Machine Learning for Healthcare

Intro to Causal Inference + why it is important in ML for Health

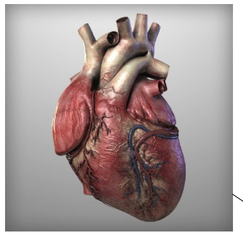
04/27/2023

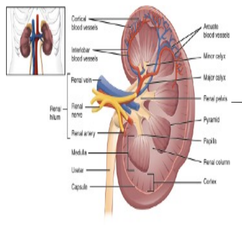
Slides inspired and adapted from David Sontag's excellent ML for healthcare course

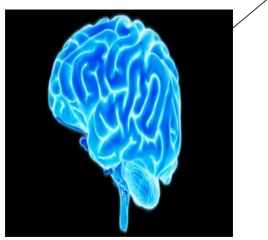
Machine Learning for Healthcare

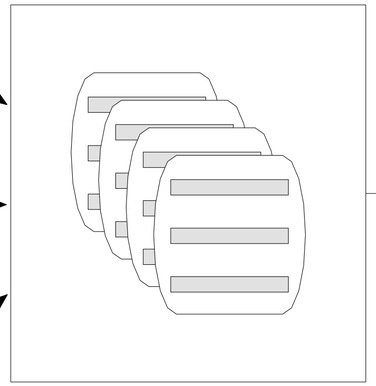


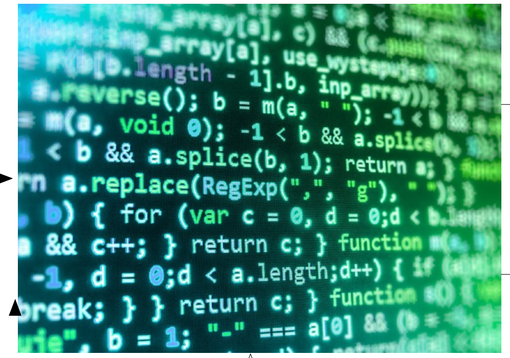
Interventions in Machine Learning for Healthcare

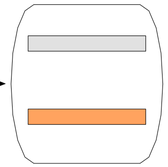










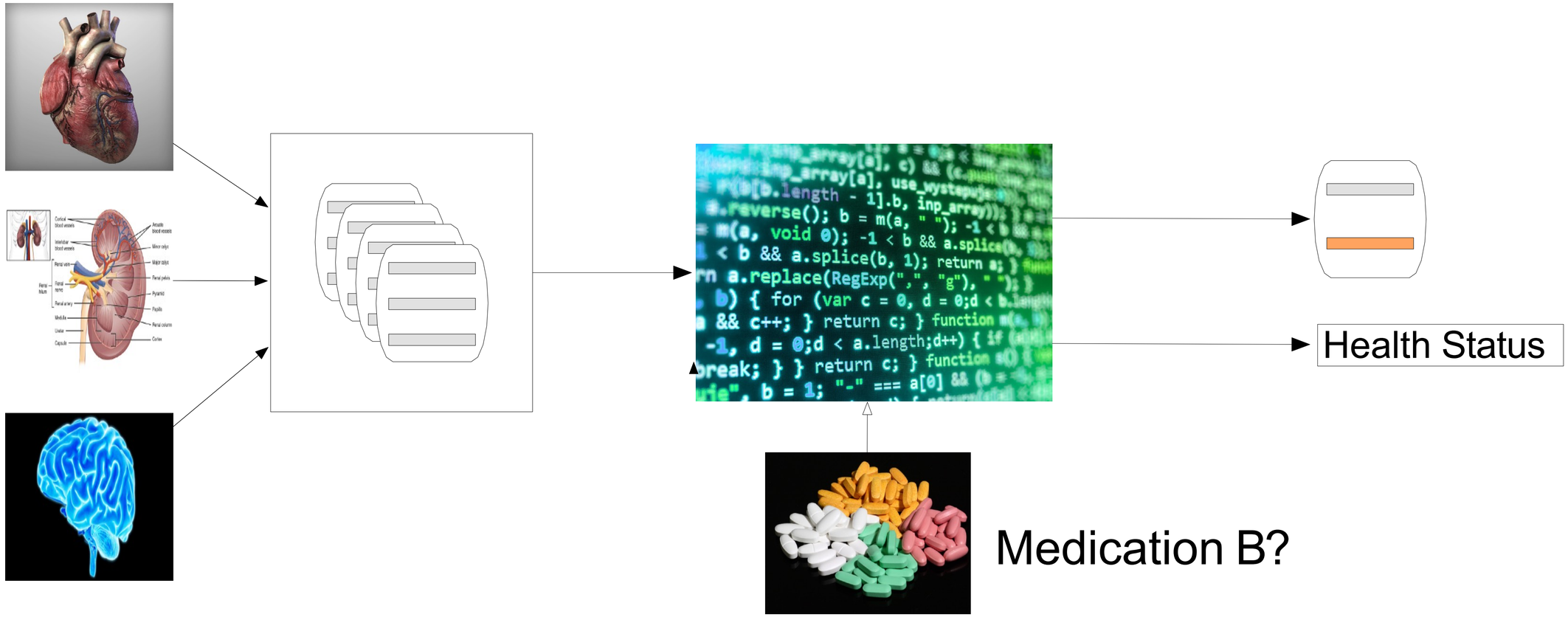


Health Status



Medication A?

Interventions in Machine Learning for Healthcare



Question: Should we give medication A or medication B?

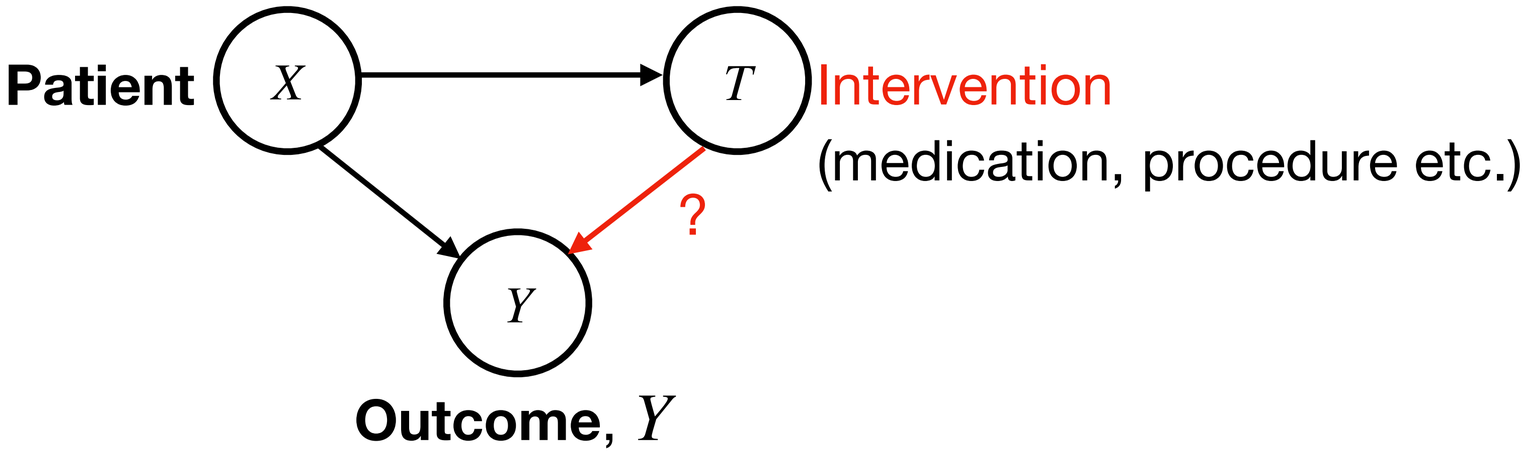
Causal Inference

Causal inference seeks to estimate the effect of an intervention

- Which statin should a patient with hyperlipidemia take?
- Which medication should be given to a depressed person?
- Which patients should hospice care be given to?

Solutions?

- Find important features to a target in survival analysis?
  - Why should this work?
- What about other interventions after feature collection?
- To treat these issues formally we need a **mathematical model for causal inference**



Potential Outcomes (Rubin-Neyman Causal Model)

- **Definition:** Each unit (individual) $\mathbf{x}_i$ has two potential outcomes
  - $Y_0(\mathbf{x}_i)$ is the potential outcome had the unit not been treated: *"control outcome"*
  - $Y_1(\mathbf{x}_i)$ is the potential outcome had the unit been treated: *"treated outcome"*
- Example:
  - $Y_0(\mathbf{x}_i)$ is PHQ-9 score after 2 months of SSRI treatment for depression
  - $Y_1(\mathbf{x}_i)$ is PHQ-9 score after 2 months of cognitive behavioral therapy for depression

Potential Outcomes (Rubin-Neyman Causal Model)

- Causal effects are functions of the potential outcomes
  - Is treatment 0 better than treatment 1 for patient i: $Y_0(\mathbf{x}_i) > Y_1(\mathbf{x}_i)$
  - What's the difference in outcomes between the treatments: $Y_0(\mathbf{x}_i) - Y_1(\mathbf{x}_i)$
- Conditional Average Treatment Effect for unit i:

$$CATE(\mathbf{x}_i) = \mathbb{E}_{Y_1 \sim p(Y_1|\mathbf{x}_i)}\left[Y_1 \mid \mathbf{x}_i\right] - \mathbb{E}_{Y_0 \sim p(Y_0|\mathbf{x}_i)}\left[Y_0 \mid \mathbf{x}_i\right]$$

- Average Treatment Effect:

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

Problem Solved?

- Equipped with the definition of potential outcomes, let us propose an estimation algorithm.

Algorithm 1: Causal Inference

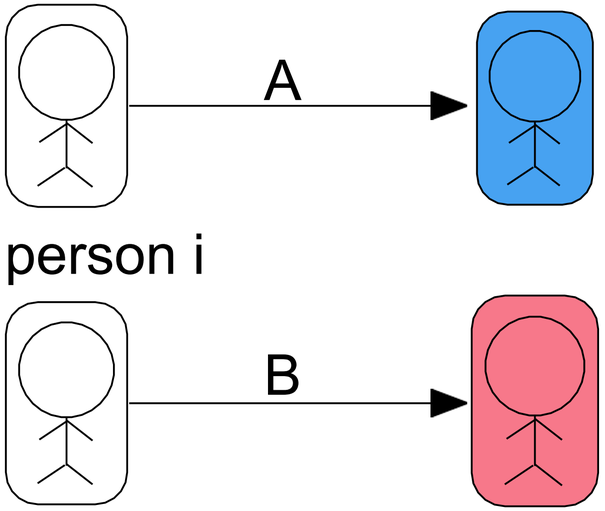
- **Input:** $\{Y_0(\mathbf{x}_i), Y_1(\mathbf{x}_i), \mathbf{x}_i\}_{i=1}^{N}$
  - Features of training unit i: $\mathbf{x}_i$
  - Potential outcomes of training unit i: $Y_0(\mathbf{x}_i), Y_1(\mathbf{x}_i)$
- **Train** a model $\hat{f}(\mathbf{x}) = [\hat{Y}_0(\mathbf{x}), \hat{Y}_1(\mathbf{x})]$ to predict $Y_0(\mathbf{x}), Y_1(\mathbf{x})$
- **At test time** predict and calculate effect, e.g. $\hat{Y}_0(\mathbf{x}) - \hat{Y}_1(\mathbf{x})$

Problem Solved?

- Equipped with the definition of potential outcomes, let us propose an estimation algorithm.

Algorithm 1: Causal Inference

- **Input:** $\{Y_0(\mathbf{x}_i), Y_1(\mathbf{x}_i), \mathbf{x}_i\}_{i=1}^{N}$
  - Features of training unit i: $\mathbf{x}_i$
  - **Potential** outcomes of training unit i: $Y_0(\mathbf{x}_i), Y_1(\mathbf{x}_i)$
- **Train** a model $\hat{f}(\mathbf{x}) = [\hat{Y}_0(\mathbf{x}), \hat{Y}_1(\mathbf{x})]$ to predict $Y_0(\mathbf{x}), Y_1(\mathbf{x})$
- **At test time** predict and calculate effect, e.g. $\hat{Y}_0(\mathbf{x}) - \hat{Y}_1(\mathbf{x})$



- **The fundamental problem of causal inference:** we cannot observe both outcomes!

Some Terminology

- The observed quantities are:
  - Treatment: t_i
  - Features: $\mathbf{x}_i$
  - Observed factual outcome: $y_i = t_i Y_1(\mathbf{x}_i) + (1 - t_i) Y_0(\mathbf{x}_i)$
- *Un*observed *Counter*factaul outcome: $y_i^{CF} = (1 - t_i) Y_1(\mathbf{x}_i) + t_i Y_0(\mathbf{x}_i)$

Estimating the ATE

- Assume we have N observed units, $\{\mathbf{x}_i, t_i, y_i\}_{i=1}^{N}$, how should we estimate the ATE?

$$ATE = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} \left[Y_1(\mathbf{x}) - Y_0(\mathbf{x}) \right]$$

- **Naïve solution:** difference between average outcome on *treated* and *untreated*

$$\widehat{ATE} = \frac{1}{N_1} \sum_{i:t_i=1} y_i - \frac{1}{N_0} \sum_{j:t_j=0} y_j$$

$$N_1 = \sum_i t_i, \; N_0 = \sum_i 1 - t_i$$

(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

(Example from Uri Shalit)

(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

(Example from Uri Shalit)

(age, gender, exercise)	Y_0: Sugar levels *had they received* medication A	Y_1: Sugar levels *had they received* 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

(Example from Uri Shalit)

(age,gender, exercise)	Sugar levels *had they received* medication A	Sugar levels *had they received* 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

mean(sugar|medication B) –
mean(sugar|medicaton A) =
?

mean(sugar|*had they received* B) –
mean(sugar|*had they received* A) =
?

(Example from Uri Shalit)

(age,gender, exercise)	Sugar levels *had they received* medication A	Sugar levels *had they received* 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

mean(sugar|medication B) –
mean(sugar|medicaton A) =
7.875 - 7.125 = 0.75

mean(sugar|*had they received* B) –
mean(sugar|*had they received* A) =
7.125 - 7.875 = -0.75

(Example from Uri Shalit)

Randomization

- Assume we have N observed units, $\{\mathbf{x}_i, t_i, y_i\}_{i=1}^{N}$, how should we estimate the ATE?
- **Naïve solution:** difference in average outcome between *treated* and *untreated*

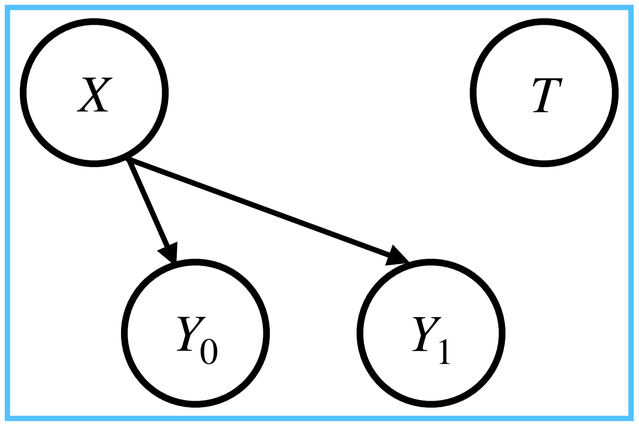
$$\widehat{ATE} = \frac{1}{N_1} \sum_{i:t_i=1} y_i - \frac{1}{N_0} \sum_{j:t_j=0} y_j$$

- When will this work?

  Randomized Control Trial!

$$\begin{aligned} ATE &= \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[Y_1(\mathbf{x}) - Y_0(\mathbf{x})\right] \\ &= \mathbb{E}_{\mathbf{x}}\left[Y_1(\mathbf{x})\right] - \mathbb{E}_{\mathbf{x}}\left[Y_0(\mathbf{x})\right] \\ &= \mathbb{E}_{\mathbf{x}}\left[Y_1(\mathbf{x}) \mid T=1\right] - \mathbb{E}_{\mathbf{x}}\left[Y_0(\mathbf{x}) \mid T=0\right] \\ &\approx \widehat{ATE} \end{aligned}$$

Randomized



Observational

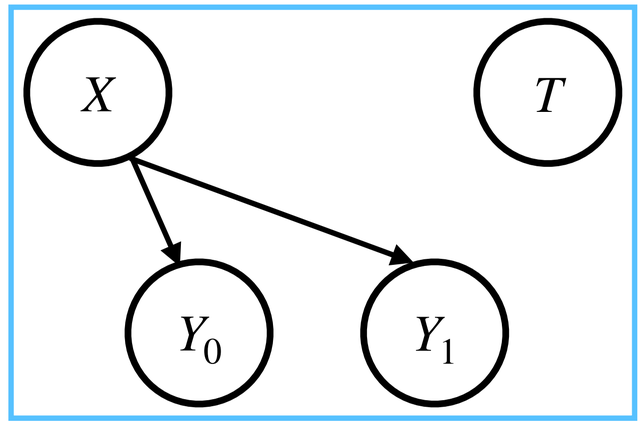


Treated people may be sicker, older, etc.

Randomization

$$ATE = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[Y_1(\mathbf{x}) - Y_0(\mathbf{x})\right] = \mathbb{E}_{\mathbf{x}}\left[Y_1(\mathbf{x}) \mid T = 1\right] - \mathbb{E}_{\mathbf{x}}\left[Y_0(\mathbf{x}) \mid T = 0\right]$$

- Gold standard for estimating effect, hard or unethical in many cases
  - Does smoking cause cancer?
  - What are the effects of multiple drugs for coronary artery disease?
  - What's the effect of a job training program?

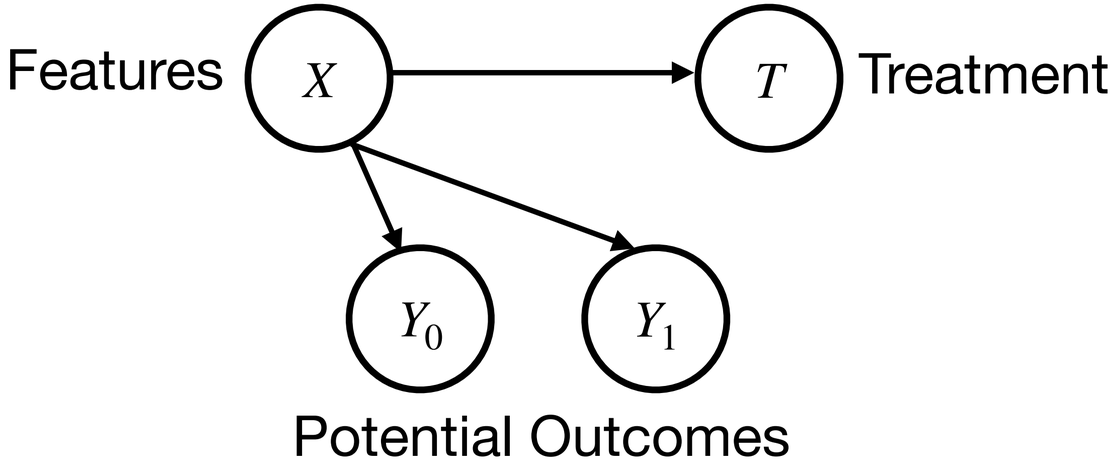


Observational Causal Inference

- We wish to estimate ATE from observational data

$$ATE = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} \left[Y_1(\mathbf{x}) - Y_0(\mathbf{x}) \right]$$

- What can we do?
  - Impossible without assumptions
  - A possible problem is *confounding* between treatment and outcomes

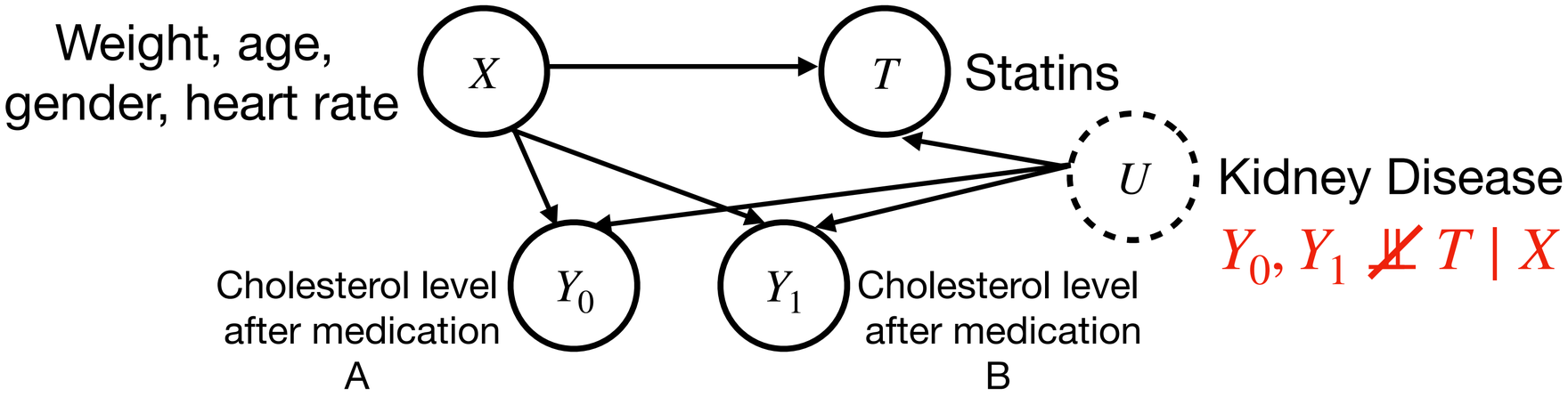


Observational Causal Inference

- We wish to estimate ATE from observational data

$$ATE = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} \left[Y_1(\mathbf{x}) - Y_0(\mathbf{x}) \right]$$

- What can we do?
  - Impossible without assumptions
  - A possible problem is *confounding* between treatment and outcomes



Ignorability

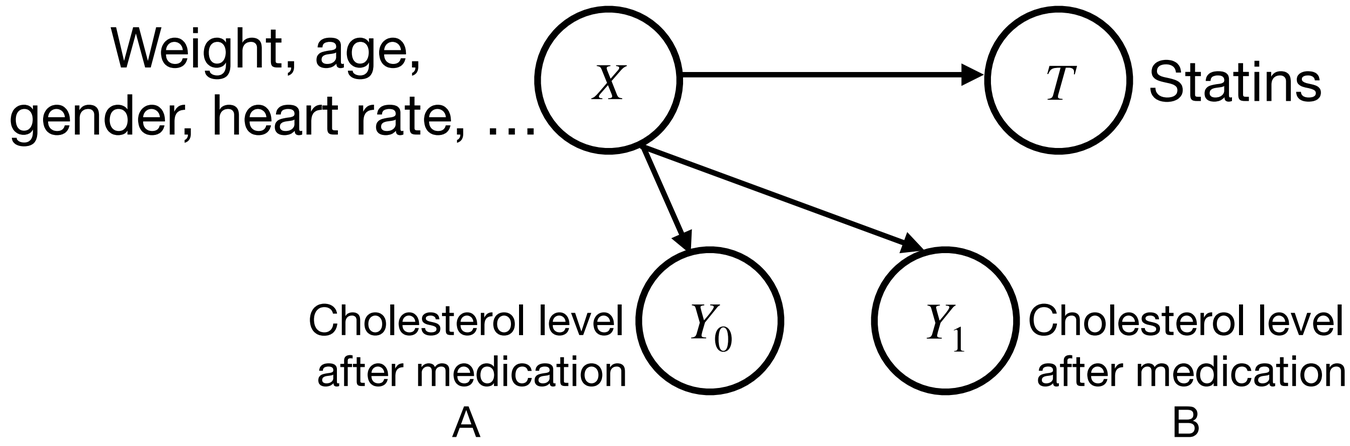
- Typical assumption - no unmeasured confounders

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

Ignorability

Y_0, Y_1: Potential outcomes for control and treated
X: Unit covariates
T: Treatment assignment

- Can we test if ignorability holds?



Common Support

- Another typical assumption: non-zero probability to assign patient to control/treatment

$$p(T = t \mid X = \mathbf{x}) > 0 \quad \forall \mathbf{x}, t$$

Common Support

- Y_0, Y_1: Potential outcomes for control and treated
  X: Unit covariates
  T: Treatment assignment

- **Next step:** Algorithms for ATE estimation under ignorability and common support
  - Derive the *adjustment formula* / *G-formula* [Pearl 2009, Hernán & Robins 2010]
  - Methods for estimation

Average Treatment Effect

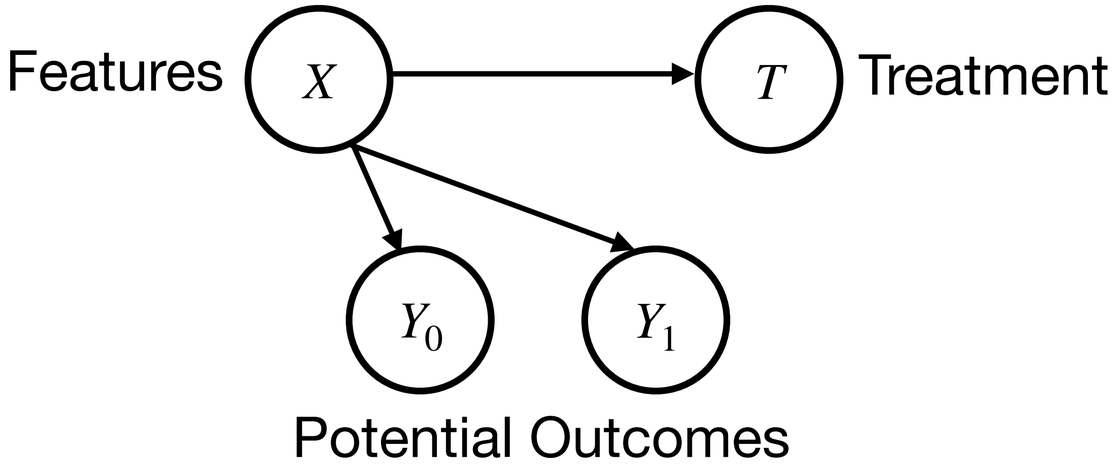
- Let us write our estimand:

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

$$\mathbb{E}\left[Y_1\right] = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\mathbb{E}_{Y_1\sim p(Y_1|X=\mathbf{x})}\left[Y_1 \mid \mathbf{x}\right]\right] \quad \text{Law of total expectation}$$

$$= \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\mathbb{E}_{Y_1\sim p(Y_1|X=\mathbf{x})}\left[Y_1 \mid \mathbf{x}, T=1\right]\right] \quad \text{Ignorability, } Y_1, Y_0 \perp\!\!\!\perp T \mid X$$

$$= \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\mathbb{E}\left[Y_1 \mid \mathbf{x}, T=1\right]\right] \quad \text{Lighter notation}$$



Average Treatment Effect

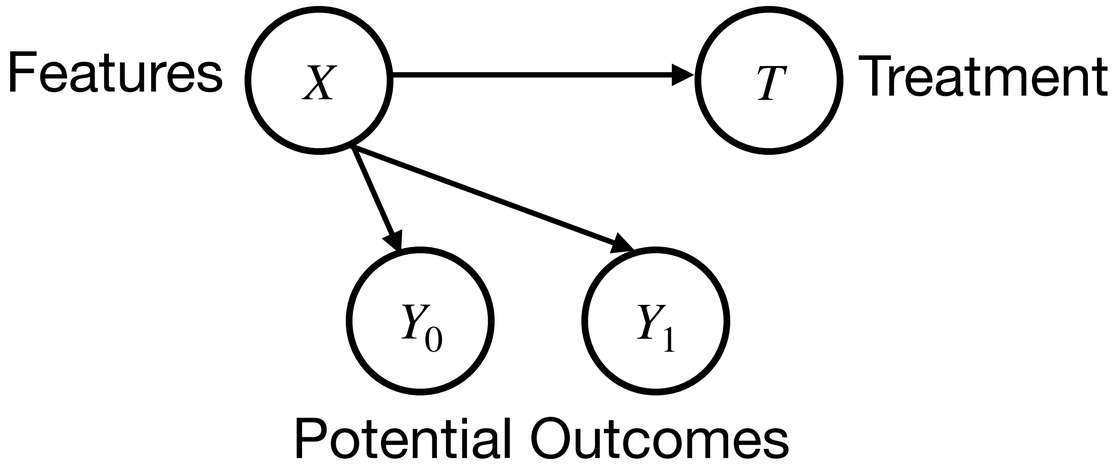
- Let us write our estimand:

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

$$\mathbb{E}\left[Y_1\right] = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})}\left[\mathbb{E}\left[Y_1 \mid \mathbf{x}, T = 1\right]\right]$$

$$\mathbb{E}\left[Y_0\right] = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})}\left[\mathbb{E}\left[Y_0 \mid \mathbf{x}, T = 0\right]\right]$$

Same derivation with $T = 0$



The Adjustment Formula

- Under ignorability we have:

$$ATE := \mathbb{E}\left[Y_1 - Y_0\right] = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\mathbb{E}\left[Y_1 \mid \mathbf{x}, T=1\right] - \mathbb{E}\left[Y_0 \mid \mathbf{x}, T=0\right]\right]$$

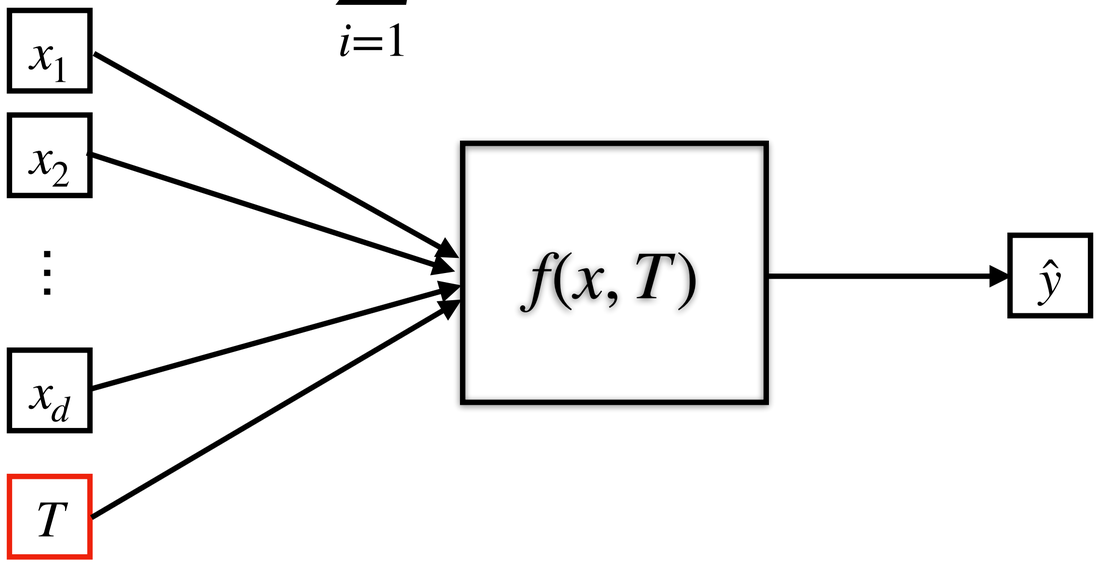
- $\mathbb{E}\left[Y_1 \mid \mathbf{x}, T=1\right]$, $\mathbb{E}\left[Y_0 \mid \mathbf{x}, T=0\right]$ } Quantities that can be *identified* from data
- How to estimate them from a finite sample?
  - Covariate adjustment
  - Propensity score re-weighting
  - Matching
  - Doubly robust estimators
  - …

Our data are samples from $p(\mathbf{x} \mid T=1)$ and $p(\mathbf{x} \mid T=0)$, wish to extrapolate to $p(\mathbf{x})$

Where does full support come in?

Covariate Adjustment

- Explicility model the relationship between T, X, Y_0, Y_1 with a regression model
  *can also be used for individual effects
- Solve $\min_{\theta} \mathbb{E}[(f_\theta(\mathbf{x}, t) - y)^2]$, obtain $f_\theta(\mathbf{x}, t) \approx \mathbb{E}\left[Y_t \mid T = t, \mathbf{x}\right]$
- Plug-in to obtain estimate $\widehat{ATE} = \sum_{i=1}^{N} f_\theta(\mathbf{x},1) - f_\theta(\mathbf{x},0)$



Needs to be handled with care!

Discussion

- Causal inference from observational data requires assumptions!
  - Requires domain knowledge and careful data collection
- How do we evaluate our fitted covariate adjustments?
- Where do questions related to causal inference come up in ML besides effect estimation?