* to go to job training
- Only under special conditions do we have equality

“The fundamental problem of
causal inference”

We only ever observe one of the
two outcomes

Non-identifiable example: the story of the smart snake-oil salesman

- Hidden confounder x :
whether a patient will recover from illness ($y = 1$) or not ($y = 0$)
 $p(y = x) = 1$
- Smart snake-oil salesman: knows x even when patients don't
- For patients with $x = 1$, he gives them his SnakeOil™ ($T = 1$)
For patients with $x = 0$, he does not give them his SnakeOil™ ($T = 0$)
- Just from observing, we cannot know whether SnakeOil™ has healing powers

Non-identifiable example: the story of the smart snake-oil salesman

	Y_0	Y_1	T	Y
$x = 0$			0	0
$x = 1$			1	1

Non-identifiable example: the story of the smart snake-oil salesman



	Y_0	Y_1	T	Y
$x = 0$	0	1	0	0
$x = 1$	0	1	1	1

Non-identifiable example: the story of the smart snake-oil salesman



	Y_0	Y_1	T	Y
$x = 0$	0	0	0	0
$x = 1$	1	1	1	1



Potential outcomes

(age, gender, exercise,treatment)			Observed sugar levels
(45, F, 0, A)			6
(45, F, 1, B)			6.5
(55, M, 0, A)			7
(55, M, 1, B)			8
(65, F, 0, B)			8
(65,F, 1, A)			7.5
(75,M, 0, B)			9
(75,M, 1, A)			8

Potential outcomes

(age, gender, exercise)			Observed sugar levels
(45, F, 0)			6
(45, F, 1)			6.5
(55, M, 0)			7
(55, M, 1)			8
(65, F, 0)			8
(65, F, 1)			7.5
(75, M, 0)			9
(75, M, 1)			8

Potential outcomes

(age, gender, exercise)	Y_0 : Sugar levels <i>had they received</i> medication A	Y_1 : Sugar levels <i>had they received</i> medication B	Observed sugar levels
(45, F, 0)	6	5.5	6
(45, F, 1)	7	6.5	6.5
(55, M, 0)	7	6	7
(55, M, 1)	9	8	8
(65, F, 0)	8.5	8	8
(65, F, 1)	7.5	7	7.5
(75, M, 0)	10	9	9
(75, M, 1)	8	7	8

Potential outcomes

(age,gender, exercise)	Sugar levels <i>had they received</i> medication A	Sugar levels <i>had they received</i> medication B	Observed sugar levels
(45, F, 0)	6	5.5	6
(45, F, 1)	7	6.5	6.5
(55, M, 0)	7	6	7
(55, M, 1)	9	8	8
(65, F, 0)	8.5	8	8
(65, F, 1)	7.5	7	7.5
(75, M, 0)	10	9	9
(75, M, 1)	8	7	8

$$\text{mean}(\text{sugar} | \text{medication B}) - \text{mean}(\text{sugar} | \text{medication A}) = 7.875 - 7.125 = 0.75$$

$$\text{mean}(\text{sugar} | \text{had they received B}) - \text{mean}(\text{sugar} | \text{had they received A}) = 7.125 - 7.875 = -0.75$$

What do we wish to estimate?

- Sample of units $i = 1, \dots, n$
- Each has potential outcomes $(Y_0^1, Y_1^1), \dots, (Y_0^n, Y_1^n)$

- Individual Treatment Effect for unit i :

$$ITE_i \equiv Y_1^i - Y_0^i$$

- Average Treatment Effect over the sample

$$ATE_{finite} \equiv \frac{1}{n} \sum_{i=1}^n Y_1^i - Y_0^i$$

- Usually: assume some joint distribution $p(Y_0, Y_1)$

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

- Define average over which population (“diabetics living in Israel over age 65”)

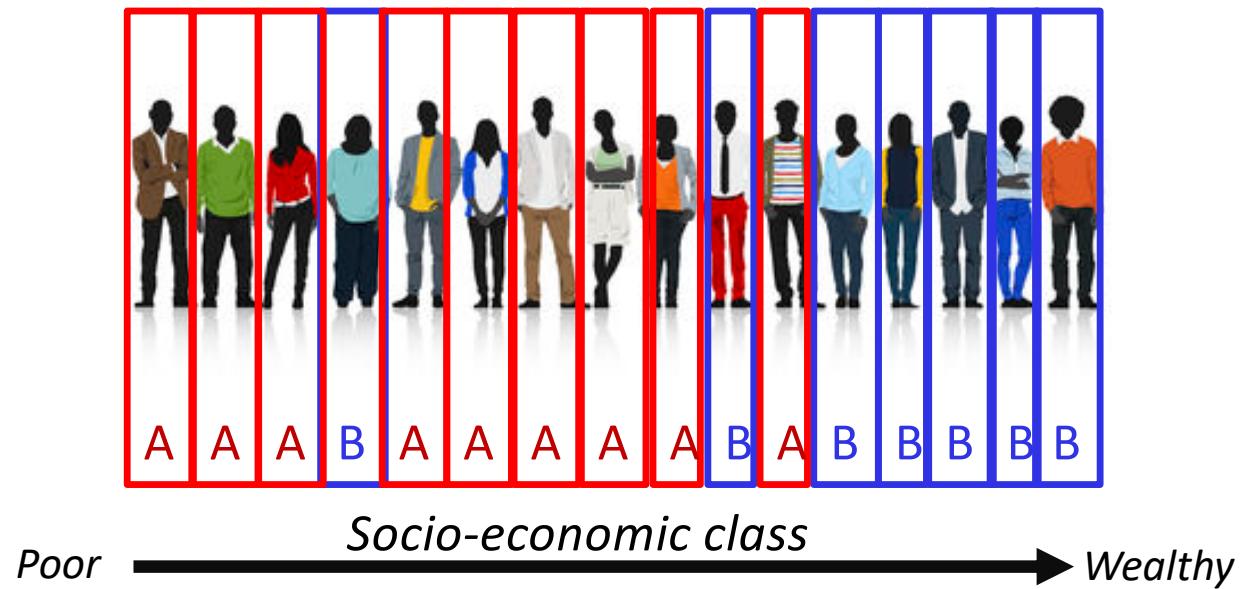
What do we wish to estimate?

- Usually: assume some joint distribution $p(Y_0, Y_1)$
$$ATE \equiv \mathbb{E}[Y_1 - Y_0]$$

- Conditional Average Treatment Effect (CATE)
for feature (covariate) X

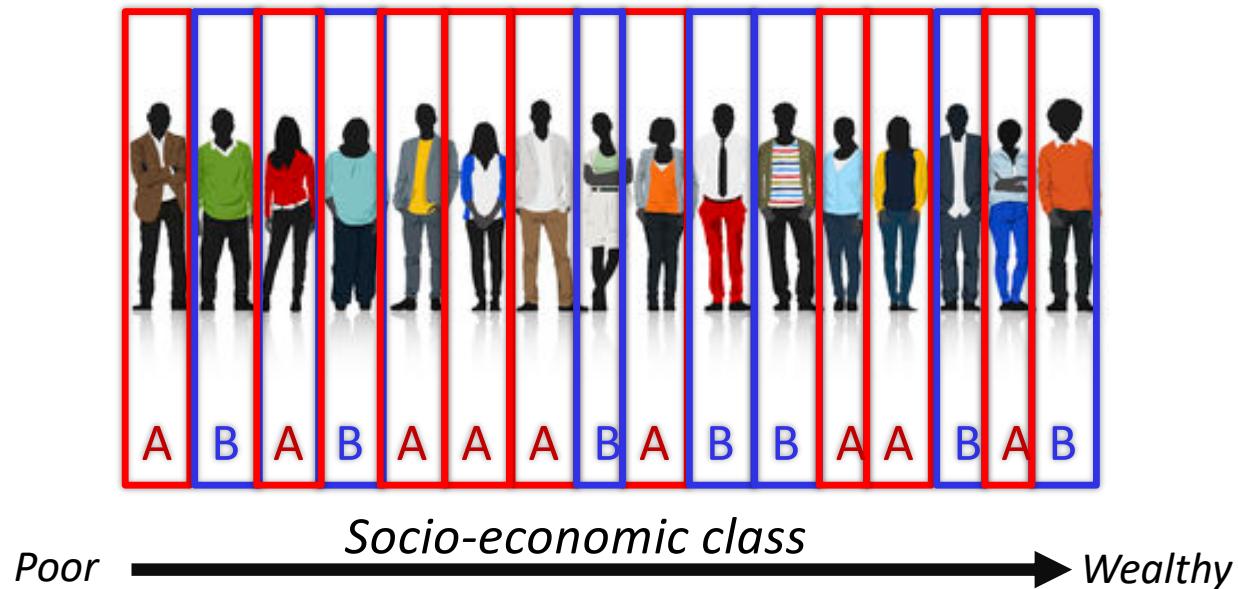
$$CATE \equiv \mathbb{E}[Y_1 - Y_0 | X]$$

Observational study



treatment
A or B

Randomized controlled trial (RCT)



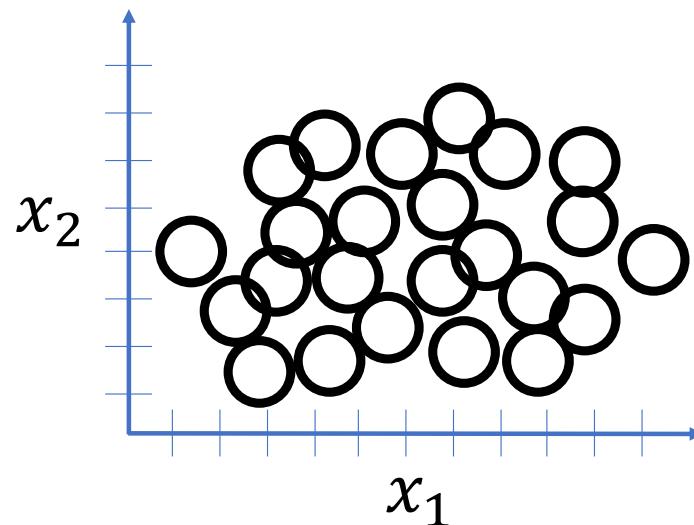
treatment
A or B

Complete randomization promises unbiased
causal estimates:
The power of randomized trials
(aka A/B testing)

- Complete randomization promises unbiased causal estimates: The power of randomized trials
- Identifying assumption: treatment is randomized

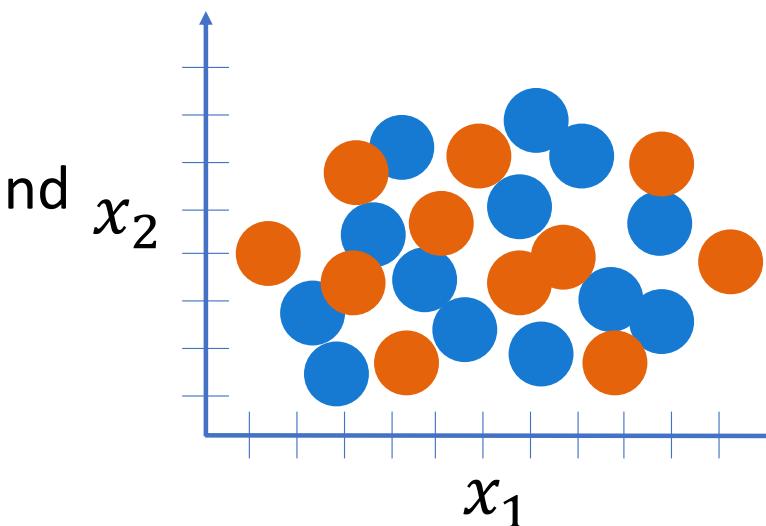
Identification

When is estimating treatment effect easier? Randomized Controlled Trials



- Control, $t = 0$
- Treated, $t = 1$

When is estimating treatment effect easier? Randomized Controlled Trials



- Control, $t = 0$
- Treated, $t = 1$

From unobservable to observable

- Y_0, Y_1 : potential outcomes
- T : binary treatment
- X : observed covariates
- $Y = Y_{obs} = TY_1 + (1 - T)Y_0$

Assume that treatment assignment is
completely random

$$Y = Y_{obs} = TY_1 + (1 - T)Y_0$$

- Treatment is random:
 $(Y_0, Y_1) \perp\!\!\!\perp T$

Assume that treatment assignment is
completely random

$$Y = Y_{obs} = TY_1 + (1 - T)Y_0$$

- Treatment is random:

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

- $\mathbb{E}[Y_1] =$ Because treatment is random \rightarrow prior knowledge!
- $\mathbb{E}[Y_1|T = 1] =$

Assume that treatment assignment is
completely random

$$Y = Y_{obs} = TY_1 + (1 - T)Y_0$$

- Treatment is random:

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

- $\mathbb{E}[Y_1] =$ Because treatment is random → prior knowledge!
- $\mathbb{E}[Y_1|T = 1] =$ Math
- $\mathbb{E}[TY_1|T = 1] =$

Assume that treatment assignment is
completely random

$$Y = Y_{obs} = TY_1 + (1 - T)Y_0$$

- Treatment is random:

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

- $\mathbb{E}[Y_1] =$ Because treatment is random \rightarrow prior knowledge!
- $\mathbb{E}[Y_1|T = 1] =$ Math
- $\mathbb{E}[TY_1|T = 1] =$ From definition of Y_{obs}
- $\mathbb{E}[Y_{obs} - (1 - T)Y_0|T = 1] =$

Assume that treatment assignment is
completely random

$$Y = Y_{obs} = TY_1 + (1 - T)Y_0$$

- Treatment is random:

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

- $\mathbb{E}[Y_1] =$ Because treatment is random → prior knowledge!
- $\mathbb{E}[Y_1|T = 1] =$ Math
- $\mathbb{E}[TY_1|T = 1] =$ From definition of Y_{obs}
- $\mathbb{E}[Y_{obs} - (1 - T)Y_0|T = 1] =$ Linearity of expectation
- $\mathbb{E}[Y_{obs}|T = 1] - \mathbb{E}[(1 - T)Y_0|T = 1] =$

Assume that treatment assignment is
completely random

$$Y = Y_{obs} = TY_1 + (1 - T)Y_0$$

- Treatment is random:

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

- $\mathbb{E}[Y_1] =$ Because treatment is random → prior knowledge!
- $\mathbb{E}[Y_1|T = 1] =$ Math
- $\mathbb{E}[TY_1|T = 1] =$ From definition of Y_{obs}
- $\mathbb{E}[Y_{obs} - (1 - T)Y_0|T = 1] =$ Linearity of expectation
- $\mathbb{E}[Y_{obs}|T = 1] - \mathbb{E}[(1 - T)Y_0|T = 1] =$
- $\mathbb{E}[Y_{obs}|T = 1]$

Assume that treatment assignment is
completely random

$$Y = Y_{obs} = TY_1 + (1 - T)Y_0$$

- Treatment is random:

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

- $\mathbb{E}[Y_1] =$
- $\mathbb{E}[Y_1|T = 1] =$
- $\mathbb{E}[TY_1|T = 1] =$
- $\mathbb{E}[Y_{obs} - (1 - T)Y_0|T = 1] =$
- $\mathbb{E}[Y_{obs}|T = 1] - \mathbb{E}[(1 - T)Y_0|T = 1] =$
- $\mathbb{E}[Y_{obs}|T = 1]$

Can be estimated from data

Assume that treatment is random

- Treatment is random:

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

- $\mathbb{E}[Y_1] =$
- $\mathbb{E}[Y_1|T = 1] =$
- $\mathbb{E}[Y_{obs}|T = 1]$

Can be estimated from data

- Treatment is random:

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

- $\mathbb{E}[Y_0] =$
- $\mathbb{E}[Y_0|T = 0] =$
- $\mathbb{E}[Y_{obs}|T = 0]$

Can be estimated from data

Assume that treatment is completely random

- Treatment is random:

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

- $\mathbb{E}[Y_1] =$
- $\mathbb{E}[Y_1|T = 1] =$
- $\mathbb{E}[Y_{obs}|T = 1]$

Can be estimated from data

- Treatment is random:

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

- $\mathbb{E}[Y_0] =$
- $\mathbb{E}[Y_0|T = 0] =$
- $\mathbb{E}[Y_{obs}|T = 0]$

Can be estimated from data

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

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

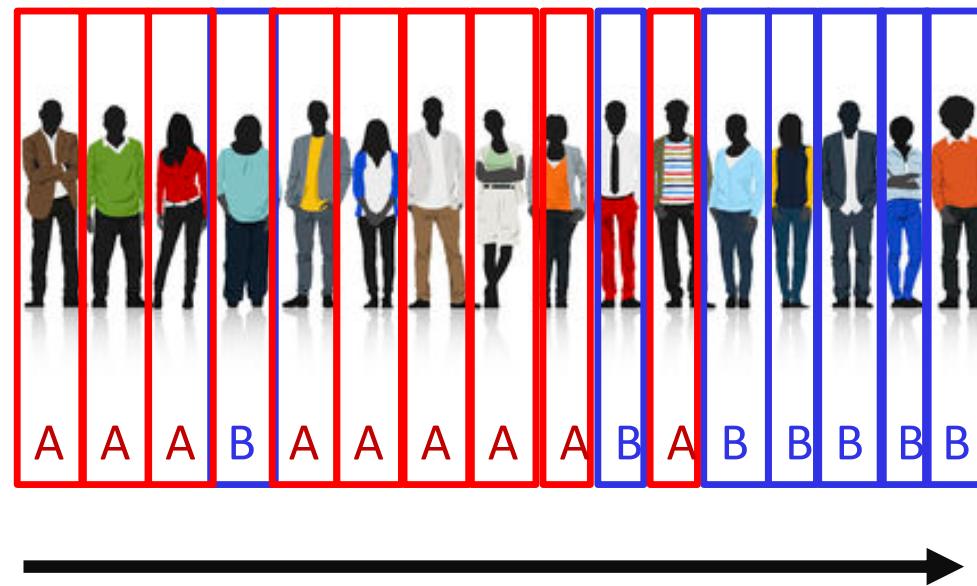
$$\mathbb{E}[Y_{obs}|T = 1] - \mathbb{E}[Y_{obs}|T = 0]$$

Under complete randomization:

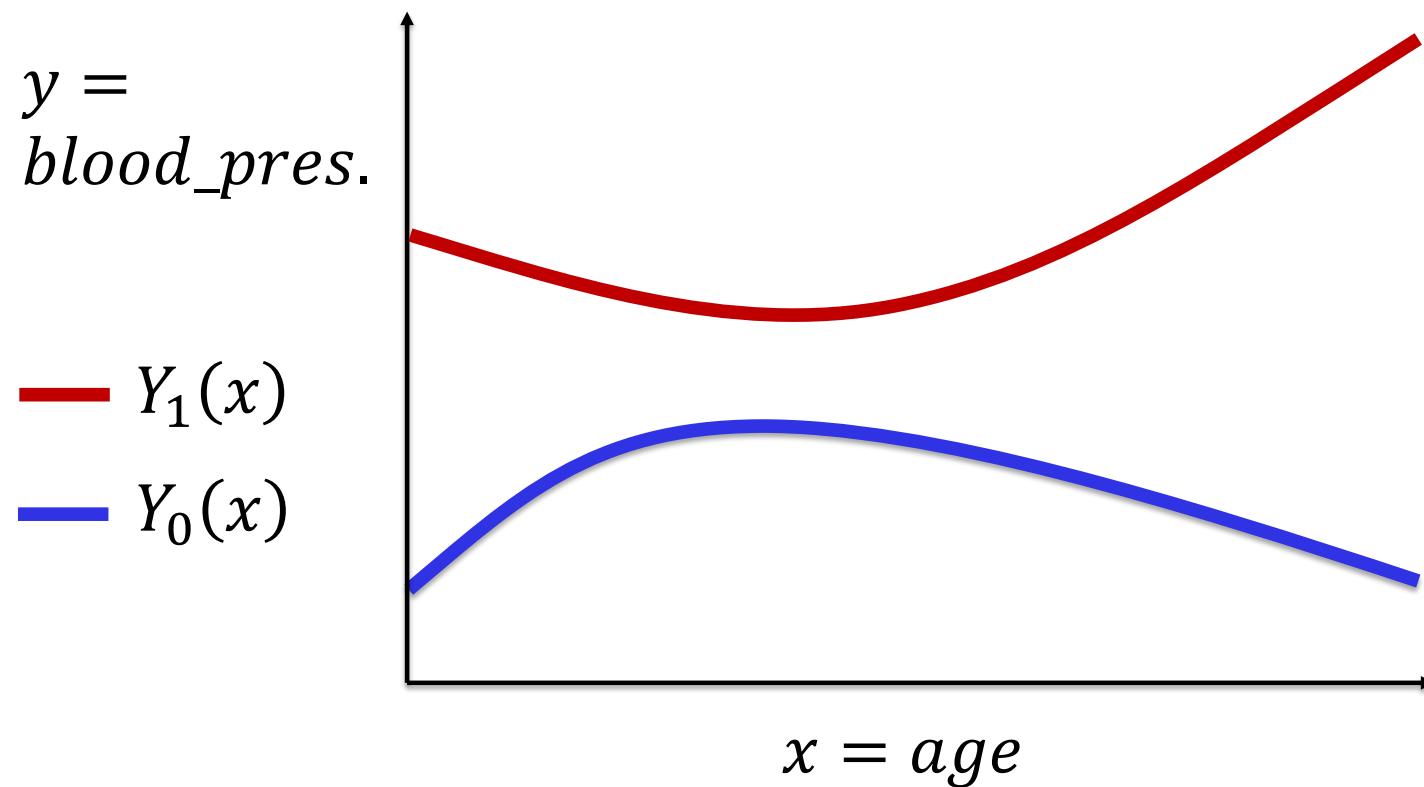
$$\begin{aligned}ATE &= \mathbb{E}[Y_1 - Y_0] = \\&\mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \\&\mathbb{E}[Y_{obs}|T = 1] - \mathbb{E}[Y_{obs}|T = 0]\end{aligned}$$

Note the difference between
unobservable quantities (potential outcomes)
and
observable quantities

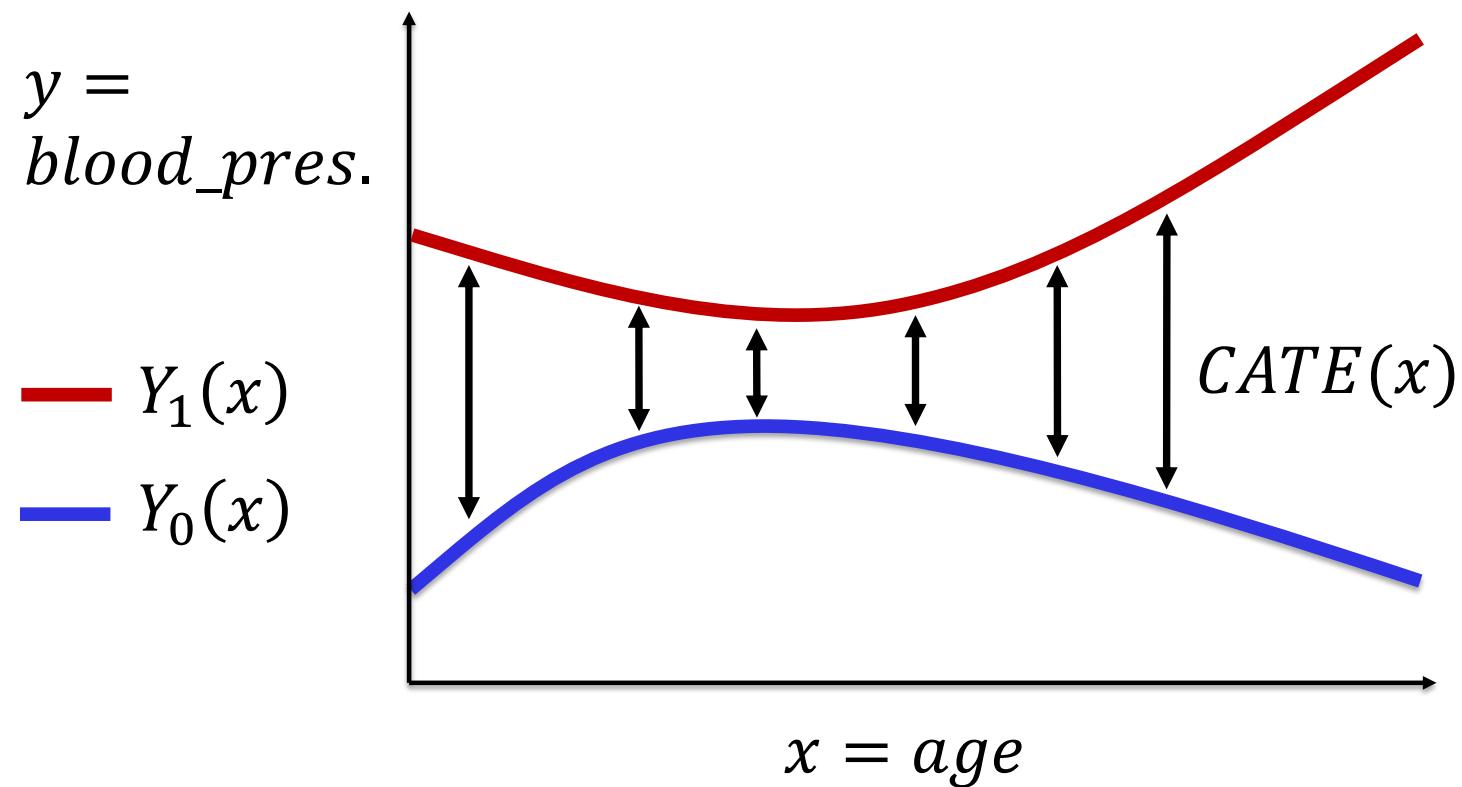
What happens when treatment isn't randomized?



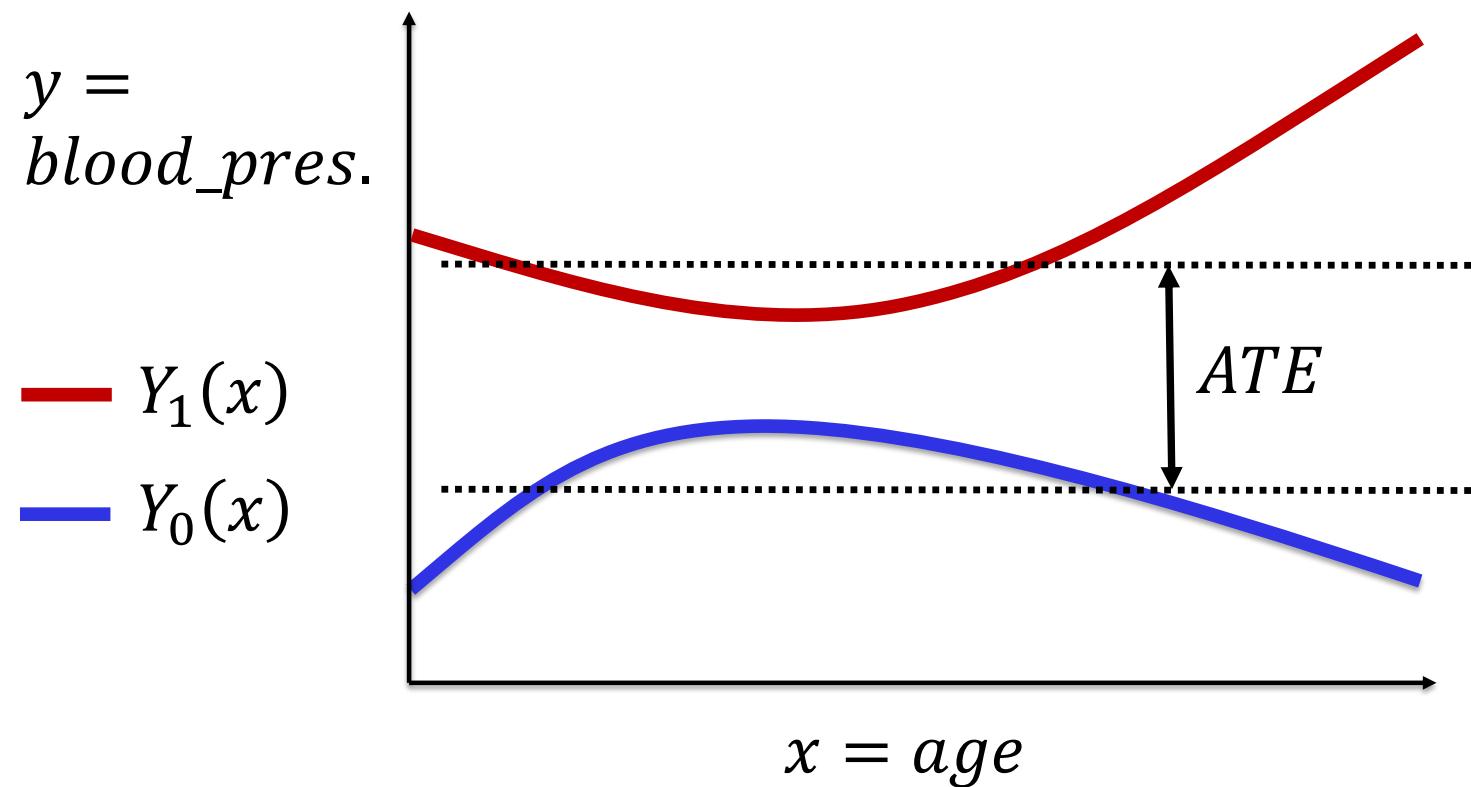
Blood pressure and age



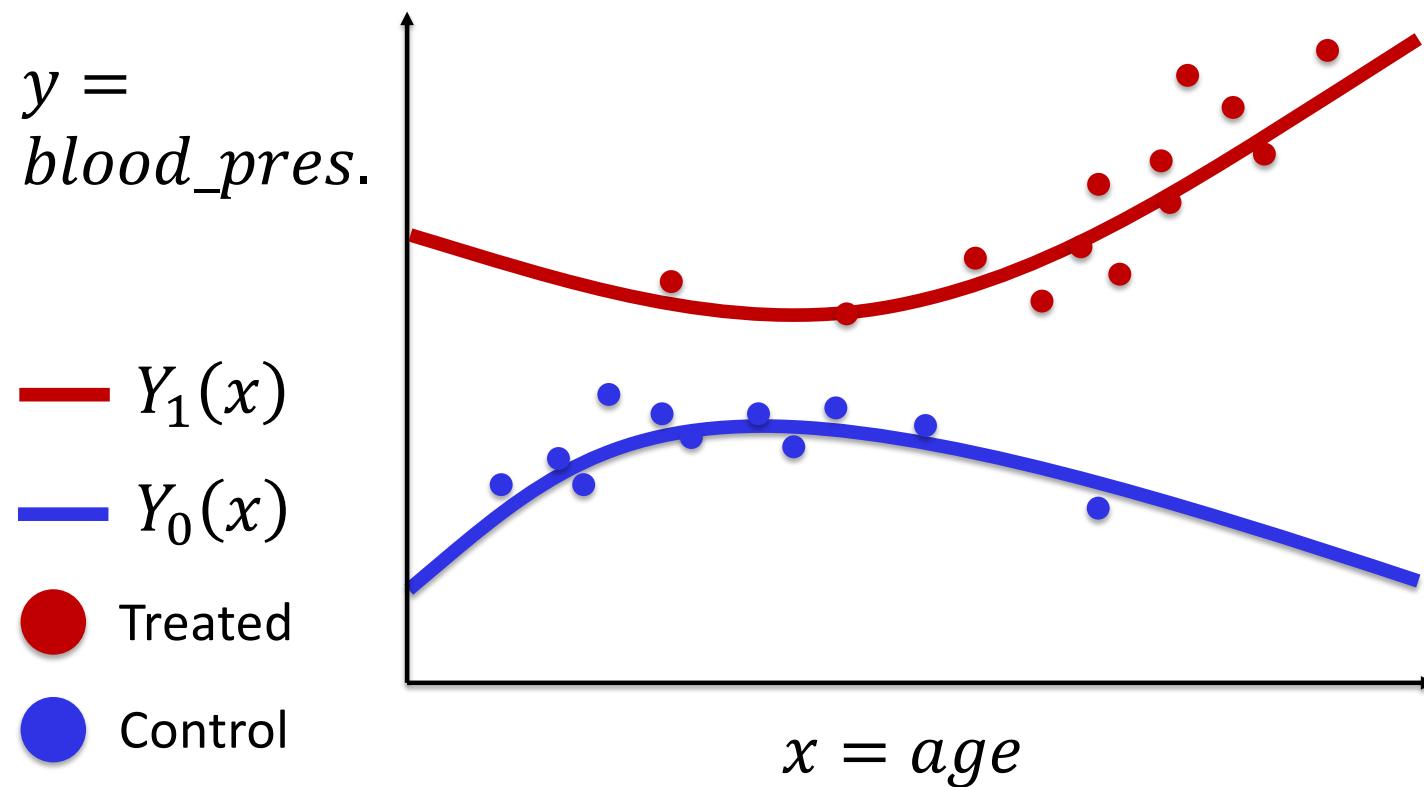
Blood pressure and age



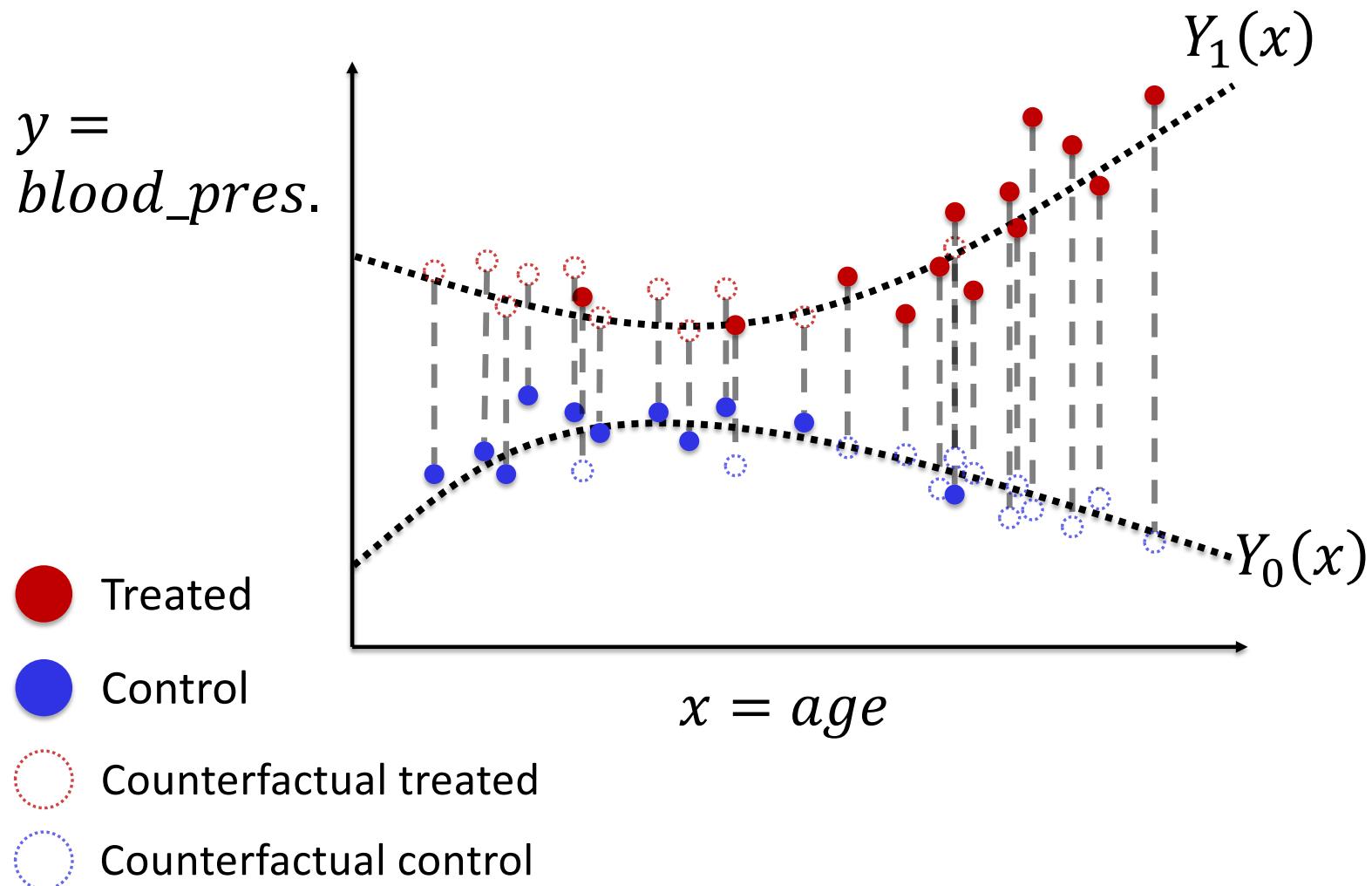
Blood pressure and age



Blood pressure and age



Blood pressure and age



Estimation

True treatment effect:

$$\mathbb{E}[Y_1 - Y_0] = 2$$

Gender	Treatment	Y_0 : Sugar levels <i>had they received treatment</i>	Y_1 : Sugar levels <i>had they received treatment</i>	Y : Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Estimation

True treatment effect:

$$\mathbb{E}[Y_1 - Y_0] = 2$$

$$\mathbb{E}[Y|t = 1] - \mathbb{E}[Y|t = 0] =$$

$$\frac{1}{4}(10 + 6 + 6 + 6) +$$

$$\frac{1}{4}(8 + 8 + 8 + 4) =$$

$$7 - 7 = 0$$

Gender	Treatment	Y_0 : Sugar levels <i>had they received treatment</i>	Y_1 : Sugar levels <i>had they received treatment</i>	Y : Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Estimation

True treatment effect:

$$\mathbb{E}[Y_1 - Y_0] = 2$$

$$\mathbb{E}[Y|t = 1] = 7$$

$$\mathbb{E}[Y|t = 0] = 7$$

Gender	Treatment	Y_0 : Sugar levels <i>had they received treatment</i>	Y_1 : Sugar levels <i>had they received treatment</i>	Y : Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Estimation

True treatment effect:

$$\mathbb{E}[Y_1 - Y_0] = 2$$

$$\mathbb{E}[Y|t = 1] = 7$$

$$\mathbb{E}[Y|t = 0] = 7$$

$$\mathbb{E}[Y|t = 0, \text{Gender} = M] = 8$$

$$\mathbb{E}[Y|t = 1, \text{Gender} = M] = 10$$

$$\mathbb{E}[Y|t = 0, \text{Gender} = F] = 4$$

$$\mathbb{E}[Y|t = 1, \text{Gender} = F] = 6$$

Gender	Treatment	Y_0 : Sugar levels <i>had they received treatment</i>	Y_1 : Sugar levels <i>had they received treatment</i>	Y : Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Estimation

True treatment effect:

$$\mathbb{E}[Y_1 - Y_0] = 2$$

$$\mathbb{E}[Y|t = 1] = 7$$

$$\mathbb{E}[Y|t = 0] = 7$$

$$\mathbb{E}[Y|t = 0, \text{Gender} = M] = 8$$

$$\mathbb{E}[Y|t = 1, \text{Gender} = M] = 10$$

$$\mathbb{E}[Y|t = 0, \text{Gender} = F] = 4$$

$$\mathbb{E}[Y|t = 1, \text{Gender} = F] = 6$$

Gender	Treatment	Y_0 : Sugar levels <i>had they received treatment</i>	Y_1 : Sugar levels <i>had they received treatment</i>	Y : Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Within each *group* we get
the true treatment effect

Treatment assignment mechanism

- $G=0$ if gender=F,
 $G=1$ if gender=M

$$Y_0 = 4 + 4 * G$$

$$Y_1 = 4 + 4 * G + 2$$

Gender	Treatment	Y_0 : Sugar levels <i>had they received treatment</i>	Y_1 : Sugar levels <i>had they received treatment</i>	Y : Observed sugar levels
		0	1	
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Treatment assignment mechanism

- $G=0$ if gender=F,
 $G=1$ if gender=M

$$Y_0 = 4 + 4 * G$$

$$Y_1 = 4 + 4 * G + 2$$

- $p(t=1 | G=1) = 0.25$
 $p(t=1 | G=0) = 0.75$

Gender	Treatment	Y_0 : Sugar levels <i>had they received treatment</i>	Y_1 : Sugar levels <i>had they received treatment</i>	Y : Observed sugar levels
		0	1	
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Treatment assignment mechanism

Gender	Treatment	Y_0 : Sugar levels <i>had they received treatment 0</i>	Y_1 : Sugar levels <i>had they received treatment 1</i>	Y: Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

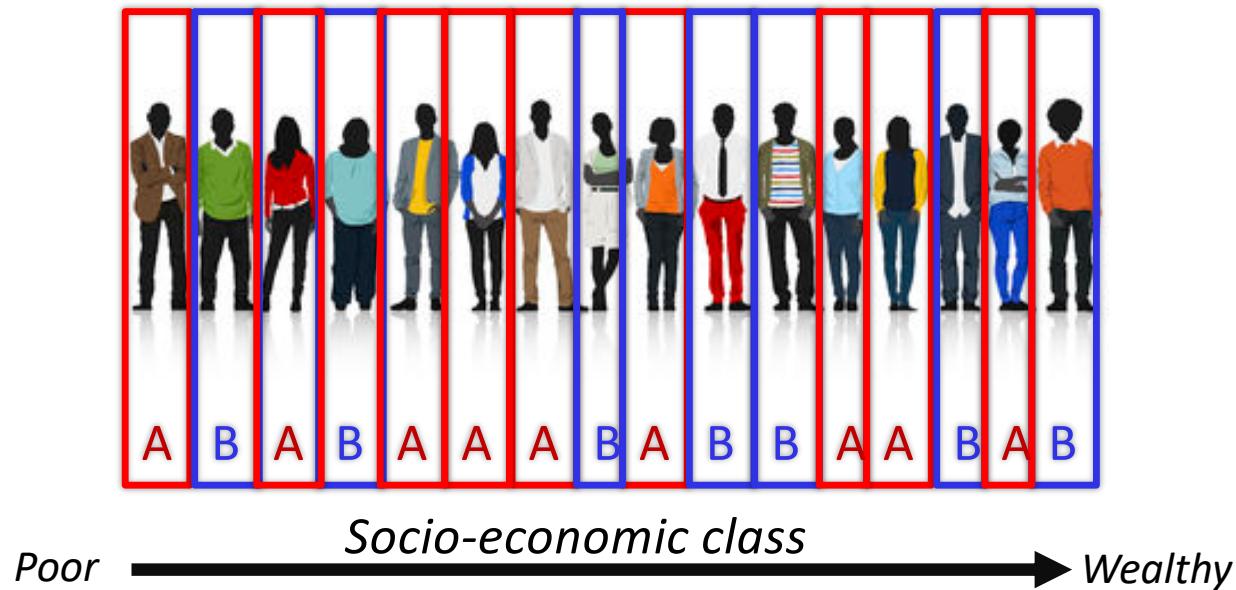
Treatment assignment mechanism

Gender	Treatment	Y_0 : Sugar levels <i>had they received treatment 0</i>	Y_1 : Sugar levels <i>had they received treatment 1</i>	Y: Observed sugar levels
M	0	8	?	8
M	0	8	?	8
M	0	8	?	8
M	1	?	10	10
F	0	4	?	4
F	1	?	6	6
F	1	?	6	6
F	1	?	6	6

Treatment assignment mechanism

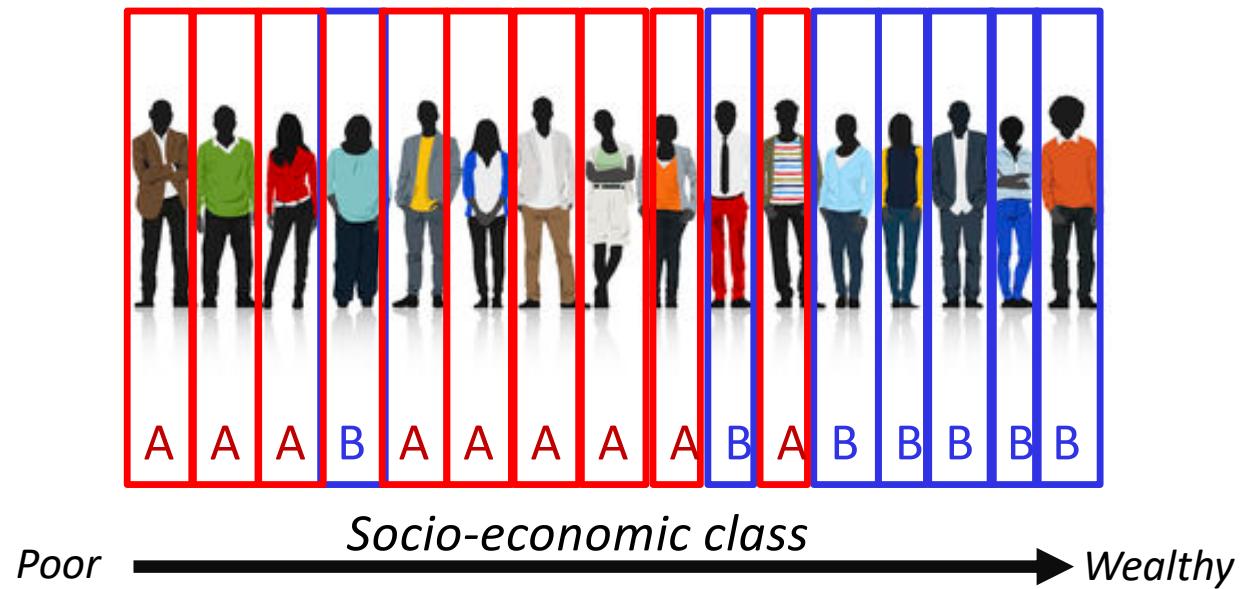
Gender	Treatment	Y_0 : Sugar levels <i>had they received treatment 0</i>	Y_1 : Sugar levels <i>had they received treatment 1</i>	Y: Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Randomized controlled trial (RCT)



treatment
A or B

Observational study



treatment
A or B

$$P(Y_0 = 8|T = 0) = 0.75$$

$$P(Y_0 = 8|T = 1) = 0.25$$

$$P(Y_1 = 10|T = 0) = 0.75$$

$$P(Y_1 = 10|T = 1) = 0.25$$

(Y_0, Y_1) **are not**
independent of T

Gender	T: Treatment	Y_0 : Sugar levels <i>had they received treatment</i>	Y_1 : Sugar levels <i>had they received treatment</i>	Y: Observed sugar levels
		0	1	
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

$$P(Y_0 = 8|T = 0, G = M) = 1$$

$$P(Y_0 = 8|T = 1, G = M) = 1$$

$$P(Y_1 = 10|T = 0, G = M) = 1$$

$$P(Y_1 = 10|T = 1, G = M) = 1$$

(Y_0, Y_1) **are independent** of T
conditioned on

$G=M$, and conditioned on $G=F$

Gender	T: Treatment	Y_0 : Sugar levels <i>had they received treatment</i>	Y_1 : Sugar levels <i>had they received treatment</i>	Y: Observed sugar levels
		0	1	
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

$$\begin{aligned}
 P(Y_0 = 4|T = 0, G = F) &= 1 \\
 P(Y_0 = 4|T = 1, G = F) &= 1 \\
 P(Y_1 = 6|T = 0, G = F) &= 1 \\
 P(Y_1 = 6|T = 1, G = F) &= 1
 \end{aligned}$$

(Y_0, Y_1) **are independent** of T **conditioned** on
 $G=M$, and conditioned on $G=F$

Gender	T: Treatment	Y ₀ : Sugar levels <i>had they received treatment</i>		Y: Observed sugar levels
		0	1	
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

$$\begin{aligned}
 P(Y_0 = 4|T = 0, G = F) &= 1 \\
 P(Y_0 = 4|T = 1, G = F) &= 1 \\
 P(Y_1 = 6|T = 0, G = F) &= 1 \\
 P(Y_1 = 6|T = 1, G = F) &= 1
 \end{aligned}$$

(Y_0, Y_1) **are independent** of T **conditioned** on
 $G=M$, and conditioned on $G=F$

$$(Y_0, Y_1) \perp\!\!\!\perp T | G$$

Gender	T: Treatment	Y_0 : Sugar levels <i>had they received treatment</i>		Y : Observed sugar levels
		0	1	
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

What if we can't randomize treatment?

- We can still succeed if the treatment assignment process is *conditionally randomized*, conditioned on an observed quantity
- This is actually just a way of saying we have *no unmeasured confounding*

What assumptions are sufficient for us to identify the causal effect in an observational study?

The target trial

- We have data collected where treatment was not explicitly randomized
- We can view it as a “flawed” randomized trial
- Are the flaws fixable?

Identifying Assumptions

- We will describe today a set of assumptions that will go with us through a big part of the course
- A big part of causal inference is understanding when these assumptions are plausible
- There is active research into relaxing each and every one of these assumptions
- We will see other “assumption sets” later in the course (e.g. instrumental variables, front-door adjustment)

“The Assumptions”

Sufficient conditions for causal inference to be possible:

- 1. Stable Unit Treatment Value Assumption**
- 2. Consistency**
- 3. *Ignorability / No unmeasured confounders***
- 4. *Common support***

Stable Unit Treatment Value Assumption SUTVA

- 1. The potential outcomes for any unit do not vary with the treatments assigned to other units*

Stable Unit Treatment Value Assumption SUTVA

1. *The potential outcomes for any unit do not vary with the treatments assigned to other units*
failure example: vaccination, network effects

Stable Unit Treatment Value Assumption SUTVA

- 1. The potential outcomes for any unit do not vary with the treatments assigned to other units*
failure example: vaccination, network effects
- 2. For each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes*

Stable Unit Treatment Value Assumption SUTVA

- 1. The potential outcomes for any unit do not vary with the treatments assigned to other units*
failure example: vaccination, network effects
- 2. For each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes*
failure example: some people get out-of-date medication

Consistency

- For a unit that receives treatment t , we observe the corresponding potential outcome Y_t

Ignorability – no unmeasured confounders

Y_0, Y_1 : potential outcomes for control and treated

X : observed unit covariates (features)

T : treatment assignment

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

The potential outcomes are independent of treatment assignment, conditioned on observed covariates x

Ignorability – no unmeasured confounders

Y_0, Y_1 : potential outcomes for control and treated

X : observed unit covariates (features)

T: treatment assignment

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

The potential outcomes are independent of treatment assignment, conditioned on observed covariates x

Weird! The outcomes obviously depend on treatment.
How can potential outcomes not depend?

Ignorability – no unmeasured confounders

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

The potential outcomes are independent of treatment assignment, conditioned on observed covariates x

Within each “strata/level” $X = x$, treatment assignment is as good as random

Ignorability – no unmeasured confounders

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

The potential outcomes are independent of treatment assignment, conditioned on observed covariates x

Within each “strata/level” $X = x$, treatment assignment is as good as random with respect to Y_0, Y_1

Failure example: treatment depends on gender, which also has an effect on the potential outcomes

Common support

Y_0, Y_1 : potential outcomes for control and treated

x : unit covariates (features)

T : treatment assignment

We assume:

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

Common support

Y_0, Y_1 : potential outcomes for control and treated

x : unit covariates (features)

T : treatment assignment

We assume:

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

Failure:

if only women receive $T=1$,
and only men receive $T=0$

Ignorability assumption is unverifiable from data!

- Remember, we never observe (Y_1^i, Y_0^i) jointly
- How do we know it holds in an RCT?

Checking the assumptions: Ignorability assumption is unverifiable from data!

- How can we convince ourselves that it is true in a given case?
- Confounders: factors that affect **both** treatment assignment and outcome
- Talk to domain experts, understand what determines treatment assignment and outcomes
- Sensitivity analysis
- Do you believe ignorability holds? If not - change the design:
 - Add relevant variables
 - Define or measure treatment differently
 - Define or measure outcome differently

Checking the assumptions - example

- Comparing effectiveness of two anti-hypertensive medications
- Treatment: first administration of medication
- Outcome: blood pressure 3 months after first treatment
 - Is outcome only measured for some of the patients?

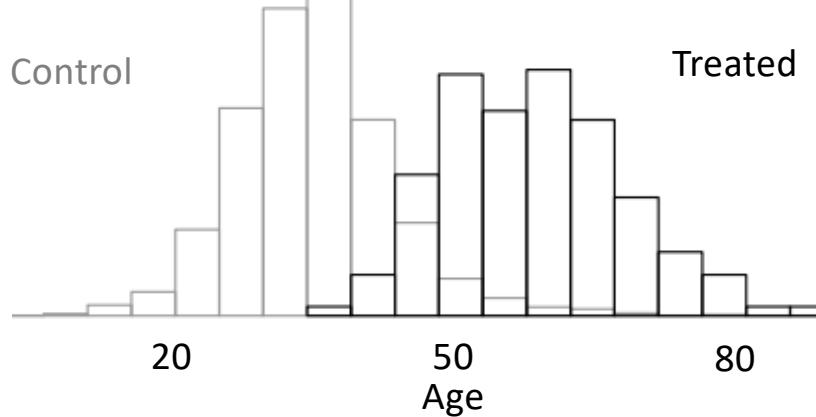
Checking the assumptions - ignorability

- Did we measure the important known causes of hypertension? Literature survey
- Example: high alcohol use is known to be a cause of hypertension
- Doctors know this, and might use this information in deciding on treatment
- If we don't measure alcohol use, it becomes hidden confounder which might bias our conclusions
- Talking to doctors to understand how they prescribe treatment

Checking the assumptions – common support

- Check for common support between treated and control:
 - Reduce dimension and plot populations
 - Check overlap on important univariate and bivariate variables, e.g. age, gender, weight in a medical study

Figure:
Hill & Gelman

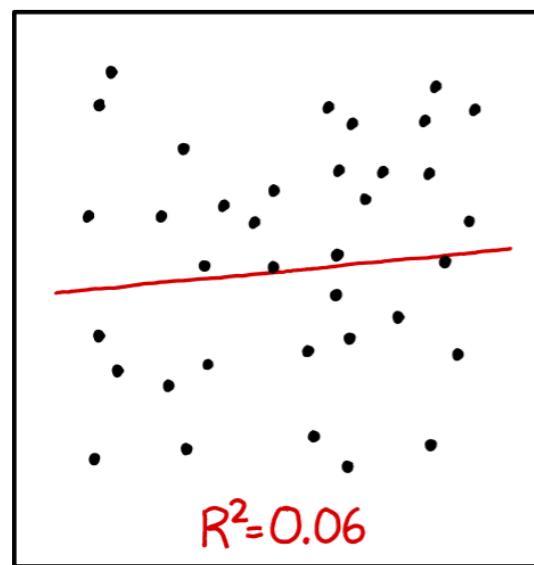


Checking the assumptions – common support

- If no common support:
 - Redefine study population, e.g. only people ages 40-60
 - More risky: check if outcome is sensitive to unbalanced variables.

Example: treated and control might differ in commute distance from hospital, but distance from hospital is not associated with any important socio-economic factors and has no observed association with outcome

How can we estimate causal
effects under ignorability and
common support?



Average Treatment Effect – the adjustment formula

- Assuming ignorability, we will derive the *covariate-adjustment formula*
- The adjustment formula is **extremely** useful in causal inference
- Strongly related to *G-formula* and *back-door adjustment*

Average Treatment Effect

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

Average Treatment Effect

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

$$\mathbb{E}[Y_1] =$$

Average Treatment Effect

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

law of total
expectation

$$\mathbb{E}[Y_1] =$$
$$\mathbb{E}_{x \sim p(x)} [\mathbb{E}_{Y_1 \sim p(Y_1|x)} [Y_1|x]] =$$

Average Treatment Effect

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

$$\mathbb{E}[Y_1] =$$

$$\mathbb{E}_{x \sim p(x)} [\mathbb{E}_{Y_1 \sim p(Y_1|x)} [Y_1|x]] = \begin{array}{l} \text{ignorability} \\ (Y_0, Y_1) \perp\!\!\!\perp T | x \end{array}$$

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

Average Treatment Effect

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

$$\mathbb{E}[Y_1] =$$

$$\mathbb{E}_{x \sim p(x)} [\mathbb{E}_{Y_1 \sim p(Y_1|x)} [Y_1|x]] =$$

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

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

shorter notation

Average Treatment Effect

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

$$\mathbb{E}[Y_1] =$$

$$\mathbb{E}_{x \sim p(x)} [\mathbb{E}_{Y_1 \sim p(Y_1|x)} [Y_1|x]] =$$

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

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

$$\mathbb{E}_{x \sim p(x)} [\mathbb{E}[Y|x, T=1]] \quad \text{consistency}$$

Average Treatment Effect

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

$$\mathbb{E} [Y_1] =$$

$$\mathbb{E}_{x \sim p(x)} [\mathbb{E}_{Y_1 \sim p(Y_1|x)} [Y_1|x]] =$$

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

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

$$\mathbb{E}_{x \sim p(x)} [\mathbb{E} [Y|x, T=1]]$$

Might be
estimated from
data!

Average Treatment Effect

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

$$\mathbb{E} [Y_0] =$$

$$\mathbb{E}_{x \sim p(x)} [\mathbb{E}_{Y_0 \sim p(Y_0|x)} [Y_0|x]] =$$

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

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

$$\mathbb{E}_{x \sim p(x)} [\mathbb{E} [Y|x, T=0]]$$

Might be
estimated from
data!

The adjustment formula

Under the assumption of ignorability,
we have that:

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

$$\left. \begin{array}{l} \mathbb{E}[Y|x, T=1] \\ \mathbb{E}[Y|x, T=0] \end{array} \right\} \quad \begin{array}{l} \text{Quantities we} \\ \text{can hope to} \\ \text{estimate} \\ \text{from data} \end{array}$$

The adjustment formula

Under the assumption of ignorability,
we have that:

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

Empirically we have samples from
 $p(x|T=1)$ or $p(x|T=0)$

The adjustment formula

Under the assumption of ignorability,
we have that:

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

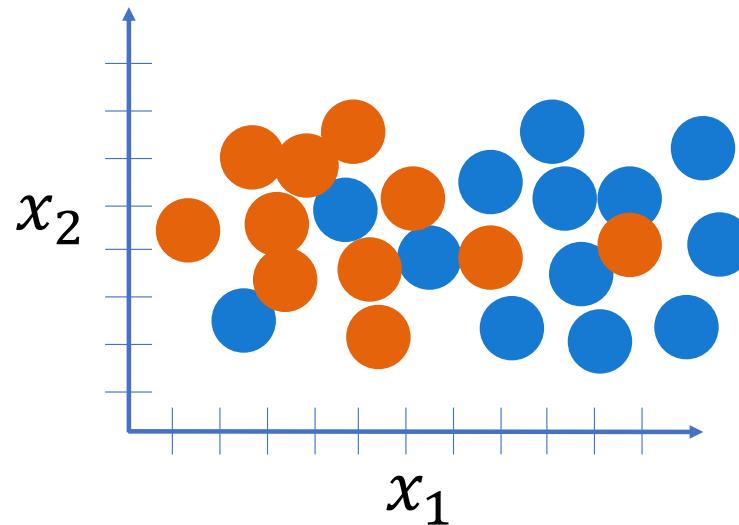
Empirically we have samples from
 $p(x|T=1)$ or $p(x|T=0)$

To extrapolate to $p(x) \rightarrow \text{overlap}$ assumption

When is estimating treatment effect harder?

Observational study

Treatment assignment non-random → counterfactual and factual have different distributions



- Control, $t = 0$
- Treated, $t = 1$

The adjustment formula

Under the assumption of ignorability,
we have that:

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

Empirically we have samples from
 $p(x|T=1)$ or $p(x|T=0)$

To extrapolate to $p(x) \rightarrow \text{overlap}$ assumption

Covariate adjustment (parametric g-formula)

- Explicitly model the relationship between treatment, confounders, and outcome
- Under ignorability, the expected causal effect of T on Y :
$$\mathbb{E}_{x \sim p(x)} [\text{red} \mathbb{E}[Y_1 | T = 1, x] - \text{blue} \mathbb{E}[Y_0 | T = 0, x]]$$
- Fit a model $f(x, t) \approx \mathbb{E}[Y_t | T = t, x]$

$$\widehat{ATE} = \frac{1}{n} \sum_{i=1}^n f(x_i, 1) - f(x_i, 0)$$

Covariate adjustment (parametric g-formula)

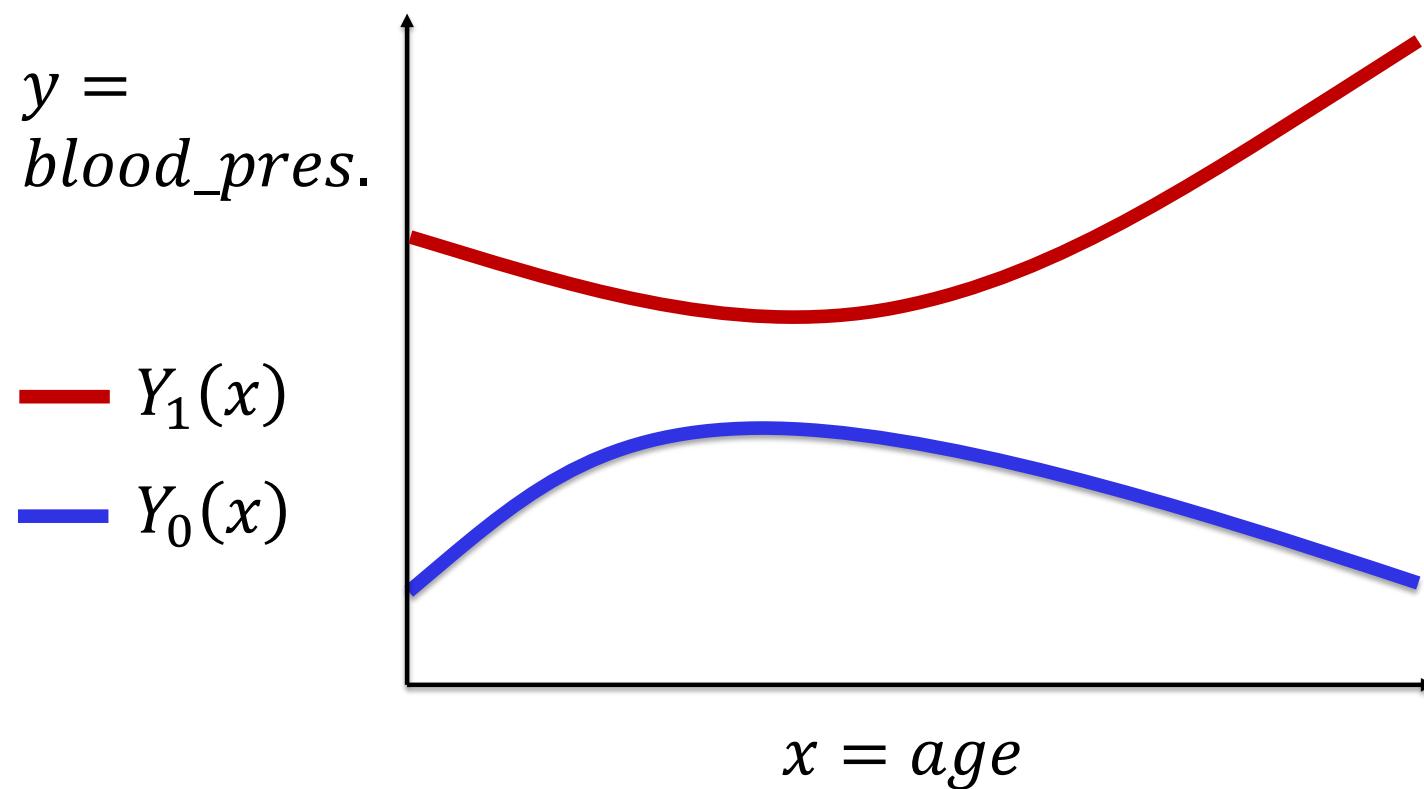
- Explicitly model the relationship between treatment, confounders, and outcome
- Under ignorability, the expected causal effect of T on Y :
$$\mathbb{E}_{x \sim p(x)} \left[\textcolor{red}{\mathbb{E}[Y_1 | T = 1, x]} - \textcolor{blue}{\mathbb{E}[Y_0 | T = 0, x]} \right]$$
- Fit a model $f(x, t) \approx \mathbb{E}[Y_t | T = t, x]$

$$\widehat{CATE}(x_i) = f(x_i, 1) - f(x_i, 0)$$

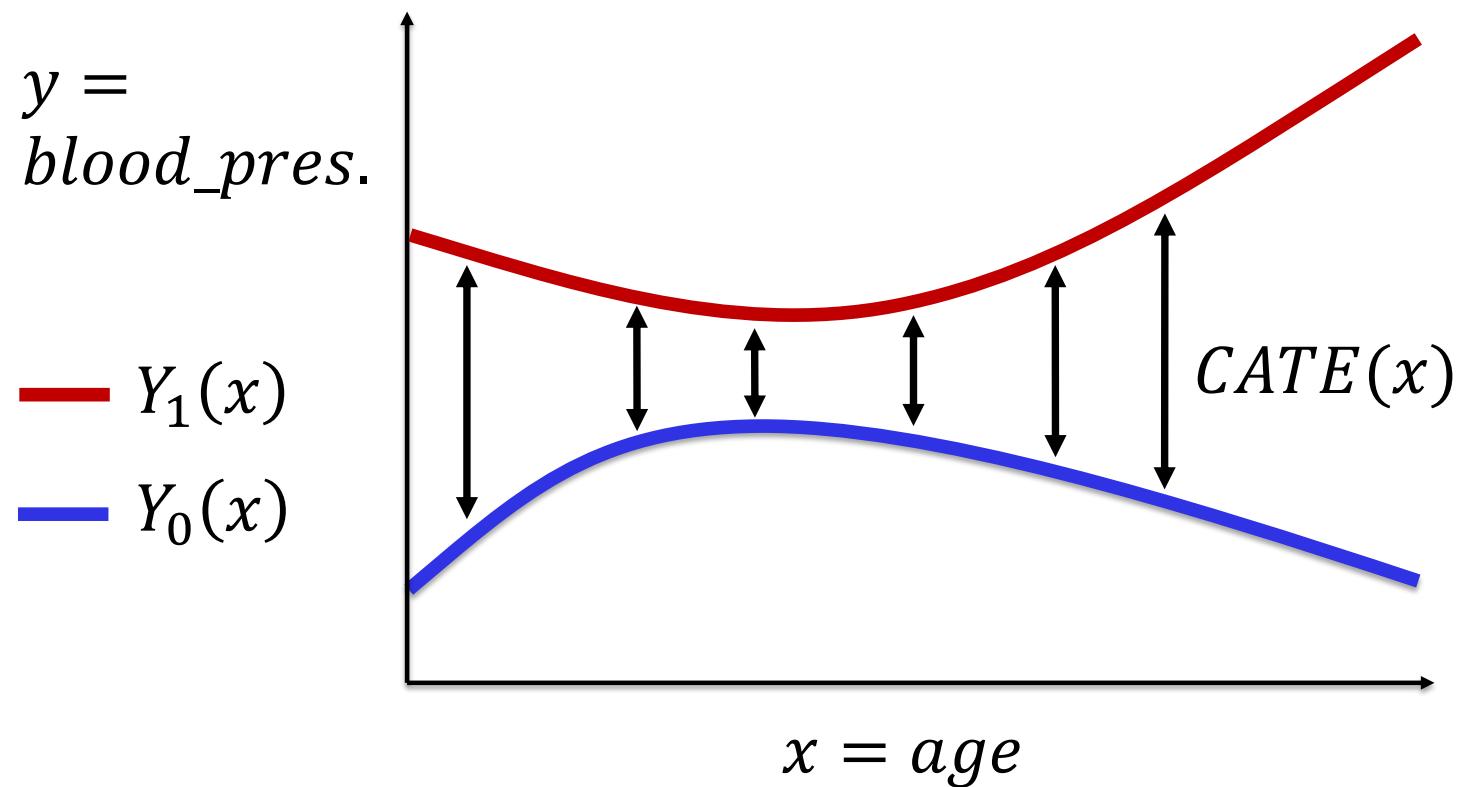
Covariate adjustment – model consistency (unrelated to our consistency assumption)

- If the model $f(x, t) \approx \mathbb{E}[Y_t | T = t, x]$ is consistent in the limit of infinite samples, then under ignorability the estimated \widehat{ATE} will converge to the true ATE
- A sufficient condition: overlap and well-specified model

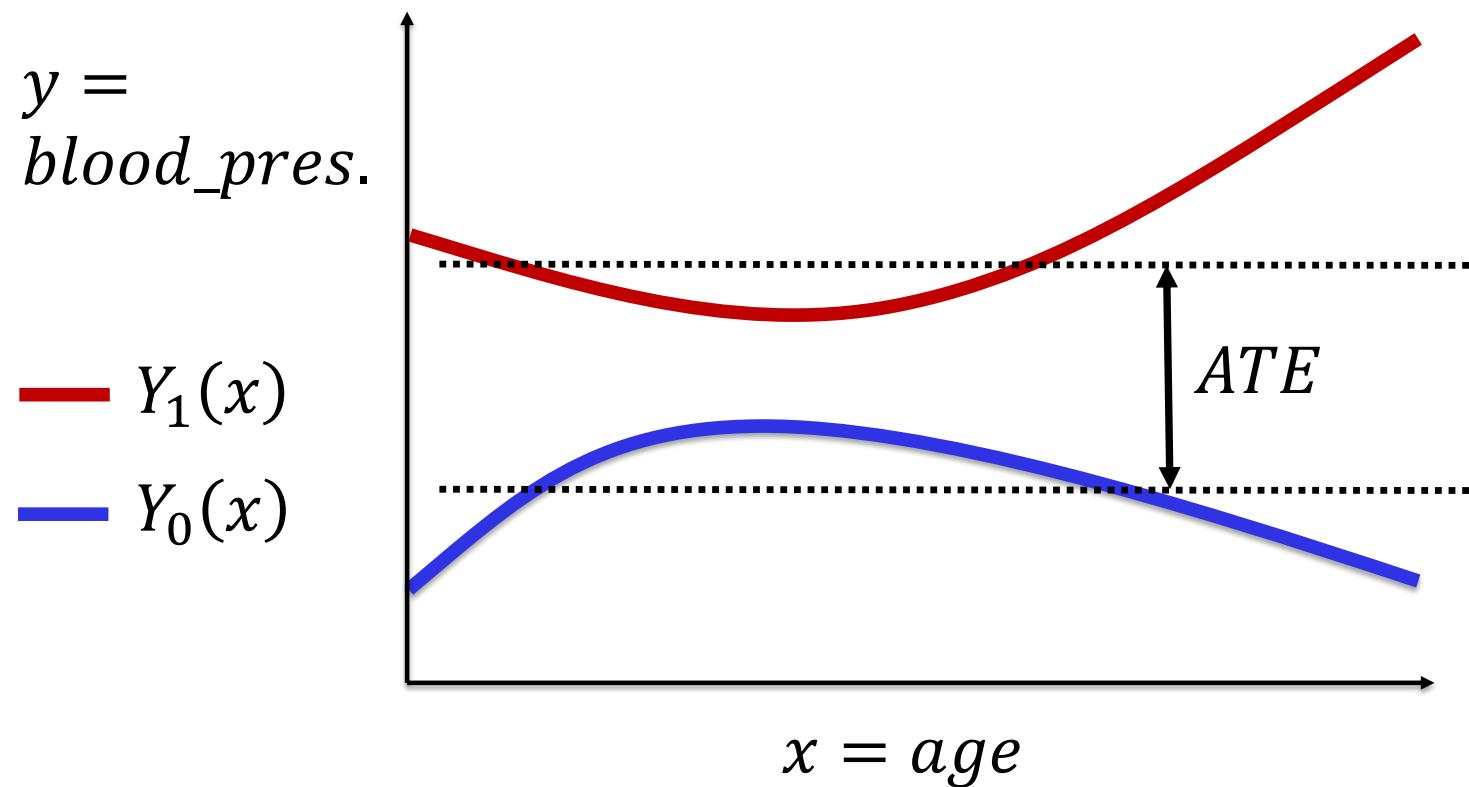
Covariate adjustment



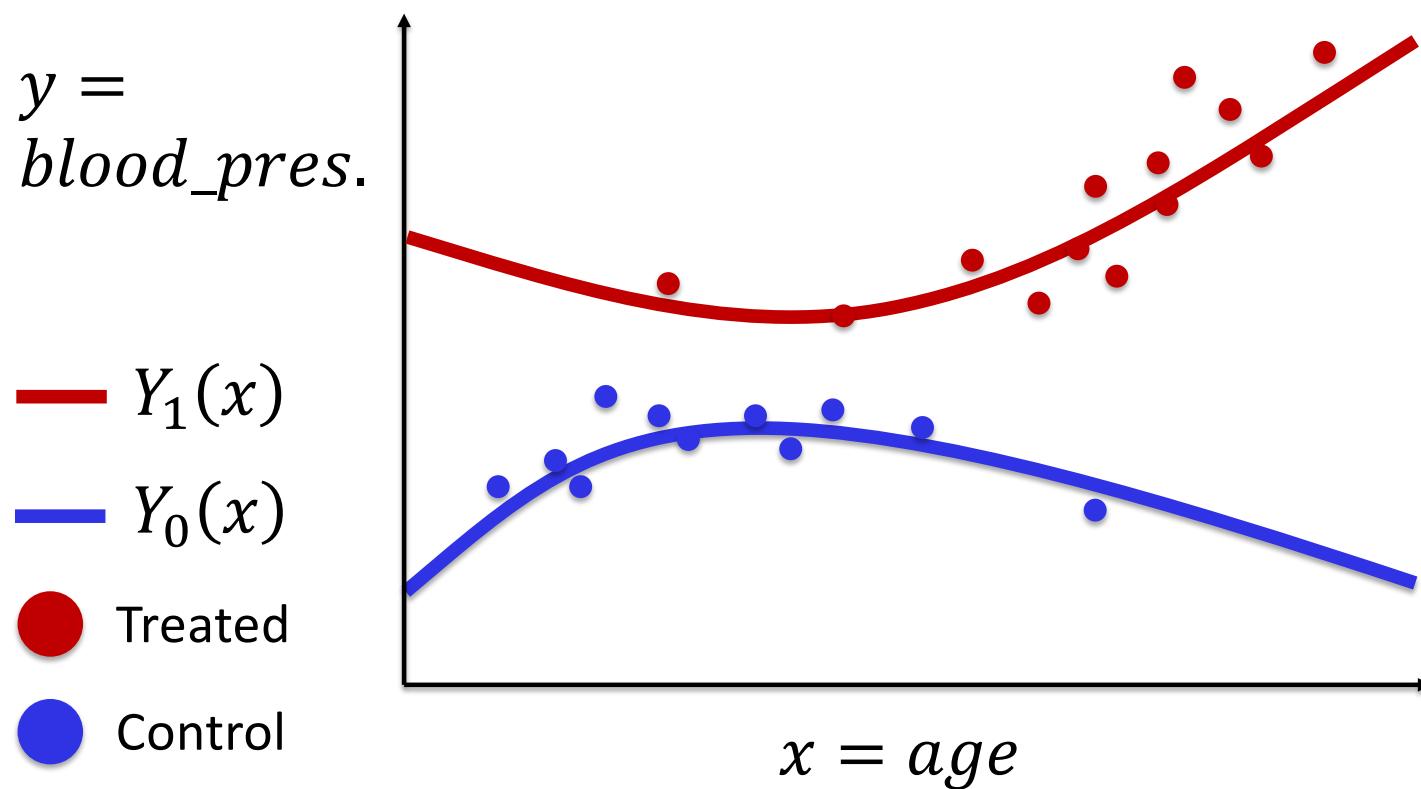
Covariate adjustment



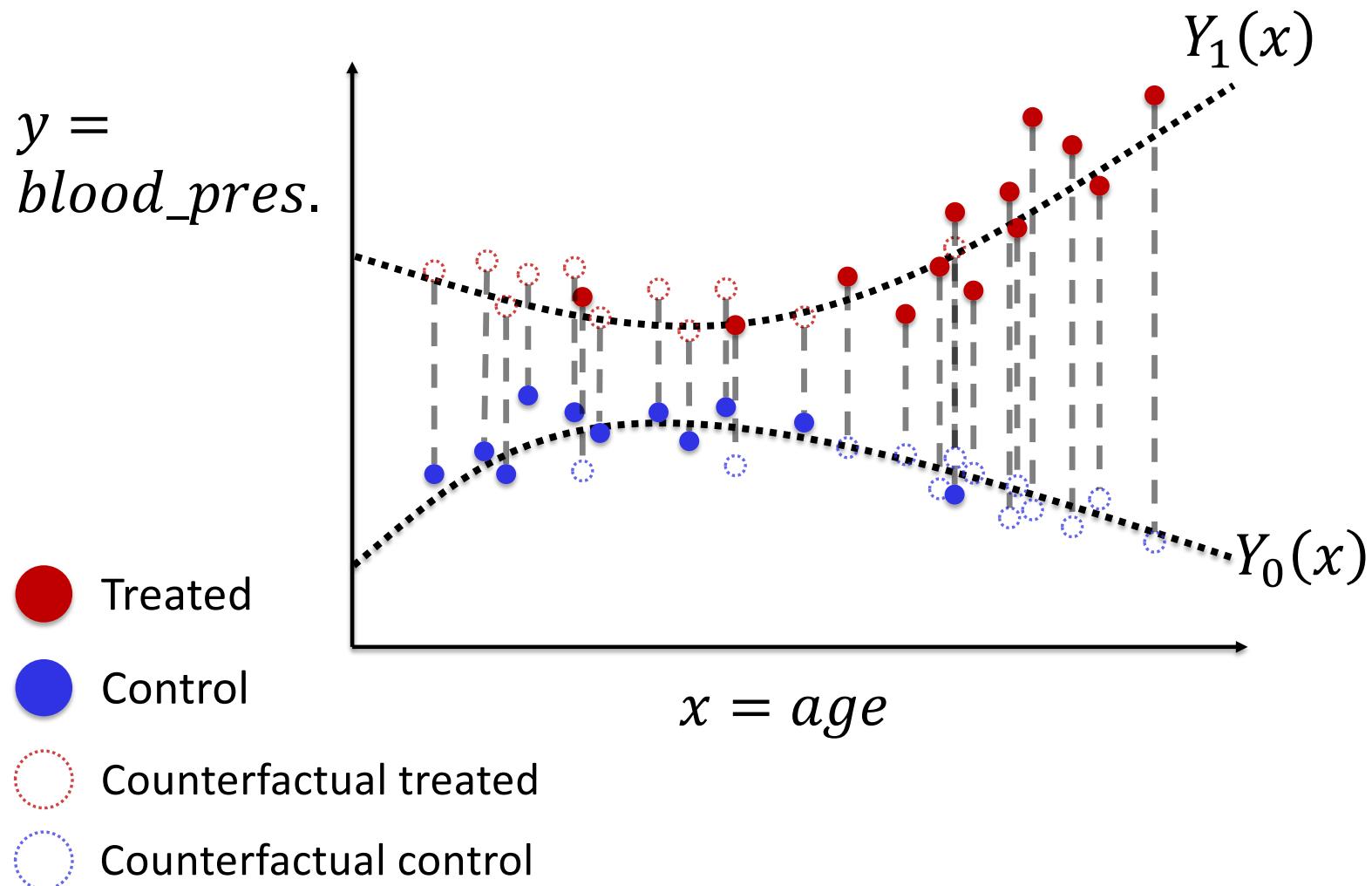
Covariate adjustment



Covariate adjustment



Covariate adjustment



Covariate adjustment in practice

- Given data:

$$(x^1, t^1, y^1), \dots, (x^n, t^n, y^n)$$

1. Linear: Regress y on (x, t) , read off coefficient for t (regularize?)
2. Non-linear: Regress y on (x, t)
3. Regress separately over set with $t=1$ and set with $t=0$?
4. Active field of research developing specialized methods

Covariate adjustment in practice

- Given data:

$$(x^1, t^1, y^1), \dots, (x^n, t^n, y^n)$$

Covariate adjustment in practice

- Given data:

$$(x^1, t^1, y^1), \dots, (x^n, t^n, y^n)$$

- Regress the observed y 's on (x, t) such that

$$y_i \approx f(x_i, t_i)$$
$$\widehat{ATE} = \frac{1}{n} \sum_{i=1 \dots n} f(x_i, 1) - f(x_i, 0)$$

Covariate adjustment in practice

- Given data:

$$(x^1, t^1, y^1), \dots, (x^n, t^n, y^n)$$

- Regress the observed y 's on (x, t) such that

$$y_i \approx f(x_i, t_i)$$

$$\widehat{ATE}_1 = \frac{1}{n} \sum_{i=1 \dots n} f(x_i, 1) - f(x_i, 0)$$

$$\widehat{ATE}_2 = \frac{1}{\sum t_i} \sum_{i \text{ s.t. } t_i=1} y_i - f(x_i, 0) + \frac{1}{\sum 1 - t_i} \sum_{i \text{ s.t. } t_i=1} f(x_i, 1) - y_i$$

Problems with covariate adjustment

- It's too easy!
- Should think thoroughly about what covariates go in – that is the most important decision (often more than which algorithm)
- Example of mistakes:
 - including post-treatment covariates (leads to zero-bias)
 - Including covariates that only influence treatment and not outcome (leads to high-variance)
 - Using a model not specified for causal inference (leads to statistical inefficiency)

Linear model

- Assume that:

$$\begin{array}{c} \text{blood pressure} \\ Y_t(x) = \beta^T x + \gamma \cdot t + \epsilon_t \\ \text{age, weight, ...} \\ \mathbb{E}[\epsilon_t] = 0 \\ \text{medication} \end{array}$$

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

- We care about γ , not about $Y_t(x)$

Estimation, not prediction

Linear model

blood pressure **age,weight,...** **medication**

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

Hypertension is affected by many variables:
lifestyle, weight, genetics, age

- Each of these often stronger **predictor** of blood-pressure, compared with type of medication taken
- Regularization (e.g. Lasso) might remove the treatment variable!
- Features → (“nuisance parameters”, “variable of interest”)

Regression - 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:

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

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

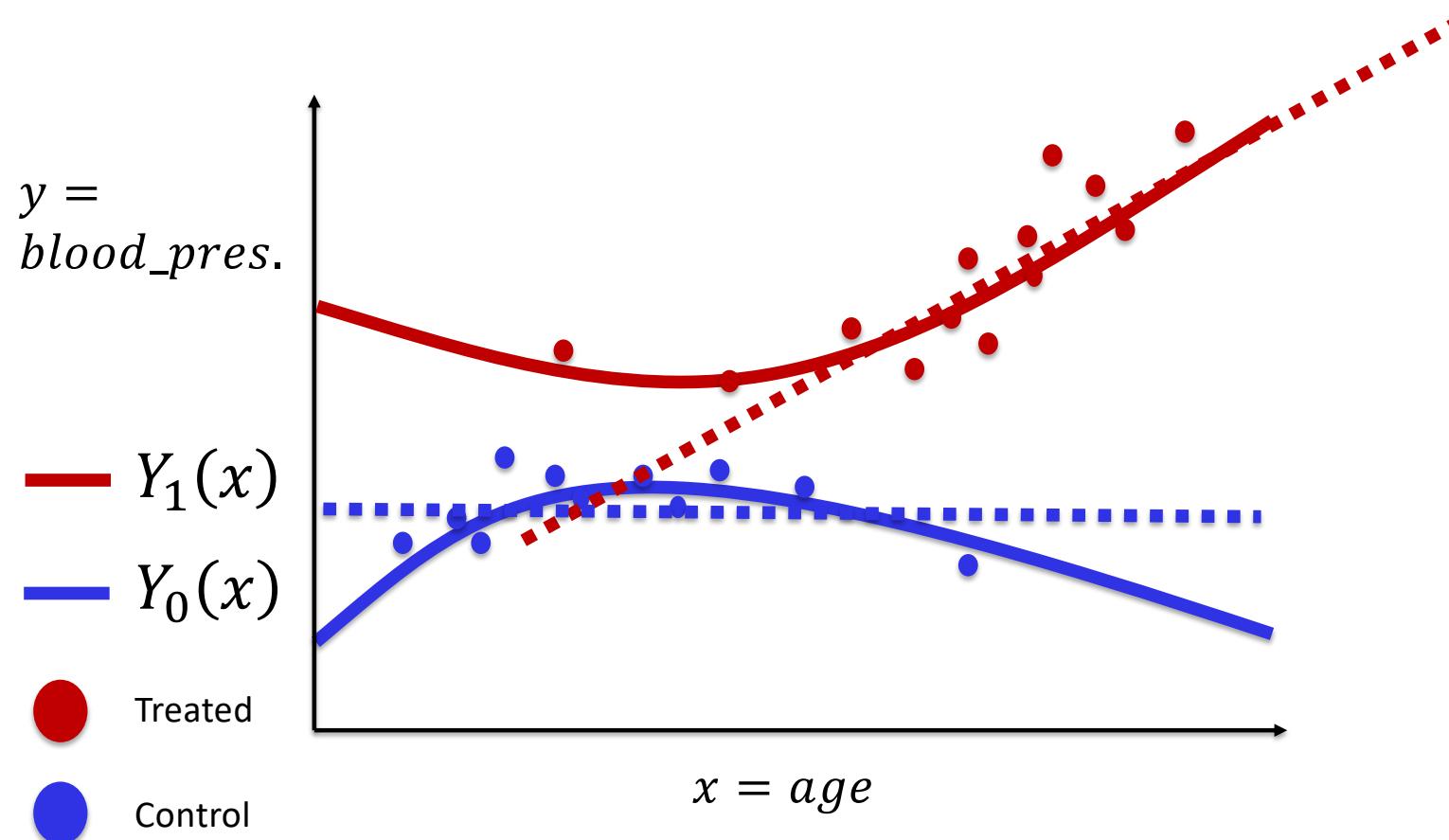
Using machine learning for causal inference

- Machine learning techniques can be very useful and have recently seen wider adoption
- Random forests and Bayesian trees
Hill (2011), Athey & Imbens (2015), Wager & Athey (2015)
- Gaussian processes
Hoyer et al. (2009), Zigler et al. (2012)
- Neural nets
Beck et al. (2000), Johansson et al. (2016), Shalit et al. (2016), Lopez-Paz et al. (2016)
- “Causal” Lasso
Belloni et al. (2013), Farrell (2015), Athey et al. (2016)

Using machine learning for causal inference

- Machine learning techniques can be very useful and have recently seen wider adoption
- How is the treatment variable used:
 - Fit two different models for treated and control?
 - Not regularized?
 - Privileged

Covariate adjustment: weak overlap



Problems with covariate adjustment

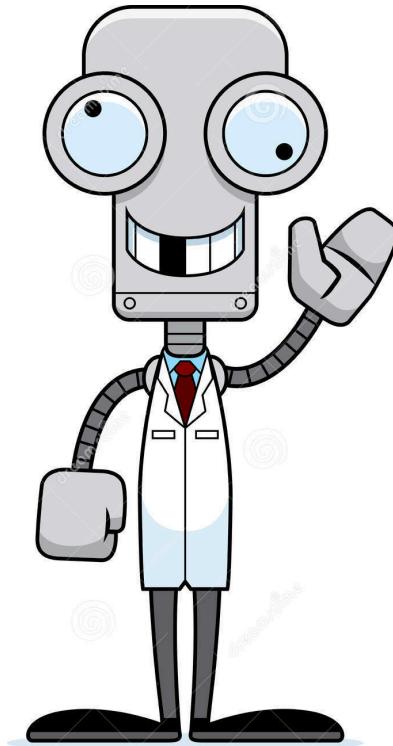
- It's too easy!
- Should think thoroughly about what covariates go in – that is the most important decision (more than which algorithm)
- Example of mistake: including post-treatment covariates

Example: post-treatment mistake

- Did we include a post-treatment covariate?
example: weight measured after treatment
- Measuring the post-treatment weight could explain away some of the causal effect, inducing bias in our estimation of the causal effect
- Conditioning on post-treatment covariates violates ignorability
- But how do we know if a variable is post-treatment?
- We will look into this issue again further in the course

Often very easy to run a regression / fit a model, and get the causal aspects wrong

A possible solution: “Target Trial” thinking





Introduction to Causal Inference - a Machine Learning Perspective

Dr. Uri Shalit

Course number 097400
2020-2021

Lesson 2b

The stages of causal inference

1. Formulate *causal assumptions* sufficient to solve the problem
 - these are mostly **untestable**
2. Under the assumptions, reduce causal problem to appropriate statistical/machine learning method
 - these methods are often specialized methods, similar but distinct from familiar methods such as regression

Identification

Estimation

Potential outcomes

- Unit: a person, a bacteria, a company, a school, a website, a family, a piece of metal, ...
- Treatments / actions / interventions
- Potential outcomes

Y_1 : the unit's outcome *had they been subjected to treatment t=1*

Y_0 : the unit's outcome *had they been subjected to treatment t=0*

If number of treatments is K, we have K potential outcomes (K possibly infinite)

Potential outcomes

- Y_0, Y_1 : potential outcomes
- T : binary treatment
- X : observed covariates
- Y : observed outcome

Consistency assumption

$$Y = TY_1 + (1 - T)Y_0:$$

Potential outcomes

- Y_0, Y_1 : potential outcomes
- T : binary treatment
- X : observed covariates
- Y : observed outcome

Consistency assumption

$$Y = TY_1 + (1 - T)Y_0:$$

Y is different from Y_0, Y_1

Potential Outcomes

- Each unit i has two potential outcomes:
 - Y_0 is the potential outcome had the unit received treatment 0
 - Y_1 is the potential outcome had the unit received treatment 1
- **Average Treatment Effect:**
$$ATE := \mathbb{E}[Y_1 - Y_0]$$
- The treatment assignment determines which of Y_0 and Y_1 we get to see

“The fundamental problem of
causal inference”

We only ever observe one of the
two outcomes

Non-identifiable example: the story of the smart snake-oil salesman

	Y_0	Y_1	T	Y
$x = 0$			0	0
$x = 1$			1	1

Non-identifiable example: the story of the smart snake-oil salesman



	Y_0	Y_1	T	Y
$x = 0$	0	1	0	0
$x = 1$	0	1	1	1

Non-identifiable example: the story of the smart snake-oil salesman



	Y_0	Y_1	T	Y
$x = 0$	0	0	0	0
$x = 1$	1	1	1	1



From unobservable to observable

- Y_0, Y_1 : potential outcomes
- T : binary treatment
- X : observed covariates
- $Y = Y_{obs} = TY_1 + (1 - T)Y_0$

Assume that treatment assignment is
completely random

$$Y = Y_{obs} = TY_1 + (1 - T)Y_0$$

- Treatment is random:

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

- $\mathbb{E}[Y_1] =$ Because treatment is random → prior knowledge!
- $\mathbb{E}[Y_1|T = 1] =$ Math
- $\mathbb{E}[TY_1|T = 1] =$ From definition of Y_{obs}
- $\mathbb{E}[Y_{obs} - (1 - T)Y_0|T = 1] =$ Linearity of expectation
- $\mathbb{E}[Y_{obs}|T = 1] - \mathbb{E}[(1 - T)Y_0|T = 1] =$
- $\mathbb{E}[Y_{obs}|T = 1]$

Assume that treatment is completely random

- Treatment is random:

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

- $\mathbb{E}[Y_1] =$
- $\mathbb{E}[Y_1|T = 1] =$
- $\mathbb{E}[Y_{obs}|T = 1]$

Can be estimated from data

- Treatment is random:

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

- $\mathbb{E}[Y_0] =$
- $\mathbb{E}[Y_0|T = 0] =$
- $\mathbb{E}[Y_{obs}|T = 0]$

Can be estimated from data

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

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

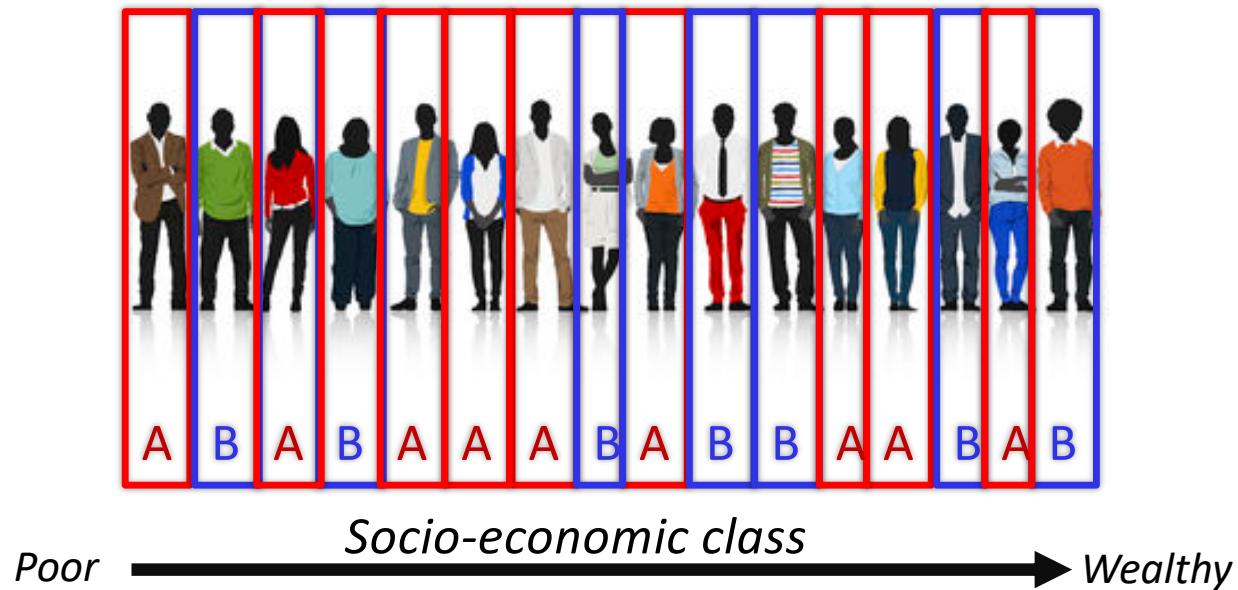
$$\mathbb{E}[Y_{obs}|T = 1] - \mathbb{E}[Y_{obs}|T = 0]$$

Under complete randomization:

$$\begin{aligned}ATE &= \mathbb{E}[Y_1 - Y_0] = \\&\mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \\&\mathbb{E}[Y_{obs}|T = 1] - \mathbb{E}[Y_{obs}|T = 0]\end{aligned}$$

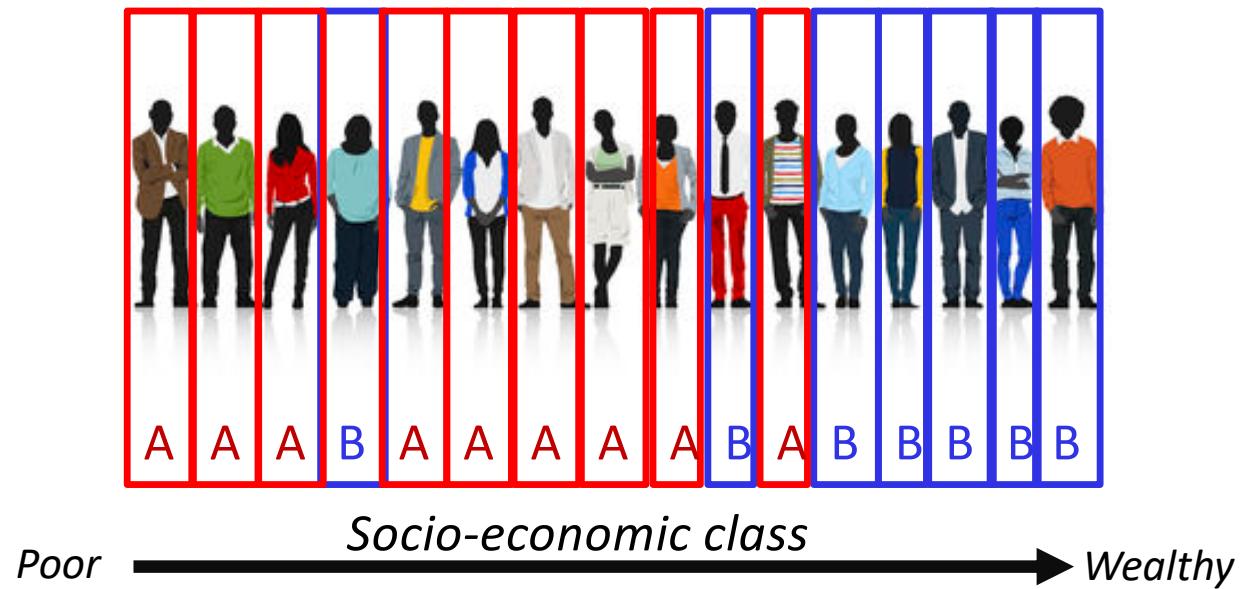
Note the difference between
unobservable quantities (potential outcomes)
and
observable quantities

Randomized controlled trial (RCT)



treatment
A or B

Observational study



treatment
A or B

Estimation

True treatment effect:

$$\mathbb{E}[Y_1 - Y_0] = 2$$

Gender	Treatment	Y_0 : Sugar levels <i>had they received treatment</i>	Y_1 : Sugar levels <i>had they received treatment</i>	Y : Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Estimation

True treatment effect:

$$\mathbb{E}[Y_1 - Y_0] = 2$$

Gender	Treatment	Y_0 : Sugar levels <i>had they received treatment</i>	Y_1 : Sugar levels <i>had they received treatment</i>	Y : Observed sugar levels
M	0	8	?	8
M	0	8	?	8
M	0	8	?	8
M	1	?	10	10
F	0	4	?	4
F	1	?	6	6
F	1	?	6	6
F	1	?	6	6

Estimation

True treatment effect:

$$\mathbb{E}[Y_1 - Y_0] = 2$$

$$\mathbb{E}[Y|t = 1] - \mathbb{E}[Y|t = 0] =$$

$$\frac{1}{4}(10 + 6 + 6 + 6) +$$

$$\frac{1}{4}(8 + 8 + 8 + 4) =$$

$$7 - 7 = 0$$

Gender	Treatment	Y_0 : Sugar levels <i>had they received treatment</i>	Y_1 : Sugar levels <i>had they received treatment</i>	Y : Observed sugar levels
M	0	8	?	8
M	0	8	?	8
M	0	8	?	8
M	1	?	10	10
F	0	4	?	4
F	1	?	6	6
F	1	?	6	6
F	1	?	6	6

Estimation

True treatment effect:

$$\mathbb{E}[Y_1 - Y_0] = 2$$

$$\mathbb{E}[Y|t = 1] = 7$$

$$\mathbb{E}[Y|t = 0] = 7$$

Gender	Treatment	Y_0 : Sugar levels <i>had they received treatment</i>	Y_1 : Sugar levels <i>had they received treatment</i>	Y : Observed sugar levels
M	0	8	?	8
M	0	8	?	8
M	0	8	?	8
M	1	?	10	10
F	0	4	?	4
F	1	?	6	6
F	1	?	6	6
F	1	?	6	6

Estimation

True treatment effect:

$$\mathbb{E}[Y_1 - Y_0] = 2$$

$$\mathbb{E}[Y|t = 1] = 7$$

$$\mathbb{E}[Y|t = 0] = 7$$

$$\mathbb{E}[Y|t = 0, \text{Gender} = M] = 8$$

$$\mathbb{E}[Y|t = 1, \text{Gender} = M] = 10$$

$$\mathbb{E}[Y|t = 0, \text{Gender} = F] = 4$$

$$\mathbb{E}[Y|t = 1, \text{Gender} = F] = 6$$

Gender	Treatment	Y_0 : Sugar levels <i>had they received treatment</i>	Y_1 : Sugar levels <i>had they received treatment</i>	Y : Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Estimation

True treatment effect:

$$\mathbb{E}[Y_1 - Y_0] = 2$$

$$\mathbb{E}[Y|t = 1] = 7$$

$$\mathbb{E}[Y|t = 0] = 7$$

$$\mathbb{E}[Y|t = 0, \text{Gender} = M] = 8$$

$$\mathbb{E}[Y|t = 1, \text{Gender} = M] = 10$$

$$\mathbb{E}[Y|t = 0, \text{Gender} = F] = 4$$

$$\mathbb{E}[Y|t = 1, \text{Gender} = F] = 6$$

Gender	Treatment	Y_0 : Sugar levels <i>had they received treatment</i>	Y_1 : Sugar levels <i>had they received treatment</i>	Y : Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Within each *group* we get
the true treatment effect

Treatment assignment mechanism

- $G=0$ if gender=F,
 $G=1$ if gender=M

$$Y_0 = 4 + 4 * G$$

$$Y_1 = 4 + 4 * G + 2$$

Gender	Treatment	Y_0 : Sugar levels <i>had they received treatment</i>	Y_1 : Sugar levels <i>had they received treatment</i>	Y : Observed sugar levels
		0	1	
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Treatment assignment mechanism

- $G=0$ if gender=F,
 $G=1$ if gender=M

$$Y_0 = 4 + 4 * G$$

$$Y_1 = 4 + 4 * G + 2$$

- $p(t=1 | G=1) = 0.25$
 $p(t=1 | G=0) = 0.75$

Gender	Treatment	Y_0 : Sugar levels <i>had they received treatment</i>	Y_1 : Sugar levels <i>had they received treatment</i>	Y : Observed sugar levels
		0	1	
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Treatment assignment mechanism

Gender	Treatment	Y_0 : Sugar levels <i>had they received treatment 0</i>	Y_1 : Sugar levels <i>had they received treatment 1</i>	Y: Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Treatment assignment mechanism

Gender	Treatment	Y_0 : Sugar levels <i>had they received treatment 0</i>	Y_1 : Sugar levels <i>had they received treatment 1</i>	Y: Observed sugar levels
M	0	8	?	8
M	0	8	?	8
M	0	8	?	8
M	1	?	10	10
F	0	4	?	4
F	1	?	6	6
F	1	?	6	6
F	1	?	6	6

$$P(Y_0 = 8|T = 0) = 0.75$$

$$P(Y_0 = 8|T = 1) = 0.25$$

$$P(Y_1 = 10|T = 0) = 0.75$$

$$P(Y_1 = 10|T = 1) = 0.25$$

(Y_0, Y_1) **are not**
independent of T

Gender	T: Treatment	Y_0 : Sugar levels <i>had they received treatment</i>	Y_1 : Sugar levels <i>had they received treatment</i>	Y: Observed sugar levels
		0	1	
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

$$P(Y_0 = 8|T = 0, G = M) = 1$$

$$P(Y_0 = 8|T = 1, G = M) = 1$$

$$P(Y_1 = 10|T = 0, G = M) = 1$$

$$P(Y_1 = 10|T = 1, G = M) = 1$$

(Y_0, Y_1) **are independent** of T
conditioned on

$G=M$, and conditioned on $G=F$

Gender	T: Treatment	Y_0 : Sugar levels <i>had they received treatment</i>	Y_1 : Sugar levels <i>had they received treatment</i>	Y: Observed sugar levels
		0	1	
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

$$\begin{aligned}
 P(Y_0 = 4|T = 0, G = F) &= 1 \\
 P(Y_0 = 4|T = 1, G = F) &= 1 \\
 P(Y_1 = 6|T = 0, G = F) &= 1 \\
 P(Y_1 = 6|T = 1, G = F) &= 1
 \end{aligned}$$

(Y_0, Y_1) **are independent** of T **conditioned** on
 $G=M$, and conditioned on $G=F$

Gender	T: Treatment	Y ₀ : Sugar levels <i>had they received treatment</i>		Y: Observed sugar levels
		0	1	
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

$$\begin{aligned}
 P(Y_0 = 4|T = 0, G = F) &= 1 \\
 P(Y_0 = 4|T = 1, G = F) &= 1 \\
 P(Y_1 = 6|T = 0, G = F) &= 1 \\
 P(Y_1 = 6|T = 1, G = F) &= 1
 \end{aligned}$$

(Y_0, Y_1) **are independent** of T **conditioned** on
 $G=M$, and conditioned on $G=F$

$$(Y_0, Y_1) \perp\!\!\!\perp T | G$$

Gender	T: Treatment	Y ₀ : Sugar levels <i>had they received treatment</i>		Y: Observed sugar levels
		0	1	
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

What if we can't randomize treatment?

- We can still succeed if the treatment assignment process is *conditionally randomized*, conditioned on an observed quantity
- This is actually just a way of saying we have *no unmeasured confounding*

What assumptions are sufficient for us to identify the causal effect in an observational study?

The target trial

- We have data collected where treatment was not explicitly randomized
- We can view it as a “flawed” randomized trial
- Are the flaws fixable?

Identifying Assumptions

- We will describe today a set of assumptions that will go with us through a big part of the course
- A big part of causal inference is understanding when these assumptions are plausible
- There is active research into relaxing each and every one of these assumptions
- We will see other “assumption sets” later in the course (e.g. instrumental variables, front-door adjustment)

“The Assumptions”

Sufficient conditions for causal inference to be possible:

- 1. Stable Unit Treatment Value Assumption**
- 2. Consistency**
- 3. *Ignorability / No unmeasured confounders***
- 4. *Common support***

Stable Unit Treatment Value Assumption SUTVA

- 1. The potential outcomes for any unit do not vary with the treatments assigned to other units*

Stable Unit Treatment Value Assumption SUTVA

1. *The potential outcomes for any unit do not vary with the treatments assigned to other units*
failure example: vaccination, network effects

Stable Unit Treatment Value Assumption SUTVA

- 1. The potential outcomes for any unit do not vary with the treatments assigned to other units*
failure example: vaccination, network effects
- 2. For each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes*

Stable Unit Treatment Value Assumption SUTVA

- 1. The potential outcomes for any unit do not vary with the treatments assigned to other units*
failure example: vaccination, network effects
- 2. For each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes*
failure example: some people get out-of-date medication

Consistency

- For a unit that receives treatment t , we observe the corresponding potential outcome Y_t

Ignorability – no unmeasured confounders

Y_0, Y_1 : potential outcomes for control and treated

X : observed unit covariates (features)

T : treatment assignment

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

The potential outcomes are independent of treatment assignment, conditioned on observed covariates x

Ignorability – no unmeasured confounders

Y_0, Y_1 : potential outcomes for control and treated

X : observed unit covariates (features)

T: treatment assignment

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

The potential outcomes are independent of treatment assignment, conditioned on observed covariates x

Weird! The outcomes obviously depend on treatment.
How can potential outcomes not depend?

Ignorability – no unmeasured confounders

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

The potential outcomes are independent of treatment assignment, conditioned on observed covariates x

Within each “strata/level” $X = x$, treatment assignment is as good as random

Ignorability – no unmeasured confounders

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

The potential outcomes are independent of treatment assignment, conditioned on observed covariates x

Within each “strata/level” $X = x$, treatment assignment is as good as random with respect to Y_0, Y_1

Failure example: treatment depends on gender, which also has an effect on the potential outcomes

Common support

Y_0, Y_1 : potential outcomes for control and treated

x : unit covariates (features)

T : treatment assignment

We assume:

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

Common support

Y_0, Y_1 : potential outcomes for control and treated

x : unit covariates (features)

T : treatment assignment

We assume:

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

Failure:

if only women receive $T=1$,
and only men receive $T=0$

Ignorability assumption is unverifiable from data!

- Remember, we never observe (Y_1^i, Y_0^i) jointly
- How do we know it holds in an RCT?

Checking the assumptions: Ignorability assumption is unverifiable from data!

- How can we convince ourselves that it is true in a given case?
- Confounders: factors that affect **both** treatment assignment and outcome
- Talk to domain experts, understand what determines treatment assignment and outcomes
- Sensitivity analysis
- Do you believe ignorability holds? If not - change the design:
 - Add relevant variables
 - Define or measure treatment differently
 - Define or measure outcome differently

Checking the assumptions - example

- Comparing effectiveness of two anti-hypertensive medications
- Treatment: first administration of medication
- Outcome: blood pressure 3 months after first treatment
 - Is outcome only measured for some of the patients?

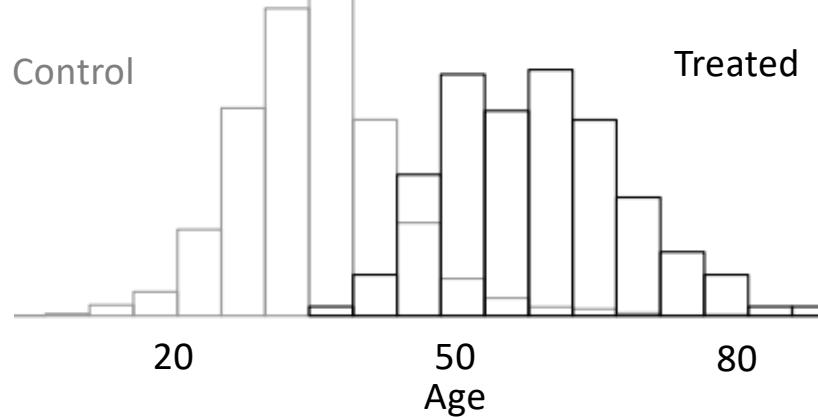
Checking the assumptions - ignorability

- Did we measure the important known causes of hypertension? Literature survey
- Example: high alcohol use is known to be a cause of hypertension
- Doctors know this, and might use this information in deciding on treatment
- If we don't measure alcohol use, it becomes hidden confounder which might bias our conclusions
- Talking to doctors to understand how they prescribe treatment

Checking the assumptions – common support

- Check for common support between treated and control:
 - Reduce dimension and plot populations
 - Check overlap on important univariate and bivariate variables, e.g. age, gender, weight in a medical study

Figure:
Hill & Gelman

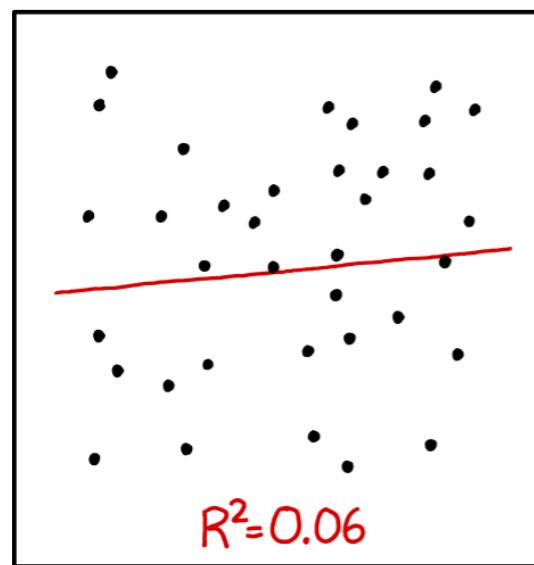


Checking the assumptions – common support

- If no common support:
 - Redefine study population, e.g. only people ages 40-60
 - More risky: check if outcome is sensitive to unbalanced variables.

Example: treated and control might differ in commute distance from hospital, but distance from hospital is not associated with any important socio-economic factors and has no observed association with outcome

How can we estimate causal
effects under ignorability and
common support?



Average Treatment Effect – the adjustment formula

- Assuming ignorability, we will derive the *covariate-adjustment formula*
- The adjustment formula is **extremely** useful in causal inference
- Strongly related to *G-formula* and *back-door adjustment*

Average Treatment Effect

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

Average Treatment Effect

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

$$\mathbb{E}[Y_1] =$$

Average Treatment Effect

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

law of total
expectation

$$\mathbb{E}[Y_1] =$$
$$\mathbb{E}_{x \sim p(x)} [\mathbb{E}_{Y_1 \sim p(Y_1|x)} [Y_1|x]] =$$

Average Treatment Effect

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

$$\mathbb{E}[Y_1] =$$

$$\mathbb{E}_{x \sim p(x)} [\mathbb{E}_{Y_1 \sim p(Y_1|x)} [Y_1|x]] = \begin{array}{l} \text{ignorability} \\ (Y_0, Y_1) \perp\!\!\!\perp T | x \end{array}$$

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

Average Treatment Effect

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

$$\mathbb{E}[Y_1] =$$

$$\mathbb{E}_{x \sim p(x)} [\mathbb{E}_{Y_1 \sim p(Y_1|x)} [Y_1|x]] =$$

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

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

shorter notation

Average Treatment Effect

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

$$\mathbb{E}[Y_1] =$$

$$\mathbb{E}_{x \sim p(x)} [\mathbb{E}_{Y_1 \sim p(Y_1|x)} [Y_1|x]] =$$

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

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

$$\mathbb{E}_{x \sim p(x)} [\mathbb{E}[Y|x, T=1]] \quad \text{consistency}$$

Average Treatment Effect

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

$$\mathbb{E} [Y_1] =$$

$$\mathbb{E}_{x \sim p(x)} [\mathbb{E}_{Y_1 \sim p(Y_1|x)} [Y_1|x]] =$$

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

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

$$\mathbb{E}_{x \sim p(x)} [\mathbb{E} [Y|x, T=1]]$$

Might be
estimated from
data!

Average Treatment Effect

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

$$\mathbb{E} [Y_0] =$$

$$\mathbb{E}_{x \sim p(x)} [\mathbb{E}_{Y_0 \sim p(Y_0|x)} [Y_0|x]] =$$

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

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

$$\mathbb{E}_{x \sim p(x)} [\mathbb{E} [Y|x, T=0]]$$

Might be
estimated from
data!

The adjustment formula

Under the assumption of ignorability,
we have that:

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

$$\left. \begin{array}{l} \mathbb{E}[Y|x, T=1] \\ \mathbb{E}[Y|x, T=0] \end{array} \right\} \quad \begin{array}{l} \text{Quantities we} \\ \text{can hope to} \\ \text{estimate} \\ \text{from data} \end{array}$$

The adjustment formula

Under the assumption of ignorability,
we have that:

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

Empirically we have samples from
 $p(x|T=1)$ or $p(x|T=0)$

The adjustment formula

Under the assumption of ignorability,
we have that:

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

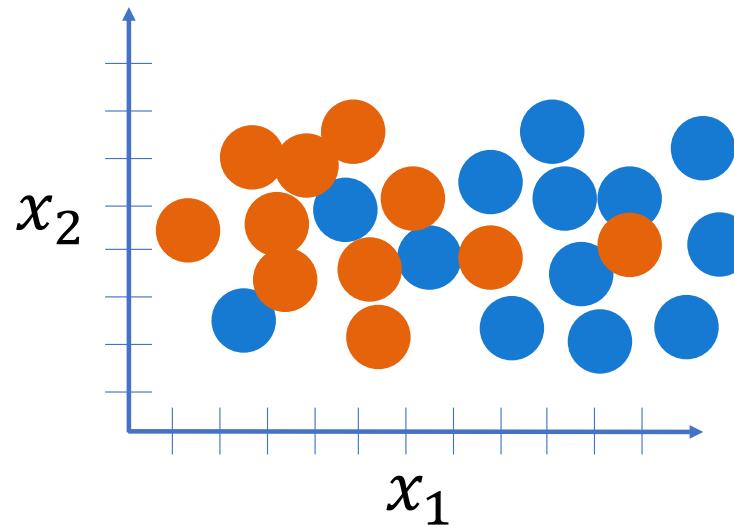
Empirically we have samples from
 $p(x|T=1)$ or $p(x|T=0)$

To extrapolate to $p(x)$ → common support assumption

When is estimating treatment effect harder?

Observational study

Treatment assignment non-random → counterfactual and factual have different distributions



- Control, $t = 0$
- Treated, $t = 1$

The adjustment formula

Under the assumption of ignorability,
we have that:

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

Empirically we have samples from
 $p(x|T=1)$ or $p(x|T=0)$

To extrapolate to $p(x) \rightarrow \text{overlap}$ assumption

Covariate adjustment (parametric g-formula)

- Explicitly model the relationship between treatment, confounders, and outcome
- Under ignorability, the expected causal effect of T on Y :
$$\mathbb{E}_{x \sim p(x)} [\text{red} \mathbb{E}[Y_1 | T = 1, x] - \text{blue} \mathbb{E}[Y_0 | T = 0, x]]$$
- Fit a model $f(x, t) \approx \mathbb{E}[Y_t | T = t, x]$

$$\widehat{ATE} = \frac{1}{n} \sum_{i=1}^n f(x_i, 1) - f(x_i, 0)$$

Covariate adjustment (parametric g-formula)

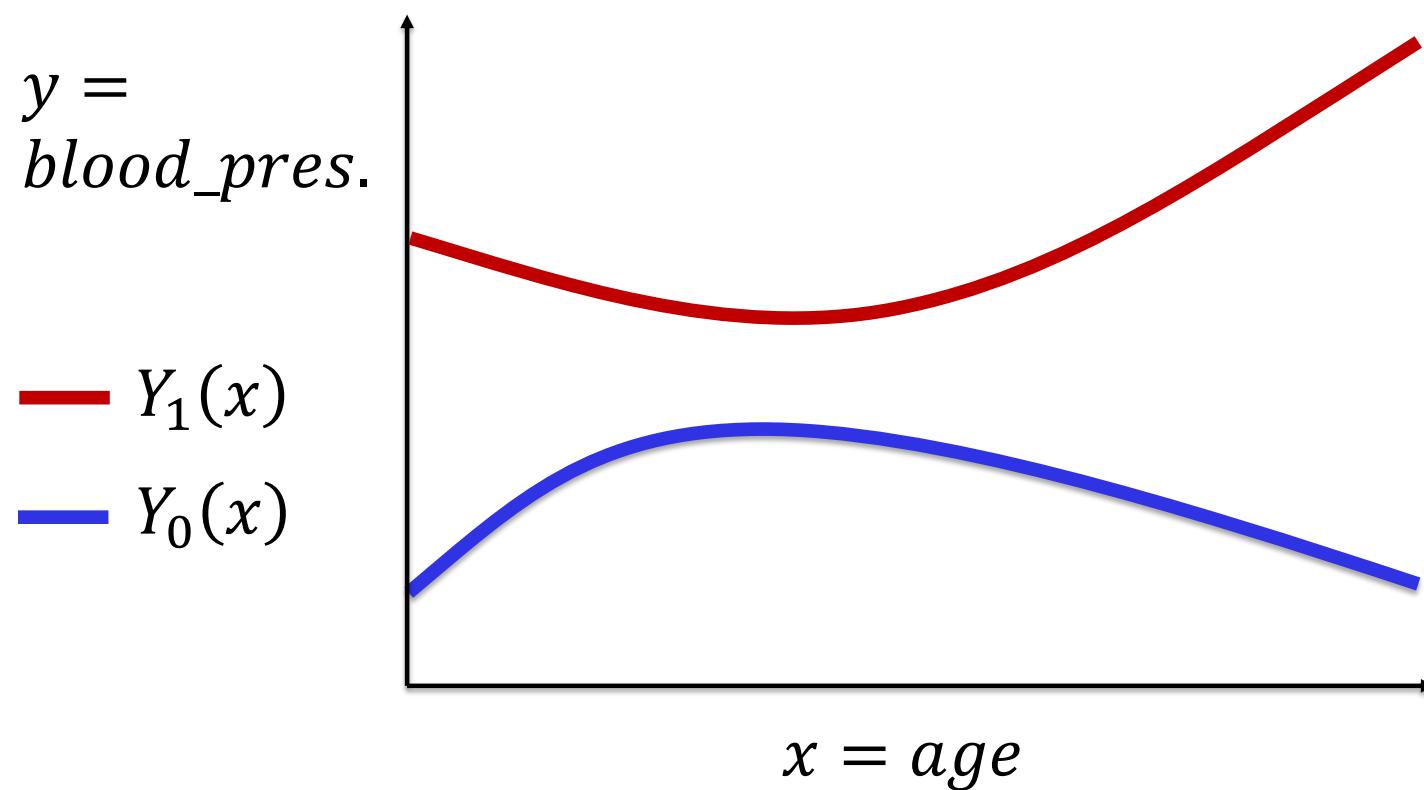
- Explicitly model the relationship between treatment, confounders, and outcome
- Under ignorability, the expected causal effect of T on Y :
$$\mathbb{E}_{x \sim p(x)} [\textcolor{red}{\mathbb{E}[Y_1 | T = 1, x]} - \textcolor{blue}{\mathbb{E}[Y_0 | T = 0, x]}]$$
- Fit a model $f(x, t) \approx \mathbb{E}[Y_t | T = t, x]$

$$\widehat{CATE}(x_i) = f(x_i, 1) - f(x_i, 0)$$

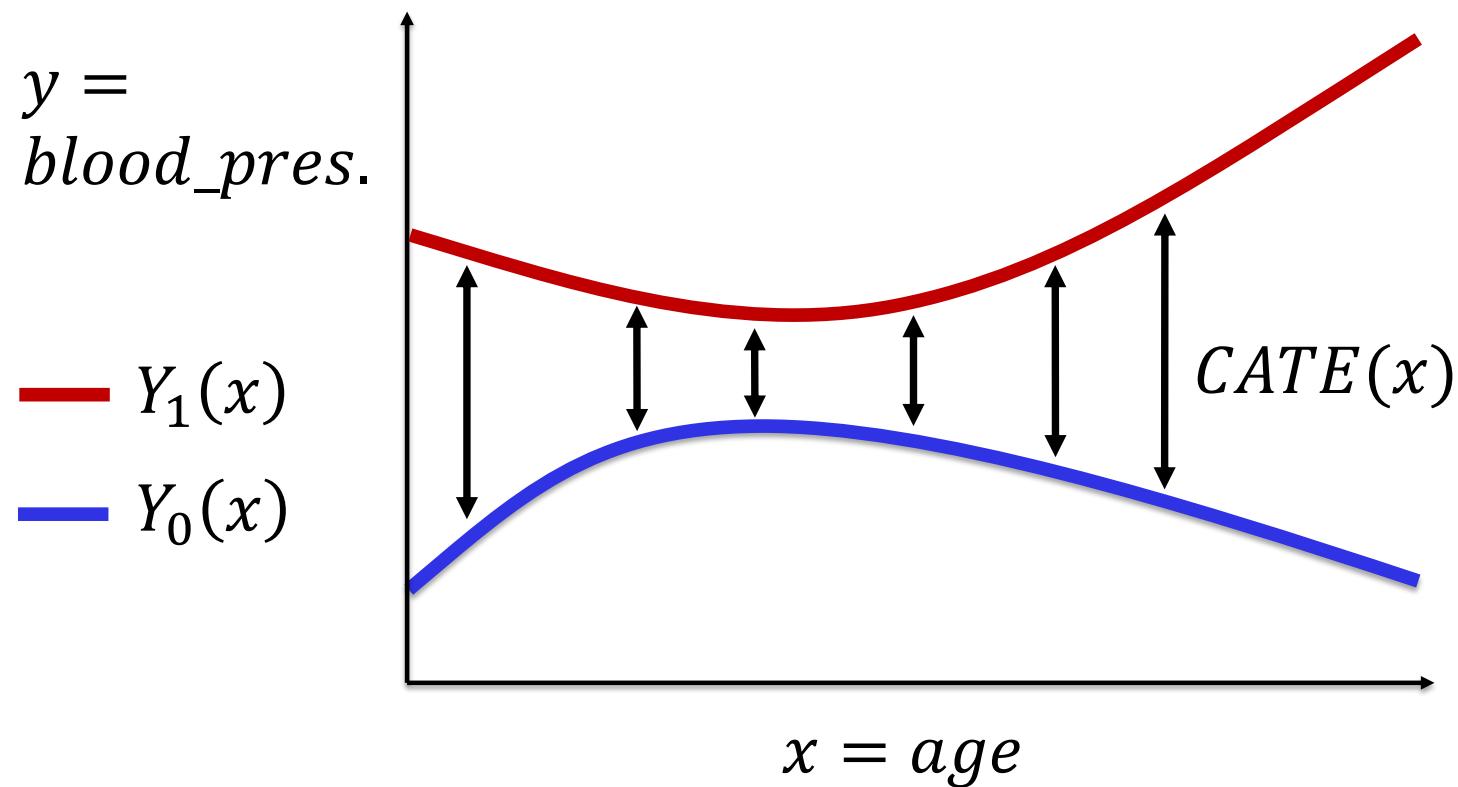
Covariate adjustment – model consistency (unrelated to our consistency assumption)

- If the model $f(x, t) \approx \mathbb{E}[Y_t | T = t, x]$ is consistent in the limit of infinite samples, then under ignorability the estimated \widehat{ATE} will converge to the true ATE
- A sufficient condition: overlap and well-specified model

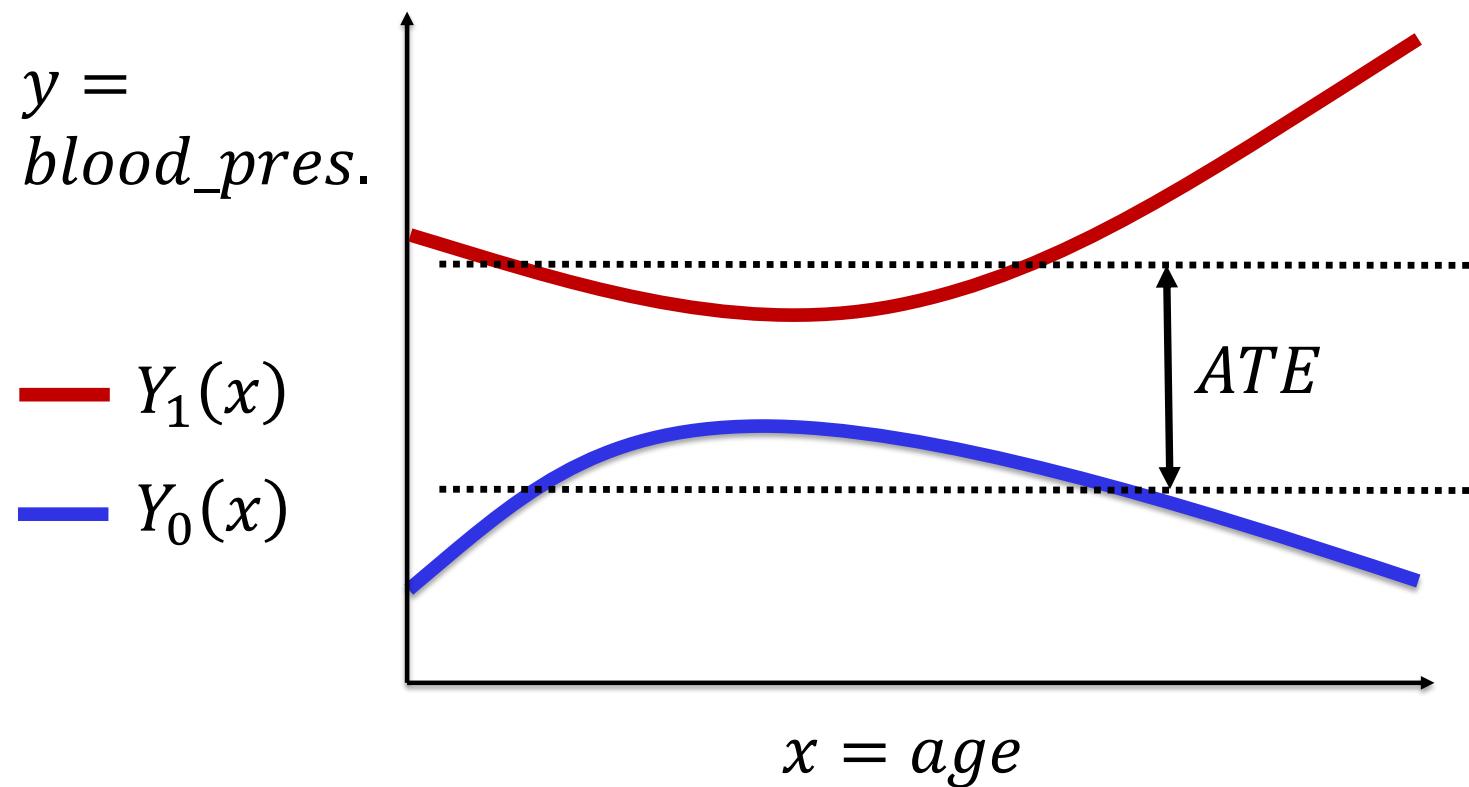
Covariate adjustment



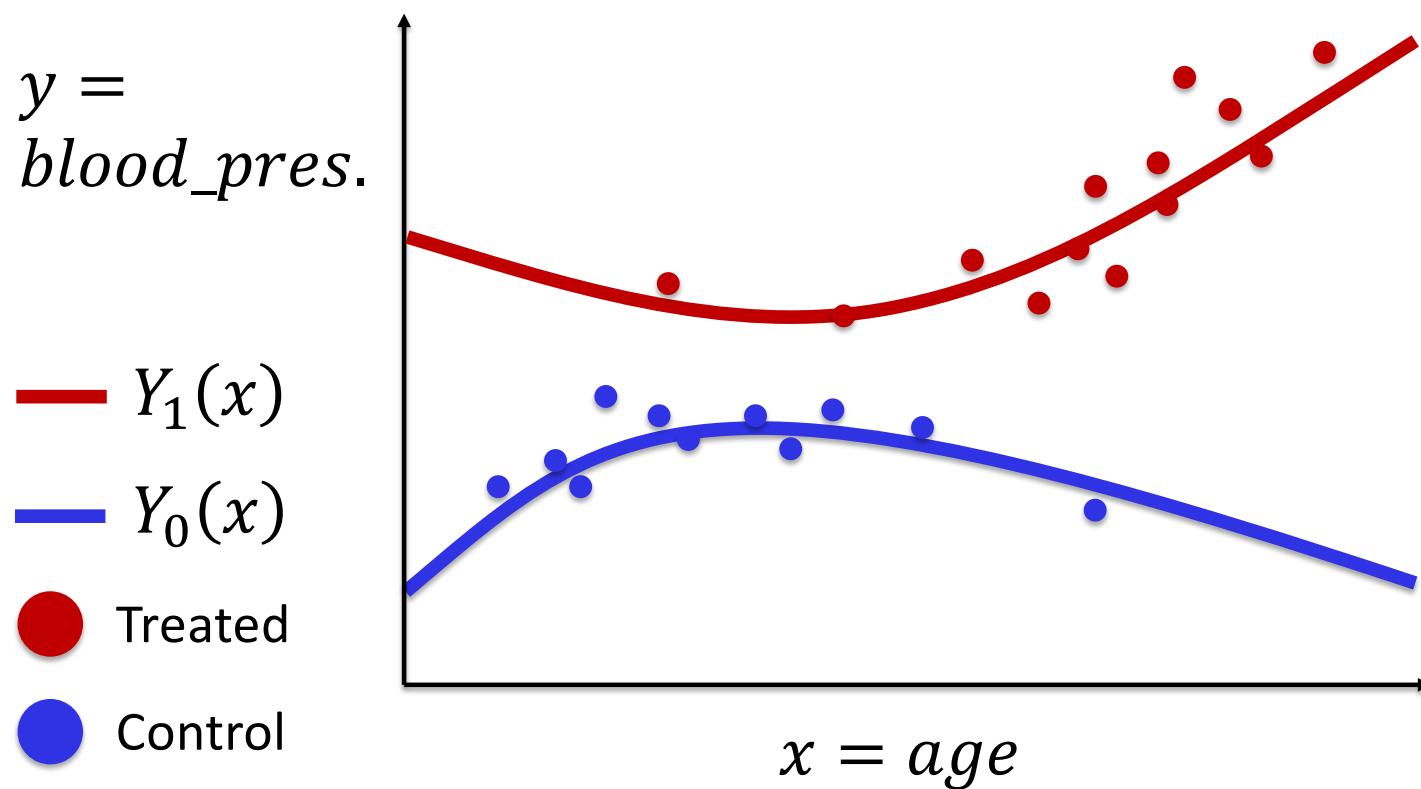
Covariate adjustment



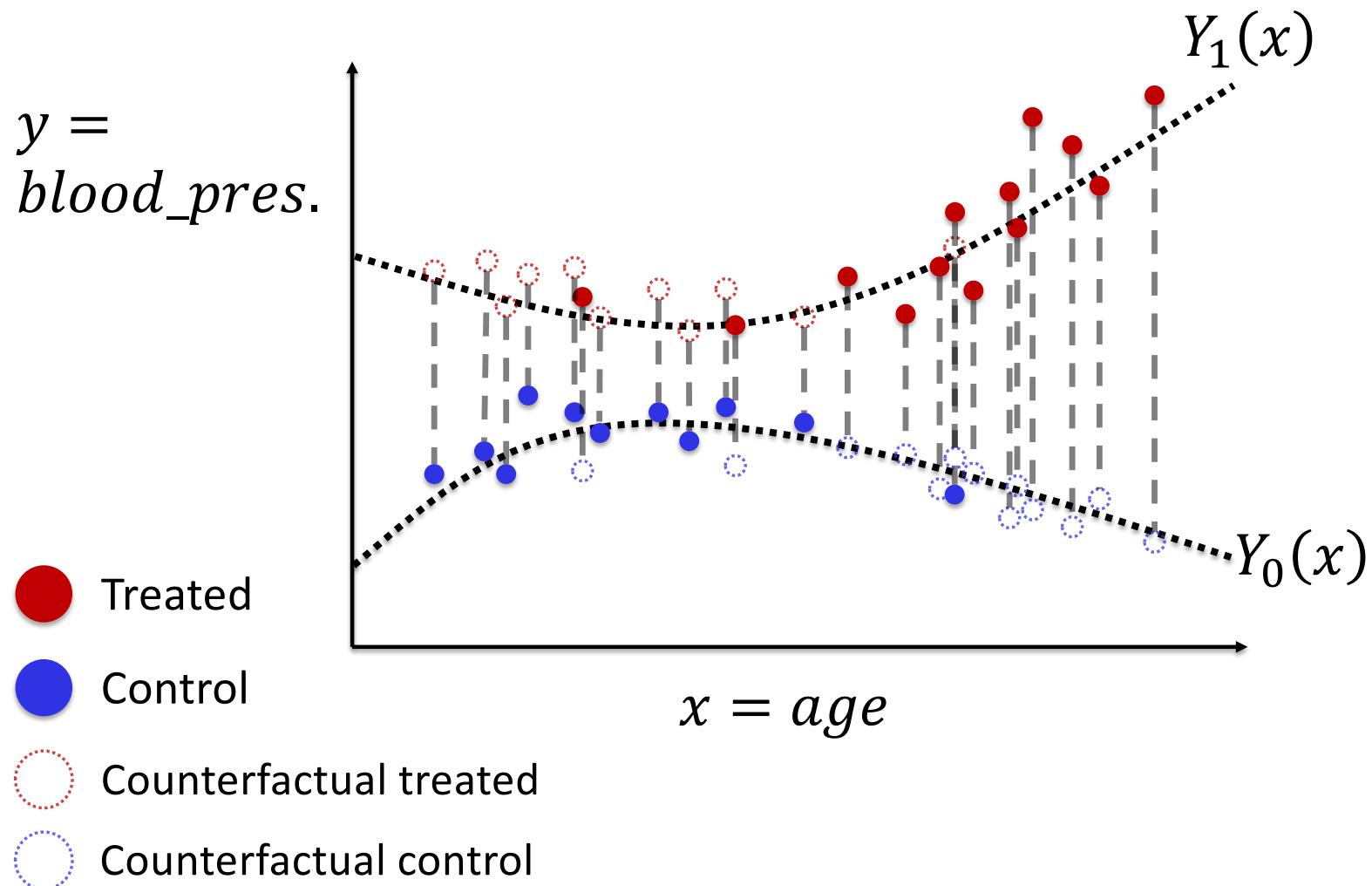
Covariate adjustment



Covariate adjustment



Covariate adjustment



Covariate adjustment in practice

- Given data:

$$(x^1, t^1, y^1), \dots, (x^n, t^n, y^n)$$

1. Linear: Regress y on (x, t) , read off coefficient for t (regularize?)
2. Non-linear: Regress y on (x, t)
3. Regress separately over set with $t=1$ and set with $t=0$?
4. Active field of research developing specialized methods

Covariate adjustment in practice

- Given data:

$$(x^1, t^1, y^1), \dots, (x^n, t^n, y^n)$$

Covariate adjustment in practice

- Given data:

$$(x^1, t^1, y^1), \dots, (x^n, t^n, y^n)$$

- Regress the observed y 's on (x, t) such that

$$y_i \approx f(x_i, t_i)$$
$$\widehat{ATE} = \frac{1}{n} \sum_{i=1 \dots n} f(x_i, 1) - f(x_i, 0)$$

Covariate adjustment in practice

- Given data:

$$(x^1, t^1, y^1), \dots, (x^n, t^n, y^n)$$

- Regress the observed y 's on (x, t) such that

$$y_i \approx f(x_i, t_i)$$

$$\widehat{ATE}_1 = \frac{1}{n} \sum_{i=1 \dots n} f(x_i, 1) - f(x_i, 0)$$

$$\widehat{ATE}_2 = \frac{1}{\sum t_i} \sum_{i \text{ s.t. } t_i=1} y_i - f(x_i, 0) + \frac{1}{\sum 1 - t_i} \sum_{i \text{ s.t. } t_i=1} f(x_i, 1) - y_i$$

Problems with covariate adjustment

- It's too easy!
- Should think thoroughly about what covariates go in – that is the most important decision (often more than which algorithm)
- Example of mistakes:
 - including post-treatment covariates (leads to zero-bias)
 - Including covariates that only influence treatment and not outcome (leads to high-variance)
 - Using a model not specified for causal inference (leads to statistical inefficiency)

Linear model

- Assume that:

$$\begin{array}{c} \text{blood pressure} \\ Y_t(x) = \beta^T x + \gamma \cdot t + \epsilon_t \\ \text{age, weight, ...} \\ \mathbb{E}[\epsilon_t] = 0 \\ \text{medication} \end{array}$$

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

- We care about γ , not about $Y_t(x)$

Estimation, not prediction

Linear model

blood pressure **age,weight,...** **medication**

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

Hypertension is affected by many variables:
lifestyle, weight, genetics, age

- Each of these often stronger **predictor** of blood-pressure, compared with type of medication taken
- Regularization (e.g. Lasso) might remove the treatment variable!
- Features → (“nuisance parameters”, “variable of interest”)

Regression - 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:

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

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

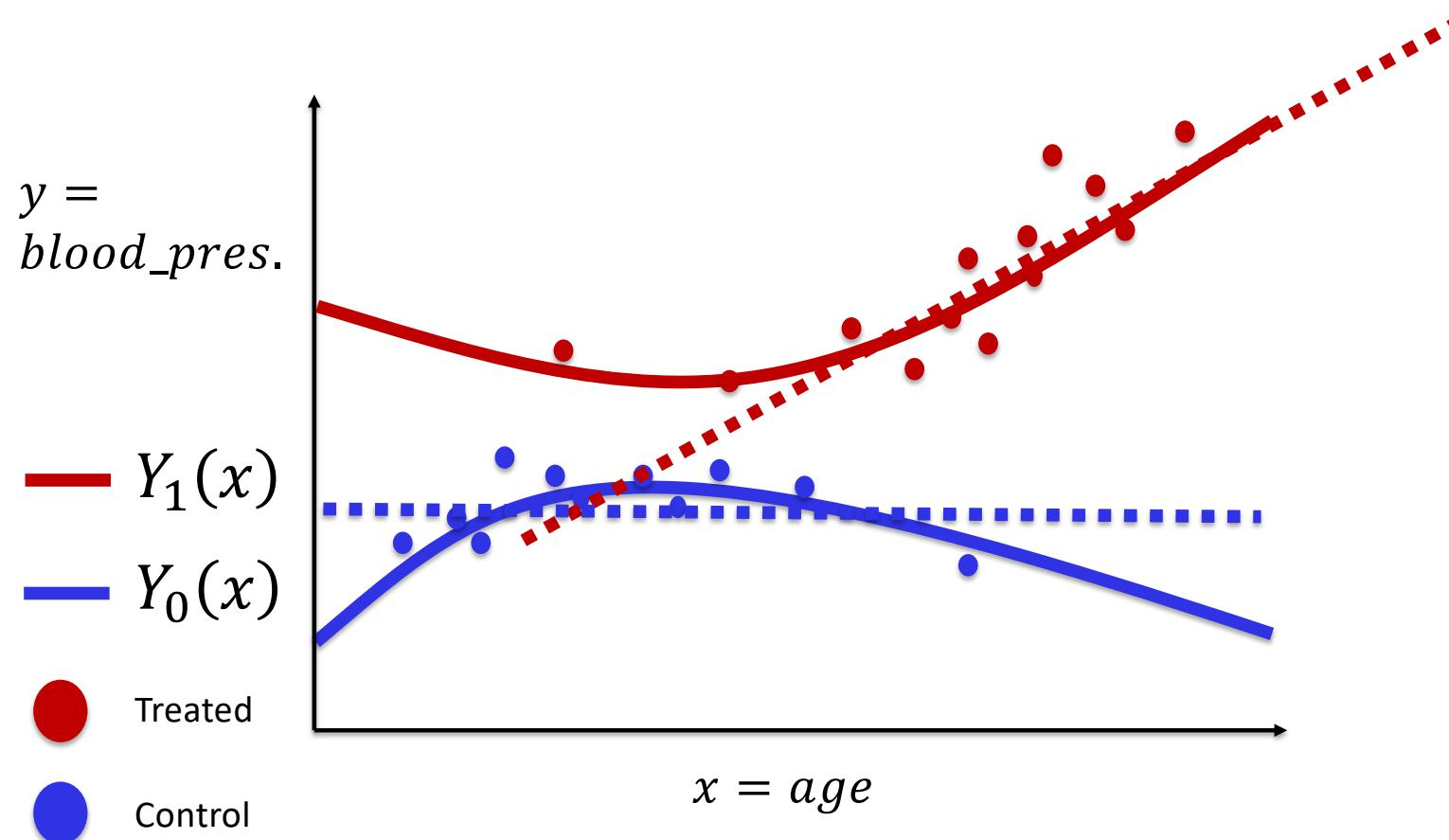
Using machine learning for causal inference under ignorability

- Machine learning techniques can be very useful and have recently seen wider adoption
- Random forests and Bayesian trees
Hill (2011), Athey & Imbens (2015), Wager & Athey (2015)
- Gaussian processes
Hoyer et al. (2009), Zigler et al. (2012)
- Neural nets
Beck et al. (2000), Johansson et al. (2016), Shalit et al. (2016), Lopez-Paz et al. (2016)
- “Causal” Lasso
Belloni et al. (2013), Farrell (2015), Athey et al. (2016)

Using machine learning for causal inference under ignorability

- Machine learning techniques can be very useful and have recently seen wider adoption
- How is the treatment variable used:
 - Fit two different models for treated and control?
 - Not regularized?
 - Privileged

Covariate adjustment: weak overlap



Problems with covariate adjustment

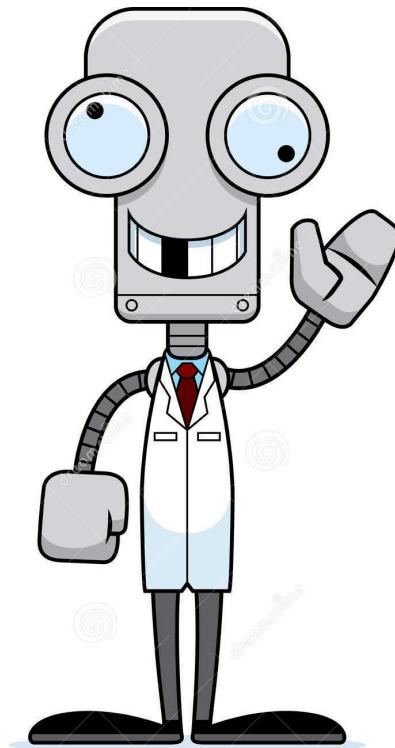
- It's too easy!
- Should think thoroughly about what covariates go in – that is the most important decision (more than which algorithm)
- Example of mistake: including post-treatment covariates

Example: post-treatment mistake

- Did we include a post-treatment covariate?
example: weight measured after treatment
- Measuring the post-treatment weight could explain away some of the causal effect, inducing bias in our estimation of the causal effect
- Conditioning on post-treatment covariates violates ignorability
- But how do we know if a variable is post-treatment?
- We will look into this issue again further in the course

Often very easy to run a regression / fit a model, and get the causal aspects wrong

A possible solution: “Target Trial” thinking





Introduction to Causal Inference - a Machine Learning Perspective

Dr. Uri Shalit

Course number 097400
2020-2021

Lesson 3

Reminder

- Y_0, Y_1 : potential outcomes
- T : binary treatment
- X : observed covariates

The classic identification conditions

- A set of conditions (no hidden confounders, overlap) which reduced inferring causal effects to regression
- No hidden confounders: hard to check, brings up the role of *study design*
- Common support (overlap): possible to check, we will see later today

Ignorability – no unmeasured confounders

Y_0, Y_1 : potential outcomes for control and treated

x : observed unit covariates (features)

T: treatment assignment

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

The potential outcomes are independent of treatment assignment, conditioned on observed covariates x

Second assumption: common support (overlap)

Y_0, Y_1 : potential outcomes for control and treated

x : unit covariates (features)

T : treatment assignment

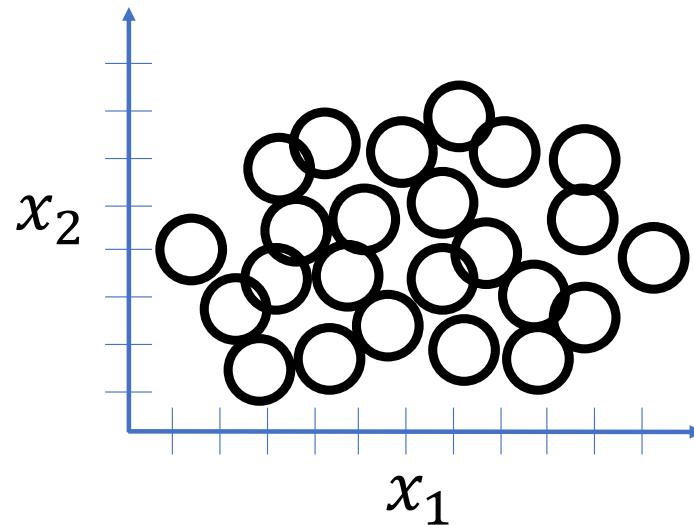
We assume:

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

Average Treatment Effect – the adjustment formula

- Assuming ignorability, we will derive the *covariate-adjustment formula*
- The adjustment formula is **extremely** useful in causal inference
- Strongly related to *G-formula* and *back-door adjustment*

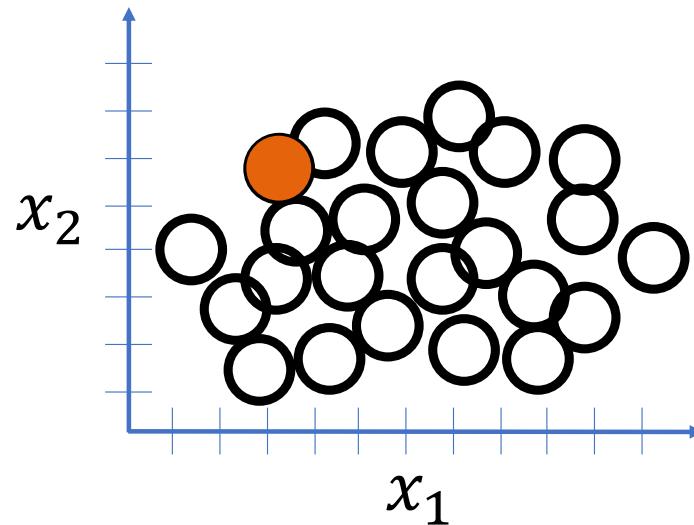
When is estimating treatment effect easier? Randomized Controlled Trials



- Control, $t = 0$
- Treated, $t = 1$

When is estimating treatment effect easier?

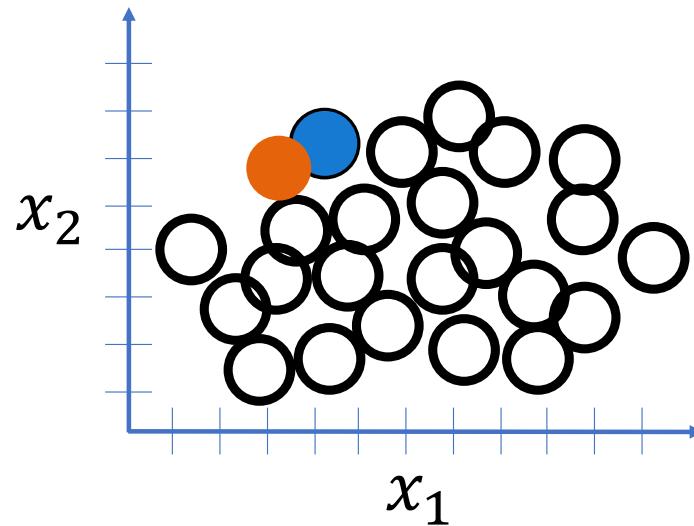
Randomized Controlled Trials



- Control, $t = 0$
- Treated, $t = 1$

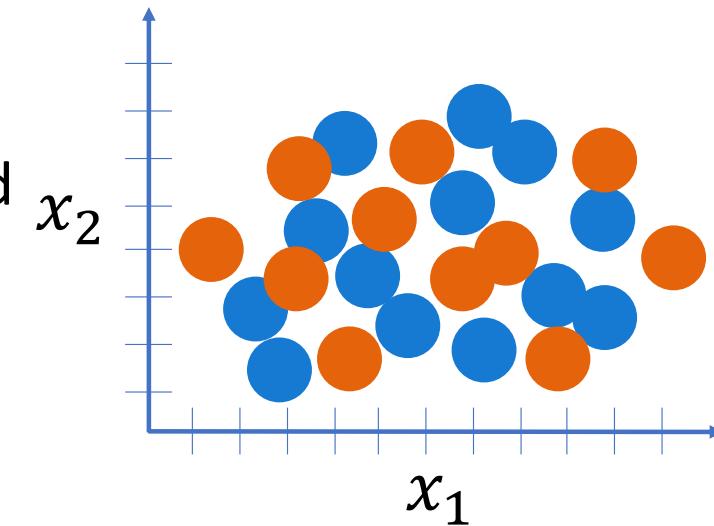
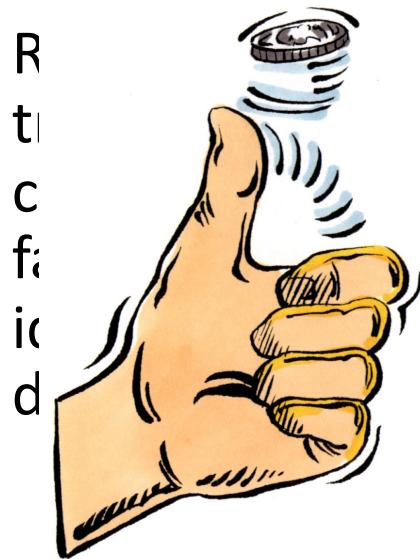
When is estimating treatment effect easier?

Randomized Controlled Trials



- Control, $t = 0$
- Treated, $t = 1$

When is estimating treatment effect easier? Randomized Controlled Trials

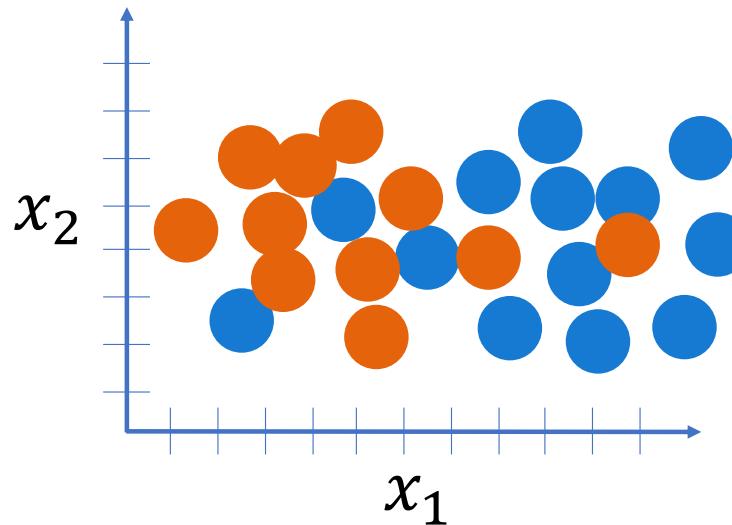


- Control, $t = 0$
- Treated, $t = 1$

When is estimating treatment effect harder?

Observational study

Treatment assignment non-random → counterfactual and factual have different distributions



- Control, $t = 0$
- Treated, $t = 1$

Propensity score

- Extremely widely used tool
- Basic idea: turn observational study into a pseudo-randomized trial by correcting for non-random sampling

Ignorability

- $(Y_0, Y_1) \perp\!\!\!\perp T | x$
- What functions of $f(x)$ will still allow $(Y_0, Y_1) \perp\!\!\!\perp T | f(x)$?
- Theorem:
Let $e(x) = p(T = 1|x)$, also called the ***propensity score***.
If ignorability holds for x , then $e(x)$ is the coarsest function of x for which ignorability still holds

Balancing score

- Balancing score:
A function $b(x)$ such that
 $x \perp\!\!\!\perp T | b(x)$
($b(x)$ not necessarily scalar valued)
- Definition:
For a pair of functions $a: \mathcal{X} \rightarrow \mathcal{H}$, $b: \mathcal{X} \rightarrow \mathcal{H}'$
 b is *finer* than a if
$$a(x_1) \neq a(x_2) \Rightarrow b(x_1) \neq b(x_2)$$
- Lemma:
If there exists a function $f: \mathcal{H} \rightarrow \mathcal{H}'$ such that $a(x) = f(b(x))$,
then b is finer than a .
- Propensity score: $e(x) = p(T = 1|x)$

- Theorem 1:

$b: \mathcal{X} \rightarrow \mathcal{H}$ is a balancing score if and only if there exists a function $f: \mathcal{H} \rightarrow \mathbb{R}$ such that $e(x) = f(b(x))$

- Theorem 2:

For any balancing score $b: \mathcal{X} \rightarrow \mathcal{H}$,
 $(Y_0, Y_1) \perp\!\!\!\perp T | x \implies (Y_0, Y_1) \perp\!\!\!\perp T | b(x)$

- Conclusion:

- Ignorability with x equivalent to ignorability with $b(x)$ for any balancing score $b(x)$
- Any balancing score is finer than the propensity score $e(x)$

- Balancing score:

A function $b(x)$ such that

$x \perp\!\!\!\perp T | b(x)$

($b(x)$ not necessarily scalar valued)

- Definition:

For a pair of functions $a: X \rightarrow H$, $b: X \rightarrow H'$ b is *finer* than a if
 $a(x_1) \neq a(x_2) \implies b(x_1) \neq b(x_2)$

- Lemma:

If there exists a function $f: H \rightarrow H'$ such that
 $a(x) = f(b(x))$, then b is finer than a .

- Propensity score: $e(x) = p(T = 1|x)$

What does the propensity score give us?

- If we have ignorability, in theory the propensity score gives all everything we need
- We can run covariate adjustment on the propensity score!
$$\mathbb{E}[Y|e(x), T = 1] - \mathbb{E}[Y|e(x), T = 0]$$
- Other method using propensity which we will see soon:
 - Inverse propensity score weighting
 - Propensity score matching
 - Stratification on the propensity score

The propensity score

- $e(x) = p(T = 1|x)$, the treatment assignment mechanism
- In most cases must be estimated from data
- Can use any machine learning method:
logistic regression, random forests, neural nets
- Unlike most ML applications, we need to get the **probability** itself accurately
- Subtle point: if we include x which are only predictive of treatment assignment but not outcome
- Hard (but not impossible) to validate models

Exact propensity scores

- RCTs
- Computational advertising
- In general whenever we know exactly what generated past actions, example an agent for which we have the true model

Another view of propensity score

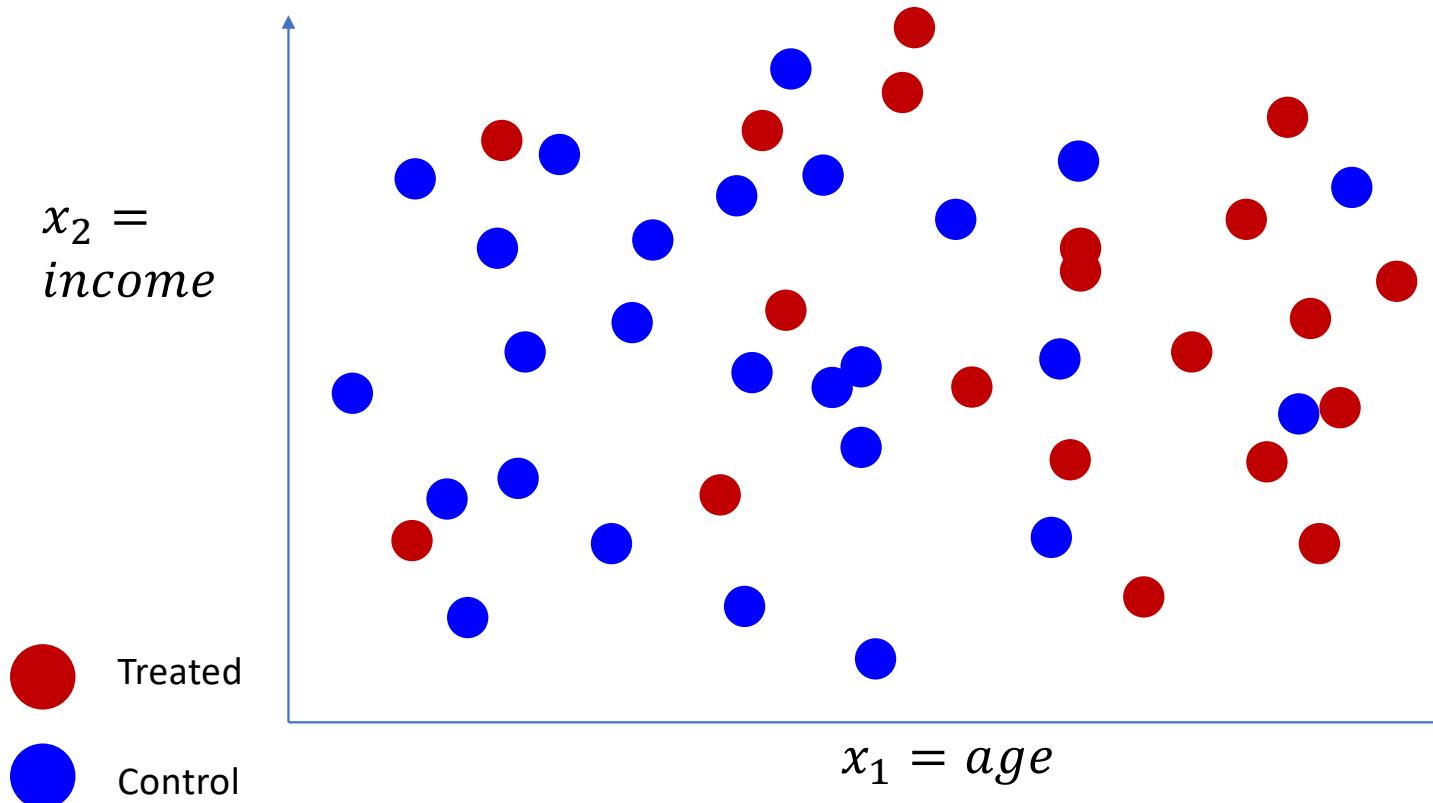
- RCTs are easy to analyze because we know the probability of each sample getting treated
- If $e(x) = 0.5$ (RCT with equal chance for treated and control) then we weight each sample by 0.5
- In observational study, we have “biased randomizations”. We can invert them to turn the observational study in a randomized study*
* with caveats

Inverse Probability Weighting (IPW) with propensity scores

Inverse propensity score re-weighting

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

reweighted control reweighted treated



Propensity scores - derivation

- Recall average treatment effect:

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

- We only have samples for:

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

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

Propensity scores - derivation

- We only have samples for:

$$\mathbb{E}_{x \sim p(x|T=1)} [\textcolor{red}{\mathbb{E}[Y|x, T=1]}]$$

$$\mathbb{E}_{x \sim p(x|T=0)} [\textcolor{blue}{\mathbb{E}[Y|x, T=0]}]$$

Propensity scores - derivation

- We only have samples for:

$$\mathbb{E}_{x \sim p(x|T=1)} [\mathbb{E}[Y|x, T=1]]$$

$$\mathbb{E}_{x \sim p(x|T=0)} [\mathbb{E}[Y|x, T=0]]$$

- We need to turn $p(x|T=1)$ into $p(x)$:

$$p(x|T=1) \cdot ? = p(x)$$

Propensity scores - derivation

- We only have samples for:

$$\mathbb{E}_{x \sim p(x|T=1)} [\mathbb{E}[Y|x, T=1]]$$

$$\mathbb{E}_{x \sim p(x|T=0)} [\mathbb{E}[Y|x, T=0]]$$

- We need to turn $p(x|T=1)$ into $p(x)$:

$$p(x|T=1) \cdot \frac{p(T=1)}{p(T=1|x)} = p(x)$$

Propensity score

Propensity scores - derivation

- We only have samples for:

$$\mathbb{E}_{x \sim p(x|T=1)} [\mathbb{E}[Y|x, T=1]]$$

$$\mathbb{E}_{x \sim p(x|T=0)} [\mathbb{E}[Y|x, T=0]]$$

- We need to turn $p(x|T=0)$ into $p(x)$:

$$p(x|T=0) \cdot \frac{p(T=0)}{p(T=0|x)} = p(x)$$

Propensity score

- We only have samples for:

$$\mathbb{E}_{x \sim p(x|T=1)} [\mathbb{E} [Y | x, T = 1]]$$

- We want:

$$\mathbb{E}_{x \sim p(x)} [\mathbb{E} [Y | x, T = 1]]$$

- We know that:

$$p(x|T = 1) \cdot \frac{p(T = 1)}{p(T = 1|x)} = p(x)$$

- Then:

$$\mathbb{E}_{x \sim p(x|T=1)} \left[\frac{p(T = 1)}{p(T = 1|x)} \mathbb{E} [Y | x, T = 1] \right] =$$

$$\mathbb{E}_{x \sim p(x)} [\mathbb{E} [Y | x, T = 1]]$$

Propensity scores – algorithm

Inverse probability of treatment weighted estimator

How to calculate ATE with propensity score

for sample $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$

1. Use any ML method to estimate $\hat{p}(T = t|x)$

2.

$$\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } \textcolor{red}{t_i=1}} \frac{y_i}{\hat{p}(t_i = 1|x_i)} - \frac{1}{n} \sum_{i \text{ s.t. } \textcolor{blue}{t_i=0}} \frac{y_i}{\hat{p}(t_i = 0|x_i)}$$

Propensity scores – algorithm

Inverse probability of treatment weighted estimator

How to calculate ATE with propensity score

for sample $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$

1. Randomized trial $p(T = t|x) = 0.5$

$$2. \hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } \textcolor{red}{t_i=1}} \frac{y_i}{\hat{p}(t_i = 1|x_i)} - \frac{1}{n} \sum_{i \text{ s.t. } \textcolor{blue}{t_i=0}} \frac{y_i}{\hat{p}(t_i = 0|x_i)}$$

Propensity scores – algorithm

Inverse probability of treatment weighted estimator

How to calculate ATE with propensity score

for sample $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$

1. Randomized trial $p(T = t|x) = 0.5$

$$2. \hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{0.5} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{0.5} =$$

Propensity scores – algorithm

Inverse probability of treatment weighted estimator

How to calculate ATE with propensity score

for sample $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$

1. Randomized trial $p = 0.5$

2.

$$\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{0.5} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{0.5} = \\ \frac{2}{n} \sum_{i \text{ s.t. } t_i=1} y_i - \frac{2}{n} \sum_{i \text{ s.t. } t_i=0} y_i$$

Propensity scores – algorithm

Inverse probability of treatment weighted estimator

How to calculate ATE with propensity score

for sample $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$

1. Randomized trial $p = 0.5$

Sum over $\sim \frac{n}{2}$ terms

2.

$$\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{0.5} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{0.5} =$$
$$\frac{2}{n} \sum_{i \text{ s.t. } t_i=1} y_i - \frac{2}{n} \sum_{i \text{ s.t. } t_i=0} y_i$$

Method

- Observational sample $(x^1, t^1, y^1), \dots, (x^n, t^n, y^n)$
- Estimate propensity scores $\widehat{e}^i = p(t^i = 1 | x^i)$
- $\widehat{ATE}_1 = \frac{1}{n} \sum_{i=1}^n \frac{y^i t^i}{\widehat{e}^i} - \frac{1}{n} \sum_{i=1}^n \frac{y^i (1-t^i)}{1-\widehat{e}^i}$
- $\widehat{ATE}_2 = \left(\sum_{i=1}^n \frac{t^i}{\widehat{e}^i} \right)^{-1} \frac{1}{n} \sum_{i=1}^n \frac{y^i t^i}{\widehat{e}^i} - \left(\sum_{i=1}^n \frac{1-t^i}{1-\widehat{e}^i} \right)^{-1} \sum_{i=1}^n \frac{y^i (1-t^i)}{1-\widehat{e}^i}$

Problems with IPW

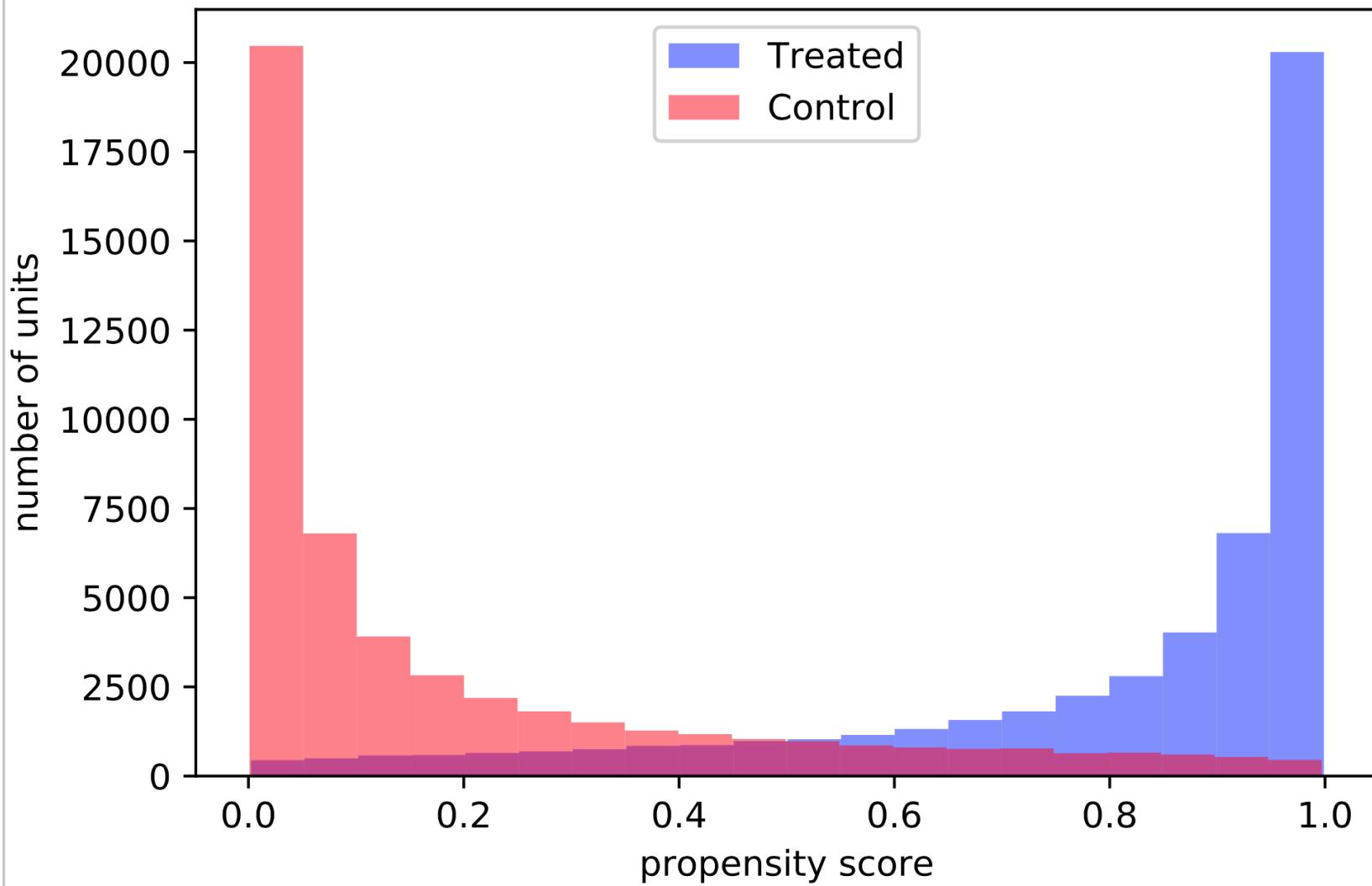
- Need to estimate propensity score (problem in all propensity score methods)
- Weighting by inverse – can create large variance and large errors for small propensity scores
 - Exacerbated when more than two treatments

Overlap

- If there's not much overlap, propensity scores become non-informative and easily mis-calibrated
- Sample variance of inverse propensity score re-weighting scales with $\sum_{i=1}^n \frac{1}{\hat{p}(T=1|x_i)\hat{p}(T=0|x_i)}$, which can grow very large when samples are non-overlapping
(Williamson et al., 2014)

Stratifying on the propensity score (sometimes called subclassifying)

- Divide sample into strata of propensity scores, say by quintiles
- Treat each strata as an RCT and obtain a strata-specific ATE estimate
- Pool the estimates together, e.g. by averaging



Adding propensity score to covariate adjustment

- You can do it
- It sometimes helps
- Sometimes it doesn't

Propensity score

- Important tool for understanding an observational study
- Widely used in practice
- Design without looking at outcomes, mimics RCTs
 - Can iterate: e.g. change models, without p-hacking/overfitting

Proof that any balancing score is finer than the propensity score

The proof below is adapted from [1].

Let $x \in \mathcal{X}$ be the covariates, (Y_0, Y_1) the potential outcomes, and $T \in \{0,1\}$ the treatment. Let $\mathcal{H}, \mathcal{H}'$ be arbitrary spaces, possibly, but not necessarily, \mathbb{R}^k .

Definition 1. For a pair of functions $a : \mathcal{X} \rightarrow \mathcal{H}$, $b : \mathcal{X} \rightarrow \mathcal{H}'$, we say that b is finer than a , if

$$a(x_1) \neq a(x_2) \Rightarrow b(x_1) \neq b(x_2).$$

Lemma 1. If there exists a function $f : \mathcal{H}' \rightarrow \mathcal{H}$ such that $a(x) = f(b(x))$, then b is finer than a .

Proof. Exercise. □

Definition 2. The propensity score is defined as:

$$e(x) = p(T = 1|x).$$

Definition 3. A balancing score is a function $b : \mathcal{X} \rightarrow \mathcal{H}$ such that for all $x \in \mathcal{X}$, $T \in \{0,1\}$:

$$x \perp\!\!\!\perp T \quad | \quad b(x).$$

Theorem 1. $b : \mathcal{X} \rightarrow \mathcal{H}$ is a balancing score if and only if there exists a function $f : \mathcal{H} \rightarrow \mathbb{R}$ such that $e(x) = f(b(x))$

Proof. We note that: $x \perp\!\!\!\perp T|b(x) \iff p(T = 1|b(x)) = p(T = 1|b(x), x)$. Because $b(x)$ is a function of x , we have that $p(T = 1|b(x), x) = p(T = 1|x) = e(x)$, where the last equality is by the definition of the propensity score $e(x)$. First, we prove that if there exists a function f such that $e(x) = f(b(x))$, then $x \perp\!\!\!\perp T|b(x)$. We have:

$$x \perp\!\!\!\perp T|b(x) \iff p(T = 1|b(x)) = e(x). \quad (1)$$

It now suffices to show that if there exists a function f such that $e(x) = f(b(x))$, then $p(T = 1|b(x)) = e(x)$. We have for all $\beta = b(x)$:

$$p(T = 1|b(x) = \beta) = \quad (2)$$

$$p(T = 1|\{x \text{ s.t. } b(x) = \beta\}) = \quad (3)$$

$$\mathbb{E}[p(T = 1|x)|b(x) = \beta] = \quad (4)$$

$$\mathbb{E}[e(x)|b(x) = \beta] =$$

$$\mathbb{E}[f(b(x))|b(x) = \beta] =$$

$$f(b(x)) = e(x).$$

Equality (4) is by the assumption $e(x) = f(b(x))$.

Now we prove that if $x \perp\!\!\!\perp T|b(x)$ then there exists a function f such that $e(x) = f(b(x))$. Specifically, let us assume that $p(T = 1|b(x)) = e(x)$ but there exists no function f such that $e(x) = f(b(x))$. If there exists no such function, then necessarily there exist $x_1 \neq x_2$ such that $e(x_1) \neq e(x_2)$, but $b(x_1) = b(x_2)$ (exercise: prove

why). Then we have that:

$$p(T = 1|x_2, b(x_2)) = \quad (5)$$

$$p(T = 1|x_2, b(x_1)) = \quad (6)$$

$$p(T = 1|x_2) \neq \quad (7)$$

$$p(T = 1|x_1) = \quad (8)$$

$$p(T = 1|x_1, b(x_1)).$$

(5) is by our assumption that $b(x_1) = b(x_2)$, (6) is by our assumption of independence of x and T conditioned on $b(x)$, (7) is by our assumption that $e(x_1) \neq e(x_2)$, and (8) is again by the conditional independence assumption. However, taken together we have that $p(T = 1|x_2, b(x_2)) \neq p(T = 1|x_1, b(x_1))$, which contradicts the conditional independence assumption $x \perp\!\!\!\perp T|b(x)$. We therefore conclude that there must exist a function f such that $e(x) = f(b(x))$. \square

Theorem 2. *For any balancing score $b : \mathcal{X} \rightarrow \mathcal{H}$, $(Y_1, Y_0) \perp\!\!\!\perp T|x \Rightarrow (Y_1, Y_0) \perp\!\!\!\perp T|b(x)$.*

Proof. We wish to prove that if $(Y_1, Y_0) \perp\!\!\!\perp T|x$, then $p(T = 1|Y_1, Y_0, b(x)) = p(T = 1|b(x))$. By (1), this is equivalent to showing that $p(T = 1|Y_1, Y_0, b(x)) = e(x)$. Recall that because b is a balancing score, by Theorem 1 there exists a function f such that $e(x) = f(b(x))$.

We have:

$$p(T = 1|Y_1, Y_0, b(x)) = \quad (9)$$

$$\mathbb{E}[p(T = 1|Y_1, Y_0, x)|Y_1, Y_0, x = b(x)] = \quad (10)$$

$$\mathbb{E}[p(T = 1|X)|Y_1, Y_0, x = b(x)] = \quad (11)$$

$$\mathbb{E}[e(x)|Y_1, Y_0, x = b(x)] = \quad (11)$$

$$\mathbb{E}[f(b(x))|Y_1, Y_0, x = b(x)] = \quad (12)$$

$$f(b(x)) = e(x) \quad (12)$$

Equality (10) is by our assumption $(Y_1, Y_0) \perp\!\!\!\perp T|x$, and equality (11) is by Theorem 1. \square

Taken together, Theorems 1 and 2 show that ignorability with x is equivalent to ignorability with $b(x)$ for any balancing score $b(x)$, and that any balancing score is finer than the propensity score $e(x)$.

References

- [1] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.



Introduction to Causal Inference - a Machine Learning Perspective

Dr. Uri Shalit

Course number 097400
2020-2021

Lesson 4

The stages of causal inference

1. Formulate *causal assumptions* sufficient to solve the problem
 - these are mostly **untestable**
2. Under the assumptions, reduce causal problem to appropriate statistical/machine learning method
 - these methods are often specialized methods, similar but distinct from familiar methods such as regression

Identification

Estimation

Potential outcomes

- Y_0, Y_1 : potential outcomes
- T : binary treatment
- X : observed covariates
- Y : observed outcome

Consistency assumption

$$Y = TY_1 + (1 - T)Y_0:$$

“The Assumptions”

Sufficient conditions for causal inference to be possible:

- 1. Stable Unit Treatment Value Assumption**
 - no interference
- 2. Consistency**
 - we see the correct potential outcome
- 3. Ignorability / No unmeasured confounders**
- 4. Common support**

The adjustment formula

Under the assumptions above we have that:

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

The adjustment formula

Under the assumptions above we have that:

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

Estimate with:

1. Covariate adjustment
2. Inverse propensity score weighting

Covariate adjustment

Covariate adjustment

Under the assumptions above we have that:

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

$$\left. \begin{array}{l} \textcolor{red}{\mathbb{E}[Y|x, T=1]} \\ \textcolor{blue}{\mathbb{E}[Y|x, T=0]} \end{array} \right\}$$

Quantities we
can hope to
estimate
from data

The adjustment formula

Under the assumption of ignorability,
we have that:

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

Empirically we have samples from
 $p(x|T=1)$ or $p(x|T=0)$

The adjustment formula

Under the assumption of ignorability,
we have that:

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

Empirically we have samples from
 $p(x|T=1)$ or $p(x|T=0)$

To extrapolate to $p(x)$ → common support assumption

Covariate adjustment (parametric g-formula)

- Explicitly model the relationship between treatment, confounders, and outcome
- Under ignorability, the expected causal effect of T on Y :
$$\mathbb{E}_{x \sim p(x)} [\text{red } \mathbb{E}[Y|T = 1, x] - \text{blue } \mathbb{E}[Y|T = 0, x]]$$
- Fit a model $f(x, t) \approx \mathbb{E}[Y|T = t, x]$

$$\widehat{ATE} = \frac{1}{n} \sum_{i=1}^n f(x_i, 1) - f(x_i, 0)$$

Covariate adjustment in practice

- Given data:

$$(x^1, t^1, y^1), \dots, (x^n, t^n, y^n)$$

1. Linear: Regress y on (x, t) , read off coefficient for t (regularize?)
2. Non-linear: Regress y on (x, t)
3. Regress separately over set with $t=1$ and set with $t=0$?
4. Active field of research developing specialized methods

Covariate adjustment in practice

- Given data:

$$(x^1, t^1, y^1), \dots, (x^n, t^n, y^n)$$

Covariate adjustment in practice

- Given data:

$$(x^1, t^1, y^1), \dots, (x^n, t^n, y^n)$$

- Regress the observed y 's on (x, t) such that

$$y_i \approx f(x_i, t_i)$$
$$\widehat{ATE} = \frac{1}{n} \sum_{i=1 \dots n} f(x_i, 1) - f(x_i, 0)$$

Covariate adjustment in practice

- Given data:

$$(x^1, t^1, y^1), \dots, (x^n, t^n, y^n)$$

- Regress the observed y's on (x,t) such that

$$y_i \approx f(x_i, t_i)$$

$$\widehat{ATE}_1 = \frac{1}{n} \sum_{i=1 \dots n} f(x_i, 1) - f(x_i, 0)$$

$$\widehat{ATE}_2 = \frac{1}{\sum t_i} \sum_{i \text{ s.t. } t_i=1} y_i - f(x_i, 0) + \frac{1}{\sum 1 - t_i} \sum_{i \text{ s.t. } t_i=1} f(x_i, 1) - y_i$$

Linear model

- Assume that:

$$\begin{array}{c} \text{blood pressure} \\ Y_t(x) = \beta^T x + \gamma \cdot t + \epsilon_t \\ \text{age, weight, ...} \\ \mathbb{E}[\epsilon_t] = 0 \\ \text{medication} \end{array}$$

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

- We care about γ , not about $Y_t(x)$

Estimation, not prediction

Linear model

blood pressure **age,weight,...** **medication**

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

Hypertension is affected by many variables:
lifestyle, weight, genetics, age

- Each of these often stronger **predictor** of blood-pressure, compared with type of medication taken
- Regularization (e.g. Lasso) might remove the treatment variable!
- Features → (“nuisance parameters”, “variable of interest”)

Regression - 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:

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

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

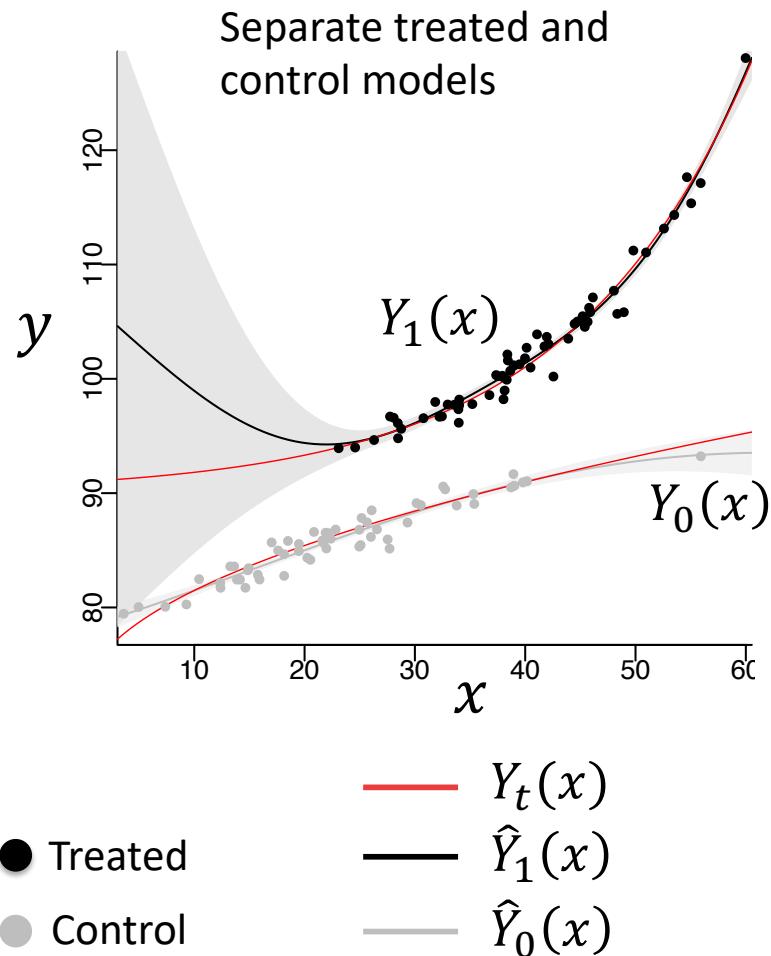
Using machine learning for causal inference under ignorability

- Machine learning techniques can be very useful and have recently seen wider adoption
- Random forests and Bayesian trees
Hill (2011), Athey & Imbens (2015), Wager & Athey (2015)
- Gaussian processes
Hoyer et al. (2009), Zigler et al. (2012)
- Neural nets
Beck et al. (2000), Johansson et al. (2016), Shalit et al. (2016), Lopez-Paz et al. (2016)
- “Causal” Lasso
Belloni et al. (2013), Farrell (2015), Athey et al. (2016)

Using machine learning for causal inference under ignorability

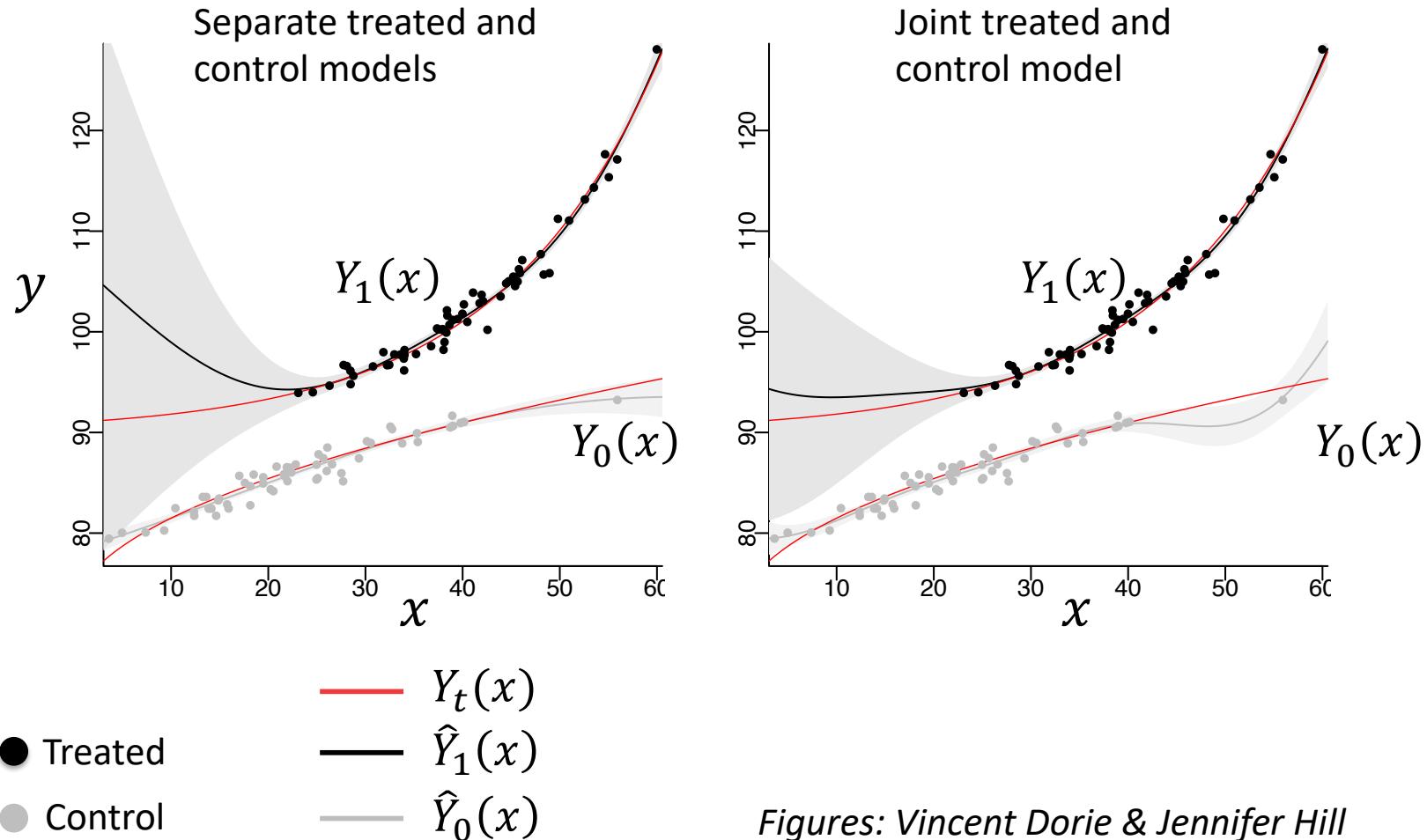
- Machine learning techniques can be very useful and have recently seen wider adoption
- How is the treatment variable used:
 - Fit two different models for treated and control?
 - Not regularized?
 - Privileged

Example: Gaussian process



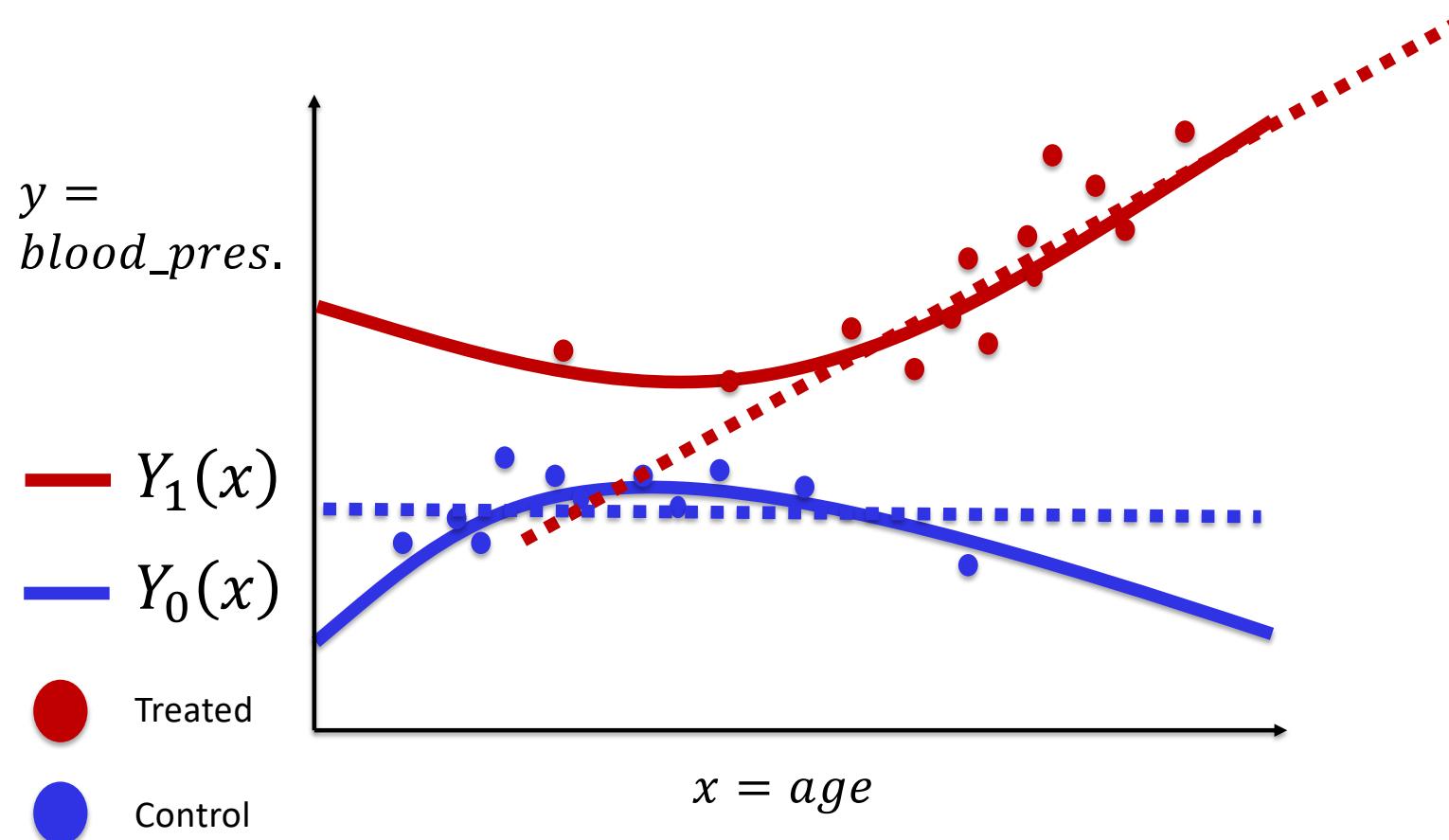
Figures: Vincent Dorie & Jennifer Hill

Example: Gaussian process



Figures: Vincent Dorie & Jennifer Hill

Covariate adjustment: weak overlap



Inverse propensity score weighting

Propensity score

- Extremely widely used tool
- Basic idea: turn observational study into a pseudo-randomized trial by correcting for non-random sampling
- The propensity score:
$$e(x) = p(T = 1|x)$$
- Theorem:
If ignorability holds for x , then $e(x)$ is the coarsest function of x for which ignorability still holds

What does the propensity score give us?

- If we have ignorability, in theory the propensity score gives all everything we need
- We can run covariate adjustment on the propensity score!
$$\mathbb{E}[Y|e(x), T = 1] - \mathbb{E}[Y|e(x), T = 0]$$
- Other method using propensity which we will see soon:
 - Inverse propensity score weighting
 - Stratification on the propensity score
 - Propensity score matching

The propensity score

- $e(x) = p(T = 1|x)$, the treatment assignment mechanism
- In most cases must be estimated from data
- Can use any machine learning method:
logistic regression, random forests, neural nets
- Unlike most ML applications, we need to get the **probability** itself accurately
- Subtle point: if we include x which are only predictive of treatment assignment but not outcome
- Hard (but not impossible) to validate models

Exact propensity scores

- RCTs
- Computational advertising
- In general whenever we know exactly what generated past actions, example an agent for which we have the true model

Another view of propensity score

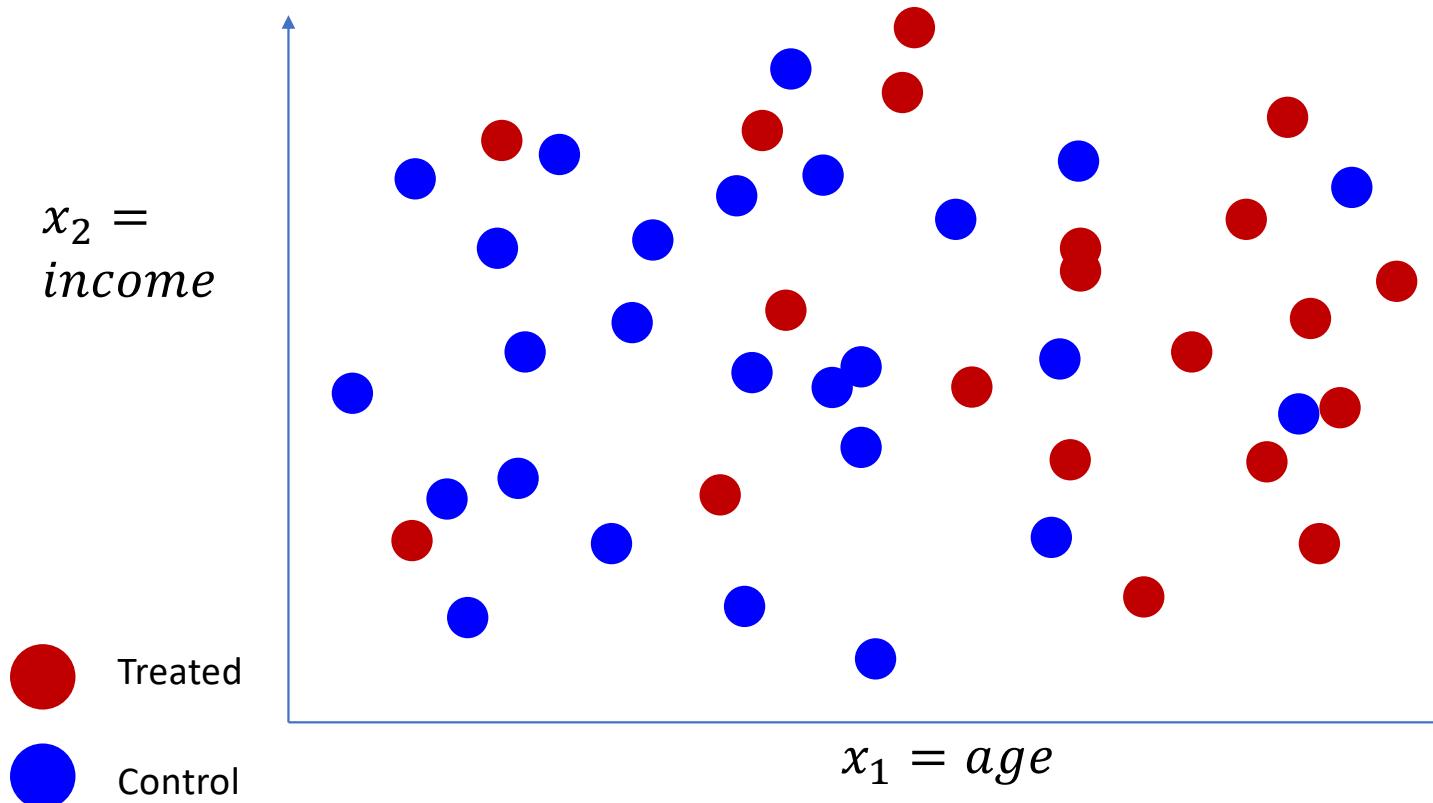
- RCTs are easy to analyze because we know the probability of each sample getting treated
- If $e(x) = 0.5$ (RCT with equal chance for treated and control) then we weight each sample by 0.5
- In observational study, we have “biased randomizations”. We can invert them to turn the observational study in a randomized study*
* with caveats

Inverse Probability Weighting (IPW) with propensity scores

Inverse propensity score re-weighting

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

reweighted control reweighted treated



Propensity scores - derivation

- Recall average treatment effect:

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

- We only have samples for:

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

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

Propensity scores - derivation

- We only have samples for:

$$\mathbb{E}_{x \sim p(x|T=1)} [\textcolor{red}{\mathbb{E}[Y|x, T=1]}]$$

$$\mathbb{E}_{x \sim p(x|T=0)} [\textcolor{blue}{\mathbb{E}[Y|x, T=0]}]$$

Propensity scores - derivation

- We only have samples for:

$$\mathbb{E}_{x \sim p(x|T=1)} [\mathbb{E}[Y|x, T=1]]$$

$$\mathbb{E}_{x \sim p(x|T=0)} [\mathbb{E}[Y|x, T=0]]$$

- We need to turn $p(x|T=1)$ into $p(x)$:

$$p(x|T=1) \cdot ? = p(x)$$

Propensity scores - derivation

- We only have samples for:

$$\mathbb{E}_{x \sim p(x|T=1)} [\mathbb{E}[Y|x, T=1]]$$

$$\mathbb{E}_{x \sim p(x|T=0)} [\mathbb{E}[Y|x, T=0]]$$

- We need to turn $p(x|T=1)$ into $p(x)$:

$$p(x|T=1) \cdot \frac{p(T=1)}{p(T=1|x)} = p(x)$$

Propensity score

Propensity scores - derivation

- We only have samples for:

$$\mathbb{E}_{x \sim p(x|T=1)} [\mathbb{E}[Y|x, T=1]]$$

$$\mathbb{E}_{x \sim p(x|T=0)} [\mathbb{E}[Y|x, T=0]]$$

- We need to turn $p(x|T=0)$ into $p(x)$:

$$p(x|T=0) \cdot \frac{p(T=0)}{p(T=0|x)} = p(x)$$

Propensity score

- We only have samples for:

$$\mathbb{E}_{x \sim p(x|T=1)} [\mathbb{E} [Y | x, T = 1]]$$

- We want:

$$\mathbb{E}_{x \sim p(x)} [\mathbb{E} [Y | x, T = 1]]$$

- We know that:

$$p(x|T = 1) \cdot \frac{p(T = 1)}{p(T = 1|x)} = p(x)$$

- Then:

$$\mathbb{E}_{x \sim p(x|T=1)} \left[\frac{p(T = 1)}{p(T = 1|x)} \mathbb{E} [Y | x, T = 1] \right] =$$

$$\mathbb{E}_{x \sim p(x)} [\mathbb{E} [Y | x, T = 1]]$$

Propensity scores – algorithm

Inverse probability of treatment weighted estimator

How to calculate ATE with propensity score

for sample $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$

1. Use any ML method to estimate $\hat{p}(T = t|x)$

2.

$$\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } \textcolor{red}{t_i=1}} \frac{y_i}{\hat{p}(t_i = 1|x_i)} - \frac{1}{n} \sum_{i \text{ s.t. } \textcolor{blue}{t_i=0}} \frac{y_i}{\hat{p}(t_i = 0|x_i)}$$

Propensity scores – algorithm

Inverse probability of treatment weighted estimator

How to calculate ATE with propensity score

for sample $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$

1. Randomized trial $p(T = t|x) = 0.5$

$$2. \hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } \textcolor{red}{t_i=1}} \frac{y_i}{\hat{p}(t_i = 1|x_i)} - \frac{1}{n} \sum_{i \text{ s.t. } \textcolor{blue}{t_i=0}} \frac{y_i}{\hat{p}(t_i = 0|x_i)}$$

Propensity scores – algorithm

Inverse probability of treatment weighted estimator

How to calculate ATE with propensity score

for sample $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$

1. Randomized trial $p(T = t|x) = 0.5$

$$2. \hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{0.5} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{0.5} =$$

Propensity scores – algorithm

Inverse probability of treatment weighted estimator

How to calculate ATE with propensity score

for sample $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$

1. Randomized trial $p = 0.5$

2.

$$\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{0.5} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{0.5} = \\ \frac{2}{n} \sum_{i \text{ s.t. } t_i=1} y_i - \frac{2}{n} \sum_{i \text{ s.t. } t_i=0} y_i$$

Propensity scores – algorithm

Inverse probability of treatment weighted estimator

How to calculate ATE with propensity score

for sample $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$

1. Randomized trial $p = 0.5$

Sum over $\sim \frac{n}{2}$ terms

2.

$$\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{0.5} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{0.5} =$$
$$\frac{2}{n} \sum_{i \text{ s.t. } t_i=1} y_i - \frac{2}{n} \sum_{i \text{ s.t. } t_i=0} y_i$$

Method

- Observational sample $(x^1, t^1, y^1), \dots, (x^n, t^n, y^n)$
- Estimate propensity scores $\widehat{e}^i = p(t^i = 1 | x^i)$
- $\widehat{ATE}_1 = \frac{1}{n} \sum_{i=1}^n \frac{y^i t^i}{\widehat{e}^i} - \frac{1}{n} \sum_{i=1}^n \frac{y^i (1-t^i)}{1-\widehat{e}^i}$
- $\widehat{ATE}_2 = \left(\sum_{i=1}^n \frac{t^i}{\widehat{e}^i} \right)^{-1} \sum_{i=1}^n \frac{y^i t^i}{\widehat{e}^i} - \left(\sum_{i=1}^n \frac{1-t^i}{1-\widehat{e}^i} \right)^{-1} \sum_{i=1}^n \frac{y^i (1-t^i)}{1-\widehat{e}^i}$

Problems with IPW

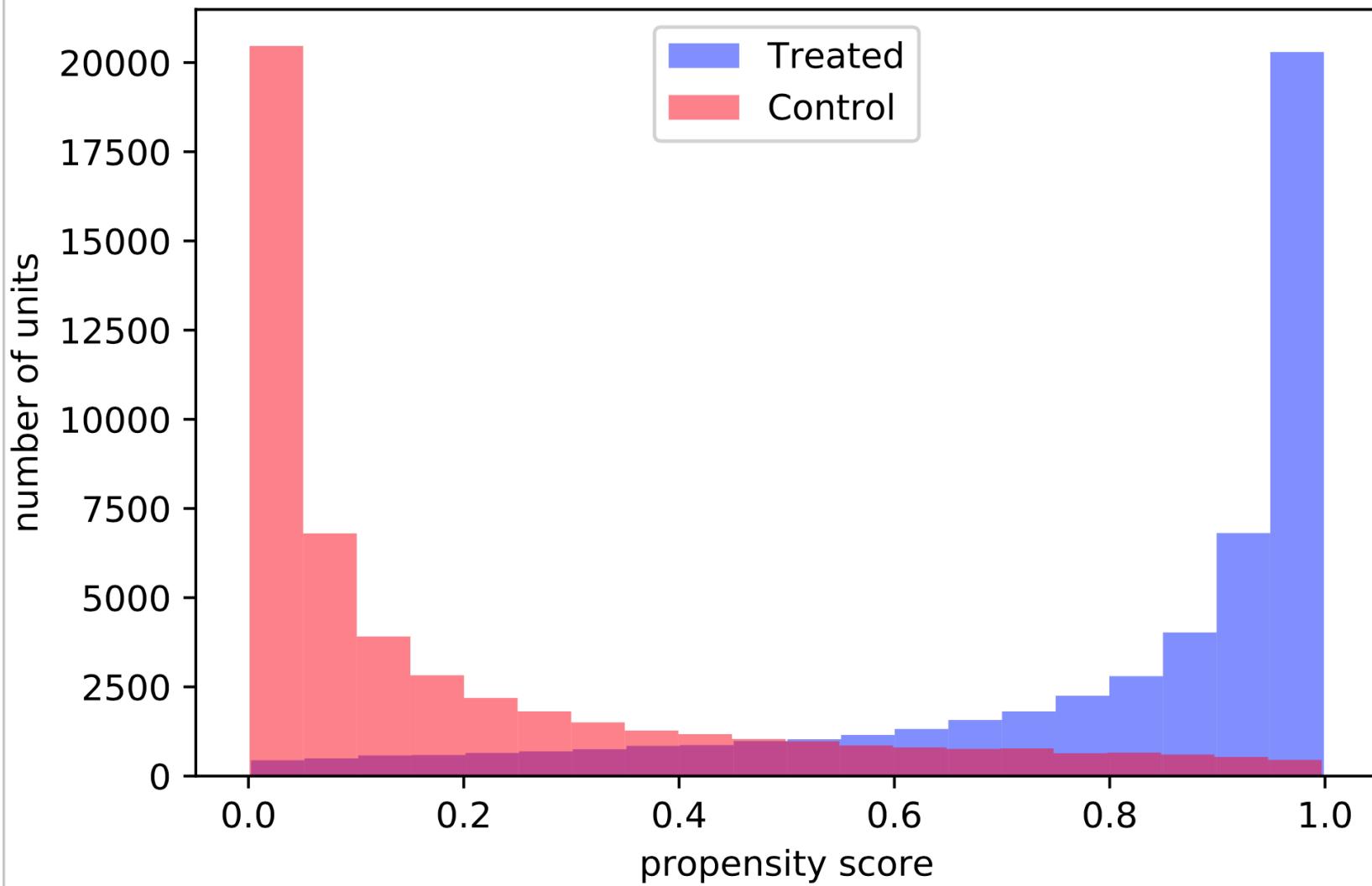
- Need to estimate propensity score (problem in all propensity score methods)
- Weighting by inverse – can create large variance and large errors for small propensity scores
 - Exacerbated when more than two treatments

Common support (Overlap)

- If there's not much overlap, propensity scores become non-informative and easily mis-calibrated
- Sample variance of inverse propensity score re-weighting scales with $\sum_{i=1}^n \frac{1}{\hat{p}(T=1|x_i)\hat{p}(T=0|x_i)}$, which can grow very large when samples are non-overlapping
(Williamson et al., 2014)

Stratifying on the propensity score (sometimes called subclassifying)

- Divide sample into strata of propensity scores, say by quintiles
- Treat each strata as an RCT and obtain a strata-specific ATE estimate
- Pool the estimates together, e.g. by averaging



Adding propensity score to covariate adjustment

- You can do it
- It sometimes helps
- Sometimes it doesn't

Problems with covariate adjustment

- It's too easy!
- Should think thoroughly about what covariates go in – that is the most important decision (often more than which algorithm)
- Example of mistakes:
 - including post-treatment covariates (leads to zero-bias)
 - Including covariates that only influence treatment and not outcome (leads to high-variance)
 - Using a model not specified for causal inference (leads to statistical inefficiency)

Example: post-treatment mistake

- Did we include a post-treatment covariate?
example: weight measured after treatment
- Measuring the post-treatment weight could explain away some of the causal effect, inducing bias in our estimation of the causal effect
- Conditioning on post-treatment covariates violates ignorability
- But how do we know if a variable is post-treatment?
- We will look into this issue again further in the course

Matching

Matching

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

Matching

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



Obama, had he gone to law school

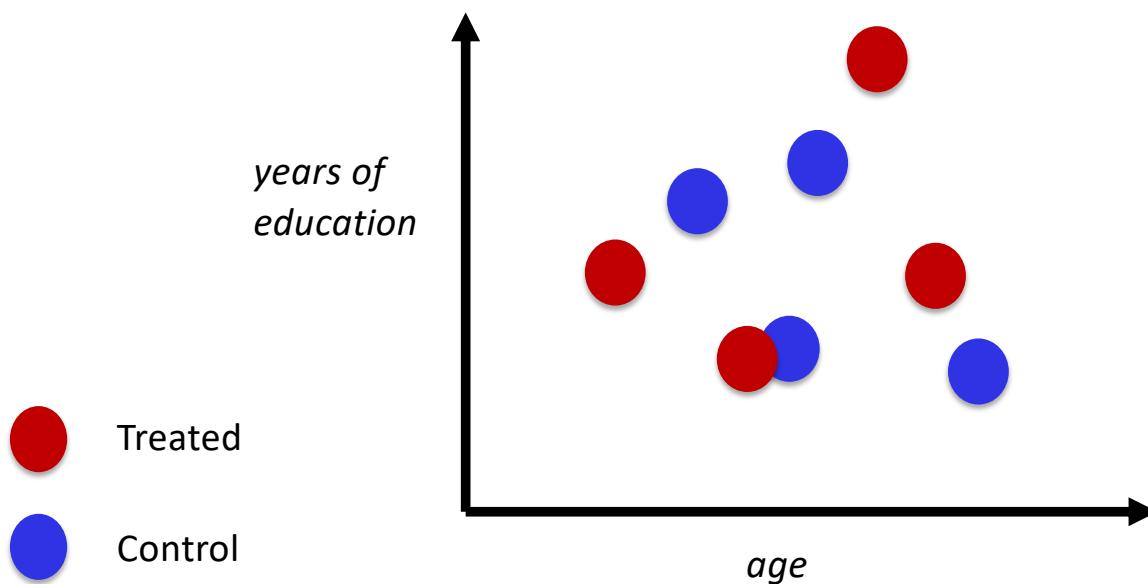


Obama, had he gone to business school

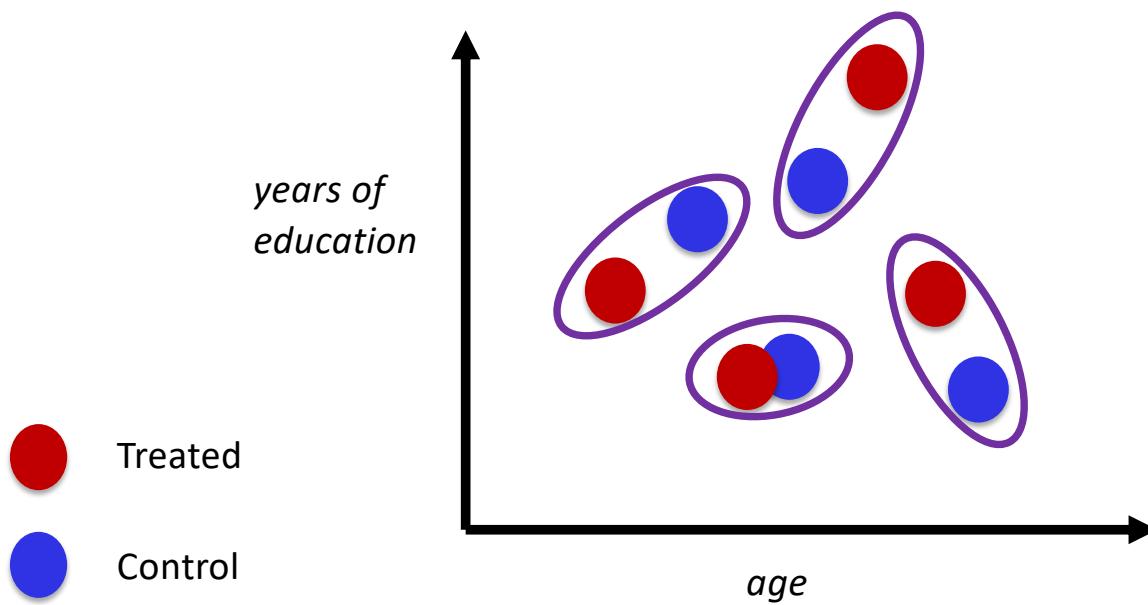
Matching

- For each unit with covariates x and treatment t , find a close partner with covariates $x' \approx x$ and treatment $1 - t$

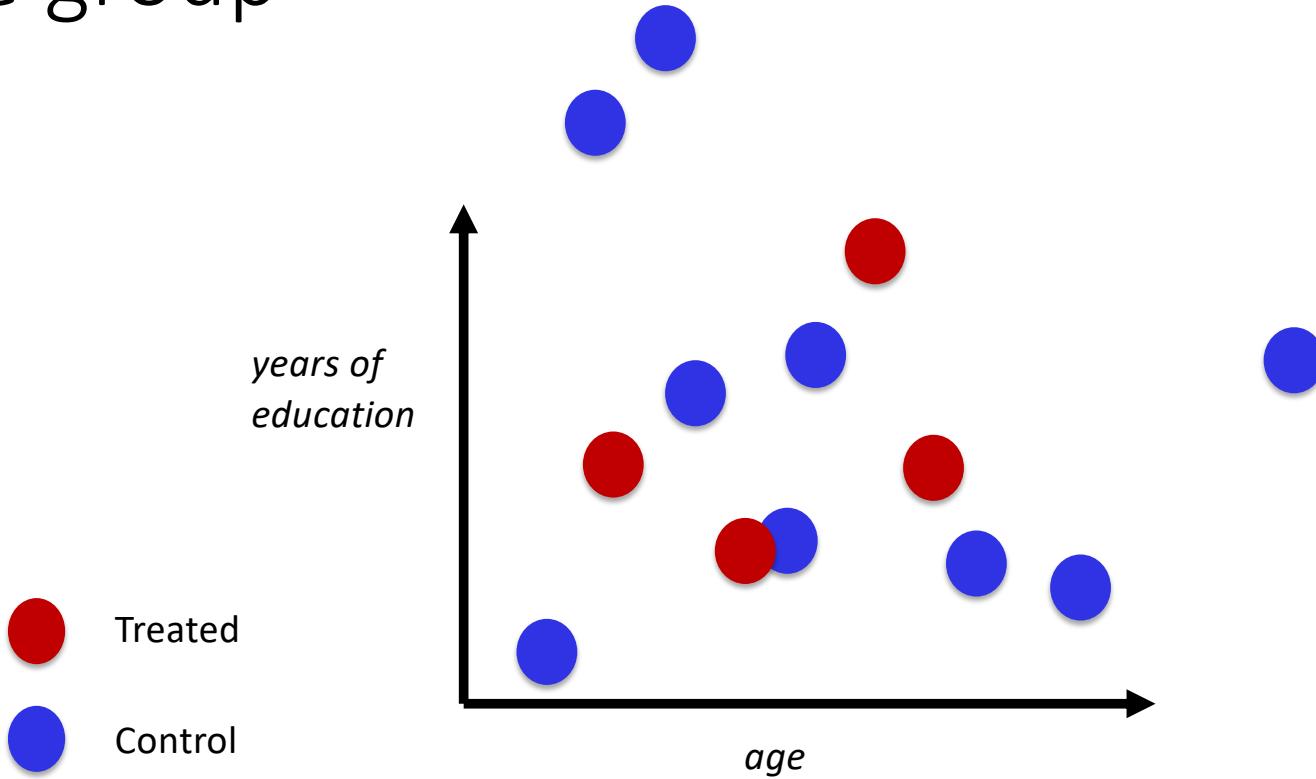
Match to nearest neighbor from opposite group



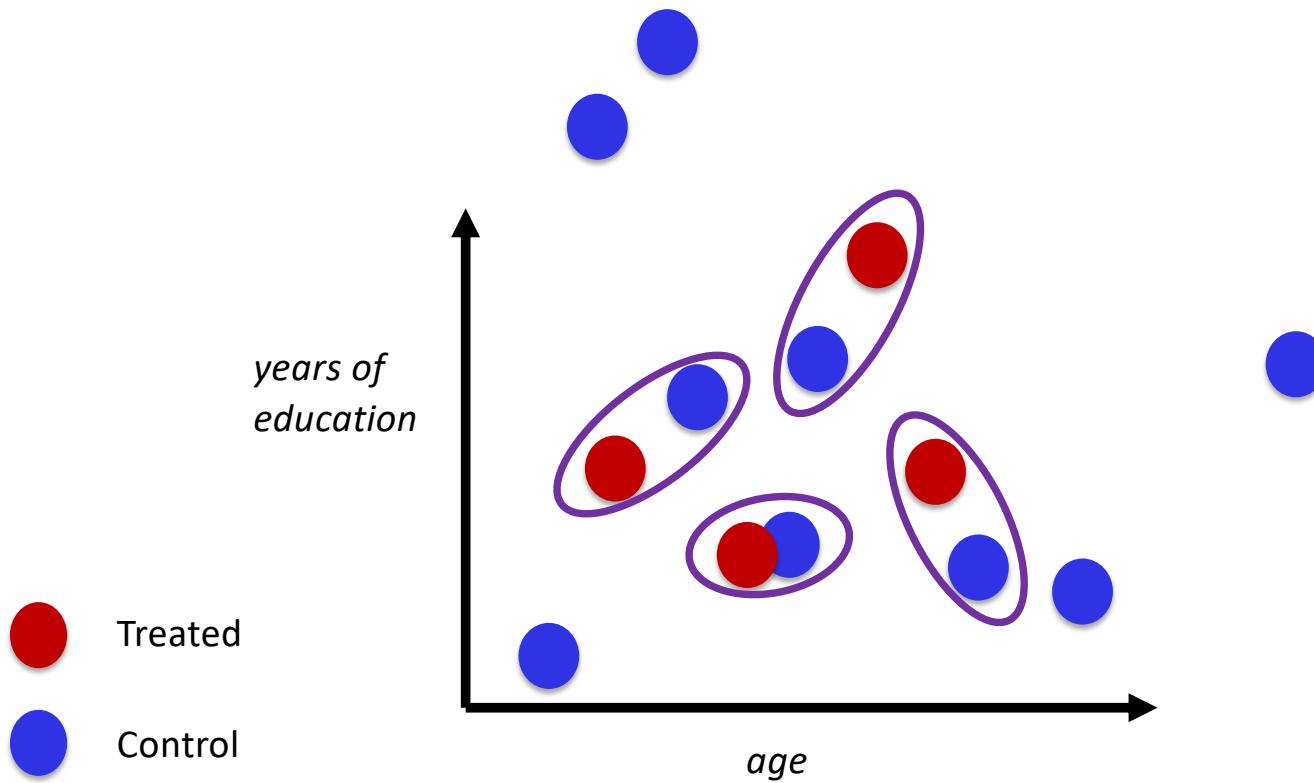
Match to nearest neighbor from opposite group



Match to nearest neighbor from
opposite group



Can remove “unnecessary” units



1-NN Matching

- Let $d(\cdot, \cdot)$ be a metric between x 's
- For each i , define $j(i) = \underset{j \text{ s.t. } t_j \neq t_i}{\operatorname{argmin}} d(x_j, x_i)$
 $j(i)$ is the nearest counterfactual neighbor of i
- $t_i = 1$, unit i is treated:
- $\widehat{ITE}(i) = y_i - y_{j(i)}$
- $t_i = 0$, unit i is control:
- $\widehat{ITE}(i) = y_{j(i)} - y_i$
- $\widehat{ATE} = \frac{1}{n} \sum_{i=1}^n \widehat{ITE}(i)$

General Matching

- For each i , define $J(i) = \{j \mid j \text{ is close to } i, t_j \neq t_i\}$
 $J(i)$ are a set close counterfactual neighbor of i
- $t_i = 1$, unit i is treated:
- $\widehat{ITE}(i) = y_i - \frac{1}{|J(i)|} \sum_{j \in J(i)} y_j$
- $t_i = 0$, unit i is control:
- $\widehat{ITE}(i) = \frac{1}{|J(i)|} \sum_{j \in J(i)} y_j - y_i$
- $\widehat{ATE} = \frac{1}{n} \sum_{i=1}^n \widehat{ITE}(i)$

Advantages of matching

- Design without looking at outcomes, mimics RCTs
 - Can iterate: e.g. change metrics, without p-hacking/overfitting
- Interpretable

Which covariates to use?

- In general everything that could be a confounder
- However, when number of covariates becomes greater than 50 or 100, some issues crop up
- Very common: matching on the propensity score itself
- However, matching on full covariates is often stronger in the sense of less variance

“Why Propensity Scores Should Not Be Used for Matching?” King & Nielsen 2016

Common metrics

- $x = (x_1, x_2, \dots, x_d), x' = (x'_1, x'_2, \dots, x'_d)$

- Euclidean

$$d(x, x') = \sqrt{\sum_i (x_i - x'_i)^2} = \sqrt{x^\top x'}$$

- L1

$$d(x, x') = \sum_i |x_i - x'_i|$$

- Mahalanobis

$$d(x, x') = \sqrt{x^\top \Sigma^{-1} x'} \text{ where } \Sigma \text{ is the data covariance matrix}$$

- Cosine distance (not a proper metric, no triangle inequality)

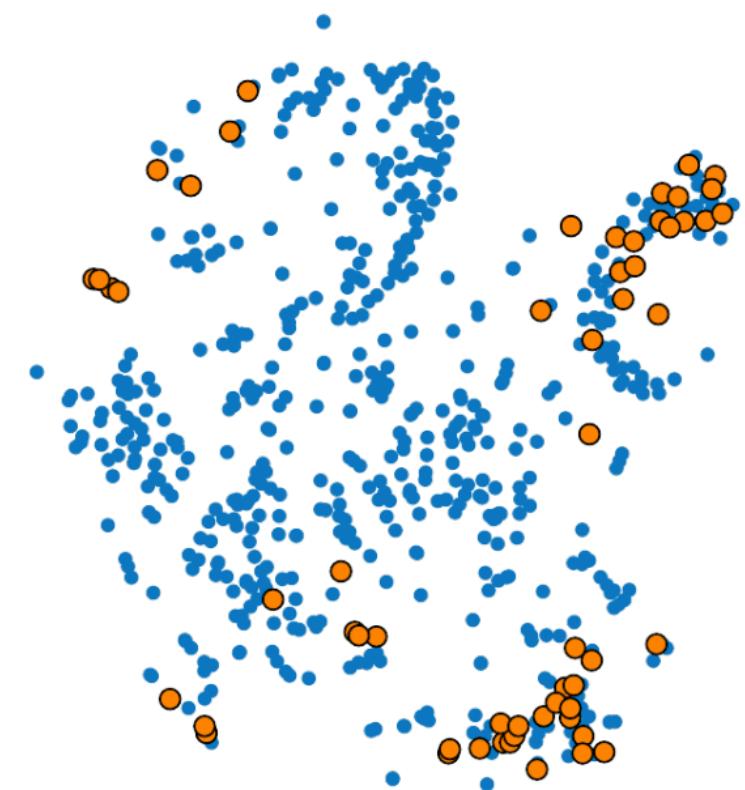
$$d(x, x') = 1 - \frac{x^\top x'}{\|x\| \cdot \|x'\|}$$

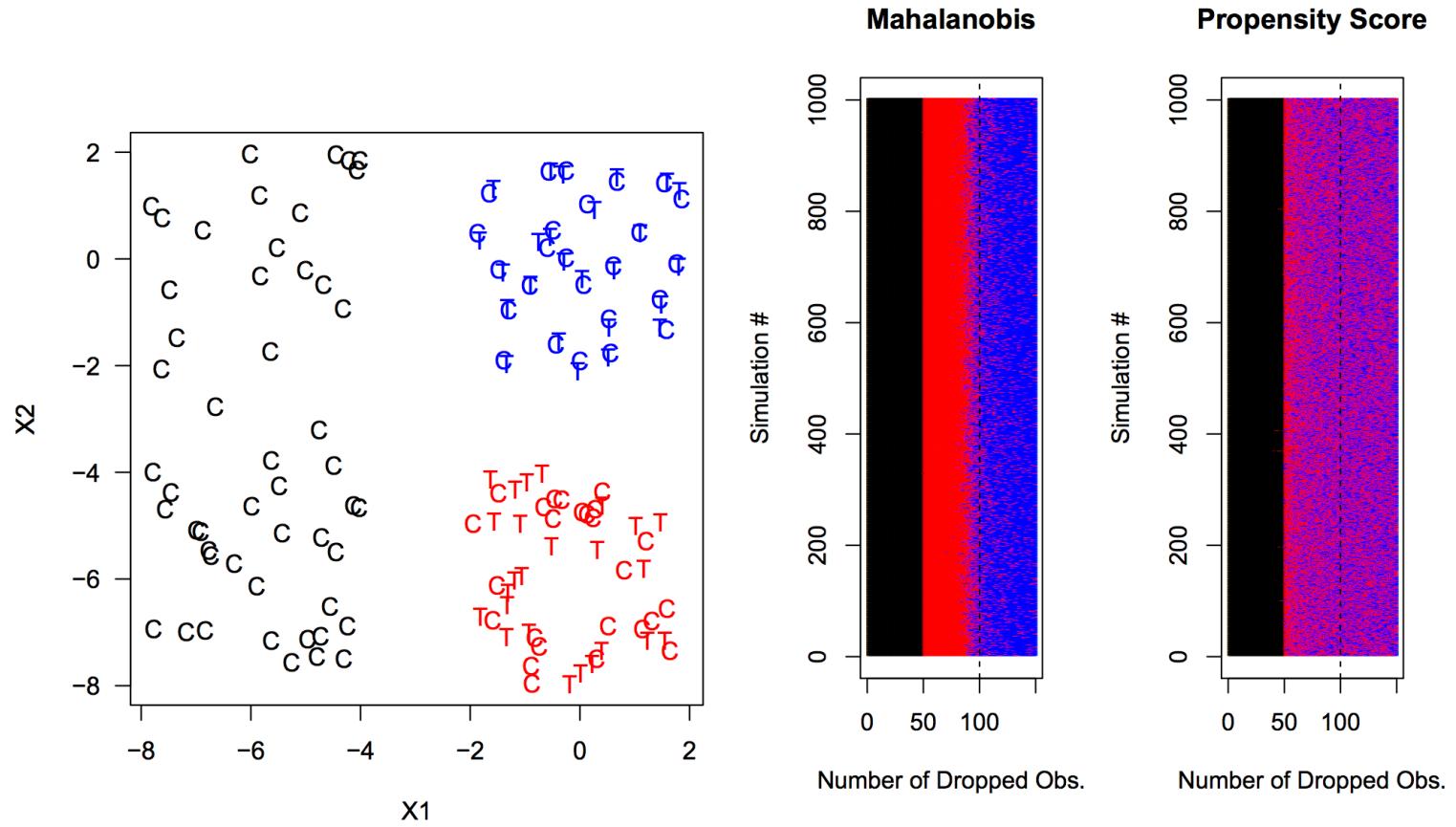
Metrics

- Make sure all covariates are on the same scale
- What to do with many weakly relevant features vs. few highly relevant?

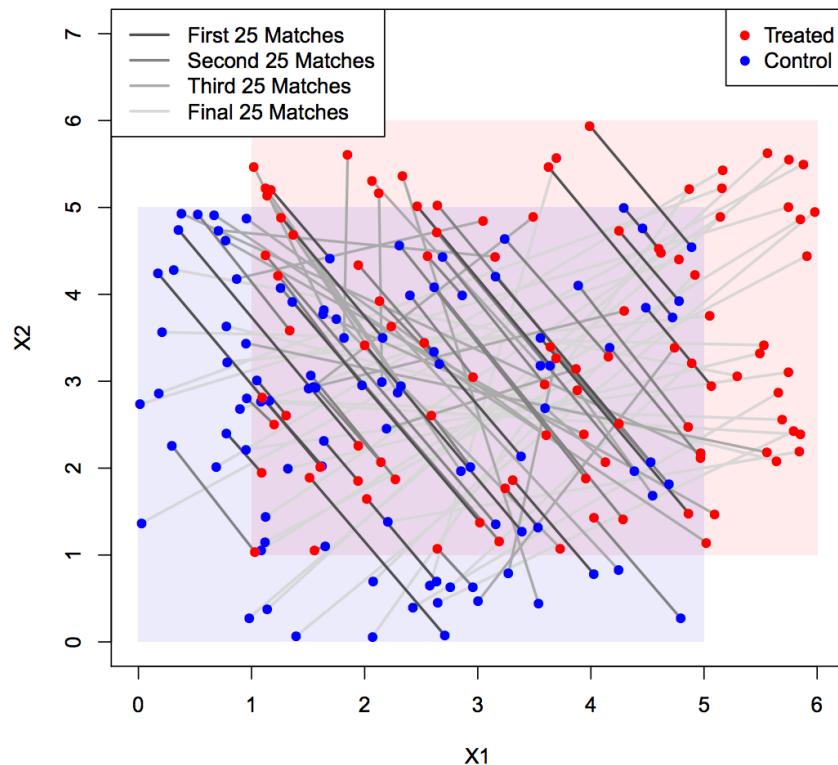
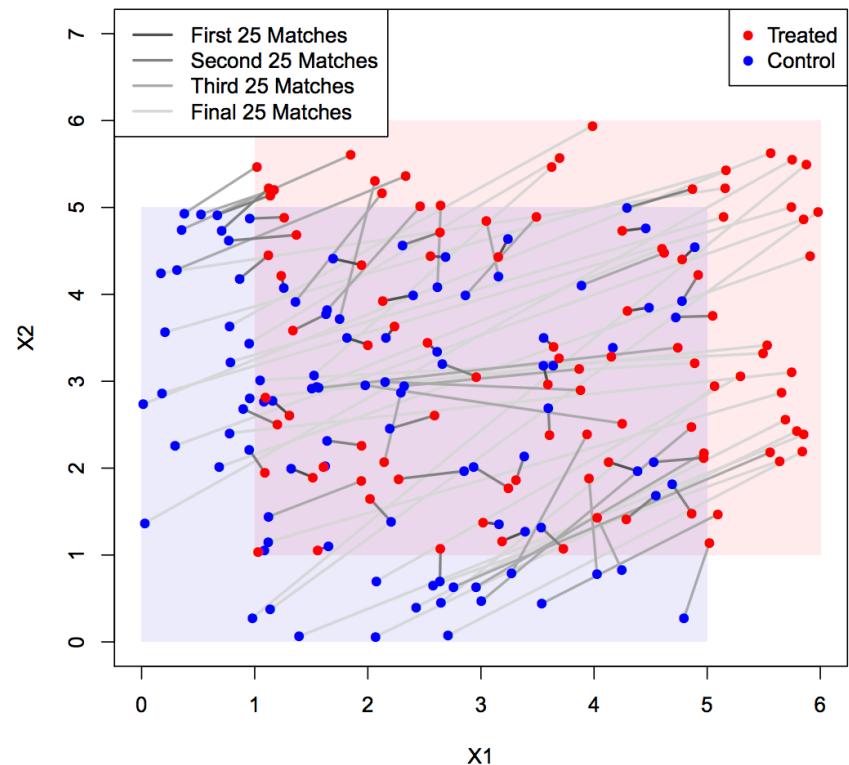
Matching methods

- k-Nearest Neighbor matching
Each treated is matched with exactly k controls, with or without replacement.
- The effect of k: bias-variance tradeoff
- Some samples might be unused, but not always a big issue



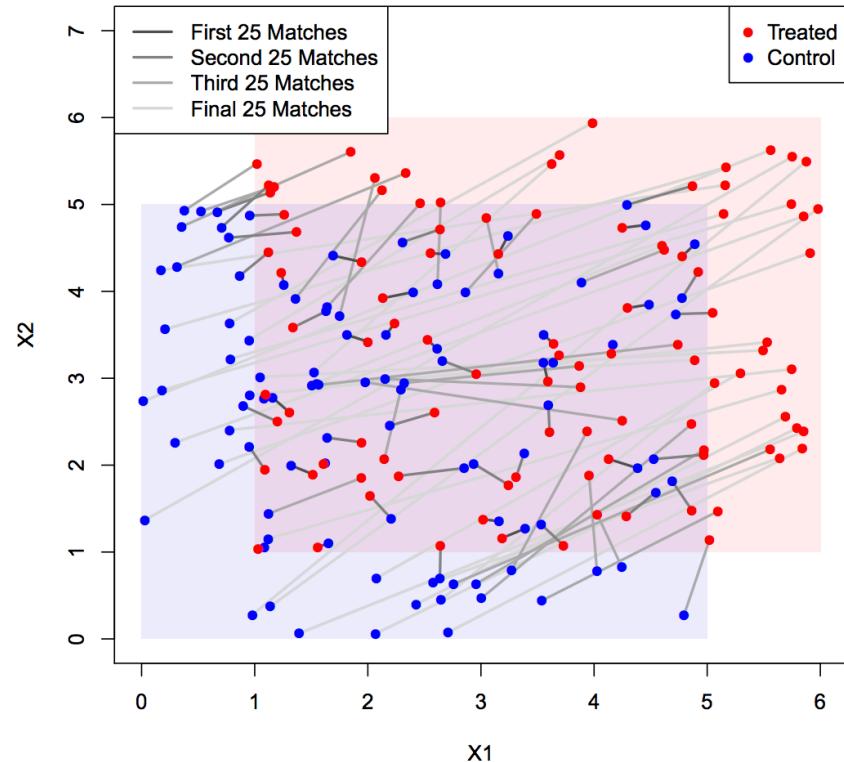


“Why Propensity Scores Should Not Be Used for Matching?”
 King & Nielsen 2016

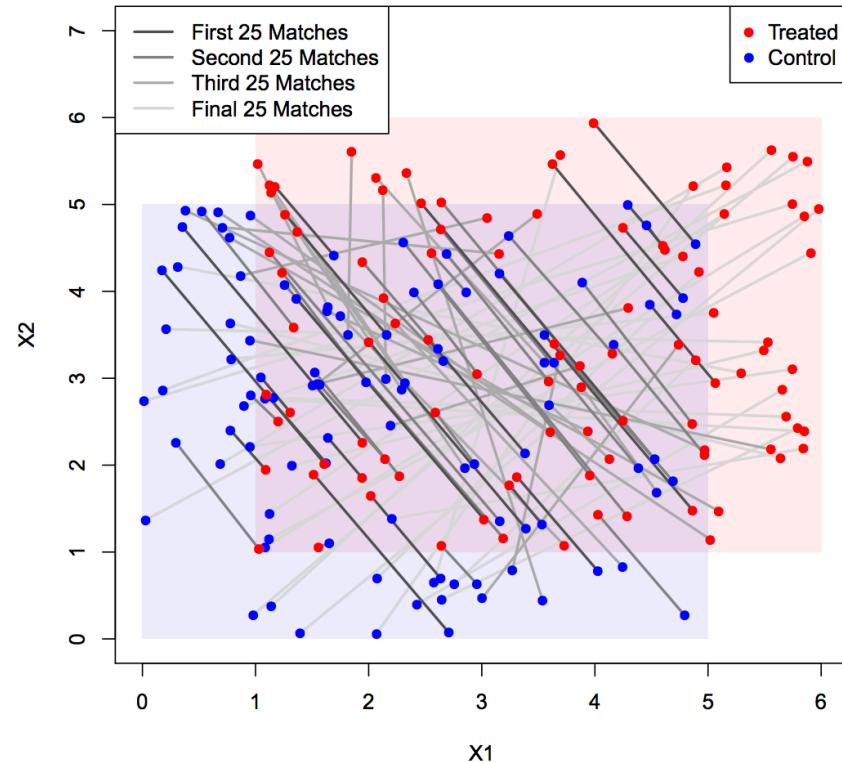


“Why Propensity Scores Should Not Be Used for Matching?”
King & Nielsen 2016

Full covariate matching with
Mahalanobis distance



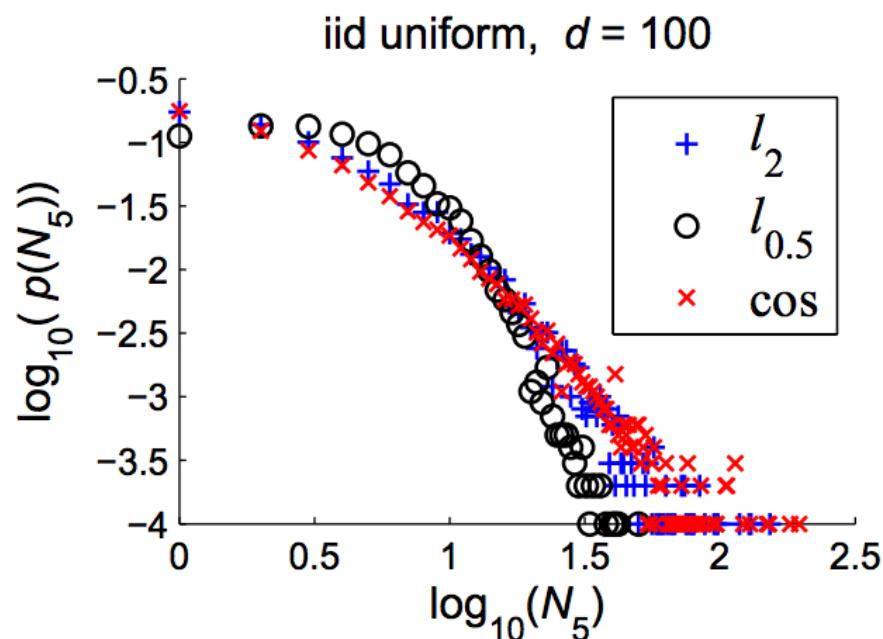
Propensity score matching



“Why Propensity Scores Should Not Be Used for Matching?”
King & Nielsen 2016

Curse of dimensionality for nearest neighbors

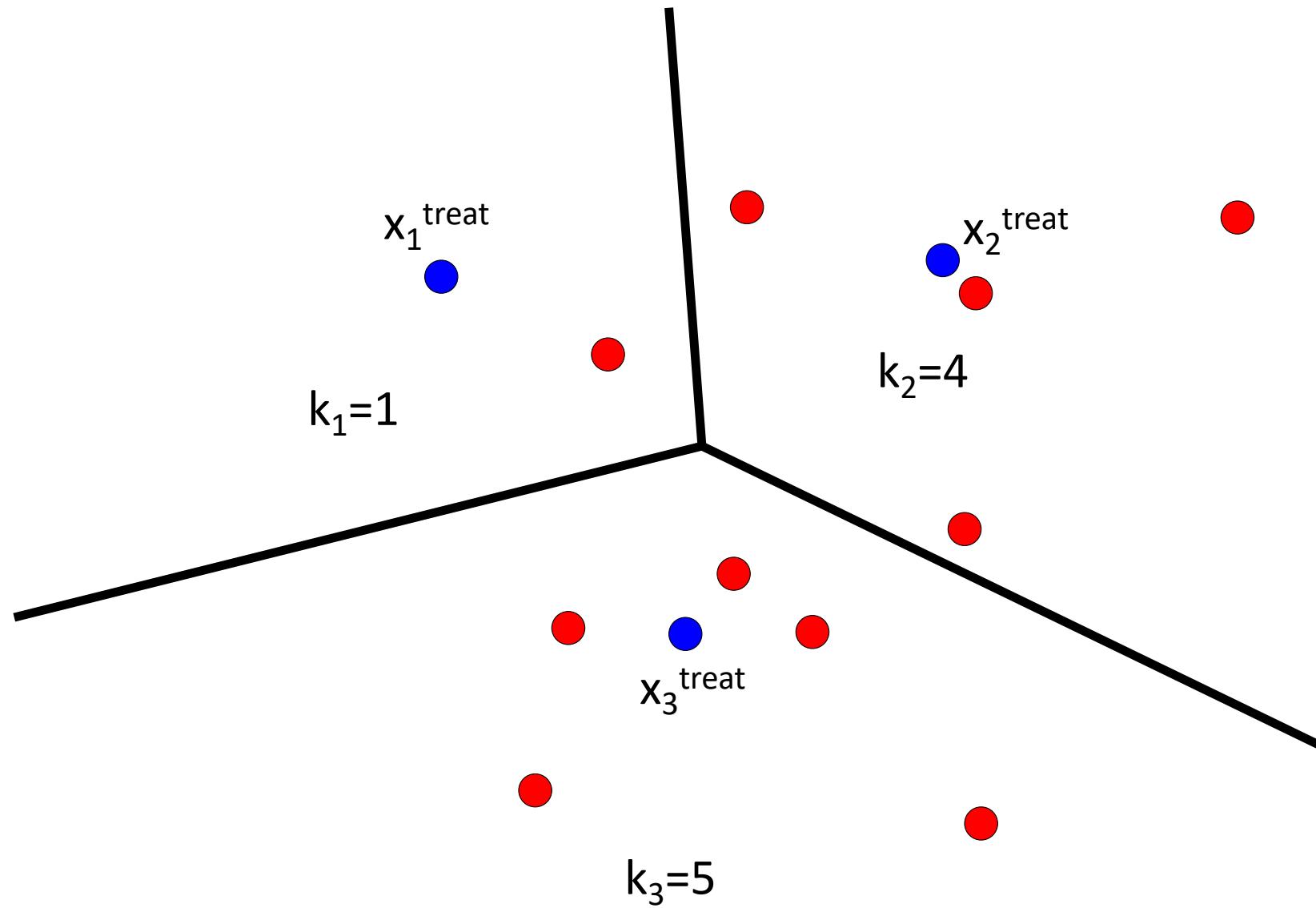
- A small number of points are neighbors of everyone
- In general distances are less informative in high-dimensions



Hubs in Space: Popular Nearest
Neighbors in High-Dimensional Data,
Radovanović et al.,
JMLR 2010

Matching methods

- Optimal matching: minimize sum of distances between all pairs
- Related to optimal transport, solvable by linear programming



How to choose a metric and matching?

- Domain knowledge, e.g. expert opinion on matches
- Relative size of control and treated groups
- Checking for “balance”

What is balance?

- Means the treated and control have the same marginal distributions for most features
- Standardized differences

*Using Mixed Integer Programming for Matching in an Observational Study of Kidney Failure after Surgery,
Zubizarreta (2012)*

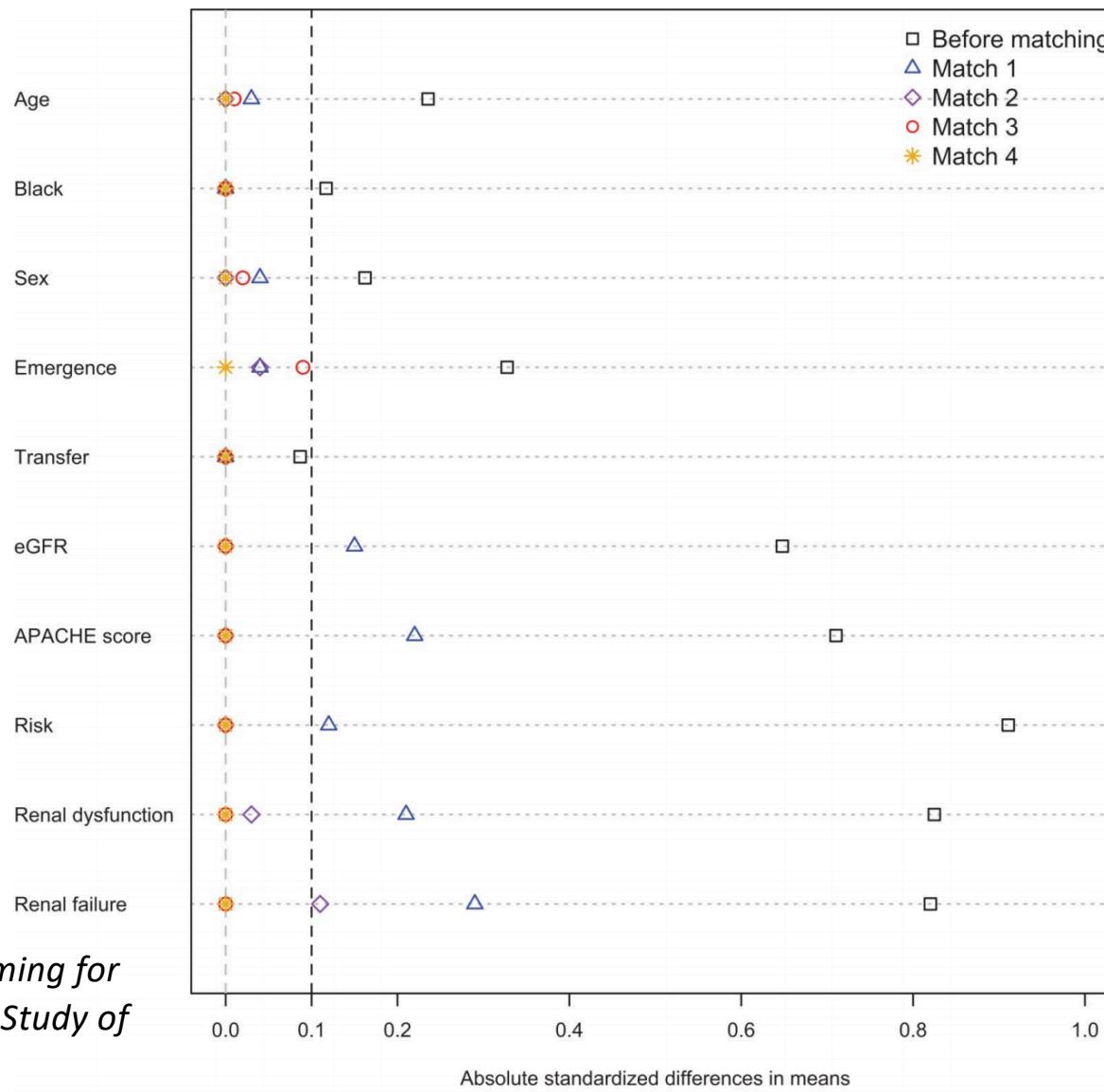
Covariate	Before matching			Match 1		
	ω_i	$\bar{x}_{T,i}$	$\bar{x}_{C,i}$	ω_i	$\bar{x}_{T,i}$	$\bar{x}_{C,i}$
Age	-	74.19	73.22	0	74.19	74.29
Black	-	0.12	0.08	0	0.12	0.12
Sex	-	0.39	0.47	0	0.39	0.37
Emergency	-	0.32	0.18	0	0.32	0.31
Transfer	-	0.02	0.01	0	0.02	0.02
eGFR	-	54.21	73.60	0	54.21	58.82
APACHE	-	35.73	28.69	0	35.73	33.56
Risk	-	-2.97	-3.82	0	-2.97	-3.08
Dysfunction	-	0.33	0.04	0	0.33	0.26
Failure	-	0.31	0.03	0	0.31	0.21

Table 2. Covariate imbalance before and after matching

Covariate	Covariate mean			Standardized difference		2-sample P-value	
	New	Ex-B	Ex-A	Before	After	Before	After
Sample size	6,260	123,846	6,260				
Age	77.883	76.992	77.926	0.116	-0.005	0.000	0.617
Male	0.345	0.358	0.346	-0.027	-0.003	0.038	0.880
ER-admit	0.538	0.323	0.537	0.444	0.003	0.000	0.886
Transfer	0.008	0.008	0.007	0.000	0.013	1.000	0.532
Risk	0.042	0.030	0.040	0.214	0.031	0.000	0.237
CHF	0.149	0.123	0.143	0.076	0.019	0.000	0.311
Liver	0.043	0.036	0.038	0.035	0.026	0.005	0.161
Cancer	0.164	0.175	0.164	-0.029	0.001	0.030	0.981
Past A	0.170	0.171	0.161	-0.002	0.024	0.880	0.178
Diabetes	0.189	0.197	0.199	-0.019	-0.024	0.145	0.198
Renal	0.069	0.058	0.064	0.046	0.020	0.000	0.282
COPD	0.167	0.147	0.160	0.055	0.019	0.000	0.298
CC	0.028	0.028	0.022	-0.006	0.031	0.691	0.075
Dementia	0.101	0.065	0.093	0.131	0.032	0.000	0.103
Paraplegia	0.019	0.011	0.015	0.063	0.031	0.000	0.114
Past MI	0.058	0.054	0.051	0.015	0.031	0.265	0.083
PPF	0.023	0.020	0.021	0.023	0.015	0.069	0.429
Stroke	0.068	0.058	0.063	0.041	0.019	0.001	0.312

NOTE: The table compares new surgeons to experienced surgeons, before and after matching, in term of covariate means, standardized differences in means as a fraction of the standard deviation before matching, and two-sample P-values. New = new surgeon, Ex-B = experienced surgeon, before matching, Ex-A = experienced surgeon, after matching. Standardized differences above 1/10th of a standard deviation are in bold.

Large, Sparse Optimal Matching With Refined Covariate Balance in an Observational Study of the Health Outcomes Produced by New Surgeons, Pimental et al. (2015)



*Using Mixed Integer Programming for
Matching in an Observational Study of
Kidney Failure after Surgery,
Zubizarreta (2012)*

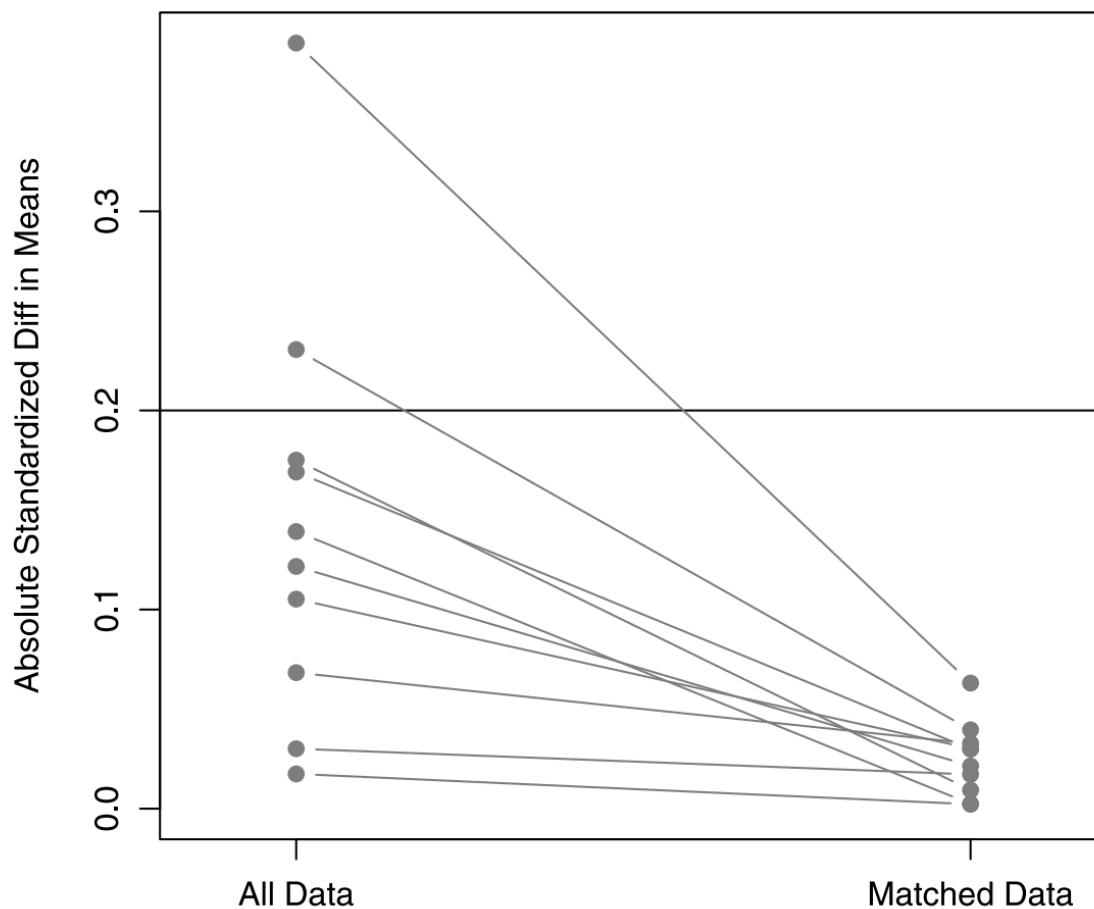


FIG. 2. *Plot of standardized difference of means of 10 covariates before and after matching. Data from Stuart and Green (2008).*

Problems with balance

- Often defined per covariate, at most for few covariate pairs
- What happens if 100's of covariates? What is "acceptable balance"? Rule of thumb is often 0.1 sd, but completely heuristic and should be viewed with some suspicion
- In general we want $p(X|T = 1) \approx p(X|T = 0)$
- Harder to use over high dimensional data
 - Maximum Mean Discrepancy distance
 - Wasserstein distance

Covariate adjustment and matching

- Matching is equivalent to covariate adjustment with two 1-NN 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)



Introduction to Causal Inference

Dr. Uri Shalit

Course number 097400
2020-2021

Lesson 5

Reminder

- Y_0, Y_1 : potential outcomes
- T : binary treatment
- X : observed covariates

“The Assumptions”

Sufficient conditions for causal inference to be possible:

- 1. Stable Unit Treatment Value Assumption**
 - no interference
- 2. Consistency**
 - we see the correct potential outcome
- 3. Ignorability / No unmeasured confounders**
- 4. Common support**

The adjustment formula

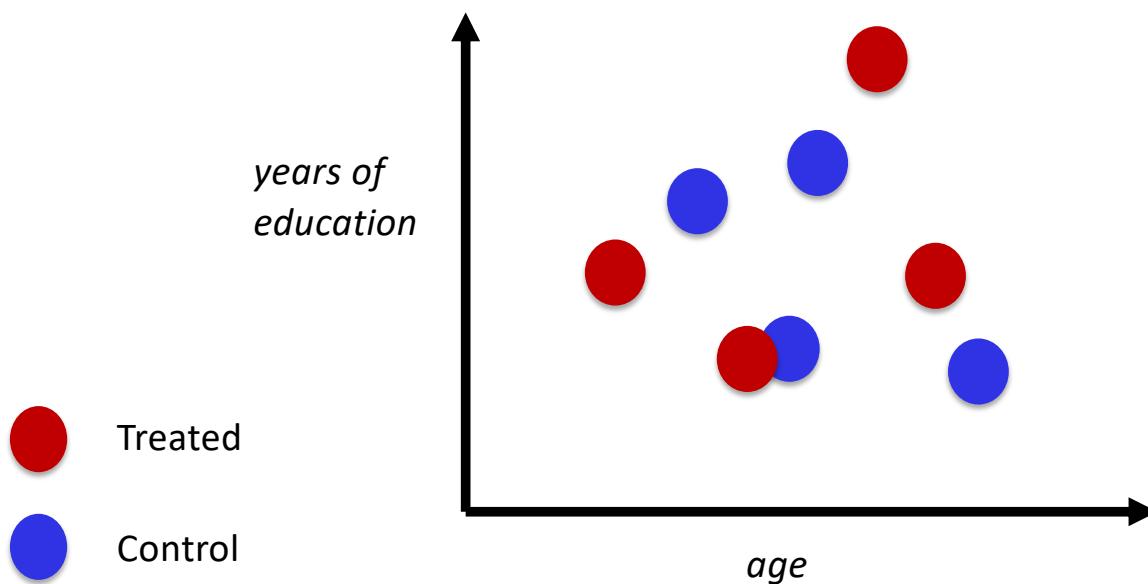
Under the assumptions above we have that:

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

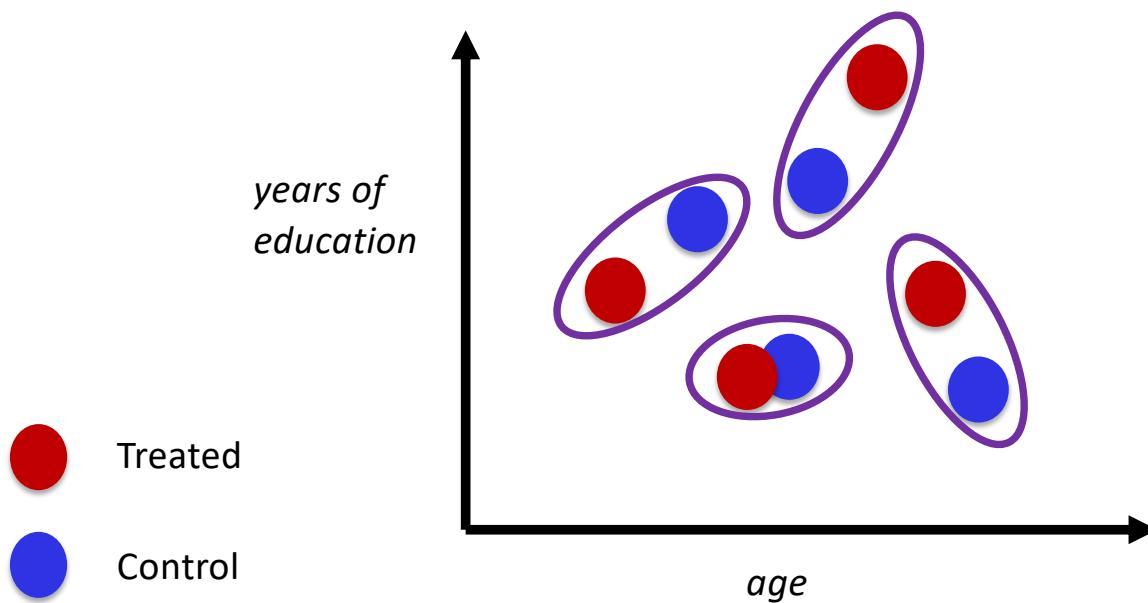
Estimate with:

1. Covariate adjustment
2. Inverse propensity score weighting

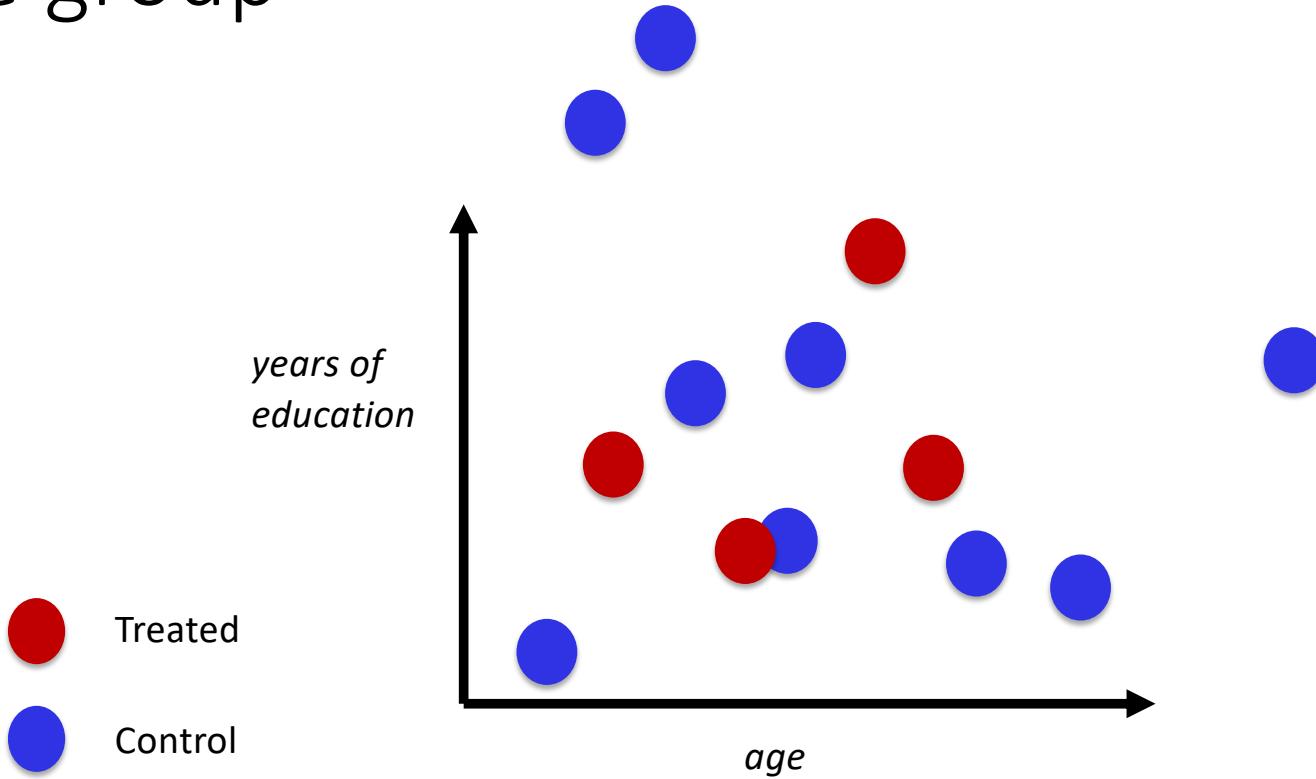
Match to nearest neighbor from opposite group



Match to nearest neighbor from opposite group



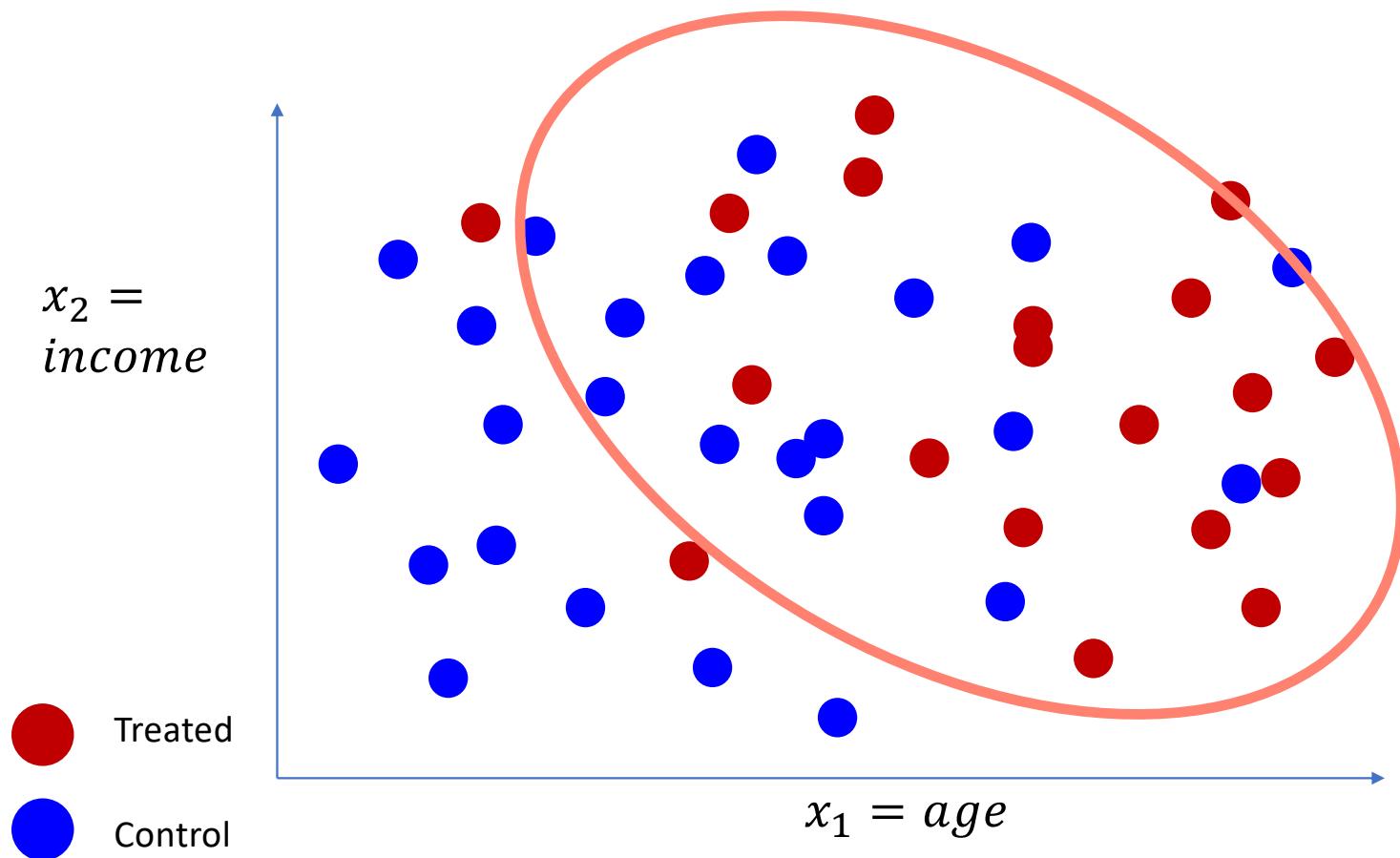
Match to nearest neighbor from
opposite group



Average treatment effect on the treated

- Often observational studies are interested in the effect of the treatment on the treated population
- $ATE = \mathbb{E}[Y_1 - Y_0]$
- $ATT = \mathbb{E}[Y_1 - Y_0 | T = 1]$

ATT



General Matching

- For each i , define $J(i) = \{j \mid j \text{ is close to } i, t_j \neq t_i\}$
 $J(i)$ are a set close counterfactual neighbor of i
- $t_i = 1$, unit i is treated:
- $\widehat{ITE}(i) = y_i - \frac{1}{|J(i)|} \sum_{j \in J(i)} y_j$
- $t_i = 0$, unit i is control:
- $\widehat{ITE}(i) = \frac{1}{|J(i)|} \sum_{j \in J(i)} y_j - y_i$
- $\widehat{ATE} = \frac{1}{n} \sum_{i=1}^n \widehat{ITE}(i)$

Advantages of matching

- Design without looking at outcomes, mimics RCTs
 - Can iterate: e.g. change metrics, without p-hacking/overfitting
- Interpretable

Common metrics

- $x = (x_1, x_2, \dots, x_d), x' = (x'_1, x'_2, \dots, x'_d)$

- Euclidean

$$d(x, x') = \sqrt{\sum_i (x_i - x'_i)^2} = \sqrt{x^\top x'}$$

- L1

$$d(x, x') = \sum_i |x_i - x'_i|$$

- Mahalanobis

$$d(x, x') = \sqrt{x^\top \Sigma^{-1} x'} \text{ where } \Sigma \text{ is the data covariance matrix}$$

- Cosine distance (not a proper metric, no triangle inequality)

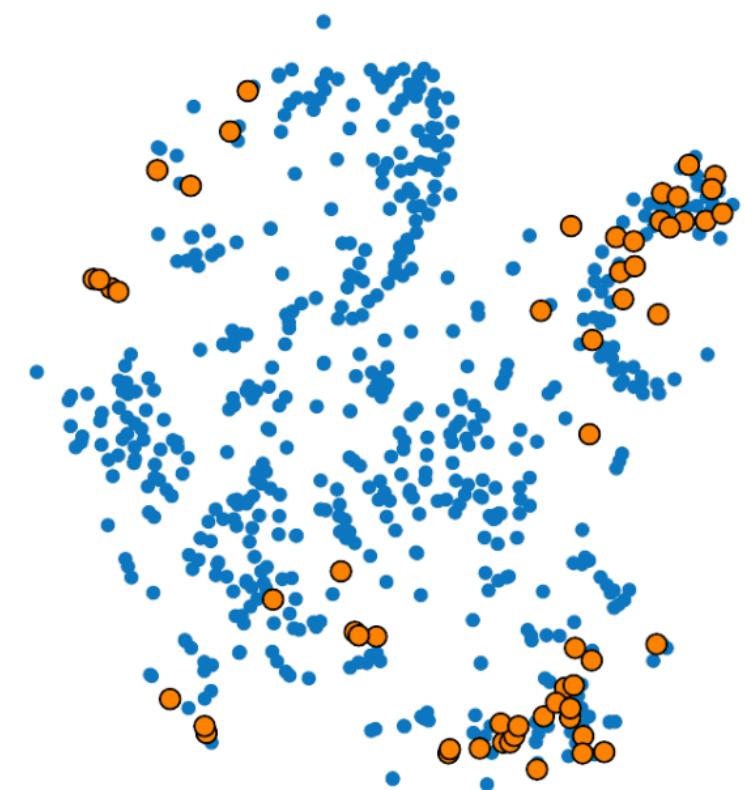
$$d(x, x') = 1 - \frac{x^\top x'}{\|x\| \cdot \|x'\|}$$

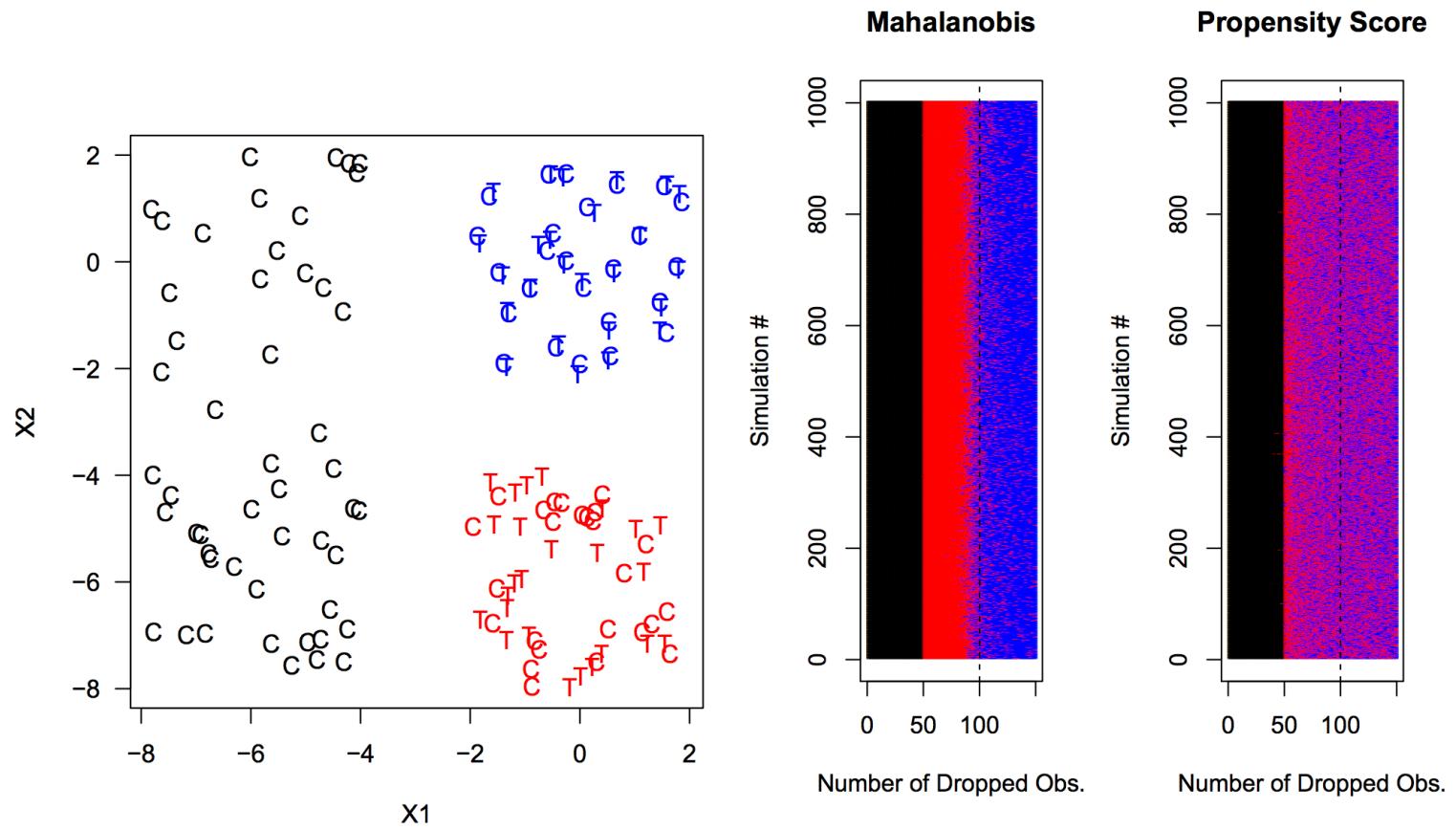
Metrics

- Make sure all covariates are on the same scale
- What to do with many weakly relevant features vs. few highly relevant?

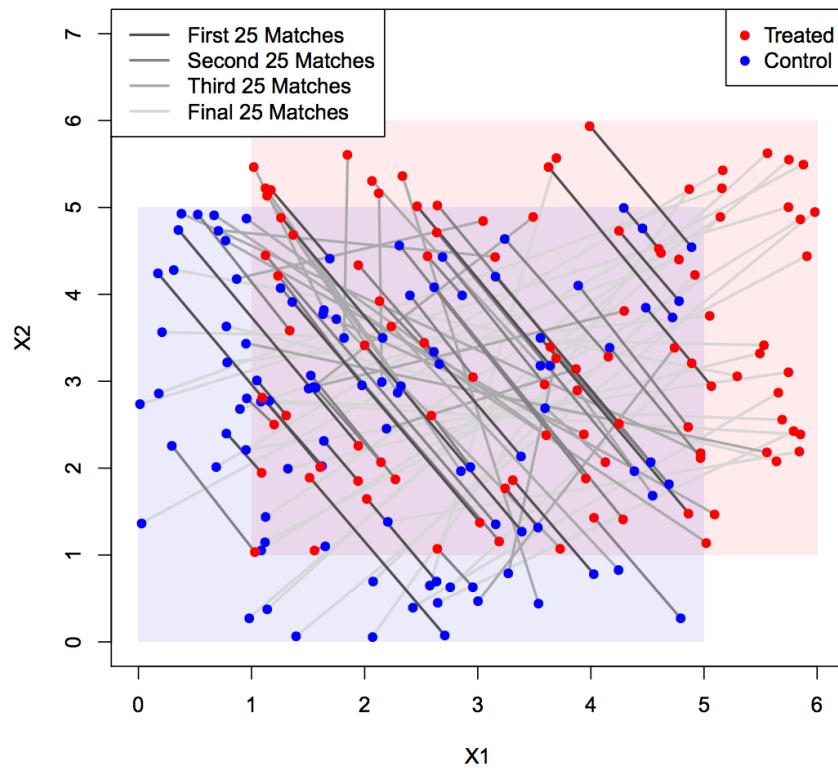
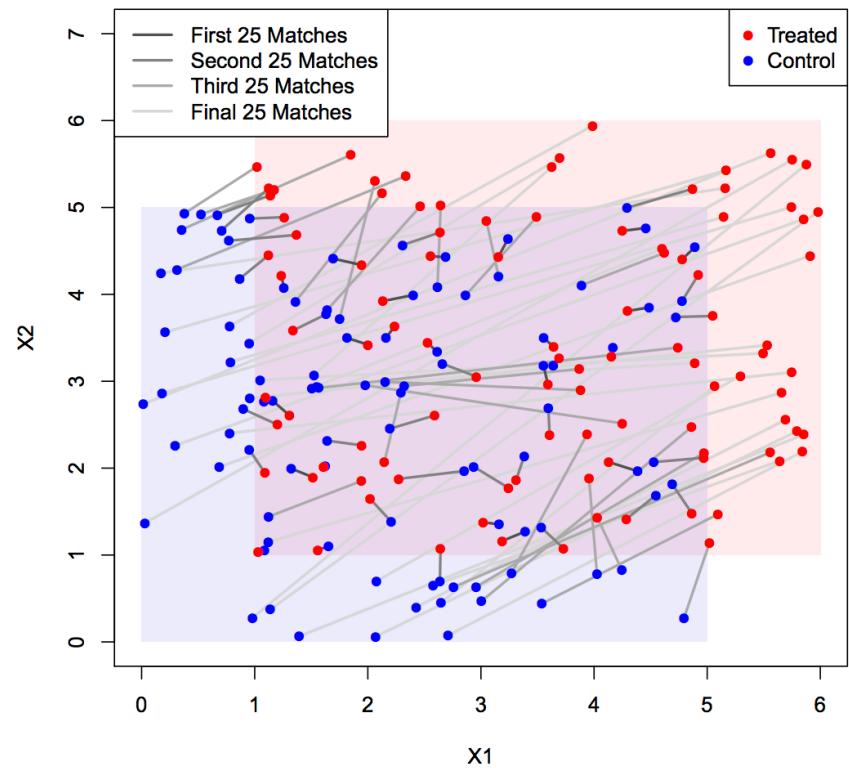
Matching methods

- k-Nearest Neighbor matching
Each treated is matched with exactly k controls, with or without replacement.
- The effect of k: bias-variance tradeoff
- Some samples might be unused, but not always a big issue



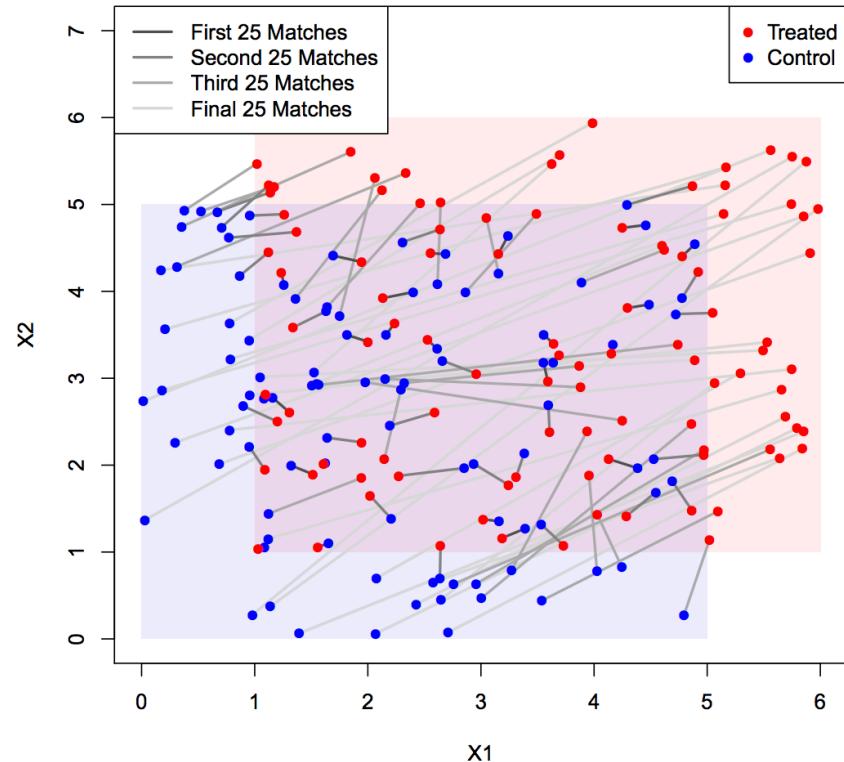


“Why Propensity Scores Should Not Be Used for Matching?”
 King & Nielsen 2016

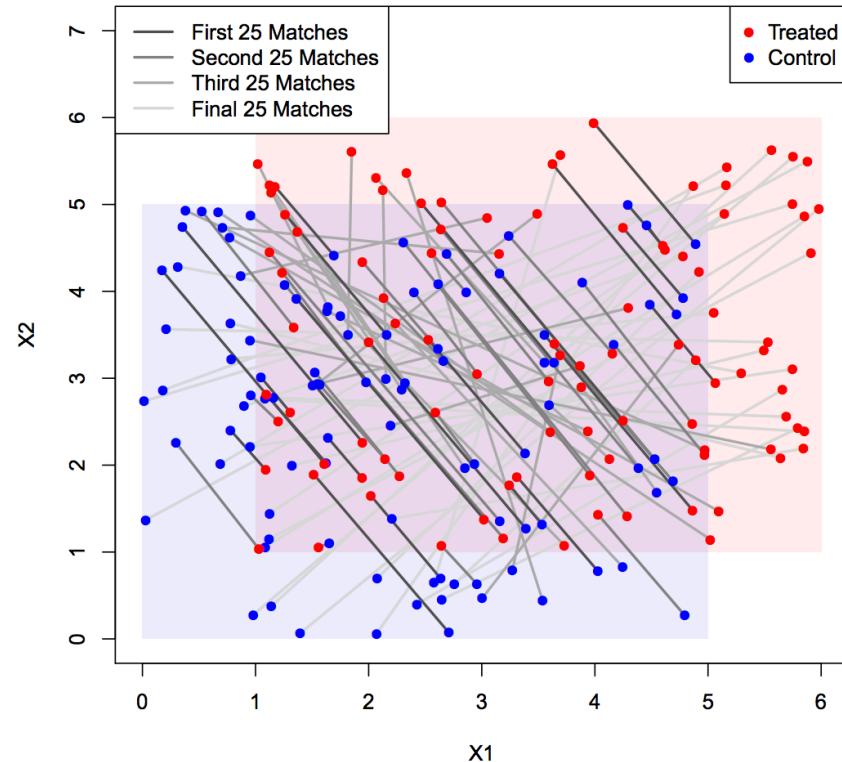


“Why Propensity Scores Should Not Be Used for Matching?”
King & Nielsen 2016

Full covariate matching with
Mahalanobis distance



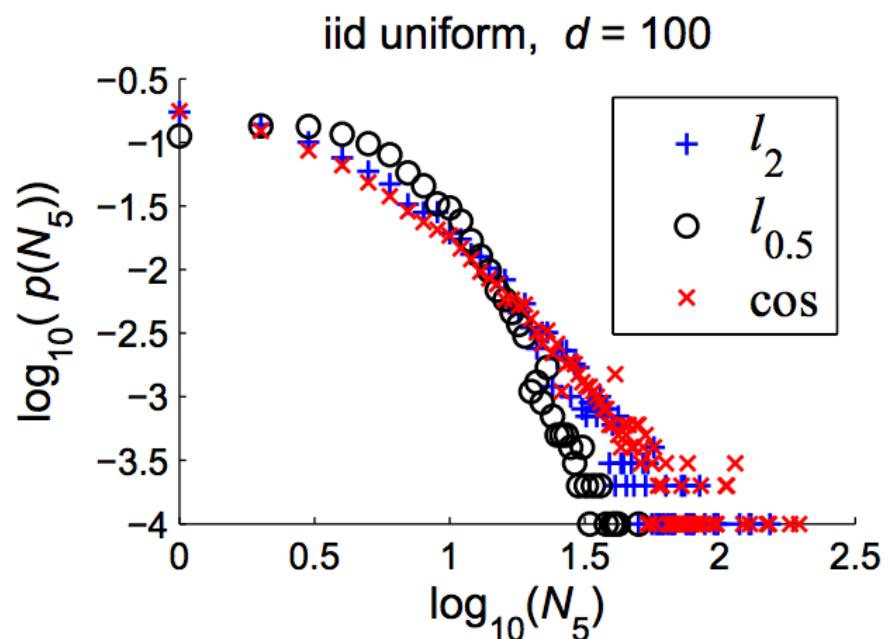
Propensity score matching



“Why Propensity Scores Should Not Be Used for Matching?”
King & Nielsen 2016

Curse of dimensionality for nearest neighbors

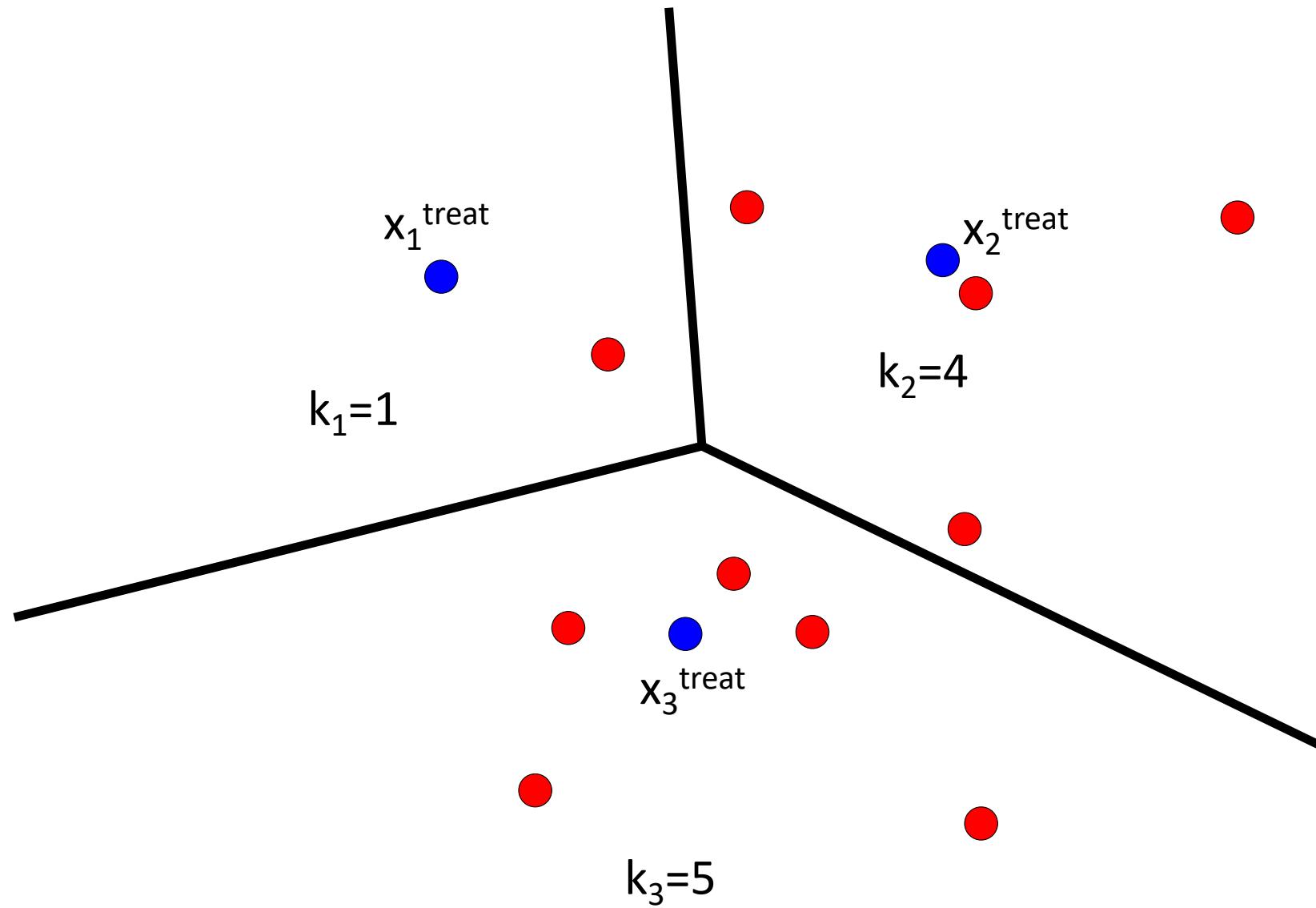
- A small number of points are neighbors of everyone
- In general distances are less informative in high-dimensions



Hubs in Space: Popular Nearest
Neighbors in High-Dimensional Data,
Radovanović et al.,
JMLR 2010

Matching methods

- Optimal matching: minimize sum of distances between all pairs
- Related to optimal transport, solvable by linear programming



How to choose a metric and matching?

- Domain knowledge, e.g. expert opinion on matches
- Relative size of control and treated groups
- Checking for “balance”

What is balance?

- Means the treated and control have the same marginal distributions for most features
- Standardized differences

*Using Mixed Integer Programming for Matching in an Observational Study of Kidney Failure after Surgery,
Zubizarreta (2012)*

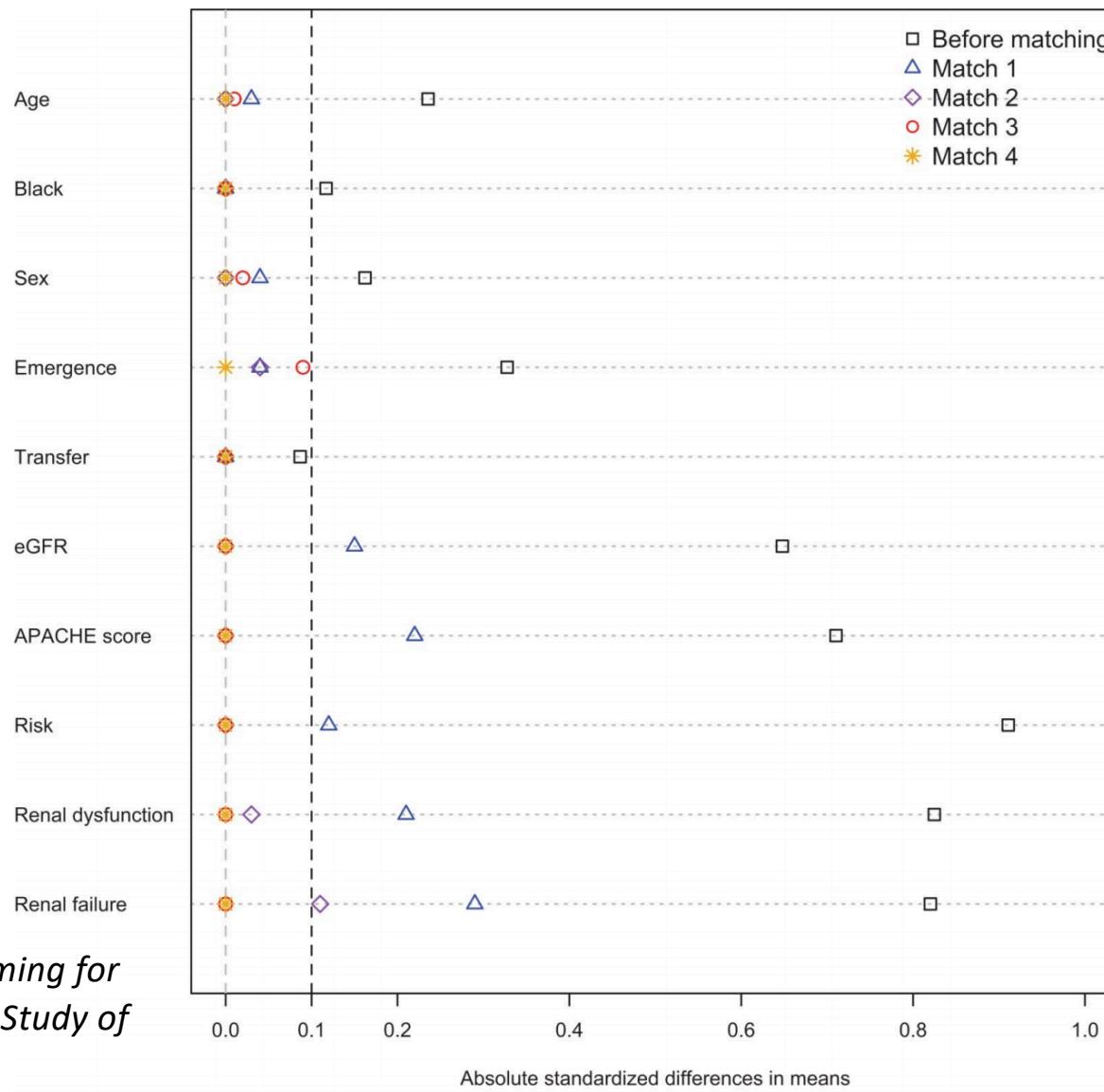
Covariate	Before matching			Match 1		
	ω_i	$\bar{x}_{T,i}$	$\bar{x}_{C,i}$	ω_i	$\bar{x}_{T,i}$	$\bar{x}_{C,i}$
Age	-	74.19	73.22	0	74.19	74.29
Black	-	0.12	0.08	0	0.12	0.12
Sex	-	0.39	0.47	0	0.39	0.37
Emergency	-	0.32	0.18	0	0.32	0.31
Transfer	-	0.02	0.01	0	0.02	0.02
eGFR	-	54.21	73.60	0	54.21	58.82
APACHE	-	35.73	28.69	0	35.73	33.56
Risk	-	-2.97	-3.82	0	-2.97	-3.08
Dysfunction	-	0.33	0.04	0	0.33	0.26
Failure	-	0.31	0.03	0	0.31	0.21

Table 2. Covariate imbalance before and after matching

Covariate	Covariate mean			Standardized difference		2-sample P-value	
	New	Ex-B	Ex-A	Before	After	Before	After
Sample size	6,260	123,846	6,260				
Age	77.883	76.992	77.926	0.116	-0.005	0.000	0.617
Male	0.345	0.358	0.346	-0.027	-0.003	0.038	0.880
ER-admit	0.538	0.323	0.537	0.444	0.003	0.000	0.886
Transfer	0.008	0.008	0.007	0.000	0.013	1.000	0.532
Risk	0.042	0.030	0.040	0.214	0.031	0.000	0.237
CHF	0.149	0.123	0.143	0.076	0.019	0.000	0.311
Liver	0.043	0.036	0.038	0.035	0.026	0.005	0.161
Cancer	0.164	0.175	0.164	-0.029	0.001	0.030	0.981
Past A	0.170	0.171	0.161	-0.002	0.024	0.880	0.178
Diabetes	0.189	0.197	0.199	-0.019	-0.024	0.145	0.198
Renal	0.069	0.058	0.064	0.046	0.020	0.000	0.282
COPD	0.167	0.147	0.160	0.055	0.019	0.000	0.298
CC	0.028	0.028	0.022	-0.006	0.031	0.691	0.075
Dementia	0.101	0.065	0.093	0.131	0.032	0.000	0.103
Paraplegia	0.019	0.011	0.015	0.063	0.031	0.000	0.114
Past MI	0.058	0.054	0.051	0.015	0.031	0.265	0.083
PPF	0.023	0.020	0.021	0.023	0.015	0.069	0.429
Stroke	0.068	0.058	0.063	0.041	0.019	0.001	0.312

NOTE: The table compares new surgeons to experienced surgeons, before and after matching, in term of covariate means, standardized differences in means as a fraction of the standard deviation before matching, and two-sample P-values. New = new surgeon, Ex-B = experienced surgeon, before matching, Ex-A = experienced surgeon, after matching. Standardized differences above 1/10th of a standard deviation are in bold.

Large, Sparse Optimal Matching With Refined Covariate Balance in an Observational Study of the Health Outcomes Produced by New Surgeons, Pimental et al. (2015)



*Using Mixed Integer Programming for
Matching in an Observational Study of
Kidney Failure after Surgery,
Zubizarreta (2012)*

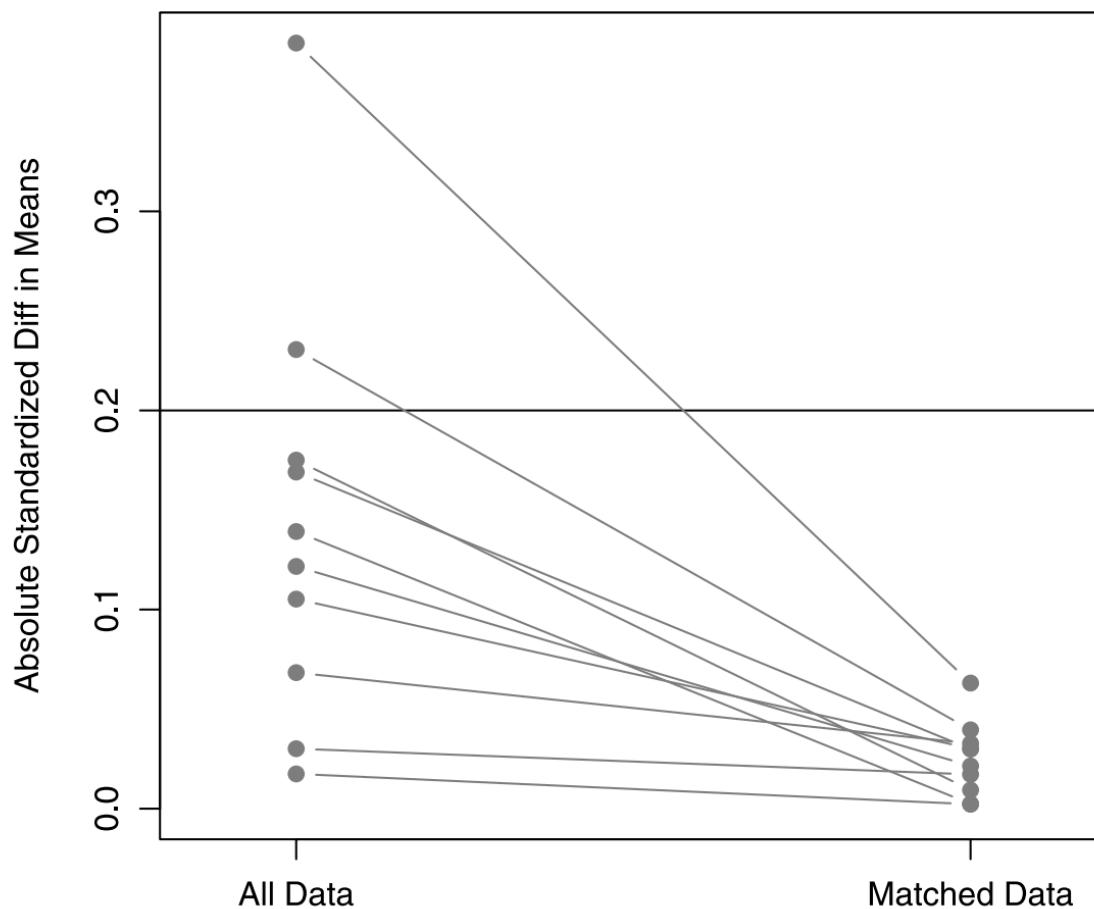


FIG. 2. *Plot of standardized difference of means of 10 covariates before and after matching. Data from Stuart and Green (2008).*

Problems with balance

- Often defined per covariate, at most for few covariate pairs
- What happens if 100's of covariates? What is "acceptable balance"? Rule of thumb is often 0.1 sd, but completely heuristic and should be viewed with some suspicion
- In general we want $p(X|T = 1) \approx p(X|T = 0)$
- Harder to use over high dimensional data
 - Maximum Mean Discrepancy distance
 - Wasserstein distance

Covariate adjustment and matching

- Matching is equivalent to covariate adjustment with two 1-NN 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)

Estimation methods so far

- Covariate adjustment
- Inverse propensity score weighting
- Matching

Causal inference – no test set

- A statistical challenge in causal inference from observational studies (even under ignorability):
You don't know if you have the right answer!
- What if you used the wrong model?
- No ultimate answer to this problem, but there are ways to improve your model

Doubly robust estimators

- Combine inverse propensity score weighting and covariate adjustment
- If at least *one* of the two models is correct, then there is no bias
- You don't need to know which of the models is correct
- Even if both models are wrong, bias roughly behaves like the one which is “less wrong”

Doubly robust estimators

- Excellent reference: Section 5 of <http://www4.stat.ncsu.edu/~davidian/double.pdf> (Marie Davidian, UNC)
- Notation
 - $\hat{e}(x)$: model for $p(t = 1|x)$
 - $\hat{m}_1(x)$: model for $\mathbb{E}[Y|X, T = 1]$
 - $\hat{m}_0(x)$: model for $\mathbb{E}[Y|X, T = 0]$
- Dataset: $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$
- Doubly robust score:

$$\hat{g}_1(x_i) = \hat{m}_1(x_i) + \frac{t_i}{\hat{e}(x_i)} (y_i - \hat{m}_1(x_i))$$

$$\hat{g}_0(x_i) = \hat{m}_0(x_i) + \frac{1 - t_i}{1 - \hat{e}(x_i)} (y_i - \hat{m}_0(x_i))$$

Doubly robust estimators

- Doubly robust score:

$$\hat{g}_1(x_i) = \hat{m}_1(x_i) + \frac{t_i}{\hat{e}(x_i)} (y_i - \hat{m}_1(x_i))$$

$$\hat{g}_0(x_i) = \hat{m}_0(x_i) + \frac{1 - t_i}{1 - \hat{e}(x_i)} (y_i - \hat{m}_0(x_i))$$

$$\widehat{ATE}_{DR} = \frac{1}{n} \sum_{i=1}^n g_1(x_i) - g_0(x_i)$$

- For units with $t_i = 1$, $g_0(x_i) = m_0(x_i)$
- For units with $t_i = 0$, $g_1(x_i) = m_1(x_i)$
- Can be shown that this estimator is consistent if *either* (\hat{m}_1, \hat{m}_0) or \hat{e} is consistent
- This property is called **double robustness**
- The above is not the only double-robust estimator, there are many

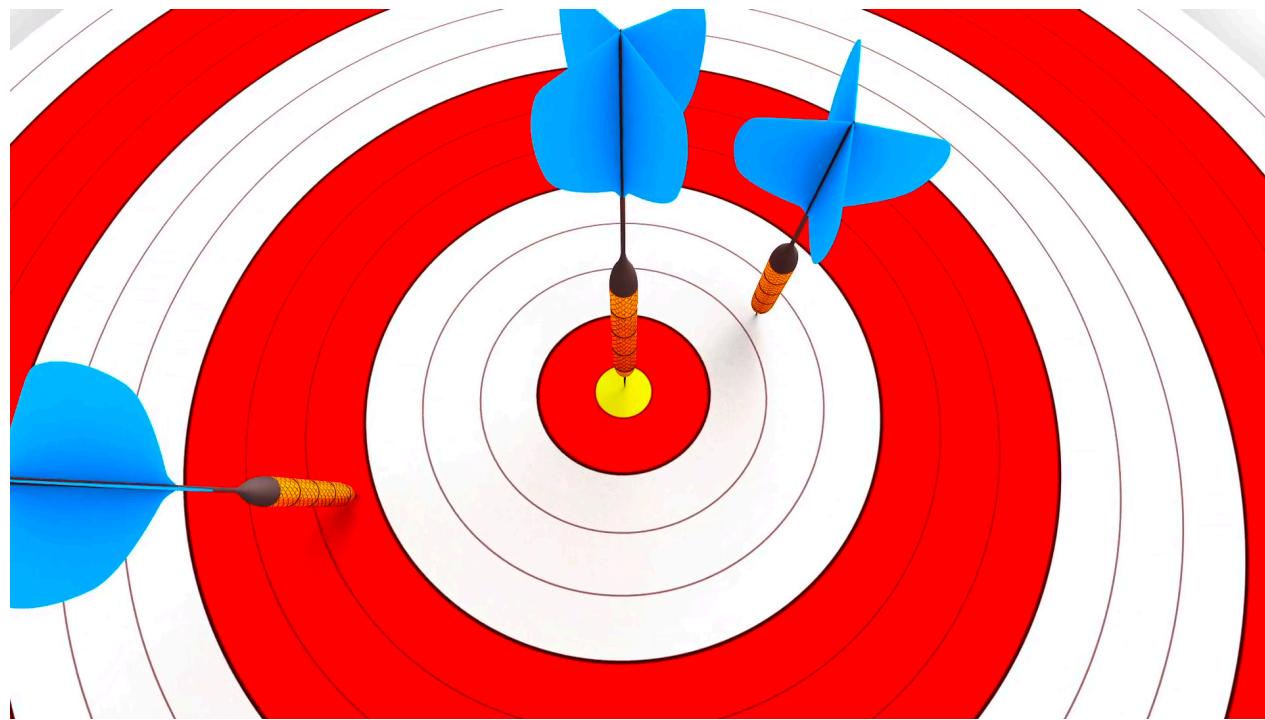
Doubly robust estimators

- Theoretically: great
 - Even if both models are mis-specified, if at least one is only slightly mis-specified the bias will be small
- In practice: not always optimal
 - Kang, Joseph DY, and Joseph L. Schafer. "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data." *Statistical science* (2007): 523-539.APA
 - Show that in practice may perform worse, on simulation studies
- 2016 Atlantic Causal Inference Challenge, a doubly-robust method called Targeted Maximum Likelihood Estimation (Susan Gruber & Mark van der Laan) came in second place
- Area of active research

Standard errors for \widehat{ATE}

- Not the focus of this course, but important nonetheless and the focus of much work
- Crucial when reporting a finding!
- ATE is an average, so many standard tools apply
- Most common approaches:
 - Standard large sample theory, using e.g. a sandwich estimator
 - Bootstrap
- Often bootstrap is easiest to implement

Target Trial



Following slides courtesy of Miguel Hernán, Harvard

We analyze observational data

because we cannot conduct a randomized trial

Observational analyses are **not** our preferred choice

- For each observational analysis for causal inference, we can imagine a hypothetical randomized trial that we would prefer to conduct
 - If only it were possible

The Target Trial

- The (hypothetical) randomized trial that we would like to conduct to answer a causal question
 - To learn what works and what harms

The Target Trial

- The (hypothetical) randomized trial that we would like to conduct to answer a causal question
 - To learn what works and what harms

- A causal analysis of observational data can be viewed as an attempt to emulate some target trial

Example of how a clinical trial is designed

- Intervention/treatment for each trial arm
- Outcome measures
 - Including start and end of follow-up
- Eligibility criteria
 - Inclusion
 - Exclusion
- Analysis plan
- <https://clinicaltrials.gov/ct2/show/NCT04470427>

Step 1

Specify Target Trial protocol

- Eligibility criteria
- Treatment strategies
- Randomized assignment
- Start/End follow-up
- Outcomes
- Causal contrast
- Analysis plan



Step 2

Emulate Target Trial protocol

- Eligibility criteria
- Treatment strategies
- Randomized assignment
- Start/End follow-up
- Outcomes
- Causal contrast
- Analysis plan



Ok, so why is this a big deal?

- Why do we need to explicitly need to emulate a target trial when using observational data to learn what works?

Ok, so why is this a big deal?

- Why do we need to explicitly need to emulate a target trial when using observational data to learn what works?
- What happens if we just analyze the data as usual?
 - That is, if we compare “exposed” vs. “unexposed” and adjust for covariates?

Ok, so why is this a big deal?

- Why do we need to explicitly need to emulate a target trial when using observational data to learn what works?
- What happens if we just analyze the data as usual?
 - That is, if we compare “exposed” vs. “unexposed” and adjust for covariates?
- Let’s see an example

EXAMPLE #1

Postmenopausal hormone therapy and heart disease

- Observational epidemiologic studies
 - >30% **lower risk** in current users vs. never users
 - e.g., hazard ratio: 0.68 in Nurses' Health Study
 - Grodstein et al. *J Women's Health* 2006
- Randomized trial
 - >20% **higher risk** in initiators vs. noninitiators
 - hazard ratio: 1.24 in Women's Health Initiative
 - Manson et al. *New England J Med* 2003

EXAMPLE #1

Postmenopausal hormone therapy and heart disease

- Observational epidemiologic studies
 - >30% **lower risk** in current users vs. never users
 - e.g., hazard ratio: 0.68 in Nurses' Health Study
 - Grodstein et al. *J Women's Health* 2006
- Randomized trial
 - >20% **higher risk** in initiators vs. noninitiators
 - hazard ratio: 1.24 in Women's Health Initiative
 - Manson et al. *New England J Med* 2003

Shocking discrepancy!

The randomized trial Women's Health Initiative (WHI)

- Double-blind
 - Placebo-controlled
 - Large
 - >16,000 U.S. women aged 50-79 yrs
 - Randomly assigned to
 - estrogen plus progestin therapy
 - placebo
 - Women followed approximately every year
 - for a maximum of 8 years
-

Effect estimates from the randomized trial

Intention-to-treat hazard ratio (95% CI) of coronary heart disease

Overall 1.23 (0.99, 1.53)

Years of
follow-up

■ 0-2 1.51 (1.06, 2.14)

■ >2-5 1.31 (0.93, 1.83)

■ >5 0.67 (0.41, 1.09)

Years since
menopause

■ <10 0.89 (0.54, 1.44)

■ 10-20 1.24 (0.86, 1.80)

■ >20 1.65 (1.14, 2.40)

Women who survived 5 years! If HRT is truly dangerous, in this sub-group the women who took HRT will be the especially healthy / hardy ones

A type of selection bias

Effect estimates from the randomized trial

Intention-to-treat hazard ratio (95% CI) of coronary heart disease

<input type="checkbox"/> Overall	1.23 (0.99, 1.53)
<input type="checkbox"/> Years of follow-up	
■ 0-2	1.51 (1.06, 2.14)
■ >2-5	1.31 (0.93, 1.83)
■ >5	0.67 (0.41, 1.09)
<input type="checkbox"/> Years since menopause	
■ <10	0.89 (0.54, 1.44)
■ 10-20	1.24 (0.86, 1.80)
■ >20	1.65 (1.14, 2.40)

This hazard ratio can be fully explained by selection bias even if no woman benefits from hormone therapy
(Stensrud et al. *Epidemiology* 2017)

Why did observational studies get it “wrong”?

□ Popular theory

- Insufficient adjustment for lifestyle and socioeconomic indicators (residual confounding)
- Corollary: causal inference from observational data is a hopeless undertaking

□ An alternative theory

- The observational studies were not emulating a target trial

The randomized trial compared women who **initiated** therapy with women who did not

Design

- Women randomly assigned to initiation of hormone therapy or placebo
- Almost all women assigned to initiation received at least a dose, that is, they are classified as initiators

Analysis

- Compared risk between initiators (**incident** users) and noninitiators of hormone therapy

This trial informs decisions about therapy initiation

Observational studies compared women **currently using** therapy with women who did not use it

Design

- Women were asked about therapy use
- They were classified as current, past, or never users

Analysis

- Compared risk between current (**prevalent**) users and never users of hormone therapy
 - Was the estimate different from that of the WHI trial?

Observational studies compared women **currently using** therapy with women who did not use it

- Design
 - Women were asked about therapy use
 - They were classified as current, past, or never users
- Analysis
 - Compared risk between current (**prevalent**) users and never users of hormone therapy
 - Was the estimate different from that of the WHI trial?
- What decision does this design/analysis inform?
 - What is the target trial?

What if we re-analyze the observational data...

... to explicitly emulate a target trial as close as possible to the WHI randomized trial?

Causal inference algorithm

- Step 1: Specify the protocol of a target trial of hormone therapy and coronary heart disease
- Step 2: Emulate it
 - Hernán et al. *Biometrics* 2005; 61(4):922–930
 - Hernán et al. *Epidemiology* 2008; 19(6):766-779

Abbreviated Target Trial Protocol: Hormone therapy and coronary heart disease

Eligibility criteria	Postmenopausal women with no history of cancer and other diseases, and no use of hormone therapy in the last 2 years.
Treatment strategies	<ol style="list-style-type: none">1. Initiate estrogen plus progestin hormone therapy at baseline and remain on it during the follow-up, unless deep vein thrombosis, pulmonary embolism, myocardial infarction, or cancer are diagnosed2. Refrain from taking hormone therapy during the follow-up
Assignment procedures	Participants will be randomly assigned to either strategy at baseline, and will be aware of the strategy they have been assigned to.
Follow-up period	Starts at randomization and ends at coronary heart disease diagnosis, death, loss to follow-up, or June 2000, whichever occurs earlier.
Outcome	Coronary heart disease diagnosed by a cardiologist
Causal contrasts	Intention-to-treat effect, per-protocol effect
Analysis plan	Intention-to-treat analysis, non-naïve per-protocol analysis

Important

Target trial must be a pragmatic trial

- Observational data cannot be used to emulate
 - a placebo-controlled trial
 - at most a trial with a “usual care” group
 - a trial with blind design
 - individuals are generally aware of the treatment they receive
 - treatment strategies that do not exist in the real world
 - enforcement of adherence to the protocol
 - tight monitoring that doesn’t happen in the real world

Observational data for emulation: The Nurses' Health Study

- Epidemiologic follow-up (cohort) study
- ~80,000 women with full data in 1980
- Information updated by questionnaire every two years
 - Use of hormone therapy
 - Diagnosis of coronary heart disease (confirmed by physician)
 - Medical diagnoses
 - Lifestyle data: diet, exercise, smoking...
 - Other risk factors for coronary heart disease

Emulation

- Eligibility criteria

- Analysis restricted to women who met the eligibility criteria of the target trial

- Treatment strategies

- 1) Initiation and continued use of oral estrogens plus progesterone therapy during the follow-up
 - 2) No hormone therapy use during the follow-up

- Outcome

- a diagnosis of coronary heart disease during the follow-up

Emulation: Randomized assignment

- This is what “adjustment for confounding” means
- If insufficient data on confounders, then emulation of random assignment fails
 - Confounding bias
- Need to adjust for baseline covariates
 - via matching, stratification or regression, standardization or inverse probability (IP) weighting, g-estimation...

Emulation: Causal contrast

- Intention-to-treat effect
 - The effect of **assignment** to therapy vs. no therapy
 - regardless of actual use

- Since the dataset doesn't include prescription dates, we estimate the effect of **initiation** of therapy
 - Analogous to a modified intention-to-treat approach in a trial
 - Including only those who take at least one dose of treatment

Emulation: “Intention-to-treat” analysis

- Compare risk between initiators and noninitiators of therapy at baseline
 - regardless of use during the follow-up
- Fit a Cox model with an indicator for treatment initiation + confounders
 - Age, past hormone use, parental history of myocardial infarction before age 60, education, husband's education, ethnicity, age at menopause, calendar time, high cholesterol, high blood pressure, diabetes, angina, stroke, coronary revascularization, osteoporosis, body mass index, smoking, aspirin use, alcohol intake, physical activity, diet score, multivitamin use, fruit/vegetable intake

Emulation summary

- We used the observational data to emulate a target trial with similar eligibility criteria, treatment strategies, outcome, causal contrast, and analysis plan as the randomized trial

 - Some differences
 - Not blinded
 - Not placebo-controlled
 - Shorter average time since menopause than WHI
 - Longer follow-up than WHI
-

Effect estimates: hazard ratios (95% CIs)

	Randomized Women's Health Initiative	Observational Nurses' Health Study
□ Overall	1.23 (0.99, 1.53)	1.05 (0.82, 1.34)
□ Years of follow-up		
■ 0-2	1.51 (1.06, 2.14)	1.43 (0.92, 2.23)
■ >2	1.07 (0.81, 1.41)	0.91 (0.72, 1.16)
□ Years since menopause		
■ <10	0.89 (0.54, 1.44)	0.88 (0.63, 1.21)
■ 10-20	1.24 (0.86, 1.80)	1.13 (0.85, 1.49)
■ >20	1.65 (1.14, 2.40)	--

When the target trial is explicitly emulated,
then a similar **causal question** is asked

- No shocking observational-randomized discrepancies
 - though wide confidence intervals in both studies

 - What about the popular hypothesis? Any residual confounding?
 - Probably, but insufficient to explain the original discrepancy
-

Interesting state of affairs

- The usual criticism of observational analyses is lack of randomization
 - Failure to emulate randomization because of insufficient data on confounders (residual confounding)
 - Hard to fix
 - Yet mounting evidence suggests another problem
 - Failure to choose a correct time zero
 - Easy to fix
-

Step 1

Specify Target Trial protocol

- Eligibility criteria

Handling time zero correctly:
The low-hanging fruit for
causal inference

- Outcomes

- Causal contrast

- Analysis plan

Step 2

Emulate Target Trial protocol

- Eligibility criteria

- Treatment strategies

- Randomized assignment

- Start/End follow-up

- Outcomes

- Causal contrast

- Analysis plan



Time zero of follow-up in the Target Trial

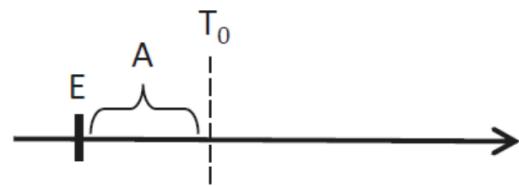
- For each person, the time when 3 things happen
 - eligibility criteria are met
 - treatment strategies are assigned
 - study outcomes begin to be counted
- The same applies to observational analyses

- Misalignment of eligibility criteria and treatment assignment leads to selection bias / immortal time bias
 - Hernán et al. *J Clin Epidemiol* 2016; 79:70-75.

Misalignment of eligibility (E) and treatment assignment (A) prevents correct emulation

Type of
emulation failure

1. T_0 after E and A



Selection of...

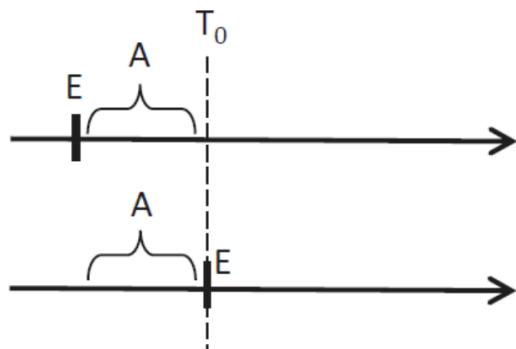
eligible individuals who initiate a treatment strategy and remain under follow-up through reset T_0

Immortal
time

No

Misalignment of eligibility (E) and treatment assignment (A) prevents correct emulation

Type of emulation failure	Selection of...	Immortal time
1. T_0 after E and A	eligible individuals who initiate a treatment strategy and remain under follow-up through reset T_0	No
2. T_0 at E but after A	individuals who initiated a treatment strategy before, and remained under follow-up until, eligibility (specified at T_0)	No



Misalignment of eligibility (E) and treatment assignment (A) prevents correct emulation

Type of emulation failure	Selection of...	Immortal time
1. T_0 after E and A	eligible individuals who initiate a treatment strategy and remain under follow-up through reset T_0	No
2. T_0 at E but after A	individuals who initiated a treatment strategy before, and remained under follow-up until, eligibility (specified at T_0)	No
3. T_0 before E and A	individuals who initiated a treatment strategy before, and remained under follow-up until, eligibility (specified at T_0)	Yes

Misalignment of eligibility (E) and treatment assignment (A) prevents correct emulation

Type of emulation failure	Selection of...	Immortal time
1. T_0 after E and A	eligible individuals who initiate a treatment strategy and remain under follow-up through reset T_0	No
2. T_0 at E but after A	individuals who initiated a treatment strategy before, and remained under follow-up until, eligibility (specified at T_0)	No
3. T_0 before E and A	individuals who initiated a treatment strategy before, and remained under follow-up until, eligibility (specified after T_0)	Yes
4. T_0 at E but before A	eligible individuals at T_0 who remained under follow-up until completing a treatment strategy	Yes

Hernán et al.
J Clin Epidemiol
2016; 79:70-75

Why is it hard to align eligibility and treatment assignment at time zero?

- Time of eligibility may not be unique
 - An individual may meet the eligibility criteria at multiple times

- Treatment group may not be known at time zero
 - An individual's treatment strategy/exposure plan will be revealed after time zero

Basic principle of trial design

(and of observational analyses that emulate a target trial)

- Treatment assignment and the determination of eligibility occur simultaneously at time zero

- Observational analyses that violated this principle yielded implausible estimates
 - Good news: correct time zero determination is always possible

2 key components of the emulation of the target trial

1. Randomization

- Emulation requires adjustment for confounding

2. Specification of time zero

- Emulation requires that time zero is synchronized with determination of eligibility and assignment of treatment strategies

- Lack of randomization is usually blamed for the failings of observational analyses, but...
- We have seen that incorrect specification of time zero is often the actual culprit

Emulation of a target trial is what we do when we cannot conduct the trial

- Reasonable people will always prefer a randomized trial, but often there is no alternative to observational studies
 - we better keep improving them
 - because people will keep using observational data to guide their decisions
- And we have identified some simple ways of improving observational analyses to learn what works and what harms

Target trial

- If you had no ethical or financial considerations, what is the RCT you would have liked to conduct?
- Now think whether you can approximate it using an observational study
- “No causation without manipulation”
- What is the causal effect of BMI on death?
- Is that true?
In most practical cases, I believe it's a good guideline



Miguel Hernán
 @_MiguelHernan

Following

Pearl believes that any causal effect we can name must also exist.

To him, the meaning of “the causal effect of A on death” is self-evident. He says we can quantify, say, the causal effect of race or the causal effect of obesity.

I don't think we can.



Análise Real @_analisereal · May 18

Replying to @_MiguelHernan

Can you quote the relevant passage you disagree with? Also, it seems you are conflating two things, quantification with definition. Quantifying the effect of race is surely hard, but do you think it's not possible to *define* it?

1 2 3 4 5 6



Miguel Hernán @_MiguelHernan · Dec 14

Your hit the nail in the head.

Are we doing science if we define quantities that we cannot quantify? Is it even possible to define a quantity that cannot be quantified, not even in principle?
ncbi.nlm.nih.gov/pubmed/27641316



Judea Pearl @yudapearl · Jul 30

For those who are hooked on $Y(x)$ go ahead and replace $P(Y|do(x))$ with $P(Y(x))$, no rule will fail. Just recall: $P(Y|do(x), do(z), W=w)$ goes into $P(Y(x,z)|W(x,z)=w)$. Also note: the #bookofwhy does not go into the do-calculus, except conceptually.
#epibookclub

Ellie Murray @EpiEllie

Replying to @yudapearl

Okay, thanks, that helps.

My last question: what are the practical implications for what can be solved —for example, do any of the 3 rules of do-calculus fail if you replace

2 3 21 5 6



Miguel Hernán @_MiguelHernan · Jul 30

Ironically...

You prefer the “do” notation but aren’t bothered by ill-defined interventions—not even in principle “doable”—and I prefer the counterfactual notation but insist on sufficiently well-defined interventions to interpret causal effect estimates

Miguel Hernán @_MiguelHernan

Pearl believes that any causal effect we can name must also exist.

To him, the meaning of “the causal effect of A on death” is self-evident. He says we can quantify, say, the causal effect of race or the causal effect of obesity.
Show this thread

5 2 9 5 6



Introduction to Causal Inference

Dr. Uri Shalit

Course number 097400
2020-2021

Lesson 6

Today

- Example study focusing on the statistical analysis methods (not the target trial aspect)
- Systematic comparisons of observational and randomized trials

Example study



American Journal of Epidemiology

© The Author 2008. Published by the Johns Hopkins Bloomberg School of Public Health.

All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org.

Vol. 168, No. 6

DOI: 10.1093/aje/kwn184

Advance Access publication August 6, 2008

Original Contribution

Adolescent Cannabis Problems and Young Adult Depression: Male-Female Stratified Propensity Score Analyses

Valerie S. Harder¹, Elizabeth A. Stuart¹, and James C. Anthony²

¹ Johns Hopkins Bloomberg School of Public Health, Baltimore, MD.

² Michigan State University College of Human Medicine, East Lansing, MI.

Received for publication December 28, 2007; accepted for publication May 22, 2008.

Does using cannabis as an adolescent cause depression as a young-adult?

- Why this question?

Concern about many adverse effects of early use of cannabis

Does using cannabis as an adolescent cause depression as a young-adult?

- Must define:
 - “using cannabis”
 - “as an adolescent”
- “depression”
- “as a young adult”

Does using cannabis as an adolescent cause depression as a young-adult?

Data

- Data: obtained from a long-term RCT testing a classroom interventions and their effect on future drug use and mental health
- Children followed from childhood until early-adulthood (25% attrition)
- RCT did not intervene on Cannabis use
- Data can still be used as observational study
- 1494 individuals (826 female, 668 male)

Does using cannabis as an adolescent cause depression as a young-adult?

Definition of treatment (exposure)

- Treated:
Occurrence of *cannabis problems* before age 17
- Control:
No cannabis use or use without problems by age 17
- Why 17? Prior literature
- Definition of problem with a standardized questionnaire
- Nitpick: self-reported

Does using cannabis as an adolescent cause depression as a young-adult?

Definition of outcome

- Depressive episode in the last year, if between the ages of 19 and 24 years
- Live interview, use DSM-IV definition of depression
- Why 19 to 24?
Most depression onset is after 19
Upper limit set to some degree by study population

Does using cannabis as an adolescent cause depression as a young-adult?

Covariates

- RCT study collected many datapoints about each individual
- Demographics, socio-economic, child and parent psychological questionnaire, concentration, shyness, parental supervision
- Psychological: before age 12
- Previous drug, alcohol and tobacco use
Taken from *before first cannabis use*, or before age 17 for non-users

Does using cannabis as an adolescent cause depression
as a young-adult?

Missing data

- Encode “missingness” as a category for each covariate type
- Merge small categories with no-overlap with nearby categories, e.g. merge race “other” with “white”

Does using cannabis as an adolescent cause depression as a young-adult?

Analysis

- Separate analysis for females and males
- Reasons: higher prevalence of cannabis problems in males, higher prevalence of depression in females
- This is an issue of heterogeneous effects

Does using cannabis as an adolescent cause depression as a young-adult? Analysis

TABLE 1. Baseline characteristics of 1,494 adolescent-onset cannabis problem users and comparison individuals from the original 2,311 individuals in the Prevention Research Center cohort, United States, 1985–2001

	Cannabis problem users		Comparison individuals		Chi square*	<i>p</i> value, two sided
	No.	%	No.	%		
Sex					63.54	<0.005
Male	151	70	517	40		
Female	66	30	760	60		
Race					19.05	<0.005
Black	132	61	948	74		
White	84	39	315	25		
Other†	1	0	14	1		
Family income					1.90	0.59
Low	19	9	115	9		
Middle	54	25	373	29		
High	71	33	381	30		
Missing	73	34	408	32		
Free lunch					1.68	0.43
No	62	29	314	25		
Yes	149	69	931	73		
Missing	6	3	32	3		
Daily tobacco smoker					336.08	<0.005
No	69	32	1,089	85		
Yes	146	67	163	13		
Missing	2	1	25	2		

TABLE 1. Continued

	Cannabis problem users		Comparison individuals		Chi square*	<i>p</i> value, two sided
	No.	%	No.	%		
Behavior problems‡						41.98 <0.005
Lower	32	15	397	31		
Low	69	32	377	30		
Moderate	30	14	171	13		
High	25	12	60	5		
Higher	10	5	21	2		
Missing	51	24	251	20		
Shyness‡						3.24 0.70
Lower	7	3	65	5		
Low	49	23	310	24		
Moderate	76	35	457	36		
High	31	14	171	13		
Higher	3	1	23	2		
Missing	51	24	251	20		
Depression symptoms§						4.40 0.22
Low	29	13	182	14		
Moderate	117	54	764	60		
High	14	6	64	5		
Missing	57	26	267	21		
Anxiety symptoms§						5.47 0.14
Low	45	21	235	18		
Moderate	94	43	657	51		
High	21	10	118	9		

Does using cannabis as an adolescent cause depression as a young-adult? Analysis

TABLE 1. Baseline characteristics of 1,494 adolescent-onset cannabis problem users and comparison individuals from the original 2,311 individuals in the Prevention Research Center cohort, United States, 1985–2001

	Cannabis problem users		Comparison individuals		Chi square*	<i>p</i> value, two sided
	No.	%	No.	%		
Sex					63.54	<0.005
Male	151	70	517	40		
Female	66	30	760	60		
Race					19.05	<0.005
Black	132	61	948	74		
White	84	39	315	25		
Other†	1	0	14	1		
Family income					1.90	0.59
Low	19	9	115	9		
Middle	54	25	373	29		
High	71	33	381	30		
Missing	73	34	408	32		
Free lunch					1.68	0.43
No	62	29	314	25		
Yes	149	69	931	73		
Missing	6	3	32	3		
Daily tobacco smoker					336.08	<0.005
No	69	32	1,089	85		
Yes	146	67	163	13		
Missing	2	1	25	2		

TABLE 1. Continued

	Cannabis problem users		Comparison individuals		Chi square*	<i>p</i> value, two sided
	No.	%	No.	%		
Behavior problems‡						41.98 <0.005
Lower	32	15	397	31		
Low	69	32	377	30		
Moderate	30	14	171	13		
High	25	12	60	5		
Higher	10	5	21	2		
Missing	51	24	251	20		
Shyness‡						3.24 0.70
Lower	7	3	65	5		
Low	49	23	310	24		
Moderate	76	35	457	36		
High	31	14	171	13		
Higher	3	1	23	2		
Missing	51	24	251	20		
Depression symptoms§						4.40 0.22
Low	29	13	182	14		
Moderate	117	54	764	60		
High	14	6	64	5		
Missing	57	26	267	21		
Anxiety symptoms§						5.47 0.14
Low	45	21	235	18		
Moderate	94	43	657	51		
High	21	10	118	9		

Does using cannabis as an adolescent cause depression as a young-adult? Analysis

TABLE 1. Baseline characteristics of 1,494 adolescent-onset cannabis problem users and comparison individuals from the original 2,311 individuals in the Prevention Research Center cohort, United States, 1985–2001

	Cannabis problem users		Comparison individuals		Chi square*	<i>p</i> value, two sided
	No.	%	No.	%		
Sex					63.54	<0.005
Male	151	70	517	40		
Female	66	30	760	60		
Race					19.05	<0.005
Black	132	61	948	74		
White	84	39	315	25		
Other†	1	0	14	1		
Family income					1.90	0.59
Low	19	9	115	9		
Middle	54	25	373	29		
High	71	33	381	30		
Missing	73	34	408	32		
Free lunch					1.68	0.43
No	62	29	314	25		
Yes	149	69	931	73		
Missing	6	3	32	3		
Daily tobacco smoker					336.08	<0.005
No	69	32	1,089	85		
Yes	146	67	163	13		
Missing	2	1	25	2		

TABLE 1. Continued

	Cannabis problem users		Comparison individuals		Chi square*	<i>p</i> value, two sided
	No.	%	No.	%		
Behavior problems‡						41.98 <0.005
Lower	32	15	397	31		
Low	69	32	377	30		
Moderate	30	14	171	13		
High	25	12	60	5		
Higher	10	5	21	2		
Missing	51	24	251	20		
Shyness‡						3.24 0.70
Lower	7	3	65	5		
Low	49	23	310	24		
Moderate	76	35	457	36		
High	31	14	171	13		
Higher	3	1	23	2		
Missing	51	24	251	20		
Depression symptoms§						4.40 0.22
Low	29	13	182	14		
Moderate	117	54	764	60		
High	14	6	64	5		
Missing	57	26	267	21		
Anxiety symptoms§						5.47 0.14
Low	45	21	235	18		
Moderate	94	43	657	51		
High	21	10	118	9		

Does using canr as a young-adul Analysis

TABLE 1. Baseline characteristics of 1,494 adolescent-onset cannabis problem users and comparison individuals from the original 2,311 individuals in the Prevention Research Center cohort, United States, 1985–2001

	Cannabis problem users		Comparison individuals		Chi square*	<i>p</i> value, two sided
	No.	%	No.	%		
Sex					63.54	<0.005
Male	151	70	517	40		
Female	66	30	760	60		
Race					19.05	<0.005
Black	132	61	948	74		
White	84	39	315	25		
Other†	1	0	14	1		
Family income					1.90	0.59
Low	19	9	115	9		
Middle	54	25	373	29		
High	71	33	381	30		
Missing	73	34	408	32		
Free lunch					1.68	0.43
No	62	29	314	25		
Yes	149	69	931	73		
Missing	6	3	32	3		
Daily tobacco smoker					336.08	<0.005
No	69	32	1,089	85		
Yes	146	67	163	13		
Alcohol abuse or dependence					376.62	<0.005
No	113	52	1,210	95		
Yes	93	43	39	3		
Missing	11	5	28	2		
Other illegal drug use					21.06	<0.005
No	206	95	1,262	99		
Yes	7	3	5	0		
Missing	4	2	10	1		

TABLE 1. Continued

	Cannabis problem users		Comparison individuals		Chi square*	<i>p</i> value, two sided
	No.	%	No.	%		
Behavior problems‡						41.98 <0.005
Lower	32	15	397	31		
Low	69	32	377	30		
Moderate	30	14	171	13		
High	25	12	60	5		
Higher	10	5	21	2		
Missing	51	24	251	20		
Shyness‡						3.24 0.70
Lower	7	3	65	5		
Low	49	23	310	24		
Moderate	76	35	457	36		
High	31	14	171	13		
Higher	3	1	23	2		
Missing	51	24	251	20		
Depression symptoms§						4.40 0.22
Low	29	13	182	14		
Moderate	117	54	764	60		
High	14	6	64	5		
Missing	57	26	267	21		
Anxiety symptoms§						5.47 0.14
Low	45	21	235	18		
Moderate	94	43	657	51		
High	21	10	118	9		
Missing	57	26	267	21		
Intervention status (classroom)						0.31 0.86
Standard setting	129	59	736	58		
Good behavior game	42	19	266	21		
Mastery learning	46	21	275	22		
Intervention status (school)						0.04 0.98

n

Does using canr as a young-adul Analysis

TABLE 1. Baseline characteristics of 1,494 adolescent-onset cannabis problem users and comparison individuals from the original 2,311 individuals in the Prevention Research Center cohort, United States, 1985–2001

	Cannabis problem users		Comparison individuals		Chi square*	<i>p</i> value, two sided
	No.	%	No.	%		
Sex					63.54	<0.005
Male	151	70	517	40		
Female	66	30	760	60		
Race					19.05	<0.005
Black	132	61	948	74		
White	84	39	315	25		
Other†	1	0	14	1		
Family income					1.90	0.59
Low	19	9	115	9		
Middle	54	25	373	29		
High	71	33	381	30		
Missing	73	34	408	32		
Free lunch					1.68	0.43
No	62	29	314	25		
Yes	149	69	931	73		
Missing	6	3	32	3		
Daily tobacco smoker					336.08	<0.005
No	69	32	1,089	85		
Yes	146	67	163	13		
Alcohol abuse or dependence					376.62	<0.005
No	113	52	1,210	95		
Yes	93	43	39	3		
Missing	11	5	28	2		
Other illegal drug use					21.06	<0.005
No	206	95	1,262	99		
Yes	7	3	5	0		
Missing	4	2	10	1		

TABLE 1. Continued

	Cannabis problem users		Comparison individuals		Chi square*	<i>p</i> value, two sided
	No.	%	No.	%		
Behavior problems‡						41.98 <0.005
Lower	32	15	397	31		
Low	69	32	377	30		
Moderate	30	14	171	13		
High	25	12	60	5		
Higher	10	5	21	2		
Missing	51	24	251	20		
Shyness‡						3.24 0.70
Lower	7	3	65	5		
Low	49	23	310	24		
Moderate	76	35	457	36		
High	31	14	171	13		
Higher	3	1	23	2		
Missing	51	24	251	20		
Depression symptoms§						4.40 0.22
Low	29	13	182	14		
Moderate	117	54	764	60		
High	14	6	64	5		
Missing	57	26	267	21		
Anxiety symptoms§						5.47 0.14
Low	45	21	235	18		
Moderate	94	43	657	51		
High	21	10	118	9		
Missing	57	26	267	21		
Intervention status (classroom)						0.31 0.86
Standard setting	129	59	736	58		
Good behavior game	42	19	266	21		
Mastery learning	46	21	275	22		
Intervention status (school)						0.04 0.98

n

Does using cannabis as an adolescent cause depression as a young-adult?

Analysis

- Understand how the treated and control differ before adjustments
- With dozens of covariates: read them one-by-one
- With \geq hundreds of covariates: grouped analysis, outliers
- Here:
 - Cannabis users and non-users had similar levels of pre-use depression
 - Cannabis users more prone to pre-exposure alcohol and tobacco use
 - Cannabis users had slightly higher pre-exposure concentration and behavioral problems at school

Does using cannabis as an adolescent cause depression as a young-adult?

Analysis

- *What does it even mean to choose "the right" parametric model estimates when any chosen model depends on assumptions we cannot verify*

Ho, Daniel E., et al. "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference." *Political analysis* 15.3 (2007): 199-236.

Does using cannabis as an adolescent cause depression as a young-adult?

Analysis

- Focus on propensity score methods:
Estimate propensity score with three different methods
 1. Logistic regression
 2. Logistic regression with “carefully chosen interaction terms”
 3. Gradient boosted decision trees
- Propensity scores used in three different ways:
 1. 1:1 matching
 2. Full matching
 3. Inverse propensity score weighting

Does using cannabis as an adolescent cause depression as a young-adult?

Analysis

- 3 prop. score models X 3 estimation methods = 9 models
- Choose between them based on:
 - Minimizing standardized difference in means across most covariates
 - Give more importance to balance “theoretically critical covariates”
 - Avoid outliers: covariates where balance is very bad
- Best models:
 - For female sample: LR+full matching, LR+IPW
 - For male sample: GB decision tree + IPW, LR+IPW
 - For combined sample: GB decision tree + IPW
- Present five analyses (2 for females, 2 for males, 1 for combined)

Does using cannabis as an adolescent cause depression
as a young-adult?

Analysis

- For each chosen method run covariate adjustment on re-weighted / matched sample, using logistic regression
- Similar idea to double-robustness

Does using cannabis as an adolescent cause depression as a young-adult?

	Males						Females†					
	Cannabis problem users		Comparison individuals			Chi square, unadjusted‡	Chi square, propensity score adjusted‡	Cannabis problem users		Comparison individuals		
	No.	% unadjusted	No.	% unadjusted	% propensity score adjusted			No.	% unadjusted	No.	% unadjusted	% propensity score adjusted
Behavior problems						20.92*	3.36					
Lower	13	9	104	20	10			19	29	293	39	33
Low	50	33	158	31	36			19	29	219	29	24
Moderate	24	16	92	18	12			6	9	79	10	7
High	22	15	37	7	12			3	5	29	4	3
Higher	10	7	15	3	7							
Missing	32	21	111	21	24			19	29	140	18	33
Shyness						3.39	7.5					
Lower	2	1	20	4	2			5	8	45	6	7
Low	36	24	102	20	19			13	20	208	27	25
Moderate	55	36	197	38	40			21	32	260	34	27
High	23	15	78	15	14			8	12	107	14	8
Higher	3	2	9	2	0							
Missing	32	21	111	21	24			19	29	140	18	32
Depression symptoms						3.09	2.06					
Low	22	15	81	16	16			11	101	13	19	
Moderate	83	55	300	58	54			34	52	464	61	45
High	10	7	18	3	5			4	6	46	6	2
Missing	36	24	118	23	25			21	32	149	20	34
Anxiety symptoms						1.30	3.87					
Low	36	24	104	20	20			9	14	131	17	10
Moderate	70	46	261	50	45			24	36	396	52	44
High	9	6	34	7	9			12	18	84	11	12
Missing	36	24	118	23	25			21	32	149	20	34



Does using cannabis as an adolescent cause depression as a young-adult?

Results

- Odds ratio:

$$\frac{\hat{p}(y = 1|x, t = 1)/\hat{p}(y = 0|x, t = 1)}{\hat{p}(y = 1|x, t = 0)/\hat{p}(y = 0|x, t = 0)}$$

- In a logistic regression model:

$$\hat{p}(y = 1|x, t = 1) = \frac{1}{1 + \exp(-\eta^\top x - \beta t)}$$

- Doing the math, for LR model:

$$OR = \exp(\beta)$$

Does using cannabis as an adolescent cause depression as a young-adult?

Results

TABLE 3. Estimated association by males and females separately, linking young adult depression with adolescent-onset cannabis problems, with covariate adjustment and use of propensity score techniques for 1,494 individuals from the Prevention Research Center cohort, United States, 1985–2001

Propensity score adjustment models	No. of adolescent cannabis problem users	Odds ratio	95% confidence interval	p value, two sided
Males				
GBM* and weighting by the odds	151	1.72	0.77, 3.86	0.19
MLR* and weighting by the odds		1.67	0.77, 3.60	0.19

Does using cannabis as an adolescent cause depression as a young-adult?

Results

TABLE 3. Estimated association by males and females separately, linking young adult depression with adolescent-onset cannabis problems, with covariate adjustment and use of propensity score techniques for 1,494 individuals from the Prevention Research Center cohort, United States, 1985–2001

Propensity score adjustment models	No. of adolescent cannabis problem users	Odds ratio	95% confidence interval	p value, two sided
Males				
GBM* and weighting by the odds	151	1.72	0.77, 3.86	0.19
MLR* and weighting by the odds		1.67	0.77, 3.60	0.19
Females				
MLR and full matching	66	0.63	0.25, 1.58	0.32
MLR and weighting by the odds		0.68	0.20, 2.34	0.54

Does using cannabis as an adolescent cause depression as a young-adult?

Results

TABLE 3. Estimated association by males and females separately, linking young adult depression with adolescent-onset cannabis problems, with covariate adjustment and use of propensity score techniques for 1,494 individuals from the Prevention Research Center cohort, United States, 1985–2001

Propensity score adjustment models	No. of adolescent cannabis problem users	Odds ratio	95% confidence interval	p value, two sided
Males				
GBM* and weighting by the odds	151	1.72	0.77, 3.86	0.19
MLR* and weighting by the odds		1.67	0.77, 3.60	0.19
Females				
MLR and full matching	66	0.63	0.25, 1.58	0.32
MLR and weighting by the odds		0.68	0.20, 2.34	0.54
Combined sample				
GBM and weighting by the odds	217	1.33	0.76, 2.33	0.32

Does using cannabis as an adolescent cause depression as a young-adult?

Results

- All five analyses do not indicate a significant association neither for females or males
- They also tried straightforward covariate adjustment (no matching or propensity score re-weighting) and found for males OR of 2.6 with p-value smaller than 0.01
- Which estimate to trust?

Study's conclusions

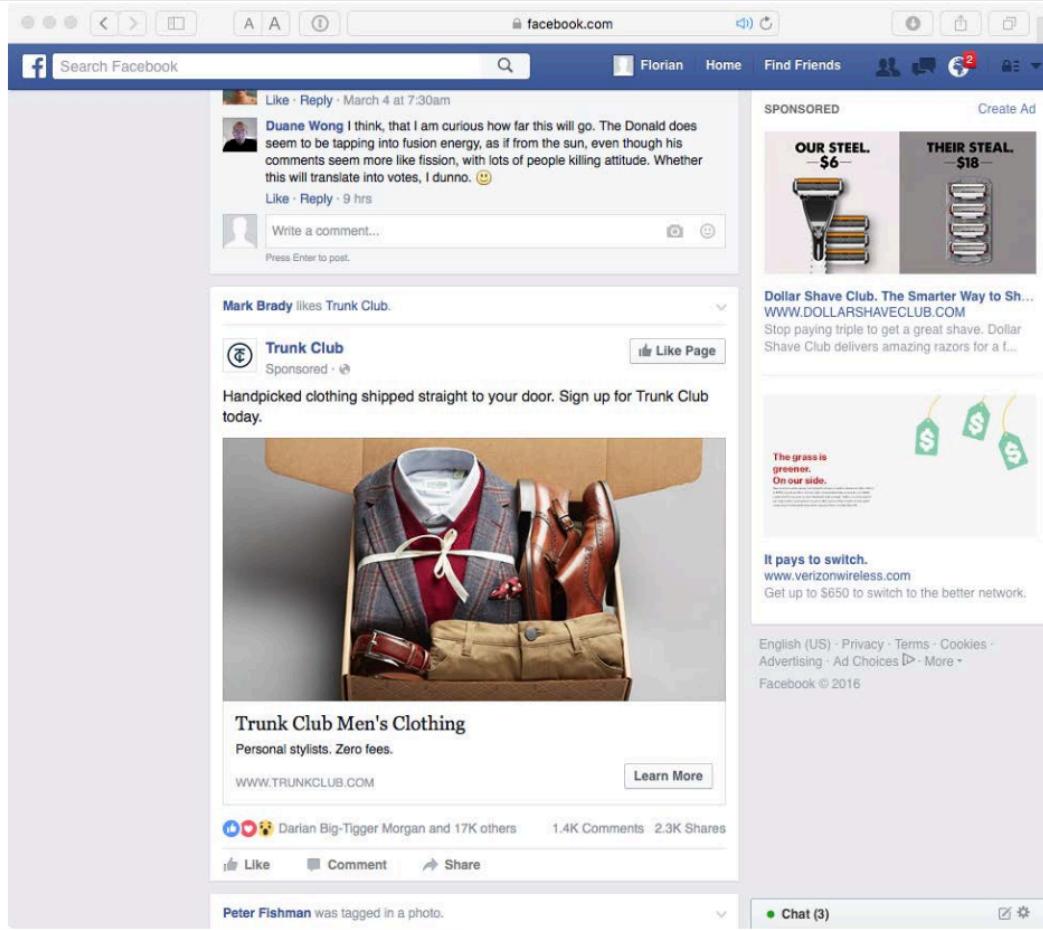
- Study **does not** support the hypothesis that adolescent-onset cannabis problem use causes young adult depression.
- Two other causal hypotheses remain:
 - Depression causes individuals to manage their symptoms through self medication by use of cannabis, and
 - A common genetic or environmental influence causes both depression and cannabis use.

Meta studies

- Facebook advertising: continuously run RCTs but can't test all hypotheses
- Study by Brett Gordon, Florian Zettelmeyer, Neha Bhargava and Dan Chapsky
- Slides by Florian Zettelmeyer, Kellogg school of management
https://www.ftc.gov/system/files/documents/public_events/945353/zettelmeyer_fb_fcc_11-3-2016_fz_slides_0.pdf

Facebook advertising show up in the newsfeed or to the right of the page

TRUNK CLUB EXAMPLE



Facebook recently built an experimentation platform

FEATURES OF OUR DATA

- 15 large-scale randomized advertising experiments across verticals**

Facebook recently built an experimentation platform

FEATURES OF OUR DATA

- 15 large-scale randomized advertising experiments across verticals
- Statistical power
 - Between 2 million and 150 million users per experiment
 - 492 million user-study observations
 - 1.5 billion total ad impressions

Facebook recently built an experimentation platform

FEATURES OF OUR DATA

- 15 large-scale randomized advertising experiments across verticals
- Statistical power
 - Between 2 million and 150 million users per experiment
 - 492 million user-study observations
 - 1.5 billion total ad impressions
- Single-user login
 - Eliminates issues with cookie-based measurement
 - Captures cross-device activity

Facebook recently built an experimentation platform

FEATURES OF OUR DATA

- **15 large-scale randomized advertising experiments across verticals**
- **Statistical power**
 - Between 2 million and 150 million users per experiment
 - 492 million user-study observations
 - 1.5 billion total ad impressions
- **Single-user login**
 - Eliminates issues with cookie-based measurement
 - Captures cross-device activity
- **Measure outcomes** (e.g., purchases, registrations) directly via conversion pixels on advertisers' websites—no ad clicks required

Serve the ad that would have been shown in the absence of the Jasper's Market ad campaign

Ad Auction

1.



2.



3.



4.



A Facebook feed on a user's timeline. The top post is from 'Brett' with 20+ likes. Below it is a 'Suggested Post' from 'Jasper's Market' (Sponsored) showing a fig tart with almonds. The caption reads: 'It's fig season! Not sure what to do with figs? Here's a great dessert recipe to share.' Below the image is the recipe title 'Fig Tart with Almonds' and a brief description. At the bottom of the post is the URL 'WWW.JASPERSMARKET.COM'. The post has 91 Likes, 4 Comments, and 14 Shares. The bottom right of the post area shows the text 'Like Comment Share'. To the right of the post is a sponsored ad for '+1 Movie Friday'.

A Facebook feed on a user's timeline. The top post is from 'Brett' with 20+ likes. Below it is a 'Suggested Post' from 'Waterford Lux Resorts' (Sponsored) showing a woman using a mobile app. The caption reads: 'Plan your next vacation and find exclusive, new deals now with our Waterford Lux Resorts travel app.' Below the image is the app description. At the bottom of the post is the URL 'WWW.WATERFORDLUXRESORTS.COM'. The post has 1,962 Likes. The bottom right of the post area shows the text 'Like Comment Share'. To the right of the post is a sponsored ad for '+1 Movie Friday'.

Test

Control

Results: ATT Lift

Average Treatment Effect on the Treated (ATT)

- Intent-to-Treat (ITT) effect = 0.012%
- 25% of users exposed in the test group
- $\text{ATT} = 0.012\% / 0.25 = 0.045\%$

Results: ATT Lift

Average Treatment Effect on the Treated (ATT)

- Intent-to-Treat (ITT) effect = 0.012%
- 25% of users exposed in the test group
- $\text{ATT} = 0.012\% / 0.25 = 0.045\%$

ATT Lift

- Conversion rate of treated (exposed) users: 0.107%
- Conversion rate if treated had not been treated: $0.107\% - 0.045\% = 0.062\%$
- $\text{Lift} = 0.045\% / 0.062\% = 73\% \quad 95\% \text{ CI} = [33, 113]$

Observational Methods

- Exact Matching (**EM**)
 - Age and gender
- Propensity Score Matching (**PSM**)
 - Logit propensity, 4 nearest neighbors
- Regression Adjustment (**RA**)
 - Inverse Probability-Weighed Regression Adjustment (**IPWRA**)
- Stratification & Regression (**STRAT**)

Unconfoundedness Assumption
 $(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \mid X_i$

Sequence of variables for the observational methods

EM: Age and gender

Sequence of variables for the observational methods

EM: Age and gender

PSM, IPWRA, STRAT:

1. Age, gender, # days on FB, FB age, friends, initiated friends, relationship status, mobile OS, tablet OS, market fixed effects, day fixed effects, etc.

Sequence of variables for the observational methods

EM: Age and gender

PSM, IPWRA, STRAT:

1. Age, gender, # days on FB, FB age, friends, initiated friends, relationship status, mobile OS, tablet OS, market fixed effects, day fixed effects, etc.
2. Same as 1 + Census/ACS data matched by zip code

Sequence of variables for the observational methods

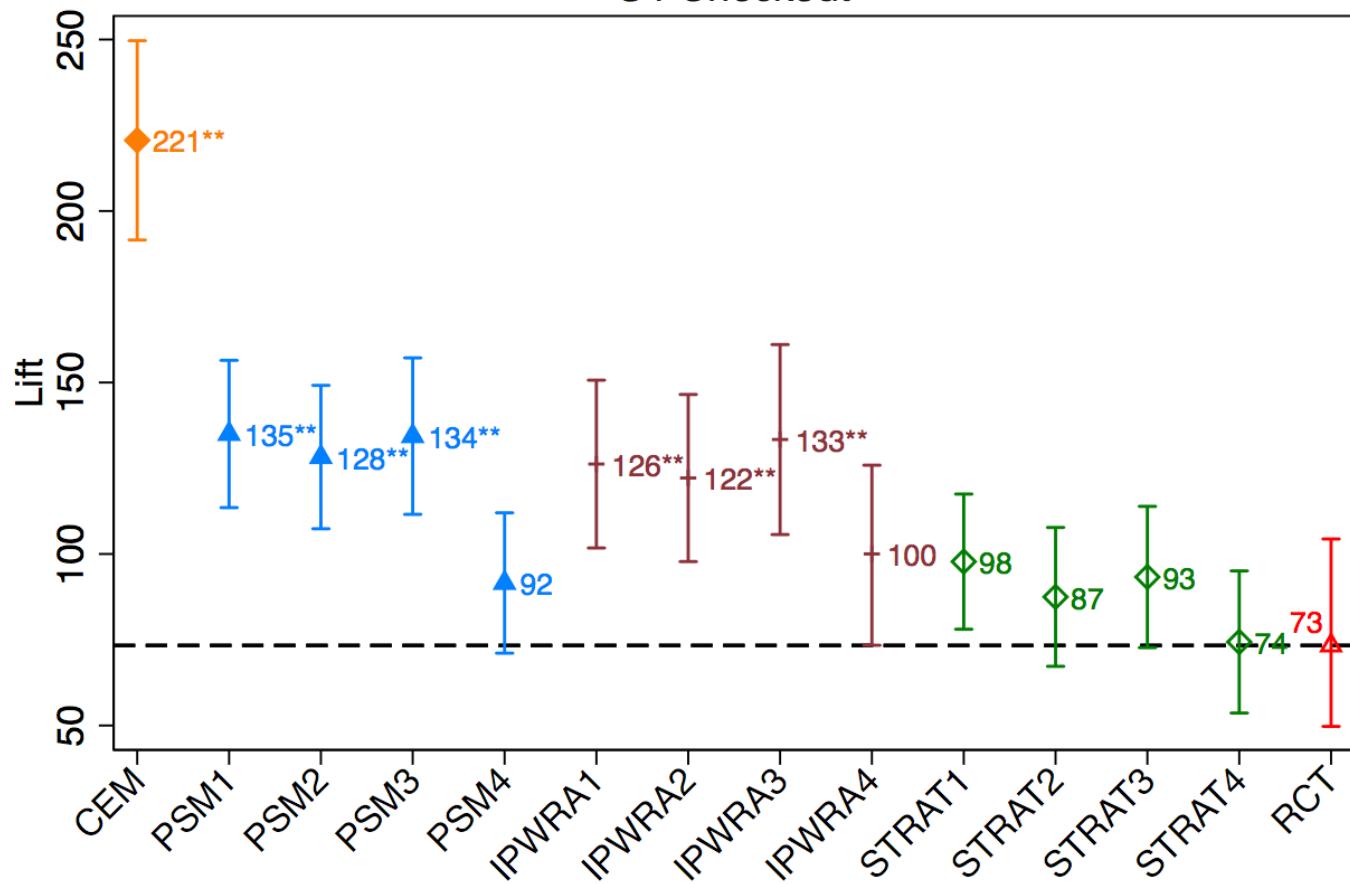
EM: Age and gender

PSM, IPWRA, STRAT:

1. Age, gender, # days on FB, FB age, friends, initiated friends, relationship status, mobile OS, tablet OS, market fixed effects, day fixed effects, etc.
2. Same as 1 + Census/ACS data matched by zip code
3. Same as 2 + Facebook User Activity (binned)
4. Same as 3 + Facebook Match Score

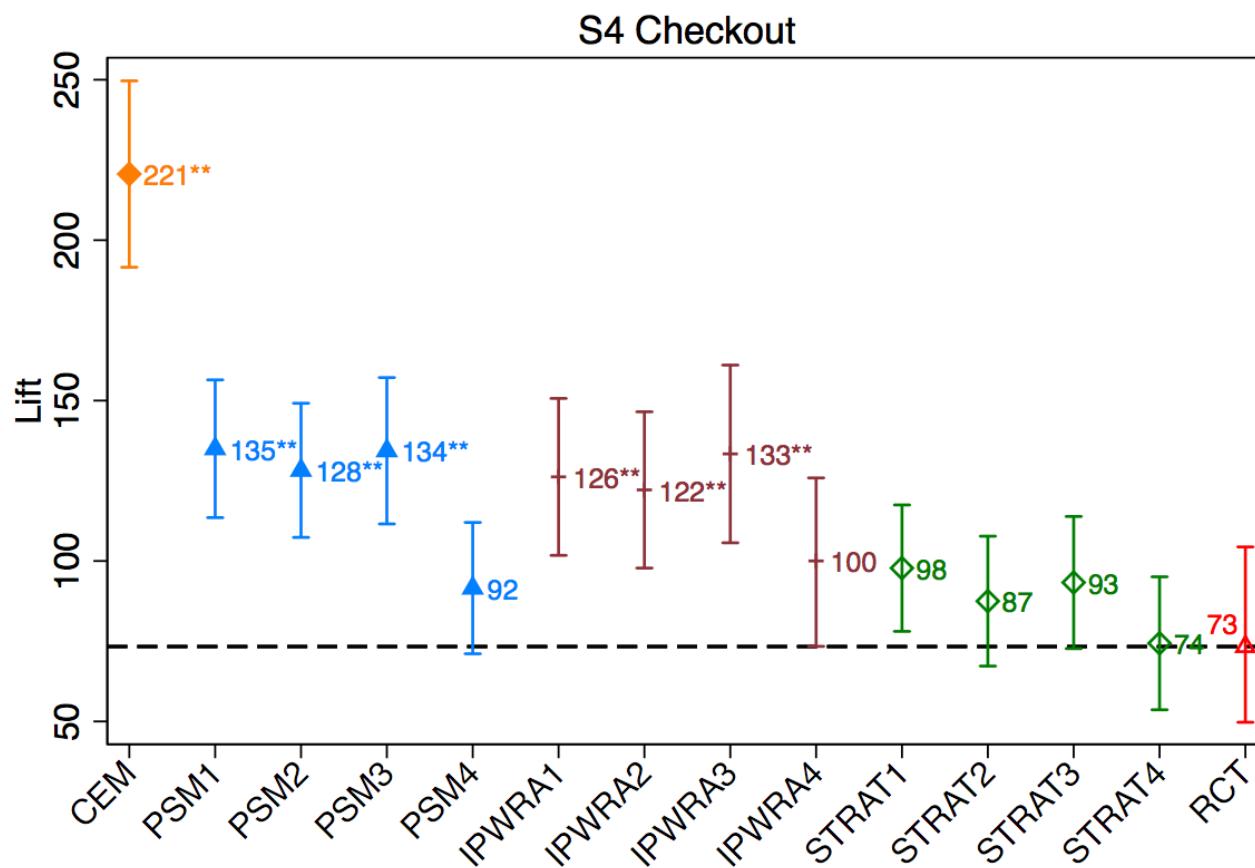
Exposed-unexposed
Lift = 416%

S4 Checkout

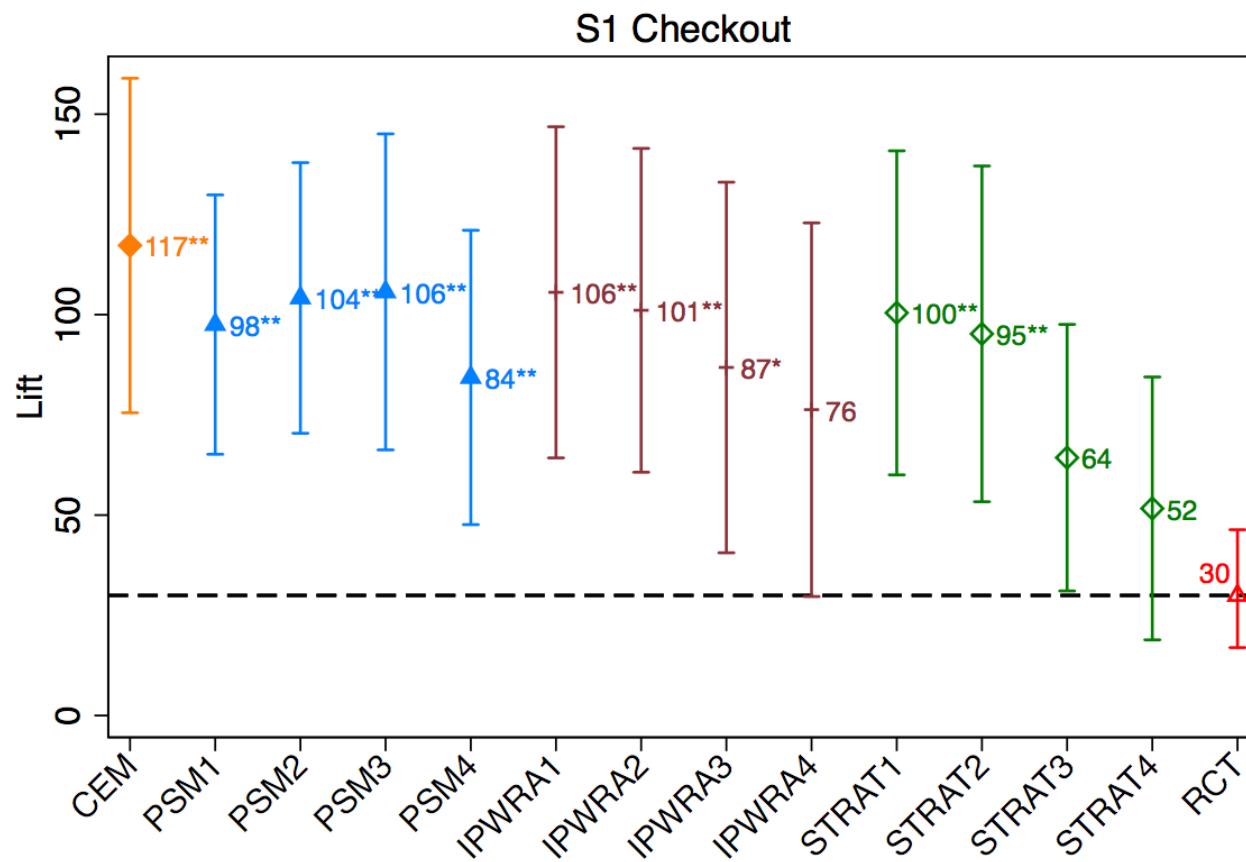


Benchmark (RCT)
Lift = 73%

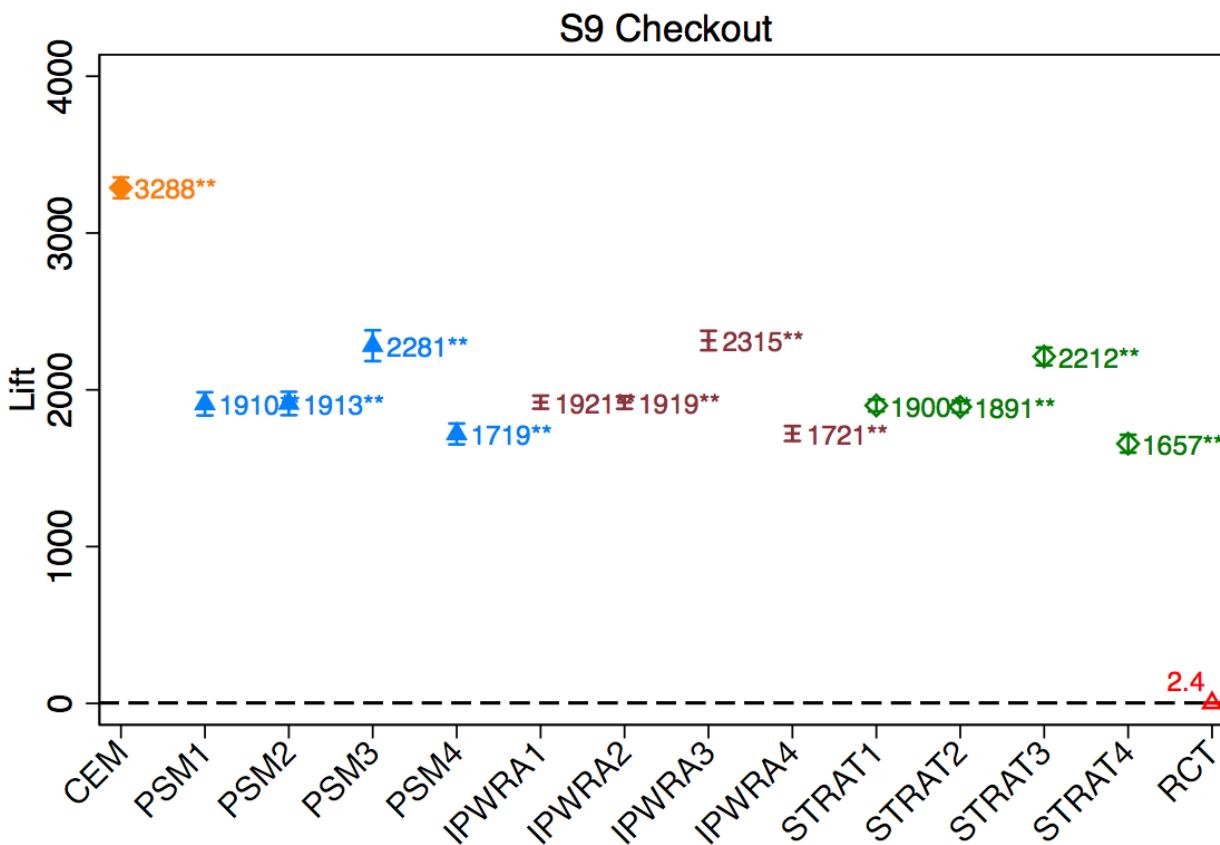
In some studies observational methods come close...



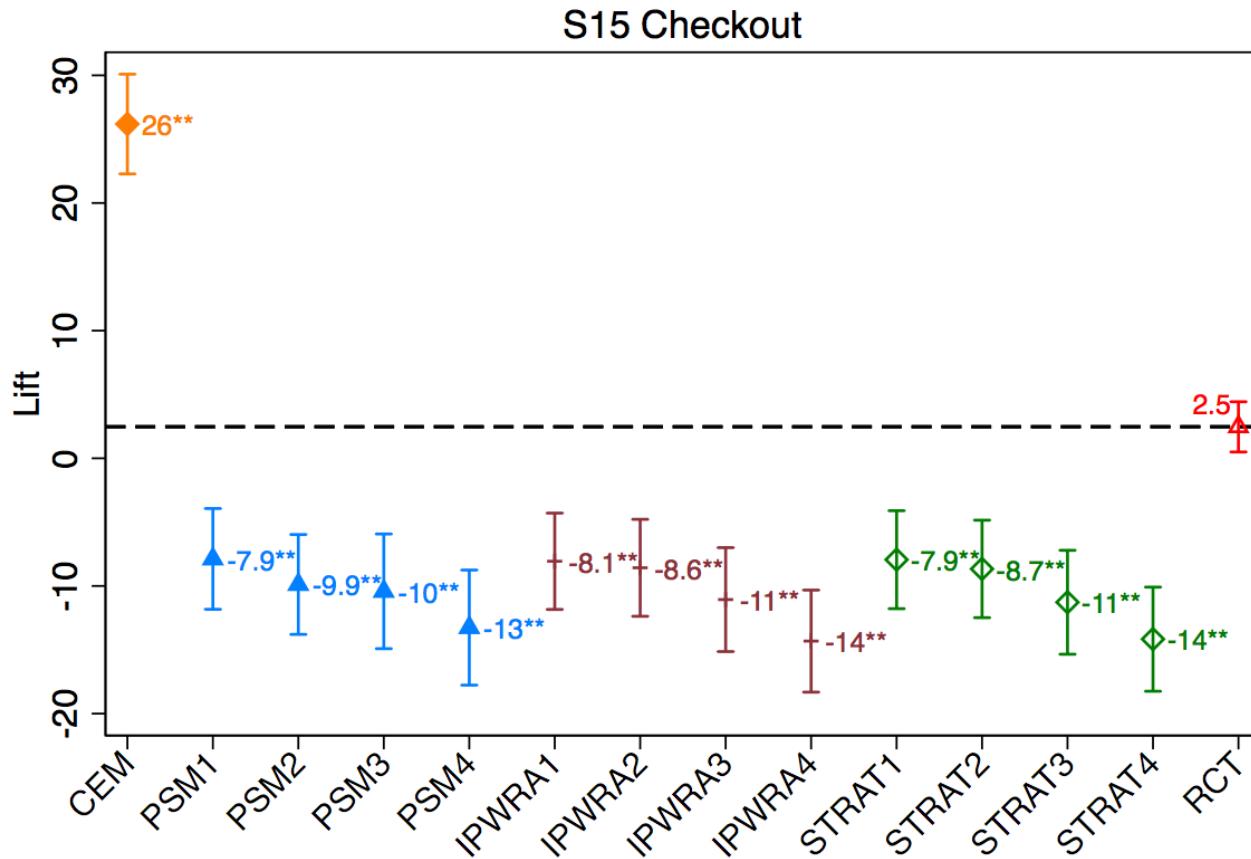
...and there might be a consistent pattern across methods



In other studies, lift estimates from observational methods widely overstate the RCT lift...



**...and sometimes the observational methods
underestimate the lift**



Conclusion

- There is a significant discrepancy between the commonly-used approaches and our true experiments in our studies

Conclusion

- There is a significant discrepancy between the commonly-used approaches and our true experiments in our studies
- While observations approaches sometimes come close to recovering the measurement from true experiments, it is difficult to predict *a priori* when this might occur

Conclusion

- There is a significant discrepancy between the commonly-used approaches and our true experiments in our studies
- While observations approaches sometimes come close to recovering the measurement from true experiments, it is difficult to predict a priori when this might occur
- Measurements are unreliable for checkout conversion outcomes

Conclusion

- There is a significant discrepancy between the commonly-used approaches and our true experiments in our studies
- While observations approaches sometimes come close to recovering the measurement from true experiments, it is difficult to predict a priori when this might occur
- Measurements are unreliable for checkout conversion outcomes
- Measurements are more reliable for registration or page view outcomes

Conclusion

- There is a significant discrepancy between the commonly-used approaches and our true experiments in our studies
- While observations approaches sometimes come close to recovering the measurement from true experiments, it is difficult to predict *a priori* when this might occur
- Measurements are unreliable for checkout conversion outcomes
- Measurements are more reliable for registration or page view outcomes
- *Many industry participants seem unaware that this is a problem*

Meta studies: comparing observational studies and randomized trials

- The OHDSI (pronounced like Odyssey, האודיסאה) collaboration: create common data model for health records
- Currently records of 600,000,000 patients worldwide
- Can perform research on medical data without violating privacy
- Open source, open science
- Present ALL results!



Meta studies: comparing observational studies and randomized trials

- Ryan, Patrick B., et al. "A comparison of the empirical performance of methods for a risk identification system." *Drug safety* 36.1 (2013): 143-158.
- Take ground truth yes/no causal links and try to recover from large-scale observational data using the methods we have learned about
- Exposures are prescribed drugs
- Outcomes are adverse side-effects like acute liver injury or gastrointestinal bleeding

A comparison of the empirical performance of methods for a risk identification system

- Ryan, Patrick B., et al. "A comparison of the empirical performance of methods for a risk identification system." *Drug safety* 36.1 (2013): 143-158.
- Take ground truth yes/no causal links and try to recover from large-scale observational data using several observational study designs
- 399 drug-outcome scenarios
 - 165 positive
 - 234 negative
 - Example positive: antibiotics and acute liver injury
 - Example negative: Beta-blockers and gastro-intestinal bleeding

A comparison of the empirical performance of methods for a risk identification system

- Evaluate separately over 5 electronic health record datasets with 1 million to 40 million persons' records
- Emphasis on *designs*
 - Case control: what we call matching
 - Self-controlled cohort: only look at the population who received drug, compare the *same people* before and after first exposure
 - Self-controlled case series: match each exposed patient with their pre-exposed self
 - Time-series modelling methods

Method

Self-controlled cohort (SCC) as implemented in the Observational Screening (OS) package

This is an extension of a traditional cohort epidemiology design where the rate of ADEs can be compared across groups of patients exposed to different medications, allowing comparisons within a cohort population, between treatments, as well as relative to the overall population at large

Self-controlled case series (SCCS)

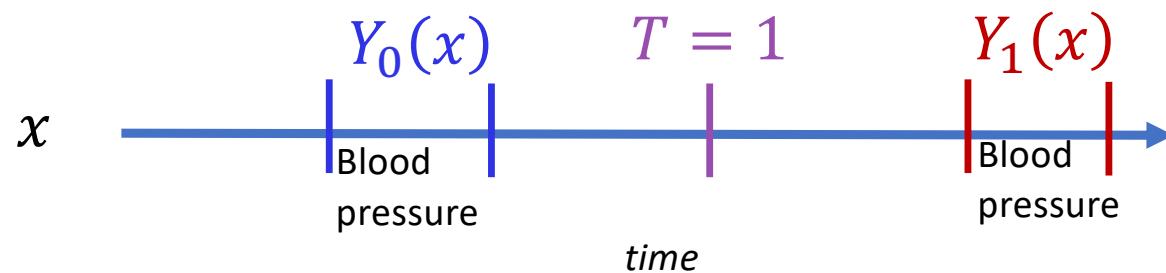
The method estimates the association between a transient exposure and adverse event using only cases; no separate controls are required because each case acts as its own control.

Case control (CC)

The program applies a case-control surveillance design to estimate odds ratios for drug-condition effects, where cases are matched to controls by age, sex, location, and race

A comparison of the empirical performance of methods for a risk identification system

- *self-control design*: look at only treated units, and compare before and after treatment:

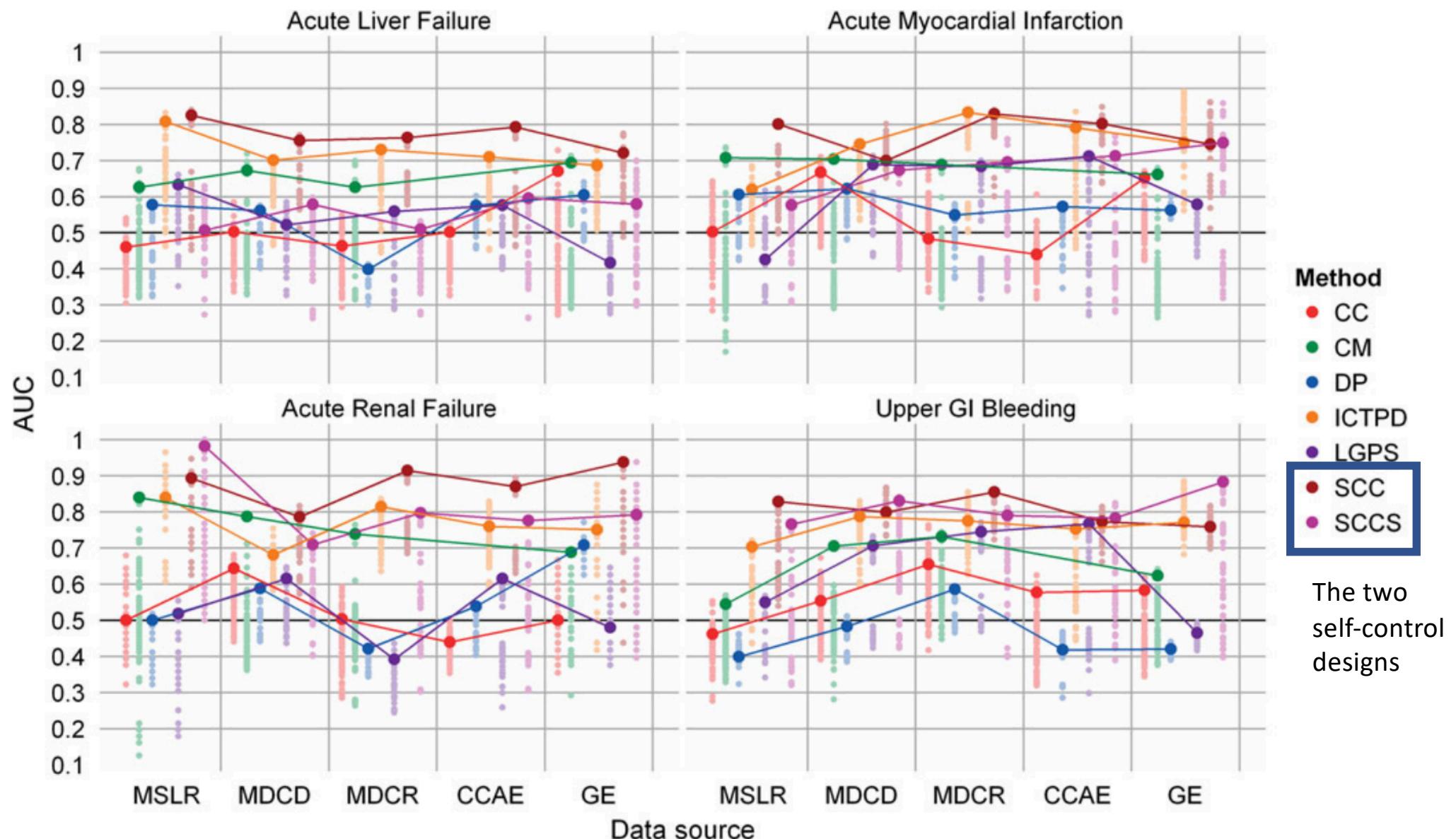


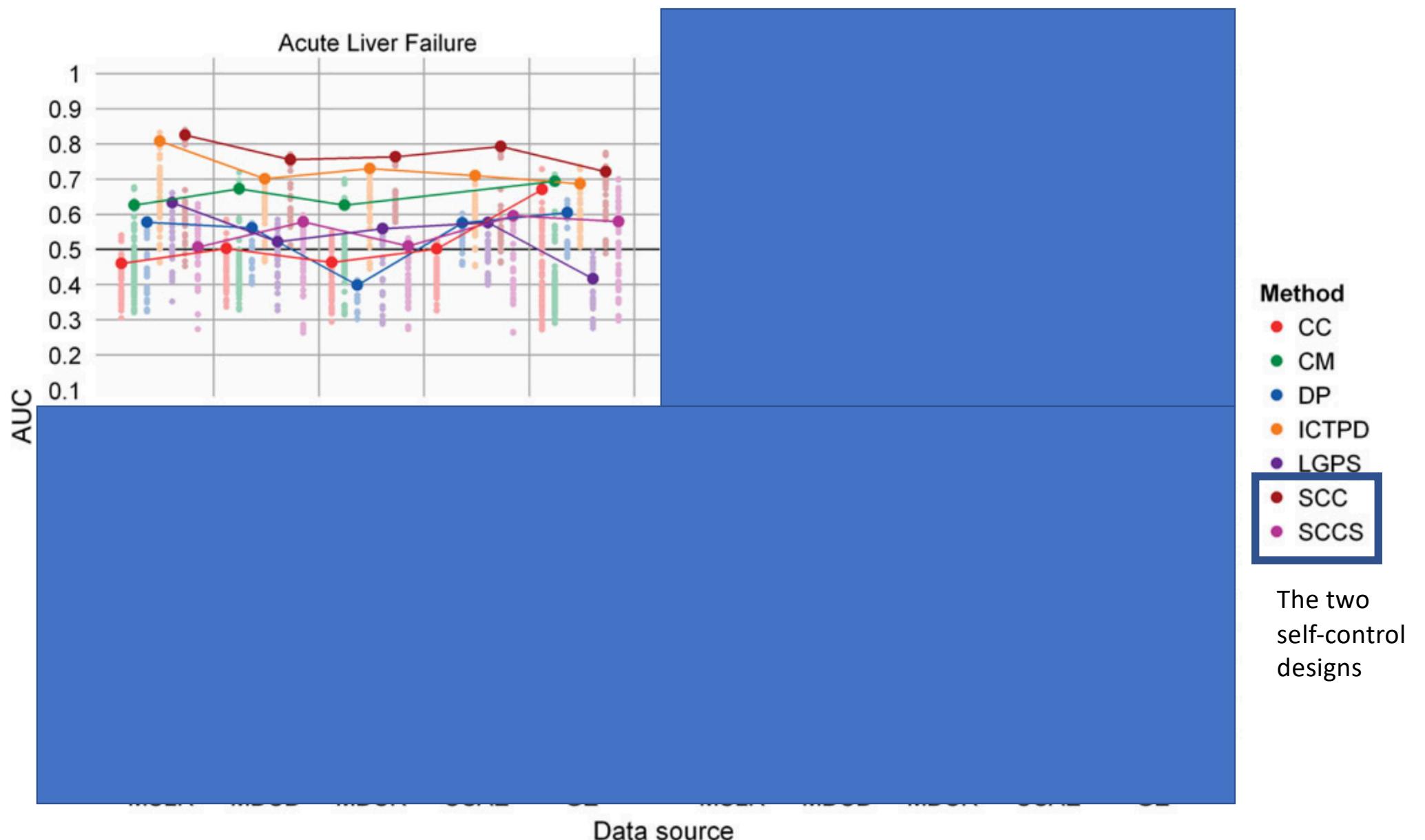
A comparison of the empirical performance of methods for a risk identification system: metrics

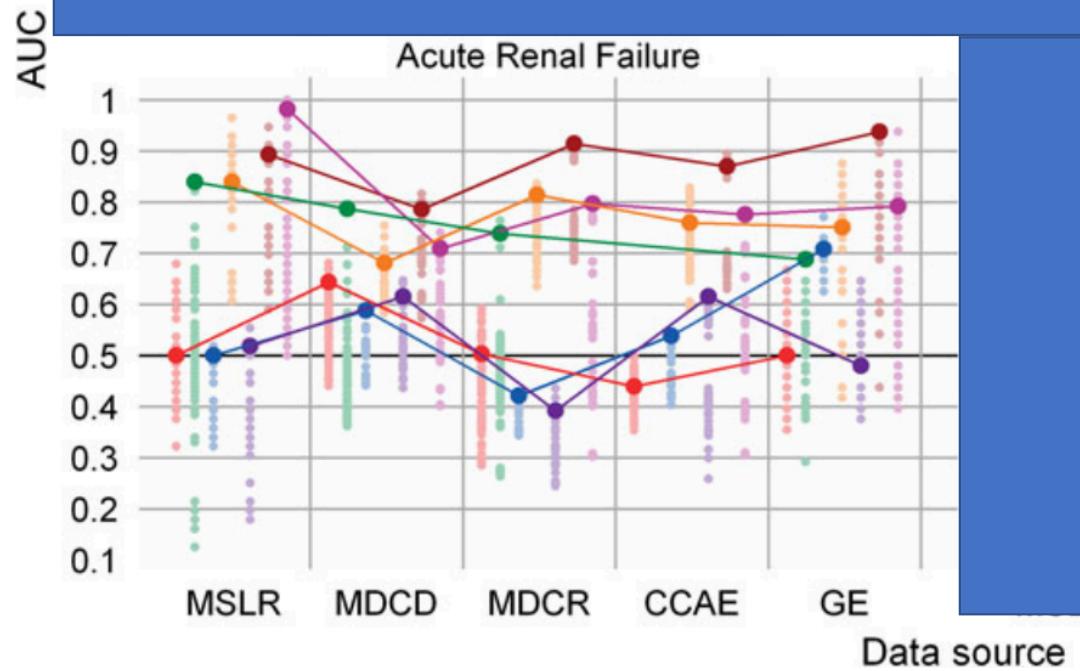
- Label each drug-risk pair as positive or negative
- For each analysis method, rank all 399 drug-risk results by estimated risk
- Calculate Area Under ROC curve
 - Equivalent to ask what proportion of all pairs of (positive, negative) are ranked correctly by the method
- For negative associations, we know ATE = 0 or in the paper's language, the relative-risk=1, $\log(\text{relative-risk})=0$
- Estimate MSE relative to known outcome

A comparison of the empirical performance of methods for a risk identification system: **results**

- Design matters more than analysis choice, but analysis choice does matter
- Several methods, including case-control, can perform badly (AUC around 0.5)
- Self-controlled and self-controlled case-series perform universally better



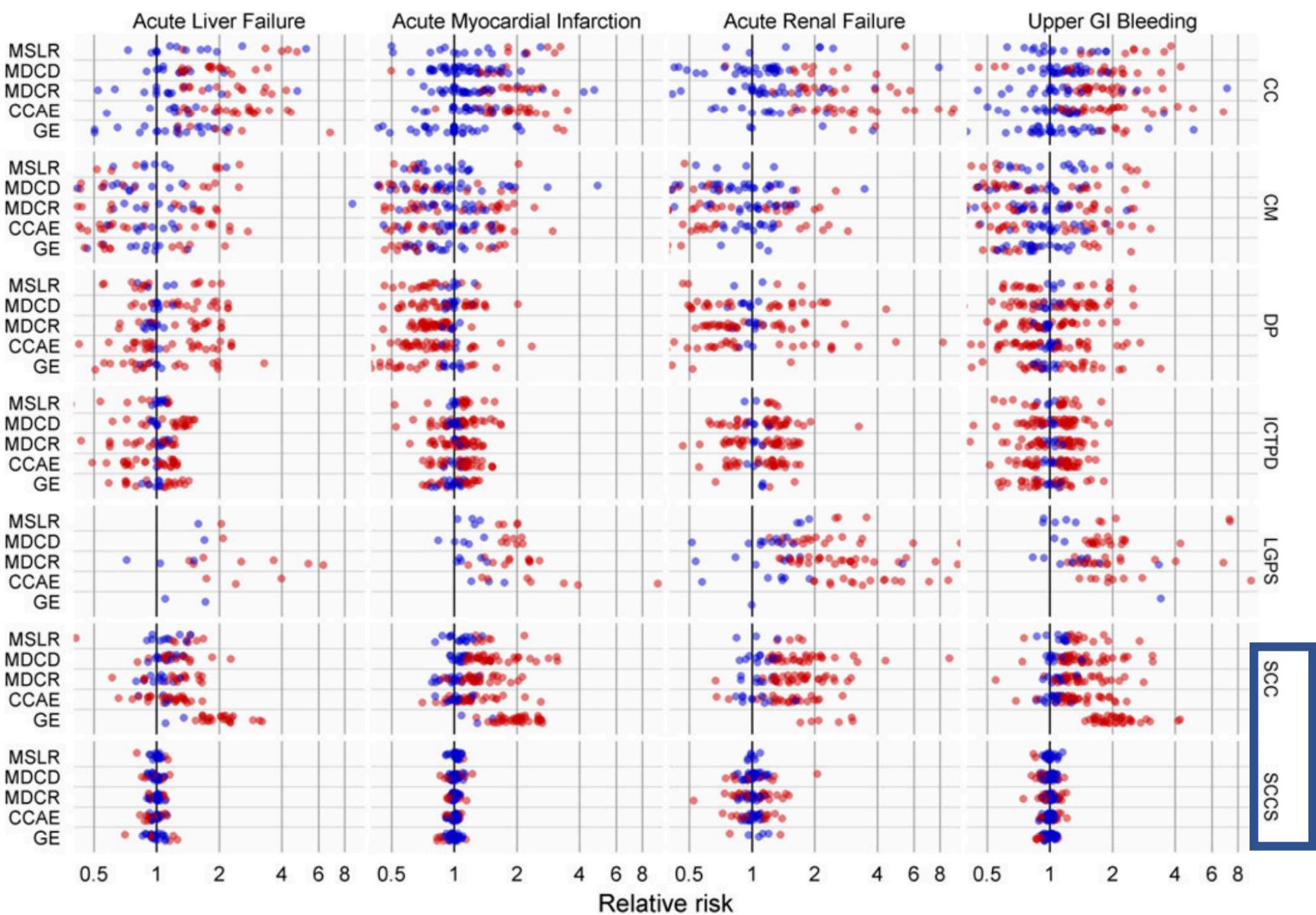




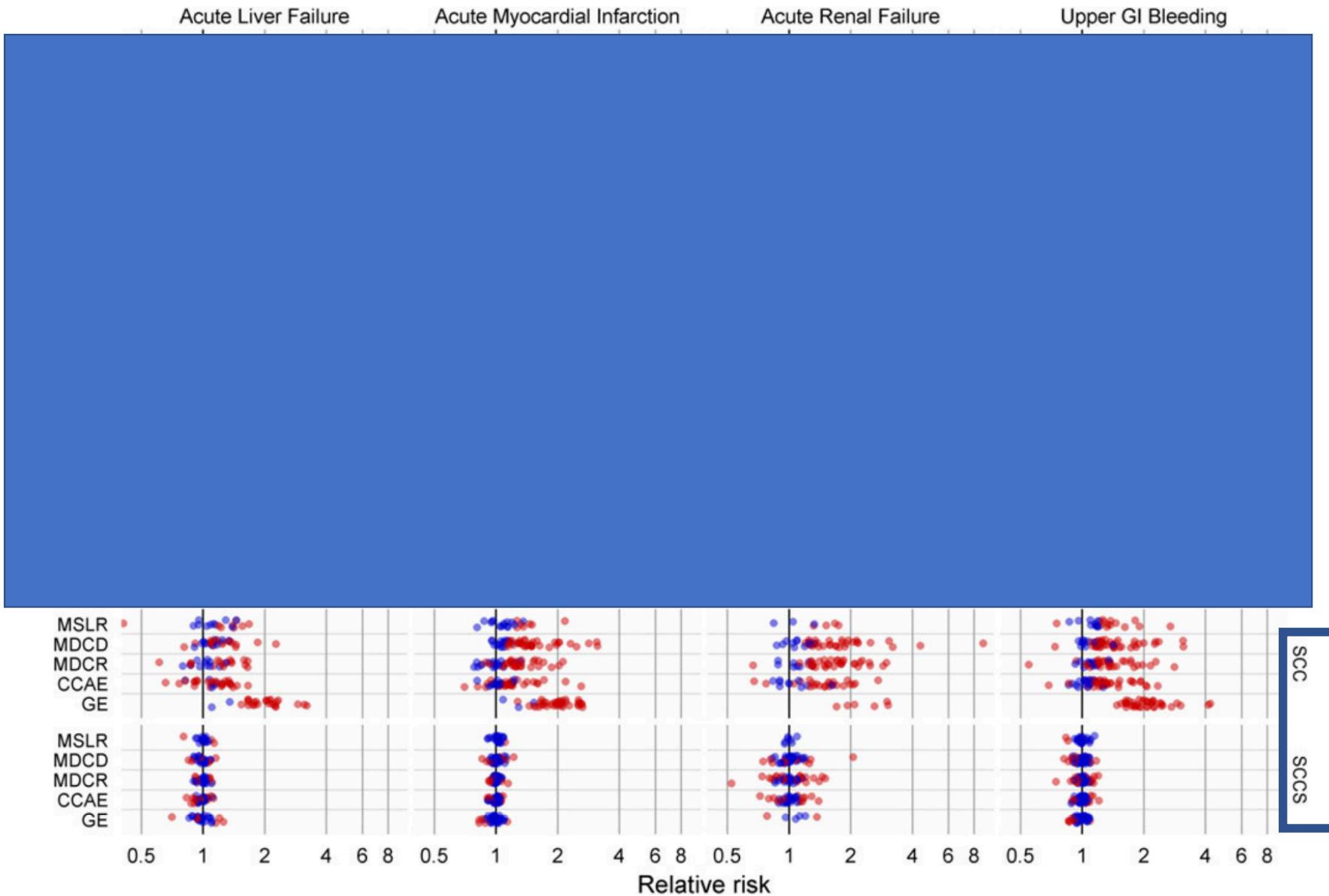
Method

- CC
- CM
- DP
- ICTPD
- LGPS
- SCC
- SCCS

The two self-control designs



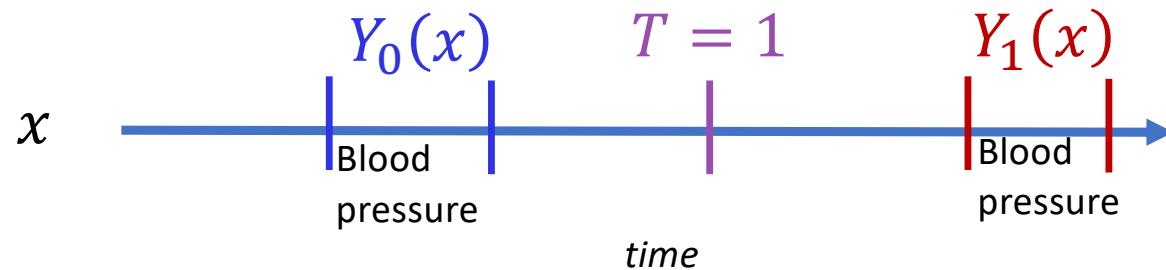
The two
self-
control
designs



The two
self-
control
designs

A comparison of the empirical performance of methods for a risk identification system: Results

- Choice of study design outweighed specific methods choices
- Best results obtained using a *self-control designs*: look at only treated units, and compare before and after treatment:



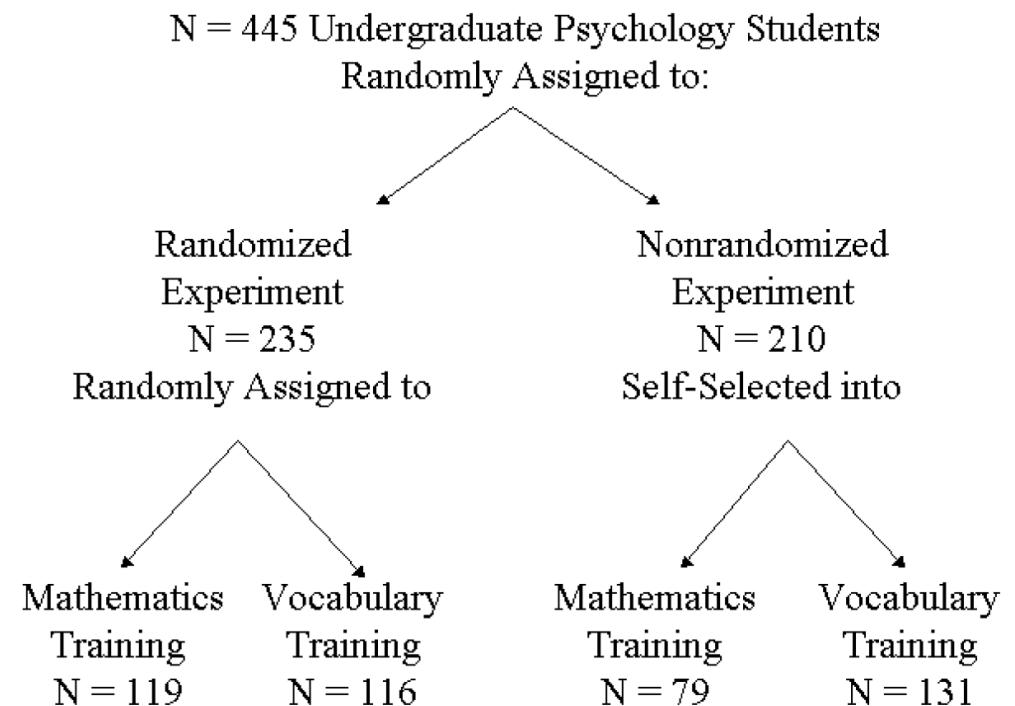
- Achieved AUCs of 0.76-0.94 over the 399 positive or negative treatment-outcome associations

A randomized trial for testing observational studies

- Shadish, William R., Margaret H. Clark, and Peter M. Steiner. "Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments." *Journal of the American statistical association* 103.484 (2008): 1334-1344.

Randomizing whether a student will be in an RCT or an Observational study (“meta-RCT”)

- Meta-treatment: RCT or Obs.
- Treatment: math training or vocabulary training
- Outcome: test results in math and vocabulary
- In the RCT group treatment was randomized
- In the Obs. group treatment was self-selected
- What problems does this raise?



Covariates

- Before assignment, student were tested extensively
 - Math and vocabulary
 - Personality questionnaires
 - Math anxiety test
 - Depression questionnaire
- In addition: age, education, marital status, major area of study, ACT and SAT scores, and grade point average (GPA) for college and high school
- Questionnaire for one of the student's guidance counselors

Results

- Math:

- RCT math training scored an average 4.01 ± 0.35 points higher (out of 18 points)

$$ATE_{math} = 4.01 \pm 0.35$$

- Obs. math training scored an average 5.01 ± 0.55 points higher (out of 18 points)

$$bias_{math} = 1$$

- Vocabulary:

- RCT vocabulary training scored an average 8.25 ± 0.37 higher (out of 30 points)

$$ATE_{vocab} = 8.25 \pm 0.37$$

- Obs. vocabulary training scored an average 9.00 ± 0.51 higher (out of 30 points)

$$bias_{vocab} = 0.75$$

Results

- Bias reduction: 100% = exactly RCT result, 0% = exactly naïve observational study
- Demographic covariates: age, sex, marital status, ethnicity

MATH	
Method	Bias reduction
Propensity score stratification (full covariates)	96%
Propensity score stratification (only demographic covariates)	-5%
Propensity score weighting	70%
Covariate adjustment with OLS	84%

VOCABULARY	
Method	Bias reduction
Propensity score stratification (full covariates)	86%
Propensity score stratification (only demographic covariates)	43%
Propensity score weighting	91%
Covariate adjustment with OLS	94%

Other comparison studies

Can We Trust Observational Studies Using Propensity Scores in the Critical Care Literature? A Systematic Comparison With Randomized Clinical Trials*

Georgios D. Kitsios, MD, PhD¹; Issa J. Dahabreh, MD, MS^{2,3}; Sean Callahan, MD⁴;
Jessica K. Paulus, PhD⁵; Anthony C. Campagna, MD⁶; James M. Dargin, MD⁶

Critical care medicine 43.9 (2015): 1870-1879.

Observational studies in critical care

- Compare 21 published observational studies based on propensity score methods with 58 RCTs for various clinical questions
- In all studies the outcome is mortality
- Conclusions:
- “... propensity score studies ... produced results that were generally consistent with the findings of randomized clinical trials.”
- However, caution is needed when interpreting propensity score studies because occasionally their results contradict those of randomized clinical trials and **there is no reliable way to predict disagreements**

A bottom line

- Sometimes it works
- Sometimes it doesn't
- Strongly depends on dataset and design possibilities
- Often – we have no choice
- Consider the alternatives: not use data?
- “Target trial” paradigm seems to correct some past mistakes
 - Hernán, Miguel A., et al. "Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses." *Journal of clinical epidemiology* 79 (2016): 70-75.
 - García-Albéniz, Xabier, John Hsu, and Miguel A. Hernán. "The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening." *European journal of epidemiology* 32.6 (2017): 495-500.
 - Dickerman, Barbra A., et al. "Avoidable flaws in observational analyses: an application to statins and cancer." *Nature Medicine* 25.10 (2019): 1601-1606.
 - Emilsson, Louise, et al. "Examining bias in studies of statin treatment and survival in patients with cancer." *JAMA oncology* 4.1 (2018): 63-70.
- Scientific challenge: How can we make sure we use the data in a way that is not detrimental?



Introduction to Causal Inference

Dr. Uri Shalit

Course number 097400
2020-2021

Lesson 7

Graphical models

- Probability distributions can factor

$$p(x_1, x_2, \dots, x_d) = \prod_{i=1}^d p(x_i | x_{i-1}, \dots, x_1)$$

- Sometimes simpler factorizations are possible, e.g. Markov chain

$$p(x_1, x_2, \dots, x_d) = \prod_{i=1}^d p(x_i | x_{i-1})$$

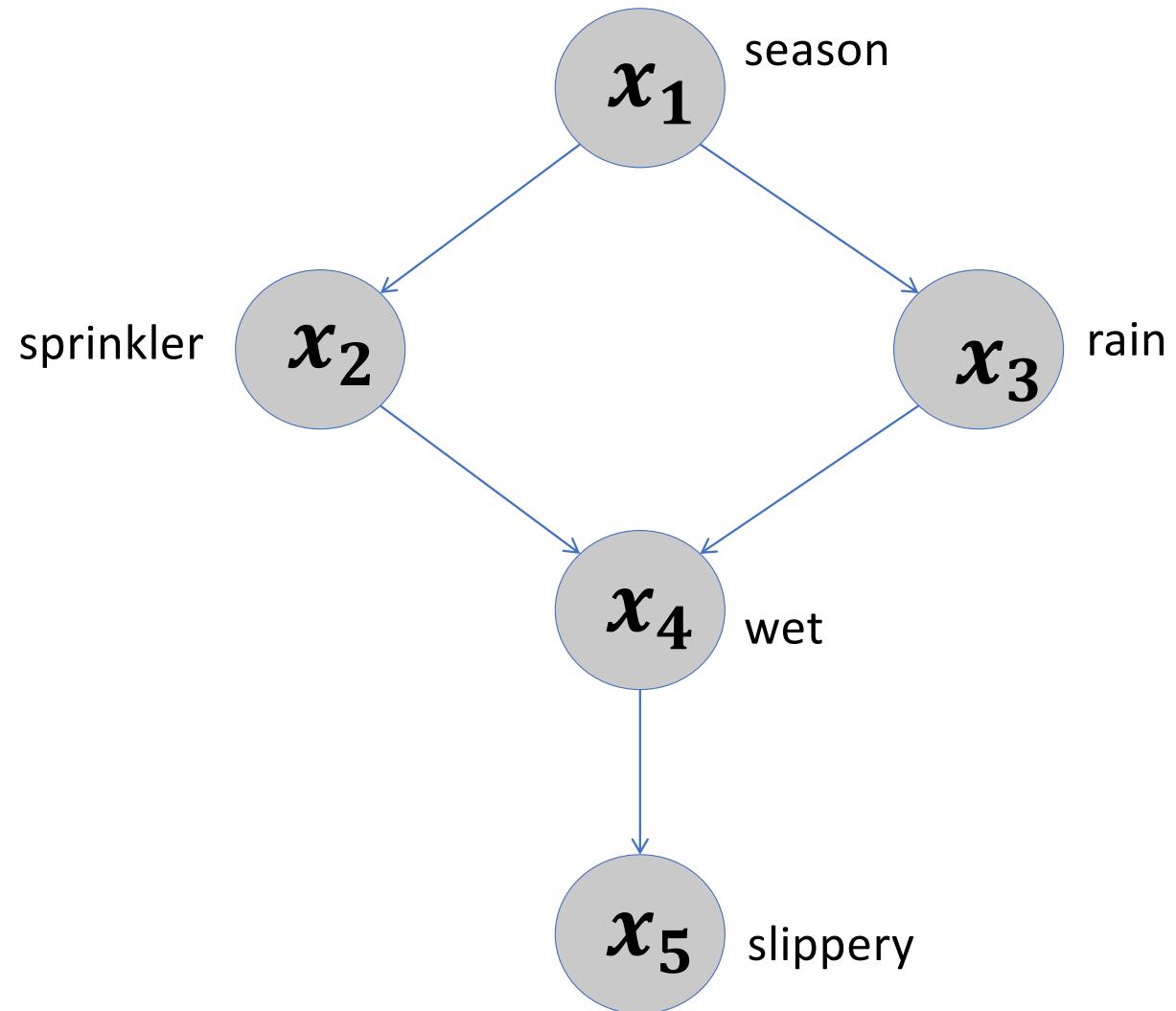
Example

- x_1 : clouds, x_2 : rain, x_3 : dry pavement, x_4 : slippery
- What is $p(\text{clouds}, \text{no-rain}, \text{dry pavement}, \text{slippery})$?
- Assume it factorizes
$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_3)$$
- Easier to estimate

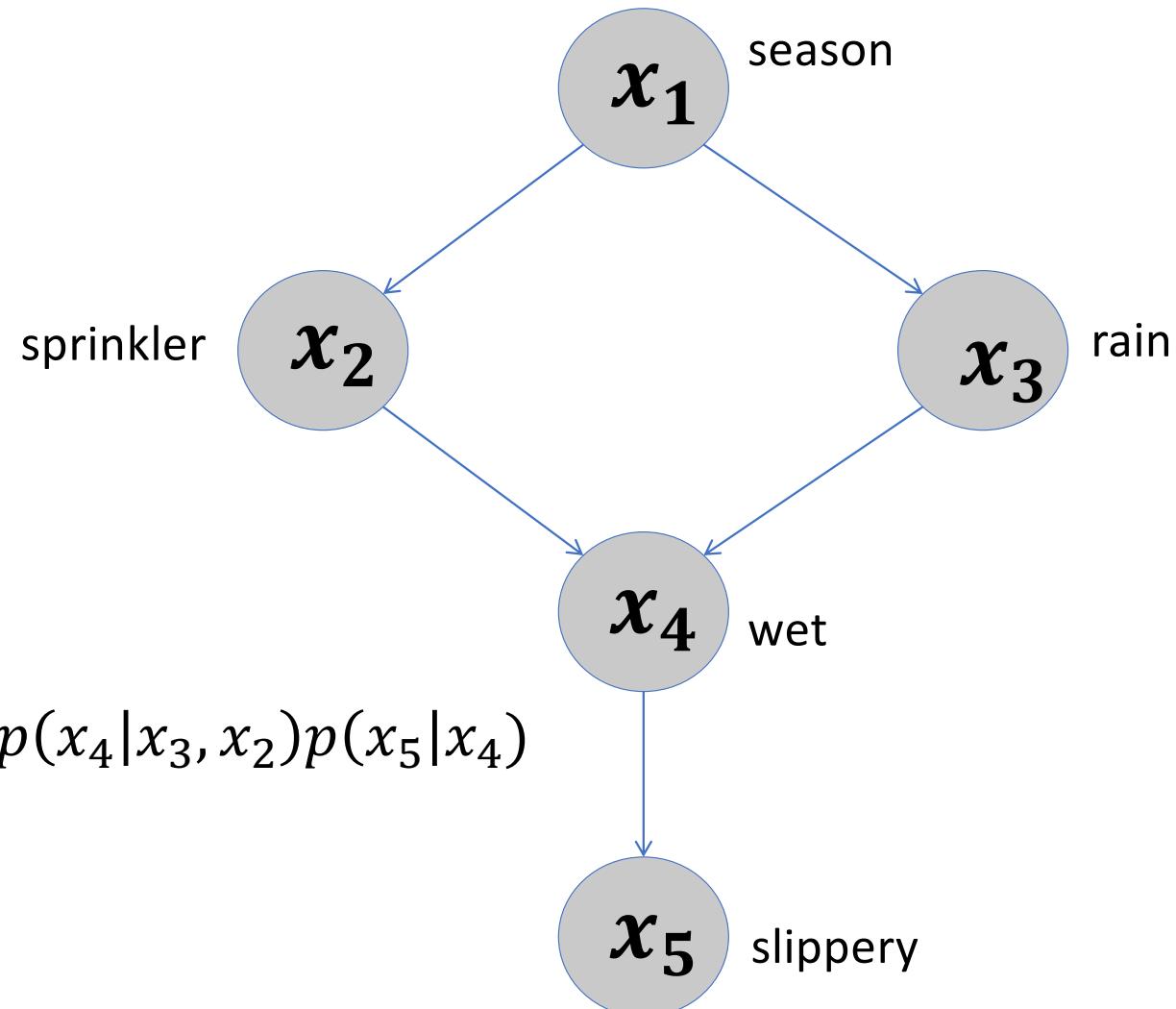
Directed acyclic graph: DAG

- Graphs with directed edges and no cycles
- Example: chains, trees
- More complex structures

Example (Pearl, 2000)

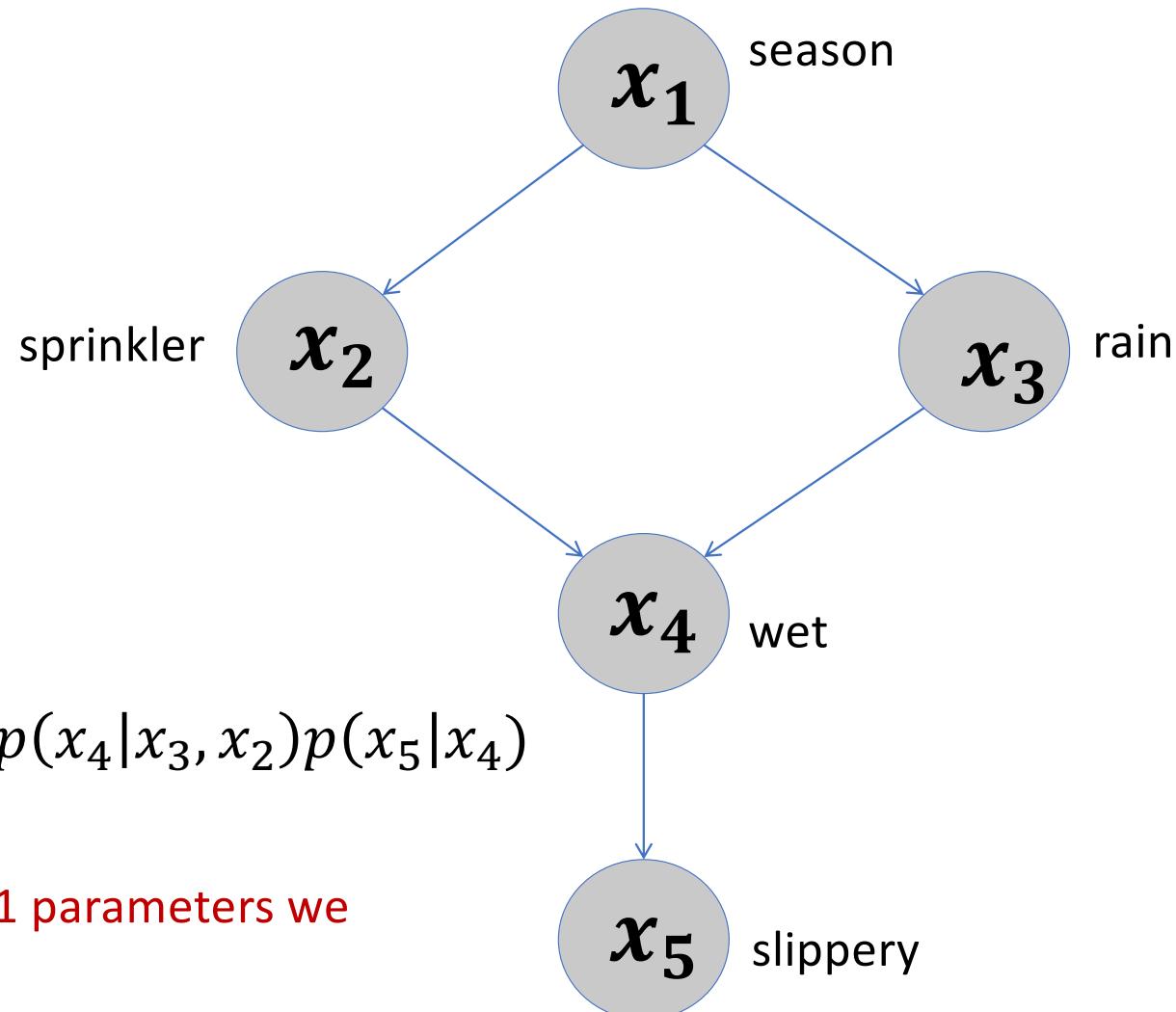


Example (Pearl, 2000)



$$\begin{aligned} p(x_1, x_2, x_3, x_4, x_5) &= \\ p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_3, x_2)p(x_5|x_4) \end{aligned}$$

Example (Pearl, 2000)



$$p(x_1, x_2, x_3, x_4, x_5) = \\ p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_3, x_2)p(x_5|x_4)$$

If all are binary, instead of 31 parameters we need only:

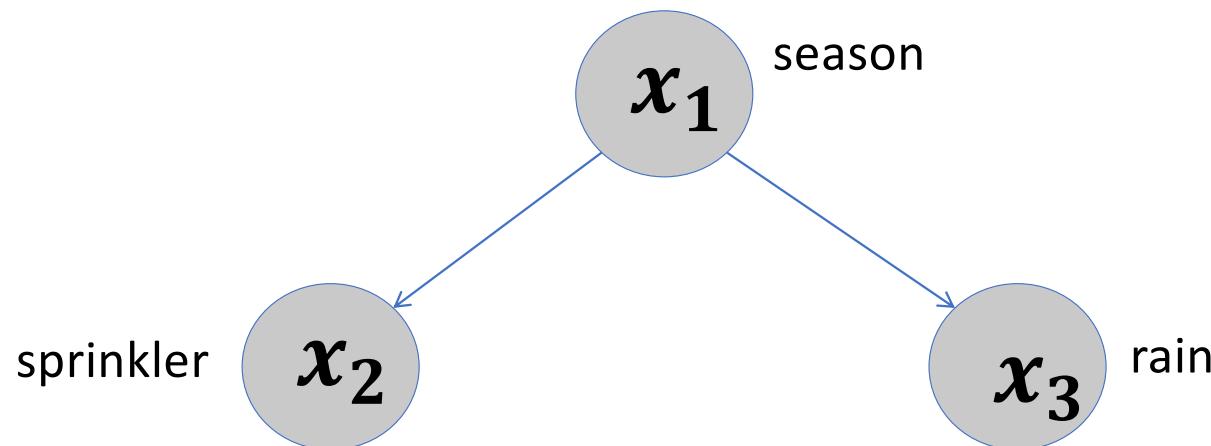
1 + 2 + 2 + 4 + 2 = 11 parameters to describe the distribution

Conditional independence

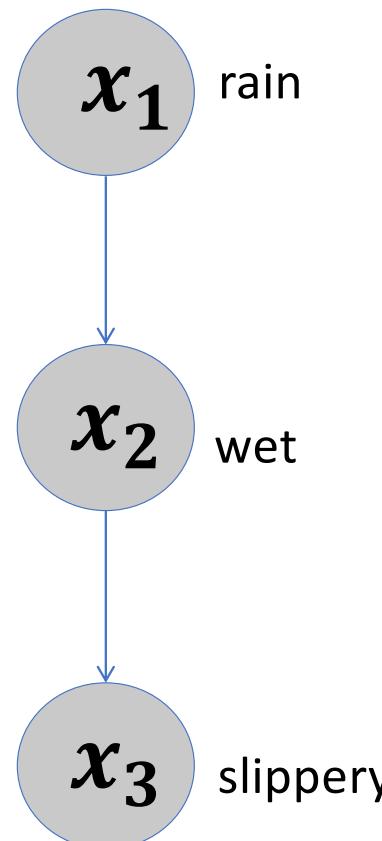
- $x_3 \perp\!\!\!\perp x_2 | x_1$
 $p(x_3, x_2 | x_1) = p(x_3 | x_1)p(x_2 | x_1)$

Conditional independence - fork

- $p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1)$
- $x_3 \perp\!\!\!\perp x_2 | x_1$
 $p(x_3, x_2|x_1) = p(x_3|x_1)p(x_2|x_1)$

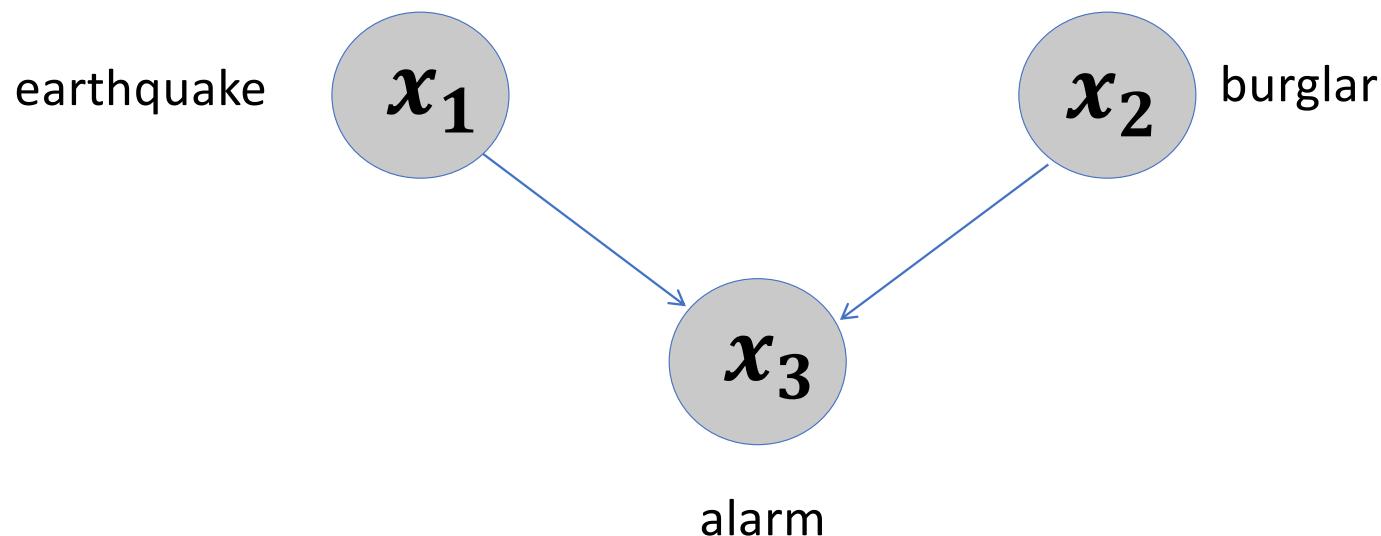


Conditional independence - chain

- $p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2)$
 - $x_3 \perp\!\!\!\perp x_1 | x_2$
 $p(x_3, x_1|x_2) = p(x_3|x_2)p(x_1|x_2)$
- 

Colliders

- $p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3|x_1, x_2)$
- $x_1 \perp\!\!\!\perp x_2$
- $x_1 \not\perp\!\!\!\perp x_2 | x_3$
- “explaining away”



Markovian parents

- Ordered set $\{x_1, x_2, \dots, x_d\}$
- A set of variables PA_j are the *Markovian Parents of x_j* if it is the minimal set of predecessors of x_j that renders x_j independent of all its predecessors

Markovian parents

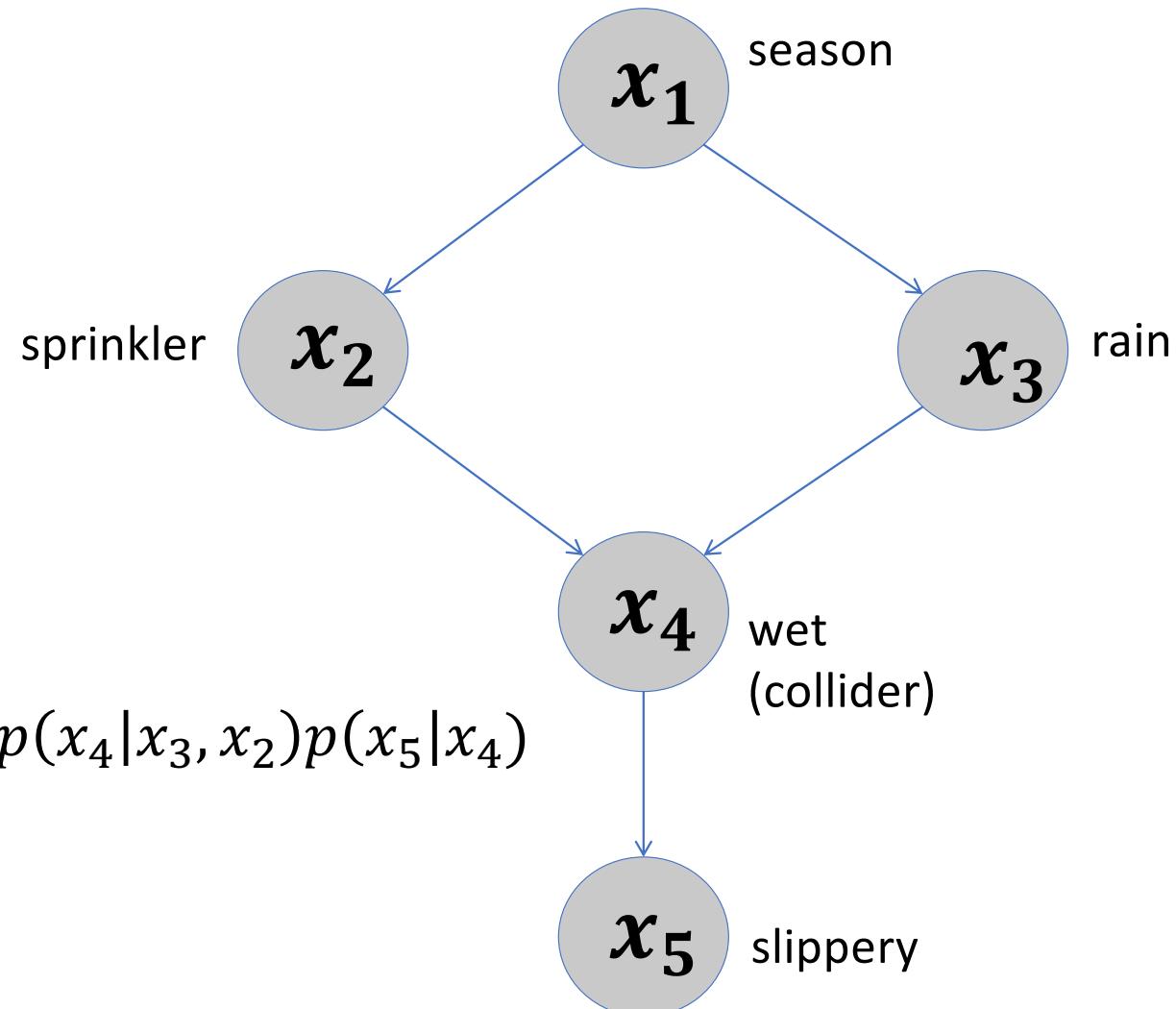
- Ordered set $\{x_1, x_2, \dots, x_d\}$
- A set of variables PA_j are the *Markovian Parents of x_j* if it is the minimal set of predecessors of x_j that renders x_j independent of all its predecessors
- Every distribution over $\{x_1, x_2, \dots, x_d\}$ can be factored as:

$$p(x_1, x_2, \dots, x_d) = \prod_{j=1}^d p(x_j | PA_j)$$

Parents and DAGs

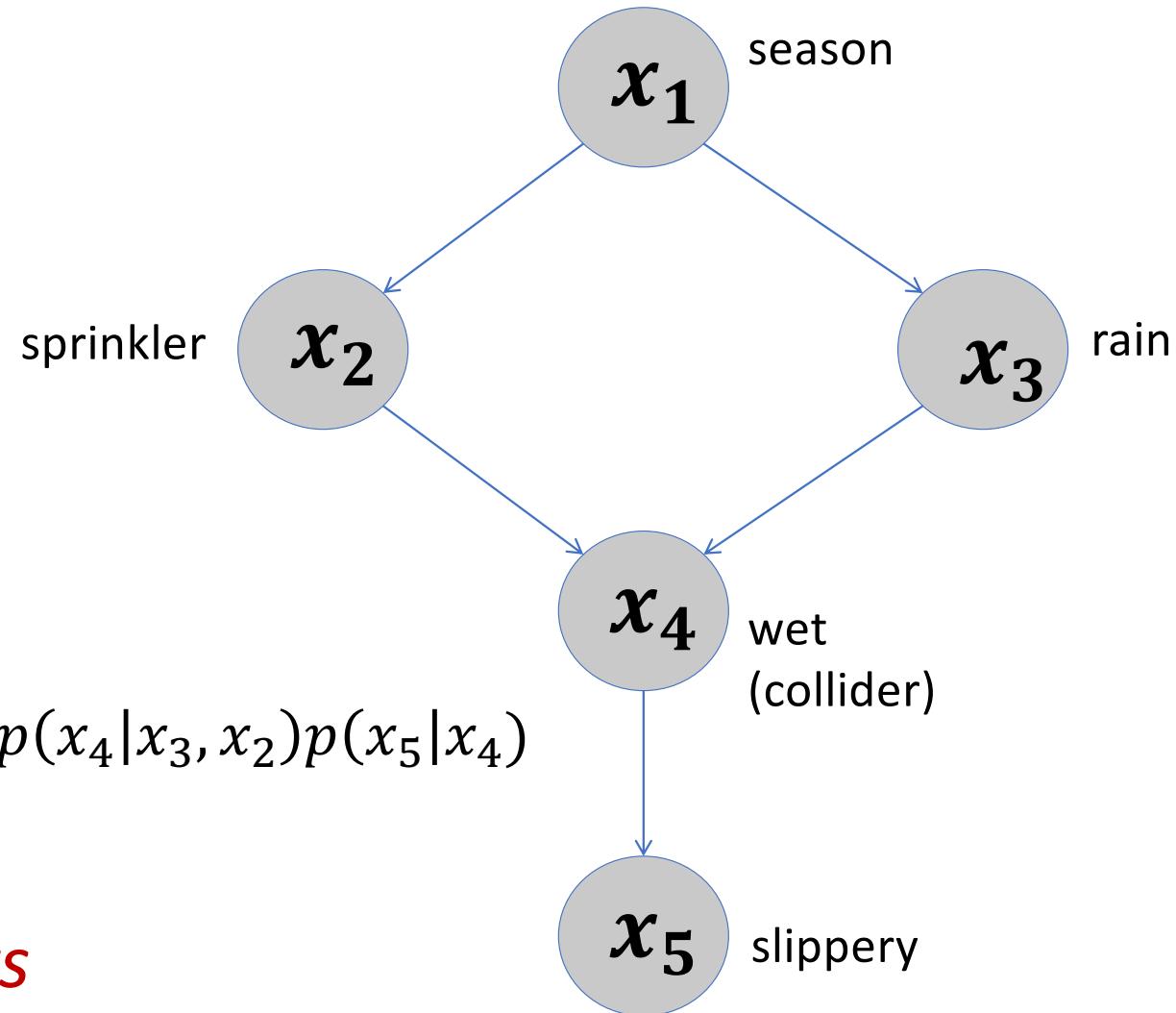
- Start from x_1, x_2
- Draw an arrow from x_1 to x_2 if they are dependent
- For x_3 , test if
 - $x_3 \perp\!\!\!\perp (x_1, x_2) : no\ arrow$
 - else
 - $x_3 \perp\!\!\!\perp x_2 | x_1 : arrow x_1 \rightarrow x_3$
 - $x_3 \perp\!\!\!\perp x_1 | x_2 : arrow x_2 \rightarrow x_3$
 - Else arrow from both x_1 and x_2 to x_3
- In general for every x_j find minimal set of predecessors that *screens them off* from all the rest
- Bayesian network is unique per ordering, assuming $p(x_1, x_2, \dots, x_d) > 0$ always (Pearl, 1988)

Example (Pearl, 2000)



$$\begin{aligned} p(x_1, x_2, x_3, x_4, x_5) &= \\ p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_3, x_2)p(x_5|x_4) \end{aligned}$$

Example (Pearl, 2000)



$$\begin{aligned} p(x_1, x_2, x_3, x_4, x_5) &= \\ p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_3, x_2)p(x_5|x_4) \end{aligned}$$

Compatibility

- Given a directed acyclic graph G with nodes $\{x_1, x_2, \dots, x_d\}$ and a probability distribution $p(x_1, x_2, \dots, x_d)$,
 p is **compatible** with G if there exists a factorization of p which agrees with the edges of G

d-separation and conditional independence

- Many conditional independence statements can be determined from graph structure alone

d-separation and conditional independence

Definition: d-separation

An undirected path p is blocked by a set of nodes Z if and only if

1. P contains a chain of nodes $A \rightarrow B \rightarrow C$ or a fork $A \leftarrow B \rightarrow C$ such that the middle node B is in Z (i.e., B is conditioned on), or
2. p contains a collider $A \rightarrow B \leftarrow C$ such that the collision node B is not in Z , and no descendant of B is in Z

d-separation and conditional independence

Definition: d-separation

An undirected path p is blocked by a set of nodes Z if and only if

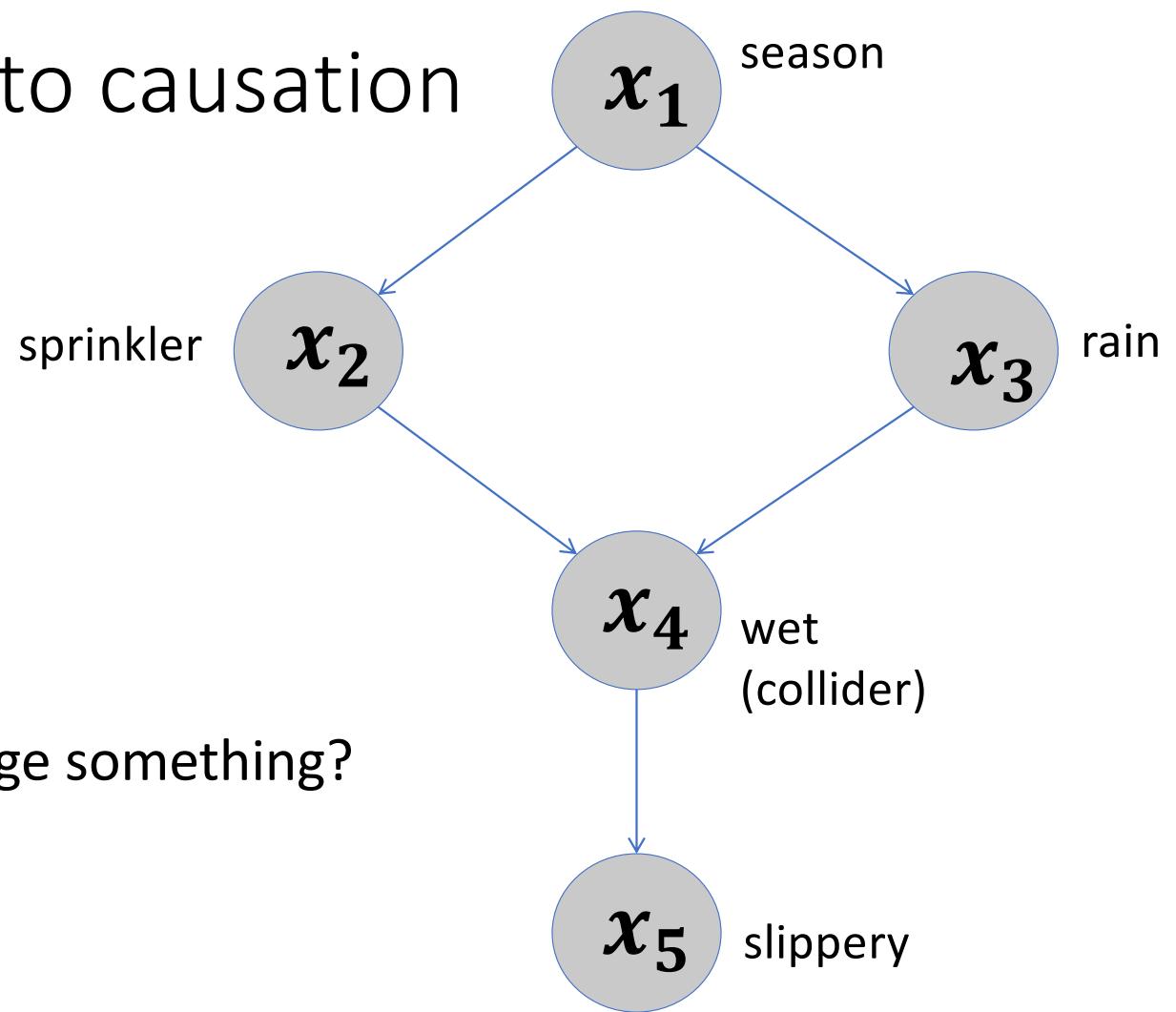
1. P contains a chain of nodes $A \rightarrow B \rightarrow C$ or a fork $A \leftarrow B \rightarrow C$ such that the middle node B is in Z (i.e., B is conditioned on), or
2. p contains a collider $A \rightarrow B \leftarrow C$ such that the collision node B is not in Z , and no descendant of B is in Z

If Z blocks every path between two nodes X and Y , then X and Y are d-separated, conditional on Z , and are independent conditional on Z .

d-separation and conditional independence

- d-separation is a graphical criterion for conditional independence which is valid in any distribution that is compatible with the graph
- Weak converse: if two sets of variables are *not* d-separated, then there exists at least one distribution compatible with the graph where conditional dependence does not hold
- Examples (board)

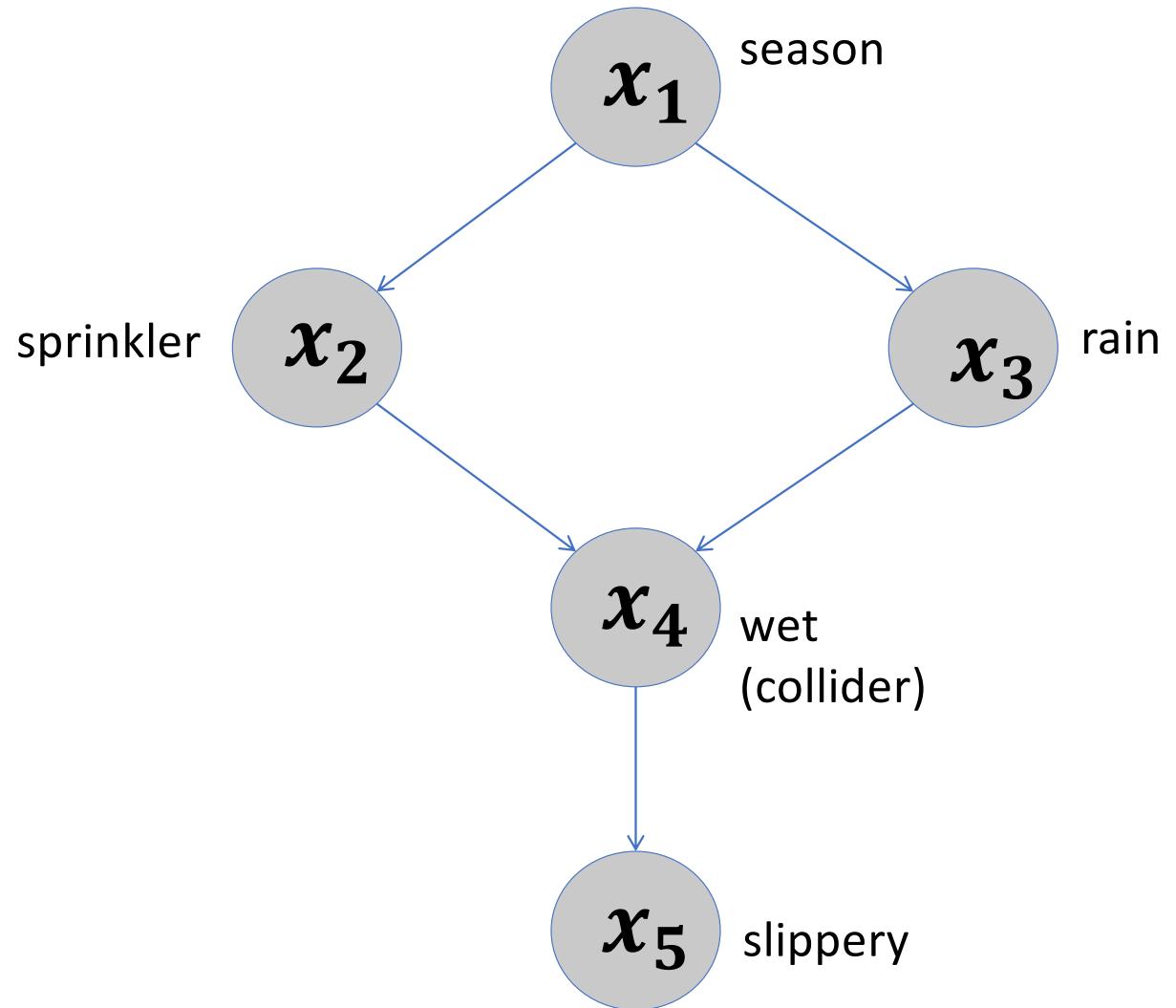
From association to causation



- Why do we accept that x_5 is independent of x_2 conditioned on x_4 ?
- What happens when we change something?

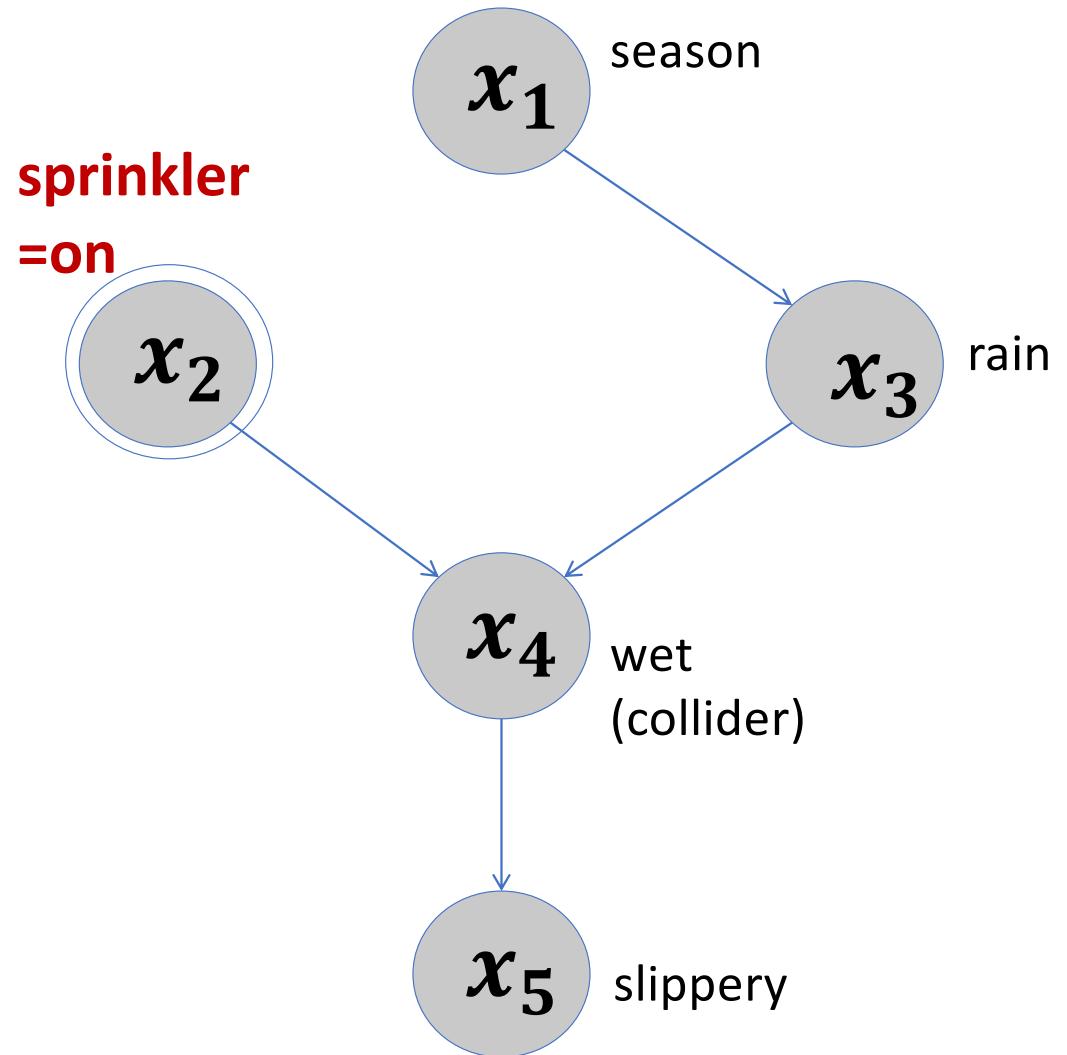
Modularity

- If each parent-child relationship in a graph dictates an autonomous mechanism, then we could turn them on or off without affecting the rest of the graph
- Turn the sprinkler on, please



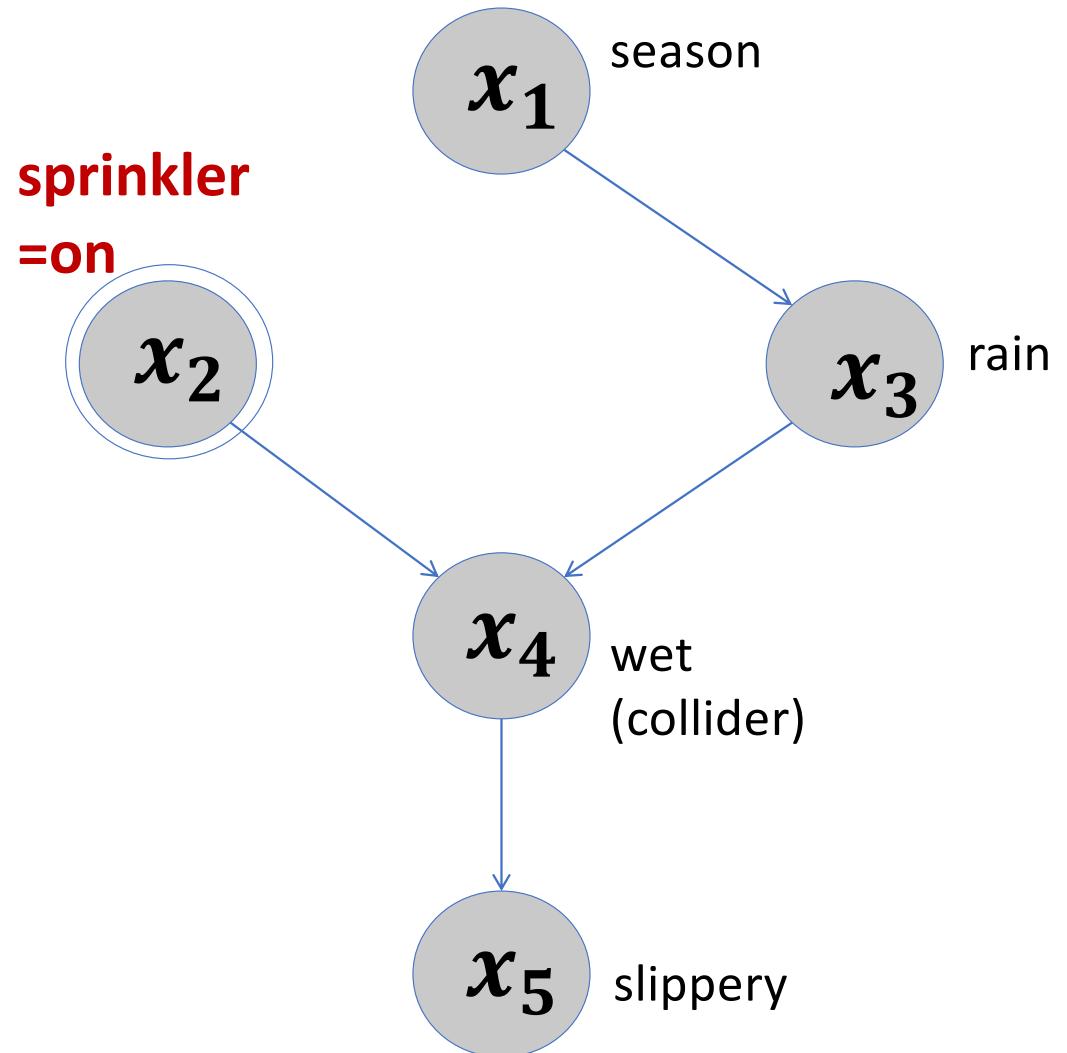
Modularity

- If each parent-child relationship in a graph dictates an autonomous mechanism, then we could turn them on or off without affecting the rest of the graph
- Turn the sprinkler on, please



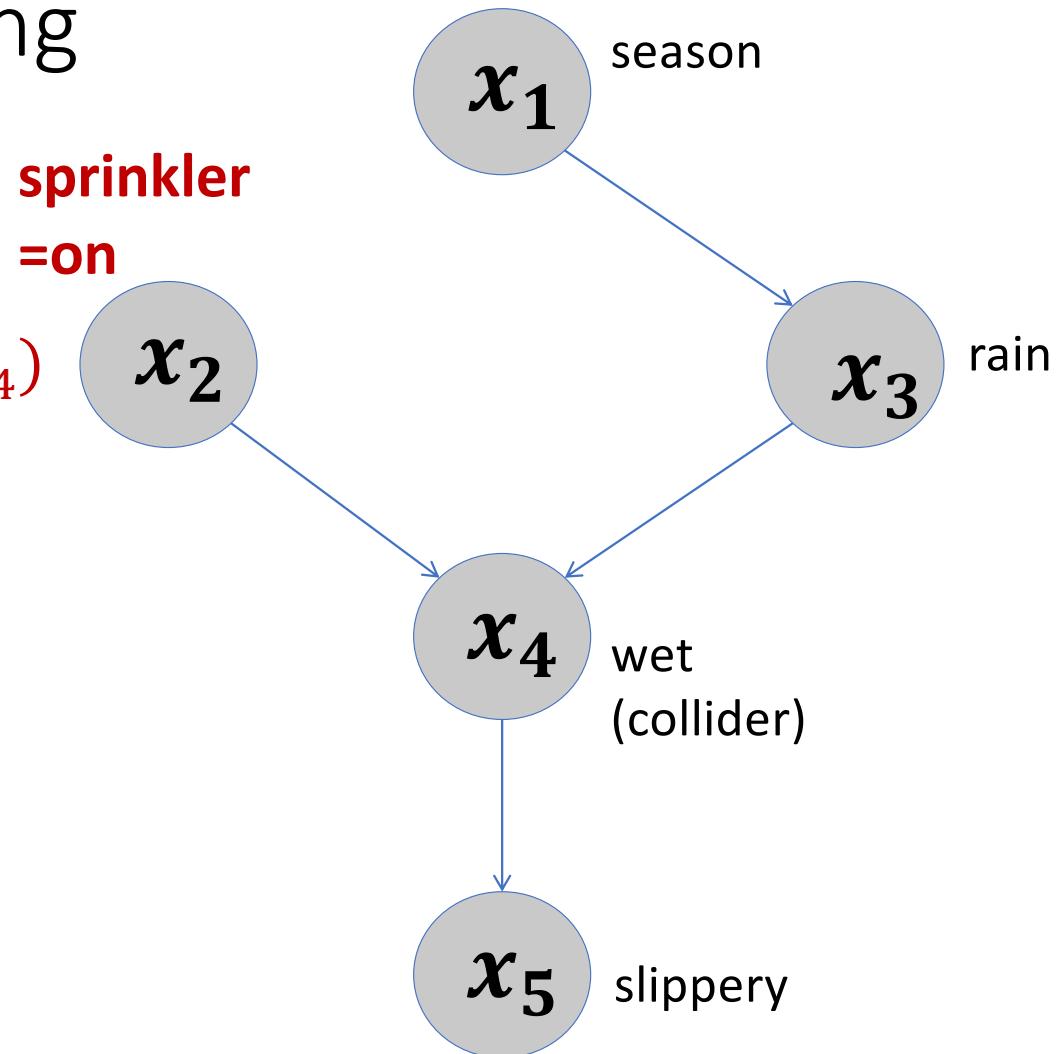
Modularity

- Turn the sprinkler on, please
- We removed the association between season and sprinkler
- We are now in a new world, where the sprinkler is set to on
- This is the *do*-operator



do-operator vs. conditioning

- $p_{do(x_2=on)}(x_1, x_3, x_4, x_5) = p(x_1)p(x_3|x_1)p(x_4|x_3, x_2 = on)p(x_5|x_4)$
- $p(x_1, x_3, x_4, x_5|x_2 = on) = p(x_1|x_2 = on)p(x_3|x_1, x_2 = on) \cdot p(x_4|x_3, x_2 = on)p(x_5|x_4, x_2 = on) =$

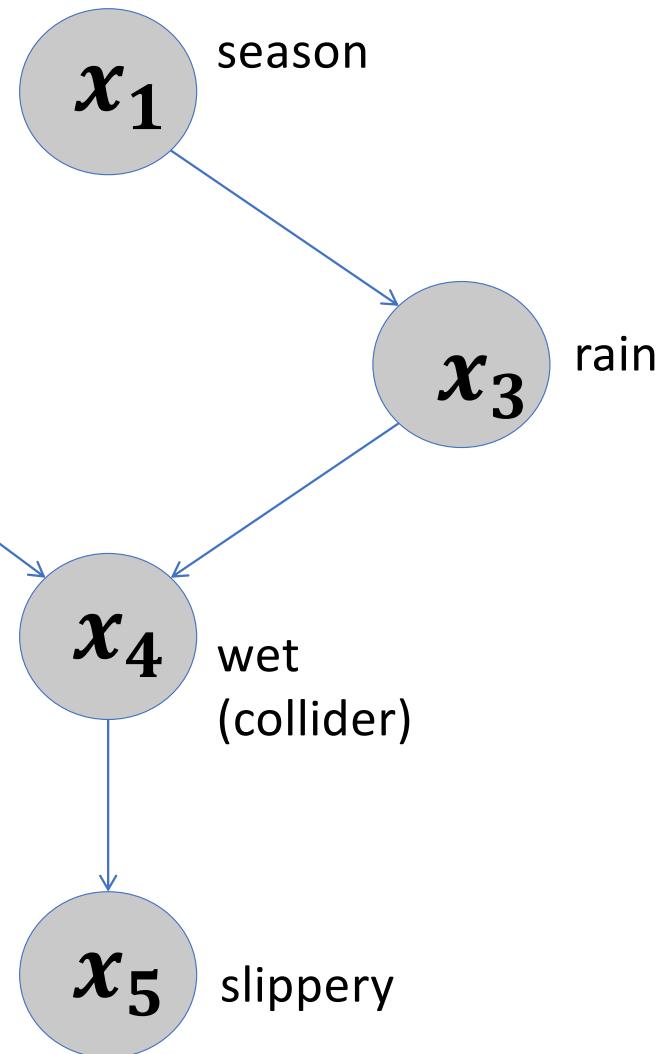


do-operator vs. conditioning

- $p_{do(x_2=on)}(x_1, x_3, x_4, x_5)$
 $= p(x_1)p(x_3|x_1)p(x_4|x_3, x_2 = on)p(x_5|x_4)$

sprinkler
=on

- $p(x_1, x_3, x_4, x_5|x_2 = on)$
 $= p(x_1|x_2 = on)p(x_3|x_1, x_2 = on)$
 $\cdot p(x_4|x_3, x_2 = on)p(x_5|x_4, x_2 = on)$
 $= p(x_1|x_2 = on)p(x_3|x_1)p(x_4|x_3, x_2 = on)p(x_5|x_4)$



notation

- $X = \{x_1, x_2, \dots, x_d\}$
- $p_{do(V=v)}(X \setminus V) \equiv p(X \setminus V | do(V = v))$
- $p_{do(x_1=x')} (x_2, \dots, x_d) \equiv p(x_2, \dots, x_d | do(x_1 = x'))$

Causal graphs

- Let $p(x_1, x_2, \dots, x_d)$ be a distribution
- $p_{do(V=v)}(x_1, x_2, \dots, x_d)$ sets a subset $V \subset \{x_1, x_2, \dots, x_d\}$ to constants v and is called an *interventional distribution*
- Denote P_* all interventional distributions derived from p , including $V = \emptyset$
- A DAG G is a *causal graph* for P_* if and only if for every $p_{do(V=v)} \in P_*$:
 1. $p_{do(V=v)}$ is compatible with G
 2. $p_{do(V=v)}(U = u) = 1$ for all $U \subset V$ whenever u is consistent with $V = v$
 3. $p_{do(V=v)}(u_i | pa_i) = p(u_i | pa_i)$ for all $u_i \notin V$ whenever pa_i is consistent with $V = v$

Truncated factorization

- $p(x_1, x_2, \dots, x_d) = \prod_{j=1}^d p(x_j | PA_j)$
- $p(x_1, x_2, \dots, x_d | do(V = v_0)) = \prod_{j|x_j \notin V} p(x_j | PA_j)_{|V=v_0}$
- Product only over variables not in V
- Conditioning on all the parents of x_j while setting their values to v_0 if they are in V
- Exercise: show that this is a valid distribution, i.e. sums to 1

Causal graphs

- Proposition 1 (setting all the parents to certain values)

$$p(u_i | pa_i = v) = p_{do(pa_i=v)}(u_i)$$

- Proposition 2 (only direct causes)

For all i and every subset S disjoint of $\{pa_i, x_i\}$

$$p_{do(pa_i=v, S=s)}(v_i) = p_{do(pa_i=v)}(v_i)$$

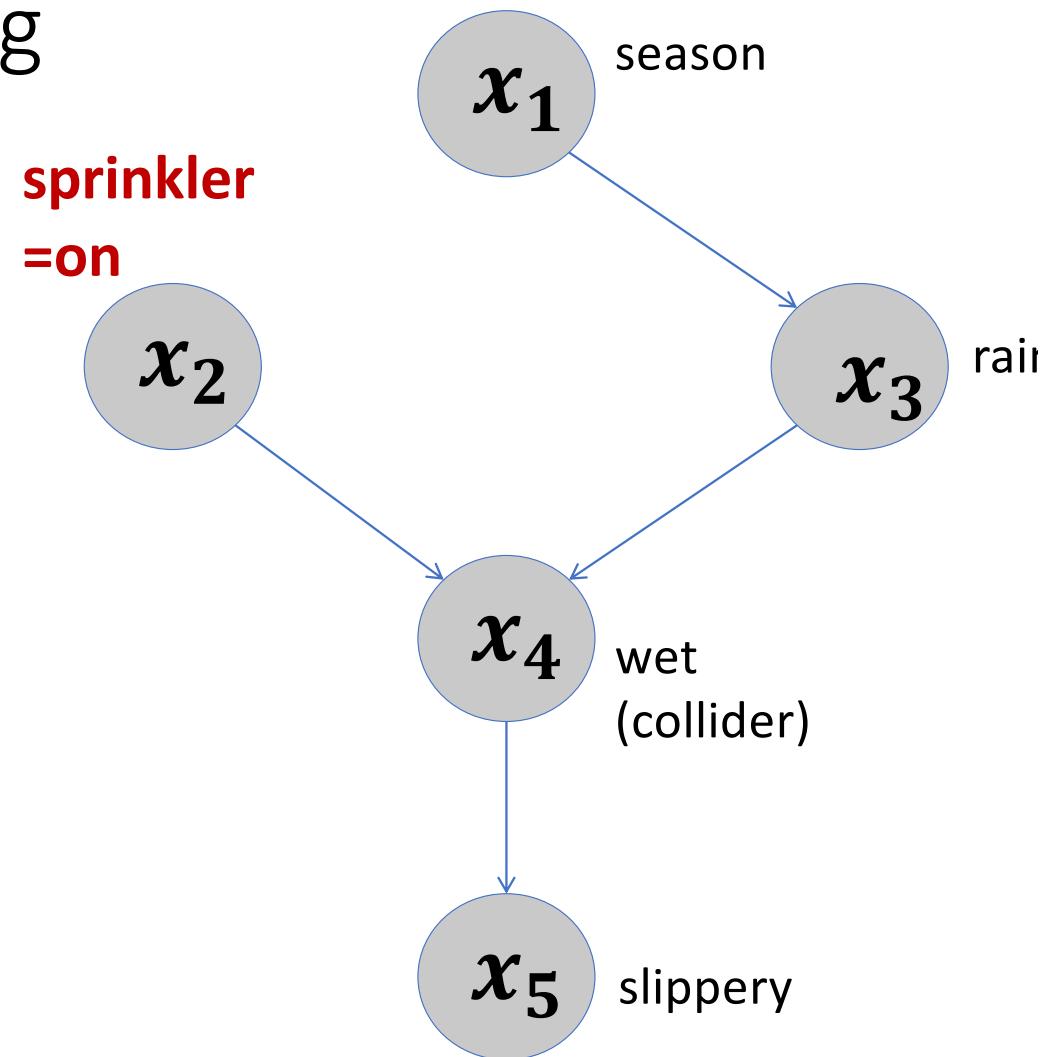
- Proposition 3 (adjustment for parents of cause)

For variable T with parents $pa_T = Z$

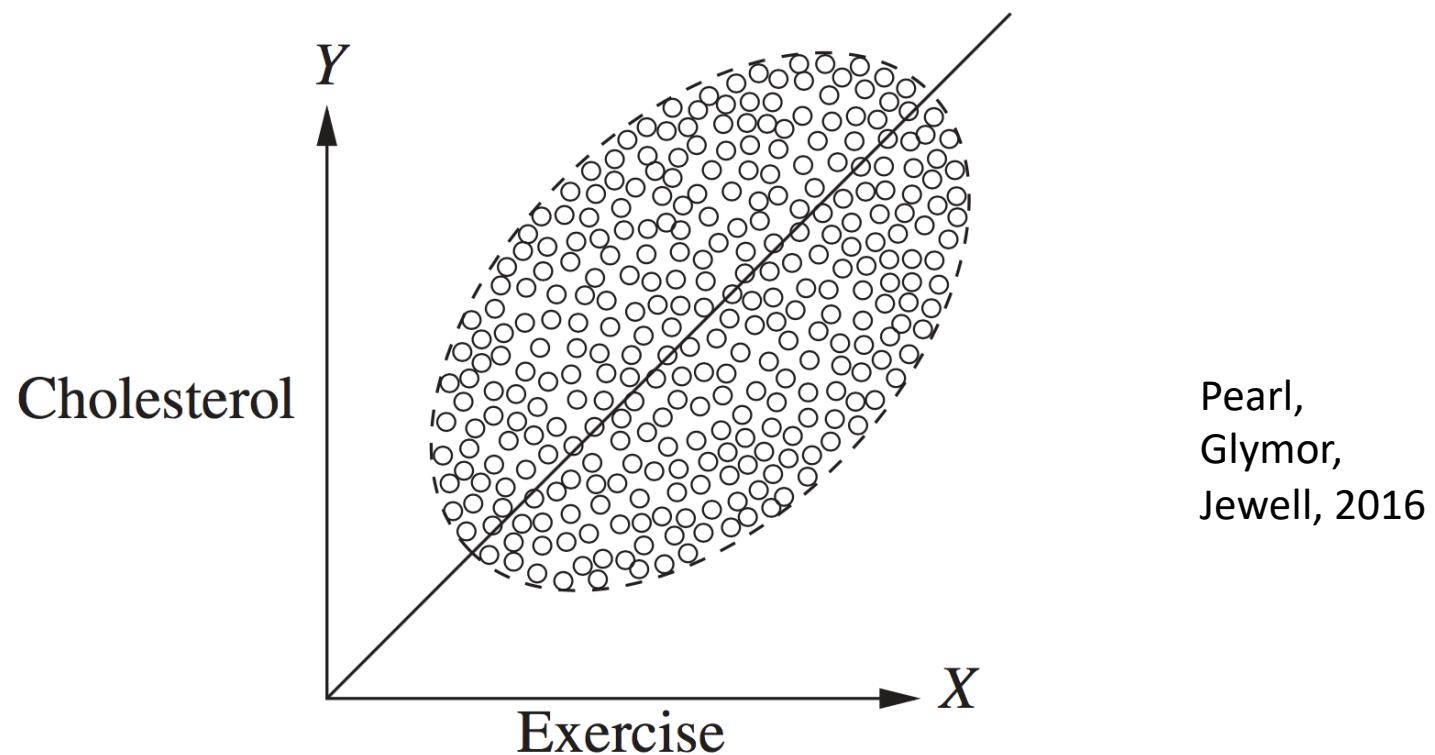
$$p_{do(T=t)}(y) = \sum_z p(y|T=t, Z=z)p(Z=z)$$

do-operator vs. conditioning

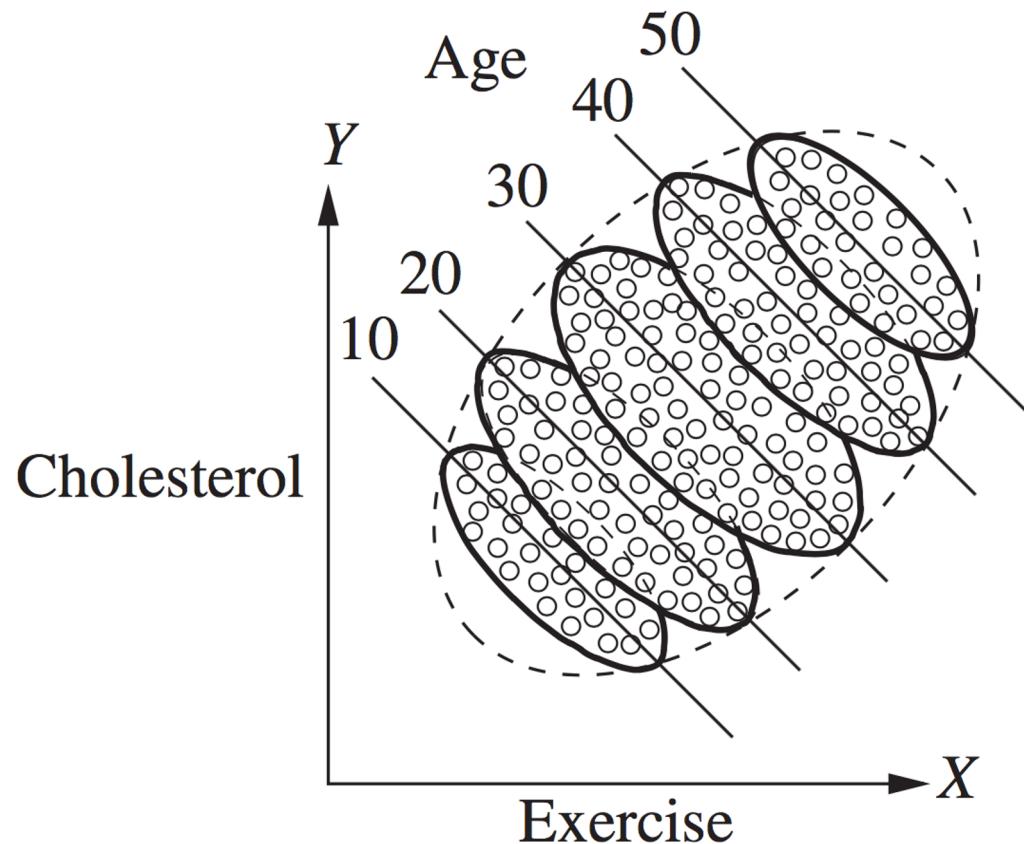
- $p(x_1, x_3, x_4, x_5 | do(x_2) = on)$
distribution under an **action**
- $p(x_1, x_3, x_4, x_5 | x_2 = on)$
distribution given **evidence**



Causal identifiability



Causal identifiability



Pearl,
Glymour,
Jewell, 2016

Average Treatment Effect

(also called Average Causal Effect in the causal graph literature)

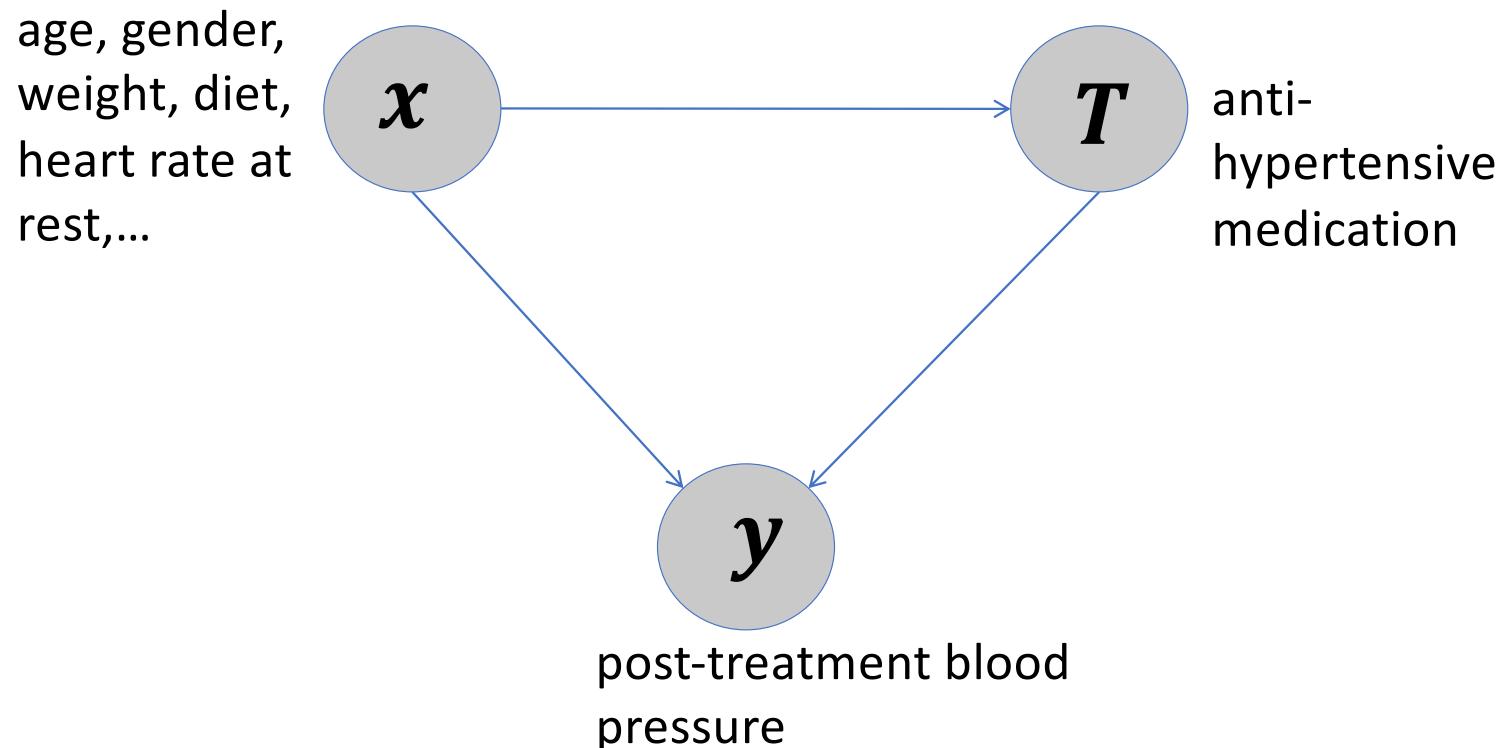
Effect of binary t on outcome y :

- $ATE = \mathbb{E}[y|do(T = 1)] - \mathbb{E}[y|do(T = 0)]$

(Sometimes we can't compute it)

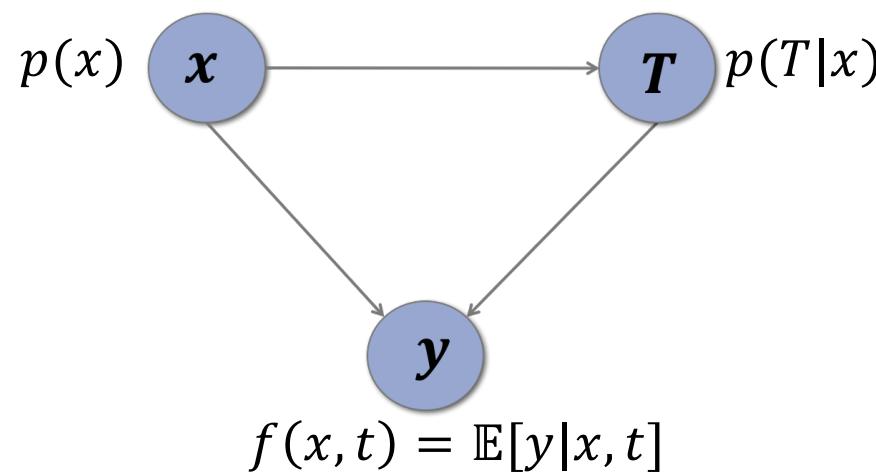
How to think about interventions: The *do* operator

(Pearl, 2009)



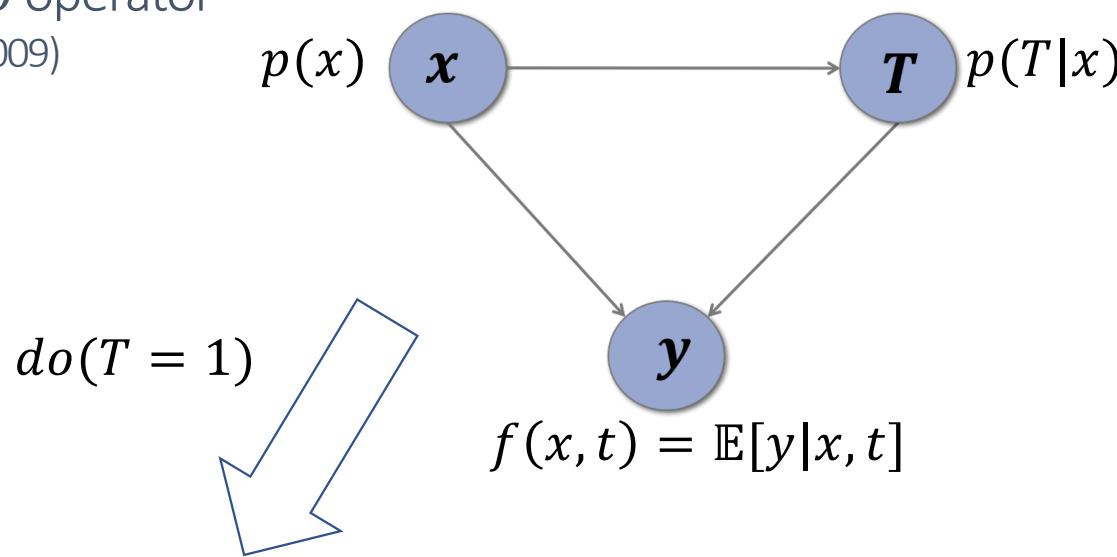
The *do* operator

(Pearl, 2009)



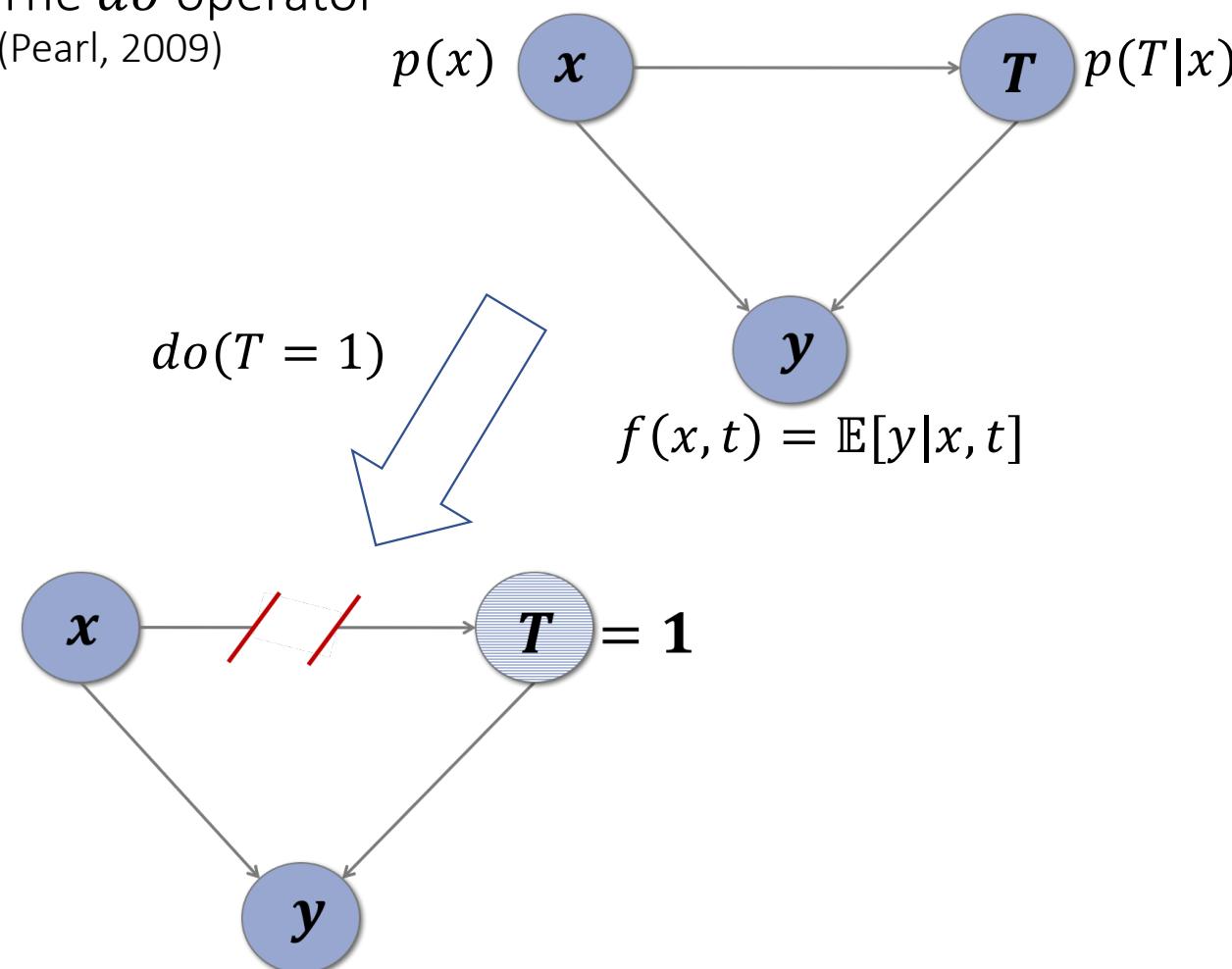
The *do* operator

(Pearl, 2009)



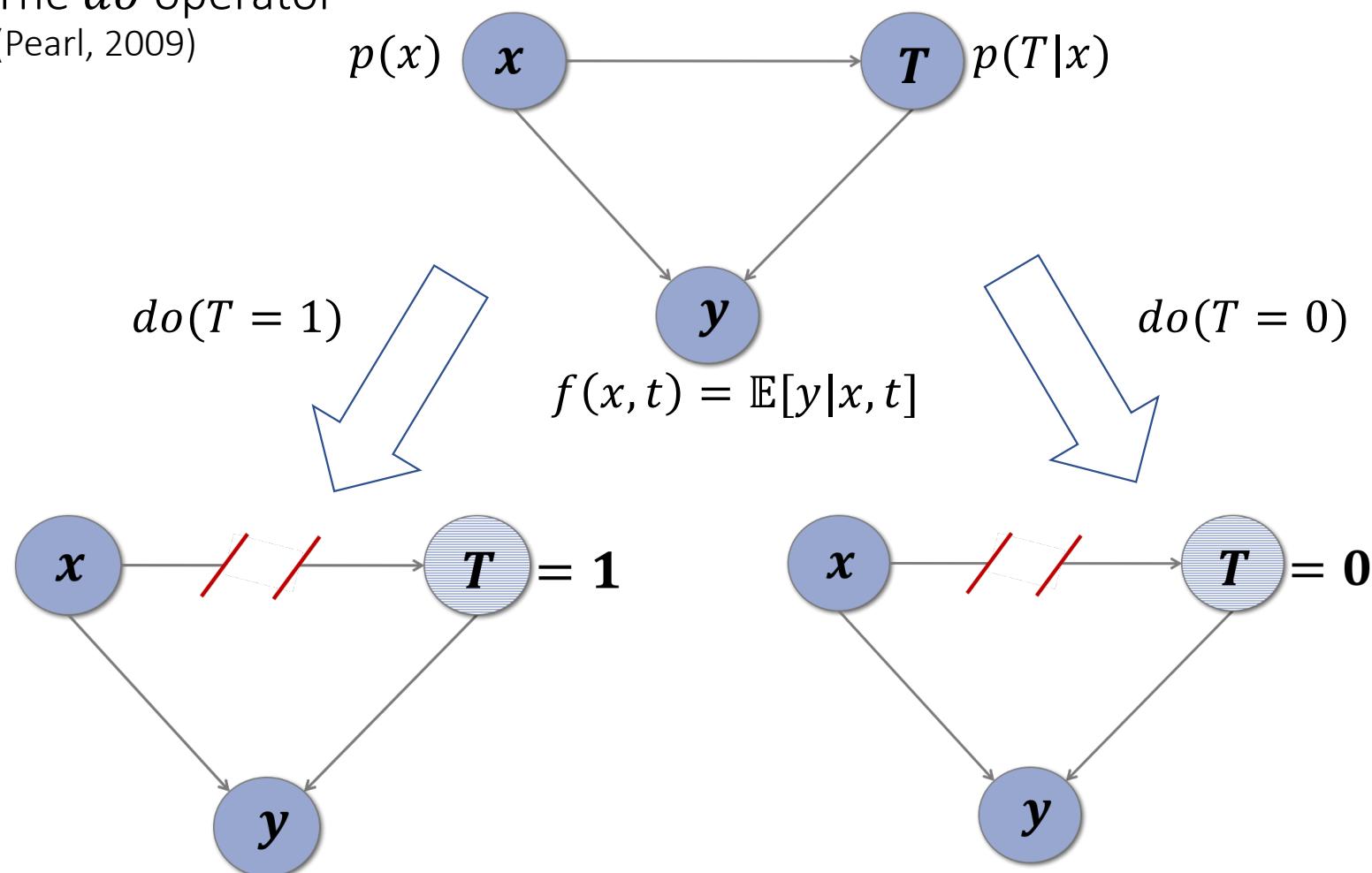
The *do* operator

(Pearl, 2009)



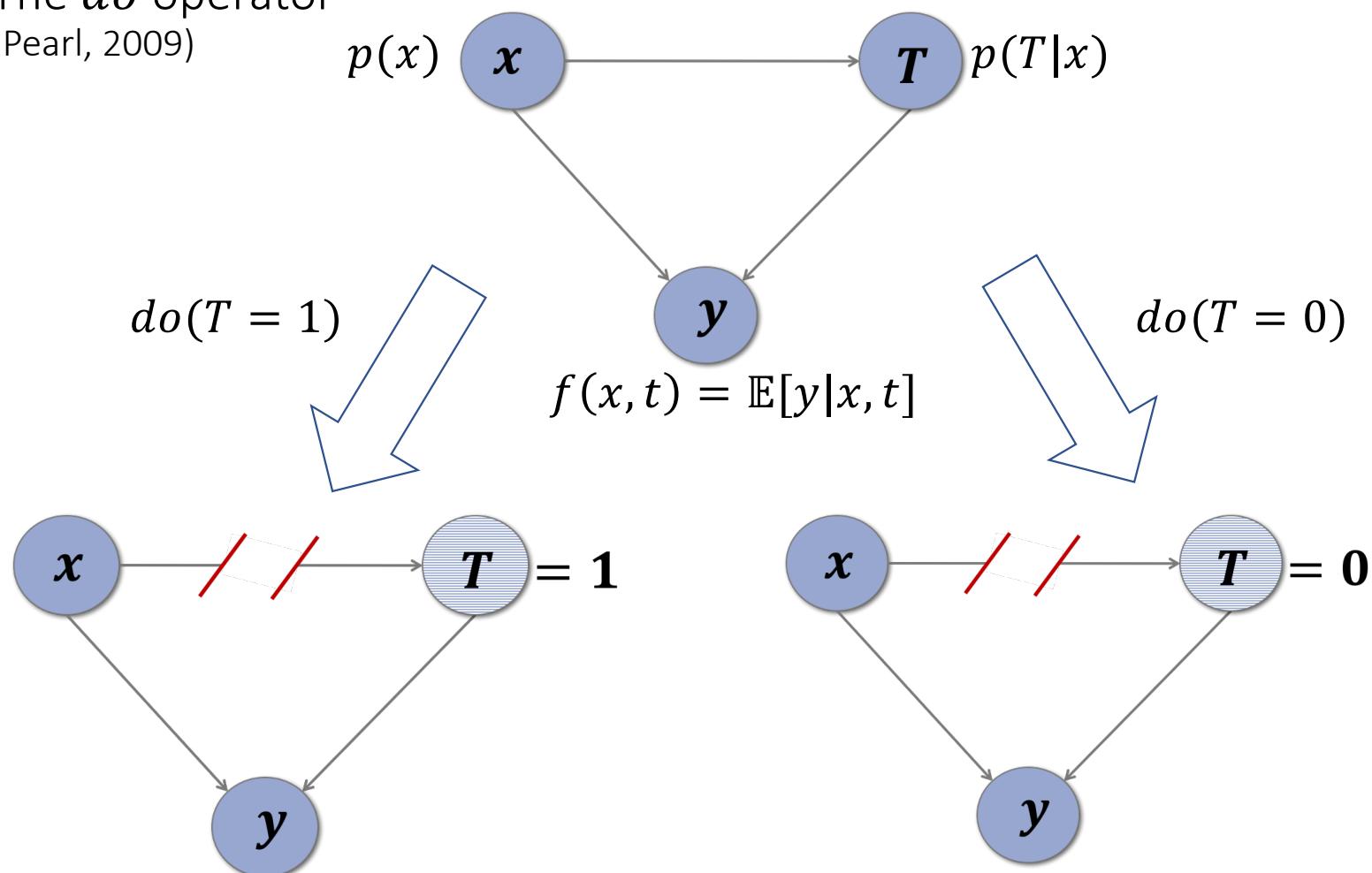
The do operator

(Pearl, 2009)



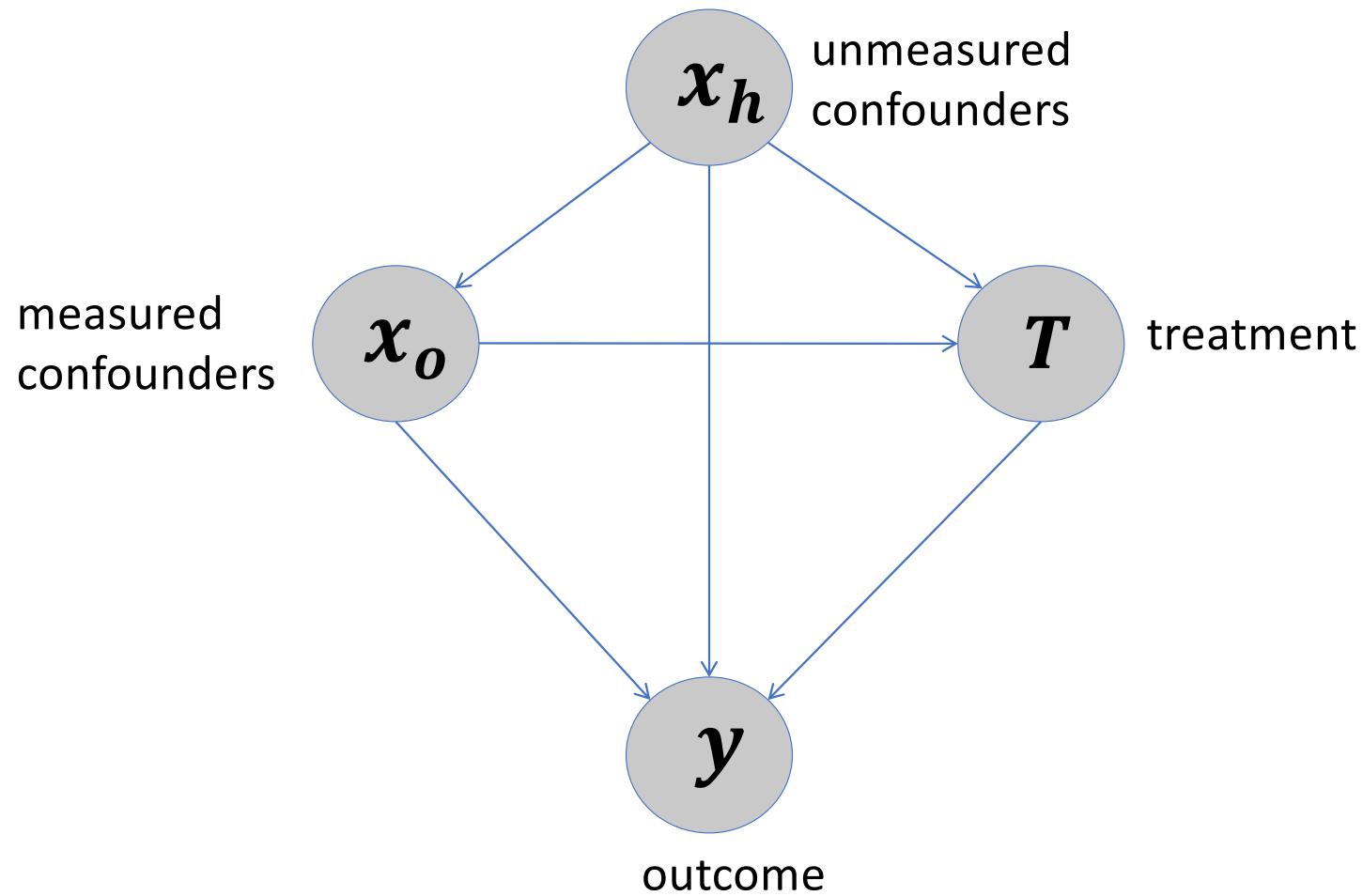
The do operator

(Pearl, 2009)



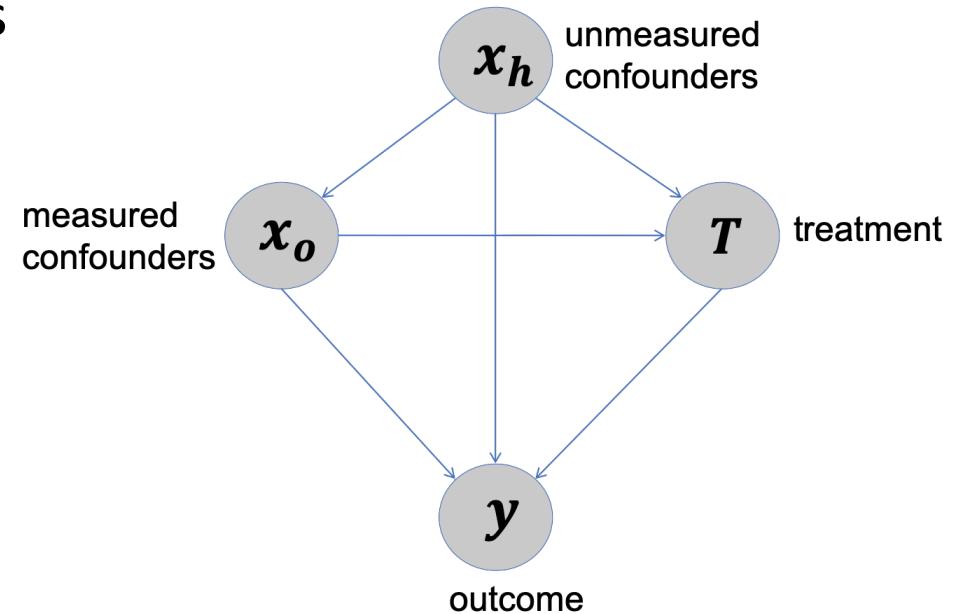
$$ATE := \mathbb{E}[y|do(T = 1)] - \mathbb{E}[y|do(T = 0)]$$

Confounding – causal version



Confounding – causal version

- Unlike potential outcomes, we only have y
- A confounder is something that affects both T and y
- Can be observed or unobserved



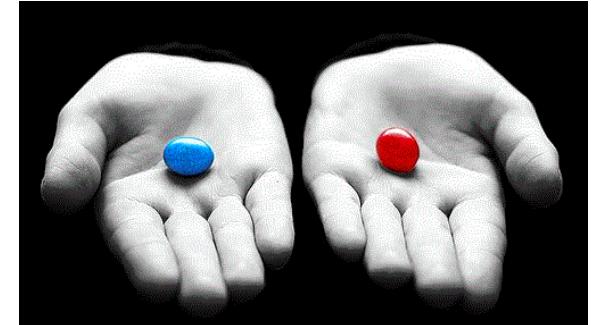
Causal identifiability

- A causal effect $p(y|do(\nu))$ is identifiable from a graph G if it can be computed uniquely from any positive probability distribution over the observable quantities which is compatible with G

Causal identifiability

- Can we infer $p(y|do(v))$ from some observed $p(y, v, x)$?
- If there are $p_1(y|do(v)) \neq p_2(y|do(v))$ that are both consistent with $p(y, v, x)$ then the answer is no
- How can we tell if $p(y|do(v))$ is uniquely determined by $p(y, v, x)$?
- Causal graphs give us many different sufficient conditions

Identifiability – an example

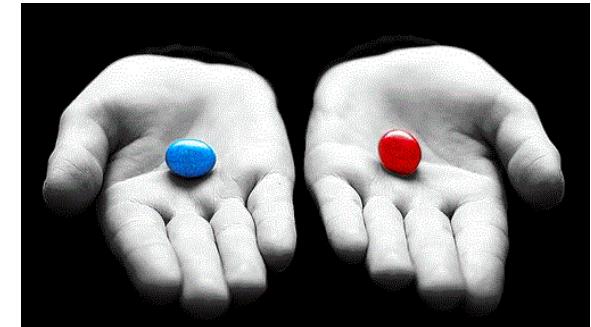


- Let $T \in \{0,1\}$, $Y \in \{0,1\}$
- Observed $p(T, Y)$:

$p(T, Y)$	$Y = 0$ (die)	$Y = 1$ (heal)
$T = 0$	0.5	0
$T = 1$	0	0.5

- $p(T = 0) = 0.5$
 $p(Y = t|T = t) = 1$

Identifiability – an example



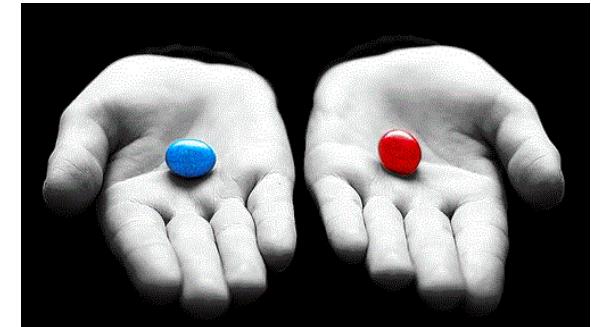
- Let $T \in \{0,1\}$, $Y \in \{0,1\}$
- Observed $p(T, Y)$:

$p(T, Y)$	$Y = 0$ (die)	$Y = 1$ (heal)
$T = 0$	0.5	0
$T = 1$	0	0.5

- If T is given randomly



Identifiability – an example



- Let $T \in \{0,1\}$, $Y \in \{0,1\}$
- Observed $p(T, Y)$:

$p(T, Y)$	$Y = 0$ (die)	$Y = 1$ (heal)
$T = 0$	0.5	0
$T = 1$	0	0.5

- If T is given randomly, $p(Y = 1|do(T = 1)) = ?$



Identifiability – an example

- Let $T \in \{0,1\}$, $Y \in \{0,1\}$

$p(T, Y)$	$Y = 0$ (die)	$Y = 1$ (heal)
$T = 0$	0.5	0
$T = 1$	0	0.5

- If T given by snake-oil salesman: see which patients will heal ($x = 1$), then gives them his Million-dollar-costing $T = 1$

- $x = 0 \rightarrow Y = 0$

$$x = 1 \rightarrow Y = 1$$

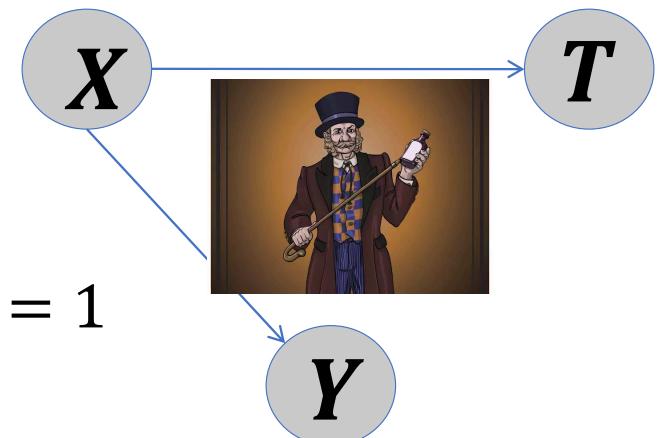
- True DGP:

$$p(x = 0) = 0.5$$

$$p(Y = x|X = x) = P(T = x|X = x) = 1$$

- $p(Y = 1|do(T = 1)) = ?$

- $p(Y = 1|do(T = 0)) = ?$

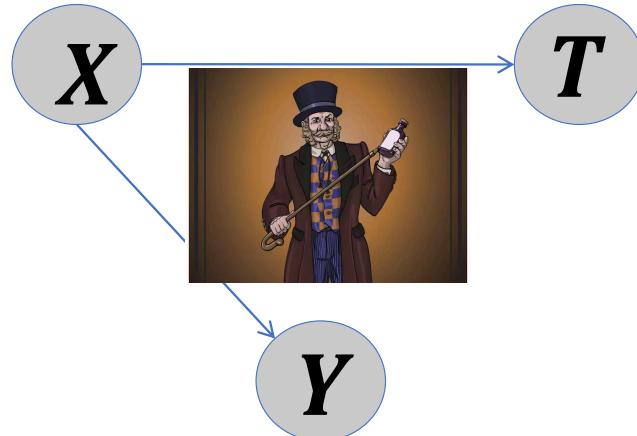


Identifiability – an example

- Let $T \in \{0,1\}$, $Y \in \{0,1\}$
- Observed $p(T, Y)$:

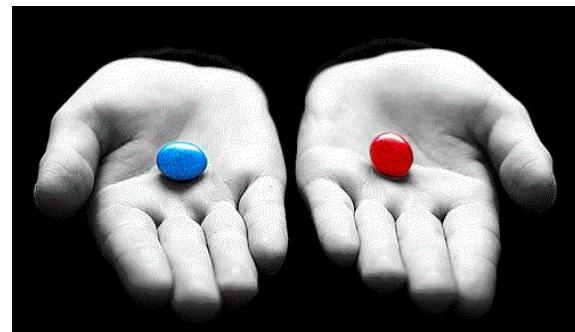
$p(T, Y)$	$Y = 0$ (die)	$Y = 1$ (heal)
$T = 0$	0.5	0
$T = 1$	0	0.5

- If T given by snake-oil salesman: see which patients will heal, then gives them $T = 1$



Identifiability

- Without knowing the causal graph, the same observable distribution can result from two very different causal processes
- Very different conclusions about which treatment we should use
- Causal graphs can give us sufficient conditions for when causal queries $p(y|do(v))$ are identifiable from an observed distribution
- Causal graphs encode extra knowledge!



Identifiability

- One way to think about it:

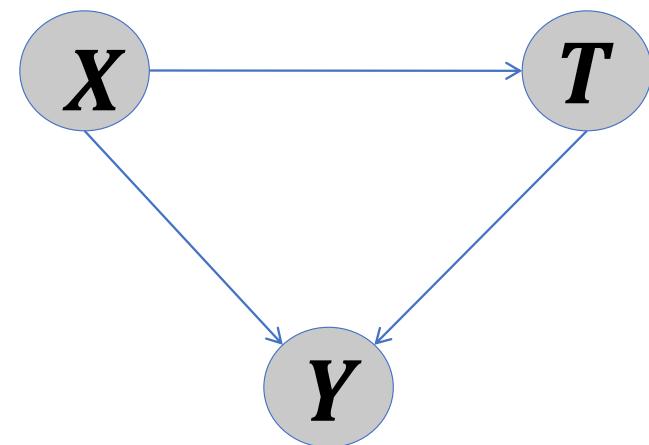
Observed correlation between T and Y
can occur both because of the direct path:

$$T \rightarrow Y$$

and the “backdoor” path:

$$T \leftarrow X \rightarrow Y$$

- If we measure X we can “block” the backdoor path



The Assumptions: causal identifiability

- A very useful sufficient condition for causal identifiability is the *back-door criterion*

Backdoor criterion

- Back-door criterion (Pearl, 1993, 2009):
The observed variables d-separate all paths between y and T that end with an arrow pointing to T
- Tells us what can we measure that will ensure causal identifiability

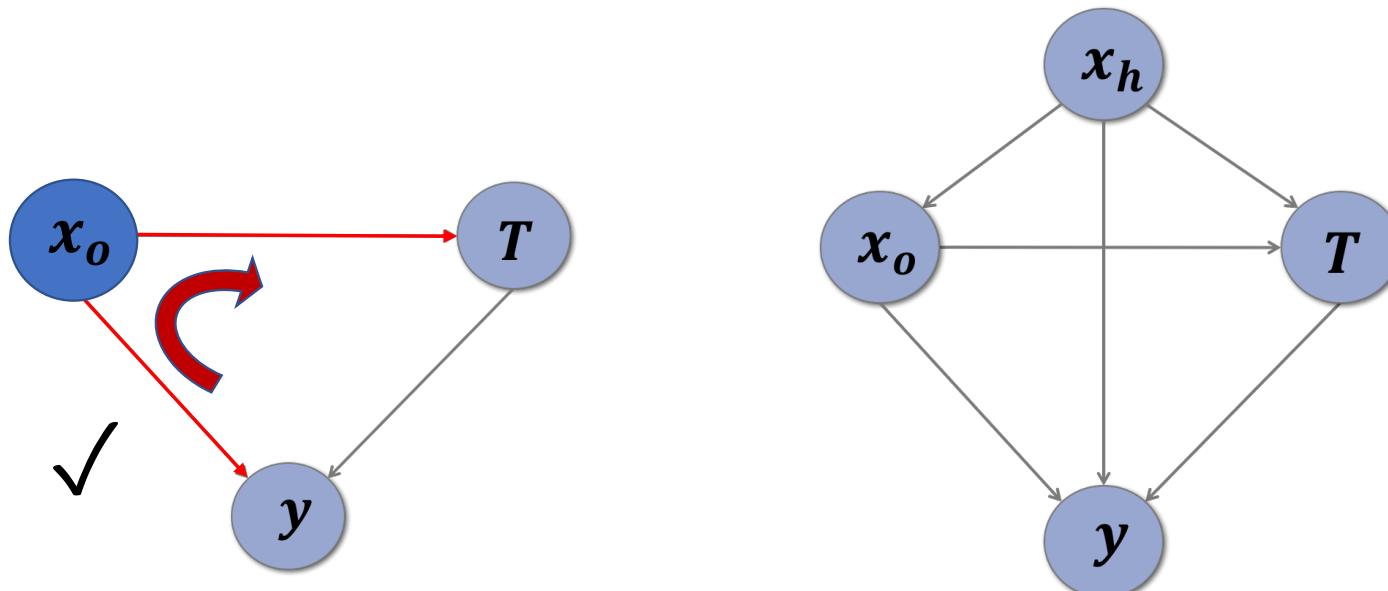
Back-door criterion: formal

A set of variables Z satisfies the *back-door criterion* relative to the ordered pair (T, Y) if:

1. No node in Z is a descendant of T ; and
2. Z blocks (in the d-separation sense) every path between T and Y that contains an arrow into T

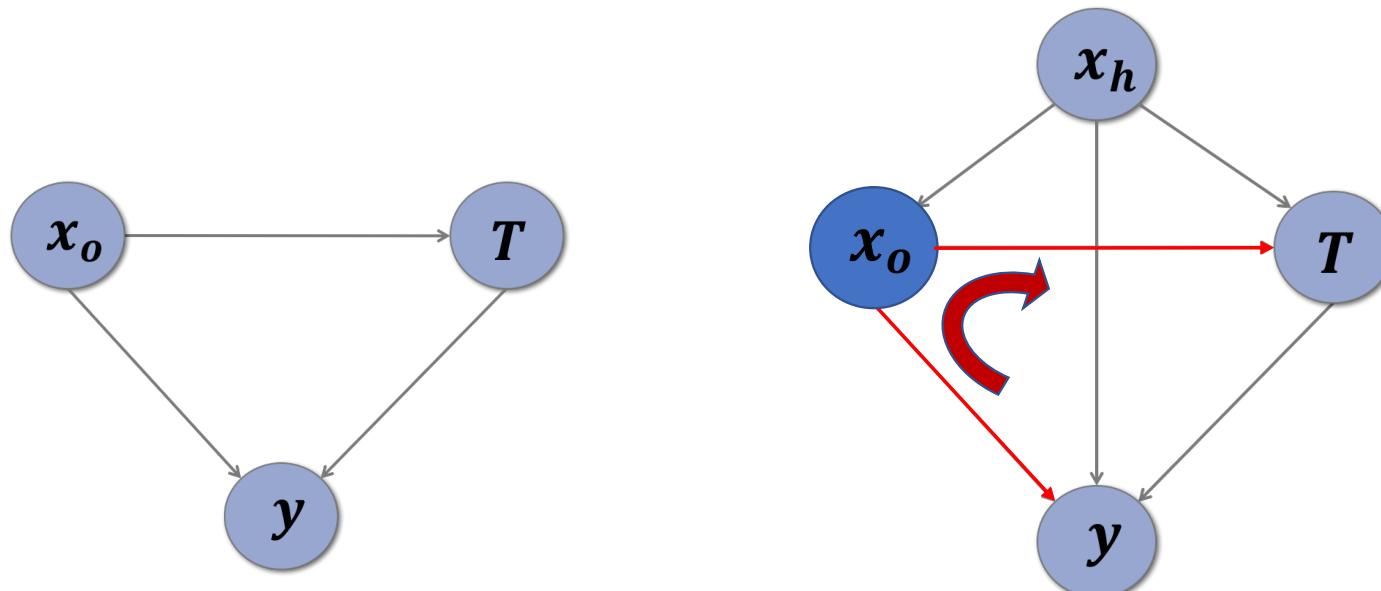
Back-door criterion

- Back-door criterion:
The observed variables d-separate all paths between y and T that end with an arrow pointing to T



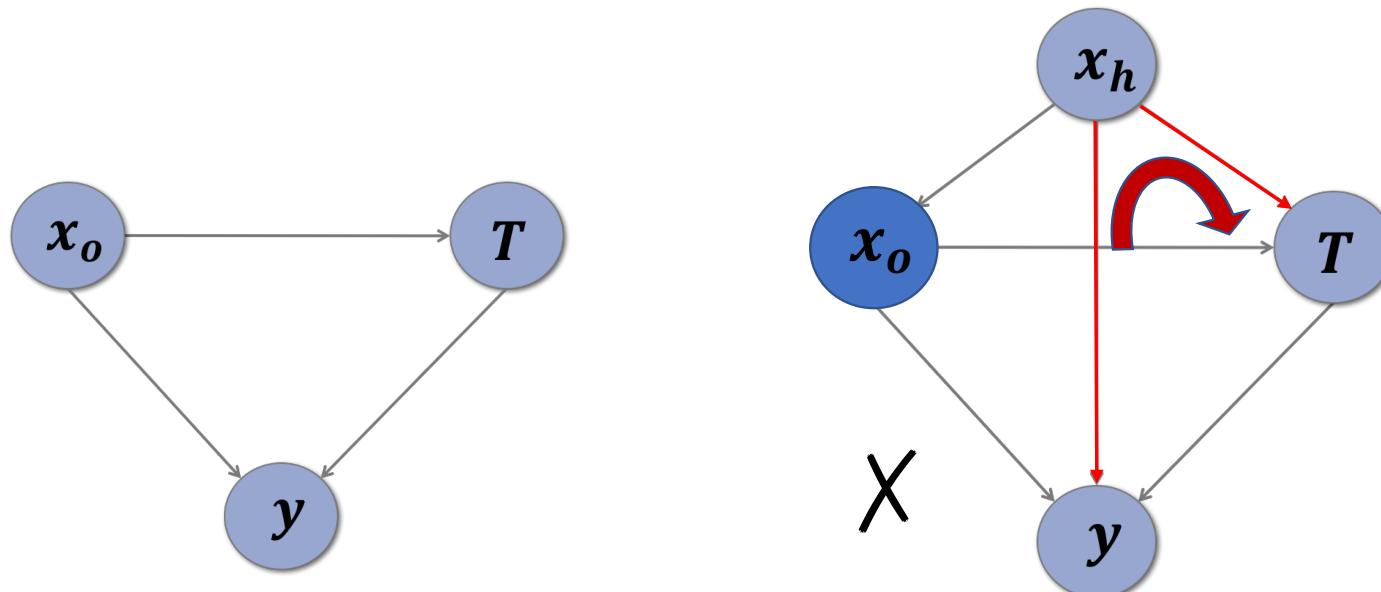
Back-door criterion

- Back-door criterion:
The observed variables d-separate all paths between y and T that end with an arrow pointing to T



Back-door criterion

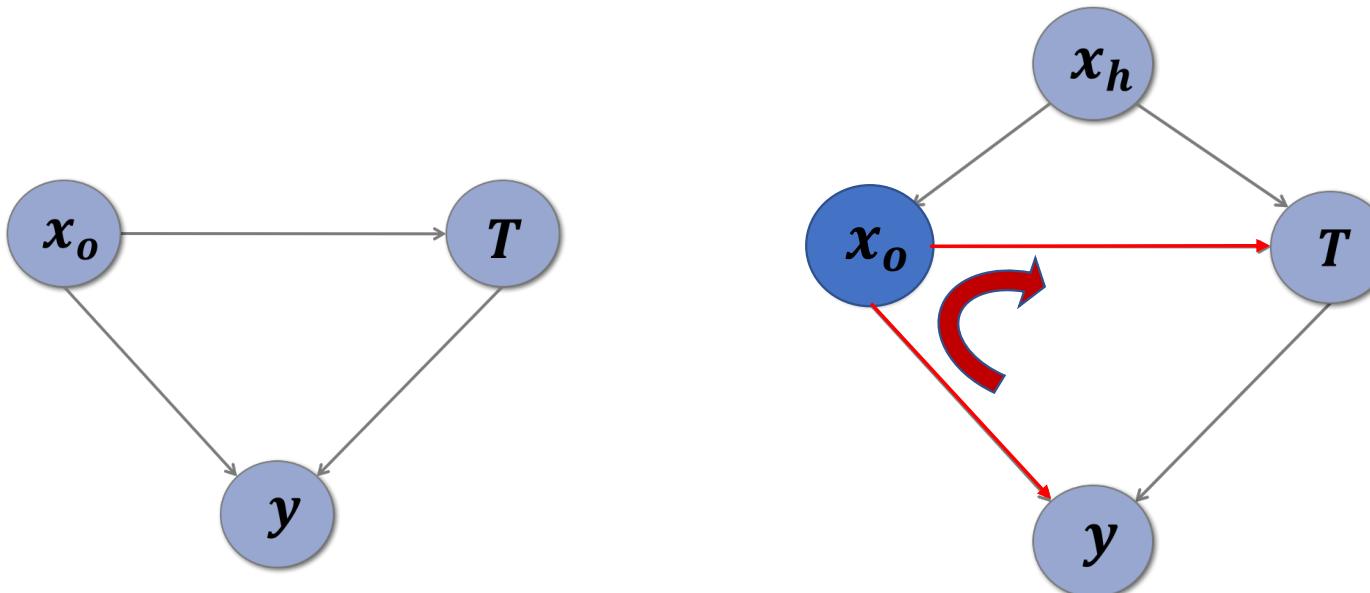
- Back-door criterion:
The observed variables d-separate all paths between y and T that end with an arrow pointing to T



Back-door criterion

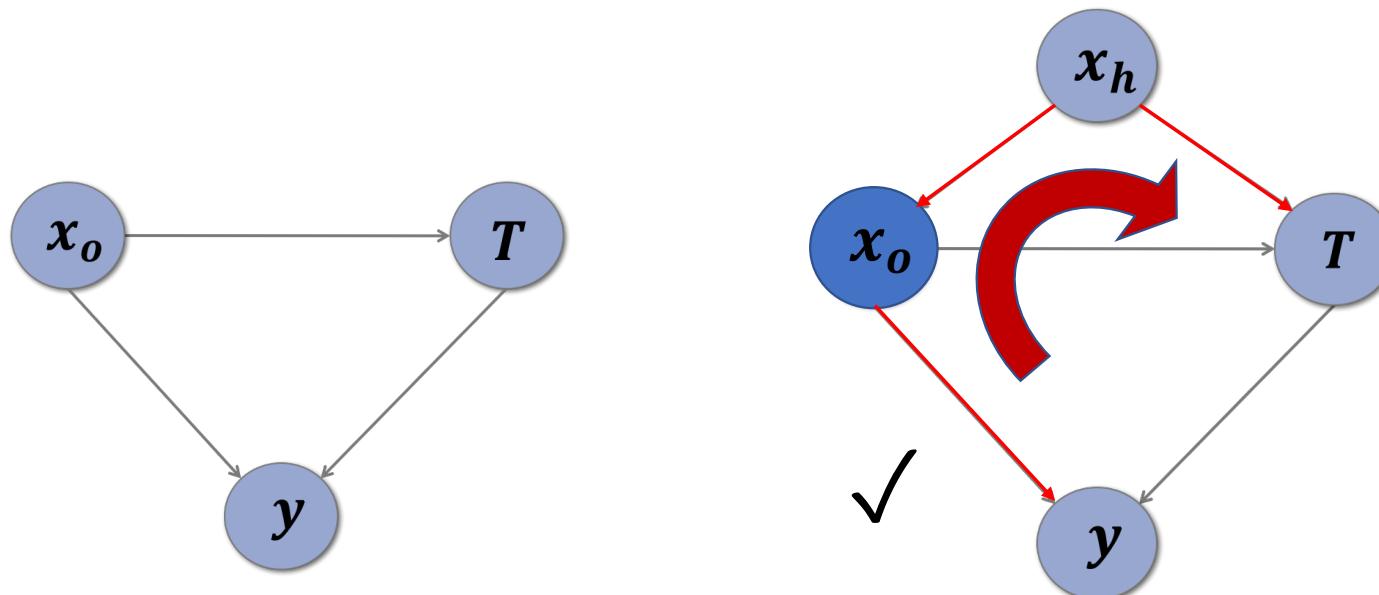
Definition:

The observed variables d-separate all paths between y and T that end with an arrow pointing to T

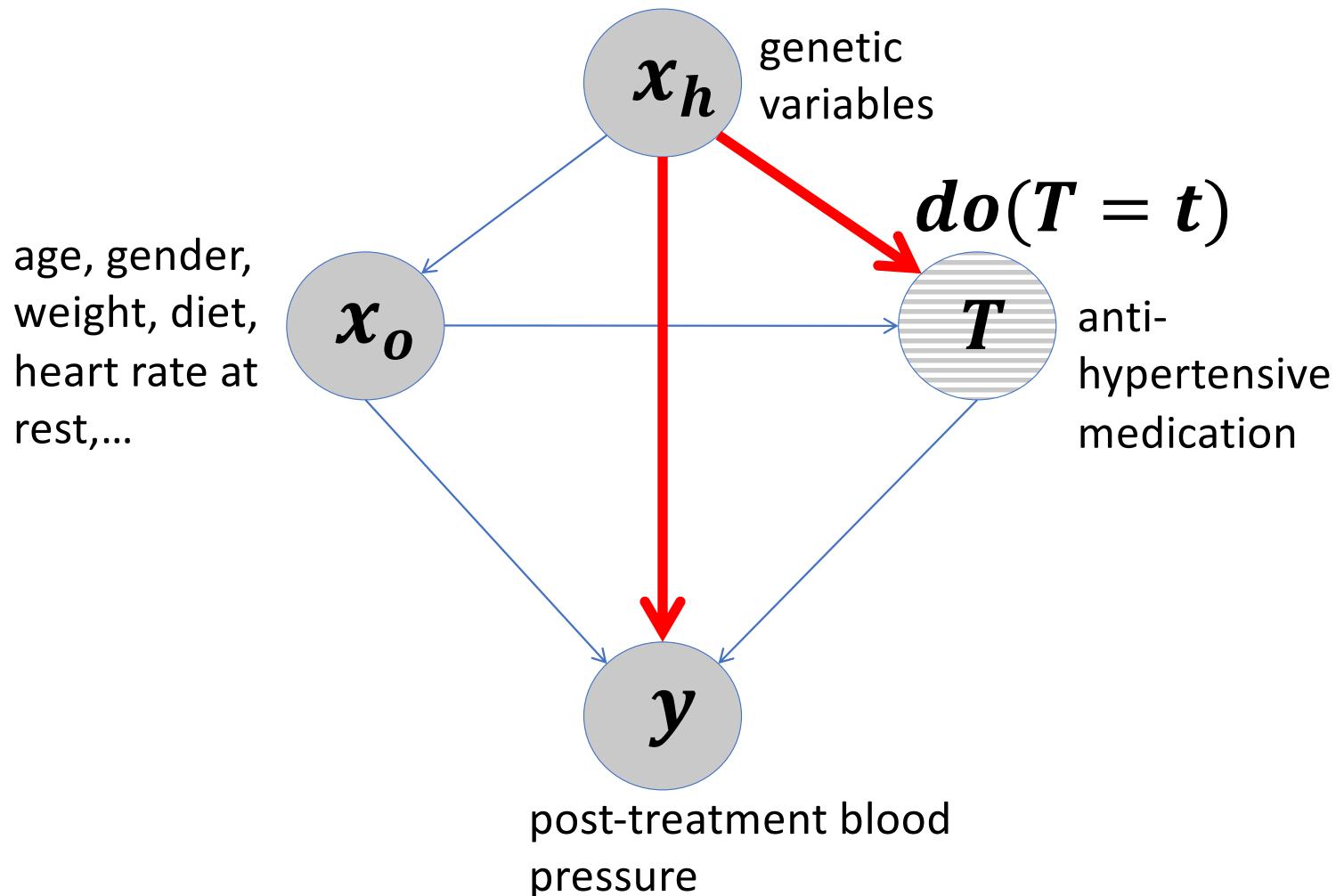


Back-door criterion

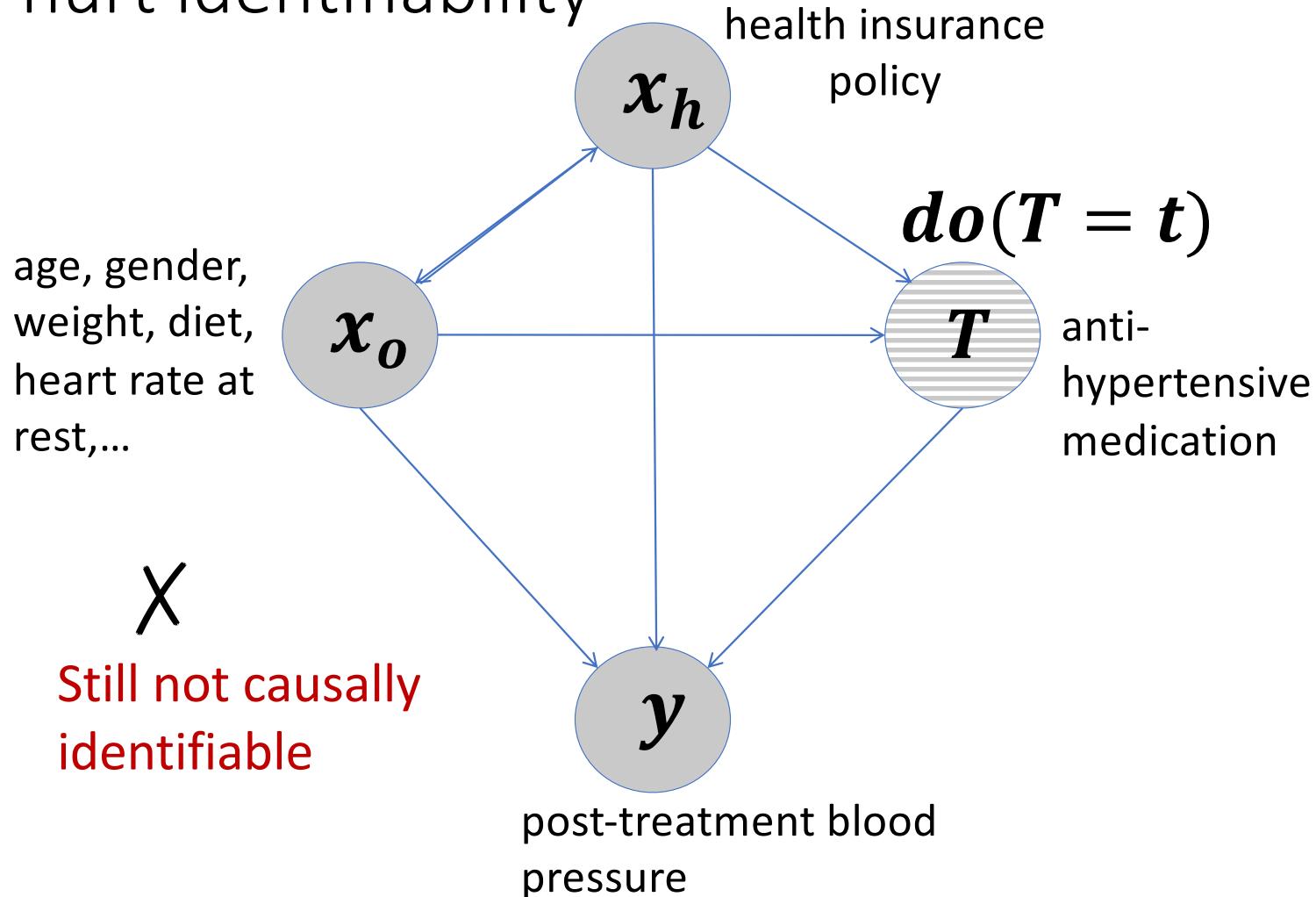
- Back-door criterion:
The observed variables d-separate all paths between y and T that end with an arrow pointing to T



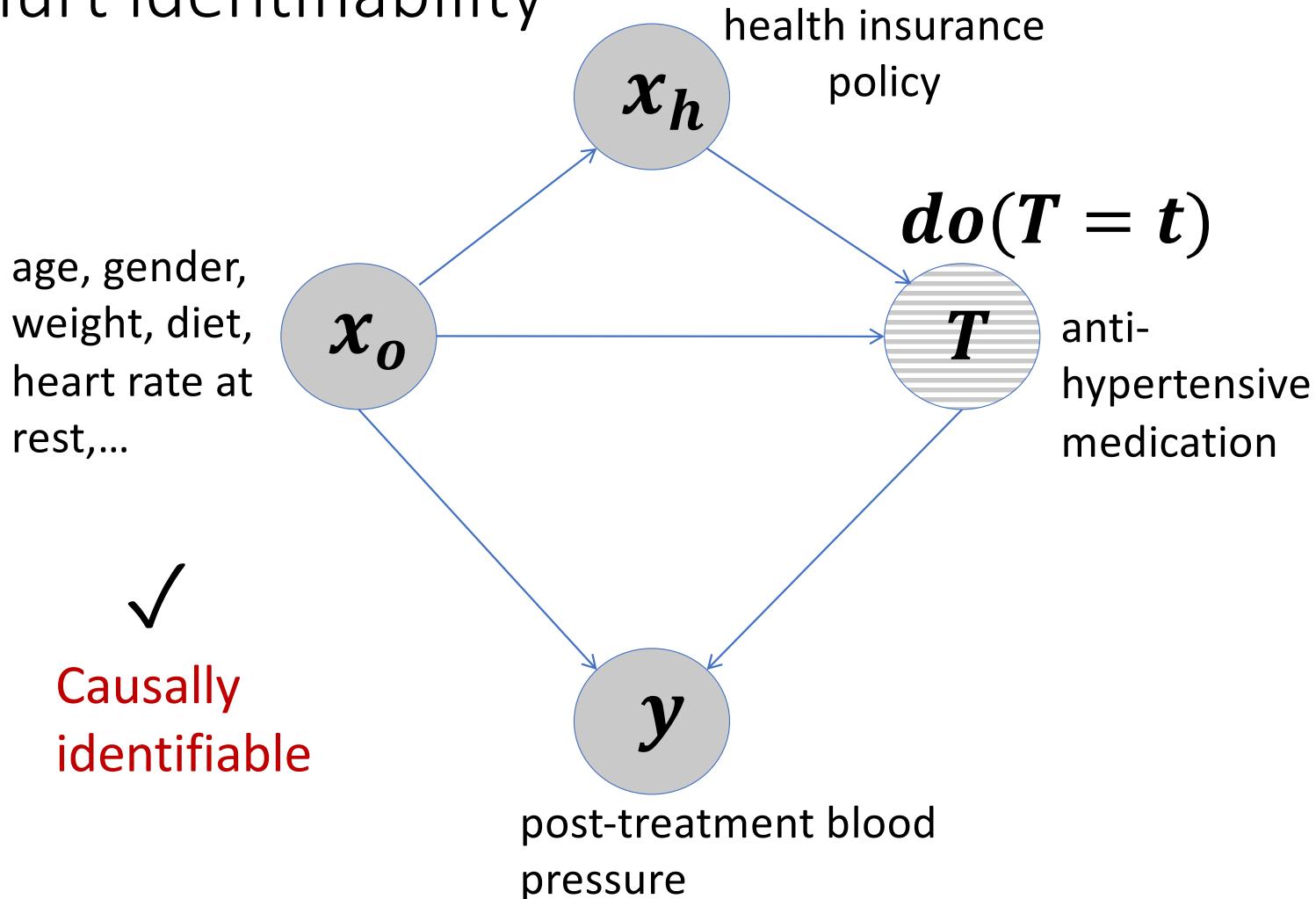
Unidentifiable causal effect



Sometimes hidden variables do not hurt identifiability

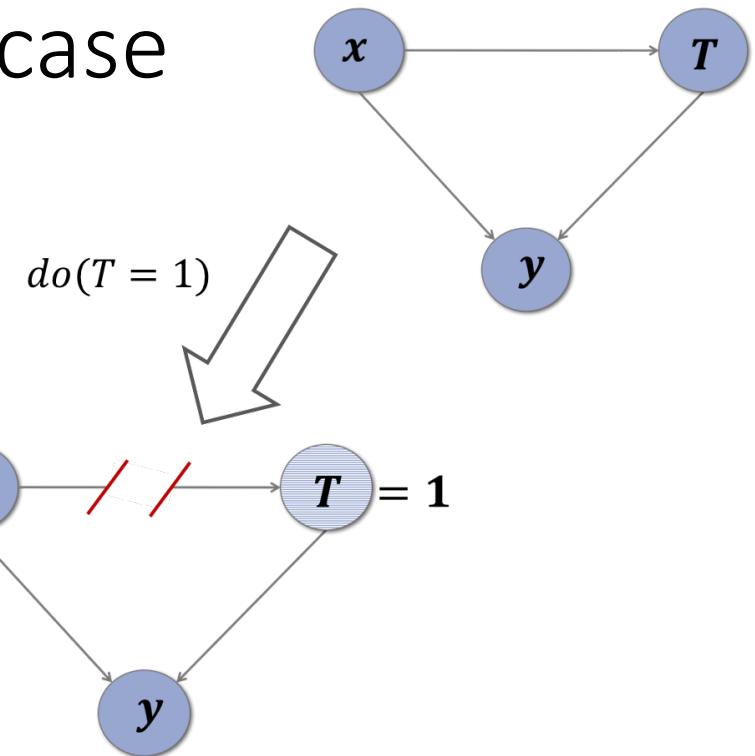


Sometimes hidden variables do not hurt identifiability



Backdoor adjustment – simple case

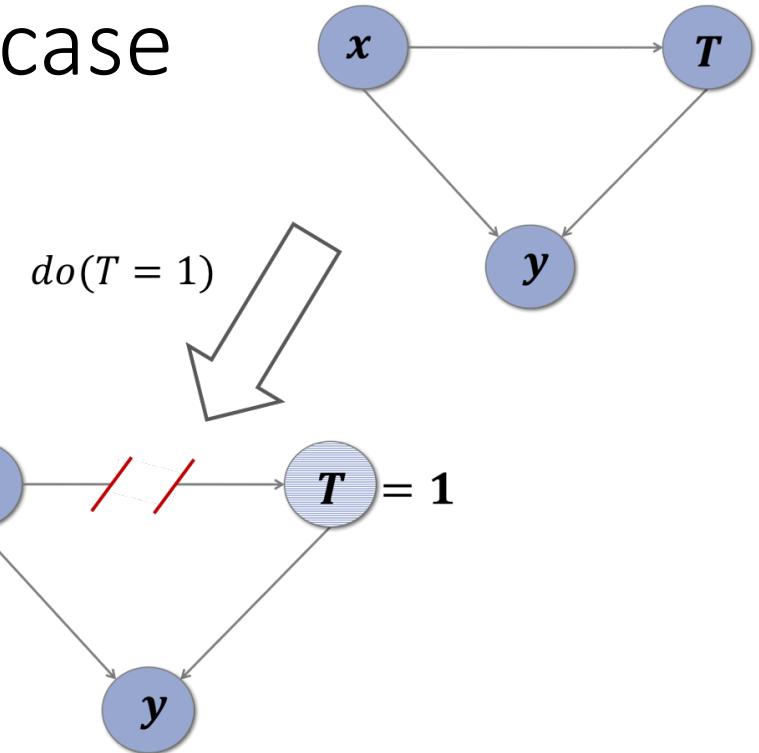
$$\mathbb{E}[y|do(T = 1)] =$$



Backdoor adjustment – simple case

$$\mathbb{E}[y|do(T = 1)] = \text{law of total expectation}$$

$$\mathbb{E}_{p(x)}[\mathbb{E}[y|do(T = 1), x]] =$$



Backdoor adjustment – simple case

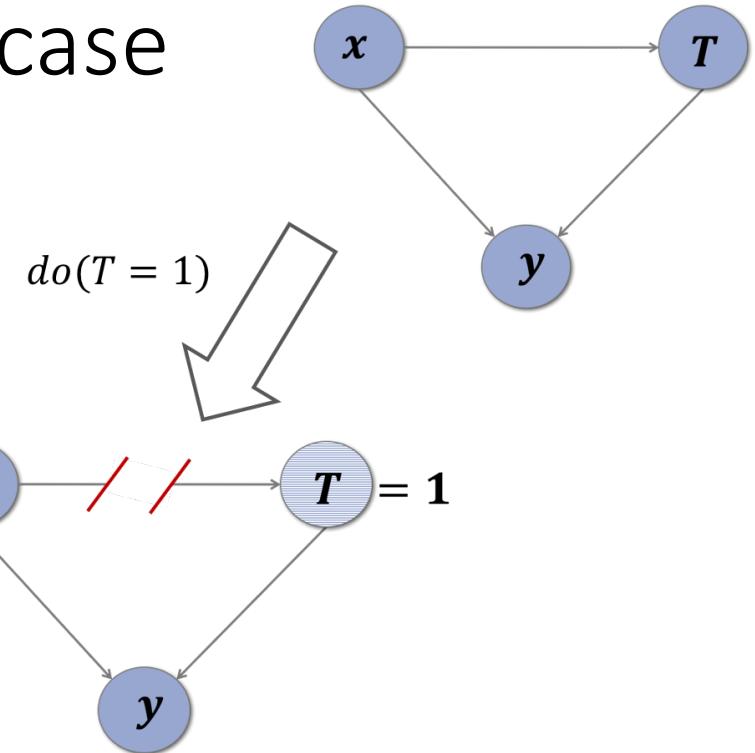
$$\mathbb{E}[y|do(T = 1)] =$$

law of total
expectation

$$\mathbb{E}_{p(x)}[\mathbb{E}[y|do(T = 1), x]] =$$

Proposition 3 of causal
graphs

$$\mathbb{E}_{p(x)}[\mathbb{E}[y|T = 1, x]]$$



Backdoor adjustment – simplest case

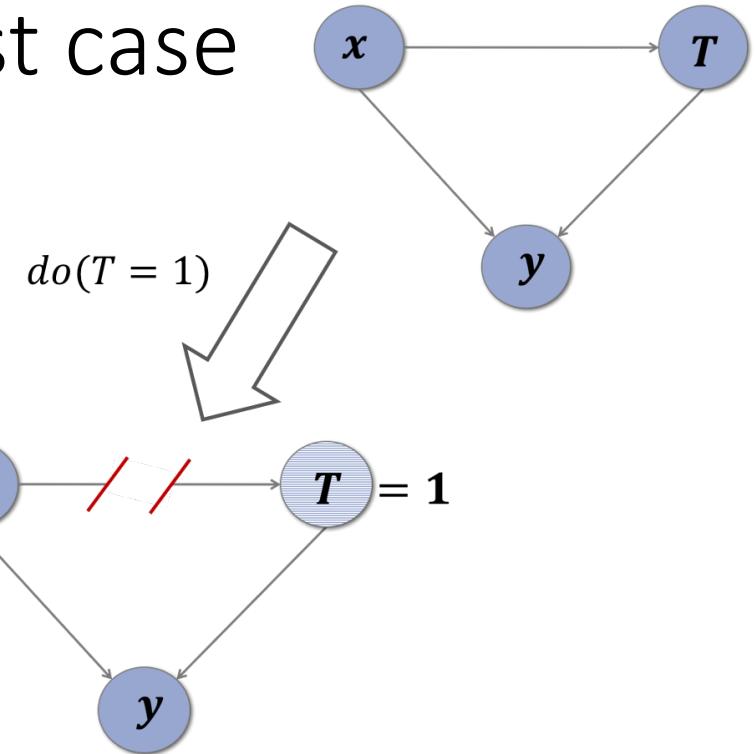
$$\mathbb{E}[y|do(T = 1)] =$$

law of total
expectation

$$\mathbb{E}_{p(x)}[\mathbb{E}[y|do(T = 1), x]] =$$

Proposition 3 of causal
graphs

$$\mathbb{E}_{p(x)}[\mathbb{E}[y|T = 1, x]]$$



- Proof above is for simple case, but the formula is same for general case

Backdoor adjustment – simplest case

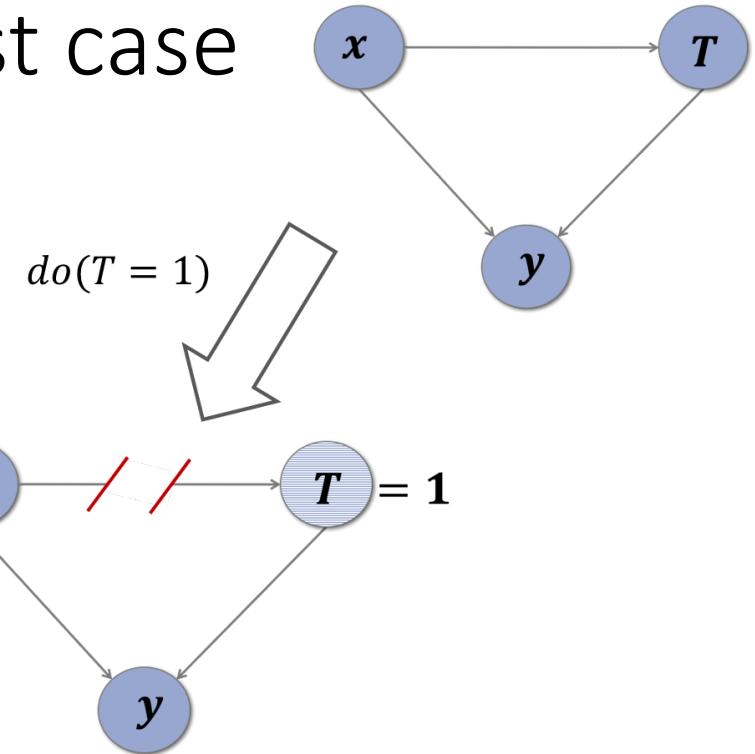
$$\mathbb{E}[y|do(T = 1)] =$$

law of total expectation

$$\mathbb{E}_{p(x)}[\mathbb{E}[y|do(T = 1), x]] =$$

Proposition 3 of causal graphs

$$\mathbb{E}_{p(x)}[\mathbb{E}[y|T = 1, x]]$$



- Proof above is for simple case, but the formula is same for general case
- Identical to the adjustment formula we saw in potential outcomes

$$\mathbb{E}[Y_1] = \mathbb{E}_{x \sim p(x)} [\mathbb{E}[Y|x, T = 1]]$$

A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks

Miguel A. Hernán, John Hsu, and Brian Healy

For much of the recent history of science, learning from data was the academic realm of statistics,^{1,2} but in the early 20th century, the founders of modern statistics made a momentous decision about what could and could not be learned from data: They proclaimed that statistics could be applied to make causal inferences when using data from randomized experiments, but not when using nonexperimental (observational) data.^{3,4,5} This decision classified an entire class of scientific questions in the health and social sciences as not amenable to formal quantitative inference.

Not surprisingly, many scientists ignored the statisticians' decree and continued to use observational data to study the unintended harms of medical treatments, health effects of lifestyle activities, or social impact of educational policies. Unfortunately, these scientists' causal questions often were mismatched with their statistical training. Perplexing paradoxes arose; for

example, the famous "Simpson's paradox" stemmed from a failure to recognize that the choice of data analysis depends on the causal structure of the problem.⁶ Mistakes occurred. For example, as a generation of medical researchers and clinicians believed that postmenopausal hormone therapy reduced the risk of heart disease because of data analyses that deviated from basic causal considerations. Even today, confusions generated by a century-old refusal to tackle causal questions explicitly are widespread in scientific research.⁷

To bridge science and data analysis, a few rogue statisticians, epidemiologists, econometricians, and computer scientists developed formal methods to quantify causal effects from observational data. Initially, each discipline emphasized different types of causal questions, developed different terminologies, and preferred different data analysis techniques. By the beginning of the 21st century, while some conceptual

discrepancies remained, a unified theory of quantitative causal inference had emerged.^{8,9}

We now have a historic opportunity to redefine data analysis in such a way that it naturally accommodates a science-wide framework for causal inference from observational data. A recent influx of data analysts, many not formally trained in statistical theory, bring a fresh attitude that does not a priori exclude causal questions. This new wave of data analysts refer to themselves as data scientists and to their activities as data science, a term popularized by technology companies and embraced by academic institutions.

Data science, as an umbrella term for all types of data analysis, can tear down the barriers erected by traditional statistics; put data analysis at the service of all scientific questions, including causal ones; and prevent unnecessary inferential mistakes. We may miss our chance to successfully integrate data analysis into all scientific

¹Tukey, J.W. 1962. The future of data analysis. *Annals of Mathematical Statistics* 33:1-67.

²Donoho, D. 2017. 50 years of data science. *Journal of Computational and Graphical Statistics* 26(4):745–66.

³Pearl, J. 2009. *Causality: Models, Reasoning, and Inference* (2nd edition). New York: Cambridge University Press.

⁴Fisher, R.A. 1925. *Statistical Methods for Research Workers*, 1st ed. Edinburgh: Oliver and Boyd.

⁵Pearson, K. 1911. *The Grammar of Science*, 3rd ed. London: Adam and Charles Black.

⁶Hernán, M.A., Clayton, D., and Keiding, N. 2011. The Simpson's paradox unraveled. *International Journal of Epidemiology* 40(3):780–5.

⁷Hernán, M.A. 2018. The C-word: Scientific euphemisms do not improve causal inference from observational data (with discussion). *American Journal of Public Health* 108(5): 616–9.

⁸Hernán, M.A., Robins J.M. 2018 (forthcoming). *Causal Inference*. Boca Raton: Chapman & Hall/CRC.

⁹Pearl, J. 2018. *The Book of Why*. New York: Basic Books.

questions, though, if data science ends up being defined exclusively in terms of technical¹⁰ activities (management, processing, analysis, visualization...) without explicit consideration of the scientific tasks.

A Classification of Data Science Tasks

Data scientists often define their work as “gaining insights” or “extracting meaning” from data. These definitions are too vague to characterize the scientific uses of data science. Only by precisely classifying the “insights” and “meaning” that data can provide will we be able to think systematically about the types of data, assumptions, and analytics that are needed. The scientific contributions of data science can be organized into three classes of tasks: description, prediction, and counterfactual prediction (see table for examples of research questions for each of these tasks).

Description is using data to provide a quantitative summary of certain features of the world. Descriptive tasks include, for example, computing the proportion of individuals with diabetes in a large healthcare database and representing social networks in a community. The analytics employed for description range from elementary calculations (a mean or a proportion) to sophisticated techniques such as unsupervised learning algorithms (cluster analysis) and clever data visualizations.

Prediction is using data to map some features of the world

(the inputs) to other features of the world (the outputs). Prediction often starts with simple tasks (quantifying the association between albumin levels at admission and death within one week among patients in the intensive care unit) and then progresses to more-complex ones (using hundreds of variables measured at admission to predict which patients are more likely to die within one week). The analytics employed for prediction range from elementary calculations (a correlation coefficient or a risk difference) to sophisticated pattern recognition methods and supervised learning algorithms that can be used as classifiers (random forests, neural networks) or predict the joint distribution of multiple variables.

Counterfactual prediction is using data to predict certain features of the world as if the world had been different, which is required in *causal inference* applications. An example of causal inference is the estimation of the mortality rate that would have been observed if all individuals in a study population had received screening for colorectal cancer vs. if they had not received screening.

The analytics employed for causal inference range from elementary calculations in randomized experiments with no loss to follow-up and perfect adherence (the difference in mortality rates between the screened and the unscreened) to complex implementations of g-methods in observational studies with

treatment-confounder feedback (the plug-in g-formula).¹¹

Note that, contrary to some computer scientists’ belief, “causal inference” and “reinforcement learning” are not synonyms. Reinforcement learning is a technique that, in some simple settings, leads to sound causal inference. However, reinforcement learning is insufficient for causal inference in complex settings (discussed below).

Statistical inference is often required for all three tasks. For example, one might want to add 95% confidence intervals for descriptive, predictive, or causal estimates involving samples of target populations.

As in most attempts at classification, the boundaries between the above categories are not always sharp. However, this trichotomy provides a useful starting point to discuss the data requirements, assumptions, and analytics necessary to successfully perform each task of data science. A similar taxonomy has traditionally been taught by data scientists from many disciplines, including epidemiology, biostatistics,¹² economics,¹³ and political science.¹⁴ Some methodologists have referred to the causal inference task as “explanation,”¹⁵ but this is a somewhat-misleading term because causal effects may be quantified while remaining unexplained (randomized trials identify causal effects even if the causal mechanisms that explain them are unknown).

Sciences are primarily defined by their questions rather than by

¹⁰Cleveland, W. 2001. Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. *International Statistical Review* 69(1):21–6.

¹¹Robins, J.M. 1986. A new approach to causal inference in mortality studies with a sustained exposure period—Application to the healthy worker survivor effect. *Mathematical Modelling* 7:1, 393–512 (1987; errata, *Mathematical Modelling* 14:917–21).

¹²Vittinghoff, E., Glidden, D.V., Shiboski, S.C., and McCulloch, C.E. 2012. *Regression Methods in Biostatistics*. New York: Springer.

¹³Mullainathan, S., and Spiess, J. 2017. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives* 31(2):87–106.

¹⁴Toshkov, D. 2016. *Research Design in Political Science*. London: Palgrave Macmillan.

¹⁵Schmueli, G. 2010. To explain or to predict? *Statistical Science* 25(3):289–310.

their tools: We define astrophysics as the discipline that learns the composition of the stars, not as the discipline that uses the spectroscope. Similarly, data science is the discipline that describes, predicts, and makes causal inferences (or, more generally, counterfactual predictions), not the discipline that uses machine learning algorithms or other technical tools. Of course data science certainly benefits from the development of tools for the acquisition, storage, integration, access, and processing of data, as well as from the development of scalable and parallelizable analytics. This data engineering powers the scientific tasks of data science.

Prediction vs. Causal Inference

Data science has excelled at commercial applications, such as shopping and movie recommendations, credit rating, stock trading algorithms, and advertisement placement. Some data scientists have transferred their skills to scientific research with biomedical applications such as Google's algorithm to diagnose diabetic retinopathy¹⁶ (after 54 ophthalmologists classified more than 120,000 images), Microsoft's algorithm to predict pancreatic cancer months before its usual diagnosis¹⁷ (using the online search histories of 3,000 users who were later diagnosed

with cancer), and Facebook's algorithm to detect users who may be suicidal¹⁸ (based on posts and live videos).

All these applications of data science have one thing in common: They are predictive, not causal. They map inputs (an image of a human retina) to outputs (a diagnosis of retinopathy), but they do not consider how the world would look like under different courses of action (whether the diagnosis would change if we operated on the retina).

Mapping observed inputs to observed outputs is a natural candidate for automated data analysis because this task only requires: 1) a large data set with inputs and outputs, 2) an algorithm that establishes a mapping between inputs and outputs, and 3) a metric to assess the performance of the mapping, often based on a gold standard.¹⁹ Once these three elements are in place, as in the retinopathy example, predictive tasks can be automated via data-driven analytics that evaluate and iteratively improve the mapping between inputs and outputs without human intervention.

More precisely, the component of prediction tasks that can be automated easily is the one that does not involve any expert knowledge. Prediction tasks require expert knowledge to specify the scientific question—what to input and what outputs—and to identify/

generate relevant data sources.²⁰ (The extent of expert knowledge varies with different prediction tasks.²¹) However, no expert knowledge is required for prediction after candidate inputs and the outputs are specified and measured in the population of interest. At this point, a machine learning algorithm can take over the data analysis to deliver a mapping and quantify its performance. The resulting mapping may be opaque, as in many deep learning applications, but its ability to map the inputs to the outputs with a known accuracy in the studied population is not in question.

The role of expert knowledge is the key difference between prediction and causal inference tasks. Causal inference tasks require expert knowledge not only to specify the question (the causal effect of what treatment on what outcome) and identify/generate relevant data sources, but also to describe the causal structure of the system under study. Causal knowledge, usually in the form of unverifiable assumptions,^{22,23} is necessary to guide the data analysis and to provide a justification for endowing the resulting numerical estimates with a causal interpretation. In other words, the validity of causal inferences depends on structural knowledge, which is usually incomplete, to supplement the information in the data. As a consequence, no algorithm

¹⁶Gulshan, V., Peng, L., Coram, M., et al. 2016. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* 316(22):2,402–10.

¹⁷Paparrizos, J., White, R.W., and Horvitz, E. 2016. Screening for Pancreatic Adenocarcinoma Using Signals From Web Search Logs: Feasibility Study and Results. *Journal of Oncological Practice* 12(8):737–44.

¹⁸Rosen, G. 2017. Getting Our Community Help in Real Time. <https://newsroom.fb.com/news/2017/11/getting-our-community-help-in-real-time/> (accessed April 26, 2018).

¹⁹Brynjolfsson, E., and Mitchell, T. 2017. What can machine learning do? Workforce implications. *Science* 358(6370):1,530–4.

²⁰Conway, D. 2010. The Data Science Venn Diagram. Accessed October 9, 2018. <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>.

²¹Beam, A.L., and Kohane I.S. 2018. Big Data and Machine Learning in Health Care. *JAMA* 319(13):1,317–8.

²²Robins, J.M. 2001. Data, design, and background knowledge in etiologic inference. *Epidemiology* 11:313–20.

²³Robins, J.M., and Greenland, S. 1986. The role of model selection in causal inference from nonexperimental data. *American Journal of Epidemiology* 123(3):392–402.

can quantify the accuracy of causal inferences from observational data. The following simplified example helps fix ideas about the different role of expert knowledge for prediction versus causal inference.

Example

Suppose we want to use a large health records database to predict infant mortality (the output) using clinical and lifestyle factors collected during pregnancy (the inputs). We have just applied our expert knowledge to decide what the output and candidate inputs are, and to select a particular database in the population of interest. The only requirement is that the potential inputs must precede the outputs temporally, regardless of the causal structure linking them. At this point of the process, our expert knowledge will not be needed any more: An algorithm can provide a mapping between inputs and outputs at least as good as any mapping we could propose and, in many cases, astoundingly better.

Now suppose we want to use the same health records database to determine the causal effect of maternal smoking during pregnancy on the risk of infant mortality. A key problem is confounding: Pregnant women who do and do not smoke differ in many characteristics (including alcohol consumption, diet, access to adequate prenatal care) that affect the risk of infant mortality. Therefore, a causal analysis must identify and adjust for those confounding factors which, by definition, are

associated with both maternal smoking and infant mortality.

However, not all factors associated with maternal smoking and infant mortality are confounders that should be adjusted for. For example, birthweight is strongly associated with both maternal smoking and infant mortality, but adjustment for birthweight induces bias because birthweight is a risk factor that is itself causally affected by maternal smoking. In fact, adjustment for birthweight results in a bias often referred to as the “birthweight paradox”: Low birthweight babies from mothers who smoked during pregnancy have a lower mortality than those from mothers who did not smoke during pregnancy.²⁴

An algorithm devoid of causal expert knowledge will rely exclusively on the associations found in the data and is therefore at risk of selecting features, like birthweight, that increase bias. The “birthweight paradox” is indeed an example of how the use of automatic adjustment procedures may lead to an incorrect causal conclusion. In contrast, a human expert can readily identify many variables that, like birthweight, should not be adjusted for because of their position in the causal structure.

A human expert also may identify features that should be adjusted for, even if they are not available in the data, and propose sensitivity analyses²⁵ to assess the reliability of causal inferences in the absence of those features. In contrast, an algorithm that ignores the causal structure will not issue an alert

about the need to adjust for features that are not in the data.

Given the central role of (potentially fallible) expert causal knowledge in causal inference, it is not surprising that researchers look for procedures to alleviate the reliance of causal inferences on causal knowledge. Randomization is the best such procedure.

When a treatment is randomly assigned, we can unbiasedly estimate the average causal effect of treatment assignment *in the absence of detailed causal knowledge about the system under study*. Randomized experiments are central in many areas of science where relatively simple causal questions are asked.²⁶ Randomized experiments are also commonly used, often under the name A/B testing, to answer simple causal questions in commercial web applications. However, randomized designs are often infeasible, untimely, or unethical in the extremely complex systems studied by health and social scientists.²⁶

A failure to grasp the different role of expert knowledge in prediction and causal inference is a common source of confusion in data science (the confusion is compounded by the fact that predictive analytic techniques, such as regression, can also be used for causal inference when combined with causal knowledge).

Both prediction and causal inference require expert knowledge to formulate the scientific question *i*, but only causal inference requires causal expert knowledge to answer the question. As a result,

²⁴Hernández-Díaz, S., Schisterman, E.F., and Hernán, M.A. 2006. The birth weight “paradox” uncovered? *American Journal of Epidemiology* 164(11):1,115–20.

²⁵Robins, J.M., Rotnitzky, A., and Scharfstein, D.O. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In Halloran E., and Berry D., eds. *Statistical Methods in Epidemiology: The Environment and Clinical Trials*. New York: Springer Verlag; 1999:1–92.

²⁶Hernán, M.A. 2015. Invited commentary: Agent-based models for causal inference-reweighting data and theory in epidemiology. *American Journal of Epidemiology* 2181(2):103–5.

the accuracy of causal estimates cannot be assessed by using metrics computed from the data, even if the data were perfectly measured in the population of interest.

Implications for Decision-making

A goal of data science is to help people make better decisions. For example, in health settings, the goal is to help decision-makers—patients, clinicians, policy-makers, public health officers, regulators—decide among several possible strategies. Frequently, the ability of data science to improve decision-making is predicated on the basis of its success at prediction.

However, the premise that predictive algorithms will lead to better decisions is questionable. An algorithm that excels at using data about patients with heart failure to predict who will die within the next five years is agnostic about how to reduce mortality. For example, a prior hospitalization may be identified as a useful predictor of mortality, but nobody would suggest that we stop hospitalizing people to reduce mortality. Identifying patients with bad prognoses is very different from identifying the best course of action for preventing or treating a disease. Worse, predictive algorithms, when incorrectly used for causal inference, may lead to incorrect confounder adjustment and therefore conclude, for example, that maternal smoking appears to be beneficial for low birthweight babies.

Predictive algorithms inform us that decisions have to be made, but they cannot help us make the

decisions. For example, a predictive algorithm that identifies patients with severe heart failure does not provide information about whether heart transplant is the best treatment option. In contrast, causal analyses are designed to help us make decisions because they tackle “what if” questions. A causal analysis will, for instance, compare the benefit-risk profile of heart transplant versus medical treatment in patients with certain severity of heart failure.

Interestingly, the distinction between prediction and causal inference (counterfactual prediction) becomes unnecessary for decision-making when the relevant expert knowledge can readily be encoded and incorporated into the algorithms. A purely predictive algorithm that learns to play Go can perfectly predict the counterfactual state of the game under different moves, and a predictive algorithm that learns to drive a car can accurately predict the counterfactual state of the car if, say, the brakes are not operated.

Because these systems are governed by a set of known game rules (in the case of games like Go) or physical laws with some stochastic components (in the case of engineering applications like self-driving cars), an algorithm can eventually predict the behavior of the entire system under a hypothetical intervention.

Take the game of Go, which has been mastered by an algorithm “without human knowledge.”²⁷ When making a move, the algorithm has access to all information that matters: game rules, current board position, and future outcomes fully determined by the

sequence of moves. Further, a reinforcement learning algorithm can collect an arbitrary amount of data by playing more games (conducting numerous experiments), which allows it to learn by trial and error. In this setting, a cleverly designed algorithm running on a powerful computer can spectacularly outperform humans—but this form of causal inference has, at this time in history, a restricted domain of applicability.

Many scientists work on complex systems with partly known and nondeterministic governing laws (the “rules of the game”), with uncertainty about whether all necessary data are available, and for which learning by trial and error—or even conducting a single experiment—is impossible. Even when the laws are known and the data available, the system may still be too chaotic for exact long-term prediction. For example, it was impossible to predict when and where the Chinese space station,²⁸ while in orbit at an altitude of about 250 km, would fall to Earth.

Consider a causal question about the effect of different epoetin strategies on the mortality of patients with renal disease. We do not understand the causal structure by which molecular, cellular, individual, social, and environmental factors regulate the effect of epoetin dose on mortality risk. As a result, it is currently impossible to construct a predictive model based on electronic health records to reproduce the behavior of the system under a hypothetical intervention on an individual. Some widely publicized disappointments in causal applications of data science, like “Watson for Oncology,”

²⁷Silver, D., Schrittwieser, J., and Simonyan, K., et al. 2017. Mastering the game of Go without human knowledge. *Nature* 550(7676):354–9.

²⁸The Data Team. 2018. An out-of-control Chinese space station will soon fall to Earth. *The Economist* March 19, 2018.

Table 1—Examples of Tasks Conducted by Data Scientists Working with Electronic Health Records

Data Science Task			
	Description	Prediction	Causal inference
Example of scientific question	How can women aged 60–80 years with stroke history be partitioned in classes defined by their characteristics?	What is the probability of having a stroke next year for women with certain characteristics?	Will starting a statin reduce, on average, the risk of stroke in women with certain characteristics?
Data	<ul style="list-style-type: none"> • Eligibility criteria • Features (symptoms, clinical parameters ...) 	<ul style="list-style-type: none"> • Eligibility criteria • Output (diagnosis of stroke over the next year) • Inputs (age, blood pressure, history of stroke, diabetes at baseline) 	<ul style="list-style-type: none"> • Eligibility criteria • Outcome (diagnosis of stroke over the next year) • Treatment (initiation of statins at baseline) • Confounders • Effect modifiers (optional)
Examples of analytics	Cluster analysis ...	Regression Decision trees Random forests Support vector machines Neural networks ...	Regression Matching Inverse probability weighting G-formula G-estimation Instrumental variable estimation ...

have arguably resulted from trying to predict a complex system that is still poorly understood and for which a sound model to combine expert causal knowledge with the available data is lacking.²⁹

The striking contrast between the cautious attitude of most traditional data scientists (statisticians, epidemiologists, economists, political scientists...) and the “can do” attitude of many computer scientists, informaticians, and others seems to be, to a large extent, the consequence of the different complexity of the causal questions

historically tackled by each of these groups. Epidemiologists and other data scientists working with extremely complex systems tend to focus on the relatively modest goal of designing observational analyses to answer narrow causal questions about the average causal effect of a variable (such as epoetin treatment), rather than try to explain the causal structure of the entire system or identify globally optimal decision-making strategies.

On the other hand, newcomers to data science have often focused on systems governed by known

laws (like board games or self-driving cars), so it is not surprising that they have deemphasized the distinction between prediction and causal inference. Bringing this distinction to the forefront is, however, urgent as an increasing number of data scientists address the causal questions traditionally asked by health and social scientists. Sophisticated prediction algorithms may suffice to develop unbeatable Go software and, eventually, safe self-driving vehicles, but causal inferences in complex systems (say, the effects of clinical strategies to

²⁹Ross, C., and Swetlitz, I. 2017. IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close. STAT. <https://www.statnews.com/2017/09/05/watson-ibm-cancer/>.

³⁰Pearl, J. 2018. Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution. Technical Report R-475 (http://ftp.cs.ucla.edu/pub/stat_ser/r475.pdf). Accessed April 26, 2018.

treat a chronic disease) need to rely on data analysis methods equipped with causal knowledge.³⁰

Processes and Implications for Teaching

The training of data scientists tends to emphasize mastering tools for data management and data analysis. While learning to use these tools will continue to play a central role, it is important that the technical training of data scientists makes it clear that the tools are at the service of distinct scientific tasks—description, prediction, and causal inference.

A training program in data science can, therefore, be organized explicitly in three components, each devoted to one of the three tasks of data science. Each component would describe how to articulate scientific questions, data requirements, threats to validity, data analysis techniques, and the role of expert knowledge (separately for description, prediction, and causal inference). This is the approach that we adopted to develop the curriculum of the Clinical Data Science core at the Harvard Medical School, which three cohorts of clinical investigators have now learned.

Our students first learn to differentiate between the three tasks of data science, then how to generate and analyze data for each task, as well as the differences between tasks. They learn that description and prediction may be affected by selection and measurement biases, but that only causal inference is affected by confounding. After learning predictive

algorithms, teams of students compete against each other in a machine learning competition to develop the best predictive model (in an application of the Common Task Framework²).

By contrast, after learning causal inference techniques, students understand that a similar competition is not possible because their causal estimates cannot be ranked automatically. Teams with different subject-matter knowledge may produce different causal estimates, and there often is no objective way to determine which one is closest to the truth using the existing data.³¹

Then students learn to ask causal questions in terms of a contrast of interventions conducted over a fixed time period as would be specified in the protocol of a (possibly hypothetical) experiment, which is the target of inference.

For example, to compare the mortality under various epoetin dosing strategies in patients with renal failure, students use subject-matter knowledge to 1) outline the design of the hypothetical randomized experiment that would estimate the causal effect of interest—the target trial, 2) identify an observational database with sufficient information to approximately emulate the target trial, and 3) emulate the target trial and therefore estimate the causal effect of interest using the observational database. We discuss why causal questions that cannot be translated into target experiments are not sufficiently well-defined,³¹ and why the accuracy of causal answers cannot be quantified using observational data. In parallel, the students also learn computer

coding and the basics of statistical inference to deal with the uncertainty inherent to any data analyses involving description, prediction, or causal inference.

A data science curriculum along the three dimensions of description, prediction, and causal inference facilitates interdisciplinary integration. Learning from data requires paying attention to the different emphases, questions, and analytic methods developed over several decades in statistics, epidemiology, econometrics, computer science, and others. Data scientists without subject-matter knowledge cannot conduct causal analyses in isolation: They don't know how to articulate the questions (what the target experiment is) and they don't know how to answer them (how to emulate the target experiment).

Conclusion

Data science is a component of many sciences, including the health and social ones. Therefore, the tasks of data science are the tasks of those sciences—description, prediction, causal inference. A sometimes-overlooked point is that a successful data science requires not only good data and algorithms, but also domain knowledge (including causal knowledge) from its parent sciences.

The current rebirth of data science is an opportunity to rethink data analysis free of the historical constraints imposed by traditional statistics, which have left scientists ill-equipped to handle causal questions. While the clout of statistics in scientific training and publishing impeded the introduction of a unified formal framework for causal inference in data

³¹Hernán, MA. 2019 (in press). Spherical cows in a vacuum: Data analysis competitions for causal inference. *Statistical Science*.

analysis, the coining of the term “data science” and the recent influx of “data scientists” interested in causal analyses provides a once-in-a-generation chance of integrating all scientific questions, including causal ones, in a principled data analysis framework. An integrated data science curriculum can present a coherent conceptual framework that fosters understanding and collaboration between data analysts and domain experts.

On the other hand, if the definitions of data science currently discussed in mainstream statistics take hold, causal inference from observational data will be once more marginalized, leaving health and social scientists on their own. The American Statistical Association statement on “The Role of Statistics in Data Science” (August 8, 2015) makes no reference to causal inference. A recent assessment of data science and statistics² did not include the word “causal” (except when mentioning the title of the course “Experiments and Causal Inference”). Heavily influenced by statisticians, many medical editors actively suppress the term “causal” from their publications.³³

A data science that embraces causal inference must (1) develop methods for the integration of sophisticated analytics with expert causal expertise, and (2) acknowledge that, unlike for prediction, the assessment of the

validity of causal inferences cannot be exclusively data-driven because the validity of causal inferences also depends on the adequacy of expert causal knowledge. Causal directed acyclic graphs^{34,35} may play an important role in the development of analytic methods that integrate learning algorithms and subject-matter knowledge. These graphs can be used to represent different sets of causal structures that are compatible with existing causal knowledge and thus to explore the impact of causal uncertainty on the effect estimates.

Large amounts of data could make expert knowledge irrelevant for prediction and for relatively simple causal inferences involving games and some engineering applications, but expert causal knowledge is necessary to formulate and answer causal questions in more-complex systems. Affirming causal inference as a legitimate scientific pursuit is the first step in transforming data science into a reliable tool to guide decision-making.

Finally, the distinction between prediction and causal inference is also crucial to defining artificial intelligence (AI). Some data scientists argue that “the essence of intelligence is the ability to predict,” and therefore that good predictive algorithms are a form of AI. From this point of view, large chunks of data science can be

rebranded as AI (and that is exactly what the tech industry is doing). However, mapping observed inputs to observed outputs barely qualifies as intelligence. Rather, a hallmark of intelligence is the ability to predict *counterfactually* how the world would change under different actions by integrating expert knowledge and mapping algorithms. No AI will be worthy of the name without causal inference. □

About the Authors

Miguel Hernán conducts research to learn what works for the treatment and prevention of cancer, cardiovascular disease, and HIV infection. With his collaborators, he designs analyses of healthcare databases, epidemiologic studies, and randomized trials. He teaches clinical data science at the Harvard Medical School, clinical epidemiology at the Harvard-MIT Division of Health Sciences and Technology, and causal inference methodology at the Harvard T.H. Chan School of Public Health, where he is the Kolokotrones Professor of Biostatistics and Epidemiology.

John Hsu is director of the Program for Clinical Economics and Policy Analysis in the Mongan Institute, Massachusetts General Hospital, and Harvard Medical School. He studies innovations in healthcare financing and delivery, and their effects on medical quality and efficiency. He primarily uses large automated and electronic health record data sets, often exploiting natural experiments from both clinical and behavioral economics perspectives.

Brian Healy is an assistant professor of neurology at the Harvard Medical School and an assistant professor in the Department of Biostatistics at the Harvard T.H. Chan School of Public Health. He is the primary biostatistician for the Partners MS Center at Brigham and Women’s Hospital and a member of the Massachusetts General Hospital (MGH) Biostatistics Center. He teaches introductory statistics in several programs and codirects the clinical data science sequence in the master of medical science and clinical investigation with Miguel Hernán.

³²Ruich, P. 2017. The Use of Cause-and-Effect Language in the JAMA Network Journals. *AMA Style Insider*. <http://amastyleinsider.com/2017/09/19/use-cause-effect-language-jama-network-journals/>. Accessed May 25, 2018.

³³Hernán, M.A., Hernández-Díaz, S., Werler, M.M., and Mitchell, A.A. 2002. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *American Journal of Epidemiology* 155:176–84.

³⁴Greenland, S., Pearl, J., and Robins, J.M. Causal diagrams for epidemiologic research. *Epidemiology* 1999; 10(1):37–48.



Introduction to Causal Inference

Dr. Uri Shalit

Course number 097400
2020-2021

Admin

- Lecturer: Dr. Uri Shalit
urishalit@technion.ac.il
- TA: Rom Gutman
romgutman@campus.technion.ac.il
- Lecture: Wednesdays 14:30 – 16:20
- Turgul: Wednesdays 16:30 – 17:20
- Course website on Moodle

Prerequisites

- Working knowledge of probability theory:
probability distributions, expectations, conditioning, Bayes' Theorem
 $p(x), p(y|x), \mathbb{E}[x], \mathbb{E}[y|x]$
- Basic knowledge of statistics and/or machine learning:
Linear and logistic regression, training and test set,
mean squared error
- Familiarity with handling data using tools such as one of the
following: Python (pandas), R, Matlab, Stata, SAS

Course requirements

- Four hand-in exercises: 20%
- Final project: 80% (10% class presentation)
 - Details at the end of today's lessons



What can we do with all this data?

What is all this data good for?

- Find which medication is best for diabetics?
- Decide which ad should I put in front of a user so they would click it?
- Decide whether higher minimum wage lead to more unemployment?
- Does a recent history of common cold infection protect from COVID-19?

Predicting diabetes

- Say we have a dataset of diabetic patients from Kupat Holim Clalit: their lab tests, diagnoses, medications
- Question: who will be *severely diabetic* in 1 year?
- Build predictive model:
features $X = [\text{lab_tests}, \text{diagnoses}, \text{medications}]$
label $y = [\text{severely_diabetic}]$
- Predict y from X
- Prediction accuracy (AUC): 0.8
- **Now what?**

We predicted diabetes onset... now what?

- You are CEO of Kupat Holim Clalit, you have a tool for predicting diabetes. What do you do with it?
- The questions we *really* care about:
 - Why do some people become diabetics?
 - How can we help them?
- Both of these are causal questions

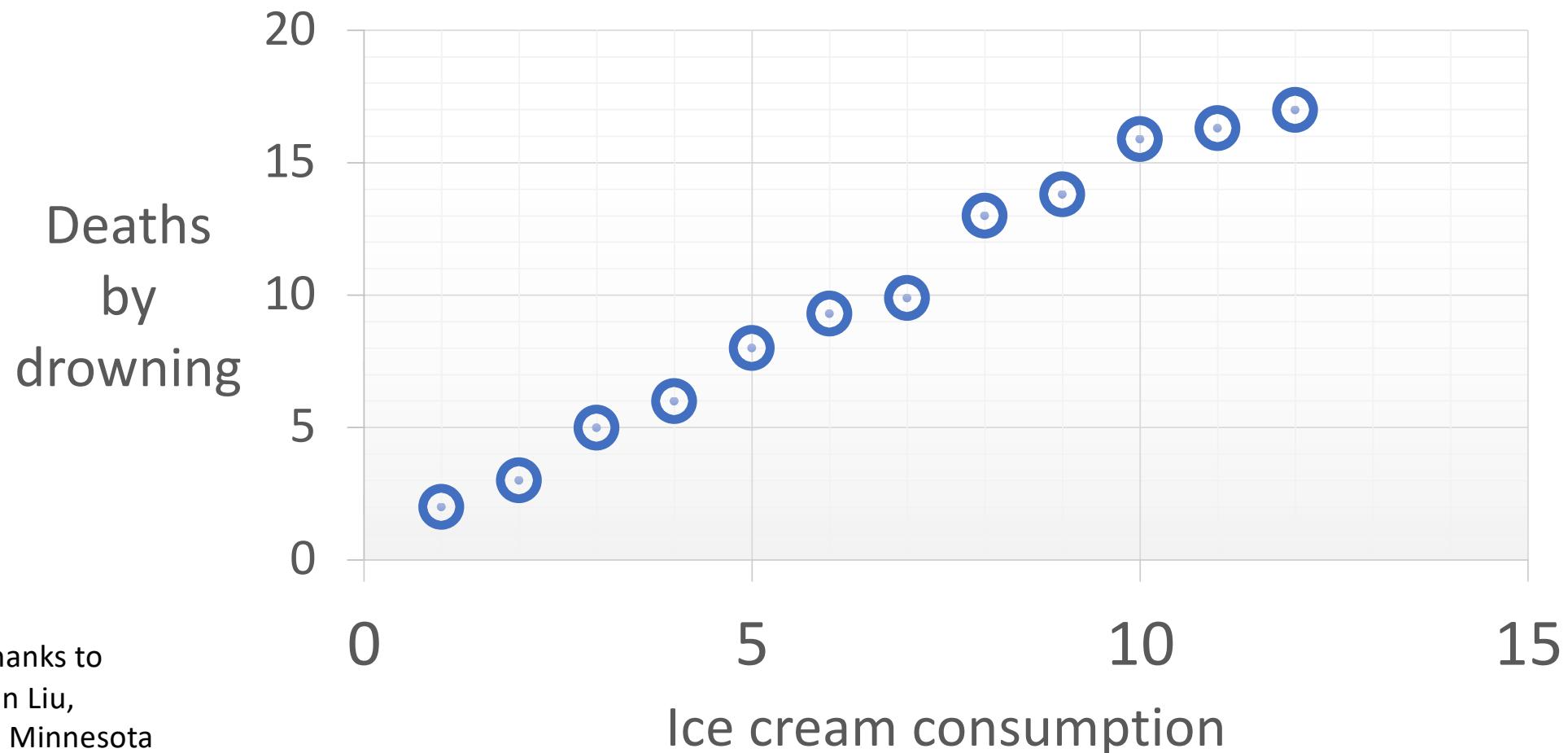
Causality: understanding the effects of actions (sometimes called interventions, treatments)

- How will patients respond to *medication*?
- How will customers respond to *ad presented*?
- Will the new *app design* lead to more engagement?
- How will students respond to *teaching method*?
- How will employers respond to *new regulations about min. wage*?
- Which customers should *receive a good loan* from the bank
- Will *working in a kindergarten* increase my odds of being infected with COVID-19?

Causality: understanding the effects of actions, interventions, treatments

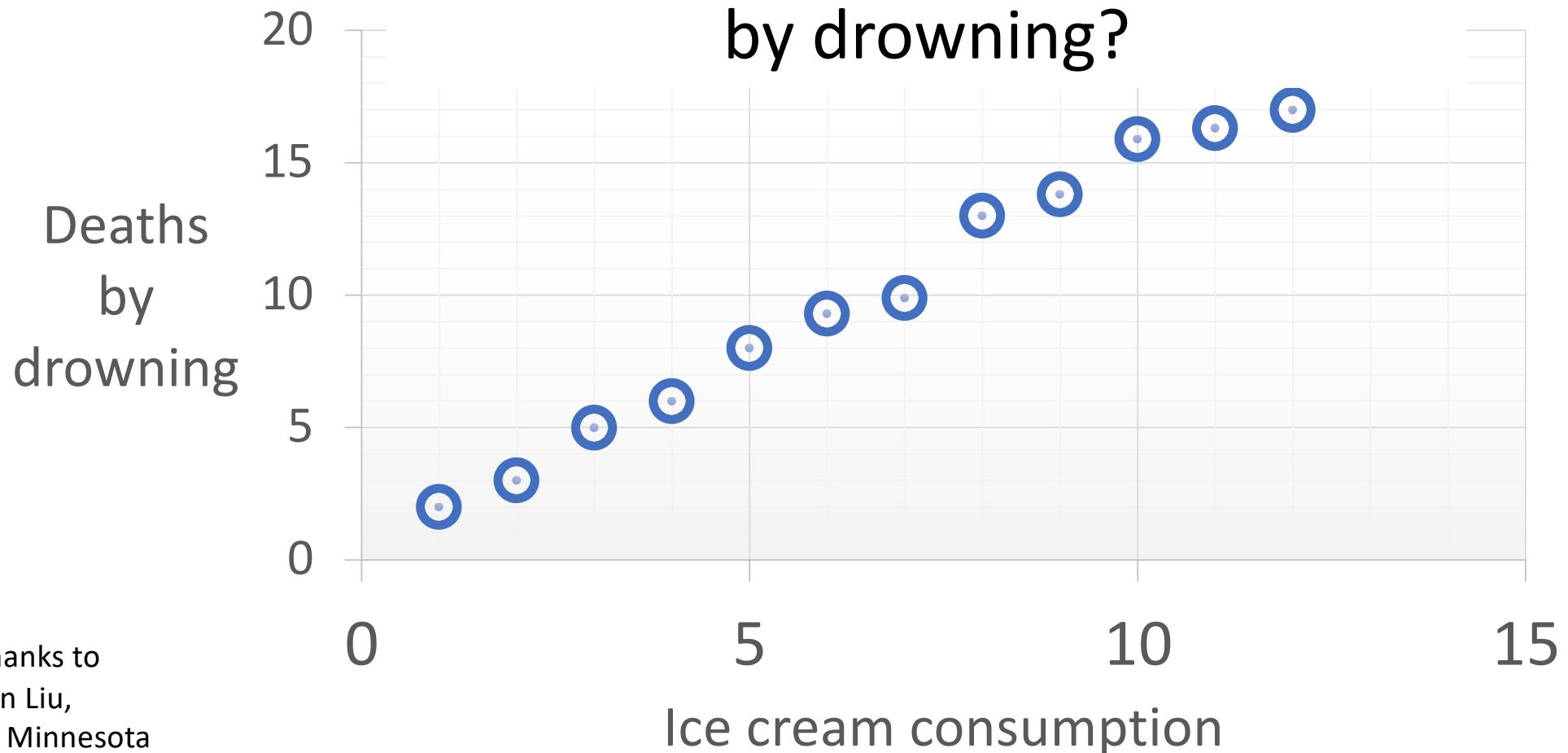
- How will patients respond to *medication*?
- How will customers respond to *ad presented*?
- Will the new *app design* lead to more engagement?
- How will students respond to *teaching method*?
- How will employers respond to *new regulations about min. wage*?
- Which customers should *receive a good loan* from the bank
- Will *working in a kindergarten* increase my odds of being infected with COVID-19?

Confounding: is ice cream deadly?



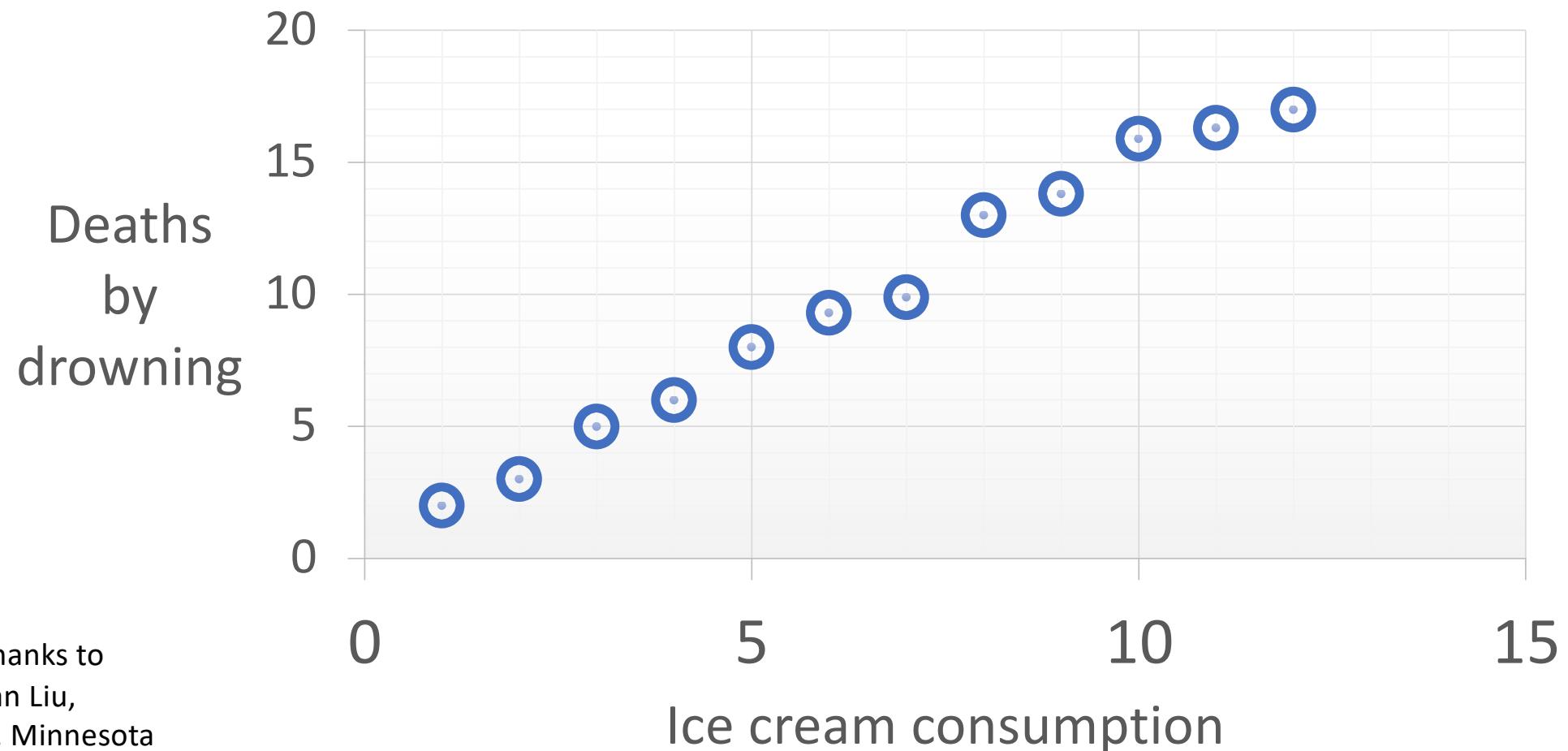
Thanks to
Lan Liu,
U. Minnesota

Should we prevent the selling of
ice cream so we can reduce death
by drowning?



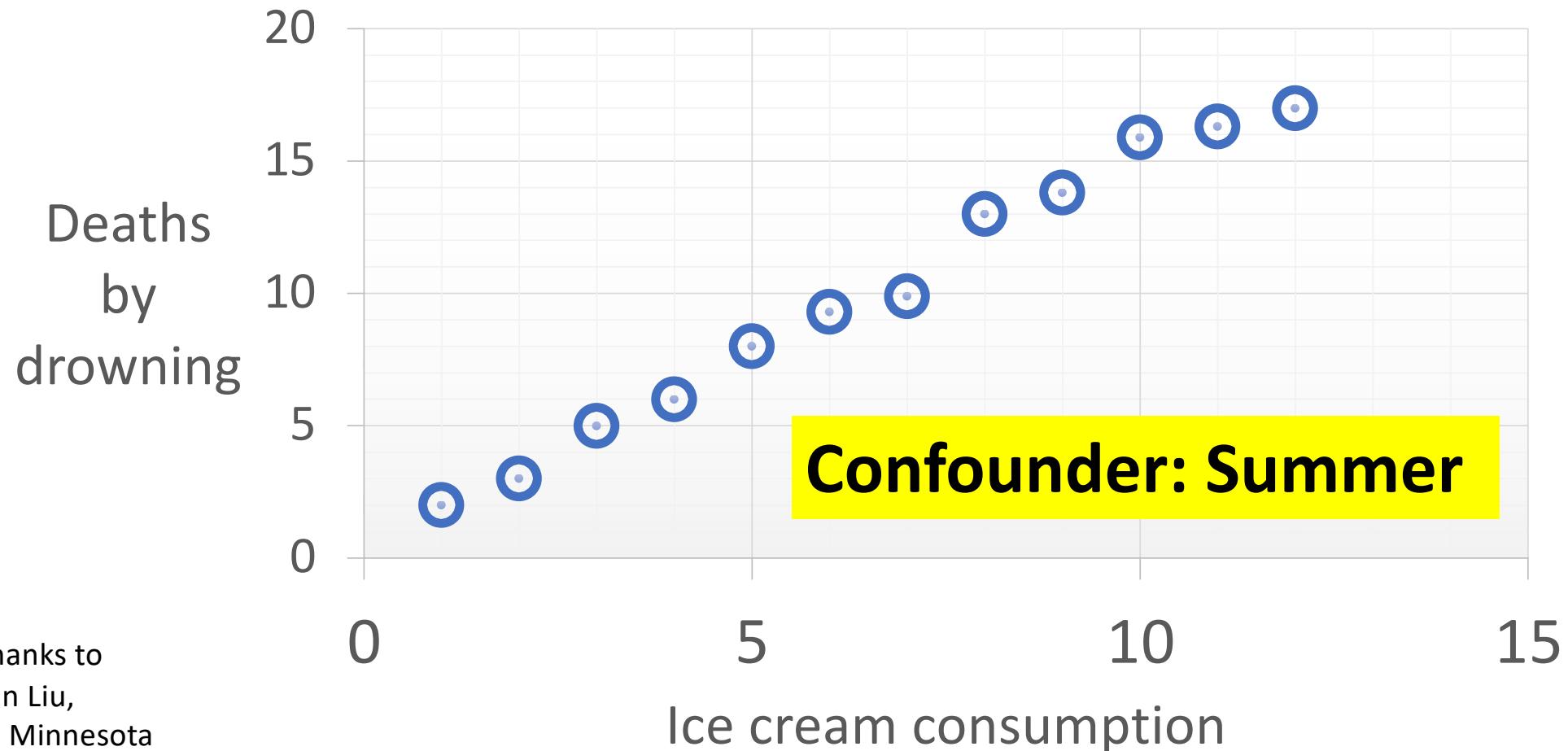
Thanks to
Lan Liu,
U. Minnesota

Will more lifesavers lower ice cream consumption??



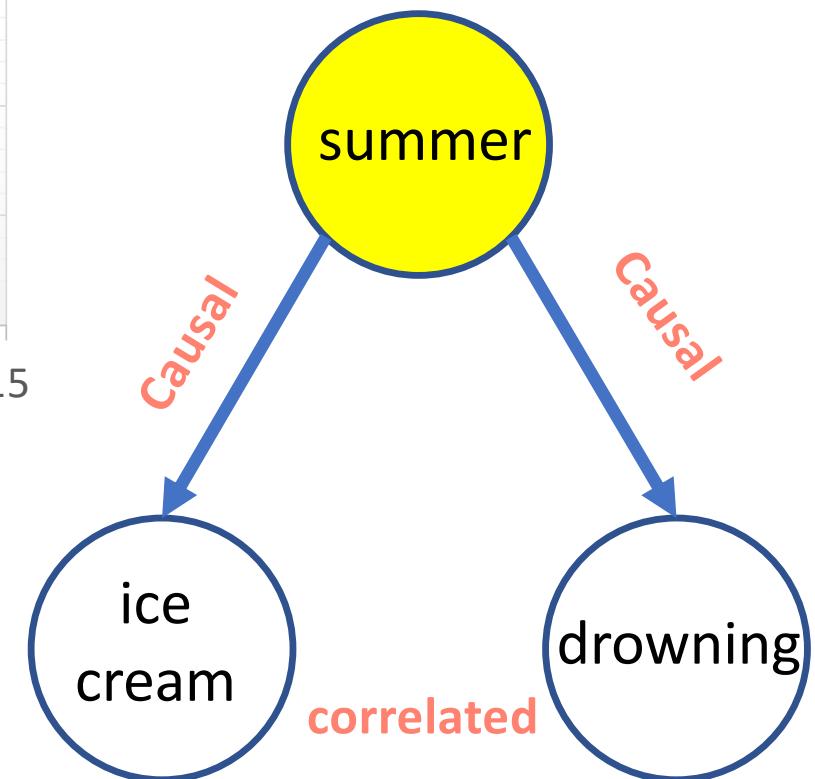
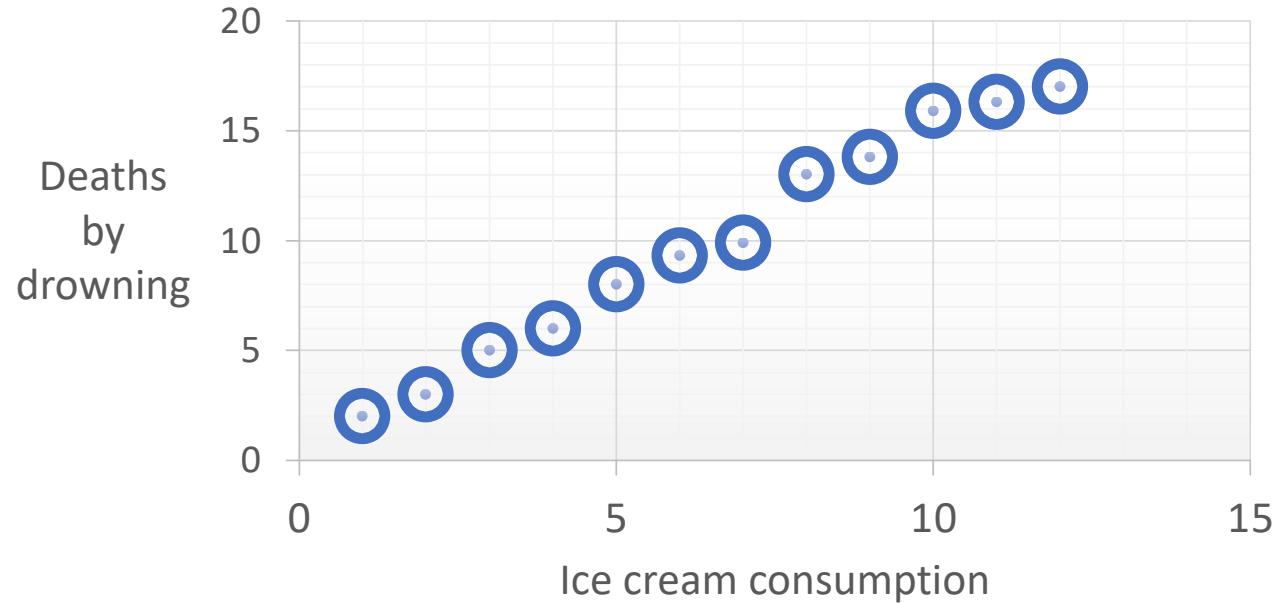
Thanks to
Lan Liu,
U. Minnesota

Confounding: is ice cream deadly?



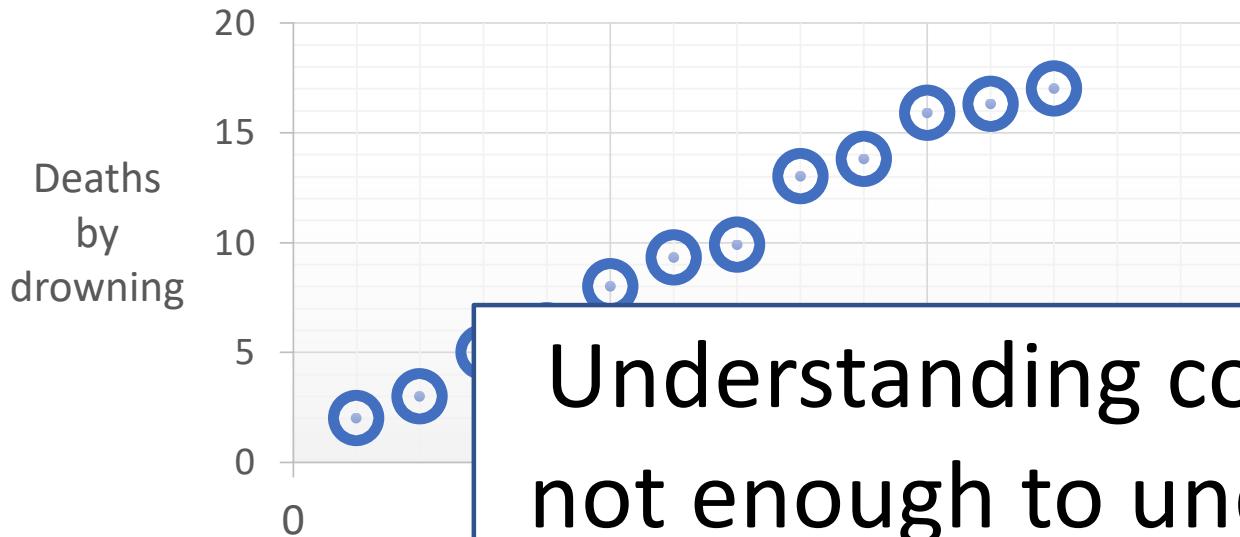
Thanks to
Lan Liu,
U. Minnesota

Confounding: is ice cream deadly?



Thanks to
Lan Liu,
U. Minnesota

Confounding: is ice cream deadly?



Understanding correlations is
not enough to understand the
effects of actions

summer

Causal

ice
cream

correlated

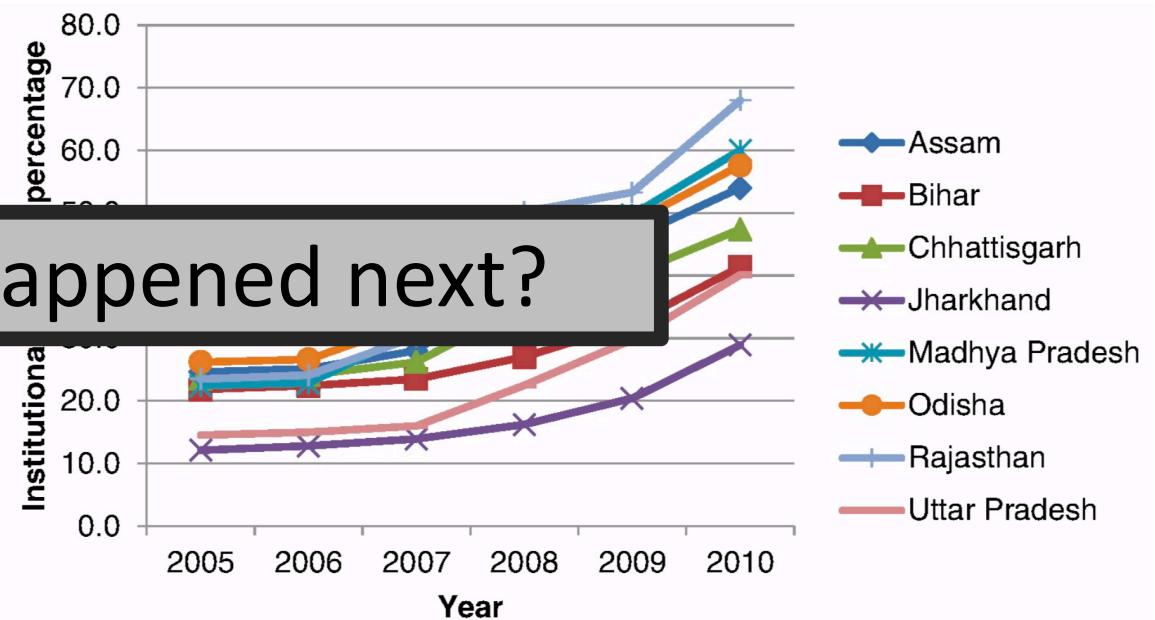
drowning

Institutional births in India

- High maternal mortality ratio
- Low institutional birth-rate

What happened next?

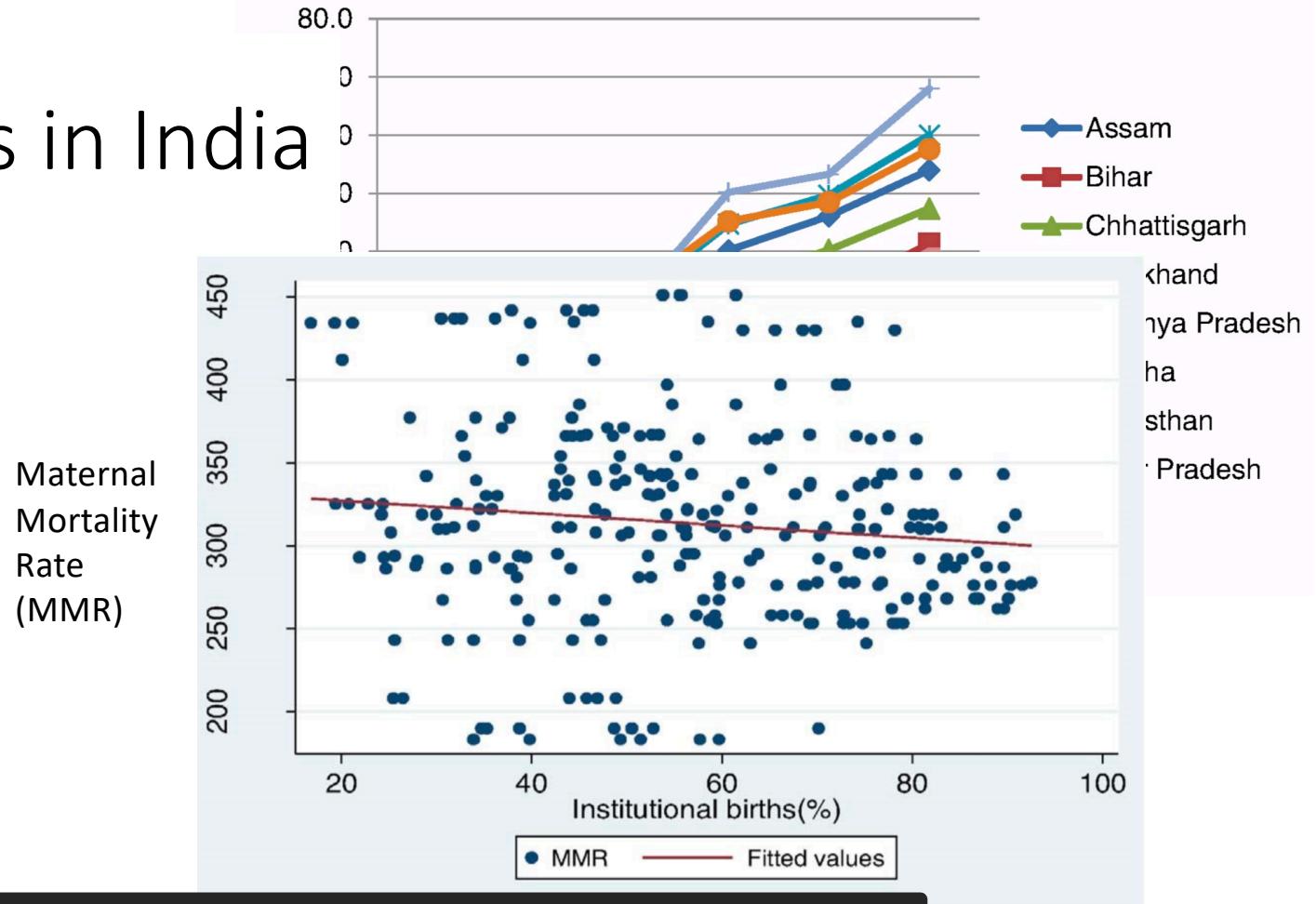
- Give cash to women who give birth in a medical institution: largest “conditional cash transfer” program in the world



(1) Randive B, Diwan V, De Costa A (2013). *India's Conditional Cash Transfer Programme (the JSY) to Promote Institutional Birth: Is There an Association between Institutional Birth Proportion and Maternal Mortality?*

(2) Ng M, Misra A, Diwan V, Agnani M, Levin-Rector A, De Costa A (2014). *An assessment of the impact of the JSY cash transfer program on maternal mortality reduction in Madhya Pradesh, India.*

Institutional births in India



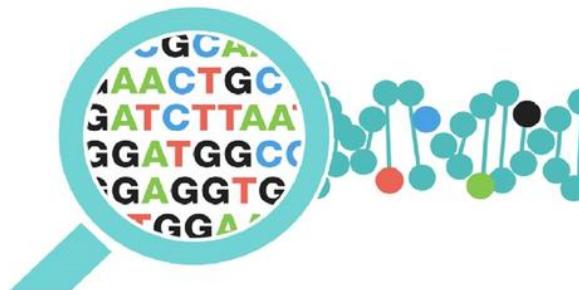
- (1) Randive B, Diwan V, De
Is There an Association bet
(2) Ng M, Misra A, Diwan V
mortality reduction in Mad

Does this mean the program
doesn't work?

nal Birth:
transfer program on materr

When supervised learning isn't enough

- Dataset of 10,000,000 patients
- Medications, blood tests, past diagnoses, doctors' notes, demographics, genetic testing

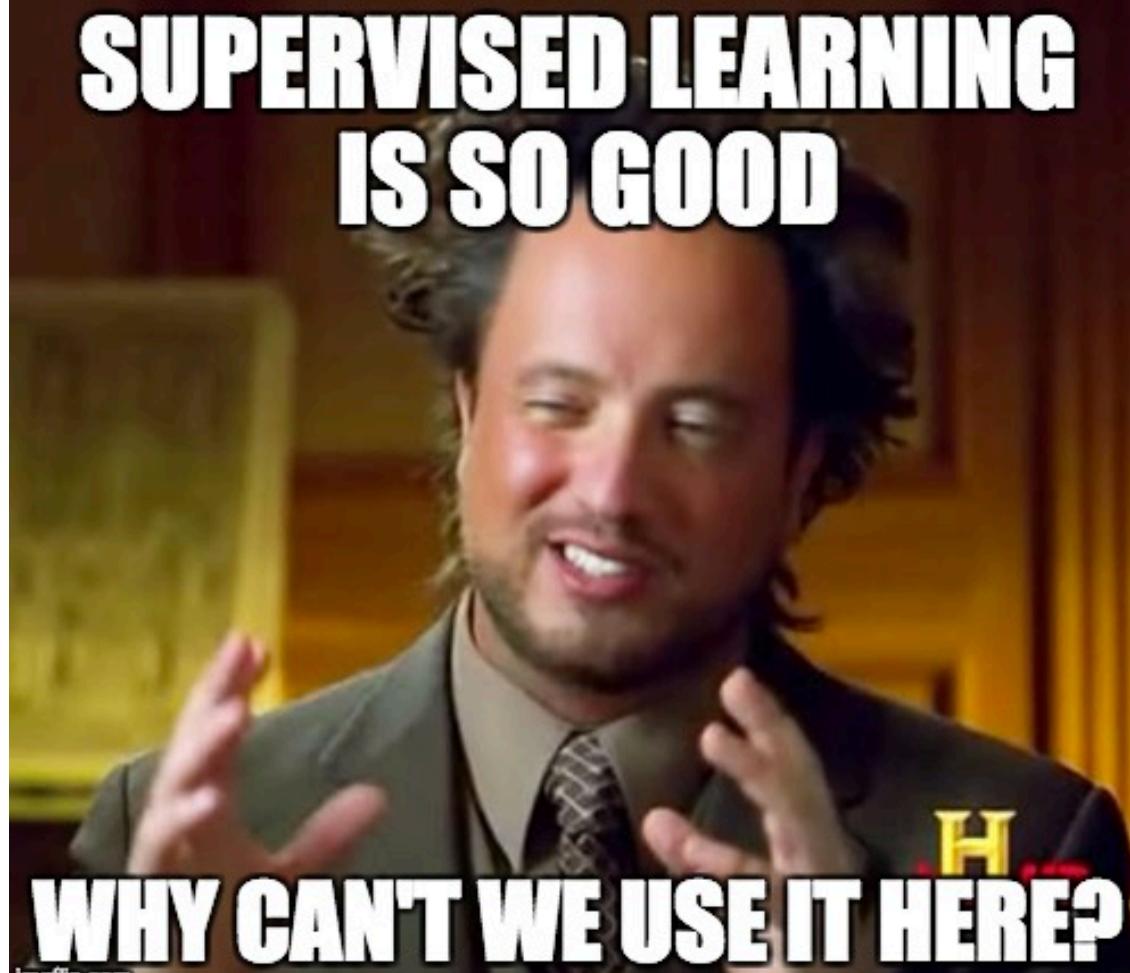


When supervised learning isn't enough

- Patient “Anna” comes in with hypertension
 - Asian, 54, history of diabetes, blood pressure 150/95, ...
- Which medication will better lower her blood pressure:
 - Calcium channel blocker (A)
 - ACE inhibitor (B)
- I have data from 10,000,000 other patients – surely that can help!



**SUPERVISED LEARNING
IS SO GOOD**



WHY CAN'T WE USE IT HERE?

SUPERVISED LEARNING IS SO GOOD

Why can't we just learn the connection between patient, treatment and outcome with supervised learning (say a neural net)?



When supervised learning isn't enough

- Patient “Anna” comes in with diabetes
 - Asian, 54, history of hypertension, blood pressure 150/95, ...
- Which medication will better lower her blood pressure:
 - DPP4 (A)
 - SGLT-2 (B)
- I have data from 10,000,000 other patients – surely that can help!



When supervised learning isn't enough

- Patient “Anna” comes in with diabetes
 - Asian, 54, history of hypertension, blood pressure 150/95, ...
- Which medication will better lower her blood pressure:
 - DPP4 (A)
 - SGLT-2 (B)
- I have data from 10,000,000 other patients – surely that can help!
- Build a regression model from patient features to blood pressure



When supervised learning isn't enough

- Build regression model from patient features to blood sugar

- Input:



Anna's features



Output:



A predicted blood sugar

—



Anna's features



B predicted blood sugar

=

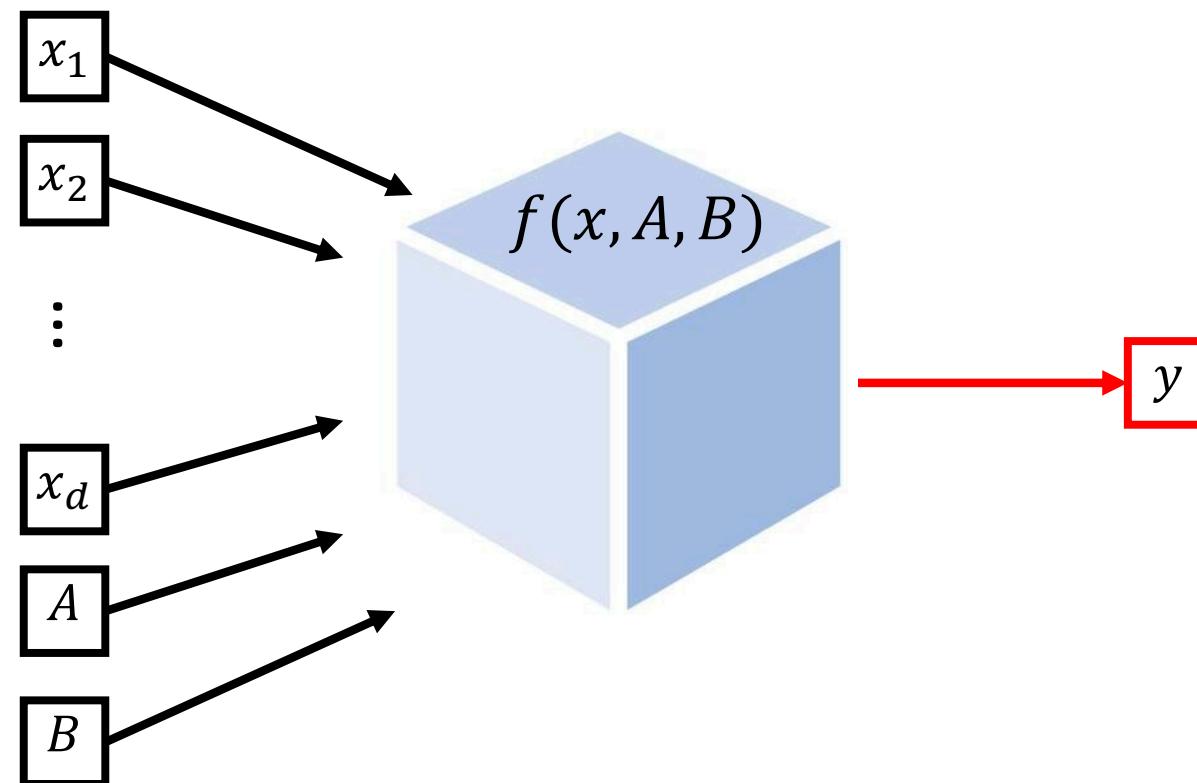
?

- Compare

Covariates
(Features)

Regression
model

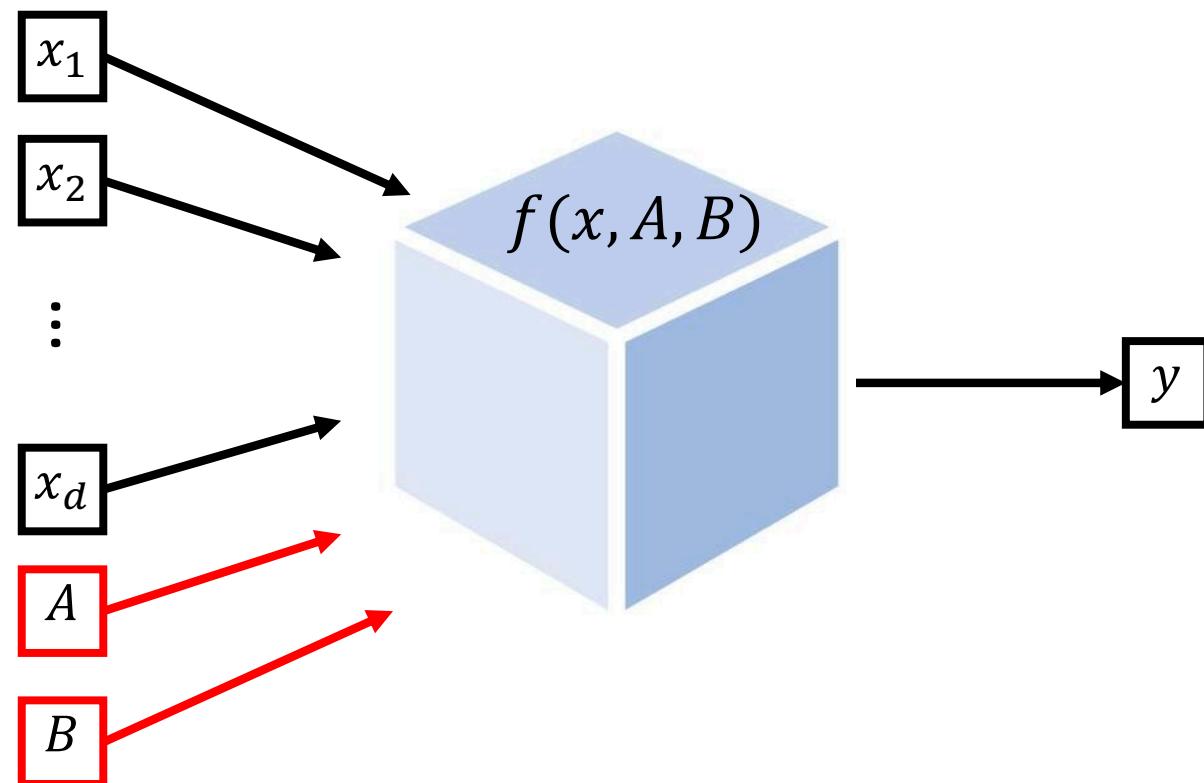
Outcome



Covariates
(Features)

Regression
model

Outcome



When supervised learning isn't enough

- This is not a classic supervised learning problem
- Our model was optimized to predict outcome, not to differentiate the influence of **A** vs. **B**
- What if our high-dimensional model threw away the feature of medication **A/B**?
- Hidden confounding:
Maybe using **B** is worse than **A**, but rich patients usually take **B** and richer people also have better health outcomes.
If we don't know whether a patient is rich or not, we might conclude **B** is better

When supervised learning isn't enough

- Hidden confounding:
Maybe using **B** is *worse* than **A**, but...

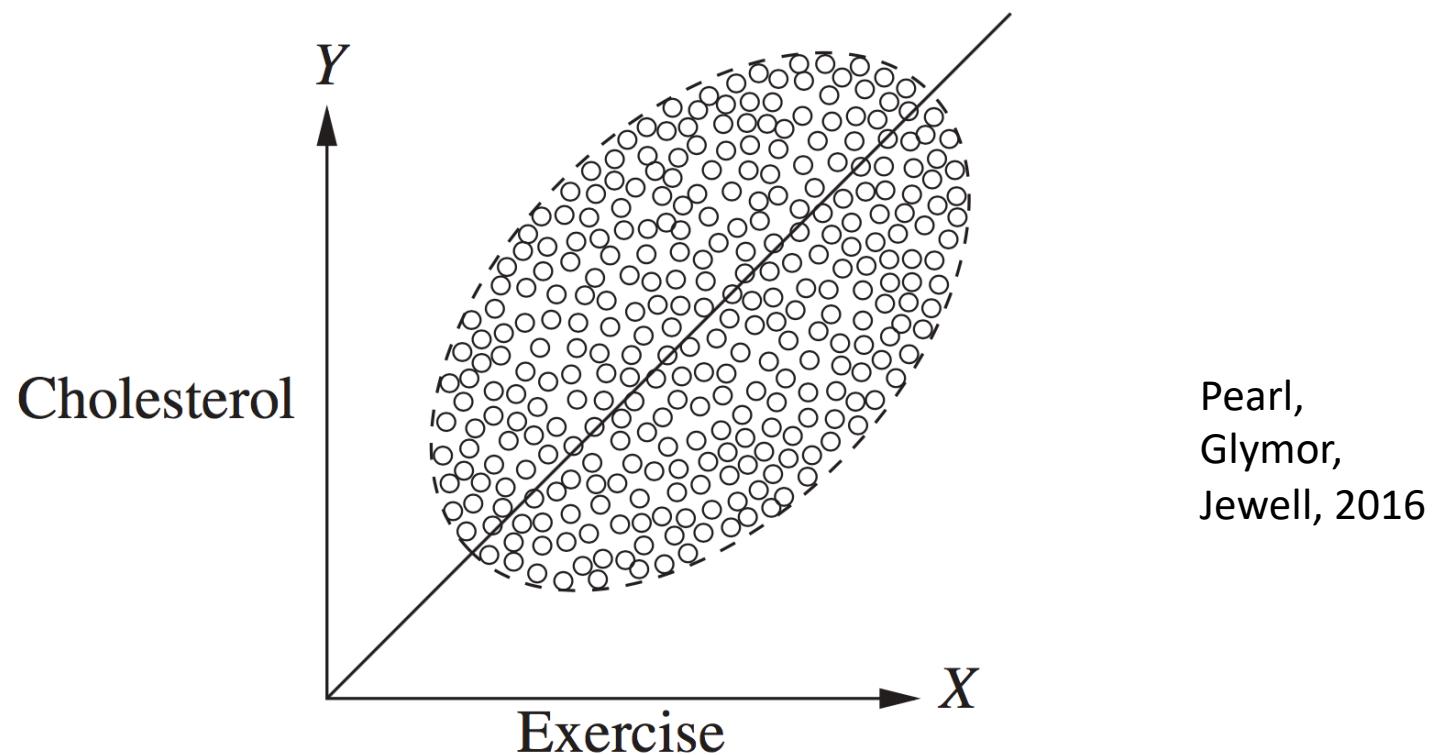
When supervised learning isn't enough

- Hidden confounding:
Maybe using B is worse than A , but...
rich patients usually take B ...

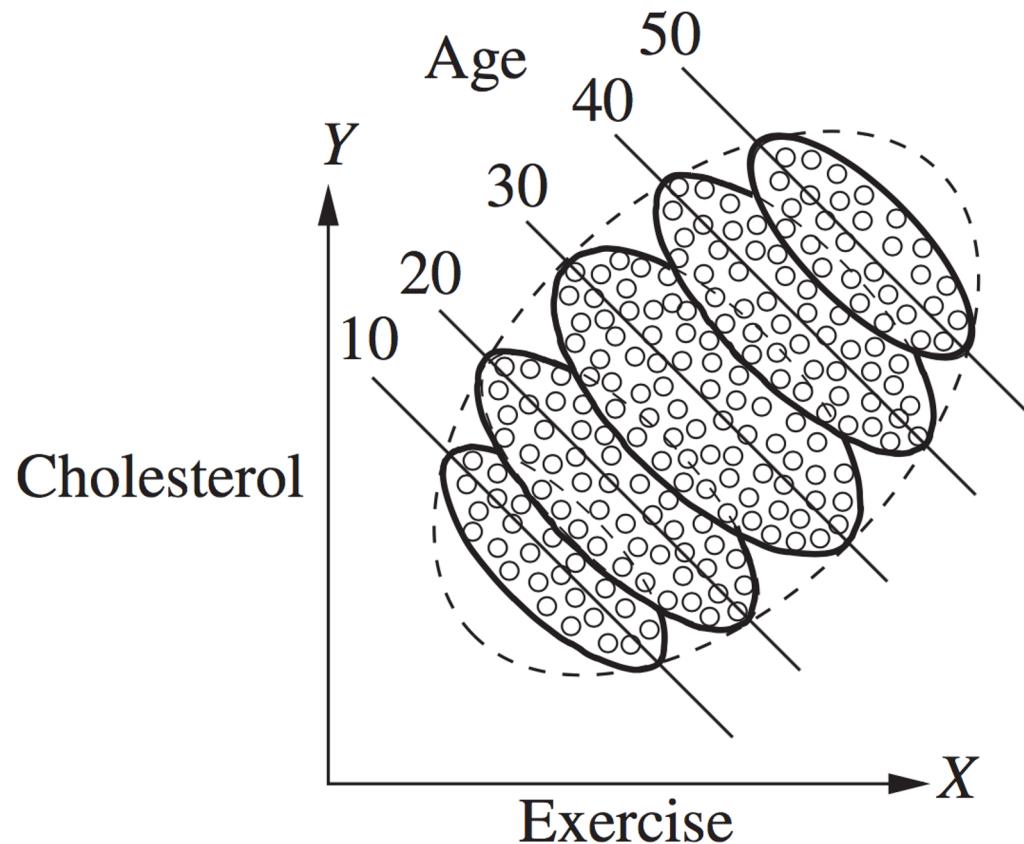
When supervised learning isn't enough

- Hidden confounding:
Maybe using ***B*** is worse than ***A***, but...
rich patients usually take ***B*** ...
therefore patients who took ***B*** exhibit better health outcomes.
- For prediction purposes this is the right thing to do:
B is telling you that the patient is probably richer and will probably be healthier
- But for optimal action this is bad

Does exercise raise your cholesterol?



Does exercise raise your cholesterol?



Pearl,
Glymour,
Jewell, 2016

OK, so how *can* we
determine the existence
and size of
causal effects?

Randomized controlled trials

- Give each patients medication A or medication B **completely at random**
- Observe which group has better outcomes
- And we're done!
- (We will prove next week why is this true)

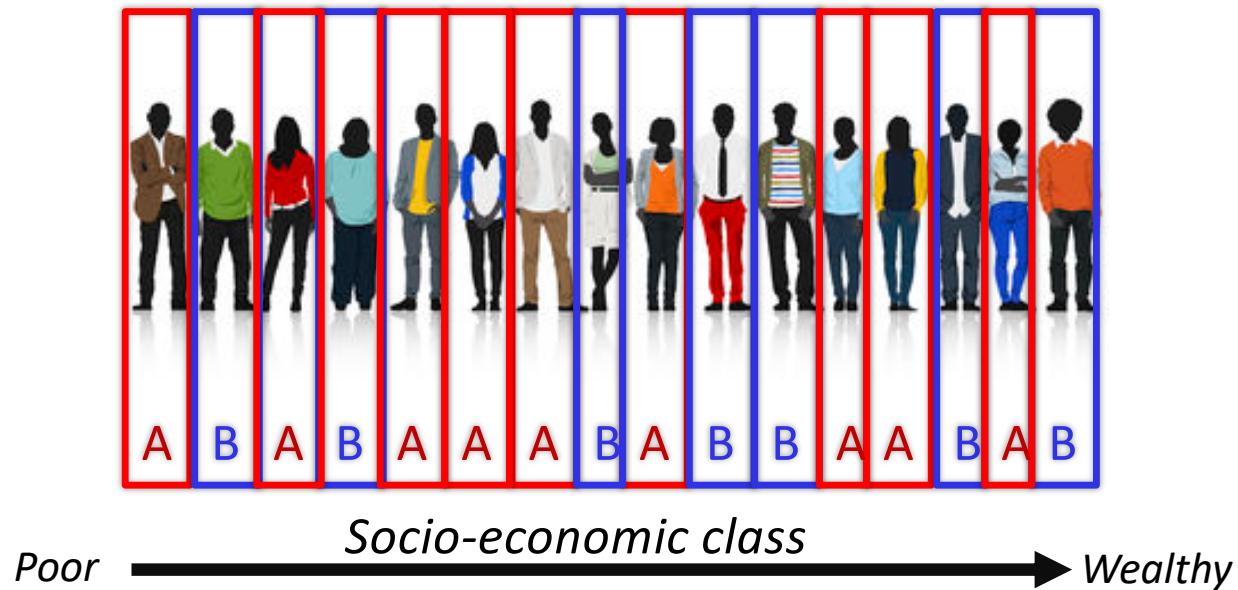
- Does inhaling Asbestos cause cancer?
- Does decreasing the interest rate reinvigorate the economy?
- We have a budget for **one new** anti-diabetic drug experiment. Can we use past health records of 100,000 diabetics to guide us?

Randomized trials vs. observational studies



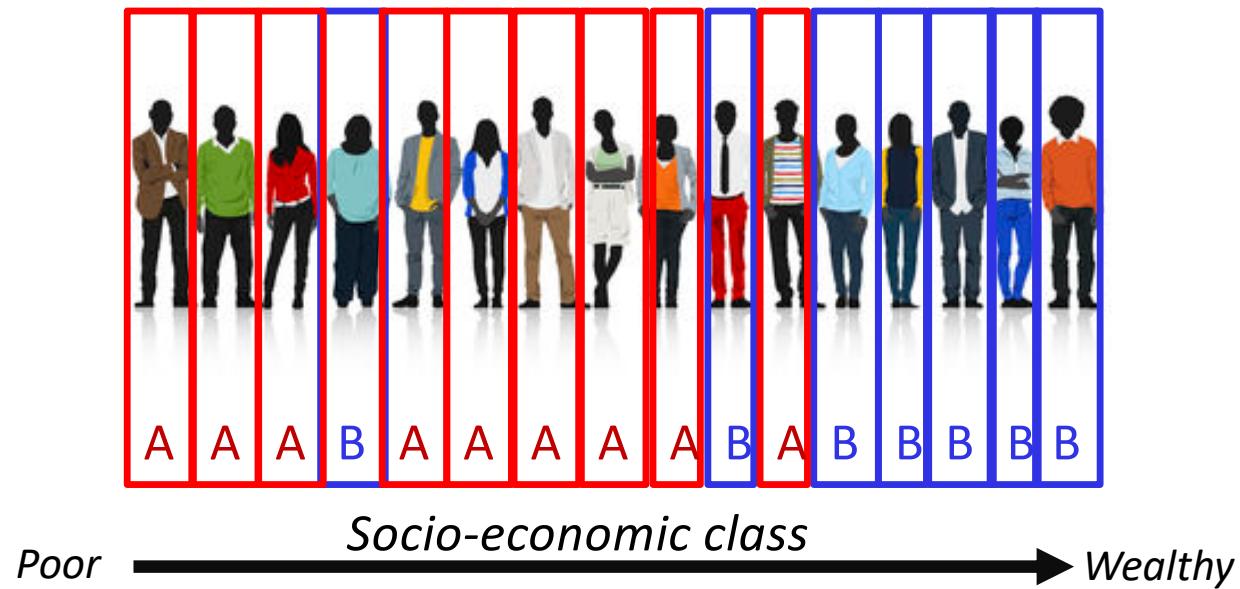
treatment
A or B

Randomized controlled trial (RCT)



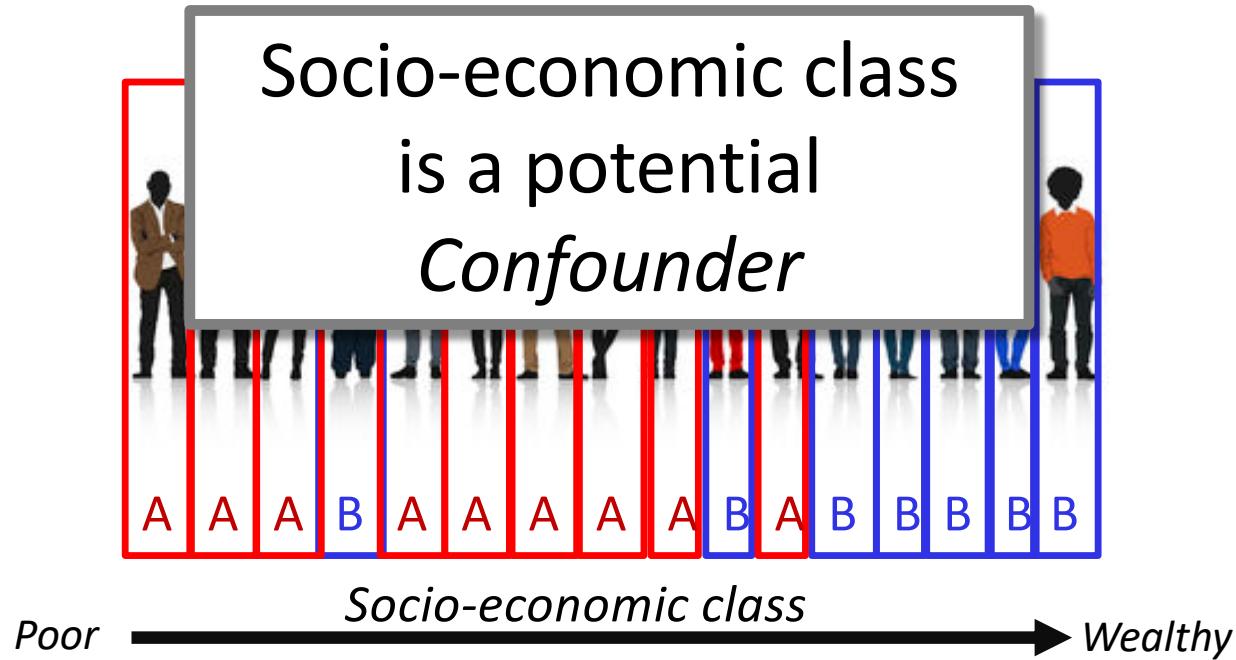
treatment
A or B

Observational study



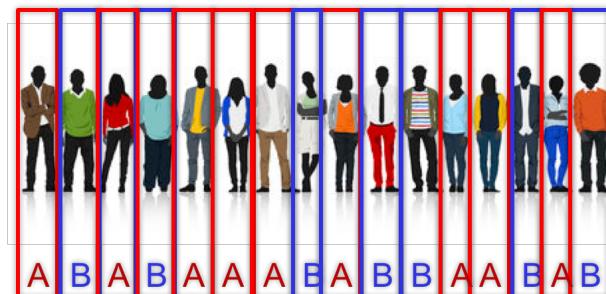
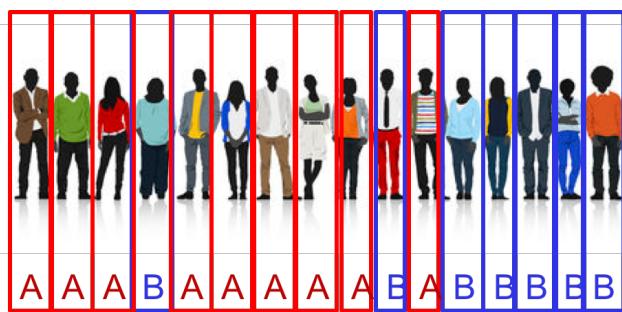
treatment
A or B

Observational study



treatment
A or B

In many fields randomized studies are
the gold standard for
causal inference, but...



Medical studies

- RCTs are also known as “clinical trials”
 - Tens of thousands every year, costing tens of billions of dollars
 - Every new medication must pass several stages of RCTs before approval for human use
- Observational study
 - Use data from Kupat Holim, tracking people’s medications and blood sugar
 - What are the possible confounders?

Economics

- RCTs:
 - Randomly give some people job training
 - Universal basic income RCTs in Kenya, Finland
- Observational study
 - Use data from government records to assess effect of job training
 - What are possible confounders?

Example:
Job training
Average Treatment Effect (ATE)



Should the government fund job-training programs?

- Existing job training programs seem to help unemployed and underemployed find better jobs
- Should the government fund such programs?
- Maybe training helps but only marginally? Is it worth the investment?
- **Average Treatment Effect (ATE)**
- Potential *confounder*: Maybe only motivated people go to job training? Maybe they would have found better jobs anyway?

Example:
Online courses



Do students who take online courses do better?

- MOOCs are becoming popular in some areas
- Do students learn better this way?
 - Might only help certain students, e.g. only strong ones
 - Might only help in certain subjects, e.g. intro courses
- Data from millions of students in MOOCs and ordinary classes
- Which study form generates better outcomes for a specific student-course combination?

**YOUR
AD
HERE**

CLICK NOW

225 x 675 | Side Bar
Every Page
excluding Home

Example:
Ad-placement

You Just Proved
Advertising Here Works!

YOUR

AD

HERE

Contact us at

[Click here for more information](#)

Ad-placement

- RCTs are often called “A/B testing” in this context
 - Choose ad randomly to present to users
- Observational study
 - Old algorithm placed ads by some criteria, how can we use logged click data to improve ad placement
 - What are the possible confounders?

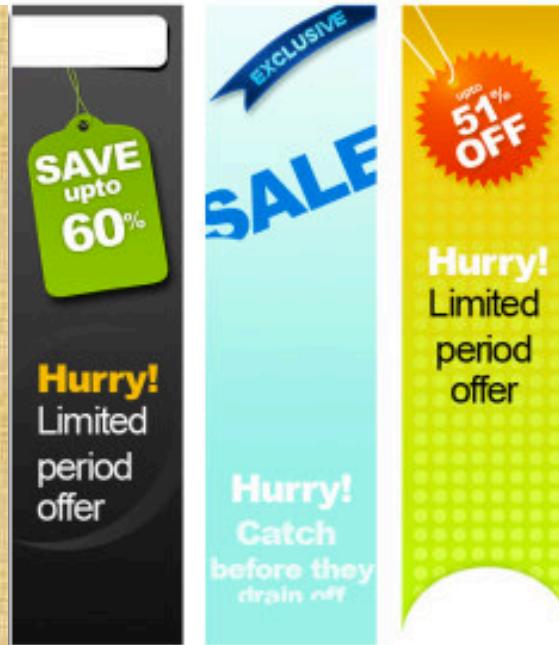
**Is smoking
dangerous?**



A photograph of a classroom filled with young children, likely preschool or kindergarten age. They are seated at several round tables, eating from white bowls. The room is brightly lit and decorated with educational materials, including a large bulletin board on the left and shelves in the background. The children are diverse in ethnicity and are dressed in casual clothing.

**Is early
childcare
beneficial for
children?**

Will running a marketing campaign increase sales?



Did a company discriminate against job applicants?



מחקר: ילדים קראו קומיקס ולמדו על מחלתם, מה שמעלה את סיכויים להבריא

חוקרים ישראלים מצאו כי ילדים שחולים בדלקת מפרקיים כרונית אשר קראו עליה בספרון קומיקס שיפורו משמעותית את הידע שלהם אודותיה, באופן שעשו להיטיב עם היענותם לטיפול

[ד"ר איתי גל](#)

מחקר: האם פופולריות בתיכון היא דבר חיובי?

מה עדיף - איכות או כמות? על פי מחקר חדש, חברות קרובות בתיכון מנबאת עליה יחסית בערך העצמי והפחיתה סימפטומים של חרדה ודיכאון, ואילו פופולריות ניבאה חרדה חברתית גבוהה יותר

לקראת שפעת: באיזו שעה ביום קיבל את החיסון לאפקטיביות מרבית?

חוקרים באנגליה מצאו כי רמת הנוגדים הייתה גבוהה יותר אצל מבוגרים שהתחסנו נגד שפעת בשעות הבוקר, לעומת אלה שהתחסנו בשעות אחר הצהרים. החוקרים ערכו לא ברור מה הוביל לשינוי בתגובהות

[ד"ר איתי גל](#) פורסם: 25.09.17 , 12:01



עד סיבת להוסף לשיקולים על
נימה רפואי Shutterstock
צילום: shutterstock

מחקרים רפואיים

לב וכלי דם • רפואי אלטרנטיבית • ידיים • מין וזוגיות • עישן • כשר • גיל הזהב • דיאטה ותזונה • סרטן

מחקר ענק: קיצור קיבה עשוי למניעת סרטן

חוקרים אמריקנים בדקו עשרות אלפי אנשים שסובלים מהשמנת יתר, חילקם עברו ניתוחים רפואיים, וגילו כי הנition הפחתת את סיכון התחלואה הסרטן בALTH. החוקרים: "הממצאים נוכנים בעיקר עבור נשים - תוצאות משמעותית מאוד נרשמו הסרטן השד ובسرطان הרחם"

[ד"ר איתי גל](#) פורסם: 08.10.17 , 09:52

מחקר ענק: האם ספורט אינטנסיבי עלול להיות מסוכן לב, עד כמה?



עודדו מפגשים עם חברים
ופעלויות חוץ
צילום: shutterstock

חוקרים ממקון הלב של מיניאפוליס בארה"ב מצאו שתחרות טרייאתлон עלולה להשפיע על

[רון קורנובסקי](#) | 10.10.2017 , 57

מחקר: תזונה מן הצומח בריאה יותר? תלוי מה אוכלים

חוקרים מהרווארד מצאו כי צריכת מזונות צמחניים רפואיים מפחיתה את הסיכון למחלת לב, בעוד צריכת מזונות לא רפואיים שכאלו מעלה אותו

[ג'ין ברודி](#), ניו יורק טימס | 16.10.2017 , 41

שיר

Observational studies are hard

- It is provably impossible to infer causal effects from an *observational* study without making strong assumptions about the data generating process
- Some of these assumptions are *unverifiable from data*

These are real-world problems!

- Observational studies with >30,000 women: Hormone Replacement Therapy (HRT) for post-menopausal women leads to **decreased** coronary heart disease
Stampfer & Colditz, *Preventive Medicine* (1991)
- Follow-up randomized controlled trials: HRT **increased** risk for heart disease and breast cancer
Million Women Study Collaborators, *The Lancet* (2003)
- There are also success stories – will see later in the course



Even randomized controlled trials have flaws

- Study population might not represent true population
 - Recruiting is hard
 - Study in one company/hospital/state/country could fail to generalize to others
- Often some people drop out of a study
 - Dropping out might be related to the treatment - “it gave me a headache so I stopped taking it”
 - This creates bias

In this course

- We will learn how to think formally and rigorously about causal questions
 - What is the mathematical language needed to prove that RCTs can determine causality?
- We will learn when and how can we answer (some) causal questions without a randomized controlled trial
 - What to do when you can't run an RCT?
- How can we answer questions such as determining individual-level effects?
- What other interesting and important causal questions can we answer from data?
- How can causal ideas help us address classic machine learning problems such as transfer learning?

Mathematical Foundations: (At least) Two ways to model causality

- Potential outcomes
(mostly identified with Rubin and Neyman)
- Causal graphs
(mostly identified with Judea Pearl)

Two major “schools” formalizing causal inference

- Potential outcomes (Rubin, Neyman)
- Causal graphs, do-calculus (Pearl)

In both formalisms we need two phases

Two major “schools” formalizing causal inference

- Potential outcomes (Rubin, Neyman)
- Causal graphs, do-calculus (Pearl)

In both formalisms we need two phases

1. Formulate *causal assumptions* sufficient to solve the problem
 - these are mostly **untestable**

Two major “schools” formalizing causal inference

- Potential outcomes (Rubin, Neyman)
- Causal graphs, do-calculus (Pearl)

In both formalisms we need two phases

1. Formulate *causal assumptions* sufficient to solve the problem
 - these are mostly **untestable**
2. Under the assumptions, reduce causal problem to appropriate statistical/machine learning method
 - these methods are often specialized methods, similar but distinct from familiar methods such as regression

Two major “schools” formalizing causal inference

- Potential outcomes (Rubin, Neyman)
- Causal graphs, do-calculus (Pearl)

Each formalism has its pros and cons, we will go over both

Potential outcomes

- Unit: a person, a bacteria, a company, a school, a website, a family, a piece of metal, ...
- Treatments / actions / interventions
- Potential outcomes

Y_1 : the unit's outcome *had they been subjected to treatment t=1*

Y_0 : the unit's outcome *had they been subjected to treatment t=0*

If number of treatments is T, we have T potential outcomes (T possibly infinite)

Counterfactuals and causal inference

- Does treatment T cause outcome Y ?
- If T had not occurred, Y would not have occurred (David Hume)
- Counterfactuals:
Kim received job training (T), and her income one year later (Y) is 20,000\$
What would have been Kim's income had she not had job training?

Counterfactuals and causal inference

- Counterfactuals:
Kim received job training (T), and her income one year later (Y) is \$20,000
What would have been Kim's income had she not had job training?
- If her income would have been \$18,000, we say that job training caused an increase of \$2,000 in Kim's income
- The problem: you never know what might have been

Sliding Doors



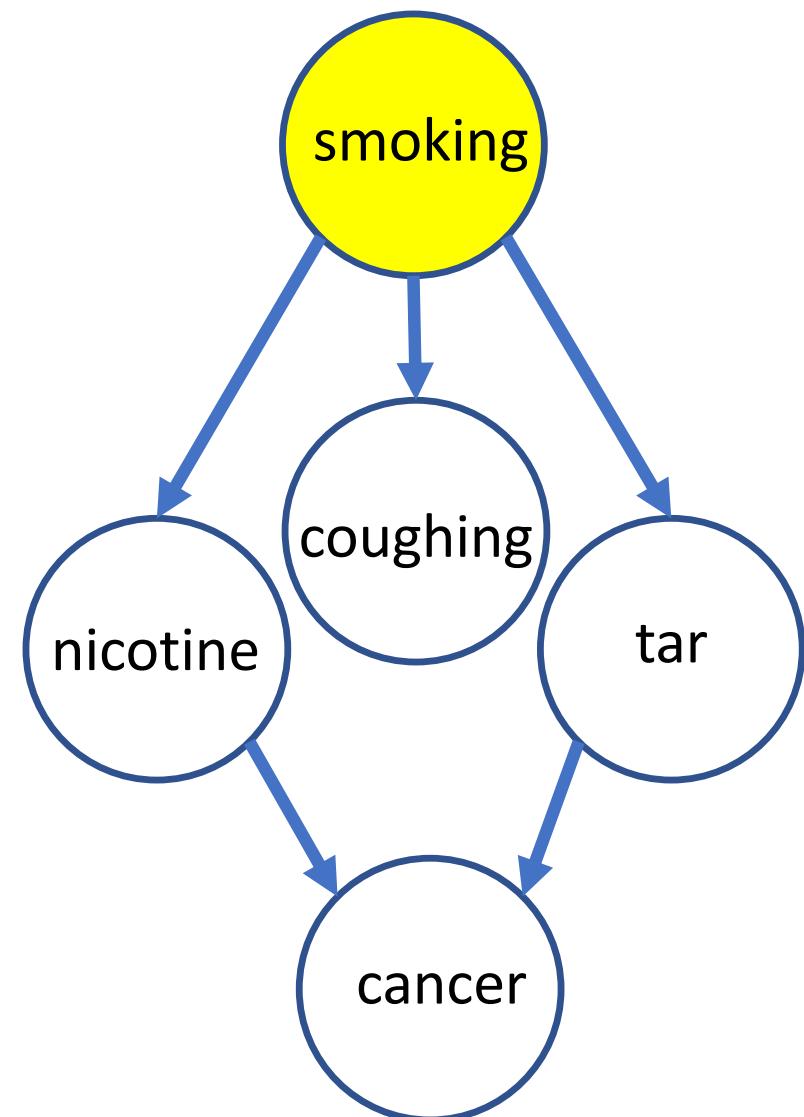
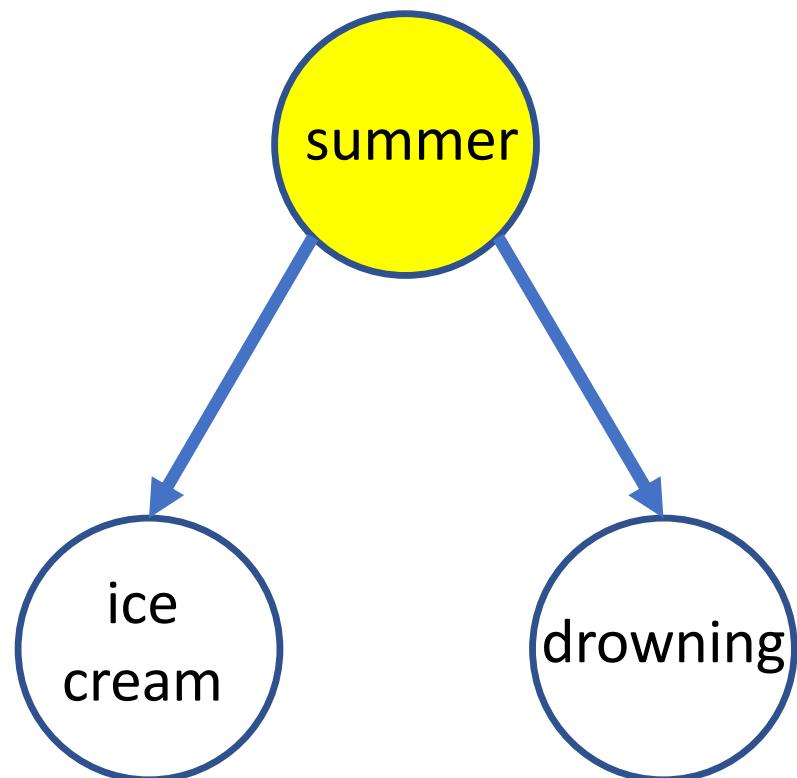
Causal graphs

- Based on the idea of Probabilistic Graphical Models
- Causal graphs strongly identified with the work of Judea Pearl

Probabilistic graphical models and causal graphs

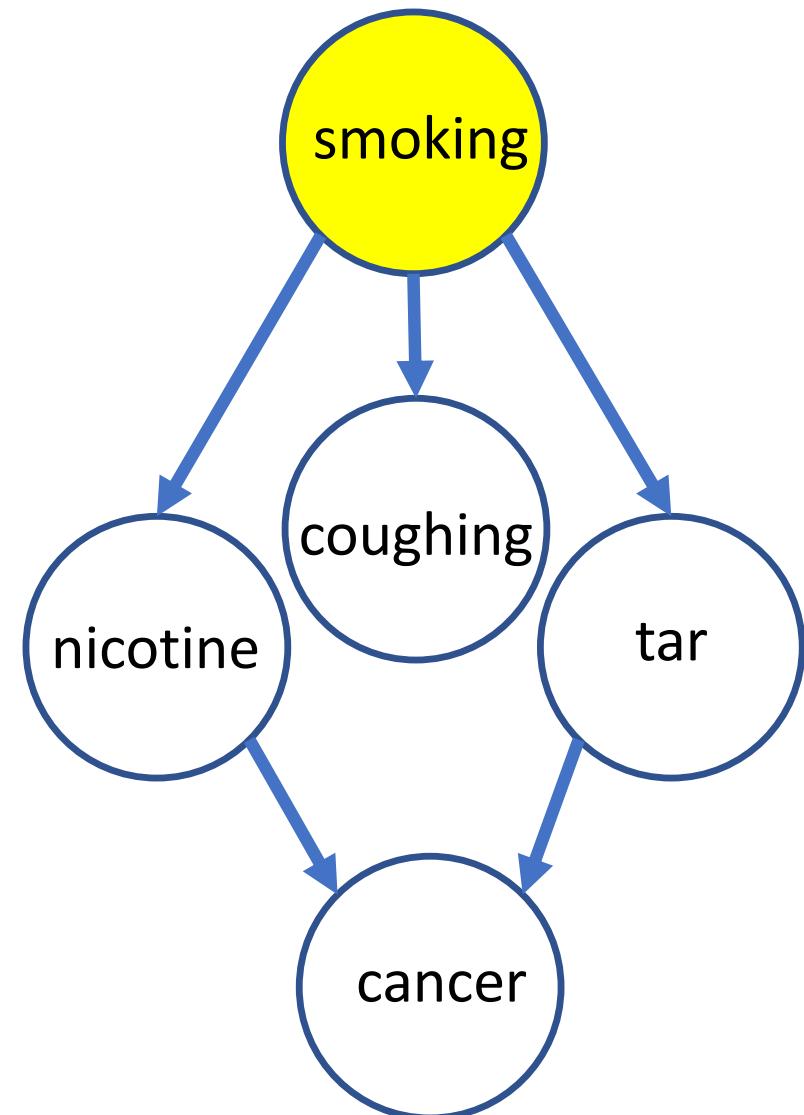
- Probabilistic graphical models are a compact and useful means for describing probability distributions
- Causal graphs extend to describing the effects of actions (interventions), through an idea called the *do*-operator

Causal graphs



Causal graphs

- What would be the effect on cancer if we eliminate tar?
- What would be the effect on cancer if we eliminate coughing?



What kind of questions are causal questions we'll be most interested in?

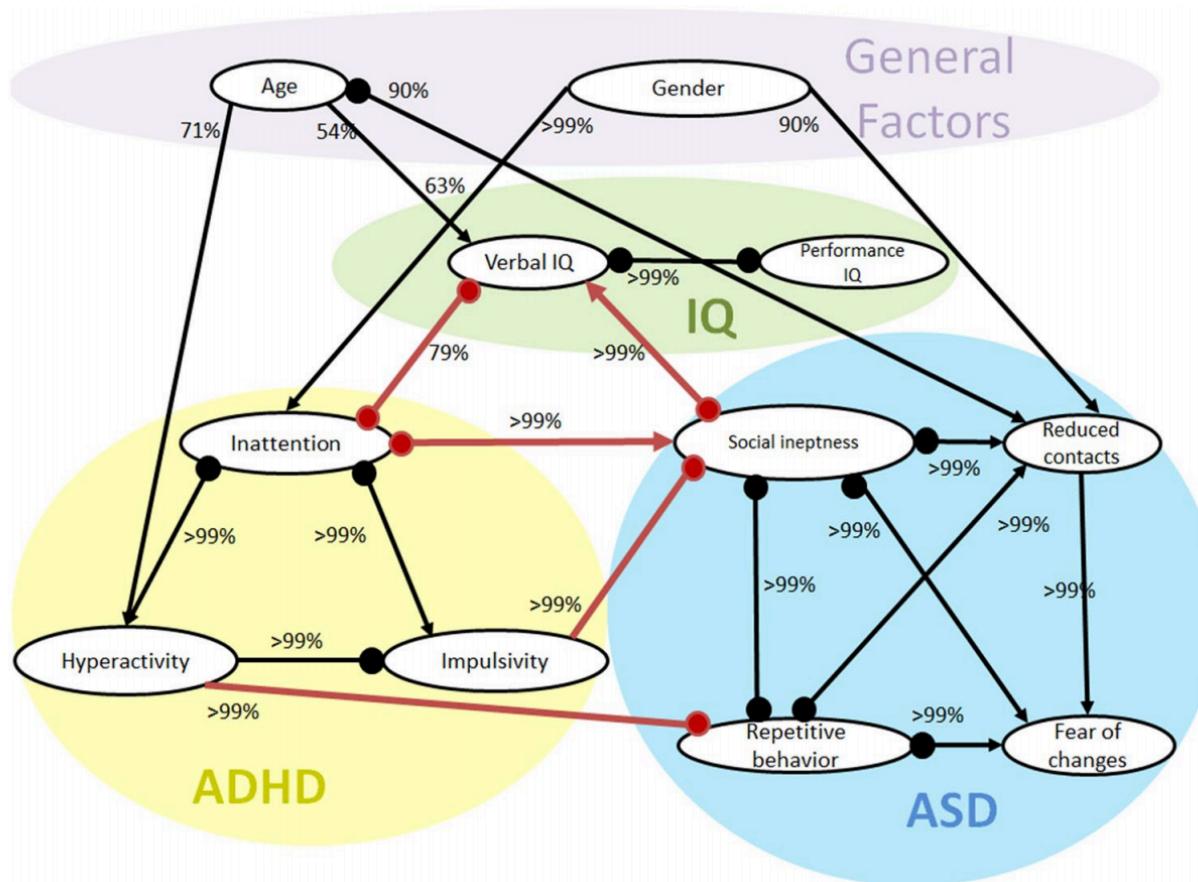
- Average treatment effects:
Which action (e.g. medication) is better *on average*
- Individual-level treatment effects:
Which action (e.g. medication) is better *for me*
- Identify heterogeneously responding subgroups

Other causal questions

- Causal mediation:
- We know that smoking causes cancer, but through which path?
smoking → tar → cancer
smoking → nicotine → cancer
- We know that university education causes better careers, but is it:
university → skills → good_job
university → prestige → good_job
university → networking → good_job

Causal discovery - structure

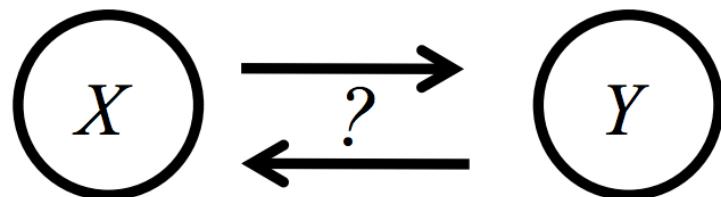
- Given multiple variable data, what is the underlying causal structure?



A Causal and Mediation Analysis
of the Comorbidity Between Attention
Deficit Hyperactivity Disorder (ADHD)
and Autism Spectrum Disorder (ASD),
Sokolova et al. (2017)

Causal discovery - direction

- Given two variables, which caused which?



Bernhard Schölkopf

- Example:
Does social isolation cause drug abuse, or does drug abuse cause social isolation?

Course outline (1-4)

1. What problems are causal inference problems?
What is causal inference from observational studies?
How is it related to supervised learning, and why is it harder?
(this lesson)
2. Conceptual foundations (I) and practical methods (I):
Rubin-Neyman potential outcomes
What are counterfactuals? Some philosophical background
The target trial. Identifiability.
Covariate adjustment.
3. Practical methods (II):
Covariate adjustment, propensity score, matching
The target trial paradigm.
4. Case studies: Successes and failures of practical methods.

Course outline (5-8)

5. Conceptual foundations (II):
Graphical models, Simpson's paradox. Pearl's causal graphs, do-calculus, the backdoor criterion,.
6. Conceptual foundations (III):
Causal graphs (cont.).
Identification by causal graphs, front door criterion.
Structural equation models
7. Natural experiments:
Instrumental variables. Regression discontinuity.
8. Individual-level treatment effects; modern methods.

Course outline (9-12)

9. How to run an observational study?
Target trial redux. Sensitivity analysis.
10. Where do causal graphs come from? Causal discovery. Connections to reinforcement learning and bandit problems.
11. Using causal ideas for better machine learning: Robust machine learning. Covariate shift, explainability.
12. Middle-of-project presentations

Course project: option 1

- Answer causal questions from data using methods introduced in the course or from the literature
- Which data?
 - Bring your own, or...
 - Find in public records e.g. <https://data.gov.il>, or...
 - I have datasets, lots of them

Course project: option 2

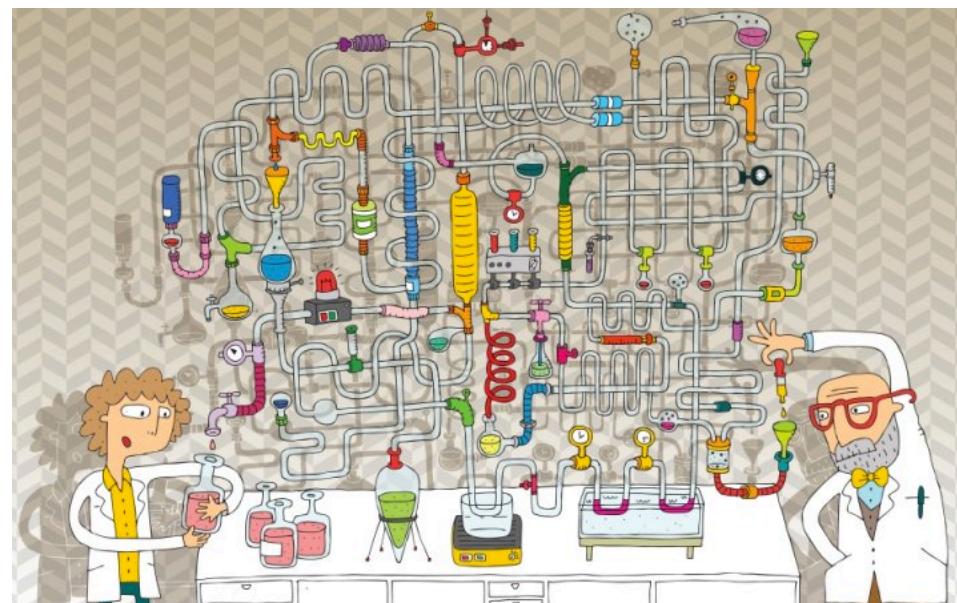
- Attempt to beat the state-of-the-art in causal inference data benchmarks
 - Atlantic causal inference challenges
 - Criteo data challenge
 - ChaLearn

Course project: option 3

- Implement and improve a modern causal inference method from recent literature, including recently introduced deep learning approaches

Course project: option 4

- Suggest your own causal inference project!



Examples of projects from previous years

- The Causal Effect of a Student's Rank on Effort
(Daphna Lipschitz & Dafna Schumacher)
- Semi-supervised Causal Inference
(Shahar Harel)
- What is the Effect of Donald Trump's Sentiment Online?
(Gal Yankovitz & Tomer Levy)
- The Effect of Mental State on Decisions: Evidence from the Language of NBA Players
(Nadav Oved & Amir Feder)
- Estimation of the causal effect of Cesarean section delivery on long term pediatric outcomes
(Ayya Keshet)

Project

- Presentation in January (10% of grade)
- Submission deadline: March 2021