JOHNS HOPKINS
UNIVERSITY

Machine Learning and Causal Inference for (Reliable) Decision Support

Suchi Saria* w/ Roy Adams, Adarsh Subbaswamy, Peter Schulam

*John C. Malone Assistant Professor,
Computer Science, Stats, and Health Policy

@suchisaria

NSF

GORDON AND BETTY MOORE FOUNDATION

Google

THE MICHAEL J. FOX FOUNDATION FOR PARKINSON'S RESEARCH

DARPA

Introduction

- Predictive analysis is the dominant analysis paradigm in machine learning and used within many decision-making applications.

  - **Movie Recommendation system**: Is this viewer likely to want to watch this movie?
  - **Advertising**: Is this site visitor likely to want to purchase this item?
  - **Churn**: Is this member likely to unsubscribe from my service?
  - **Personalized Medicine**: Is this patient likely to be high risk and should be prescribed aggressive treatment?
  - **Social Justice**: Is this individual likely to have committed a crime?

Introduction

- Predictive analysis is the dominant analysis paradigm in machine learning and used within many decision-making applications.
  - **Movie Recommendation system**: Is this viewer likely to want to watch this movie?

Human-Centered AI or Augmented Intelligence:
How can we use data-driven tools to augment
human decision-making?

service?

  - **Personalized Medicine**: Is this patient likely to be high risk and should be prescribed aggressive treatment?
  - **Social Justice**: Is this individual likely to have committed a crime?

Introduction

Algorithmic fairness

- **Goal:**
  - To make predictions about whether an applicant is suitable for a job.

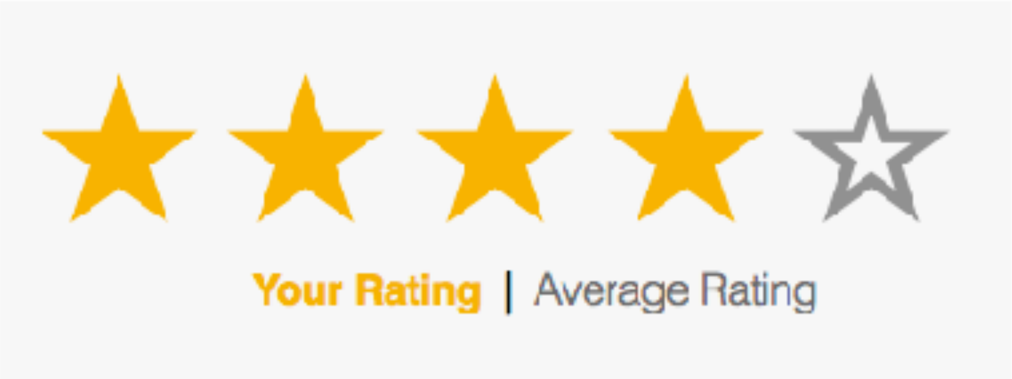
- **Why current approaches for predictive analysis fail:**
  - Retrospective data is effected by societal biases that we would like to remove from the decision making process.
    - Predictions should not depend on certain variables (e.g. race, gender, income).
    - Example: Women may have been less likely to be CEOs

- **Unintended consequence:** Enforcing historical bias

Introduction

Recommendation systems

- **Goal:**
  - Predict how a user will rate an item (e.g. a movie).
- **Naive approach:** Use retrospective data of items this user has rated to predict rating.
- **Why current approaches for predictive analysis fail:**
  - Retrospective data is biased by which user was shown which item, what they tend to rate and so on.
  - Users are more likely to rate items they like.



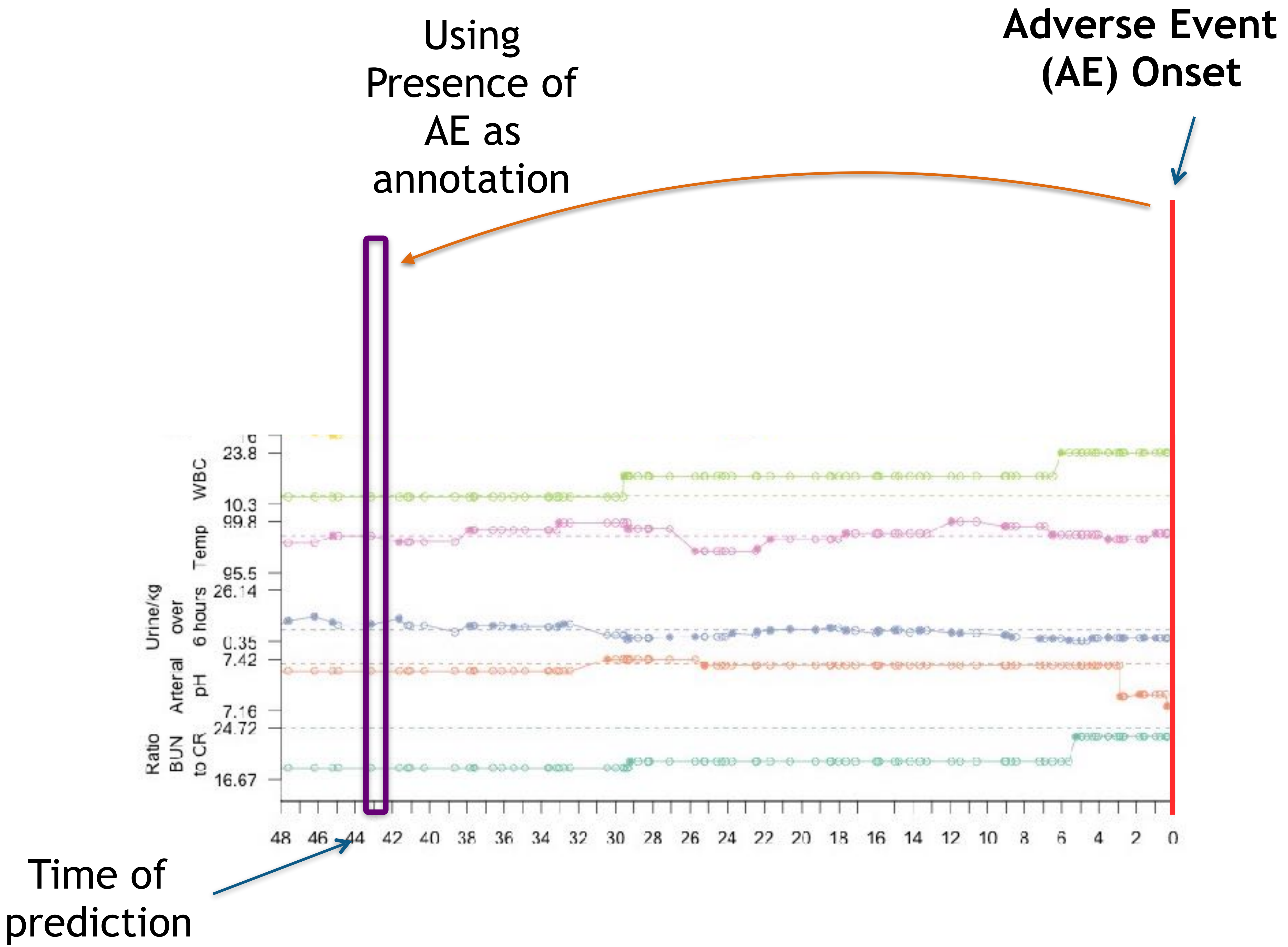
- **Unintended consequence:** Certain users are never exposed to certain movies

Introduction

Medical prognostication

- **Goal:**
  - To predict future outcomes for a patient given their medical history.
- **Why current approaches for predictive analysis fail:**

Use supervised learning for distinguishing patients with AE from those without

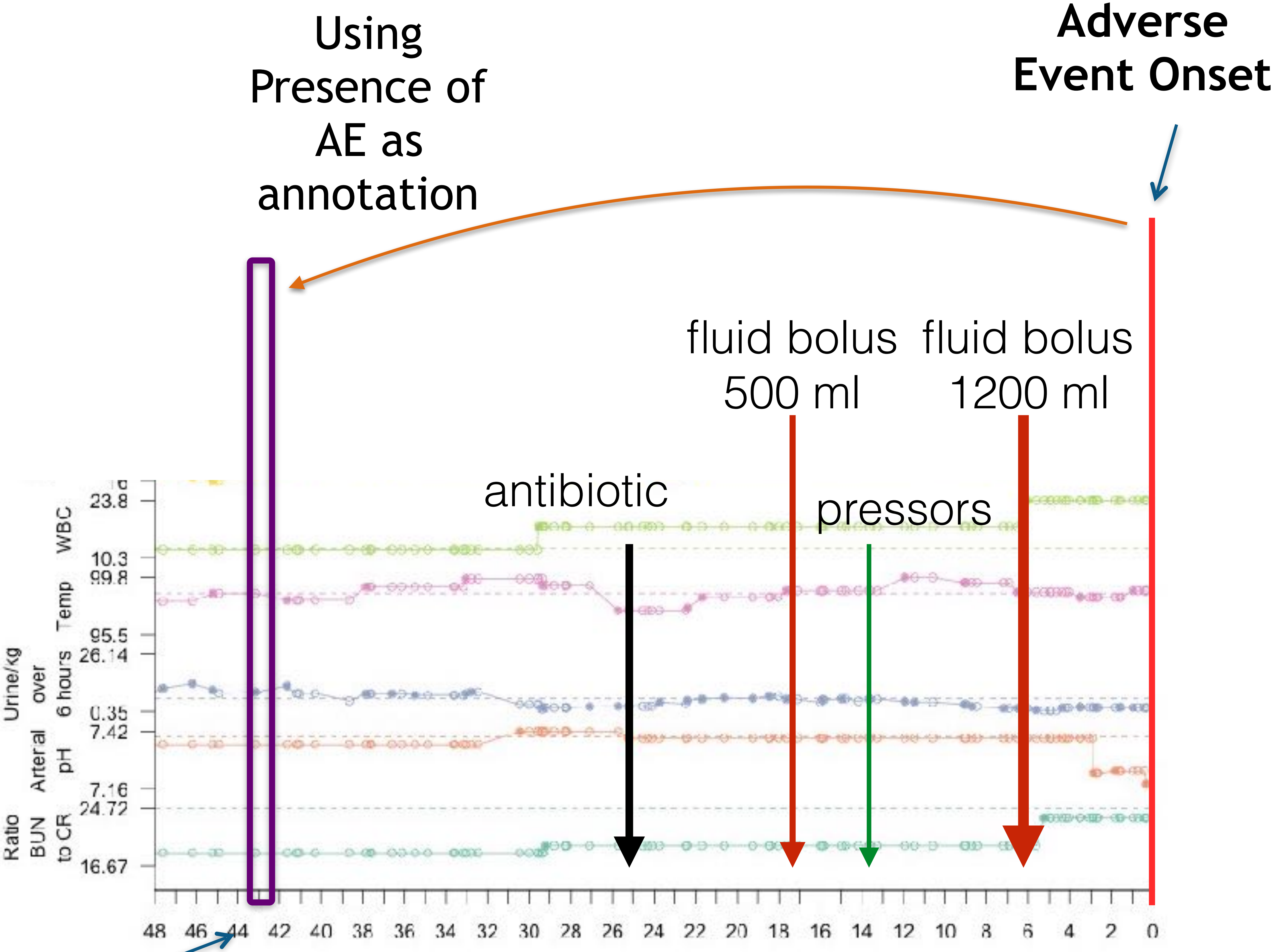


Using Presence of AE as annotation

Adverse Event (AE) Onset

Time of prediction

But, interventions *censor* the true label.



Using Presence of AE as annotation

Adverse Event Onset

fluid bolus 500 ml

fluid bolus 1200 ml

antibiotic

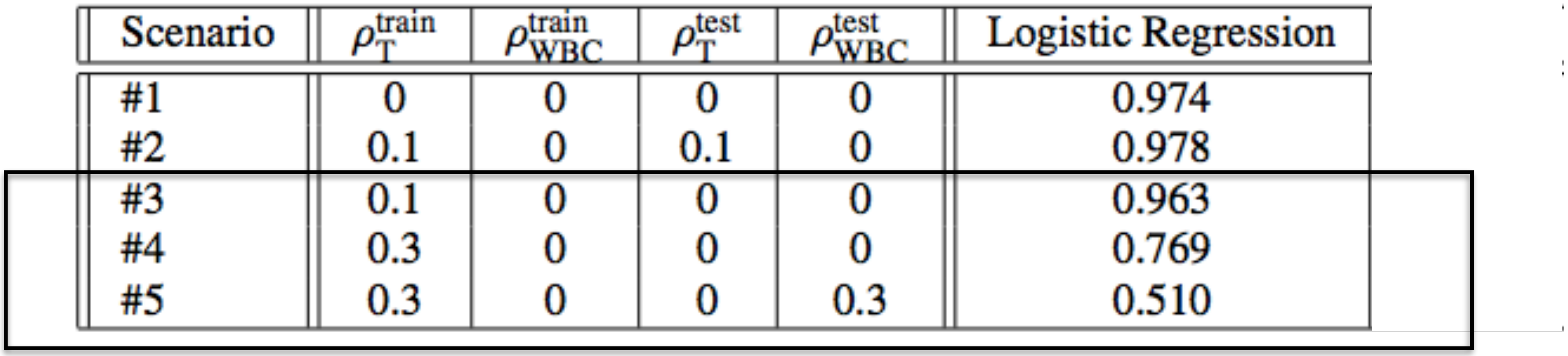
pressors

Time of prediction

Dyagilev et al., Machine Learning 2015

Bias Due to Interventional Confounds

Vary provider practice patterns between train and test:



Scenario	$\rho_{\mathrm{T}}^{\mathrm{train}}$	$\rho_{\mathrm{WBC}}^{\mathrm{train}}$	$\rho_{\mathrm{T}}^{\mathrm{test}}$	$\rho_{\mathrm{WBC}}^{\mathrm{test}}$	Logistic Regression
#1	0	0	0	0	0.974
#2	0.1	0	0.1	0	0.978
#3	0.1	0	0	0	0.963
#4	0.3	0	0	0	0.769
#5	0.3	0	0	0.3	0.510

Increase probability of treating for rising temperature

Increasing discrepancy in physician prescription behavior in train vs. test environment

Learned risk scores are high sensitive to choice of treatment practices in the training dataset

Dyagilev et al., Machine Learning 2015

Naive application of predictive tools can give counterintuitive results

Example: (**Caruana et al., KDD, 2015**)

- ML method learned that patients with pneumonia with asthma history have lower mortality risk than the general population.
- This is counterintuitive — patients with asthma history have much higher risk *if not hospitalized*
- Pneumonia patients with asthma history were admitted to the ICU, and the intensive care lowered their risk of dying

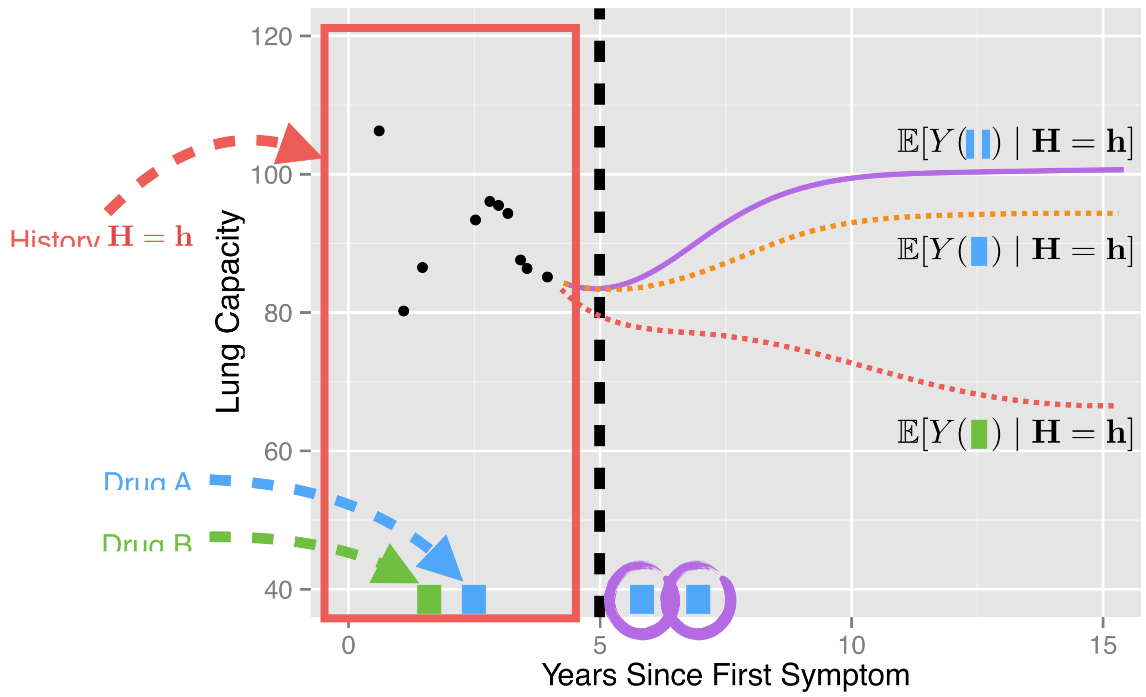
If applied naively and without considering clinical context, machine learning methods may yield **counterintuitive predictions** and models with **unintended consequences.**

Caruana et al., KDD 2015

Introduction

Medical prognostication

- **Goal:**
  - To predict future outcomes for a patient given their medical history.
- **Why current approaches for predictive analysis fail:**
  - Outcomes in retrospective data are affected by existing treatment policies and other environmental characteristics.



- **Unintended consequence:** High-risk patients maybe considered low-risk and miss receiving necessary treatment.

“Causal Predictions”

- Recasting the problem as answering **“what if”** questions.

- “What if” we gave this patient aggressive treatment versus not?
- “What if” exposed this user to this movie?
- “What if” we gave this candidate the job?

- How is the above different from how we were previously approaching the problem?
  - Existing predictive analysis techniques are good for detecting associations.
  - “Correlation is not causation”
  - “What if” formulation requires more discipline when learning models —> explicitly reason about factors that do not generalize from train to deployment.

See for further discussion/motivation: **Schulam et al., NIPS 2017**

Goals of this tutorial

- See example applications to develop a deeper appreciation for the challenge at hand.
- Introduce concepts from causal inference that we will use as building blocks for developing solutions
- Describe example approaches that address this challenge of lack of reliability.
- Revisit applications to illustrate the idea in practice.

Day 1

Formalizing “what if…” - Potential outcomes

- Suppose you are concerned about your blood pressure. And, you are interested in asking whether to start exercising so as to manage your blood pressure.

- Formulation 1: “What if” I were to exercise, would it help manage my blood pressure?
- Formulation 2: What is the effect of exercise on the blood pressure of individuals like myself?
- Formulation 3: What is the effect of exercise on blood pressure?

Question: Can learning a predictive model from retrospective data to determine whether to exercise give the right answer?

Formalizing “what if…” - Potential outcomes

- Suppose you are concerned about your blood pressure. And, you are interested in asking whether to start exercising so as to manage your blood pressure.

- Formulation 1: “What if” I were to exercise, would it help manage my blood pressure?
- Formulation 2: What is the effect of exercise on the blood pressure of individuals like myself?
- Formulation 3: What is the effect of exercise on blood pressure?

Question: Can learning a predictive model from retrospective data to determine whether to exercise give the right answer?
Answer: Depends…

Formalizing “what if…” - Potential outcomes

- Formulation 3: What is the effect of exercise on blood pressure?

- The causal effect of exercise on blood pressure (BP).
  - Exercise is called our **treatment** and denoted as A
  - BP is called our **outcome** and denoted as Y

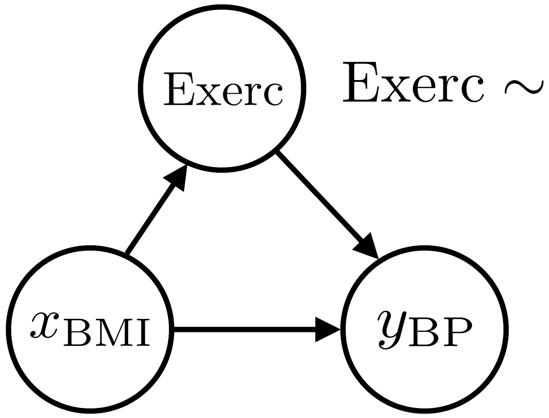
- Our goal is to estimate effect from a retrospective dataset. In causal inference, retrospective data are also called **observational data** because the learner only gets to observe but cannot control the data collection protocol. Analysis from retrospective data is significantly more challenging than a prospectively collected dataset because one cannot proactively design the collection protocol to remove biases that complicate the analyses.

Example: Exercise and Blood Pressure

- A = Exercise
- Y = Blood pressure (BP)
- X = Body mass index (BMI)
- Question: What is the effect of exercise on BP?
- Approach: Grab an existing dataset and average the BP among people who exercise and those who don't to estimate effect
- Is the resulting effect correct?
  - **Depends…** requires understanding the data generating mechanism.

Scenario #1: Observational Data w/ selection bias

- Dataset generative model:



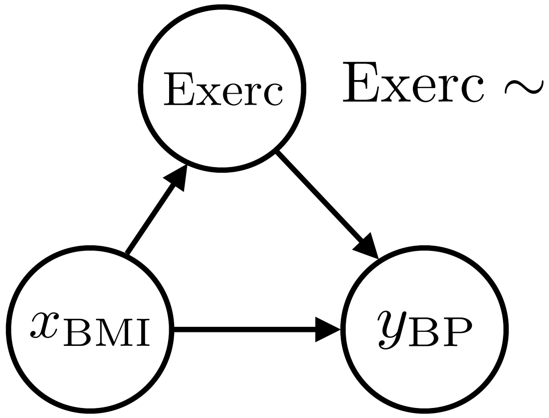
$$\text{Exerc} \sim \text{Bern}\left(\frac{1}{1+e^{-2x_{\text{BMI}}}}\right)$$

$$x_{\text{BMI}} \sim \mathcal{N}(0,1)$$

$$y_{\text{BP}} \sim \mathcal{N}(x_{\text{BMI}}, 0.4) - 0.8 \cdot \text{Exerc}$$

Scenario #1: Observational Data w/ selection bias

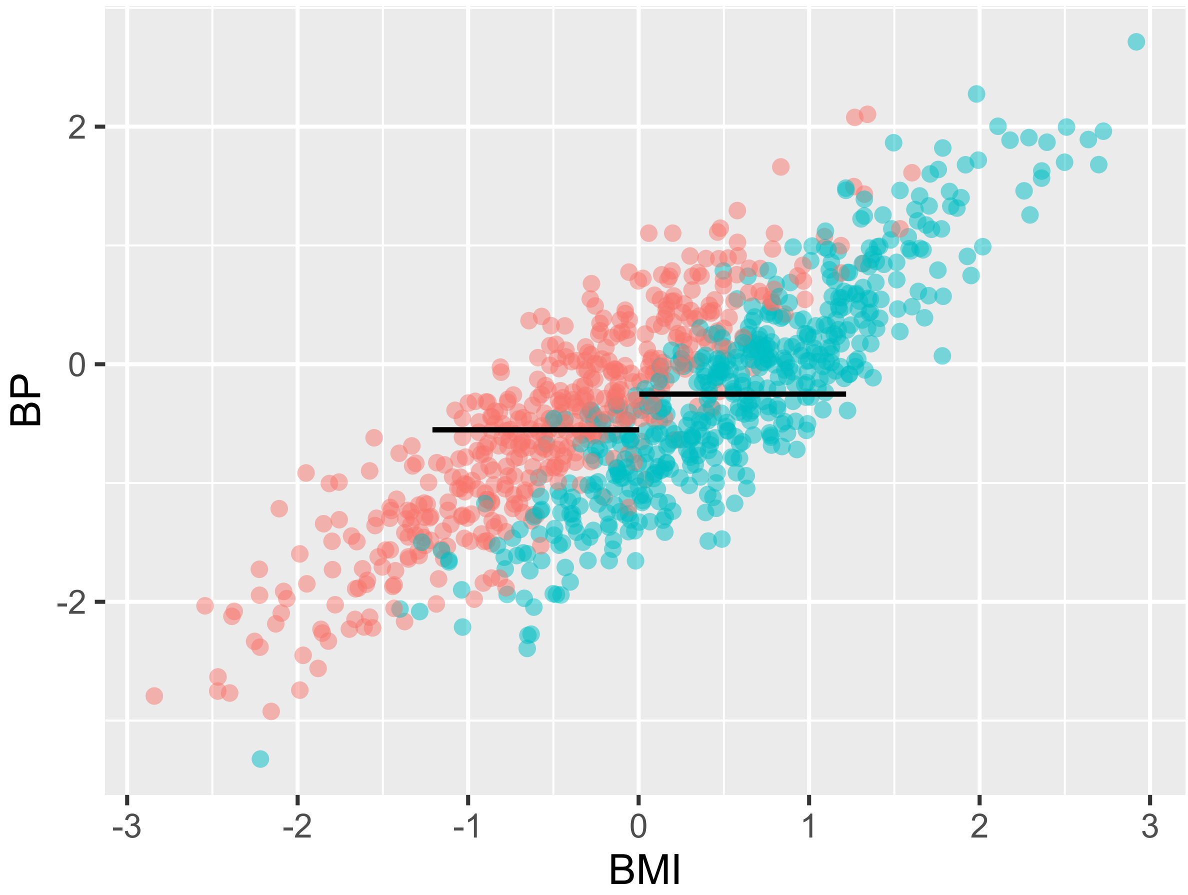
- If we estimate the causal effect of exercise on BP by simply averaging (i.e. **ignoring BMI**) BP in both treatment groups, we get the wrong answer! Why?

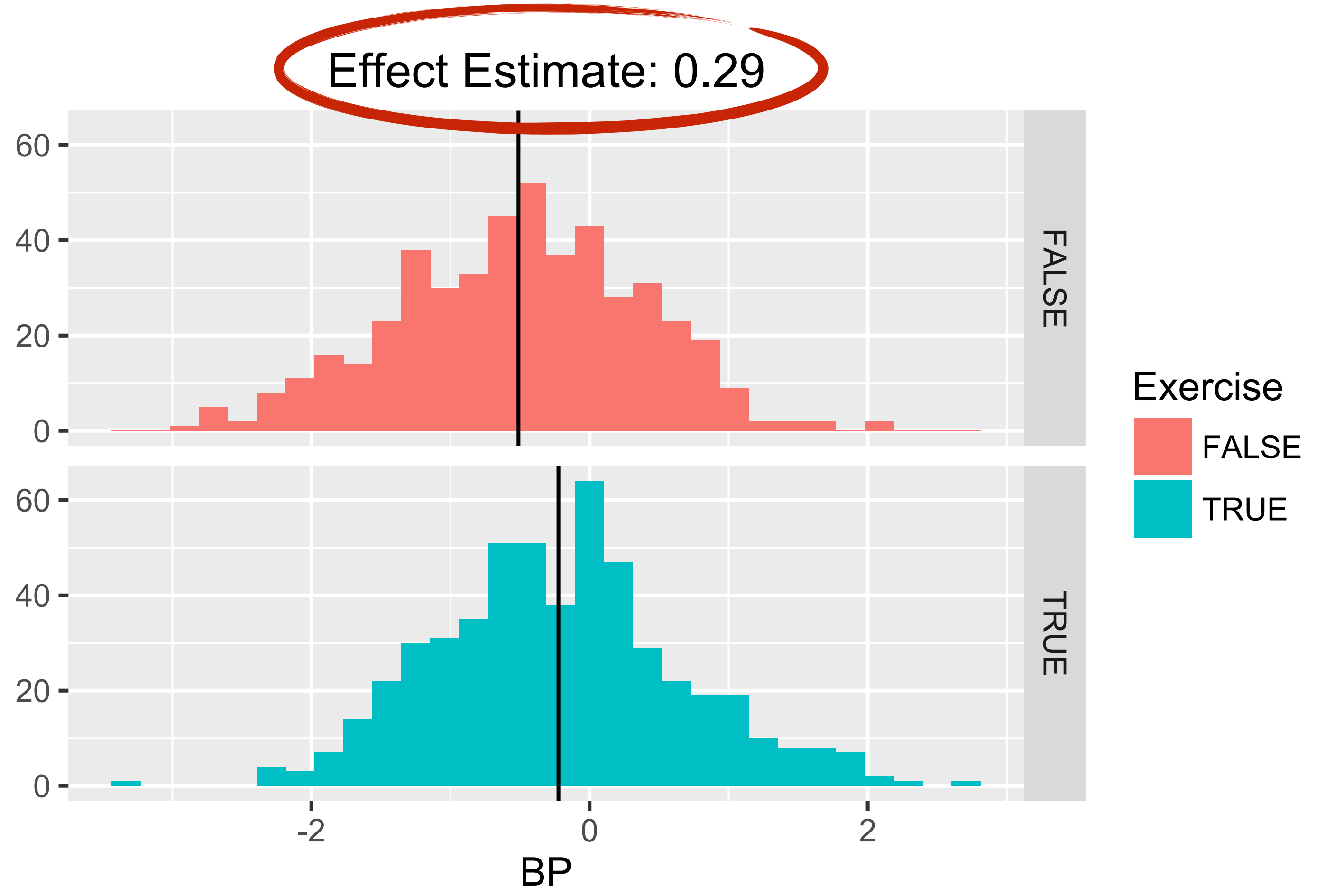


$$\text{Exerc} \sim \text{Bern}\left(\frac{1}{1+e^{-2x_{\text{BMI}}}}\right)$$

$$x_{\text{BMI}} \sim \mathcal{N}(0,1)$$

$$y_{\text{BP}} \sim \mathcal{N}(x_{\text{BMI}}, 0.4) - 0.8 \cdot \text{Exerc}$$



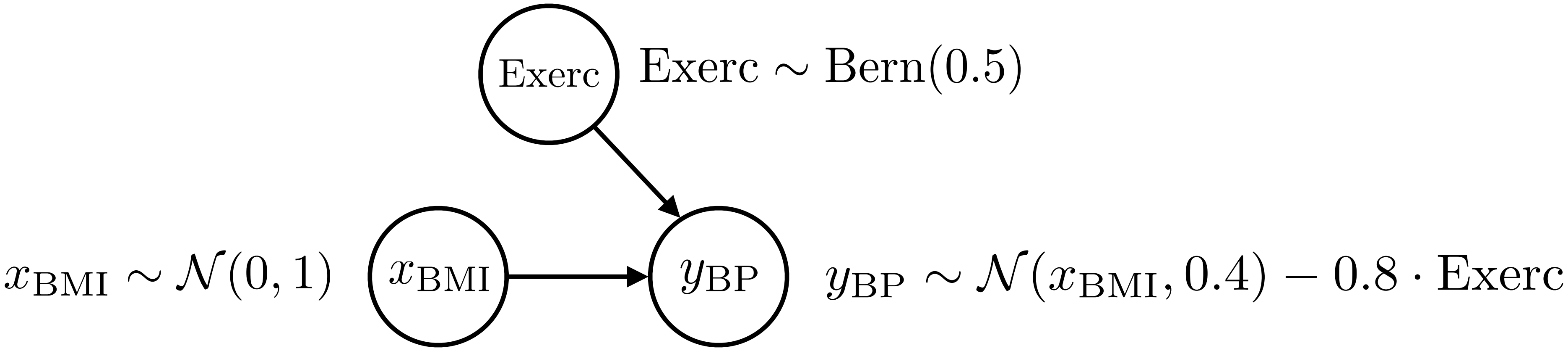


Effect Estimate: 0.29

Scenario #2: Randomized Controlled Trial (RCT)

- What happens if we assigned subjects randomly to the exercise and non-exercise arm.

- Dataset generative model:



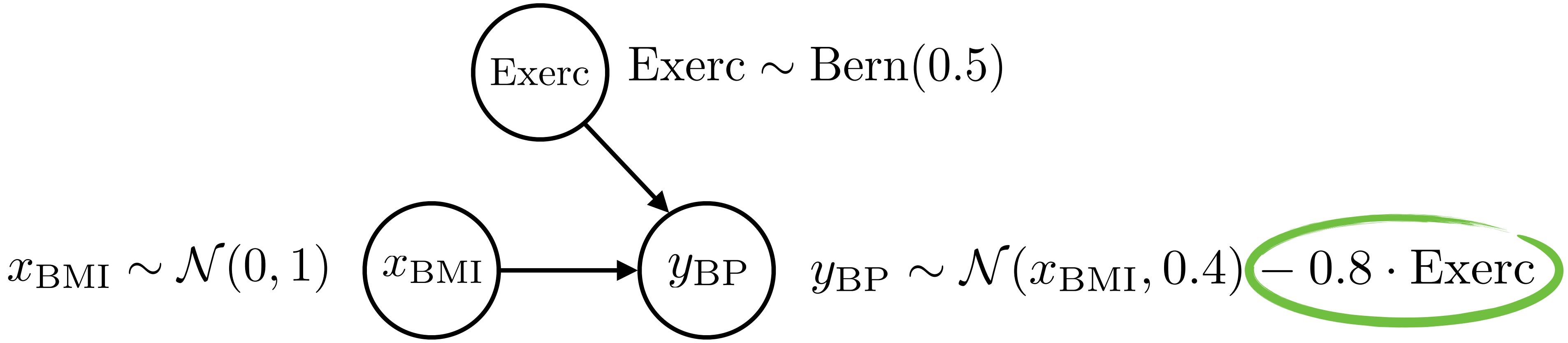
$\text{Exerc} \sim \text{Bern}(0.5)$

$x_{\text{BMI}} \sim \mathcal{N}(0, 1)$

$y_{\text{BP}} \sim \mathcal{N}(x_{\text{BMI}}, 0.4) - 0.8 \cdot \text{Exerc}$

Scenario #2: Randomized Controlled Trial (RCT)

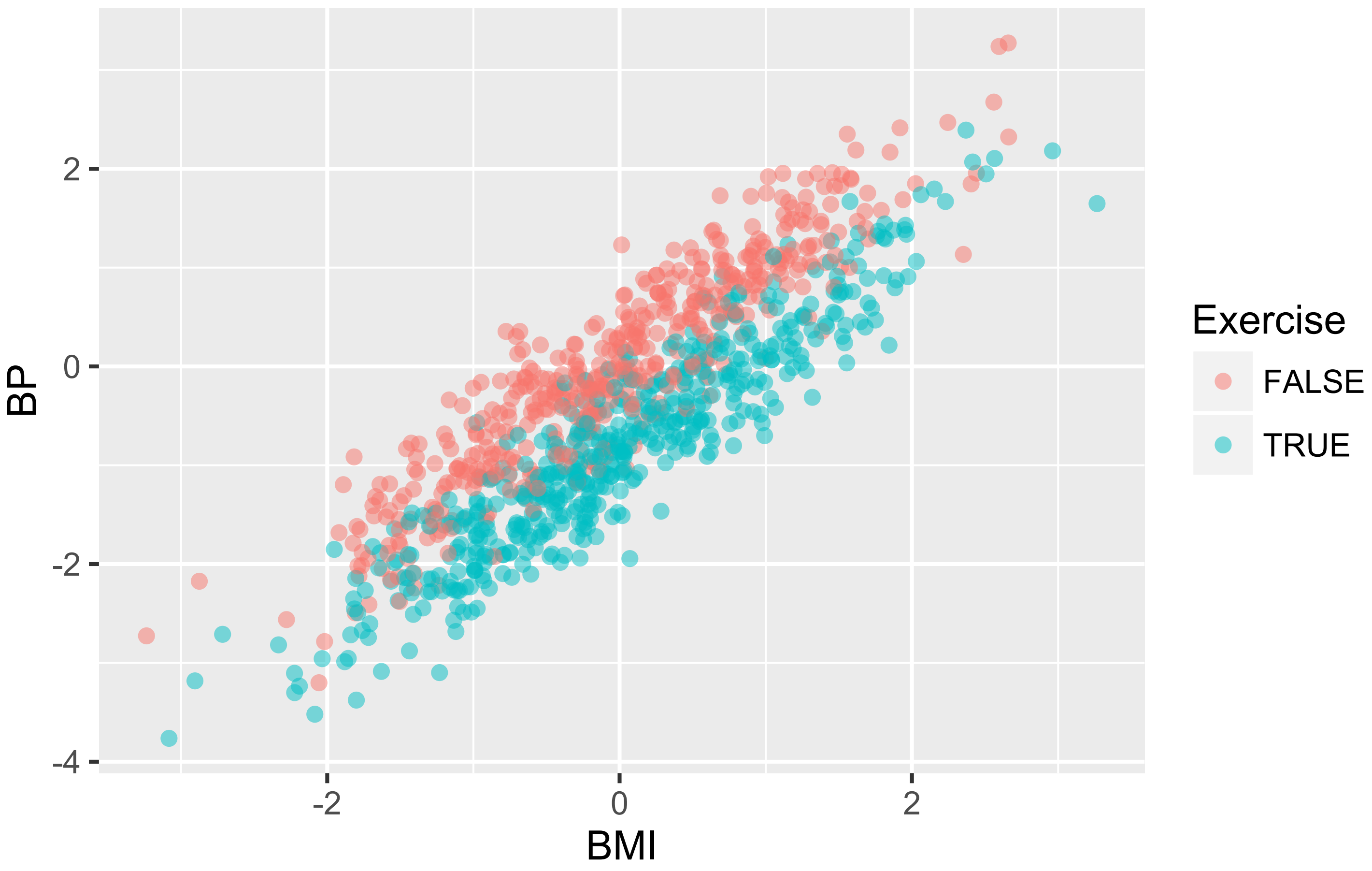
- Dataset generative model:



$$\text{Exerc} \sim \text{Bern}(0.5)$$

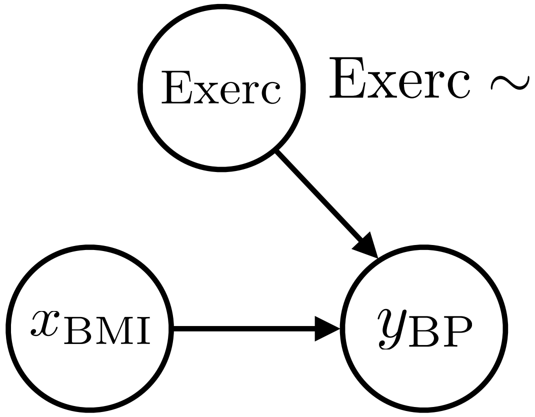
$$x_{\text{BMI}} \sim \mathcal{N}(0, 1)$$

$$y_{\text{BP}} \sim \mathcal{N}(x_{\text{BMI}}, 0.4) - 0.8 \cdot \text{Exerc}$$



Scenario #2: Randomized Controlled Trial (RCT)

- Dataset generative model:

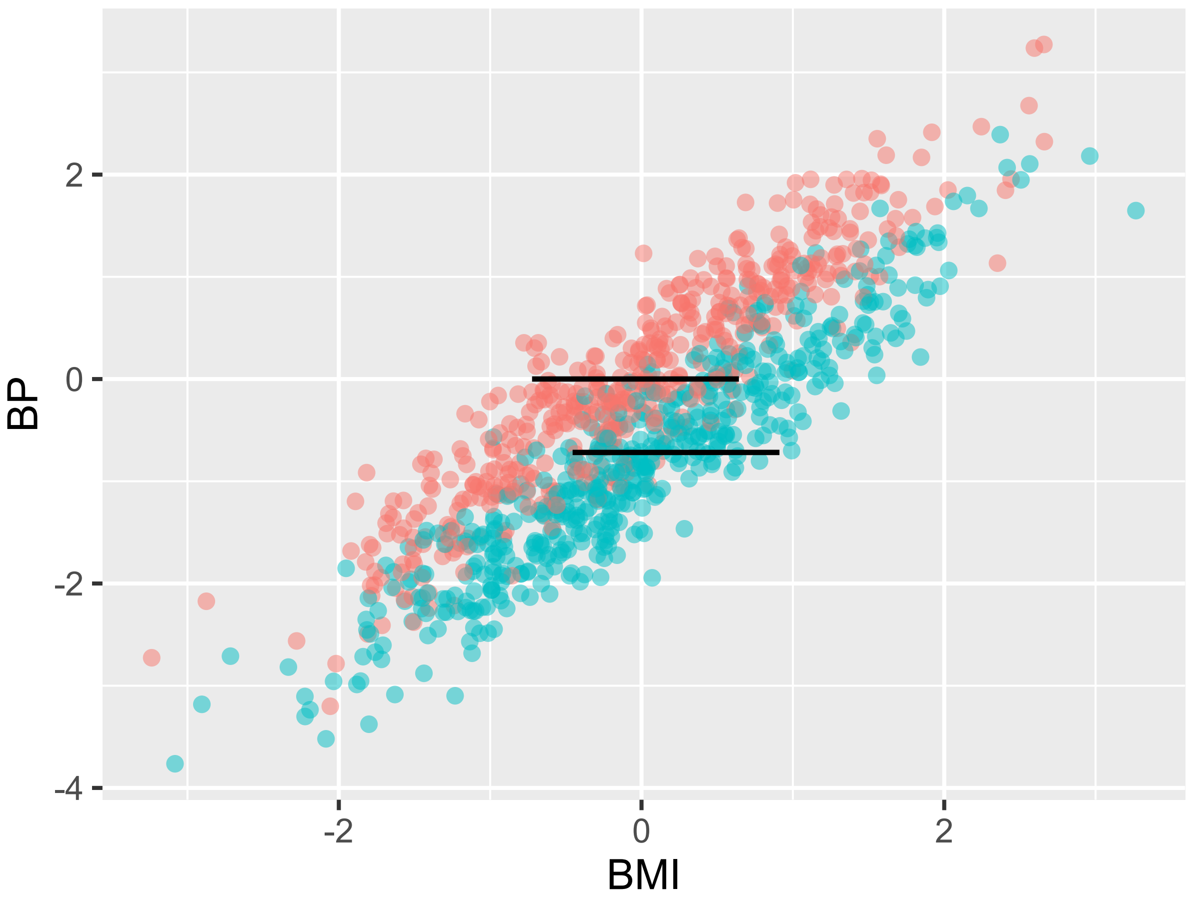


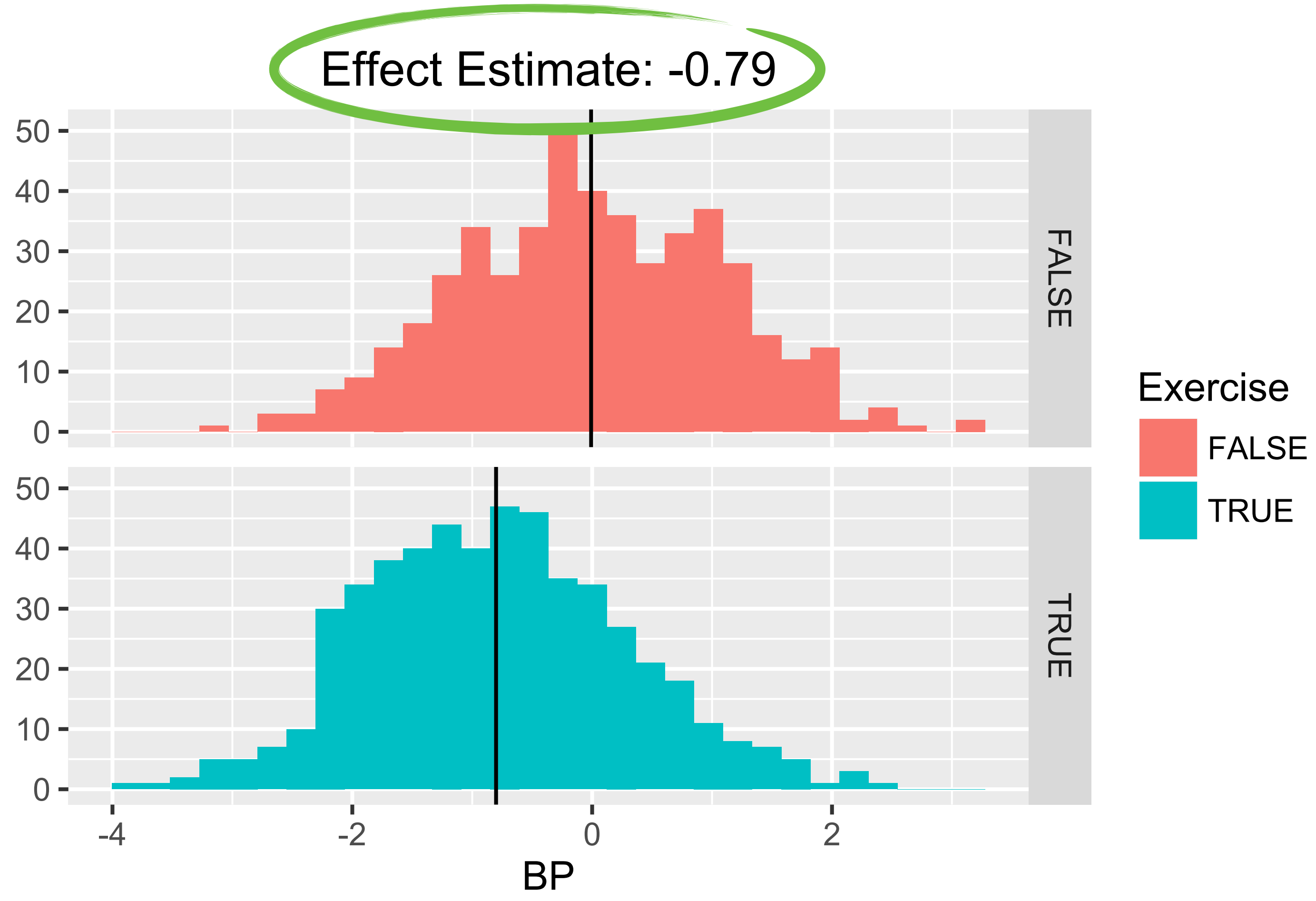
$\text{Exerc} \sim \text{Bern}(0.5)$

$x_{\text{BMI}} \sim \mathcal{N}(0,1)$

$y_{\text{BP}} \sim \mathcal{N}(x_{\text{BMI}}, 0.4) - 0.8 \cdot \text{Exerc}$

- Now computing simple averages will work! Why?



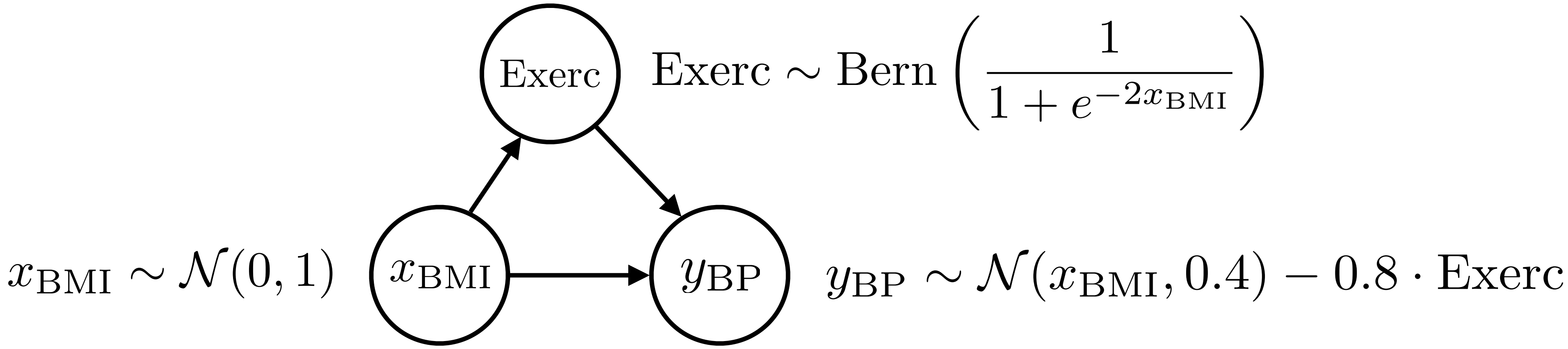

Effect Estimate: -0.79

What is a confounder?

- A **confounder** is any covariate that has a causal effect on both the treatment and outcome.

- In scenario #1, BMI serves as a confounder.

  - Individuals assigned to the exercise vs not-exercise arm are not similar: individuals in the exercise arm tend to have higher BMI. This needs to be adjusted for.



$$\text{Exerc} \sim \text{Bern}\left(\frac{1}{1+e^{-2x_{\text{BMI}}}}\right)$$

$$x_{\text{BMI}} \sim \mathcal{N}(0,1)$$

$$y_{\text{BP}} \sim \mathcal{N}(x_{\text{BMI}}, 0.4) - 0.8 \cdot \text{Exerc}$$

Data Generating Model for Scenario #1

Core assumptions: No unobserved confounders

- In order to correctly infer causal effect from an observational dataset, we must assume that all confounders are observed i.e. there should be **no unobserved confounding.**
  - We need to adjust for observed confounders (we will discuss shortly).
  - If there are unobserved confounders, it may not be possible to estimate the correct effect using the provided retrospective dataset alone. That is, the effect is not **identifiable**.

- Another presentation commonly used: each potential outcome is independent of treatment assignment given the features (revisit using SWIGs time-permitting):

$$Y^a \perp A \mid X$$

Rubin, 1974 **Neyman et al., 1923** **Rubin, 2005**

Other assumptions for identifiability: Positivity

- Every subject has non-zero probability of receiving every treatment:

$$P(A = a \mid X = x) > 0 \text{ for all } x \text{ and } a$$

- Example: If people above or below a certain BMI never exercise, then we cannot reason about the effect of exercise on this group.

Rubin, 1974 Neyman et al., 1923 Rubin, 2005

Other assumptions for identifiability: Consistency

- If the observed treatment is a, then the observed outcome Y is equal to the potential outcome for treatment a:

$$\text{If } A = a \text{ then } Y^a = Y$$

- In other words, had you intervened and administered A=a, the outcome observed in the data is what you would have observed.
- Common-senseIssues to consider: If there are multiple ways to deliver treatment (e.g. running, weight lifting, or multiple doses or administration modalities), we need a clear definition for what it means to administer treatment. Further, the outcome of interest is being recorded correctly in the data.

Rubin, 1974 **Neyman et al., 1923** **Rubin, 2005**

Randomized trials - Assumptions

- Let’s recap how the RCT satisfies our assumptions
- **Consistency:** treatment must be well defined since it is being administered
- **Positivity:** Each subject has non-zero probability of being assigned to each treatment arm by construction
- **No unobserved confounders:** The choice of which arm the individual is assigned to—e.g., treatment vs no treatment—is randomized and independent of covariates that affect the potential outcome. Therefore, there are no paths from treatment to the potential outcome via confounders. In other words,

$$Y^a \perp A \,|\, X$$

Observational Data vs. RCT

- Randomized trials may be impossible for many reasons:
  - Ethical e.g. high risk of harm
  - Hard/impossible to intervene e.g. genetics
  - Impractical size requirements e.g. rare side effects
- In many cases we can collect observational data easily. But can we infer the desired causal effects?
  - Yes, only when certain assumptions about the data hold (e.g, positivity, no unmeasured confounding (NUC)).
  - **Assumptions are not always testable from data**
  - **No escape:** Must rely on domain knowledge

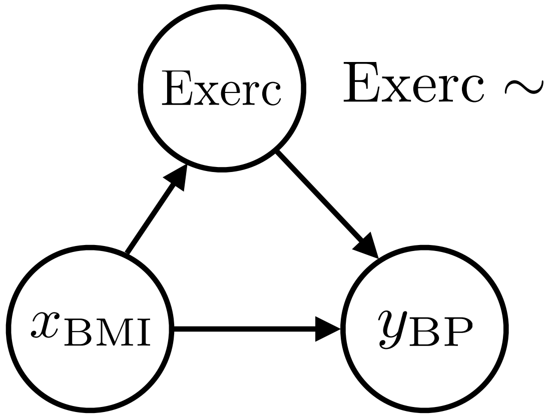
Exercise: Movie recommendation

- You're netflix. Your goal is to determine: if you recommend a movie to a user, does that influence their probability of watching the movie.
- Extract retrospective data of individuals. Collect those to whom movie M was recommended versus those without. Compute differences in viewing rates.

- Will this analysis produce the right answer?

Causal inference in observational data

- Coming back to our exercise example:



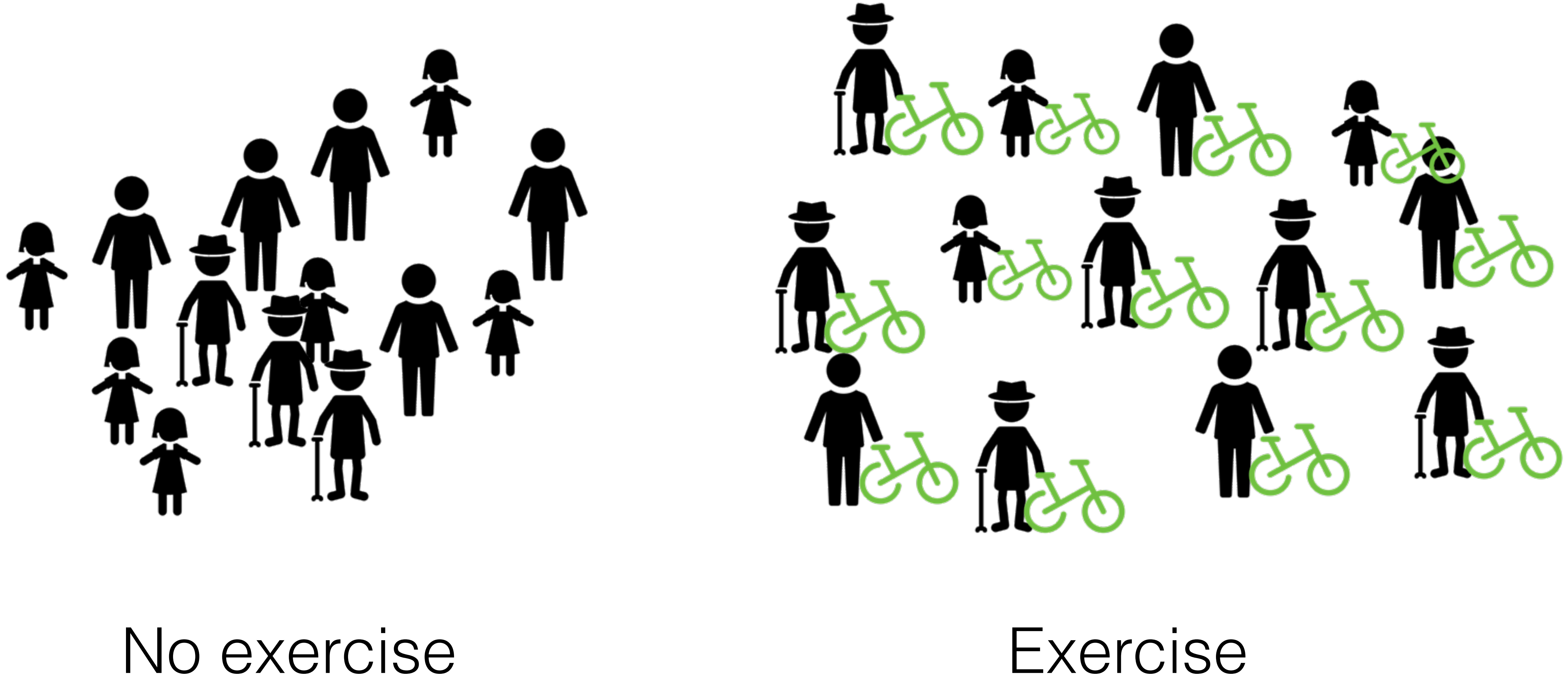
$$\text{Exerc} \sim \text{Bern}\left(\frac{1}{1+e^{-2x_{\text{BMI}}}}\right)$$

$$x_{\text{BMI}} \sim \mathcal{N}(0,1)$$

$$y_{\text{BP}} \sim \mathcal{N}(x_{\text{BMI}}, 0.4) - 0.8 \cdot \text{Exerc}$$

- We assume that BMI is our only confounder, but how do we account for it in our effect estimates?
- Many candidate methods, we will talk about three:
  - Matching/stratification
  - Weighting
  - Standardization/Potential Outcomes

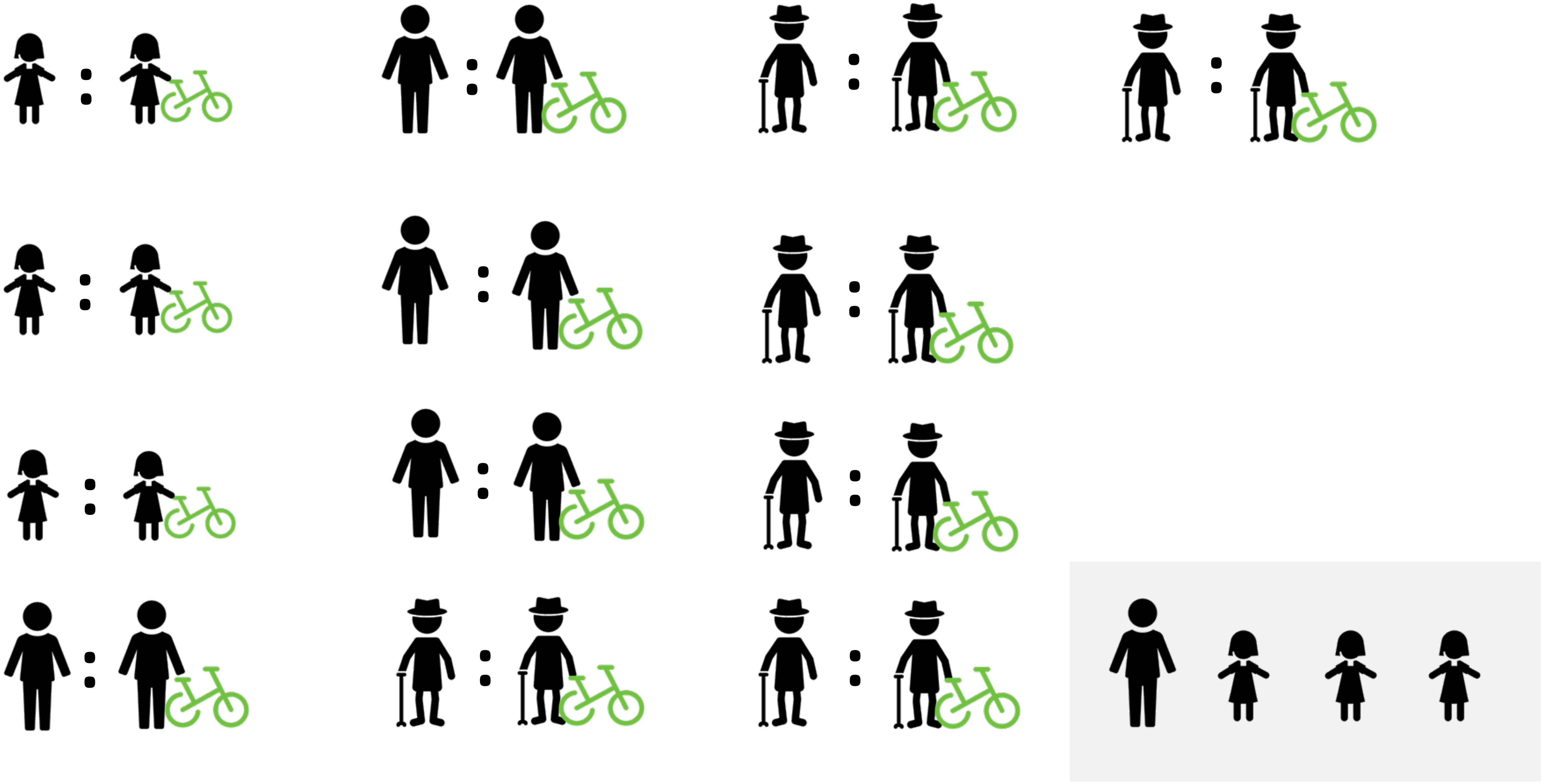
Observed confounders: Matching



Throughout, slides borrowed from or inspired by are denoted with this ref **Sharma and Kiciman (2018)**

Stuart (2010)

Observed confounders: Matching

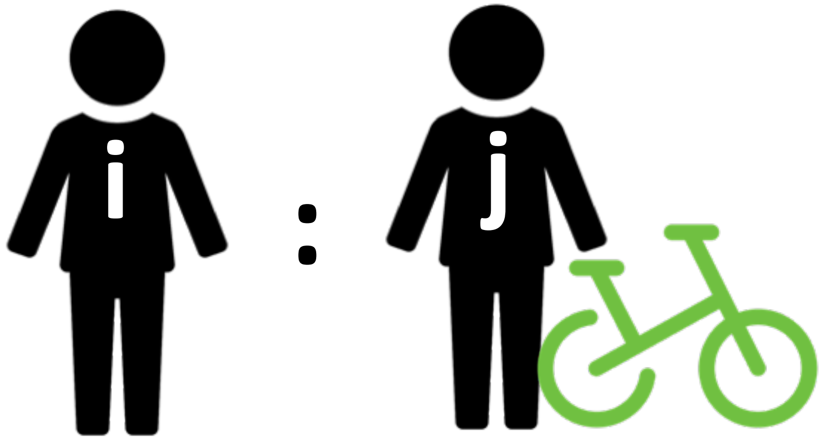


Sharma and Kiciman (2018)

Stuart (2010)

Observed confounders: Matching

- Identify pairs of treated and untreated individuals who are very similar or even identical to each other:



Very similar $::= Distance(x_i, x_j) < \epsilon$

- Paired individuals provide the counterfactual estimate for each other.
- Average the difference in outcomes within pairs to estimate the additive treatment effect.

Sharma and Kiciman (2018)

Stuart (2010)

Observed confounders: Matching

Exact matching:

$$Distance(x_i, x_j) = \begin{cases} 0 & x_i = x_j \\ \infty & x_i \neq x_j \end{cases}$$

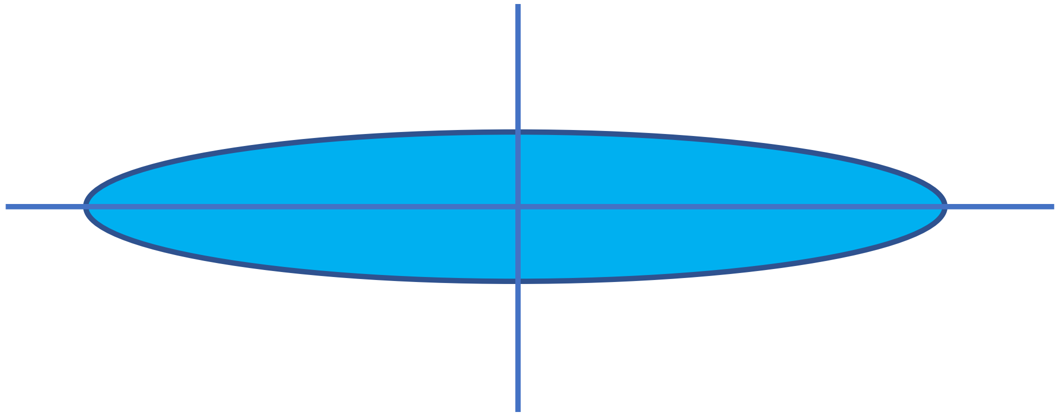
- Use in low dimensional cases with discrete features
- Fails in high dimensions

Sharma and Kiciman (2018)

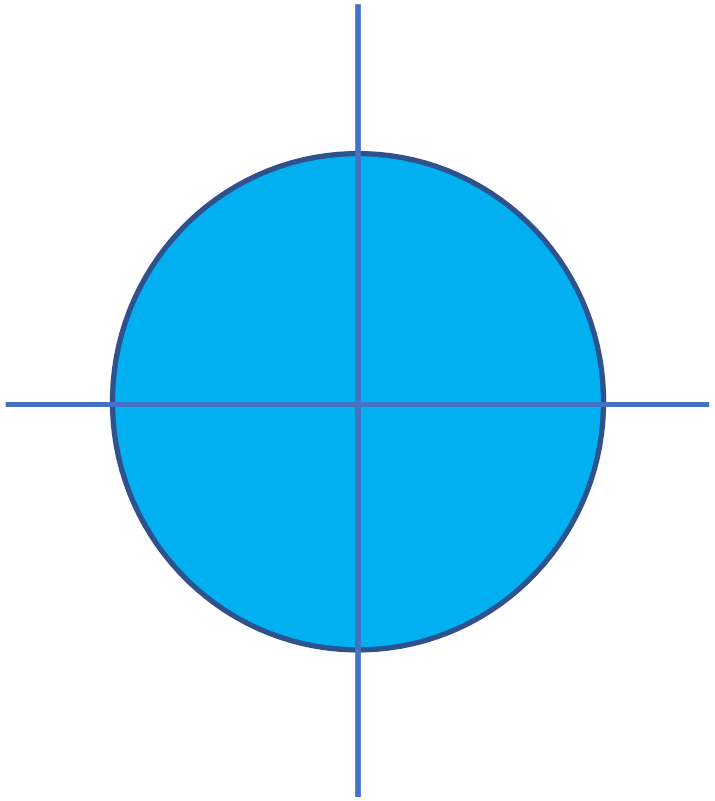
Stuart (2010)

Observed confounders: Matching

Mahalanobis distance:



$$Distance(x_i, x_j) = \sqrt{(x_i - x_j)^T S^{-1} (x_i - x_j)}$$



- S is the feature covariance matrix
- Accounts for unit differences by normalizing each dimension by its standard deviation

Sharma and Kiciman (2018)

Stuart (2010)

Observed confounders: Matching

- **Question:** What happens if we include features that have little effect on the treatment?
- **Answer:** We may include or exclude pairs based on irrelevant information. Incomplete covariate set will produce incorrect results.
- **Propensity score matching** allows us to focus on the features that determine treatment assignment
- A propensity score is an individuals probability of treatment:

$$e(x) = P(A = 1 | X = x)$$

Sharma and Kiciman (2018)

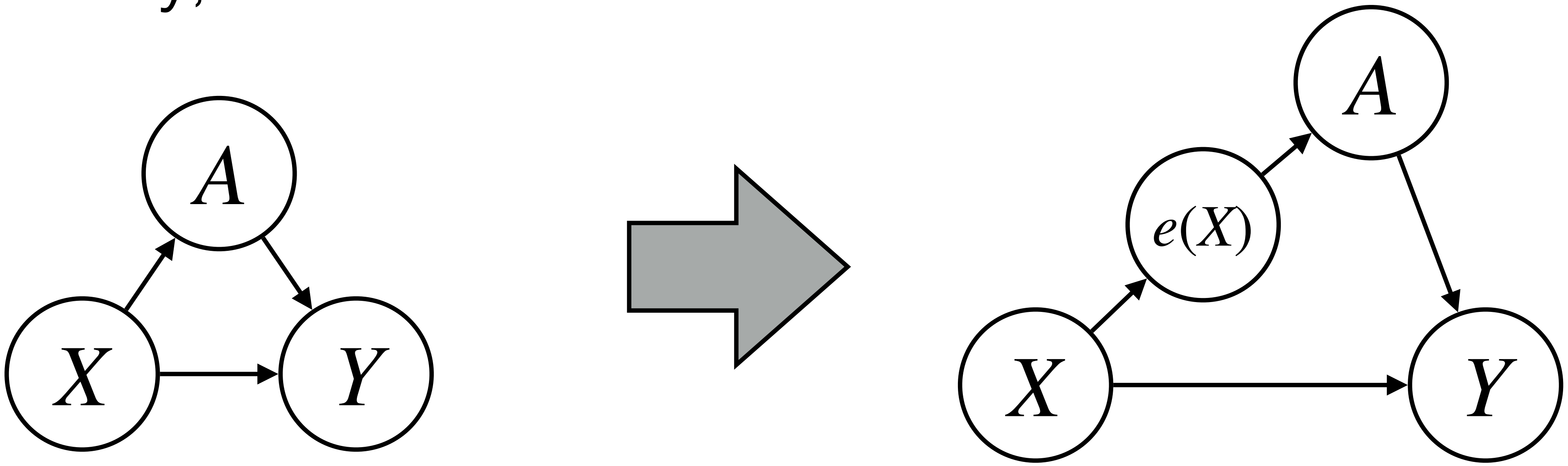
Stuart (2010)

Observed confounders: Matching

- Propensity scores break the path between features and treatment. That is,

$$A \perp X | e(X)$$

- Graphically,



- Except in rare cases, propensity scores are modeled or estimated

Sharma and Kiciman (2018)

Stuart (2010)

Observed confounders: Matching

Propensity score matching:

1. Estimate e(X) using supervised learning
   - Conventionally, logistic regression is used, but other models are fine…
   - But, the score must be **well-calibrated**. That is, it is more important to correctly estimate the probability of treatment than to achieve the highest possible accuracy.
2. Distance is the difference between the propensity scores:

$$Distance(x_i, x_j) = |\hat{e}(x_i) - \hat{e}(x_j)|$$

Sharma and Kiciman (2018)

Stuart (2010)

Observed confounders: Matching

- **Question:** What is my propensity scores are not accurate? (i.e. we can't distinguish treated and untreated)
- **Answer:** That's ok. The role of the propensity score is to balance covariates, not predict treatment.

- **Question:** What is my propensity scores are very accurate? (i.e. we can distinguish treated and untreated)
- **Answer:** This implies a potential positivity violation. Any effect we observe could be due to either the treatment or the correlated covariates.
  - **Don't** dumb down your model or exclude features.

Sharma and Kiciman (2018)

Stuart (2010)

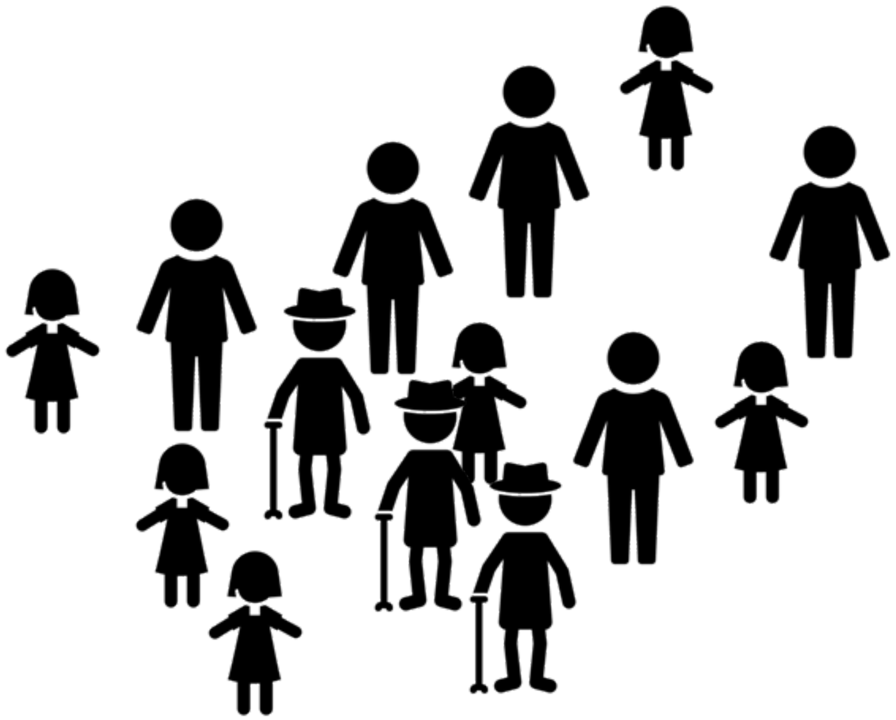
Observed confounders: Matching

- **Feature matching vs propensity score matching**:
  - Feature matching requires specifying a distance, while propensity score matching requires a model.
    - Both may introduce bias.

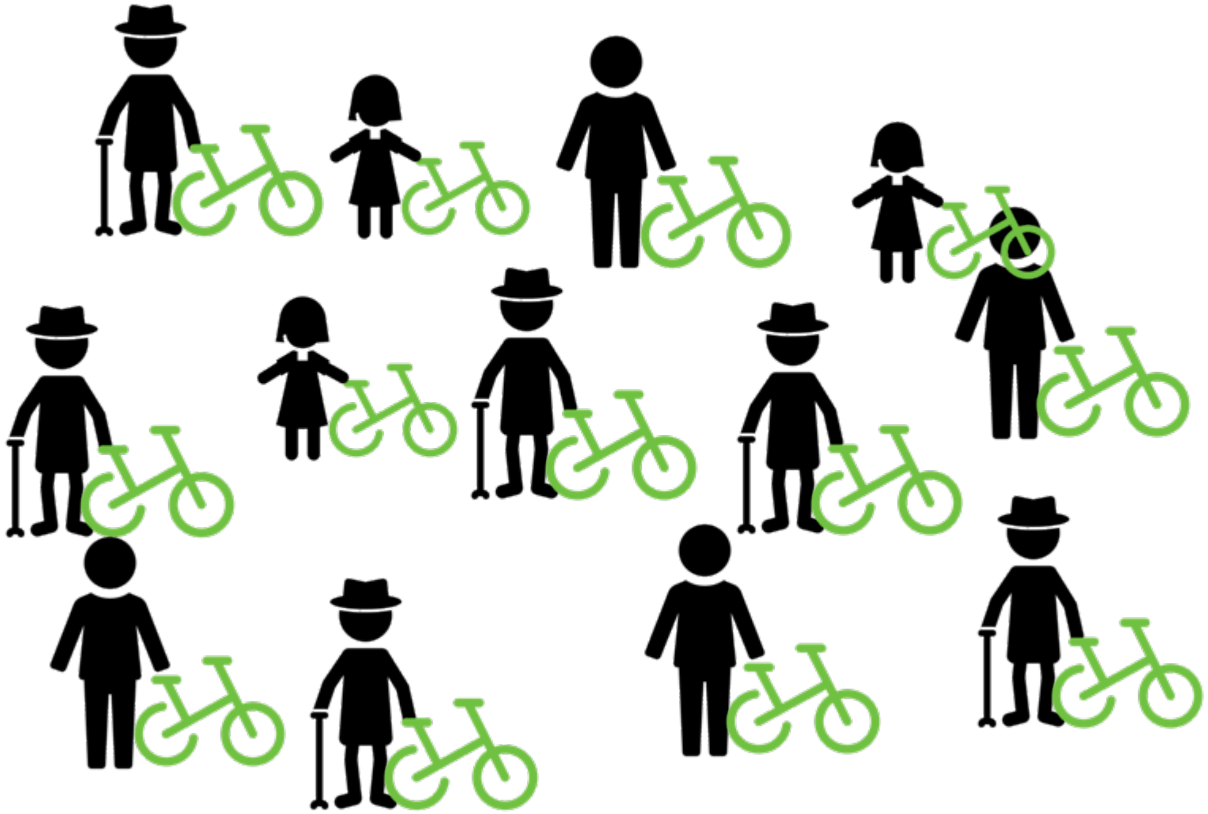
Sharma and Kiciman (2018)

Stuart (2010)

Observed confounders: Stratification



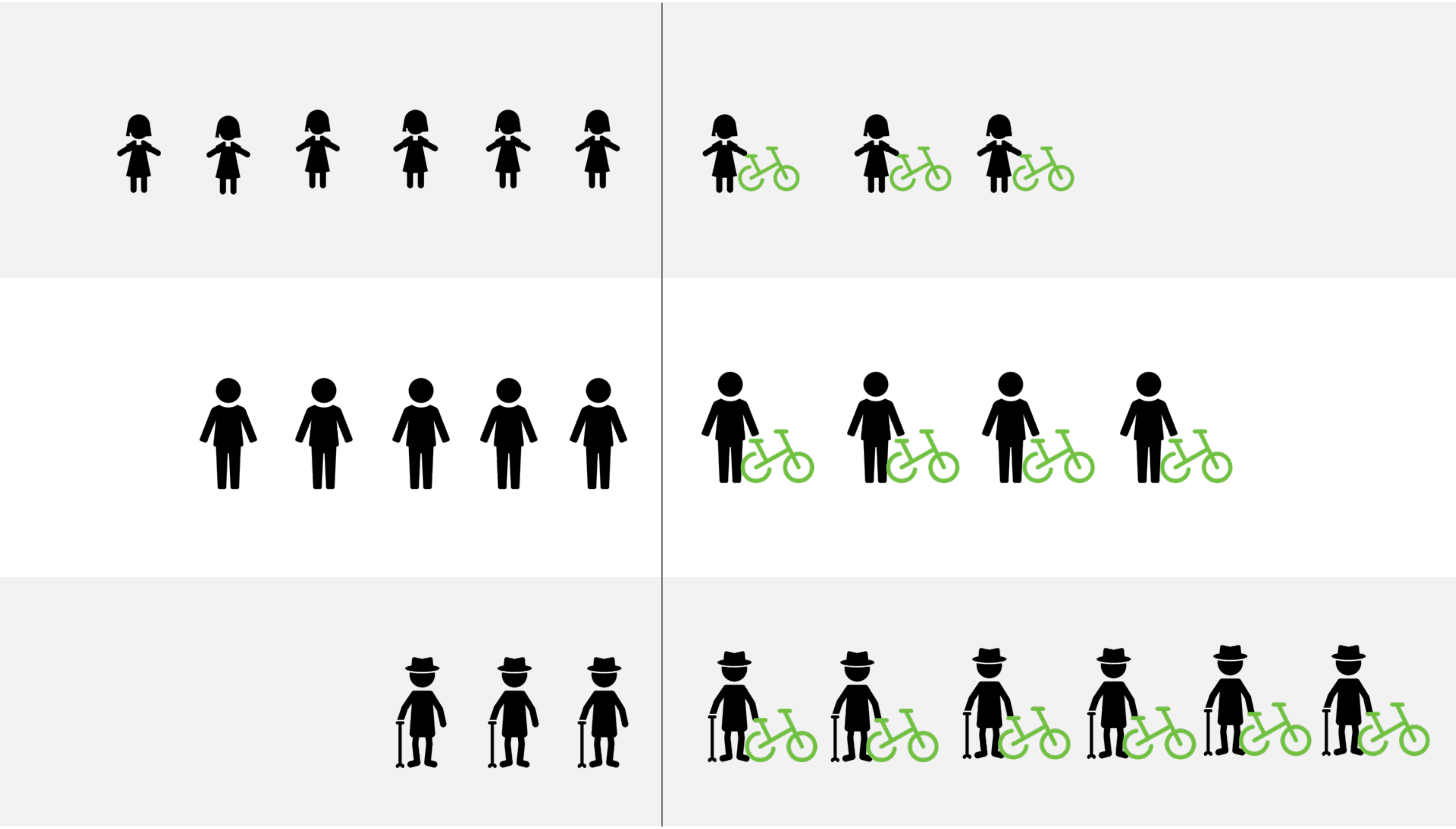
No exercise



Exercise

Sharma and Kiciman (2018)

Observed confounders: Stratification



No exercise

Exercise

Sharma and Kiciman (2018)

Observed confounders: Stratification

- Matching individuals generalizes into matching subpopulations

- Stratification identifies subpopulations with similar covariate distributions

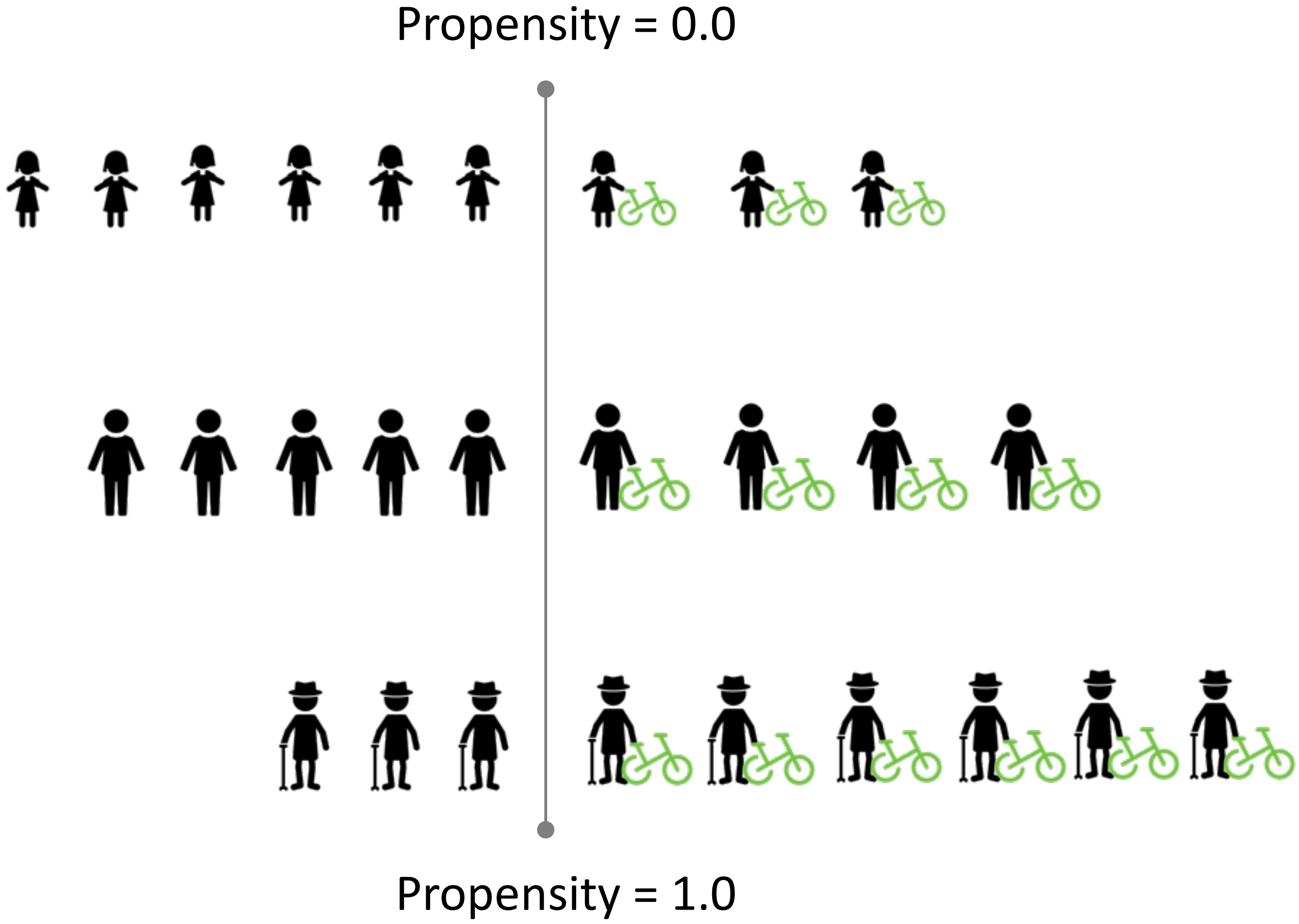
- **Question:** How do we pick strata?
  - Strata should have equal sizes
  - Each stratum should contain enough examples to reliably estimate treatment effect within it

Sharma and Kiciman (2018)

Observed confounders: Stratification

Propensity score stratification:

1. Estimate propensity score
2. Split sample into equal-sized groups based on propensity scores
3. Calculate treatment effect as the average of within strata treatment effects



Sharma and Kiciman (2018)

Observed confounders: Weighting

- **Intuition:** Count an individual more if she was unlikely to receive treatment (probability is low —> weight is high) and vice versa
- **Use case:** When we know (or can estimate) the probability of treatment P(A|X)
- **Inverse Probability Weighted Estimator (IPWE):**

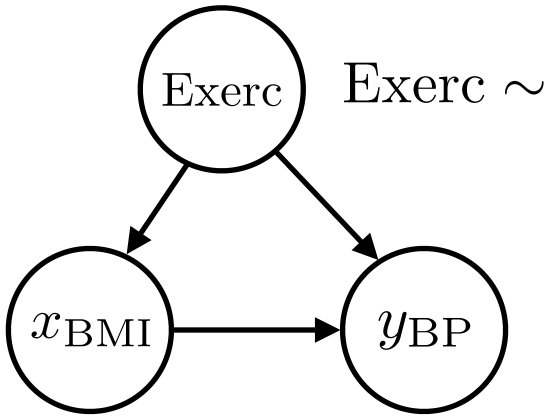
$$\mu^{a}_{weight} = \frac{1}{n} \sum_{i} \frac{\mathbf{1}[a_i = a] y_i}{\hat{P}(A = a_i \,|\, X = x_i)}$$

- **Assumption:** The estimated treatment model is correct
- **Warning:** Has high variance when probability of treatment estimates are close to zero or one

Hernán and Robins (2018)

Observed confounders: Weighting

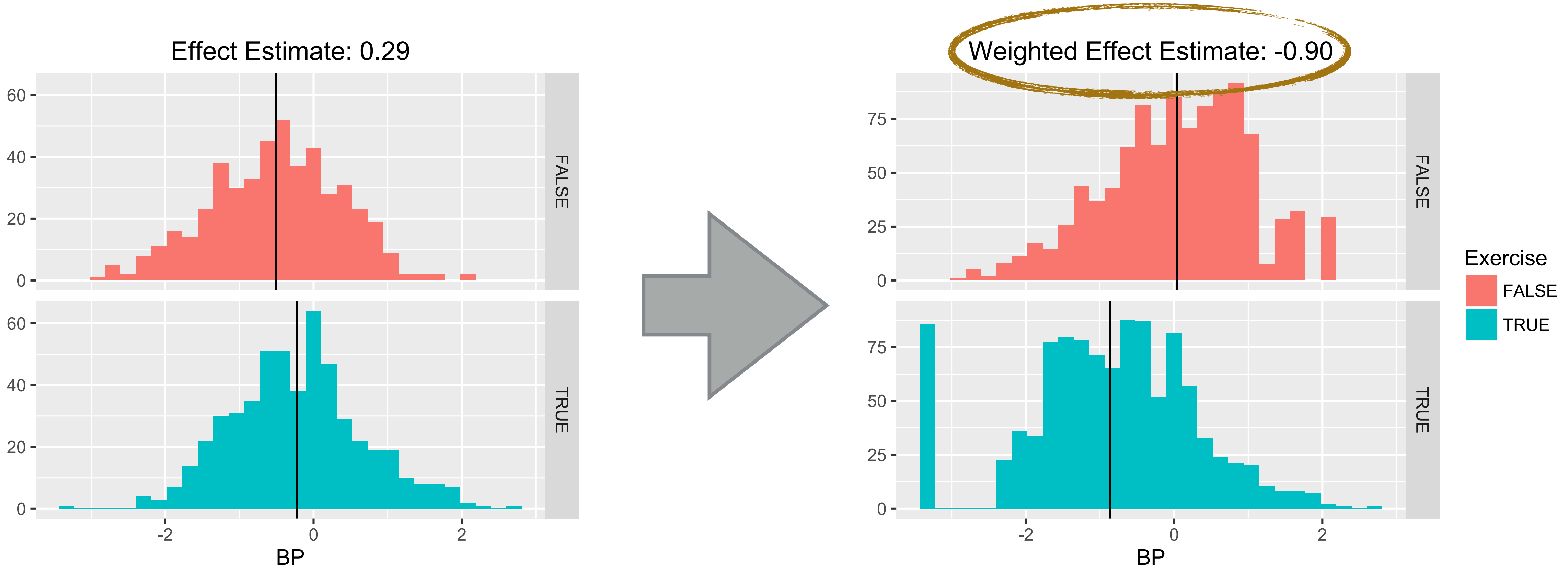
- Exercise example:



$$\text{Exerc} \sim \text{Bern}\left(\frac{1}{1+e^{-2x_{\text{BMI}}}}\right)$$

$$x_{\text{BMI}} \sim \mathcal{N}(0,1)$$

$$y_{\text{BP}} \sim \mathcal{N}(x_{\text{BMI}}, 0.4) - 0.8 \cdot \text{Exerc}$$



Observed confounders: Potential Outcomes

- To formalize, define two distinct random variables:
  - Y(a) : blood pressure *with* exercise
  - Y(b) : blood pressure *without* exercise
- More generally, we can index a set of random variables using a set of actions/treatments:

$$\{Y(a) : a \in \mathcal{A}\}$$

- Offers a way to reason about *counterfactuals*.
- **Goal: learn statistical models to estimate potential outcomes**

Potential Outcomes: Use models to adjust for bias

- Assume models of potential outcomes given covariates

$$\{\mathrm{P}(Y(a) \mid \mathbf{X} = \mathbf{x}) : a \in \mathcal{A}\}$$

- We can use them to adjust for bias in observational data
- Key idea: use models to “simulate” an RCT

Rubin 1977 **Robins 1986**

Recall: Critical Assumptions

- To learn the potential outcome models, we will use three important assumptions:
- (1) Consistency
  - Links observed outcomes to potential outcomes
- (2) Treatment Positivity
  - Ensures that we can learn potential outcome models
- (3) No unmeasured confounders (NUC)
  - Ensures that we do not learn biased models

Potential Outcomes: Learning models from data

- To simulate data from a new policy, we need to learn the potential outcome models
  - If we have an observational dataset where assumptions 1-3 hold, then this is possible!
- Assumptions allow estimation of potential outcomes from (observational) data:

$$
\begin{aligned}
\mathrm{P}(Y(a) \mid \mathbf{X} = \mathbf{x}) &= \mathrm{P}(Y(a) \mid \mathbf{X} = \mathbf{x}, A = a) && \text{(A3)} \\
&= \boxed{\mathrm{P}(Y \mid \mathbf{X} = \mathbf{x}, A = a)} && \text{(A1)}
\end{aligned}
$$

Estimation requires a statistical model for estimating conditionals

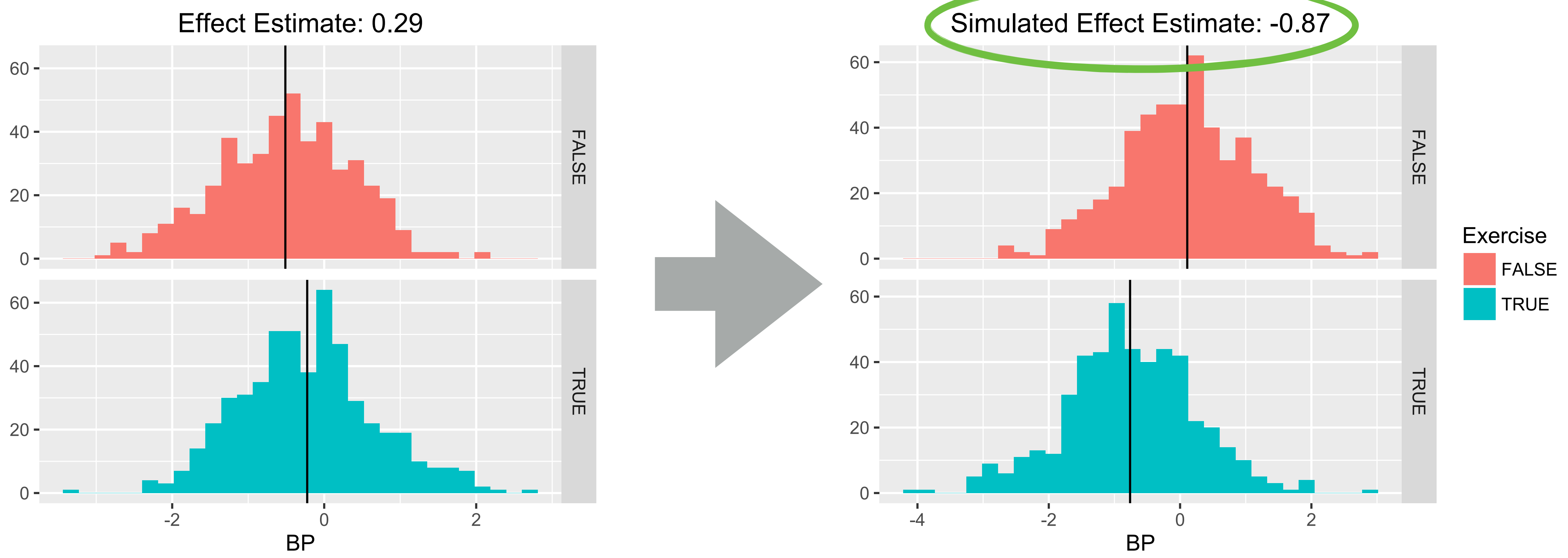
Exercise and Blood Pressure

- Returning to our exercise and blood pressure example

- We fit a model for blood pressure given exercise and BMI

- With estimated models, treatment effects are estimated as:

$$\mathbb{E}[Y(1) - Y(0)] = \frac{1}{N}\sum_{n=1}^{N}(Y_n(1) - Y_n(0))$$



Choosing a method

- **Question:** All three methods allow us to account for the effect of observed confounders, so which one should you use?
- **Answer:** Depends on the problem.
- Assuming **correct models** all three are equivalent, but **models are always wrong**.
- If you have more confidence in one model over another, then you should use the corresponding method.

Choosing a method

- **Question:** But what if I’m not sure which model is best?
- **Answer:** Use all three and compare the results. If all three methods result in similar effect estimates, then this gives you further evidence that your models are (approximately) correct.
- If the methods disagree, then you have bias somewhere and need to think more carefully about your models.
- This is an example of **sensitivity analysis** (more on this later).

Hernán and Robins (2018)

Doubly robust estimation

- Both IPWE and standardization require specifying models and both methods can fail if the models are misspecified.
- **Doubly robust (DR)** methods combine both approaches.
- If either the propensity score model **or** the outcome model are correct, DR will be correct.

Doubly robust estimation

1. Estimate a propensity model: $\hat{e}(x)$

2. Estimate an outcome model: $\hat{E}[Y|A,X]$

3. Combine as:

$$\mu^a_{DR} = \frac{1}{n}\sum_i \left(\frac{\mathbf{1}[a_i = a]y_i}{\hat{e}(x_i)} - \frac{\mathbf{1}[a_i = a] - \hat{e}(x_i)}{\hat{e}(x_i)}\hat{E}[Y|A = a, X = x_i] \right)$$

- If either model is correct, then DR is correct

- If both models are wrong, then DR may be <u>more</u> biased than either individually

Davidian (2017)

Doubly robust estimation

- Why does it work?
- By the law of large numbers, μ^a_{DR} estimates:

$$
\begin{aligned}
& E\left[\frac{\mathbf{1}[A=a]Y}{\hat{e}(X)} - \frac{\mathbf{1}[A=a]-\hat{e}(X)}{\hat{e}(X)}\hat{E}[Y|A=a,X]\right] \\
&\quad = E\left[\frac{\mathbf{1}[A=a]Y^a}{\hat{e}(X)} - \frac{\mathbf{1}[A=a]-\hat{e}(X)}{\hat{e}(X)}\hat{E}[Y|A=a,X]\right] \\
&\quad = E[Y^a] + \boxed{E\left[\frac{\mathbf{1}[A=a]-\hat{e}(X)}{\hat{e}(X)}(Y^a - \hat{E}[Y|A=a,X])\right]}
\end{aligned}
$$

- So, if the term on the right goes to zero, then μ^a_{DR} is a consistent (correct in expectation) estimator.
- We will show this for both cases: when the propensity model is correct or the outcome model is correct.

Davidian (2017)

Doubly robust estimation

- First, assume $\hat{e}(x) = P(T = a \,|\, X = x)$, then:

$$E\left[\frac{\mathbf{1}[A=a] - \hat{e}(X)}{\hat{e}(X)}(Y^a - \hat{E}[Y|A=a,X])\right]$$

$$= E\left[\frac{E[\mathbf{1}[A=a]|Y^a,X] - \hat{e}(X)}{\hat{e}(X)}(Y^a - \hat{E}[Y|A=a,X])\right]$$

$$= E\left[\frac{P(A=a|X) - \hat{e}(X)}{\hat{e}(X)}(Y^a - \hat{E}[Y|A=a,X])\right]$$

$$= E\left[\frac{\hat{e}(X) - \hat{e}(X)}{\hat{e}(X)}(Y^a - \hat{E}[Y|A=a,X])\right] = 0$$

Davidian (2017)

Doubly robust estimation

- Next, assume $\hat{E}[Y|A=a,X] = E[Y|A=a,X]$, then:

$$E\left[\frac{\mathbf{1}[A=a]-\hat{e}(X)}{\hat{e}(X)}(Y^a - \hat{E}[Y|A=a,X])\right]$$

$$= E\left[\frac{\mathbf{1}[A=a]-\hat{e}(X)}{\hat{e}(X)}(E[Y^a|A,X] - \hat{E}[Y|A=a,X])\right]$$

$$= E\left[\frac{\mathbf{1}[A=a]-\hat{e}(X)}{\hat{e}(X)}(E[Y^a|X] - E[Y^a|X])\right] = 0$$

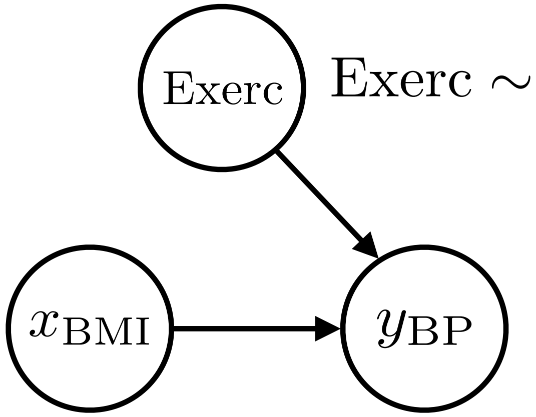
- So, the DR estimator is correct if either the treatment or outcome model is correct.

Davidian (2017)

Conditional causal inference

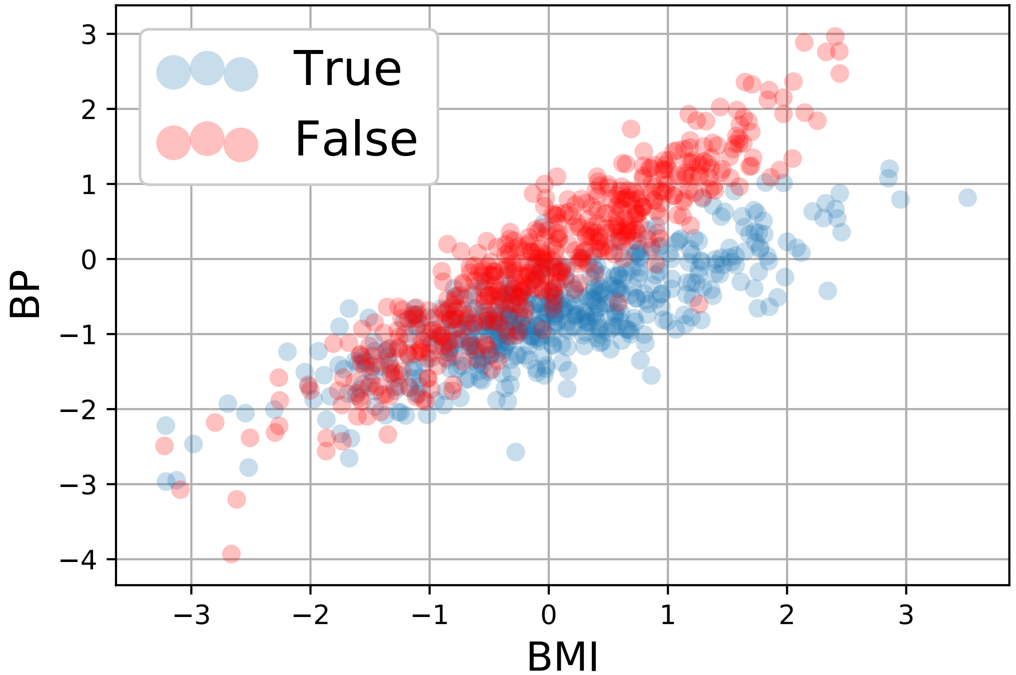
- Sometimes we want to reason about the causal effect of treatment **A** on outcome **Y** within a subpopulation defined by **X**.
- **Example:** Does exercise have a different effect on BP for people with high BMI versus people with low BMI?

$\text{Exerc} \sim \text{Bern}(0.5)$



$x_{\text{BMI}} \sim \mathcal{N}(0,1)$

$y_{\text{BP}} \sim \mathcal{N}(x_{\text{BMI}}, 0.4) - 0.8 \cdot \text{Exerc} - \boxed{0.5 \cdot \text{Exerc} \cdot x_{\text{BMI}}}$



Conditional causal inference

Our primary quantities of interest generalize to the case where we want to **condition on a subset of features $\mathbf{X^c}$**:

- Conditional expected outcome:

Hernán and Robins (2018)

$$E[Y^a \mid X^c]$$

- Conditional additive treatment effect:

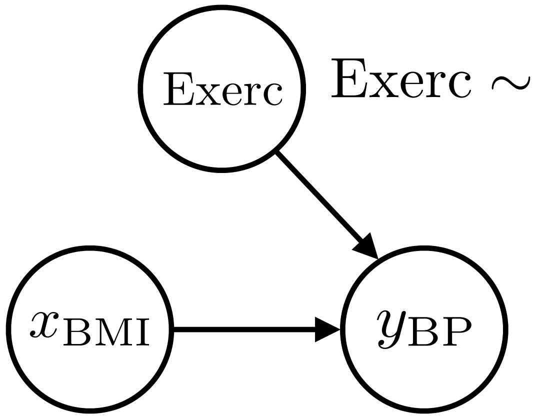
$$E\left[Y^a \mid X^c\right] - E\left[Y^{a'} \mid X^c\right]$$

- Conditional relative treatment effect:

$$E\left[Y^a \mid X^c\right] / E\left[Y^{a'} \mid X^c\right]$$

Conditional causal inference

- **Example:** Does exercise have a different effect on BP for people with high BMI versus people with low BMI?
- Data generated by an RCT

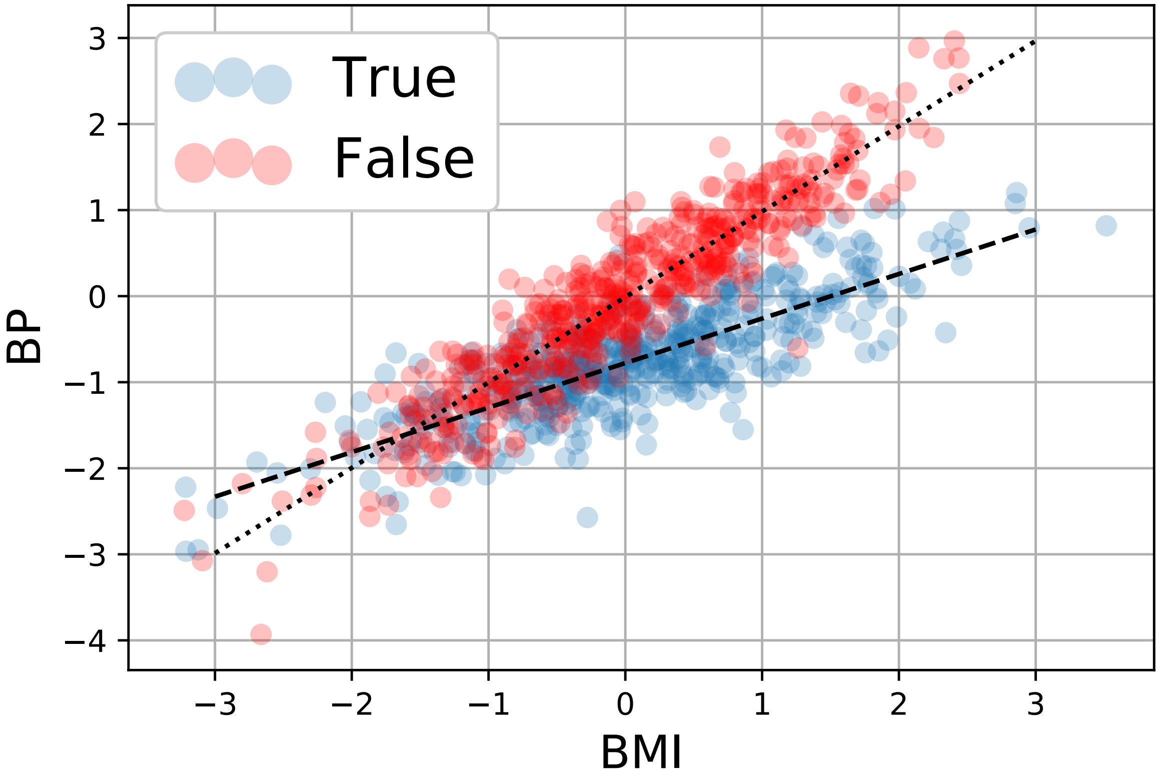


$\text{Exerc} \sim \text{Bern}(0.5)$

$x_{\text{BMI}} \sim \mathcal{N}(0,1)$

$y_{\text{BP}} \sim \mathcal{N}(x_{\text{BMI}}, 0.4) - 0.8 \cdot \text{Exerc} - 0.5 \cdot \text{Exerc} \cdot x_{\text{BMI}}$

- Conditional expected outcomes are functions of X:



Conditional causal inference

- Similarly, the methods for handling observed confounders also generalize.
- **Example:** Standardization / Marginalizing w.r.t. inputs into the potential outcome model.

$$\mu^{a,x^c}_{stand} = \frac{1}{n} \sum_i \hat{E}\left[Y | A = a, X^c = x^c, X^u = x^u_i\right]$$

- $\boldsymbol{X^u}$ is the vector of features we are <u>not</u> conditioning on.
- All methods may now require an outcome model:

$$\hat{E}\left[Y | A = a, X^c = x^c\right]$$

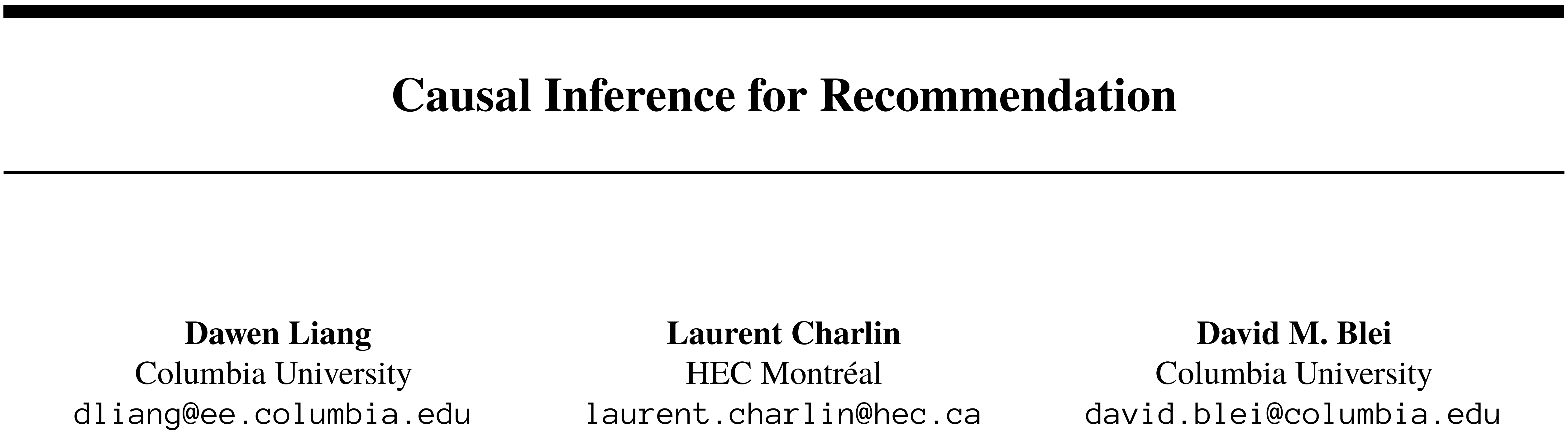
Hernán and Robins (2018)

Example Machine Learning applications…

- A few examples of applying the above methods to challenging decision-making applications…

Example Machine Learning applications…

- **Application:** Recommendation systems
- **Treatment:** Recommending a specific item.
- **Outcome:** If a user click on or buys the item.
- **Why is it complex?**
    - Recommendation typically performed using matrix factorization.
    - The data is generated using an existing recommendation algorithm.



Causal Inference for Recommendation

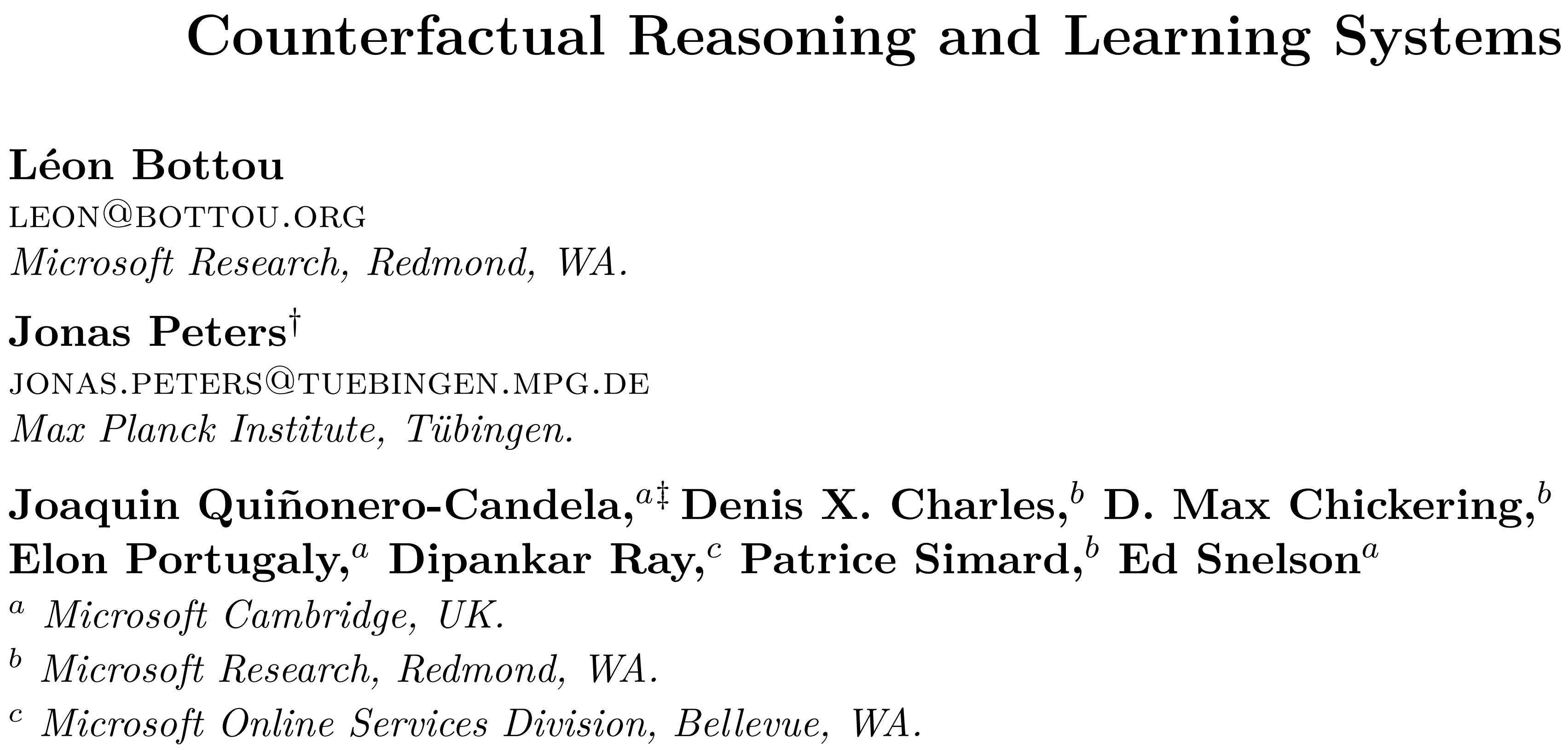
Dawen Liang
Columbia University
dliang@ee.columbia.edu

Laurent Charlin
HEC Montréal
laurent.charlin@hec.ca

David M. Blei
Columbia University
david.blei@columbia.edu

Example Machine Learning applications…

- **Application:** Selecting and placing ads.
- **Treatment:** Ad choice and placement.
- **Outcome:** Whether a user clicks an ad.
- **Why is it complex?**
  - Treatment involves
  - Covariates include complex structures such as search histories, text from emails/social media, etc.
  - Decision making context evolves over time.



Counterfactual Reasoning and Learning Systems

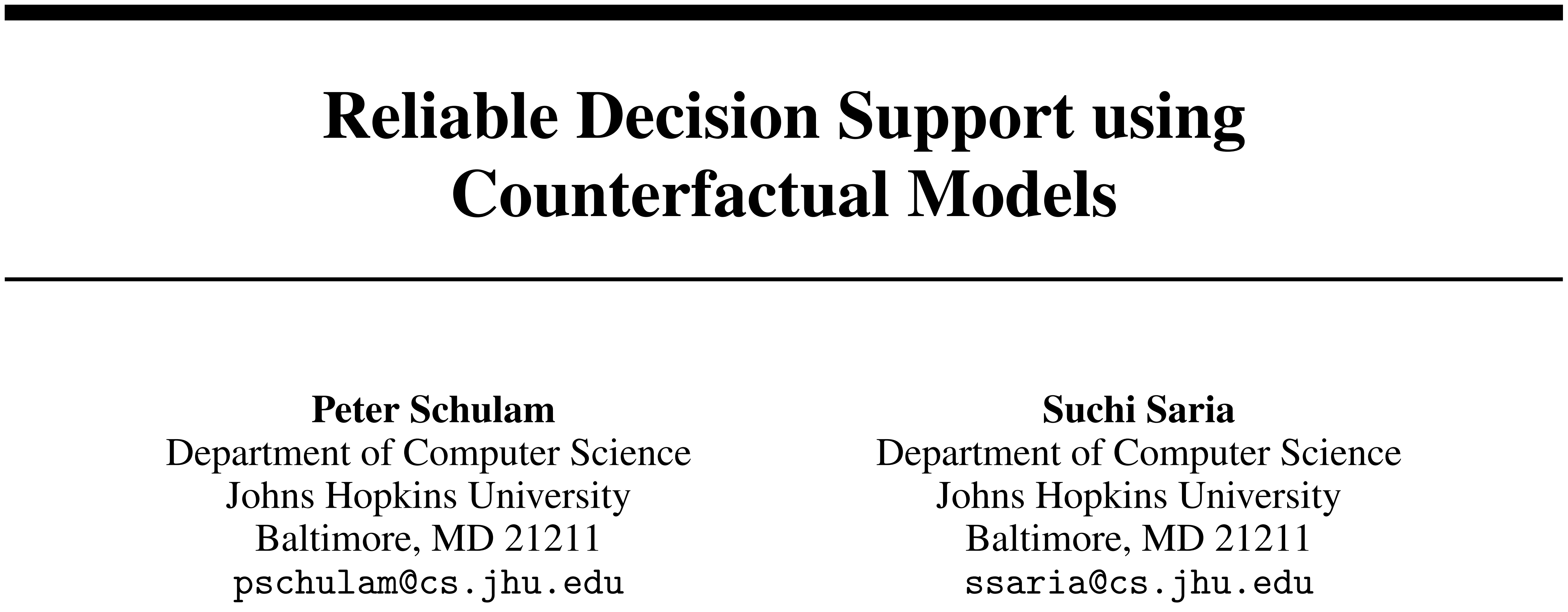
Léon Bottou
LEON@BOTTOU.ORG
Microsoft Research, Redmond, WA.

Jonas Peters[†]
JONAS.PETERS@TUEBINGEN.MPG.DE
Max Planck Institute, Tübingen.

Joaquin Quiñonero-Candela,[a‡] **Denis X. Charles,**[b] **D. Max Chickering,**[b]
Elon Portugaly,[a] **Dipankar Ray,**[c] **Patrice Simard,**[b] **Ed Snelson**[a]
[a] *Microsoft Cambridge, UK.*
[b] *Microsoft Research, Redmond, WA.*
[c] *Microsoft Online Services Division, Bellevue, WA.*

Example Machine Learning applications…

- **Application:** Making medical treatment decisions.
- **Treatment:** Type and timing of medications.
- **Outcome:** Physiologic state.
- **Why is it complex?**
  - Both the timing and type of treatment matters.
  - The data is generated under a specific treatment plan, but inferences should generalize.



Reliable Decision Support using Counterfactual Models

Peter Schulam
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21211
pschulam@cs.jhu.edu

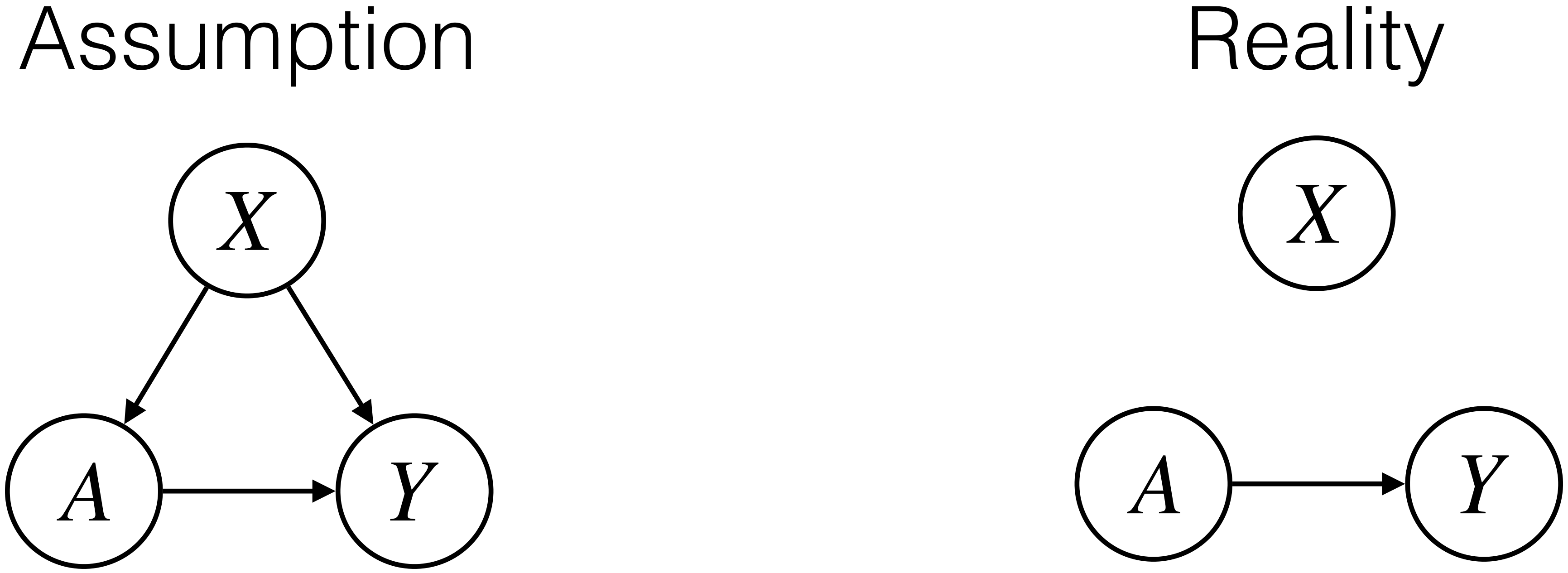
Suchi Saria
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21211
ssaria@cs.jhu.edu

Testing your assumptions: Sensitivity analysis

- Remember, we can use the previous methods on any data.
- Only our assumptions allow us to interpret the results as causal effects.
- It is critical to verify our assumptions.
- Sensitivity analysis allows us to falsify our assumptions.
- **Basic idea:** Modify the data in a way that should have predictable effects, then test whether the results match our expectations.

Testing your assumptions: Sensitivity analysis

- **Potential problem:** Overfitting our features
- **Test:** Add in random observed features.



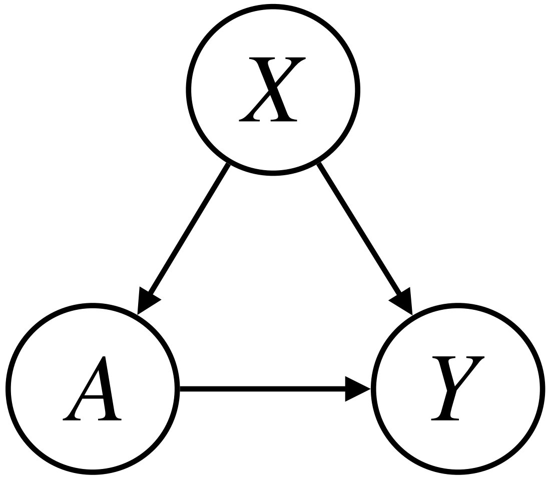
- **What assumption is wrong?**
  - That X is a confounder
- **What do we expect?**
  - Our effect estimates should not change significantly. If it does, we are overfitting to the data.

Sharma and Kiciman (2018)

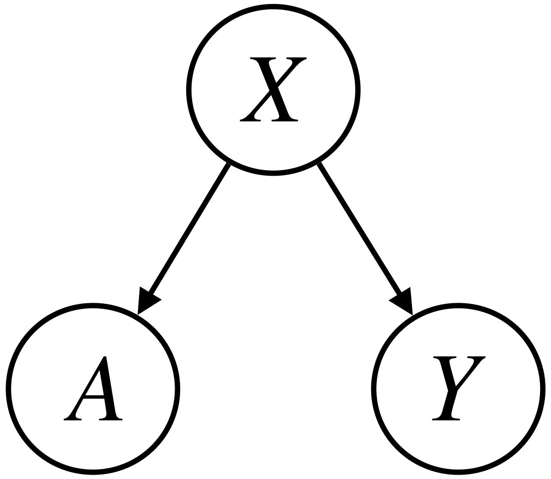
Testing your assumptions: Sensitivity analysis

- **Potential problem:** Overfitting to variation in the outcome
- **Test:** Replace the outcome with a placebo.

Assumption



Reality



- **What assumption is being violated?**
  - That A has a causal effect on Y
- **What do we expect?**
  - No causal effect detected. If we did detect an effect, the result suggests that we are overfitting to variation in the outcome.

Sharma and Kiciman (2018)

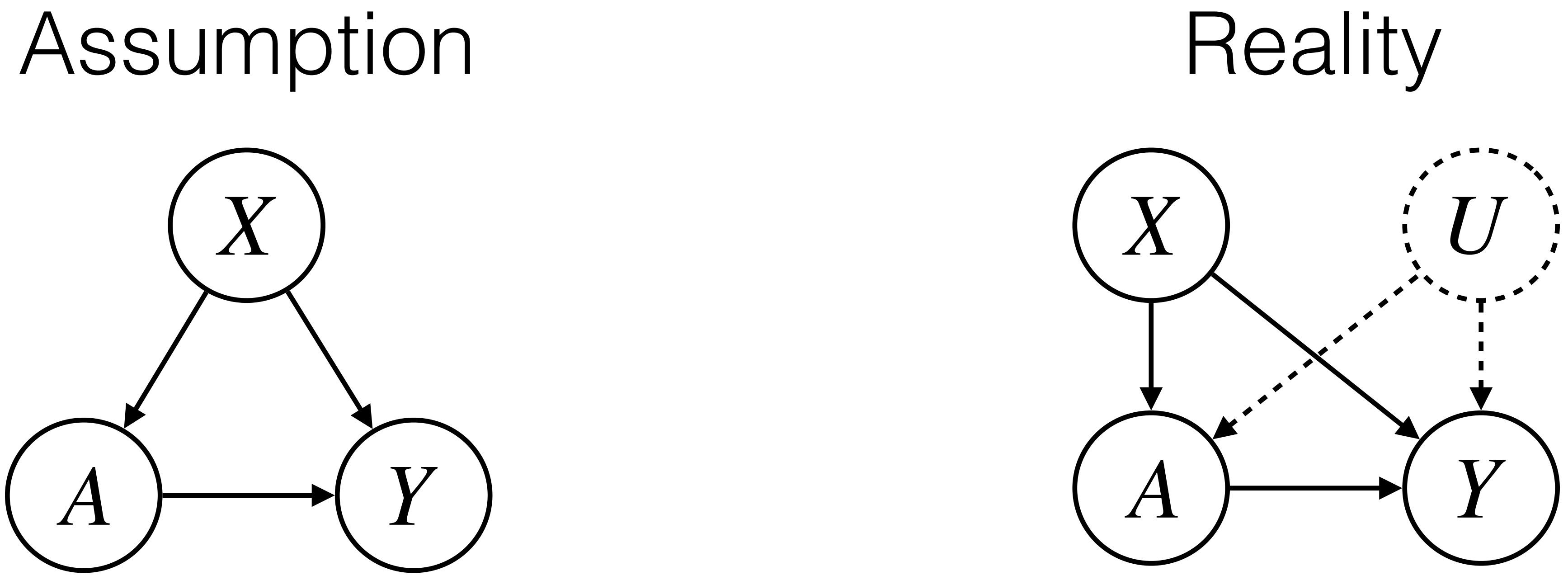
Testing your assumptions: Sensitivity analysis

- **Potential problem:** Sensitivity to the particular sample.
- **Test:** Re-run analysis on random bootstrap samples of the data
- This is a change in the sample, not the underlying model
- **What do we expect?**
  - Small changes to the effect estimate. Large variation implies wide confidence intervals.

Sharma and Kiciman (2018)

Testing your assumptions: Sensitivity analysis

- **Potential problem:** Unobserved confounder
- **Test:** Add synthetic **unobserved** confounders and vary their causal effect on **A** and **Y**.



- **What assumption is wrong?**
  - No unobserved confounders

Rosenbaum (2002)

Carnegie, Harada, and Hill (2016)

Testing your assumptions: Sensitivity analysis

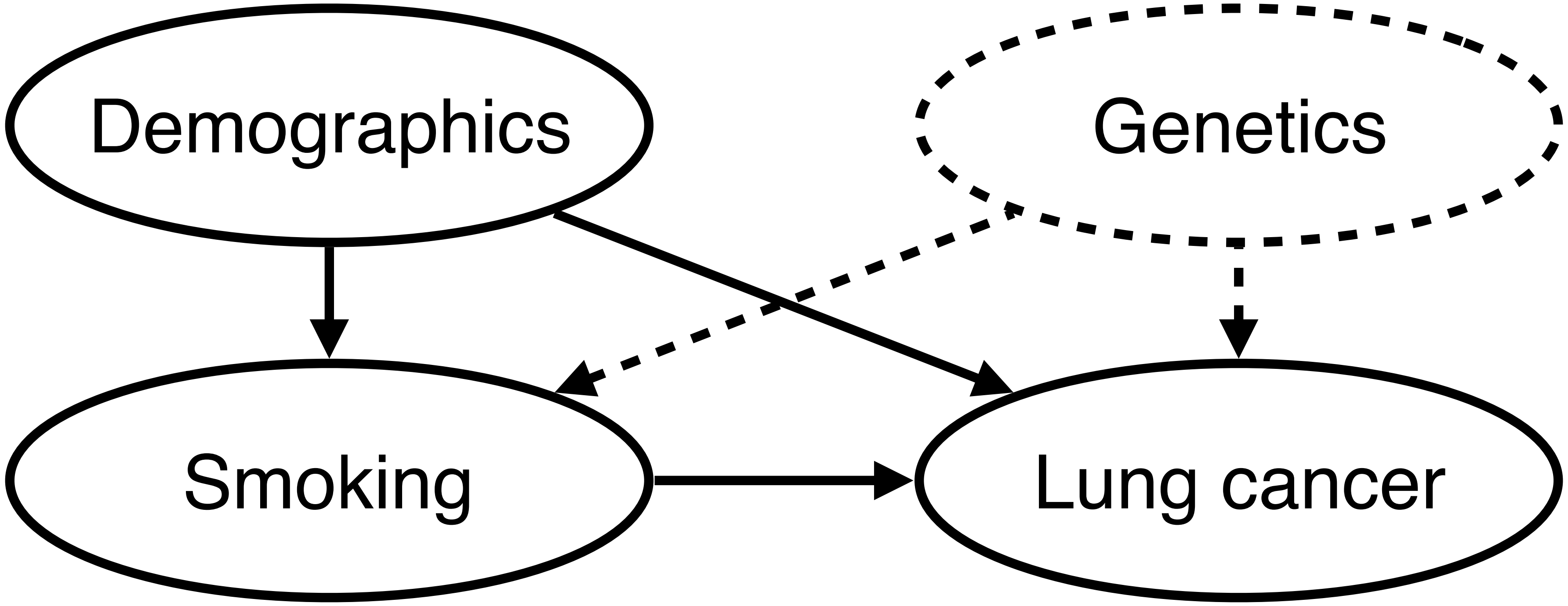
- Let τ_A be a measure of the effect size of U on A.
- Let τ_Y be a measure of the effect size of U on Y.
- τ_A and τ_Y will often be regression coefficients.
- **Basic idea:**
  - For different values of τ_A and τ_Y:
    1. Simulate a confounder with these effect sizes.
    2. Rerun your analysis, now including U as a confounder in the model.
    3. Check if A still has a causal effect on Y.
  - Output: Minimum possible effect sizes for U that lead to no causal effect of A on Y
- The specifics of simulating U will depend on the form of the outcome model, $E[Y \mid A, X]$.

Rosenbaum (2002)

Carnegie, Harada, and Hill (2016)

Testing your assumptions: Sensitivity analysis

Example:



Cornwell (1959) showed that the effect of Genes had to be **8 times** that of any known confounder for the effect to go to zero.

Cornwell (1959)

Sharma and Kiciman (2018)

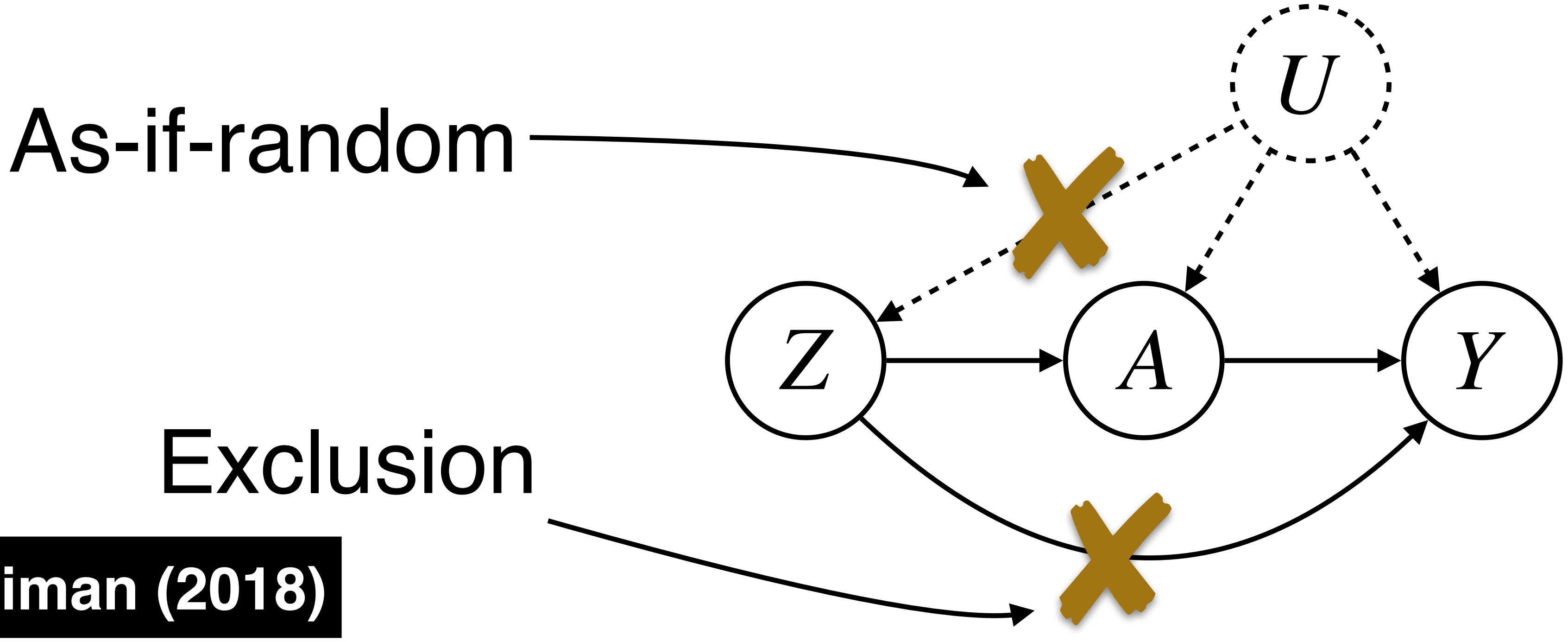
Natural experiments

- What can we do if we can't run a trial and we can't assume NUC?
- Sometimes, we can find observational data that approximates an experiment. This is called a **natural experiment**.
- **Example:** The Oregon insurance experiment
  - Oregon's Medicaid expansion was administered by lottery
  - Equivalent to a RCT for the effects of receiving Medicaid
  - Perfect natural randomization like this is uncommon

Baiker et al. (2013)

Natural experiments: Instrumental variables

- This idea can be formalized using **instrumental variables (IVs)**
- IVs have a causal effect on the treatment that allows us to emulate an RCT
- IVs must satisfy two assumptions:
  1. As-if-random: the IV must not be effected by unobserved confounders
  2. Exclusion: the IV cannot effect the outcome except through the treatment

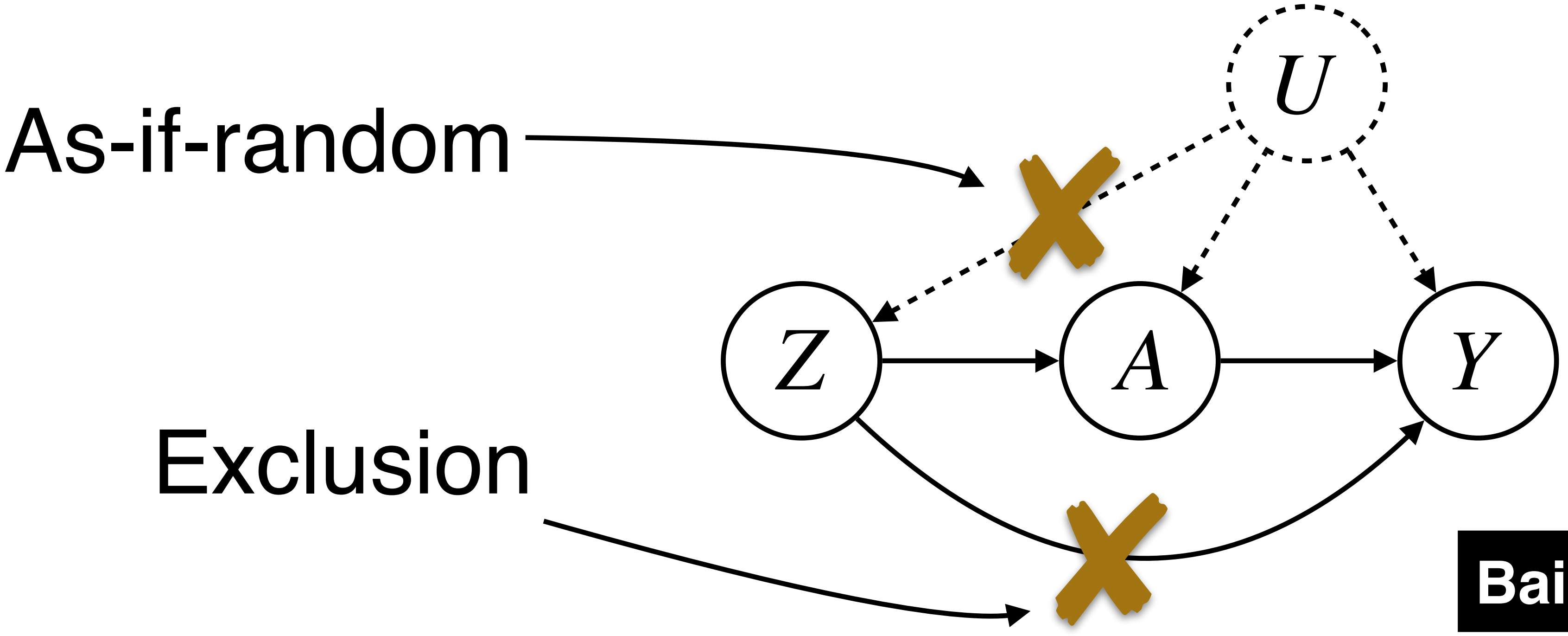


Pearl (2009)

Sharma and Kiciman (2018)

Natural experiments: Instrumental variables

- **Example:** The Oregon insurance experiment
  - In this case, assignment in the lottery is the instrumental variable.
  - It trivially satisfies both as-if-random and exclusion assumptions because it was completely random.



Baiker et al. (2013)

Natural experiments: Regression discontinuity

- One particularly common kind type of IV is called a **regression discontinuity**.
- Regression discontinuities happen when an arbitrary threshold is used to determine the treatment variable.
- Samples just above and just below the threshold are assumed to be equivalent, except for the treatment.
- **Example:** Families above a certain income receive health insurance.
- The difference between being above vs. below the threshold is assumed to be so small that it is caused by natural variation (as-if-random) and it does not effect the outcome (exclusion).

Sharma and Kiciman (2018)

Takeaways

- Many decision support applications employ predictive modeling but we're currently ignoring fundamental issues associated with undesirable biases in the learned models—> this leads to poor decisions
- Reformulate as "what-if" questions

- Challenges associated w/ answering "what-if" questions from retrospective data
  - A few simple methods for estimating causal effects
  - A few simple tricks for testing whether your estimates are good and understand conditions under which you can estimate causal effect
  - Understand what a potential outcome model is
- Next: we will use these principles to tackle machine learning applications of prediction and decision support

Day 2

Takeaways from Day 1

- Many decision support applications employ predictive modeling but we're currently ignoring fundamental issues associated with undesirable biases in the learned models—> this leads to poor decisions
- Reformulate as "what-if" questions

- Challenges associated w/ answering "what-if" questions from retrospective data
  - A few simple methods for estimating causal effects
  - A few simple tricks for testing whether your estimates are good and understand conditions under which you can estimate causal effect
  - Understand what a potential outcome model is
- Next: we will use these principles to tackle machine learning applications of prediction and decision support

Day 2

- When learning from retrospective datasets, models may encode **unintended dataset-specific biases** that hurts quality of decision-making at test time. For example, the model may **learn relationships that are *unstable*—**associations that exist in the training data but do not hold and change at test time.
  - See examples (e.g., policy creep, domain-dependent confounding, selection bias)

- See how **knowledge of the data generating process** (i.e. causal DAG) allows us to explicitly **reason about scenarios under which we can learn stable models**. **Can we identify relationships that are stable and only learn these**?

  - Deep dive: Potential outcome models for what-if reasoning over temporal trajectories —> learns relationships between predictors and outcome that are stable across environments. Requires certain assumptions to hold.

  - Deep dive: Feature augmentation procedure that identifies and learns relationships that are stable. Applicable in settings with unmeasured confounding. Requires certain other assumptions to hold.

- Broadly, frame generalization in **terms of differences in the data generating process across environments**.

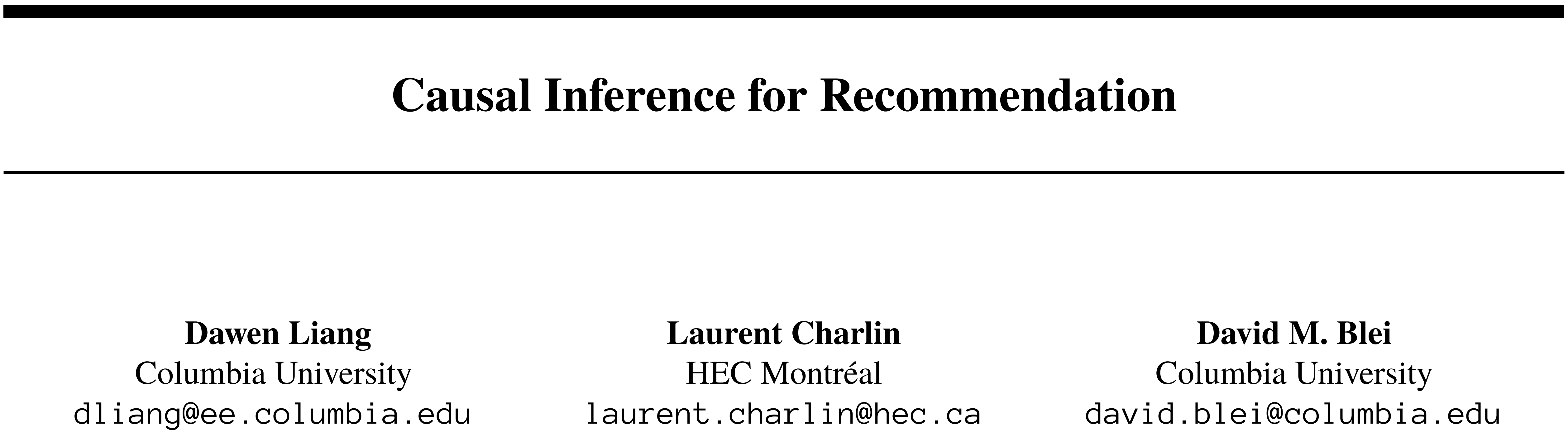
Subbaswamy and Saria, UAI 2018

Example Machine Learning applications…

- A few examples of applying the above methods to challenging decision-making applications…

Example Machine Learning applications…

- **Application:** Recommendation systems
- **Treatment:** Recommending a specific item.
- **Outcome:** If a user click on or buys the item.
- **Why is it complex?**
  - Recommendation typically performed using matrix factorization.
  - The data is generated using an existing recommendation algorithm.



Causal Inference for Recommendation

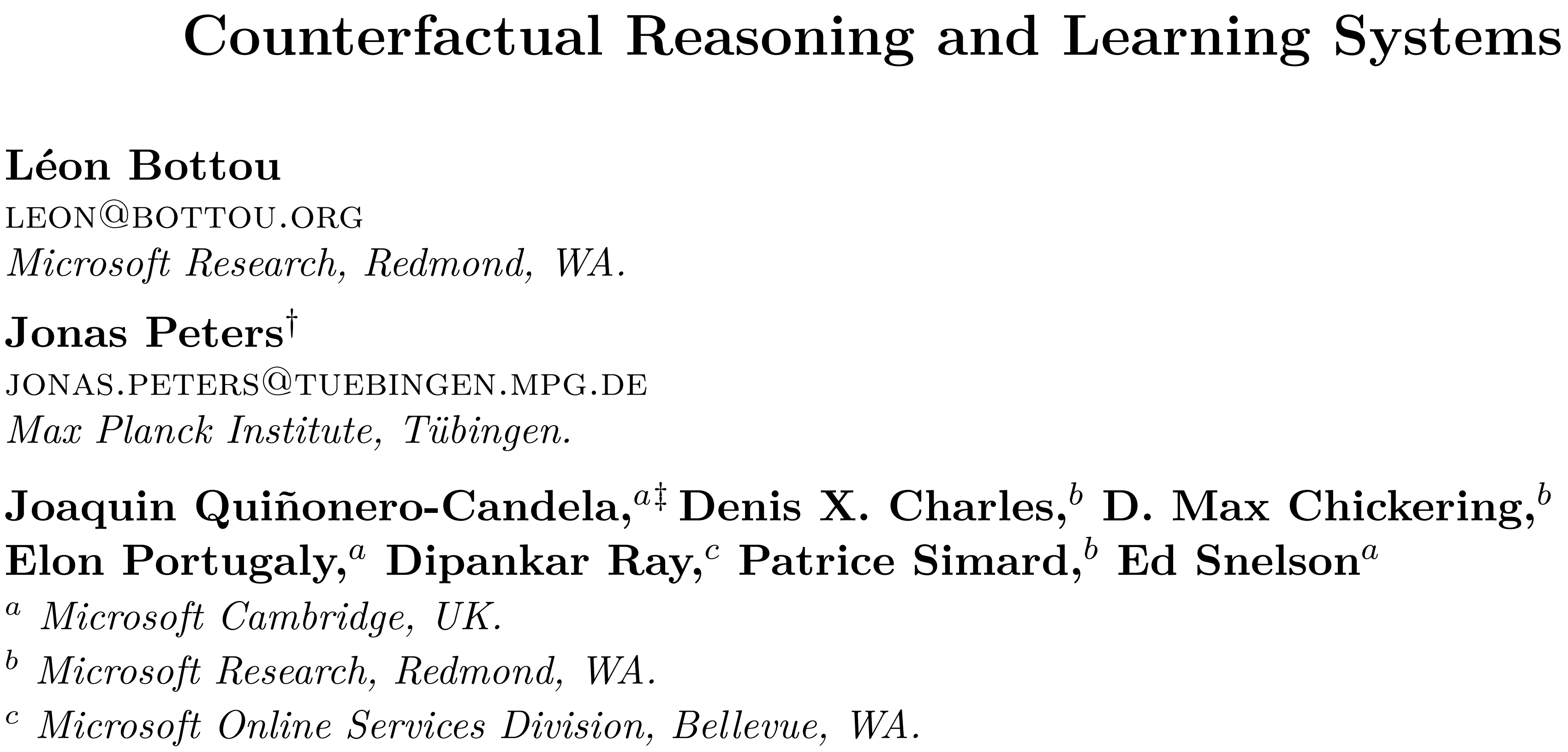
Dawen Liang
Columbia University
dliang@ee.columbia.edu

Laurent Charlin
HEC Montréal
laurent.charlin@hec.ca

David M. Blei
Columbia University
david.blei@columbia.edu

Example Machine Learning applications…

- **Application:** Selecting and placing ads.
- **Treatment:** Ad choice and placement.
- **Outcome:** Whether a user clicks an ad.
- **Why is it complex?**
  - Treatment involves
  - Covariates include complex structures such as search histories, text from emails/social media, etc.
  - Decision making context evolves over time.



Counterfactual Reasoning and Learning Systems

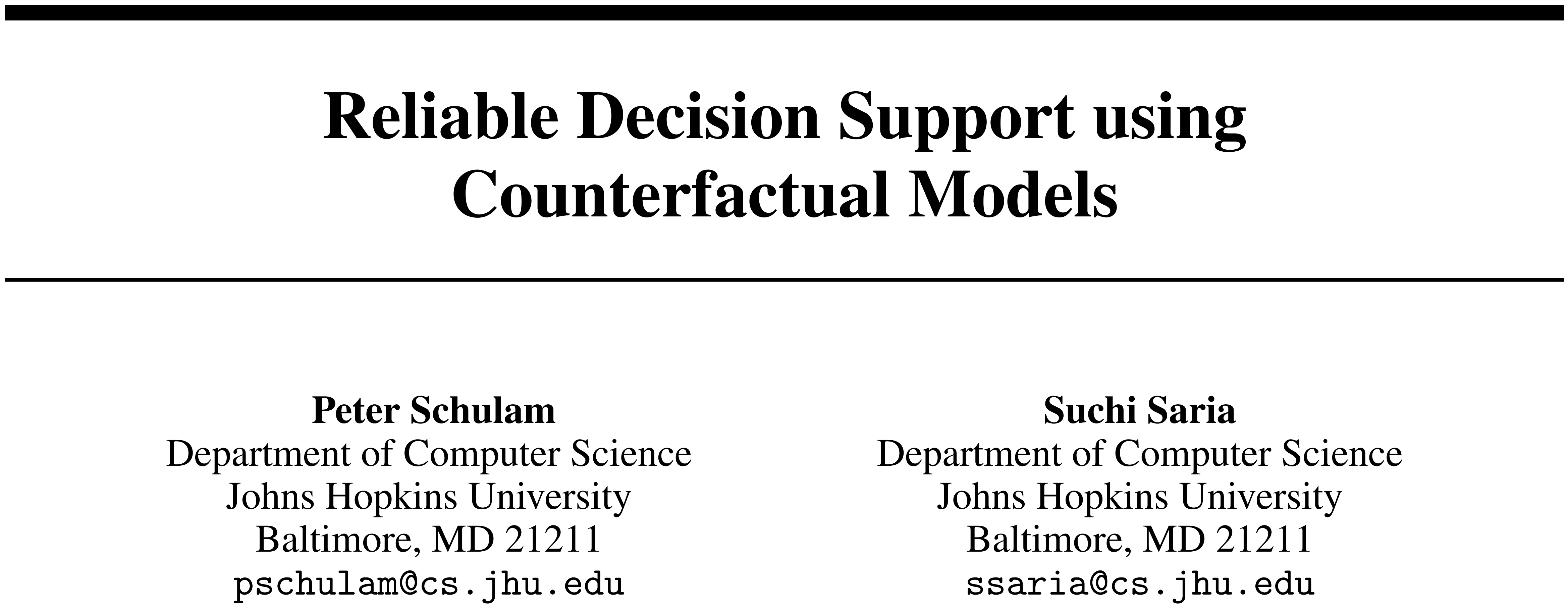
Léon Bottou
LEON@BOTTOU.ORG
Microsoft Research, Redmond, WA.

Jonas Peters[†]
JONAS.PETERS@TUEBINGEN.MPG.DE
Max Planck Institute, Tübingen.

Joaquin Quiñonero-Candela,[a‡] **Denis X. Charles,**[b] **D. Max Chickering,**[b]
Elon Portugaly,[a] **Dipankar Ray,**[c] **Patrice Simard,**[b] **Ed Snelson**[a]
[a] *Microsoft Cambridge, UK.*
[b] *Microsoft Research, Redmond, WA.*
[c] *Microsoft Online Services Division, Bellevue, WA.*

Example Machine Learning applications…

- **Application:** Making medical treatment decisions.
- **Treatment:** Type and timing of medications.
- **Outcome:** Physiologic state.
- **Why is it complex?**
  - Both the timing and type of treatment matters.
  - The data is generated under a specific treatment plan, but inferences should generalize.

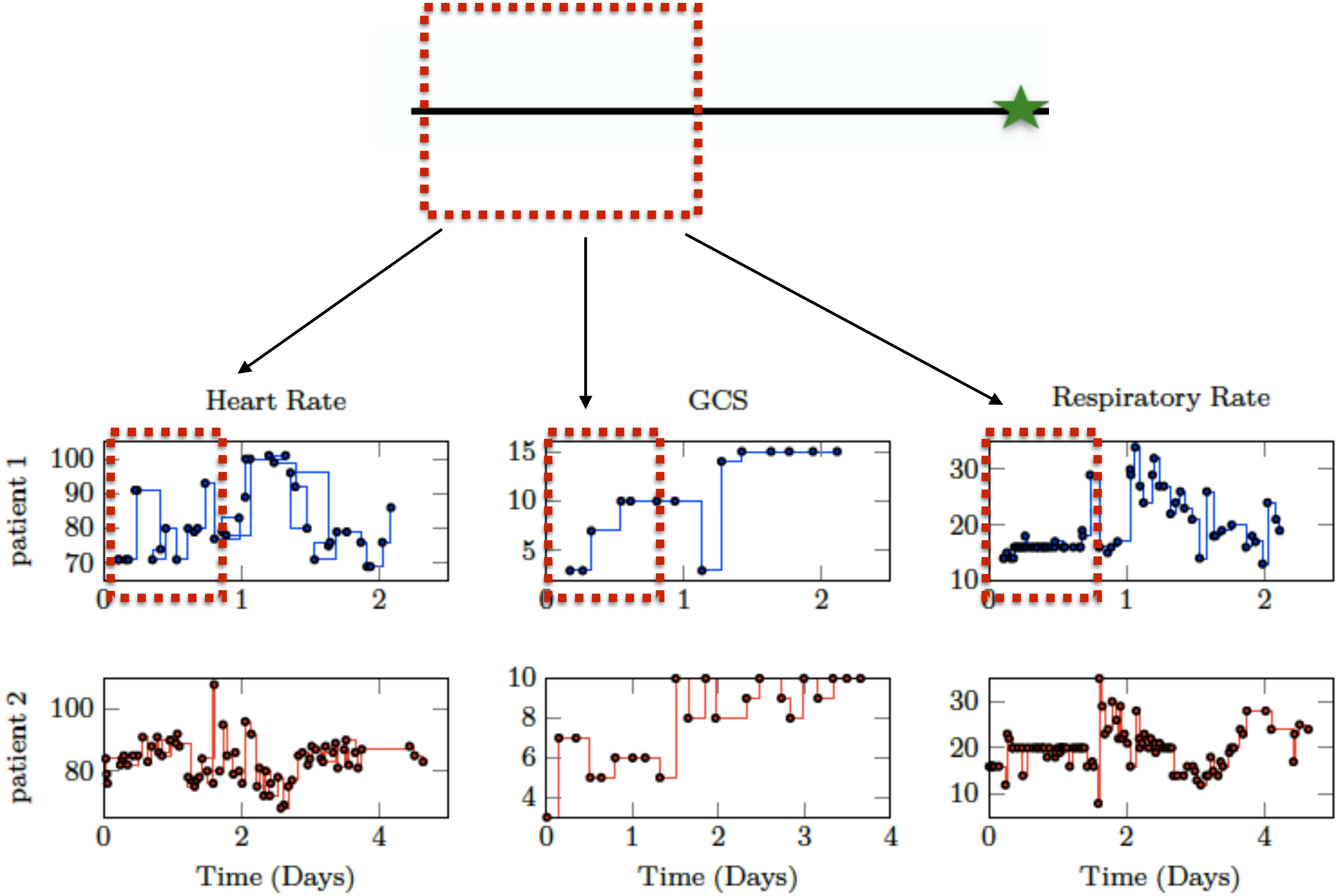


Reliable Decision Support using Counterfactual Models

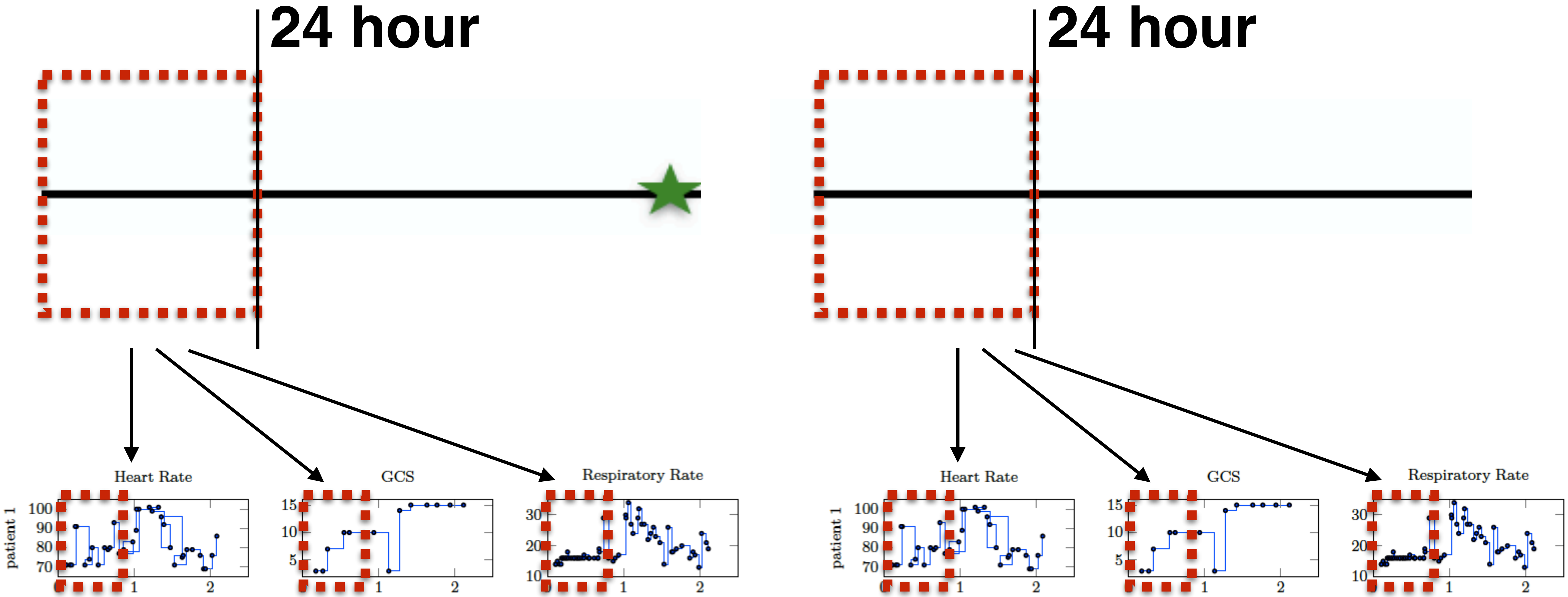
Peter Schulam
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21211
`pschulam@cs.jhu.edu`

Suchi Saria
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21211
`ssaria@cs.jhu.edu`

Deep Dive: Risk Prediction

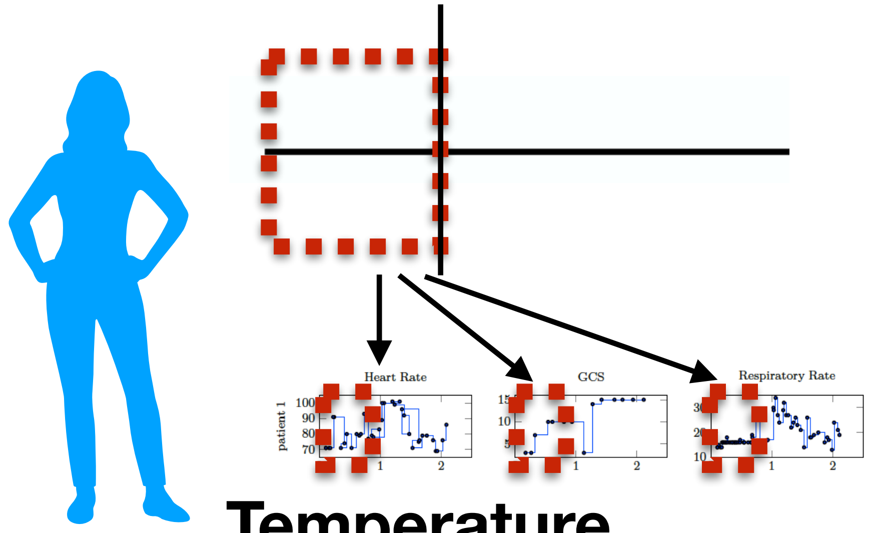


Mortality Risk Prediction as a Supervised Learning Task:



- Unreliable risk estimates leading to patient harm

Is this patient at risk?



Temperature

Heart Rate

Blood Pressure

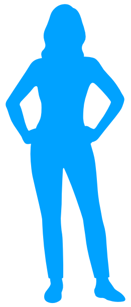
$= \; r?$

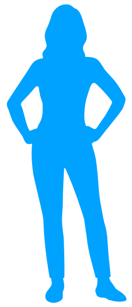
Dataset $\mathcal{D}_A$

Dataset $\mathcal{D}_B$

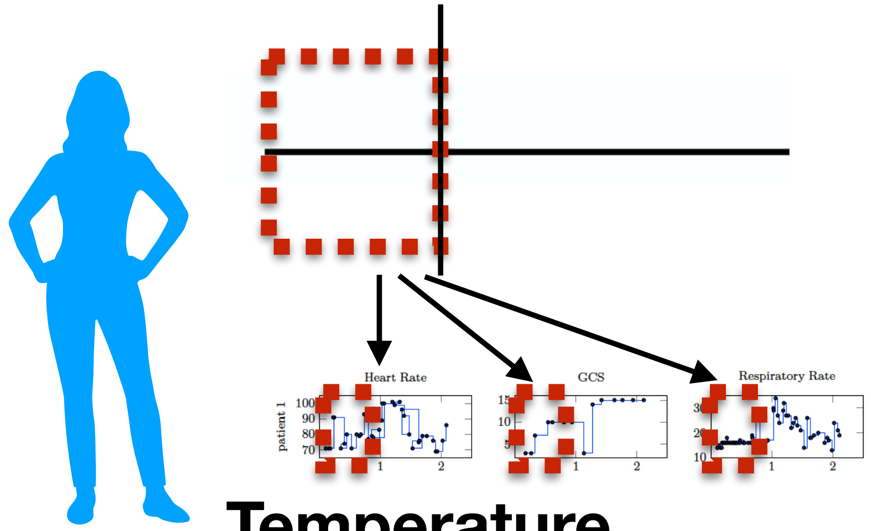
$\downarrow$

$\downarrow$

$\hat{r}_A($  $)$

$\hat{r}_B($  $)$

Is this patient at risk?



Temperature

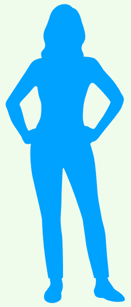
Heart Rate

$= r?$

Models trained on two different datasets, gives different contradicting risk estimates for the same patient.

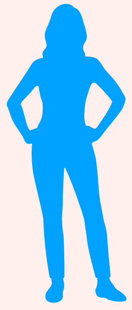
Dataset D_A

$\downarrow$

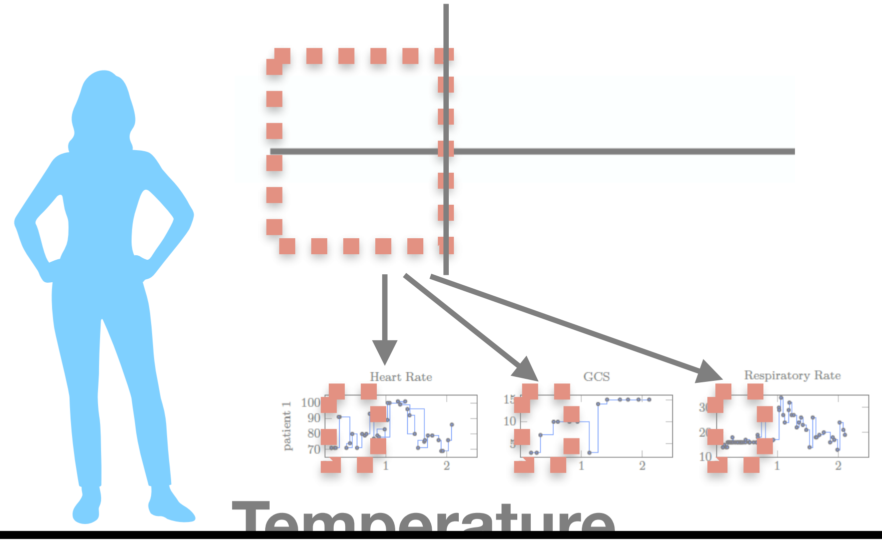
$\hat{r}_A($  $) =$ Low Risk

Dataset D_B

$\downarrow$

$\hat{r}_B($  $) =$ High Risk

Is this patient at risk?



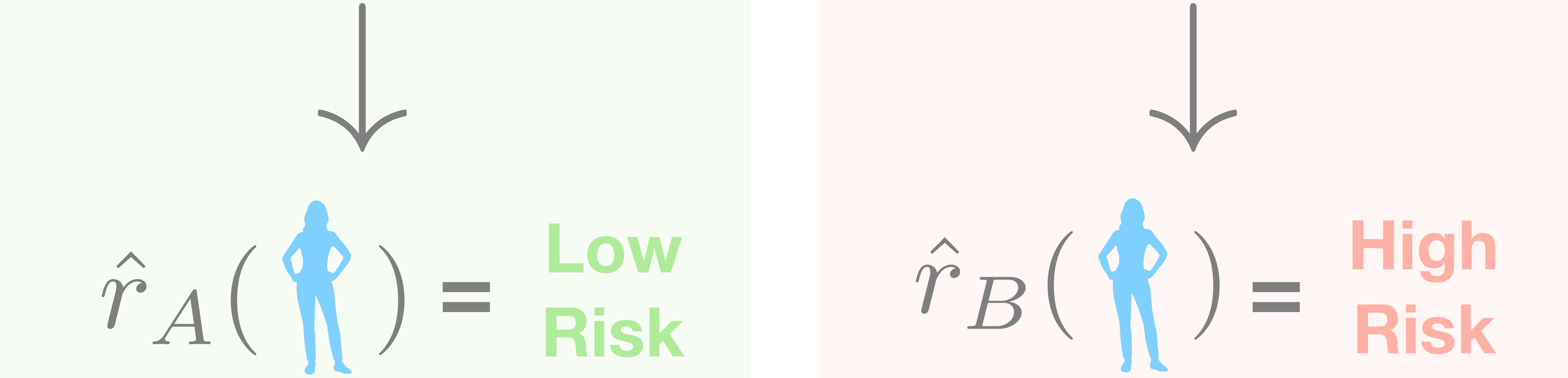
$= r?$

Models trained on two different datasets, gives different contradicting risk estimates for the same patient.

- Are the populations different? No.

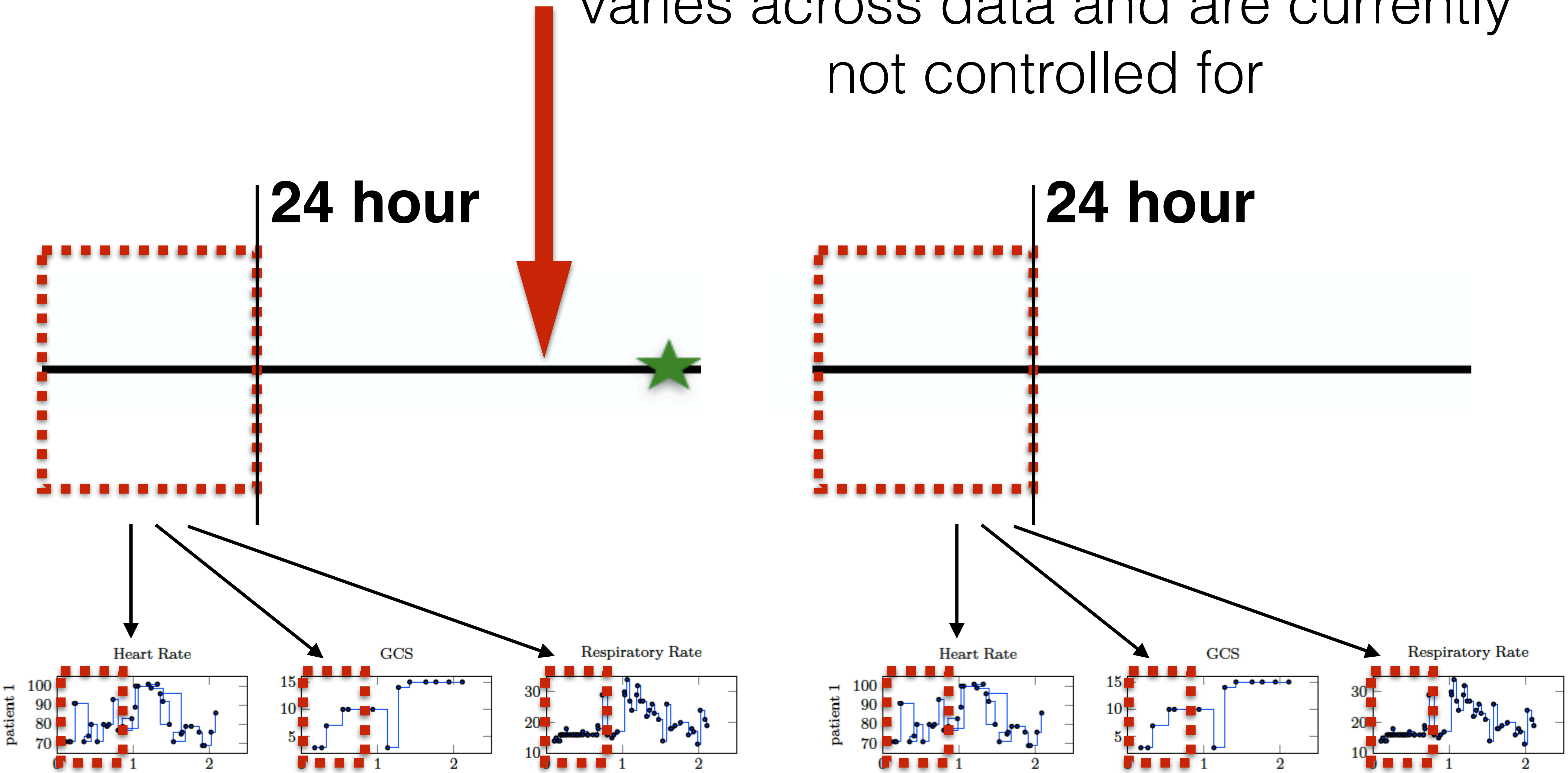
- Maybe overfitting? No.

- Do the features make sense? Yes.

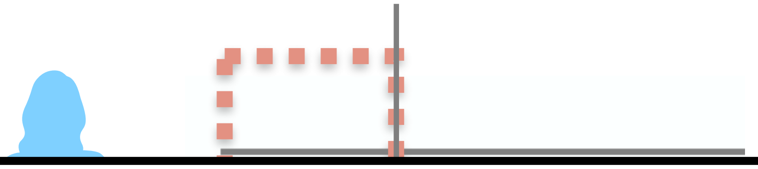


Mortality Risk Prediction as a Supervised Learning Task:

Outcome influenced by factors that varies across data and are currently not controlled for



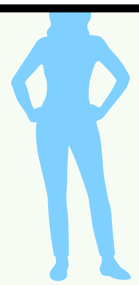
Is this patient at risk?



Naive approach suffers from "Policy Creep": learns policy-dependent relationships between variables that do not generalize when the policy changes.

↓

Unsafe & Unreliable Decisions

$\hat{r}_A($  $) =$ Low Risk

$\hat{r}_B($ $) =$

Schulam and Saria, NIPS 2017

Dyagilev and Saria, Machine Learning 2015

Potential Outcomes: Simulating an experiment

Set of actions

Random variable

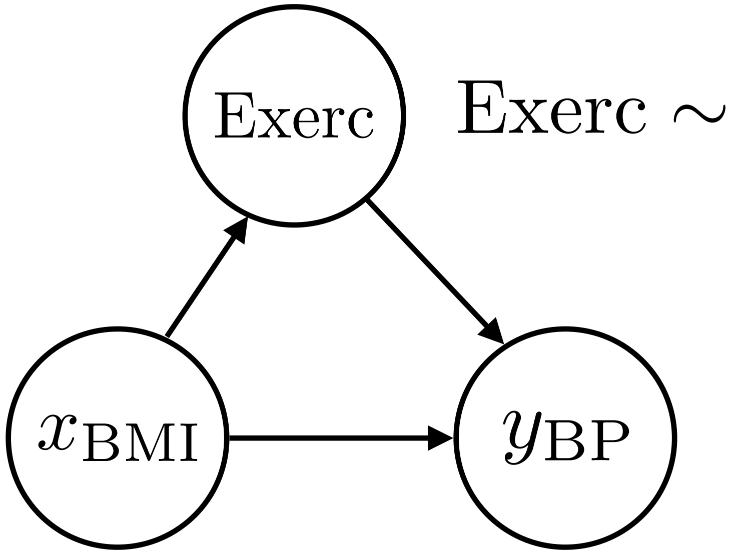
$$\{P(Y(a) \mid X) : a \in \mathcal{A}\}$$

Action

Features

Create a model of the target outcome for each possible action

Learning from data w/ non-random action assignment



$$\text{Exerc} \sim \text{Bern}\left(\frac{1}{1+e^{-2x_{\text{BMI}}}}\right)$$

$$x_{\text{BMI}} \sim \mathcal{N}(0,1)$$

$$y_{\text{BP}} \sim \mathcal{N}(x_{\text{BMI}}, 0.4) - 0.8 \cdot \text{Exerc}$$

- Goal: Learn outcome under exercise / no-exercise.
- Explicitly understand and state your sources of confounding and see if these can be adjusted for

Review: Assumptions

- To learn potential outcome models, recall that we will use three important assumptions:
- (1) Consistency
  - Links observed outcomes to potential outcomes
- (2) Treatment Positivity
  - Ensures that we can learn potential outcome models
- (3) No unmeasured confounders (NUC)
  - Ensures that we do not learn biased models

Rubin, 1974 **Neyman et al., 1923** **Rubin, 2005**

(1) Consistency

- Consider a dataset containing observed outcomes, observed treatments, and covariates:

$$\{y_i, a_i, \mathbf{x}_i\}_{i=1}^n$$

  - E.g.: blood pressure, exercise, BMI

- Consistency allows us to replace the observed response with the potential outcome of the observed treatment

$$Y \triangleq Y(a) \mid A = a$$

- Under consistency our dataset satisfies

$$\{y_i, a_i, \mathbf{x}_i\}_{i=1}^n \triangleq \{y_i(a_i), a_i, \mathbf{x}_i\}_{i=1}^n$$

(2) Positivity

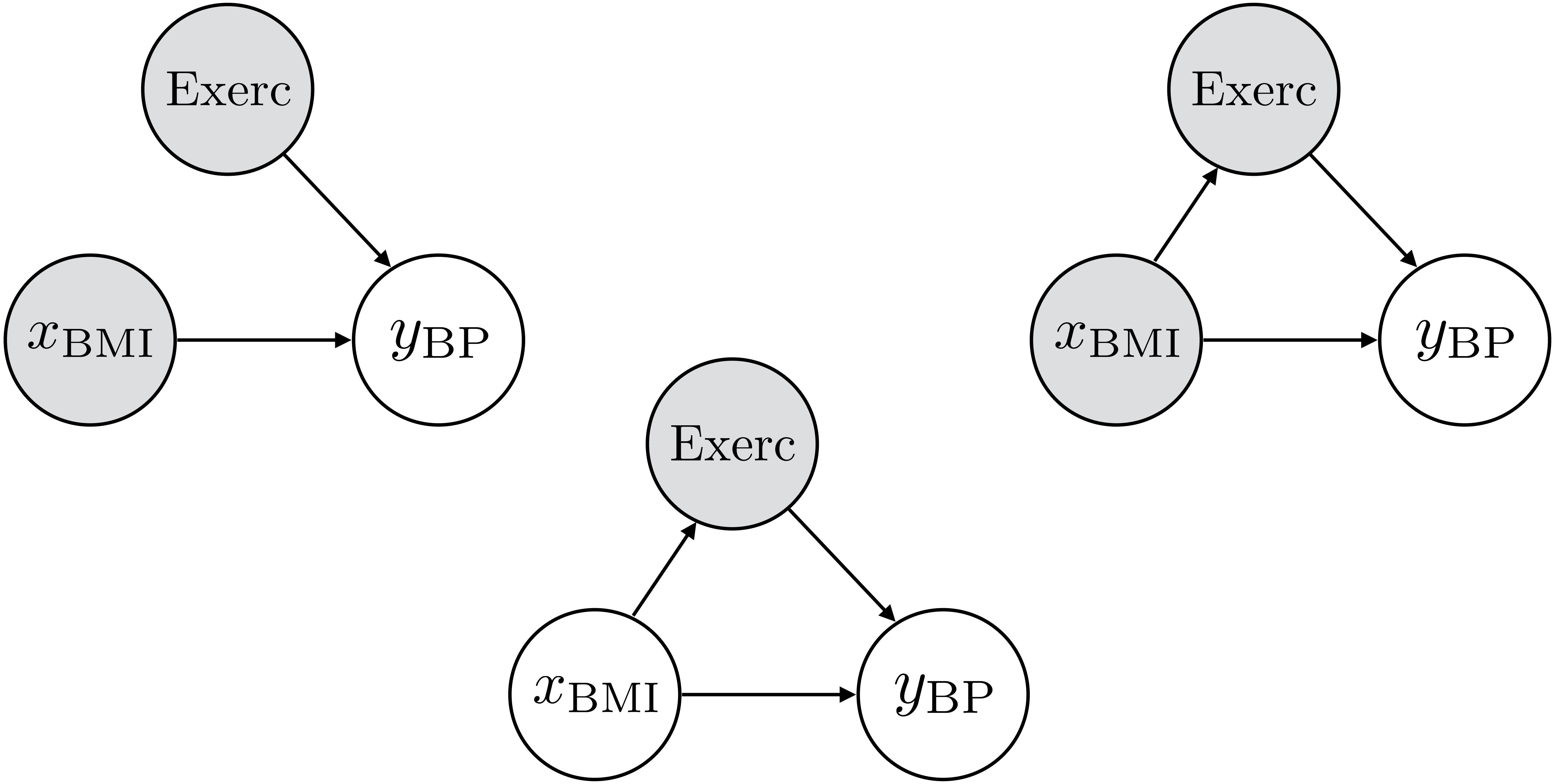
- When working with observational data, for any set of covariates $\mathbf{x}$ we need to **assume a non-zero probability of seeing each treatment**
    - Otherwise, in general, cannot learn a conditional model of the potential outcomes given those covariates
- Formally, we assume that

$$\mathrm{P}_{\mathrm{Obs}}(A = a \mid \mathbf{X} = \mathbf{x}) > 0 \;\; \forall a \in \mathcal{A}, \forall \mathbf{x} \in \mathcal{X}$$

(3) No Unmeasured Confounders (NUC)

- Formally, NUC is an statistical independence assertion:

$$Y(a) \perp A \mid \mathbf{X} = \mathbf{x} \quad : \quad \forall a \in \mathcal{A}, \forall \mathbf{x} \in \mathcal{X}$$



Potential Outcomes: Learning models from data

- To simulate data from a new policy, we need to learn the potential outcome models
  - If we have an observational dataset where assumptions 1-3 hold, then this is possible!
- Assumptions allow estimation of potential outcomes from (observational) data:

$$\begin{aligned} \mathrm{P}(Y(a) \mid \mathbf{X} = \mathbf{x}) &= \mathrm{P}(Y(a) \mid \mathbf{X} = \mathbf{x}, A = a) \quad \text{(A3)} \\ &= \boxed{\mathrm{P}(Y \mid \mathbf{X} = \mathbf{x}, A = a)} \quad \text{(A1)} \end{aligned}$$

Estimation requires a statistical model for estimating conditionals

Using Potential Outcomes Framework to Simulate RCT (e.g., 1-time step)

- Our observational data is drawn from

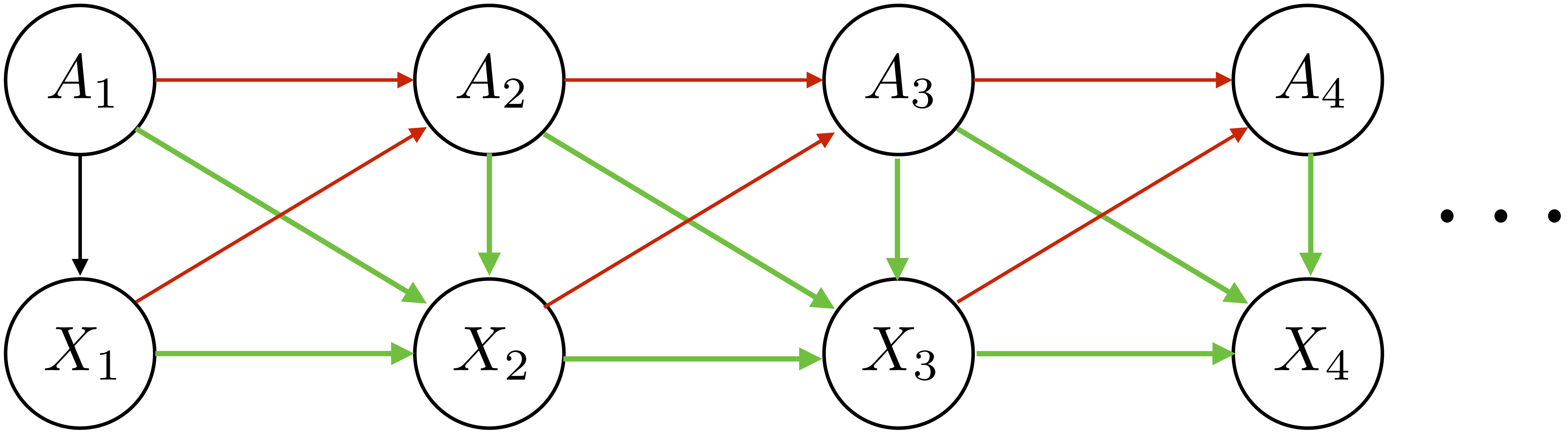
$$Q \triangleq \mathrm{P}(\mathbf{X})\mathrm{P}_{\mathrm{Obs}}(A \mid \mathbf{x})\mathrm{P}(Y \mid a, \mathbf{x}) = \mathrm{P}(\mathbf{X})\mathrm{P}_{\mathrm{Obs}}(A \mid \mathbf{x})\boxed{\mathrm{P}(Y(a) \mid \mathbf{x})}$$

- We want experimental data drawn from

$$P \triangleq \mathrm{P}(\mathbf{X})\mathrm{P}_{\mathrm{Exp}}(A)\mathrm{P}(Y \mid a, \mathbf{x}) = \mathrm{P}(\mathbf{X})\mathrm{P}_{\mathrm{Exp}}(A)\mathrm{P}(Y(a) \mid \mathbf{x})$$

- If we know potential outcome models:
  - Draw from empirical covariate distribution: $\mathbf{X} \sim \{\mathbf{x}_i\}_{i=1}^{n}$
  - Flip fair coin to assign treatment: $A \sim \mathrm{Bern}(0.5)$
  - Simulate outcome from model: $\mathrm{P}(Y(a) \mid \mathbf{X} = \mathbf{x})$

Potential Outcomes in Sequential Setting

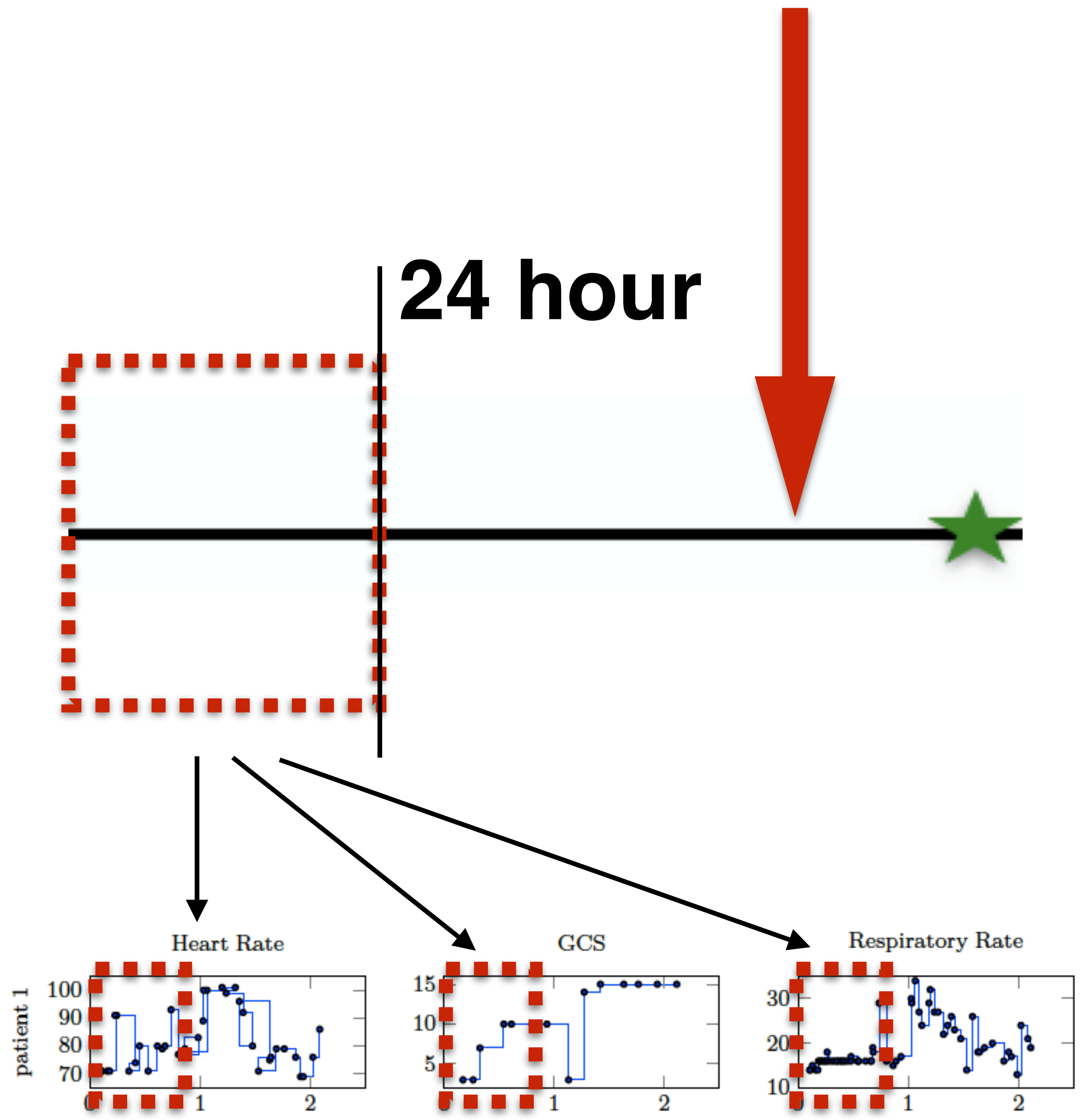


green = generalizes across datasets
red = changes across data

Can you extend idea on previous slide to this sequential setting?

Robins 1986 **Robins and Hernan 1990**

Returning to our example: Mortality Risk Prediction



$$P(\{Y_s : s > t\} \mid \mathcal{H}_t)$$

vs.

$$P(\{Y_s(\varnothing) : s > t\} \mid \mathcal{H}_t)$$

Schulam and Saria, NIPS 2017

Related Work

Neyman 1923 **Rubin 2005**
Robins 1986 **Robins and Hernan 1990**

Potential outcomes framework

Brodersen et al., 2015
Bottou et al., 2013

ads; single intervention

Taubman et al.,2009

epidemiology; multiple sequential interventions

Lok et al., 2008
Xu, Xu, Saria, 2016

sparse, irregularly sampled longitudinal data; functional outcomes

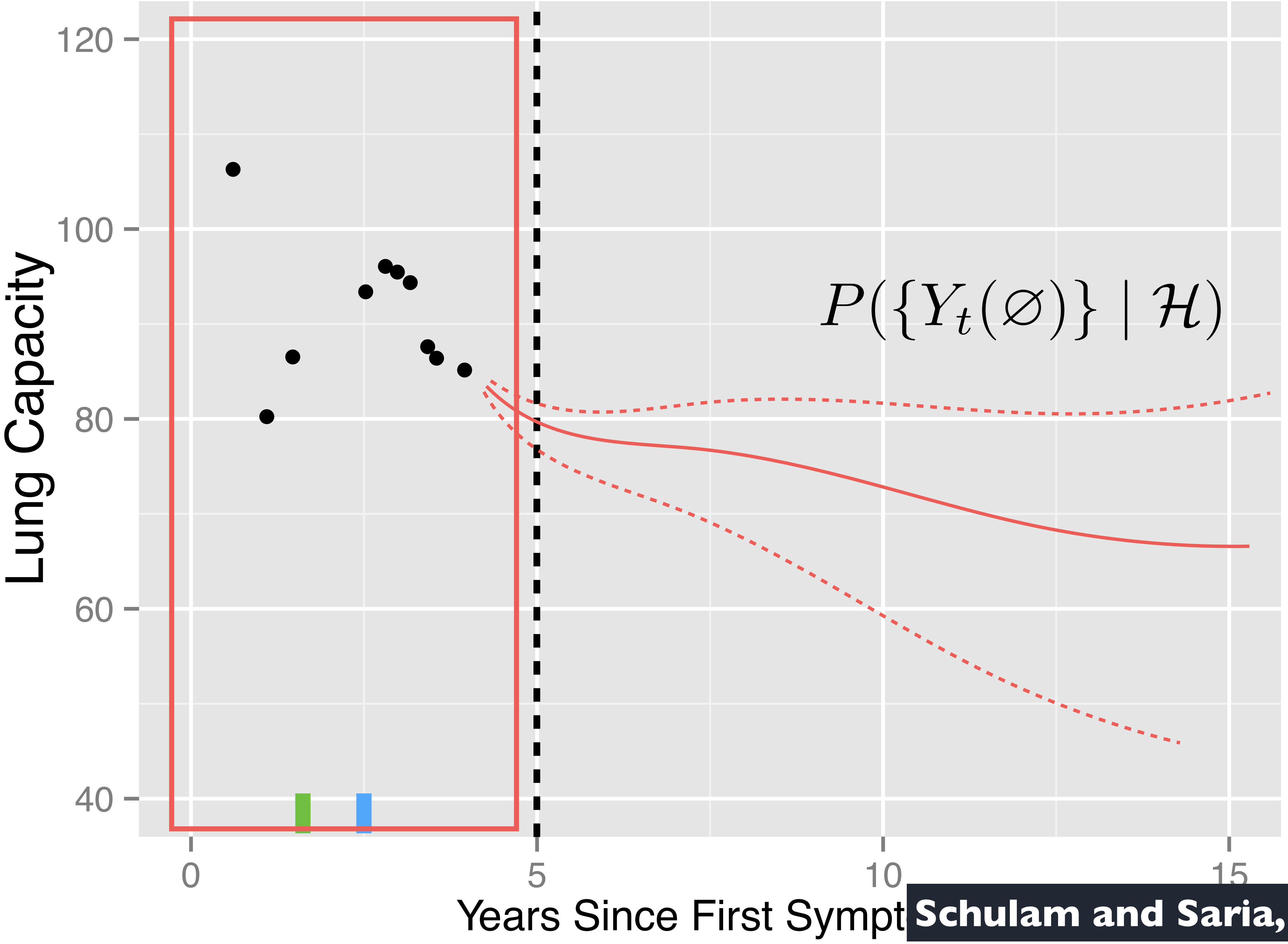
---

- Off-policy evaluation: Re-weighting to evaluate reward for a policy when learning from offline data.

  e.g. **Dudik et al., 2011** **Jiang and Li, 2016** **Paduraru et al. 2013**

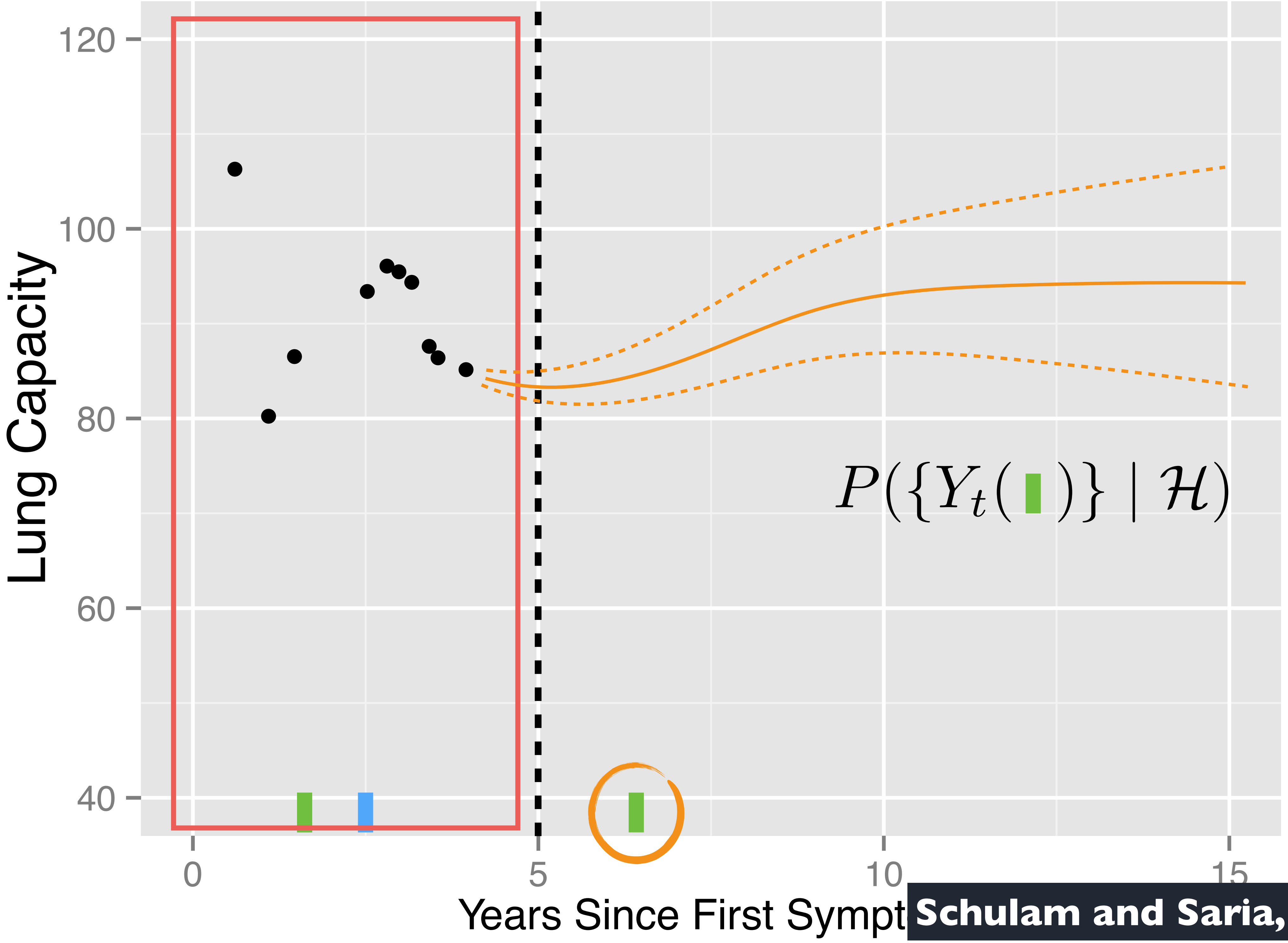
- For detailed discussion of related work, see **Schulam Saria, 2017**

Desiderata: Forecast “what-if” trajectories given history for different candidate interventions



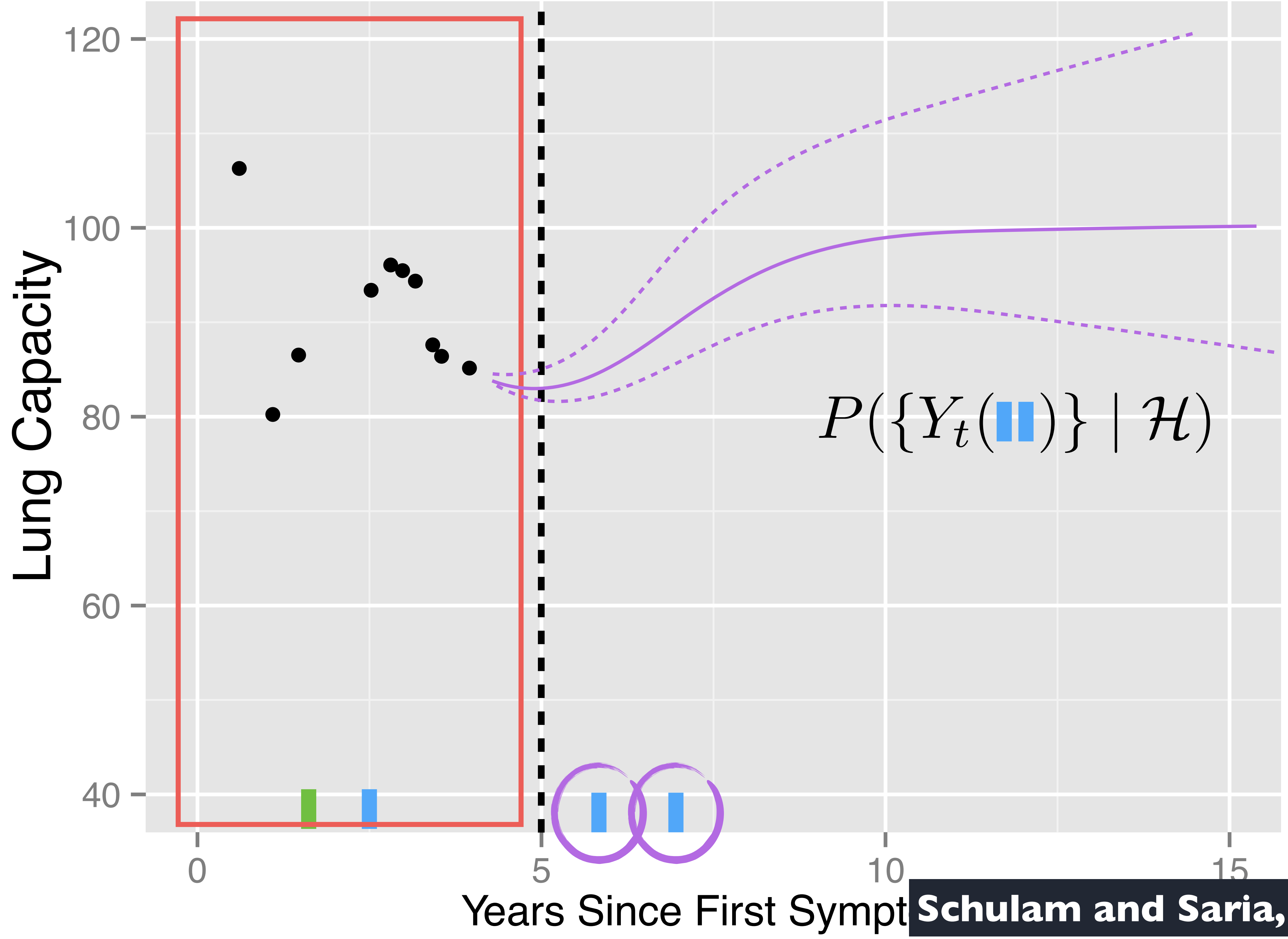
Schulam and Saria, NIPS 2017

Counterfactual GP



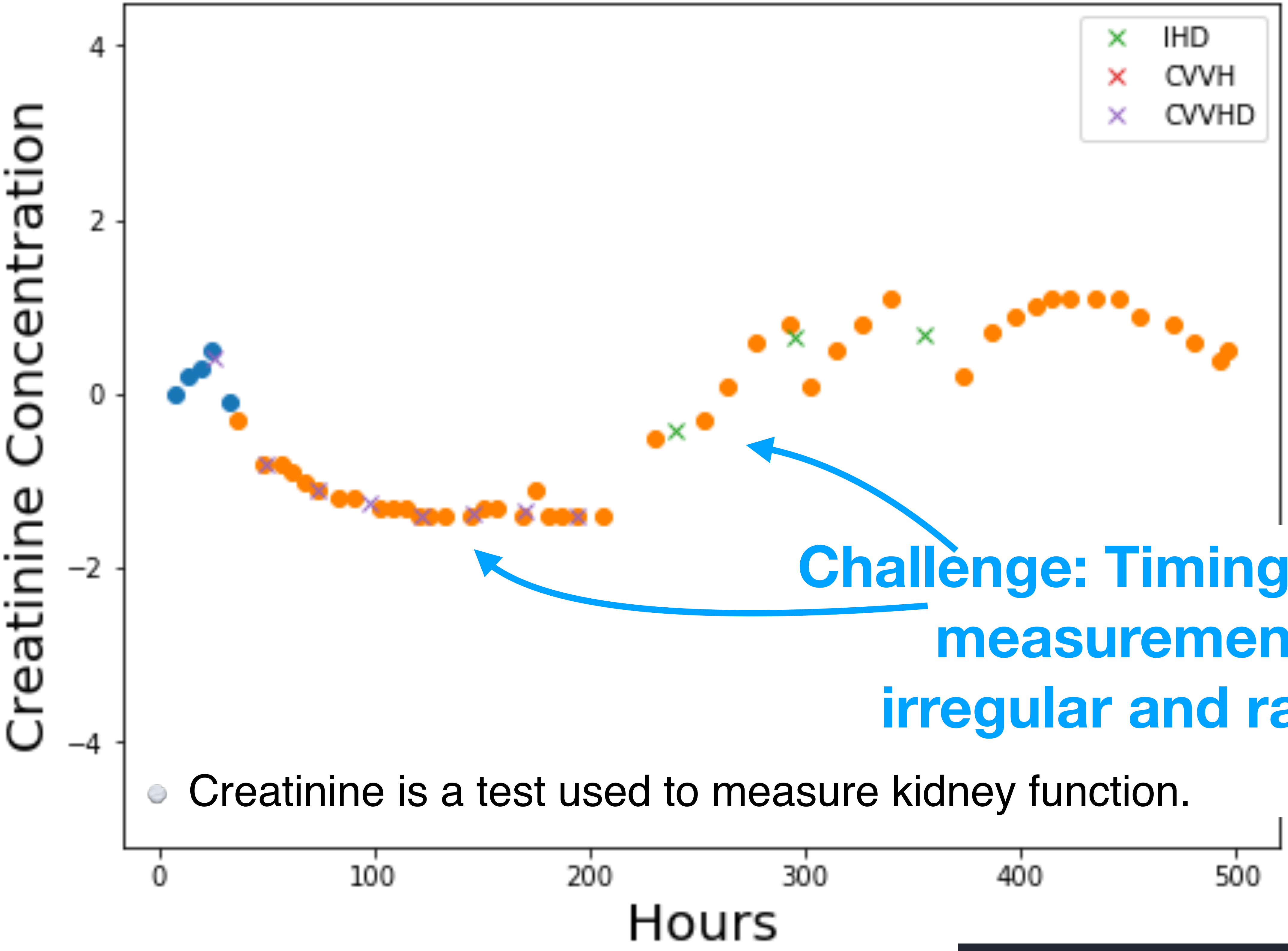
Schulam and Saria, NIPS 2017

Counterfactual GP



Schulam and Saria, NIPS 2017

Develop potential outcome model for observational Longitudinal Traces

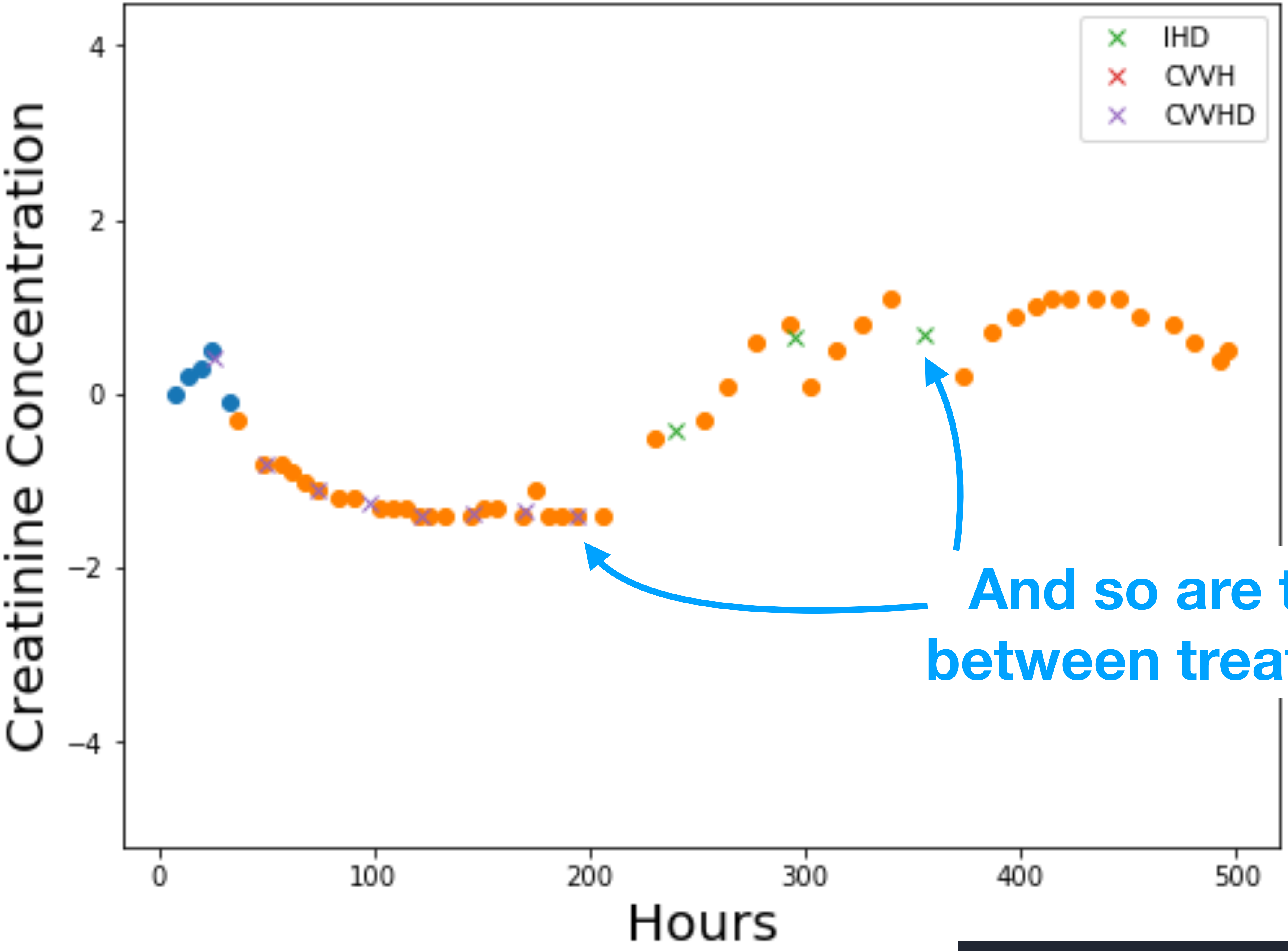


Challenge: Timing between measurements is irregular and random

- Creatinine is a test used to measure kidney function.

Schulam and Saria, NIPS 2017

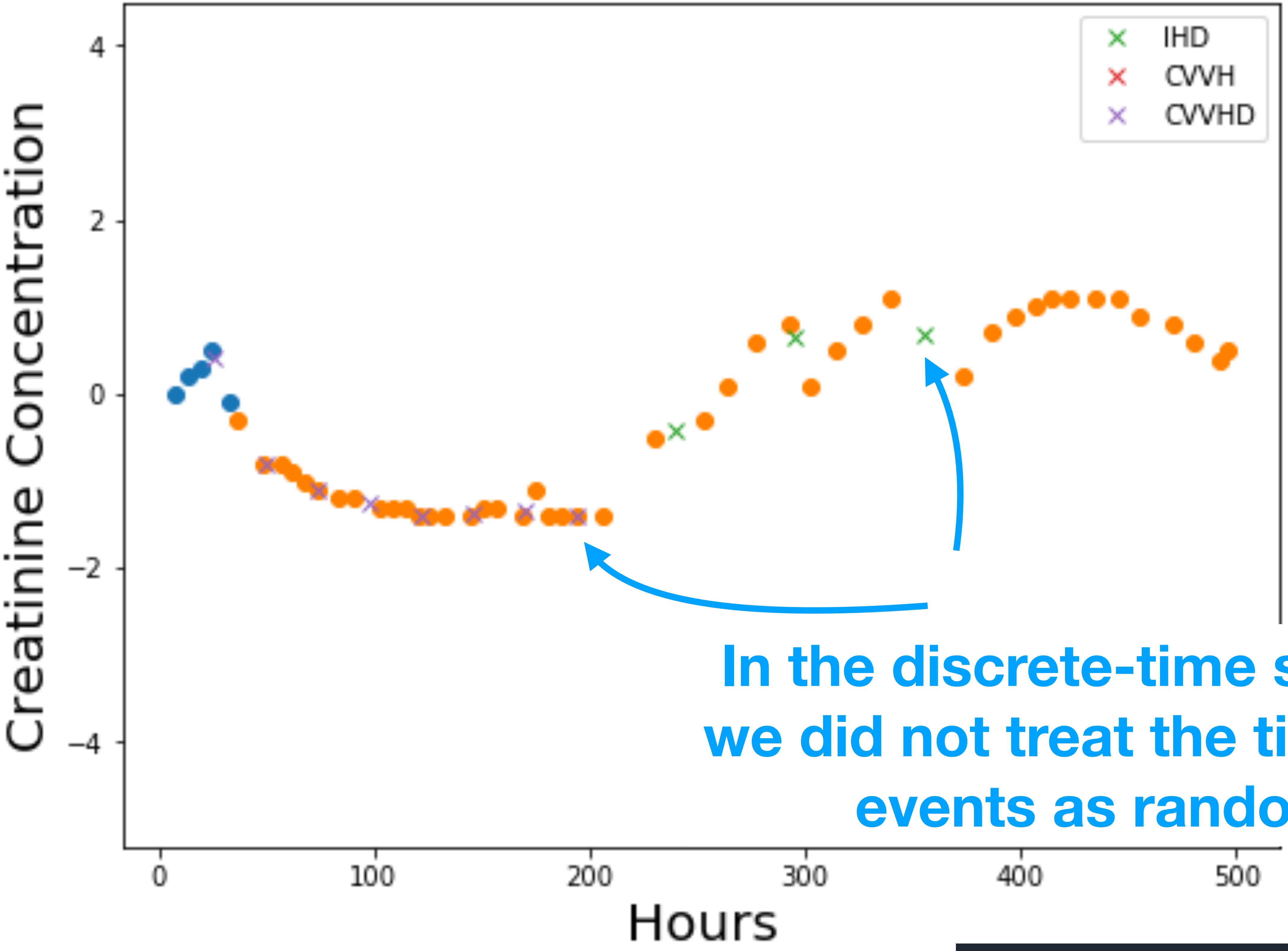
Observational Longitudinal Traces



And so are times between treatments

Schulam and Saria, NIPS 2017

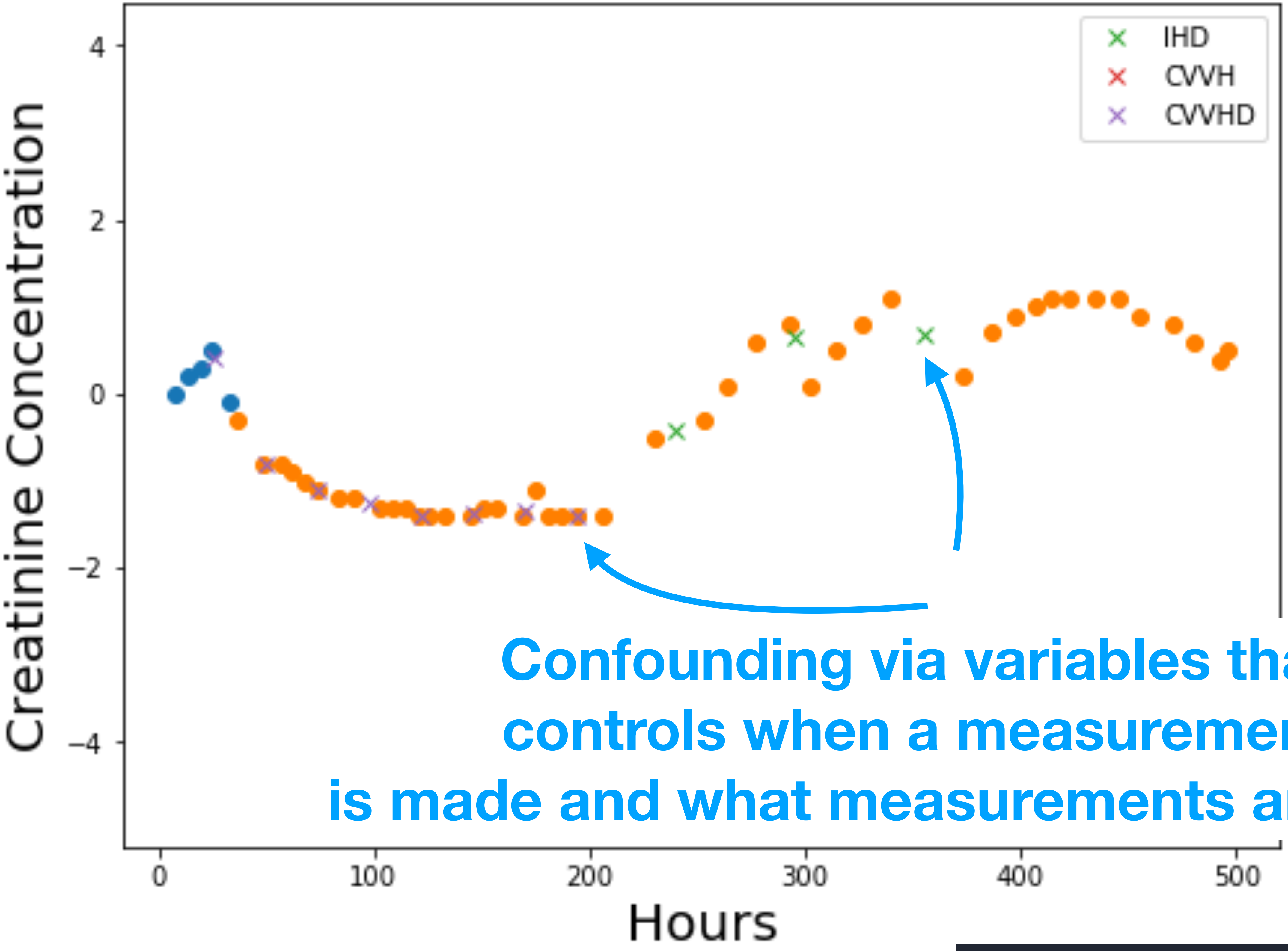
Observational Longitudinal Traces



In the discrete-time setting, we did not treat the timing of events as random

Schulam and Saria, NIPS 2017

Observational Longitudinal Traces



Confounding via variables that controls when a measurement is made and what measurements are made

Schulam and Saria, NIPS 2017

Learning Models from Observational Traces

$$\mathcal{D} \triangleq \left\{ \mathbf{h}_i = \{(t_{ij}, y_{ij}, a_{ij})\}_{j=1}^{n_i} \right\}_{i=1}^{m}$$

- (1) Posit probabilistic model of observational traces

  Posit a model for *when* a measurement is made or actions are taken and *what* the value of the measurements and actions are.

- (2) Derive maximum likelihood estimator

- (3) Establish assumptions that connect probabilistic of observational traces to *target counterfactual model*

$$P(\{Y_s[\mathbf{a}] : s > t\} \mid \mathcal{H}_t)$$

Schulam and Saria, NIPS 2017

Modeling Observational Traces

- We use a marked point process (MPP):

$$\{(T_i, X_i)\}_{i=1}^{\infty}$$

- Points model the *event times*: measurements or actions
- Mark models the type of event

$$\mathcal{X} = (\mathbb{R} \cup \{\varnothing\}) \times (\mathcal{C} \cup \{\varnothing\}) \times \{0, 1\} \times \{0, 1\}$$

Modeling Observational Traces

- We use a marked point process (MPP):

$$\{(T_i, X_i)\}_{i=1}^{\infty}$$

- Points model the *event times*: measurements or actions

- Mark models the type of event

$$\mathcal{X} = (\mathbb{R} \cup \{\varnothing\}) \times (\mathcal{C} \cup \{\varnothing\}) \times \underset{\nearrow\, z_y}{\{0, 1\}} \times \{0, 1\}$$

Did we measure an outcome?

Modeling Observational Traces

- We use a marked point process (MPP):

$$\{(T_i, X_i)\}_{i=1}^{\infty}$$

- Points model the *event times*: measurements or actions
- Mark models the type of event

$$\mathcal{X} = (\mathbb{R} \cup \{\varnothing\}) \times (\mathcal{C} \cup \{\varnothing\}) \times \underset{z_y}{\{0, 1\}} \times \underset{z_a}{\{0, 1\}}$$

Did we take an action?

Modeling Observational Traces

- We use a marked point process (MPP):

$$\{(T_i, X_i)\}_{i=1}^{\infty}$$

- Points model the *event times*: measurements or actions

- Mark models the type of event

$$\mathcal{X} = \underset{\nearrow\, y}{(\mathbb{R} \cup \{\varnothing\})} \times (\mathcal{C} \cup \{\varnothing\}) \times \underset{z_y}{\{0, 1\}} \times \underset{z_a}{\{0, 1\}}$$

What is the value of the outcome?

Modeling Observational Traces

- We use a marked point process (MPP):

$$\{(T_i, X_i)\}_{i=1}^{\infty}$$

- Points model the *event times*: measurements or actions

- Mark models the type of event

$$\mathcal{X} = \underset{y}{(\mathbb{R} \cup \{\varnothing\})} \times \underset{a}{(\mathcal{C} \cup \{\varnothing\})} \times \underset{z_y}{\{0, 1\}} \times \underset{z_a}{\{0, 1\}}$$

What action did we take?

Modeling Observational Traces

- Parameterize MPP using hazard and mark density:

$$\lambda^*(t, x) = \lambda^*(t)p^*(x \mid t)$$

Modeling Observational Traces

- Parameterize MPP using hazard and mark density:

$$\lambda^*(t, x) = \lambda^*(t)p^*(x \mid t)$$

Probability of event
happening at this time

Probability of mark
given event time

Modeling Observational Traces

- Parameterize MPP using hazard and mark density:

$$\lambda^*(t, x) = \lambda^*(t)p^*(x \mid t)$$

Star denotes dependence on history

Probability of event happening at this time

Probability of mark given event time

Recovering the CGP

- When does the MPP GP recover the CGP?
- In addition to Consistency, we define two assumptions

Schulam and Saria, NIPS 2017

Recovering the CGP

- When does the MPP GP recover the CGP?
- In addition to Consistency, we define two assumptions
- Continuous-time No Unmeasured Confounding (NUC)
    - Analogue of NUC for MPP
- Conditionally Non-informative measurement times
    - Measurement and action times are conditionally independent of potential outcomes

Schulam and Saria, NIPS 2017

Experiments

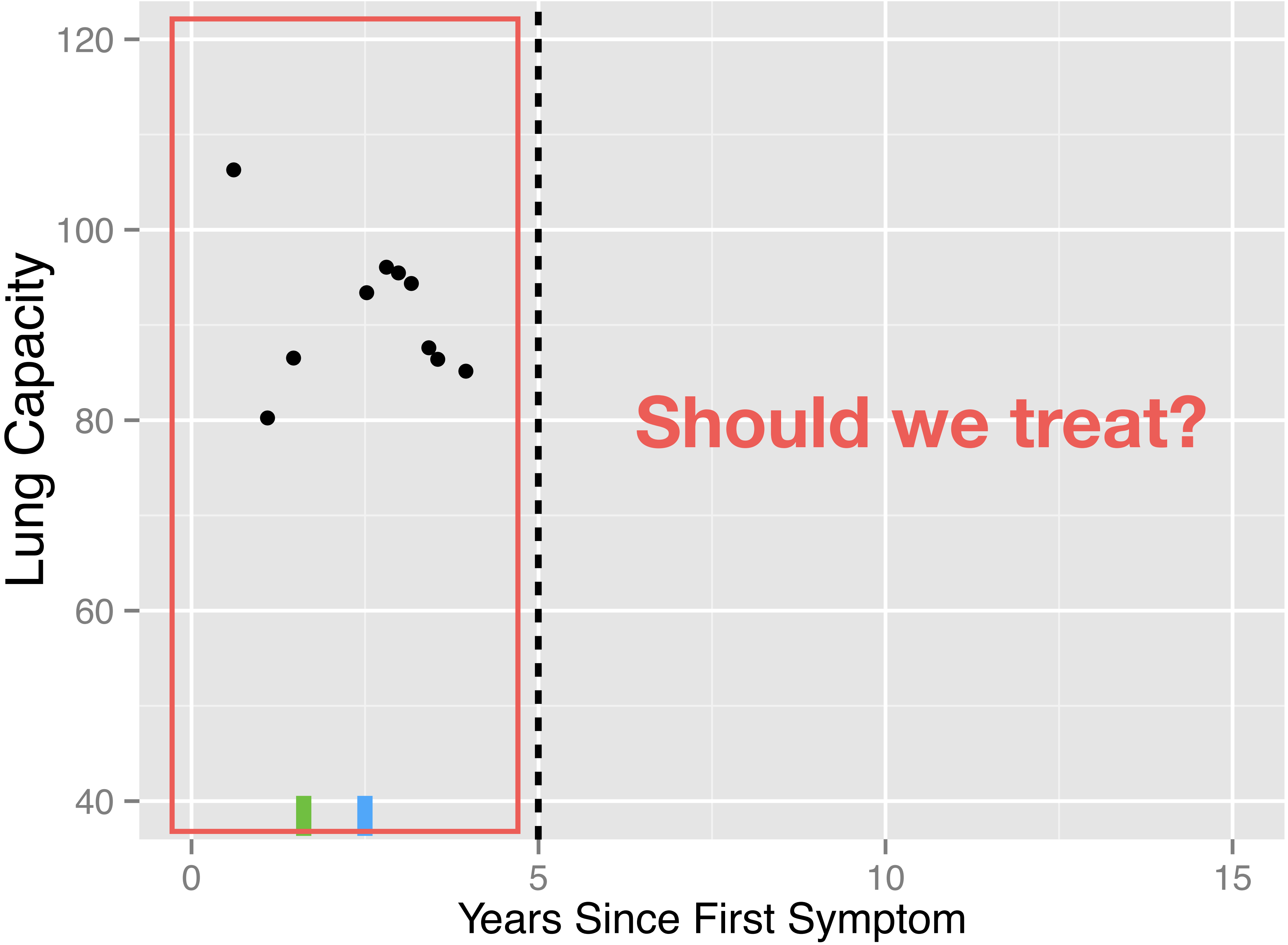
Simulated Data

- Simulate observational traces from three regimes
- Traces are treated by policies unknown to learners
- In regimes A and B, policies satisfy our assumptions
- In regime C, policy violates our assumptions
- Simulate three training sets (regimes A, B, and C)
- Simulate one common test set (regime A)

Results

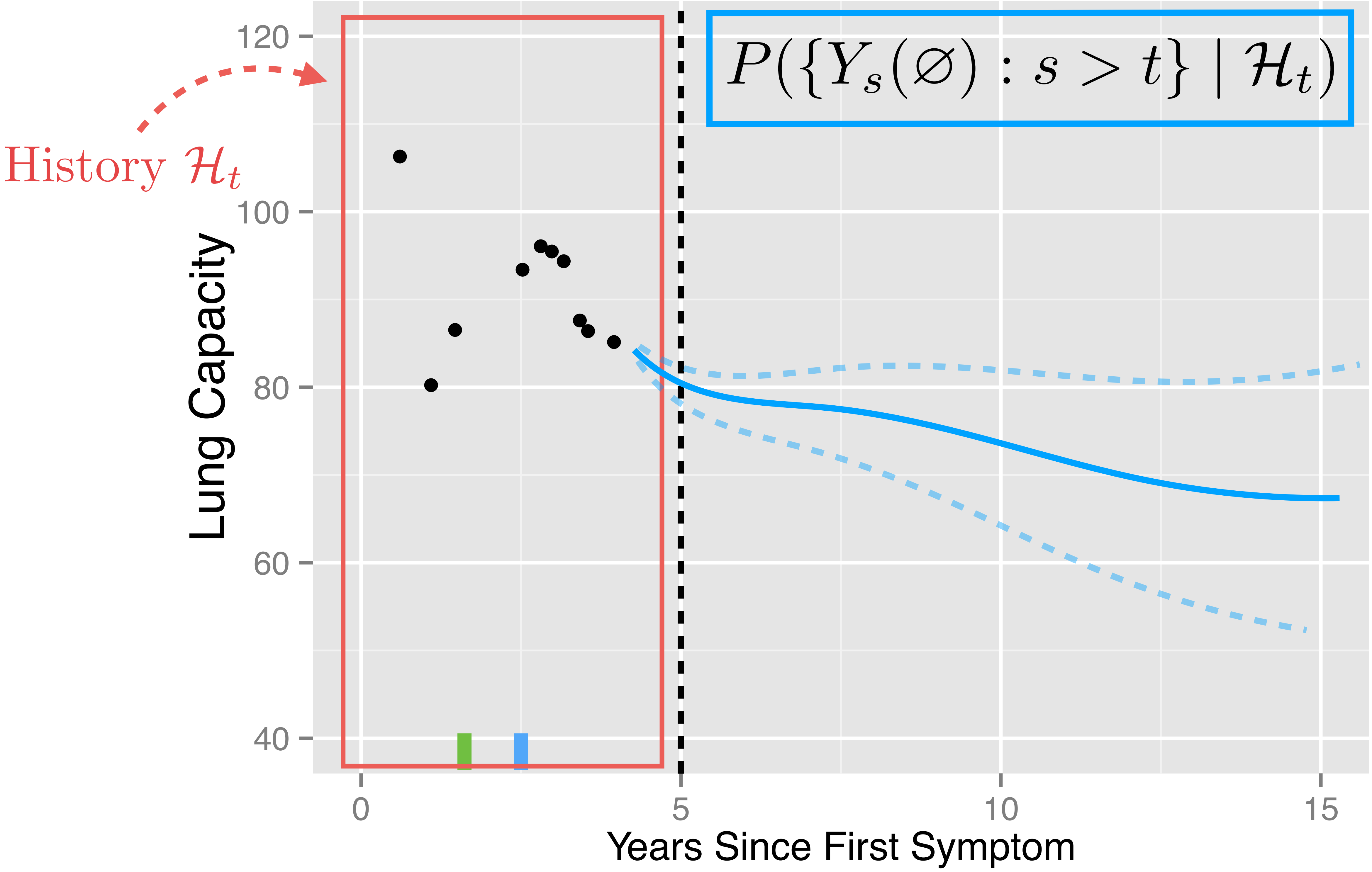
- Risk scores:
  - Use Baseline and CGP to predict final severity marker
  - Negate predictions and normalize to [0, 1]

Reliable Decisions with CGPs

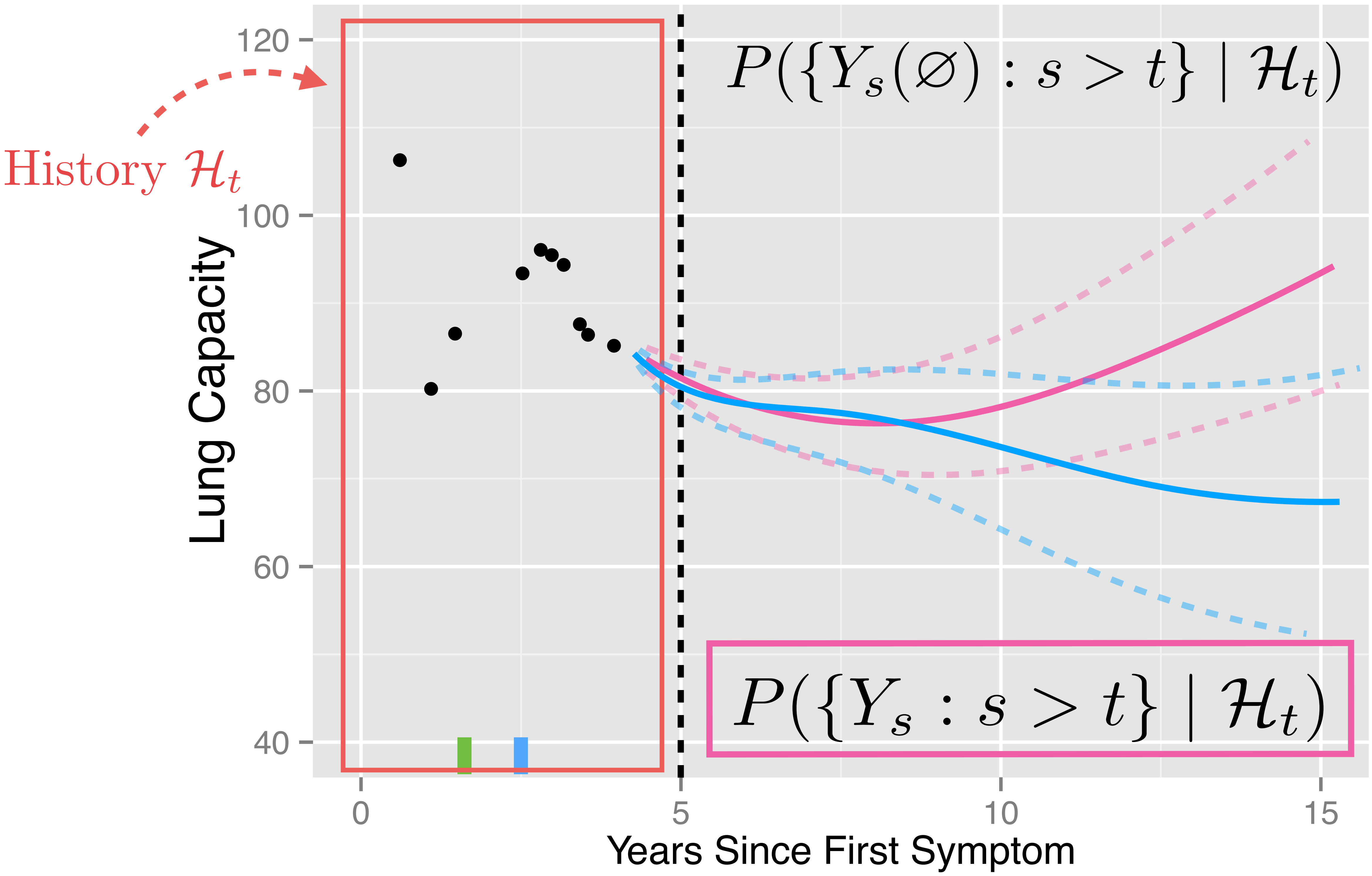


Should we treat?

Counterfactual GP

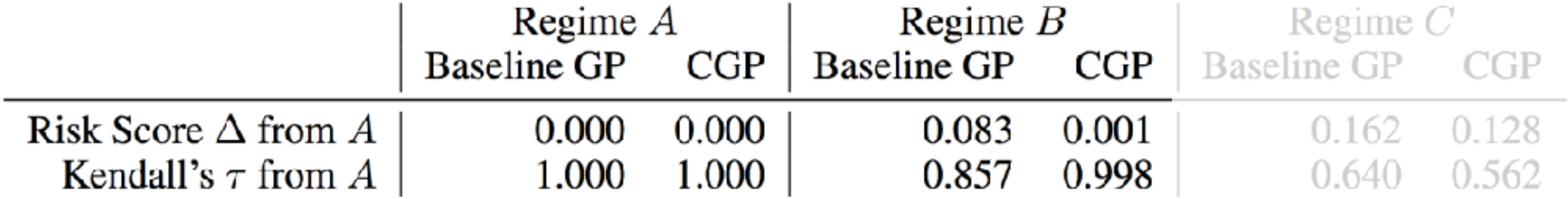


Classical Supervised Model



Risk Score Stability

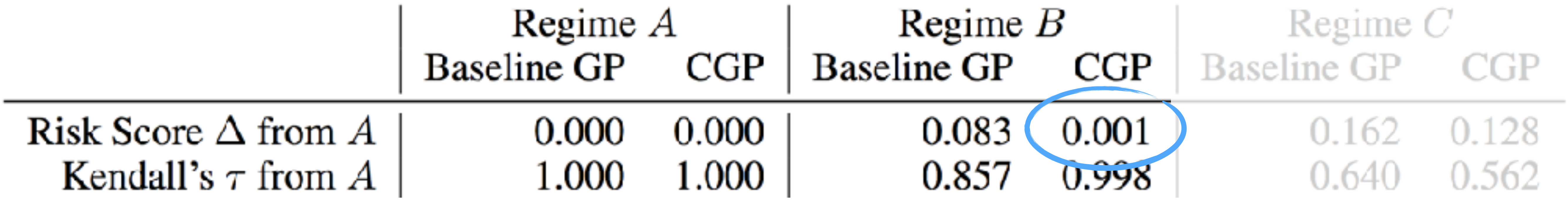
- Use Baseline and CGP to predict final severity marker
- Transform to risk score in [0, 1], where higher is riskier



	Regime A		Regime B		Regime C	
	Baseline GP	CGP	Baseline GP	CGP	Baseline GP	CGP
Risk Score Δ from A	0.000	0.000	0.083	0.001	0.162	0.128
Kendall's τ from A	1.000	1.000	0.857	0.998	0.640	0.562

Risk Score Stability

- Use Baseline and CGP to predict final severity marker
- Transform to risk score in [0, 1], where higher is riskier

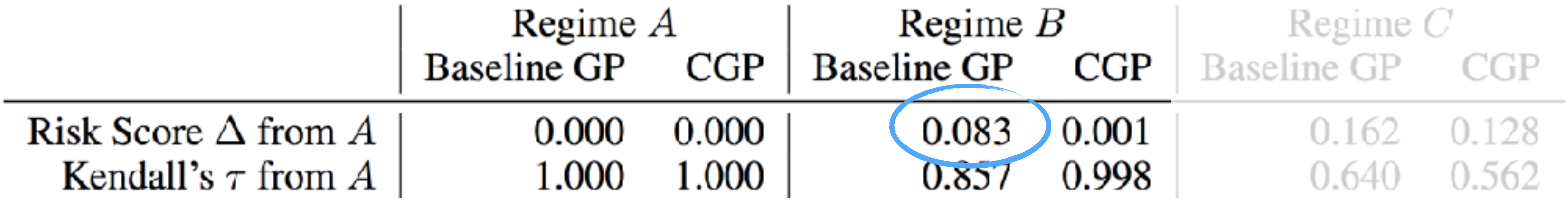


	Regime A		Regime B		Regime C	
	Baseline GP	CGP	Baseline GP	CGP	Baseline GP	CGP
Risk Score Δ from A	0.000	0.000	0.083	0.001	0.162	0.128
Kendall's τ from A	1.000	1.000	0.857	0.998	0.640	0.562

Counterfactual GP scores are stable

Risk Score Stability

- Use Baseline and CGP to predict final severity marker
- Transform to risk score in [0, 1], where higher is riskier

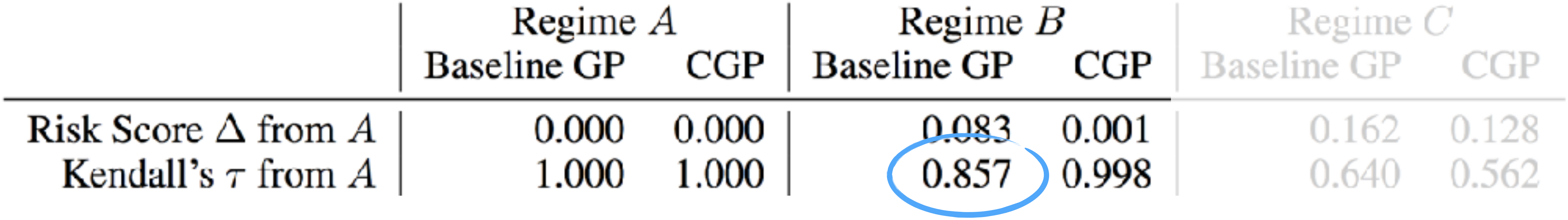


	Regime A		Regime B		Regime C	
	Baseline GP	CGP	Baseline GP	CGP	Baseline GP	CGP
Risk Score Δ from A	0.000	0.000	0.083	0.001	0.162	0.128
Kendall's τ from A	1.000	1.000	0.857	0.998	0.640	0.562

Baseline GP scores change

Risk Score Stability

- Use Baseline and CGP to predict final severity marker
- Transform to risk score in [0, 1], where higher is riskier

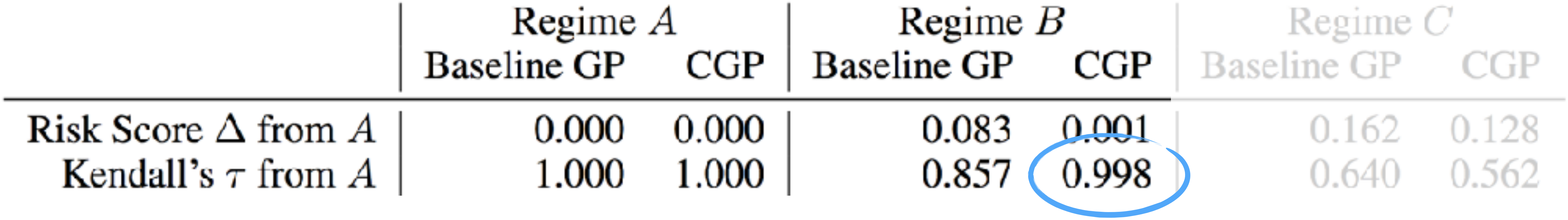


	Regime A		Regime B		Regime C	
	Baseline GP	CGP	Baseline GP	CGP	Baseline GP	CGP
Risk Score Δ from A	0.000	0.000	0.083	0.001	0.162	0.128
Kendall's τ from A	1.000	1.000	0.857	0.998	0.640	0.562

Rank correlation shows considerable change in relative risk

Risk Score Stability

- Use Baseline and CGP to predict final severity marker
- Transform to risk score in [0, 1], where higher is riskier

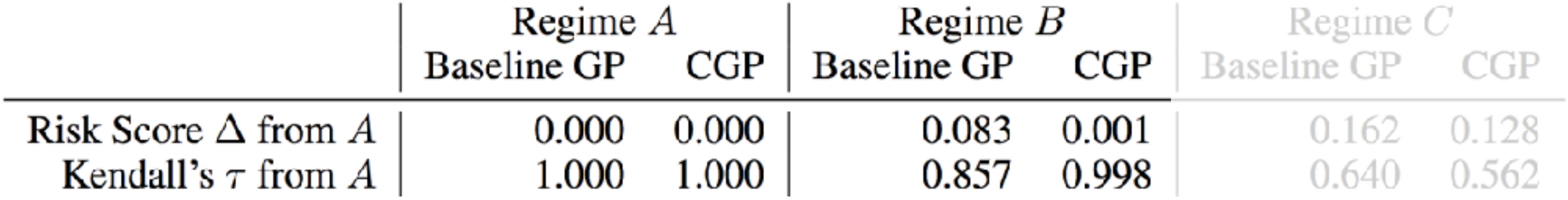


	Regime A		Regime B		Regime C	
	Baseline GP	CGP	Baseline GP	CGP	Baseline GP	CGP
Risk Score Δ from A	0.000	0.000	0.083	0.001	0.162	0.128
Kendall's τ from A	1.000	1.000	0.857	0.998	0.640	0.562

CGP ranking is stable

Risk Score Stability

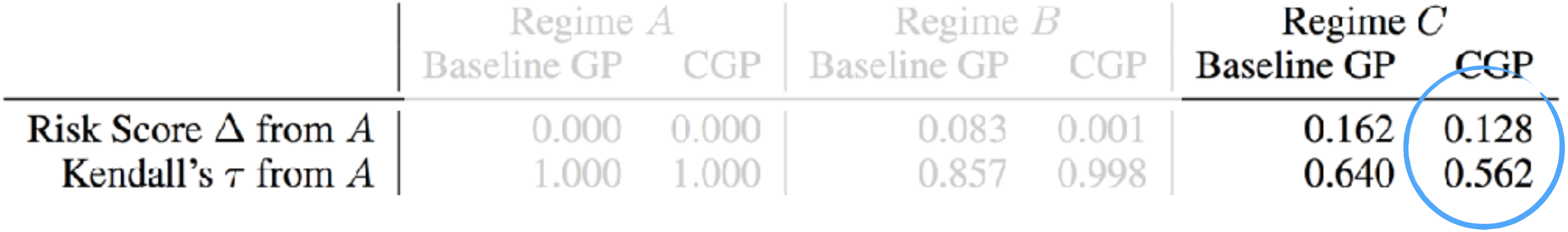
- **Key takeaways**
- **Baseline GP risk depends on why treatments were given in the training data**
- **CGP is stable to this irrelevant information**



	Regime A		Regime B		Regime C	
	Baseline GP	CGP	Baseline GP	CGP	Baseline GP	CGP
Risk Score Δ from A	0.000	0.000	0.083	0.001	0.162	0.128
Kendall's τ from A	1.000	1.000	0.857	0.998	0.640	0.562

Risk Score Stability

- Key takeaways
- With baseline GP, risk of new patient depends on why treatments were given in the training data
- CGP is stable to this irrelevant information



	Regime A		Regime B		Regime C	
	Baseline GP	CGP	Baseline GP	CGP	Baseline GP	CGP
Risk Score Δ from A	0.000	0.000	0.083	0.001	0.162	0.128
Kendall's τ from A	1.000	1.000	0.857	0.998	0.640	0.562

CGP is no longer stable if assumptions are violated

Takeaways

- Classical supervised learning algorithms yield models that are not stable to shifts in policy changes —> as action selection mechanism (policy) changes between train and deployment environments, models fail to generalize.

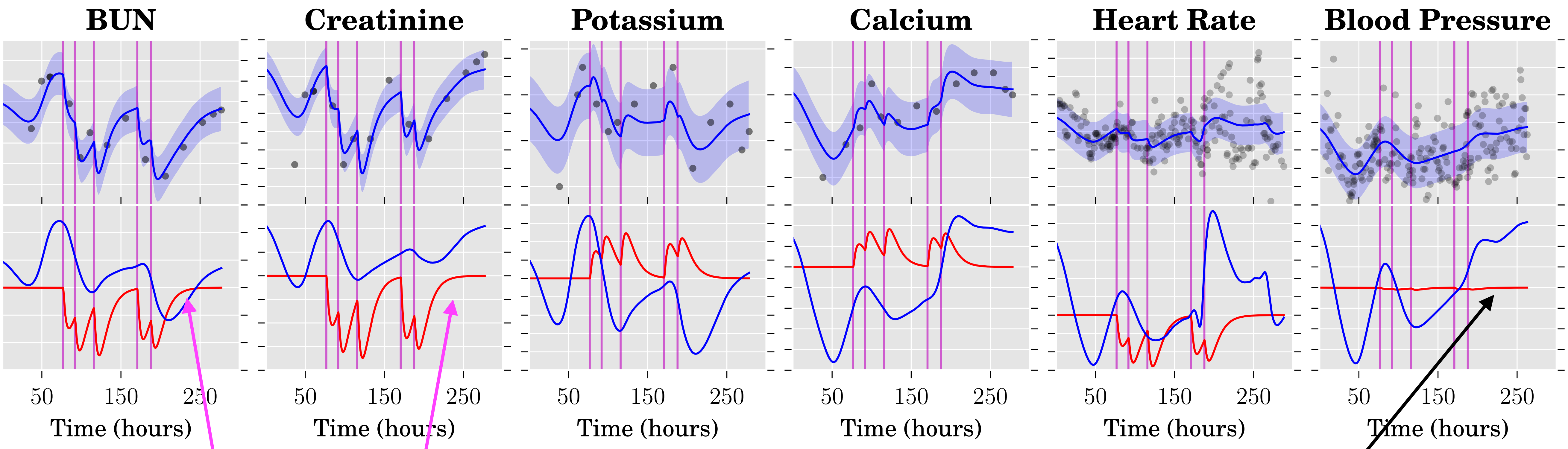
- Propose learning using a different learning objective that predicts potential outcomes.

- Develop a potential outcome model for forecasting trajectories from longitudinal traces. (See next slide for example with multiple longitudinal streams.)

- Under certain assumptions, the Counterfactual Gaussian Process (CGP) makes predictions that are invariant to policy changes in the training data.

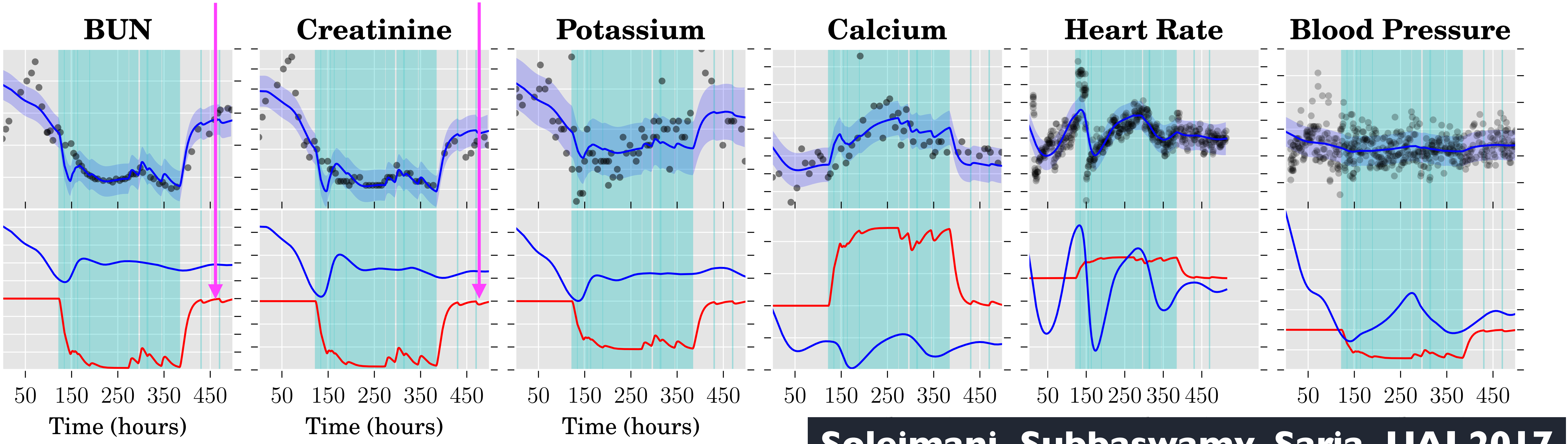
Schulam and Saria, NIPS 2017

Modeling multivariate data in the ICU



BUN and creatinine decrease during treatment, increase again after treatment is discontinued

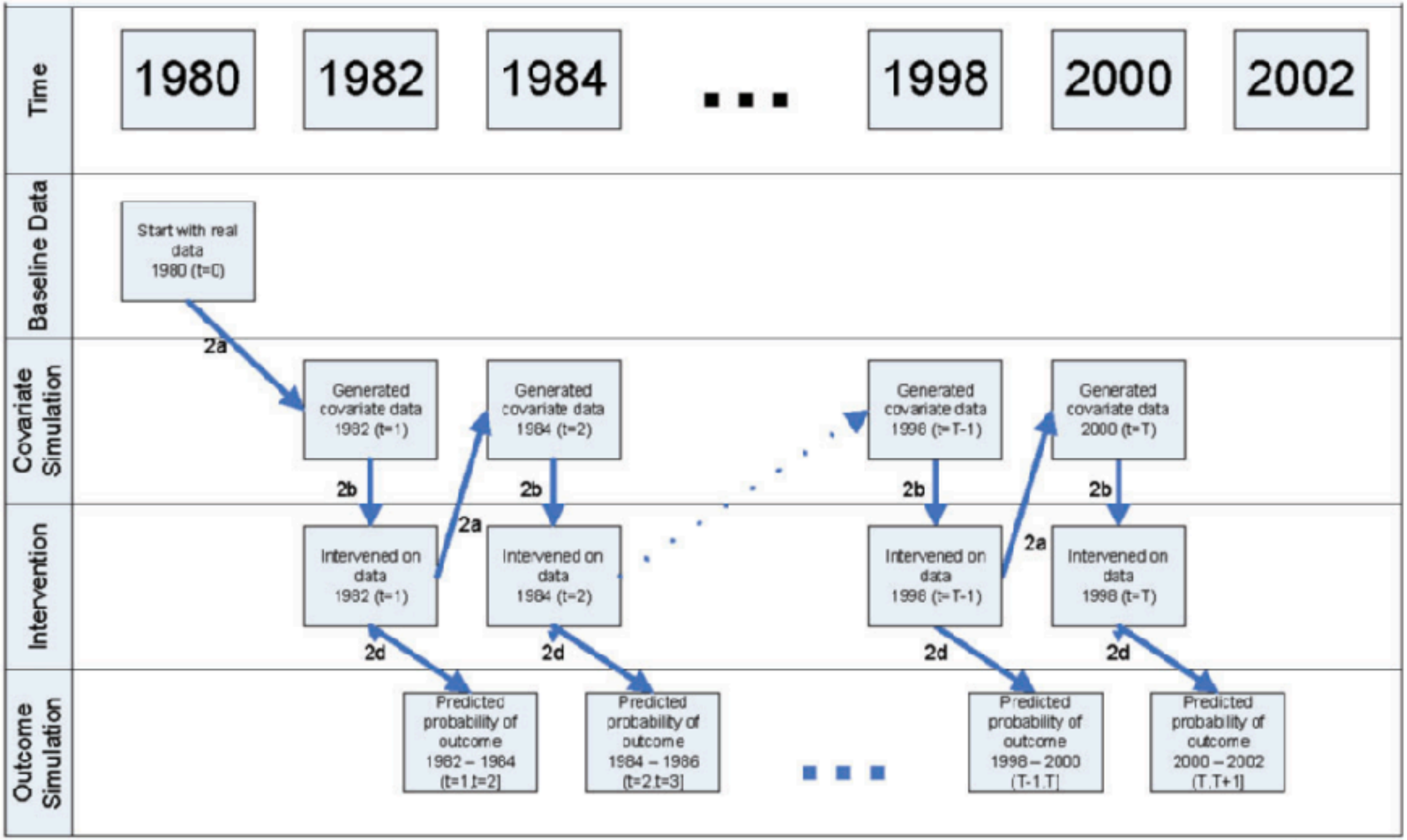
Negligible treatment response for BP



Soleimani, Subbaswamy, Saria, UAI 2017

Some other examples: Intervening on Coronary Heart Disease

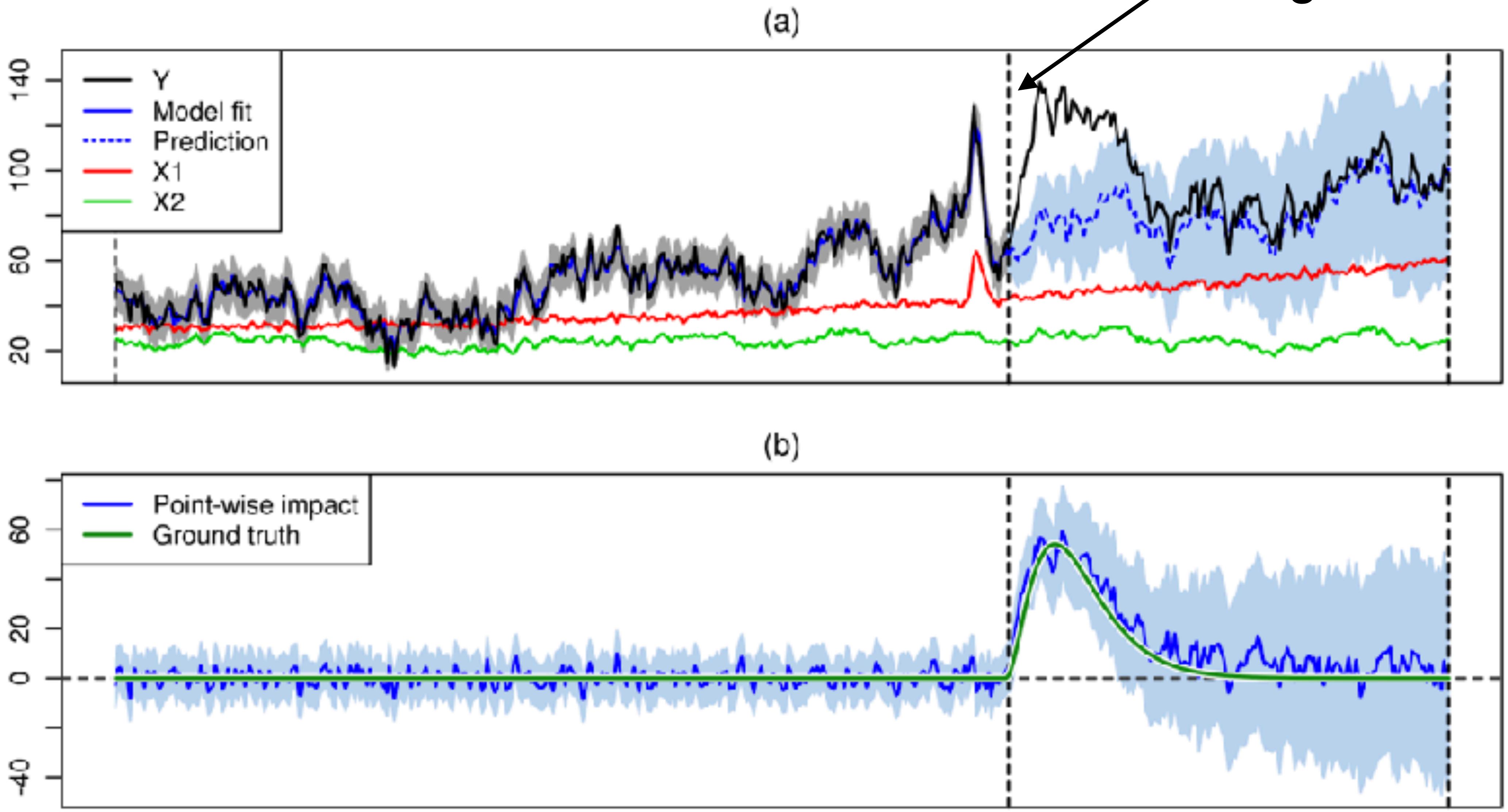
Estimate the population risk of coronary heart disease (CHD) under interventions such as quit smoking, maintain BMI < 25.


Time
Baseline Data
Covariate Simulation
Intervention
Outcome Simulation
1980
1982
1984
...
1998
2000
2002
Start with real data 1980 (t=0)
2a
Generated covariate data 1982 (t=1)
Generated covariate data 1984 (t=2)
Generated covariate data 1998 (t=T-1)
Generated covariate data 2000 (t=T)
2b
2b
2b
2b
2a
2a
Intervened on data 1982 (t=1)
Intervened on data 1984 (t=2)
Intervened on data 1998 (t=T-1)
Intervened on data 1998 (t=T)
2d
2d
2d
2d
Predicted probability of outcome 1982 – 1984 (t=1,t=2]
Predicted probability of outcome 1984 – 1986 (t=2,t=3]
...
Predicted probability of outcome 1998 – 2000 (T-1,T]
Predicted probability of outcome 2000 – 2002 (T,T+1]

Taubman et al. 2009

Potential Outcome Model for Estimating Effect of Ad Exposure

- Google's "Causal Impact"
  - Target time series Y: receives intervention
  - Control time series X1, X2. (Do not receive intervention.)
    - These are predictive of Y.
  - The relation between Y and (X1, X2) remains the same pre and post intervention.
  - Predict the counterfactual of Y (under no treatment) using X1 and X2.



Intervention begins on Y

Brodersen et al. 2014

Revisiting stability, robustness and bias

- More generally, given a problem, can we identify which relationships generalize i.e. are stable across datasets, which relationships change i.e. are unstable, and learn only the former?

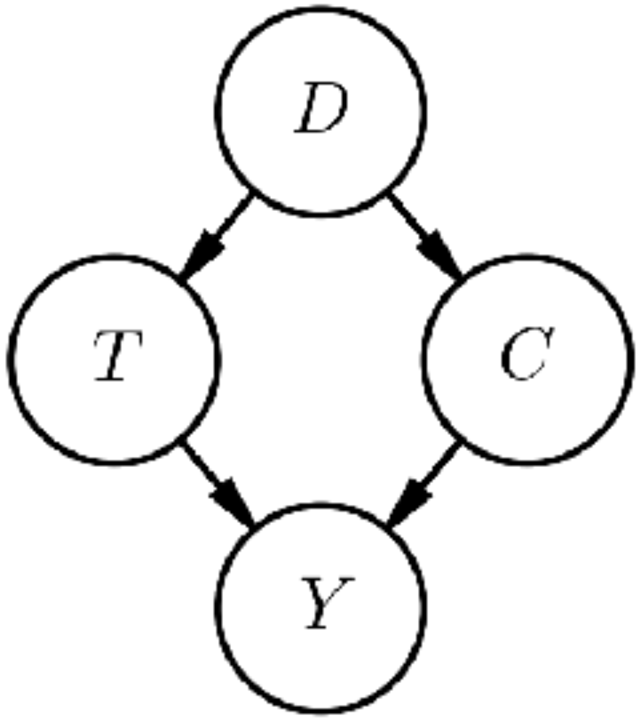
- The previous work relied on certain assumptions, specifically, the no unobserved confounders assumption that may not hold in practice. Can we relax these assumptions and if so, what can we recover?

- Beyond confounding bias, other types of biases exist in practice (e.g., selection bias).

- We use DAGS to reason about dependencies between variables; see Joris Mooij’s tutorial (or any introductory primer on causal graphs) if unfamiliar.

Diagnosis Example

- Goal: predict **T** from available features.



T: Meningitis
D: Smoking
C: Beta-blockers
Y: Blood pressure

- Some of these mechanisms will be stable across environments, others are unstable and more likely to change

- Ex: Effect of beta blockers and meningitis on blood pressure is likely stable. Ex: Policy for prescribing beta blockers to smokers is unstable—will vary from hospital to hospital.

- **A generalizable model should learn to predict using the stable relationships.**

Key idea: T|C,Y leads to an unstable model

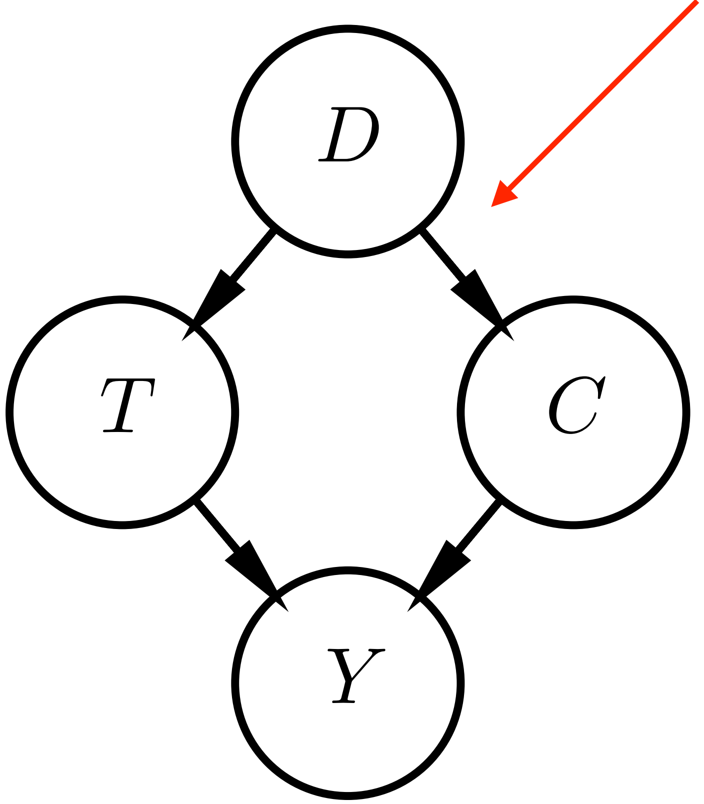
- Ideal: T | C, D, Y

  Stable to changes in P(C|D)
  Contains predictive information from D

- Naive: T | C, Y if **D is unobserved**

  Not stable to changes in P(C|D)
  Contains predictive information from D

This is an example with unobserved domain-dependent confounding

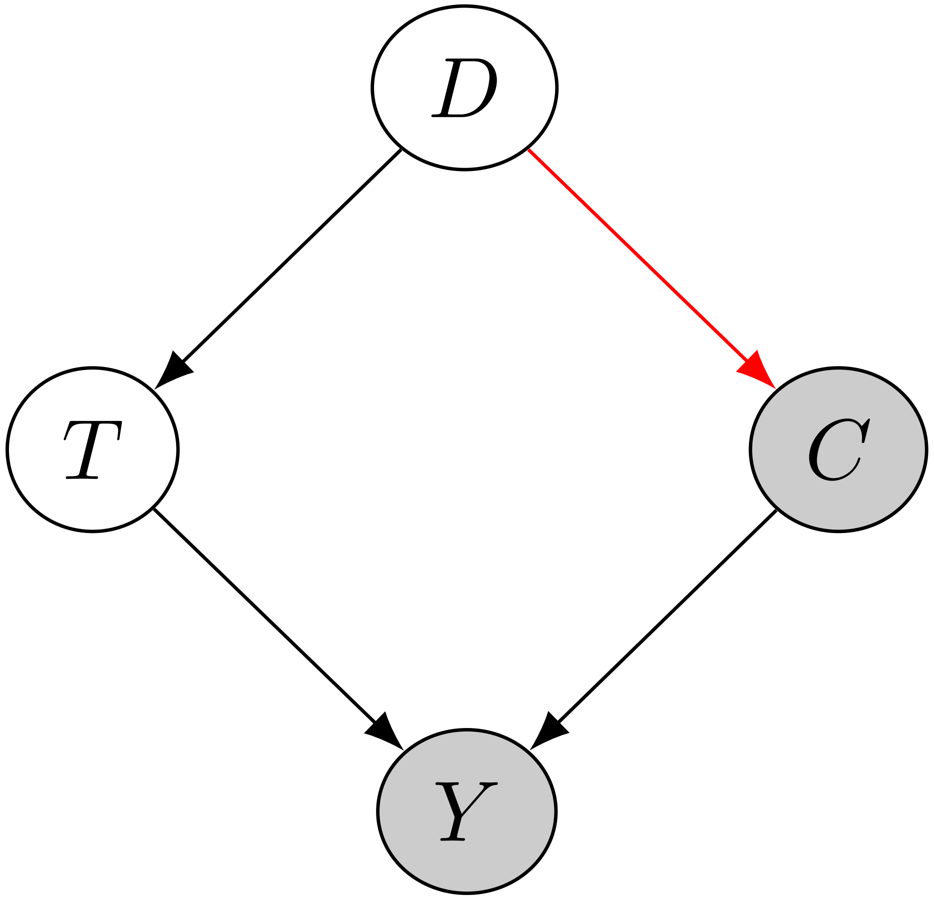


Policy-dependent

Unstable Paths

- Consider naive discriminative model P(**T**|**C**,**Y**)

- Two active paths from **C** to **T** when conditioned on **Y**:
  - $C \leftarrow D \rightarrow T$
  - $C \rightarrow Y \leftarrow T$

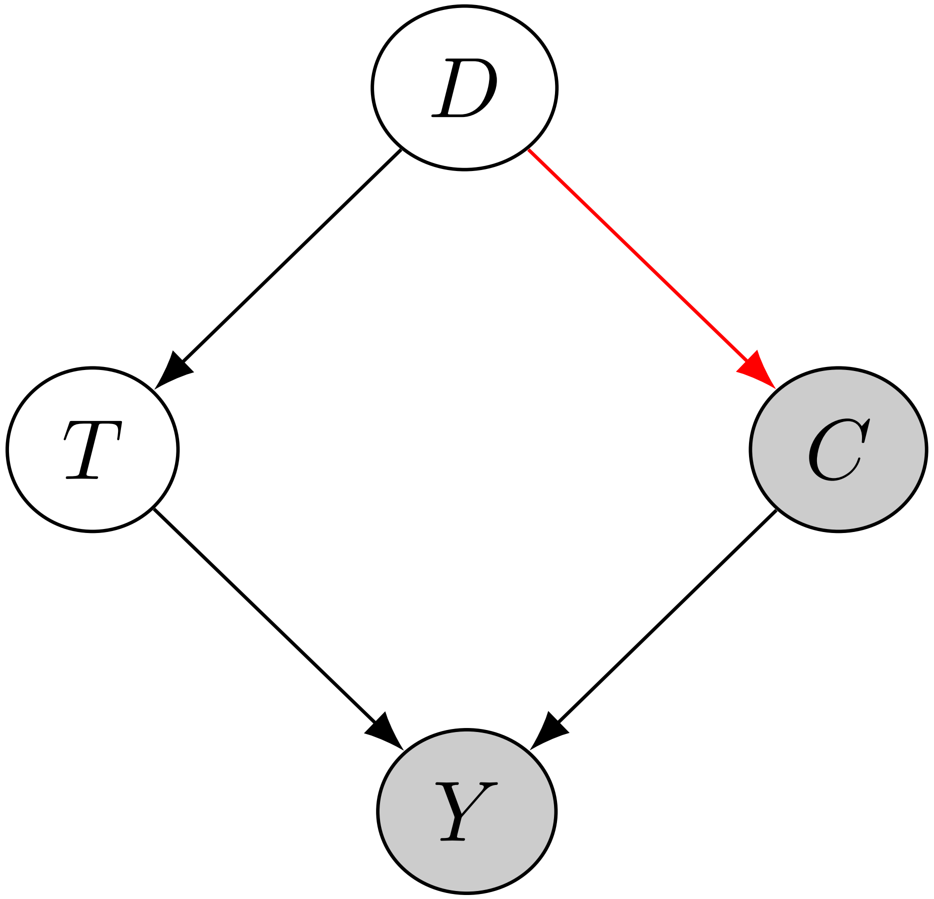


Determine active paths using d-separation

Unstable Paths

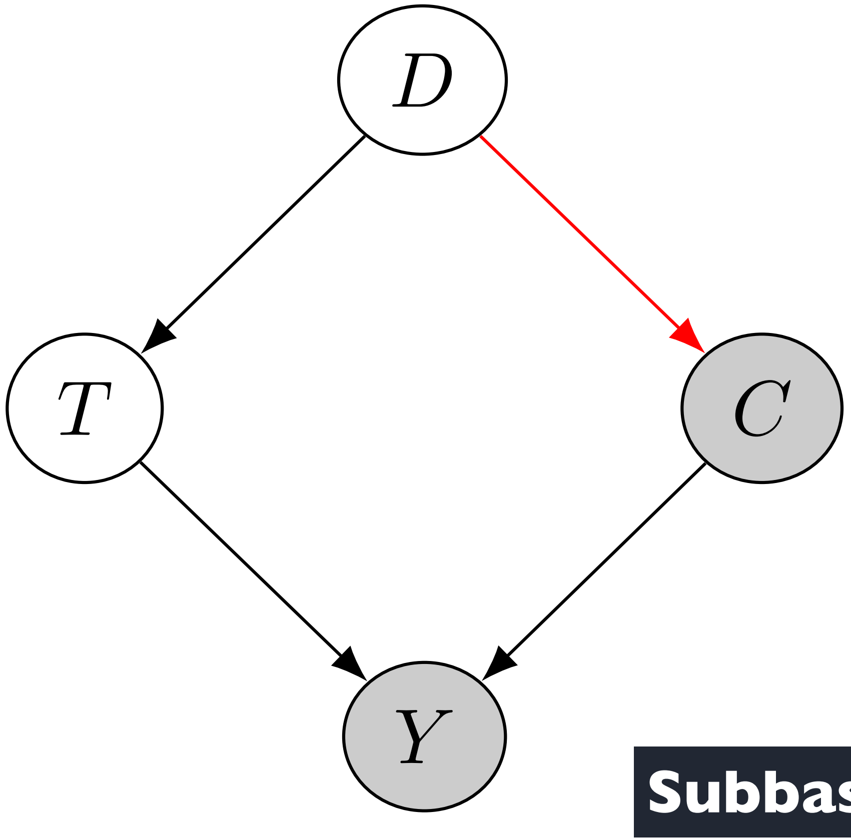
- Consider naive discriminative model P(**T**|**C**,**Y**)

- Two active paths from **C** to **T** when conditioned on **Y**:
  - $C \leftarrow D \rightarrow T$ **Unstable path**: encodes relationship that changes across domains
  - $C \rightarrow Y \leftarrow T$



Vulnerable Variables

- Consider naive discriminative model P(**T**|**C**,**Y**)
- **C** is vulnerable because it has an active unstable path to **T**
- Using **C** as a feature means we will learn relationship along both the stable and unstable paths
- Model will be unreliable and will not generalize



Subbaswamy and Saria, UAI 2018

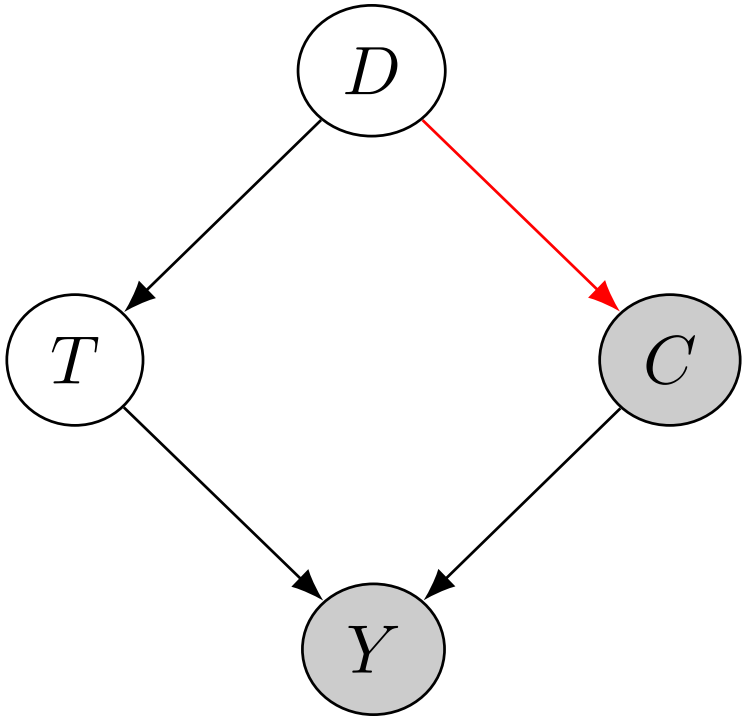
The Ideal Case

- Suppose that we were able to observe **D**. Then we could model P(**T**|**C**,**Y**,**D**).

- This model will be stable to changes in P(**C**|**D**)

- Why? The unstable path is not active: $C \leftarrow D \rightarrow T$

- Whether or not a feature is vulnerable (e.g., **C**) depends on what is unstable about data generating process and what we can condition upon.



Commentary: Tackling Dataset Shift

- We have given two example causes of dataset shift: differences in train and test distributions

- Typical machine learning approaches are reactive: use unlabeled samples from the test distribution to reweight training data. **Storkey, 2009** **Gretton et al., 2009**

- Similar problem and methods for transportability of causal effect estimates from one environment to another. **Pearl & Bareinboim, AAAI 2011**
    - "External validity": Causal models should generalize

- **Proactive Methods** which do not use test samples?

Subbaswamy and Saria, UAI 2018

Commentary: Addressing Dataset Shift

- Distributional robustness
  - Intuition: train predictive models that are optimal on distributions “close” to (empirical) training distribution
  - Takes the form of a regularizer in learning objective
  - Protects against perturbations of bounded strength (hyperparameter)
  - Guards against adversarial attacks and can improve generalization

Sinha et al., ICLR 2018 **Rothenhäusler et al., 2018**

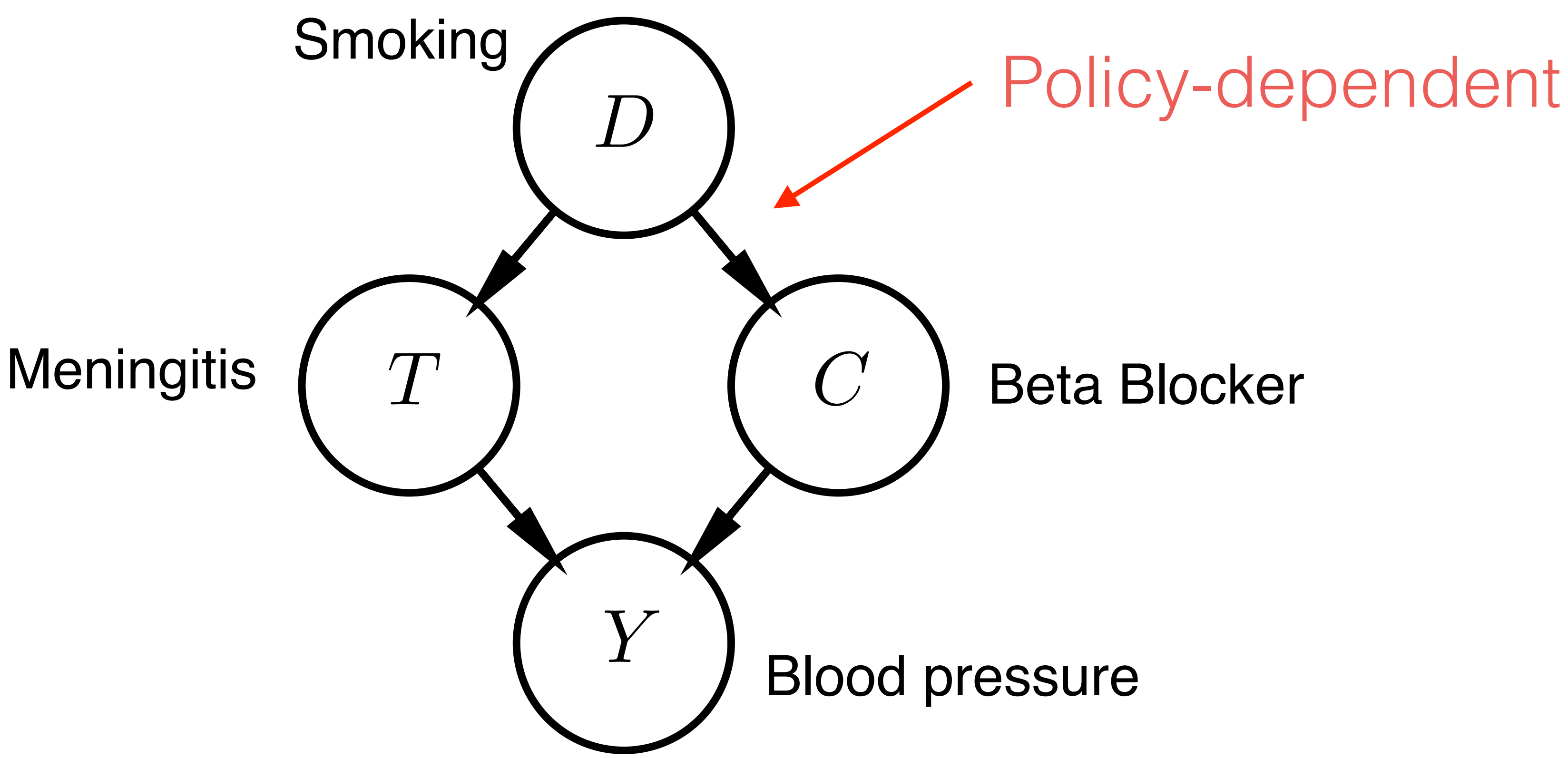
- While these methods are general purpose and easy to use, can be difficult to understand how they affect learned model.

- We propose using graphical knowledge of causal mechanisms to specify which changes to be invariant to.

  - Counterfactual Normalization

Subbaswamy and Saria, UAI 2018

Intuition



- Identify vulnerable variables —> variables that contain an active trail to T where one or more distributions along path maybe perturbed across datasets (unstable paths). Do not condition only vulnerable variables.

- More broadly, we want to only learn influence along stable paths and remove influence via unstable paths. How?

Walk you through a sketch of an algorithm …

Subbaswamy and Saria, UAI 2018

Example Solution: Graph Pruning

- Given a graph we can determine which components are stable and which are unstable.

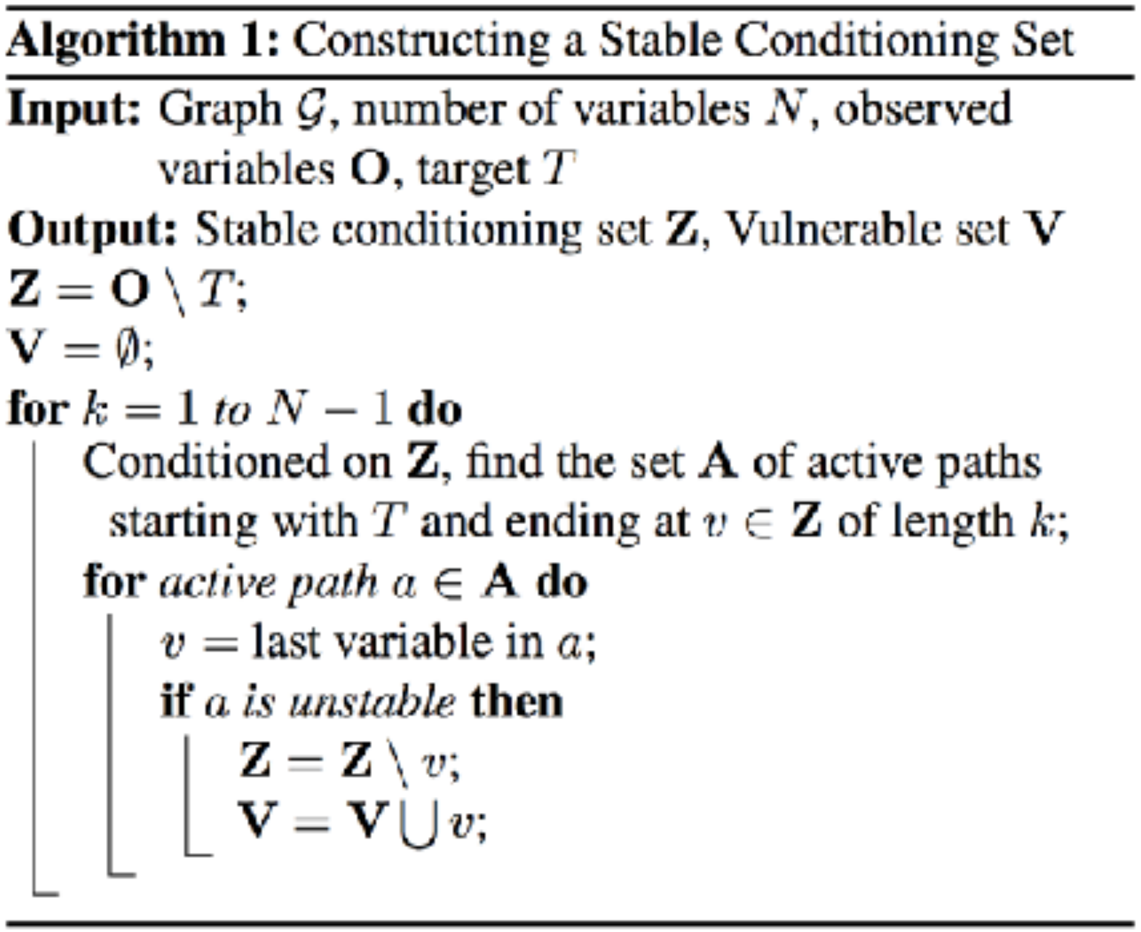
- Idea: Pick conditioning set (i.e., features in a discriminative model) that prunes the graph of unstable paths

- However, this will also prune stable paths.

- **Counterfactual Normalization**: consider adding counterfactual (potential outcome) features that retain some of the stable paths we removed during pruning.

Step 1: Constructing a Stable Conditioning Set via Graph Pruning

- Goal: Find set of observed variables that contains no active unstable paths while maximizing number of stable paths.
- First: Find a stable set **Z**
- Start by conditioning on all observed variables.
- Consider active unstable paths starting at **T** of increasing length and remove ending variable from conditioning set

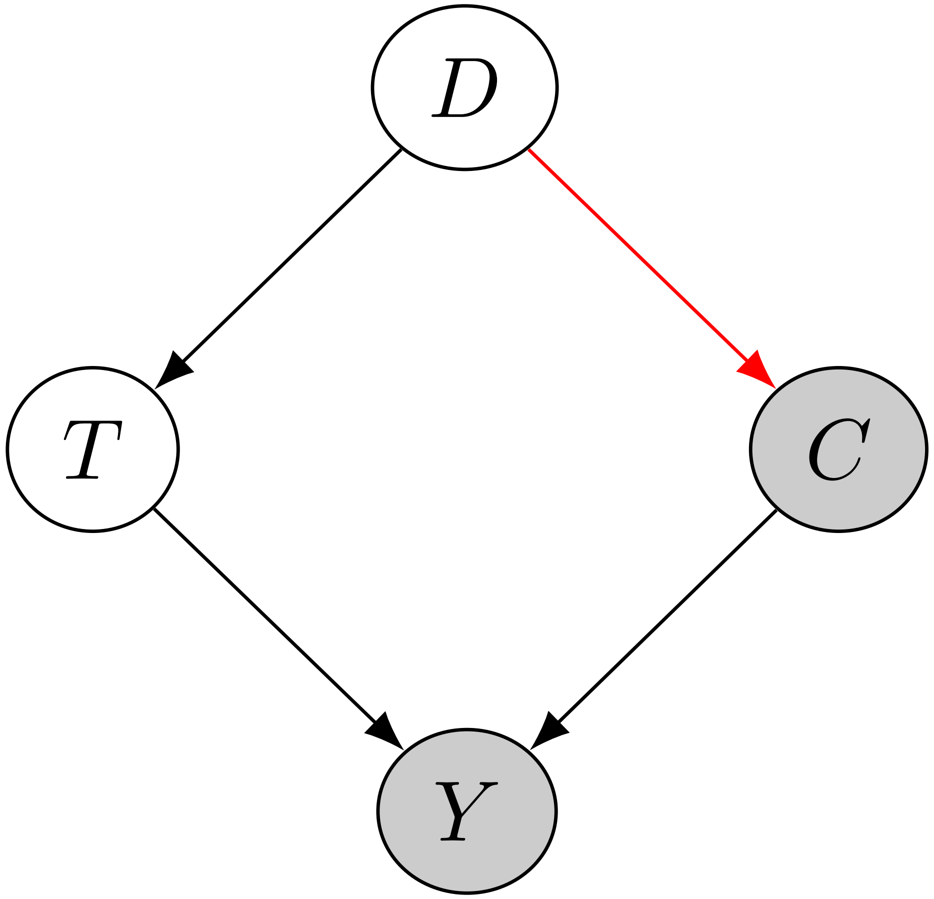


```
Algorithm 1: Constructing a Stable Conditioning Set
Input: Graph $\mathcal{G}$, number of variables $N$, observed
       variables $\mathbf{O}$, target $T$
Output: Stable conditioning set $\mathbf{Z}$, Vulnerable set $\mathbf{V}$
$\mathbf{Z} = \mathbf{O} \setminus T$;
$\mathbf{V} = \emptyset$;
for $k = 1$ to $N - 1$ do
    Conditioned on $\mathbf{Z}$, find the set $\mathbf{A}$ of active paths
      starting with $T$ and ending at $v \in \mathbf{Z}$ of length $k$;
    for active path $a \in \mathbf{A}$ do
        $v =$ last variable in $a$;
        if $a$ is unstable then
            $\mathbf{Z} = \mathbf{Z} \setminus v$;
            $\mathbf{V} = \mathbf{V} \bigcup v$;
```

Subbaswamy and Saria, UAI 2018

Example: Graph Pruning

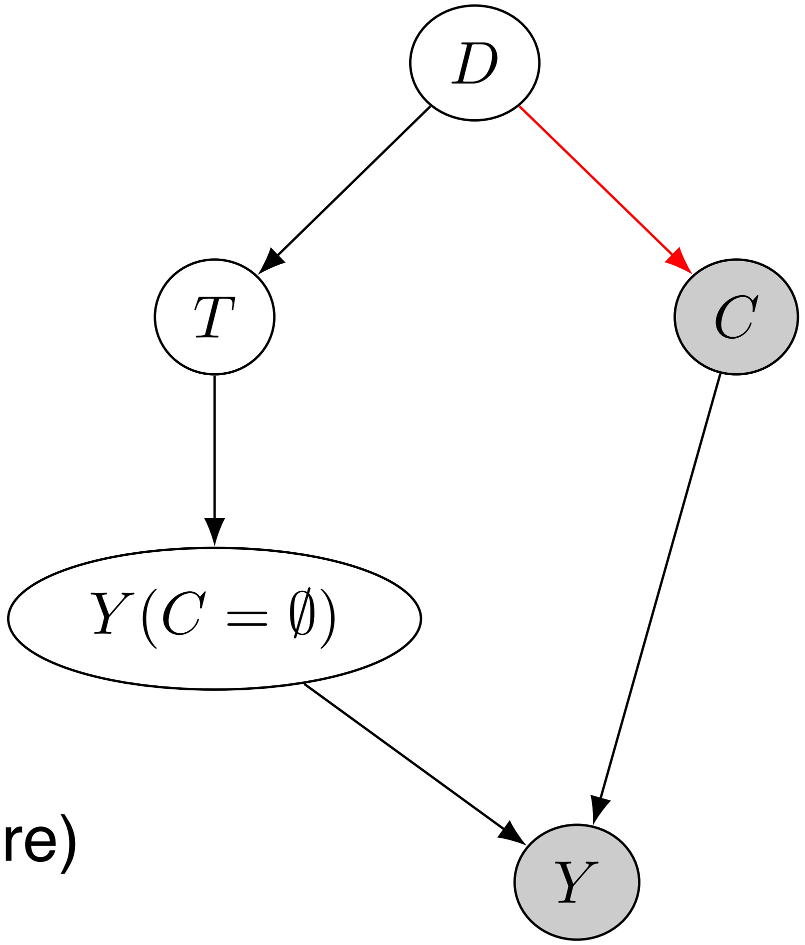
- **Z** = {**C**, **Y**}; **V** = {}
- Unstable path: $T \leftarrow D \rightarrow C$
- **Z** = {**Y**}; **V** = {**C**}
- Unstable path: $T \leftarrow D \rightarrow C \rightarrow Y$
- **Z** = {}; **V** = {**C**, **Y**}
- Stable conditioning set is empty!



Retaining Stable Paths

- Can we expand the stable conditioning set?
  - Include some of the vulnerable variables (vars that were removed) that may no longer have active unstable paths.
  - Or can we include adjusted versions of the vulnerable variables?

- If a variable has both stable and unstable paths to **T**, can we isolate its stable paths from the unstable paths?
  - **Y**: Observed/factual blood pressure
    C: Whether or not patient takes beta blockers
  - $Y(C = \emptyset)$: Patient's blood pressure if we removed the effects of beta blockers (i.e., untreated blood pressure)

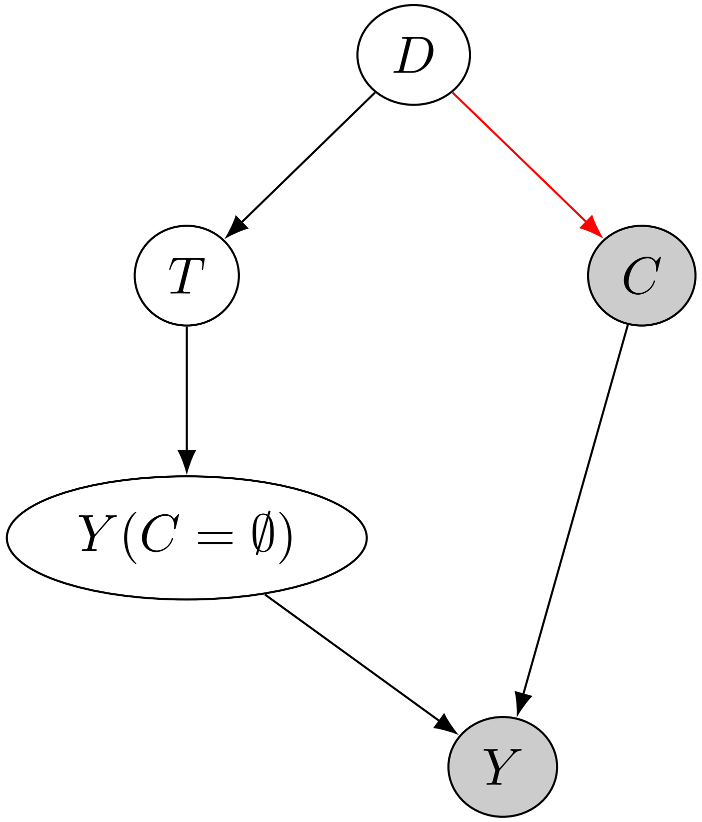


Implications of Node-splitting

- If a variable has an unstable path through its observed parent, intervening on the parent results in a counterfactual without this unstable path.

- Factual version of variable acts as collider for unstable path

- $T \leftarrow D \rightarrow C \rightarrow Y \leftarrow Y(C = \emptyset)$ unstable path to counterfactual is blocked if we do not condition on **Y**.



The Three Cases

- Ideal: T | C, D, Y

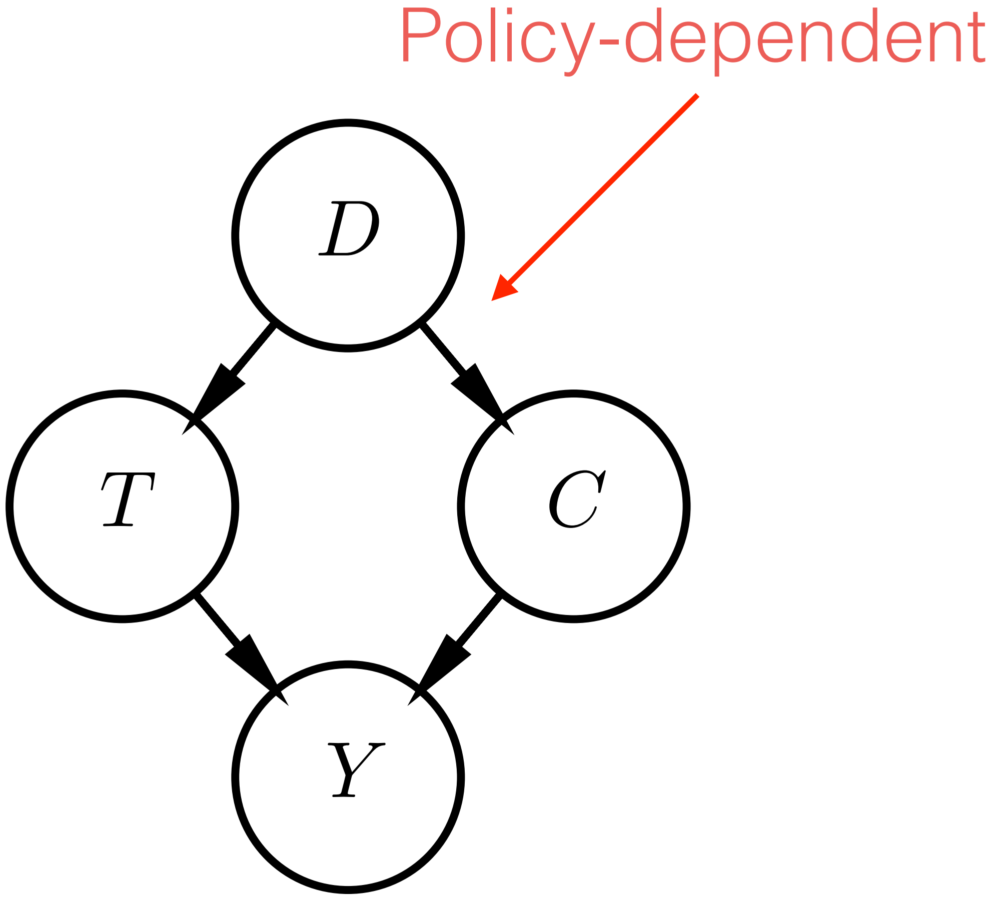
  Stable to changes in P(C|D)
  Contains predictive information from D

- Naive: T | C, Y

  Not stable to changes in P(C|D)
  Contains predictive information from D

- Counterfactually Normalized (CN): T | Z = Y(C=0)

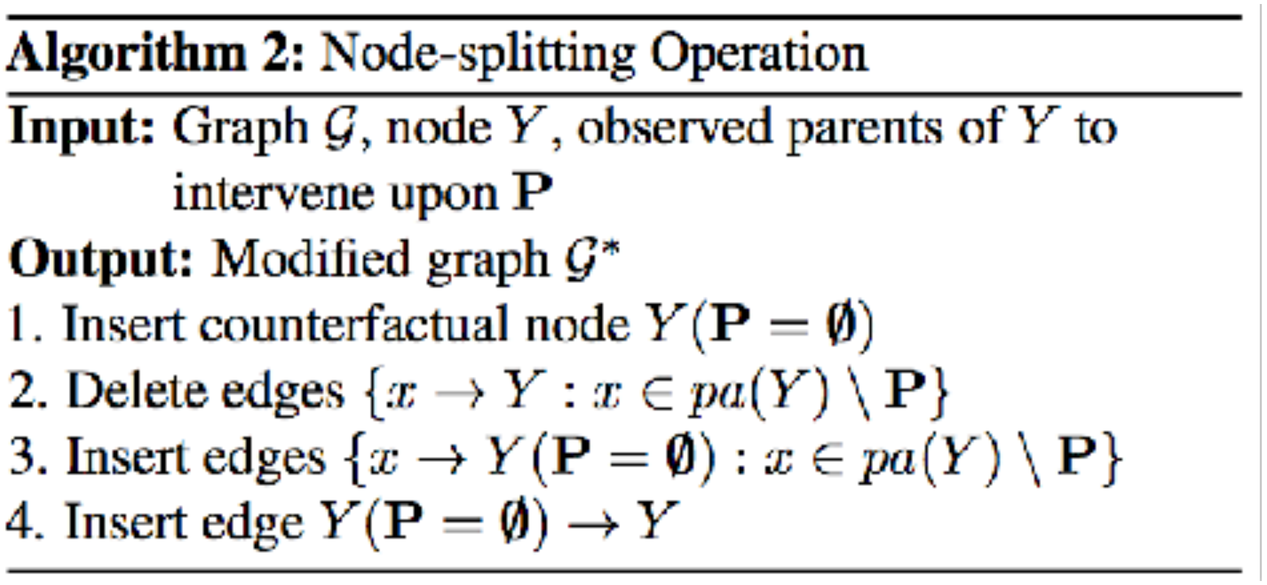
  Stable to changes in P(C|D)
  Contains no information from D



Subbaswamy and Saria, UAI 2018

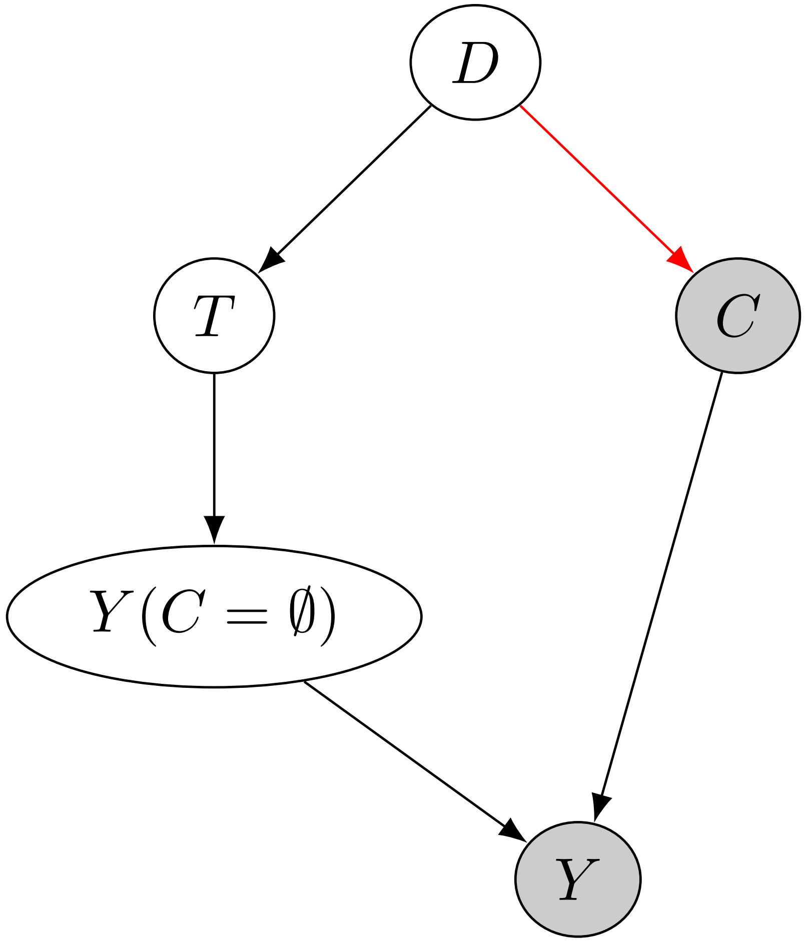
Step 2: Node-splitting—Graphically Representing Counterfactuals

- Intermediate Counterfactual: parent of factual version
  - Generatively, represents value of variable before effects of “null” parents occurred
  - Counterfactual takes parents that were not intervened upon (including ε_v).



Algorithm 2: Node-splitting Operation

```
Input: Graph $\mathcal{G}$, node $Y$, observed parents of $Y$ to
       intervene upon $\mathbf{P}$
Output: Modified graph $\mathcal{G}^*$
1. Insert counterfactual node $Y(\mathbf{P} = \emptyset)$
2. Delete edges $\{x \to Y : x \in pa(Y) \setminus \mathbf{P}\}$
3. Insert edges $\{x \to Y(\mathbf{P} = \emptyset) : x \in pa(Y) \setminus \mathbf{P}\}$
4. Insert edge $Y(\mathbf{P} = \emptyset) \to Y$
```



Counterfactual Normalization: Retaining Stable Paths

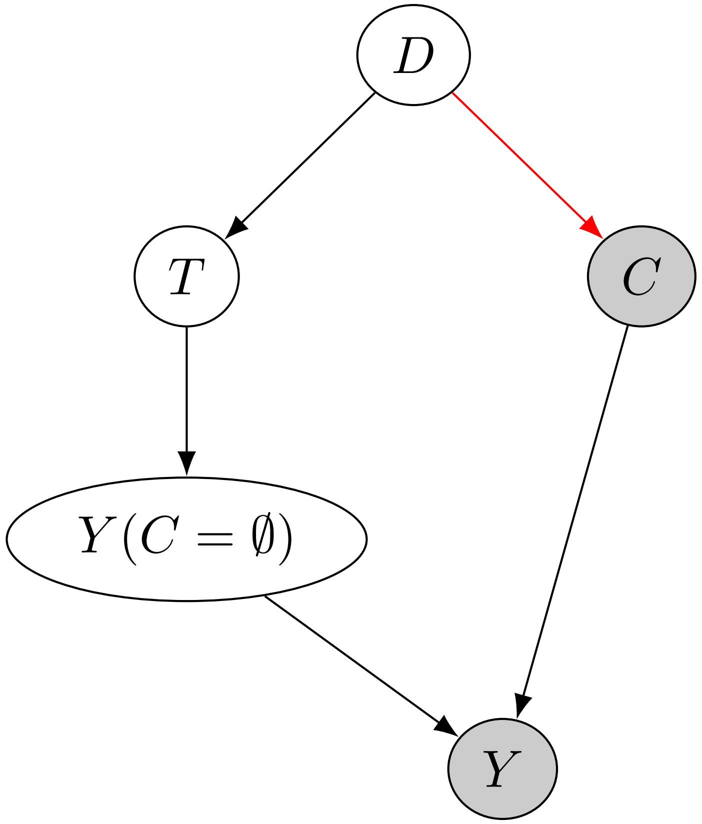
- For each vulnerable variable, try adding variable or counterfactual version to stable conditioning set.

- **Z** = {}; **V** = {**C**, **Y**}

- **C** has unstable path through unobserved parent.

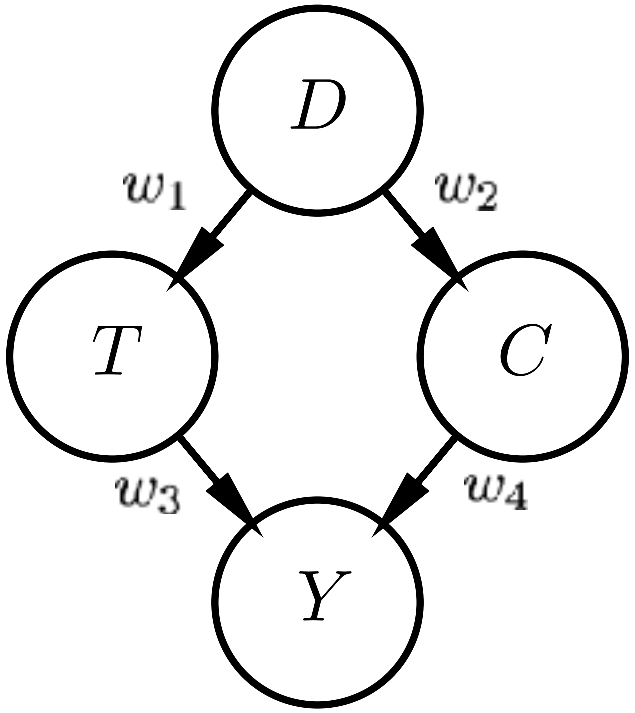
- **Y** has unstable path through observed parent! Node-split and add counterfactual.

- **Z** = {**Y**(**C**= Ø)}; **V** = {**C**, **Y**}



- Estimate **Y**(**C**= Ø) and predict by modeling P(**T**|**Y**(**C**= Ø))

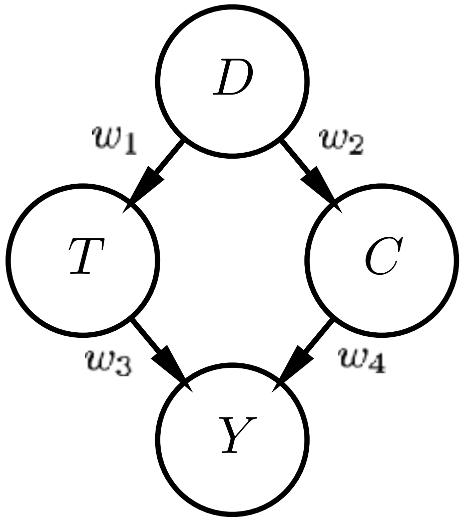
Linear Gaussian Example

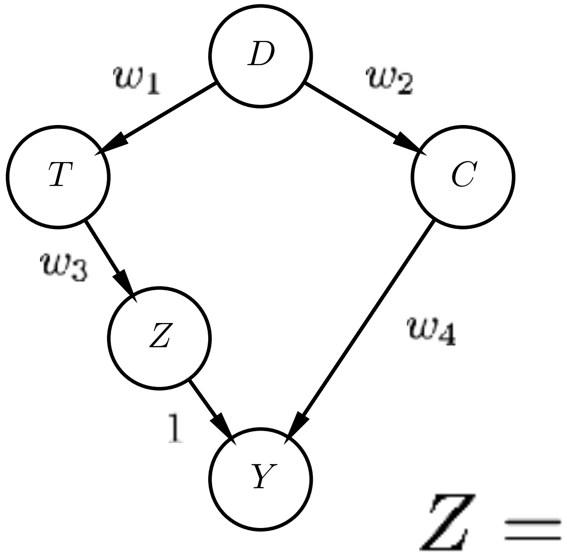


$$D = \varepsilon_D$$
$$T = w_1 D + \varepsilon_T$$
$$C = w_2 C + \varepsilon_C$$
$$Y = w_3 T + w_4 C + \varepsilon_Y$$
$$\varepsilon_D, \varepsilon_T, \varepsilon_C, \varepsilon_Y \sim \mathcal{N}(0, 0.1^2)$$

- Changing w_2 corresponds to changing P(**C|D**)
- w_2 is fixed in the training data, but in different target populations (e.g., hospitals) value may change arbitrarly

Linear Gaussian Example: Node-splitting





$$Z = Y(C = \emptyset)$$

- These are equivalent models of the data generating process.

$$Z = w_3 T + \varepsilon_Y$$
$$Y = Z + w_4 C$$

- **Y** is now a deterministic function of **C** and **Z**.

- Can easily estimate counterfactual as $Z = Y - w_4 C$

Three Ways of Predicting

- Ideal: T | C, D, Y

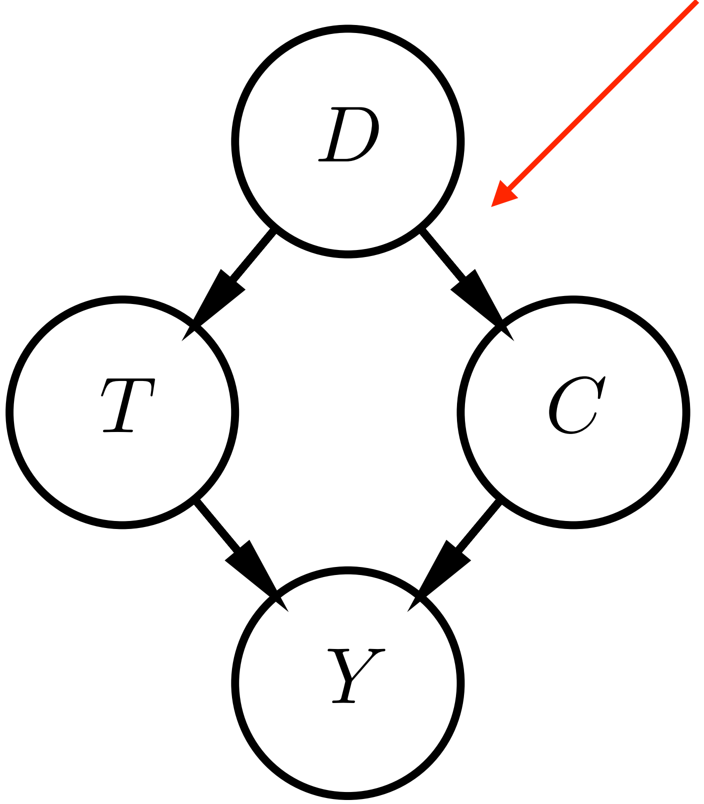
  Stable to changes in P(C|D)
  Contains predictive information from D

- Naive: T | C, Y

  Not stable to changes in P(C|D)
  Contains predictive information from D

- Counterfactually Normalized (CFN):
  T | Z = Y(C=0)

  Stable to changes in P(C|D)
  Contains no information from D



Policy-dependent

Three Ways of Predicting

- Ideal:

$$\hat{T} = \alpha_1 D + \alpha_2 C + \alpha_3 Y$$

Stable to changes in
Contains predictive information from D

- Naive:

$$\hat{T} = \beta_1 C + \beta_2 Y$$

Not stable to changes in
Contains predictive information from D

- Counterfactually Normalized (CFN):

$$\hat{T} = \gamma_1 \hat{Z}$$

Stable to changes in
Contains no information from D

Experiment

- Generate N=30000 training data points from SCM with w_2=2 in training domain.

- Train Least Squares (LS) models for

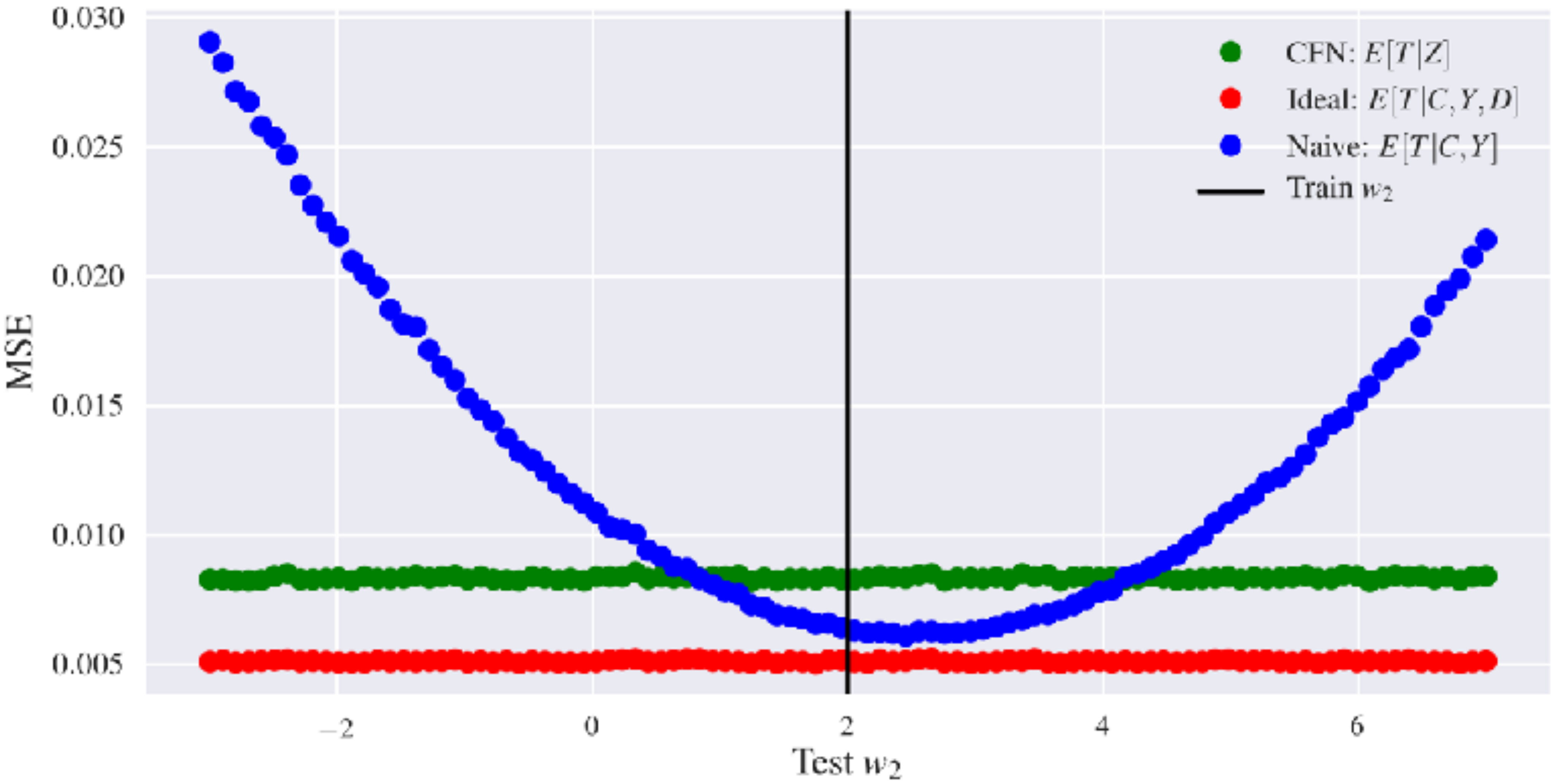
$$E[T|Y,C],\ E[T|Z],\ E[T|Y,C,D]$$

- Generate 100 test datasets

- Vary w_2 from -3 to 7 in test datasets

- Plot MSE of naive, counterfactually normalized, and ideal cases for each test w_2

Results



- CFN is stable to variations in P(**C**|**D**)

- Naive model that uses all observed features does not have stable performance.

Beyond Confounding Bias

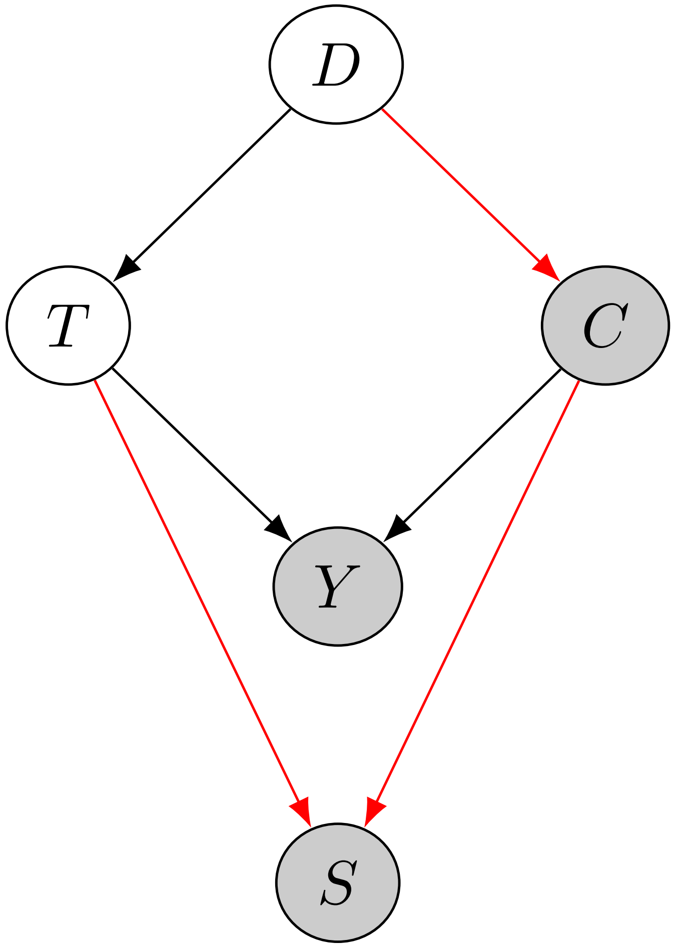
- How else do unstable paths arise?

- Selection bias: training data generated according to some selection mechanism, P(**S**|pa(**S**))

- Patients without meningitis (**T**=0) who take beta blockers (**C**=1) for their chronic condition may be underrepresented in the hospital training data (**S**=1) because of a local chronic care clinic.

- New unstable path in training data due to selection collider

$$C \rightarrow S \leftarrow T$$



Beyond Confounding Bias

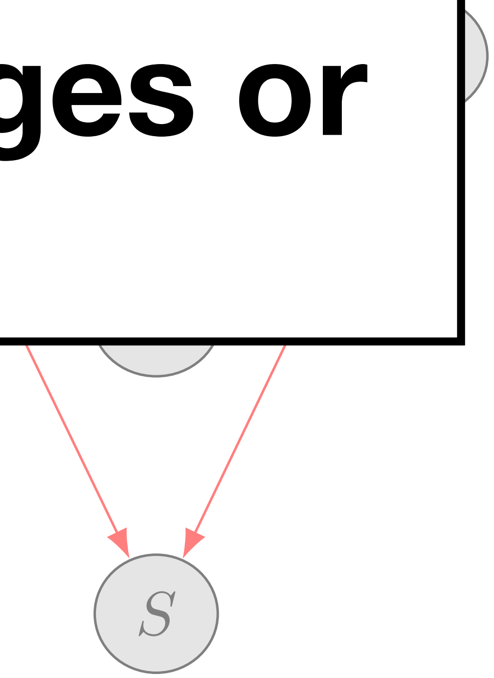
- How else do unstable paths arise?

- Selection bias: training data generated according to some selection mechanism, P(**S**|pa(**S**))

- Learned relationship between C and T through S will not generalize when the selection mechanism changes or is no longer present.

- New unstable path in training data due to selection collider

$C \rightarrow S \leftarrow T$



CFN Takeaways

- When learning from retrospective datasets, models may encode **unintended dataset-specific biases** that hurts quality of decision-making at test time. For example, the model may **learn relationships that are *unstable*—**associations that exist in the training data but do not hold or change at test time.
    - Takeaway #1: **Can we identify relationships that are stable and only learn these**? Yes.
    - Takeaway #2: How do we identify these? Use knowledge of the causal DAG to proactively identify and remove variables w/ unstable paths of influence.
        - For example, when training a discriminative model, conditioning on a variable learns influences via all active paths from that variable to the target outcome variable.
        - Investigate paths b/w predictor and outcome in the causal DAG to **identify paths that are unstable**.
        - Mark predictors with unstable paths to outcome as *vulnerable*. Conditioning on these will produce models that capture unstable relationships.
    - Takeaway #3: **Safe to condition on predictor variables with no unstable paths** (non *vulnerable* variables) —> resulting model will generalize across datasets. But, is not optimal.

    - Takeaway #4: For predictors w/ both stable and unstable paths, can we learn influence only via stable paths?
    - Yes, perhaps…**augment conditioning set to add new counterfactual features.**
- Takeaway #5: Above method for correction **applicable in graphs where the no unobserved confounding assumption is not satisfied**.

Subbaswamy and Saria, UAI 2018

Conclusions: Big Picture

- We can frame generalization in **terms of differences in the data generating process across environments**.

- When learning from retrospective datasets, models may encode **unintended dataset-specific biases** that hurts quality of decision-making at test time. For example, the model may **learn relationships that are *unstable*—**associations that exist in the training data but do not hold o r change at test time.

- **Knowledge of the data generating process** (i.e. causal DAG) allows us to explicitly **reason about scenarios under which we can learn stable models**.

- Further, we can **constrain learning so that the resulting models are invariant to unstable relationships**.

  - Example: Discussed potential outcome models for what-if reasoning over temporal trajectories —> learns relationships between predictors and outcome that are stable across environments. Requires certain assumptions to hold.

  - Example: Discussed counterfactual normalization, feature augmentation procedure that only learns relationships that are unstable. Applicable in settings with unmeasured confounding. Requires certain other assumptions to hold.

- Contrast above **ideas as proactive methods for adjusting for dataset-specific bias** as opposed to reactive methods that correct via reweighing when samples from the target distribution become available.

Reading List

- Example papers on the use of counterfactual reasoning for decision-making

Taubman et al. 2009 Bottou et al., 2013

Brodersen et al. 2014 Schulam et al., NIPS 2017 Soleimani et al. UAI 2017

- Papers discussing the issue of lack of model reliability / need for robustness to certain perturbations in prediction

Dyagilev et al., Machine Learning 2015 Caruana et al., KDD 2015

Schulam et al., NIPS 2017

Sinha et al., ICLR 2018 Rothenhäusler et al., 2018

Subbaswamy and Saria, UAI 2018

- Tutorials on DAGs and assessing independence assertions on a graph

You will need to understand the following concepts: DAG, Bayes-ball theorem, D-separation. Coursera has multiple classes that teaches these. Most will teach you a lot more. It is most beneficial is to learn how to construct a graph that captures a given set of independence assertions for a given problem. I recommend taking this on as an exercise and running your work by someone else who is familiar and can critique your graph.

Thank you!
ssaria@cs.jhu.edu

@suchisaria

All references throughout the slides are active links and clickable.

Appendix

(3) No Unmeasured Confounders (NUC)

- In our exercise example, BMI is a *confounder*
  - BMI induces a statistical dependency between the observed treatment and observed outcome
- In general, unless we observe all confounders, we cannot learn unbiased models of potential outcomes from observational data
- Formally, NUC is an statistical independence assertion:

$$Y(a) \perp A \mid \mathbf{X} = \mathbf{x} \quad : \quad \forall a \in \mathcal{A}, \forall \mathbf{x} \in \mathcal{X}$$

Making NUC intuitive using Single-World Intervention Graphs

- **SWIGs extend graphical models to explicitly represent potential outcomes**

- To obtain a SWIG, we define a causal graphical model and specify the set of treatment variables

- We apply *node-splitting* operations to treatment variables to represent interventions

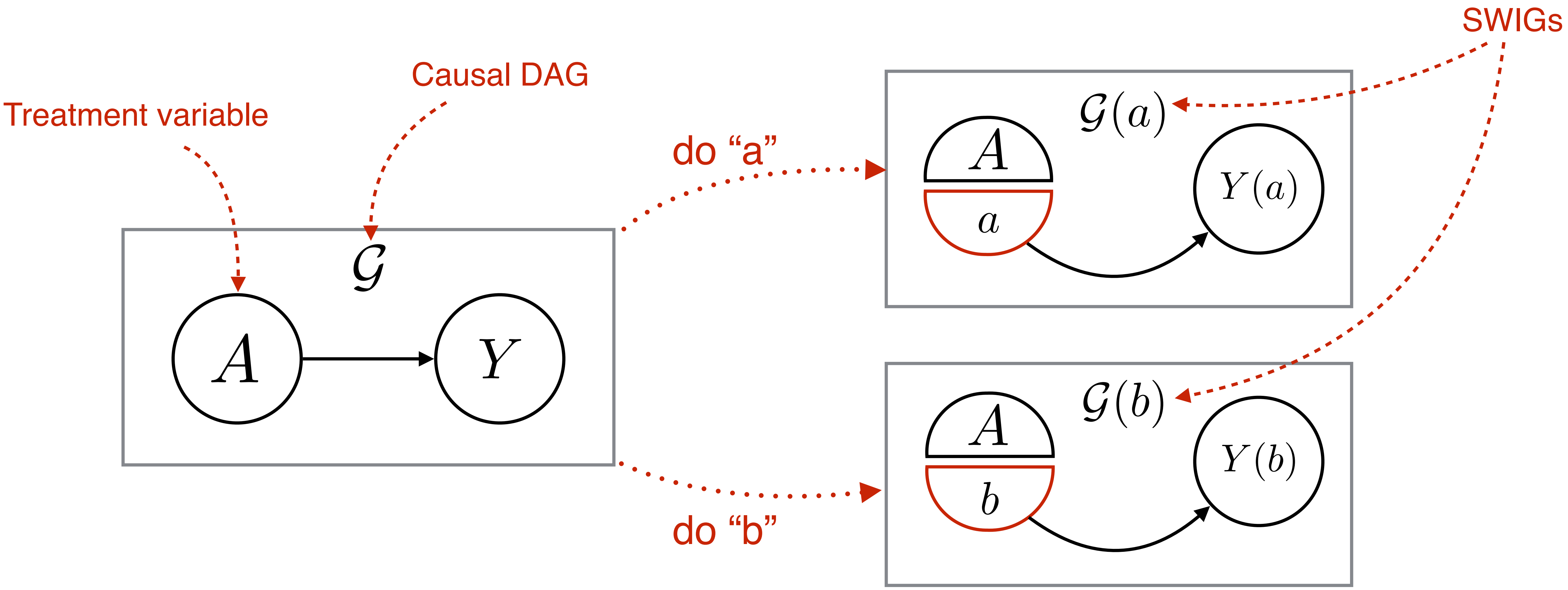
- Useful tool to determine which conditional distributions you need, and how to simulate trial

NIPS tutorial: Richardson, 2014

Richardson and Robins, 2014

Example SWIG

- We apply ***node-splitting*** **operations to treatment variables to represent interventions**
- A simple “a” vs “b” example:

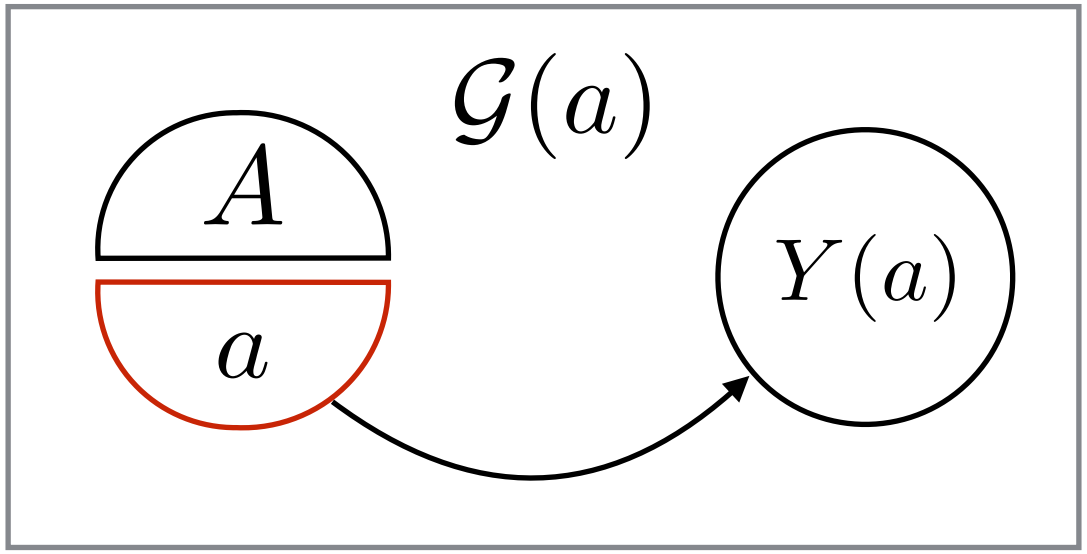


Richardson, 2014

Richardson and Robins, 2014

Interpreting SWIGs

- Treat SWIGs as standard causal graphs
  - Semi-circle nodes are just reminders that we have applied a node-splitting operation
- From this graph, can read that Y(a) is independent of the observed treatment A



Richardson, 2014

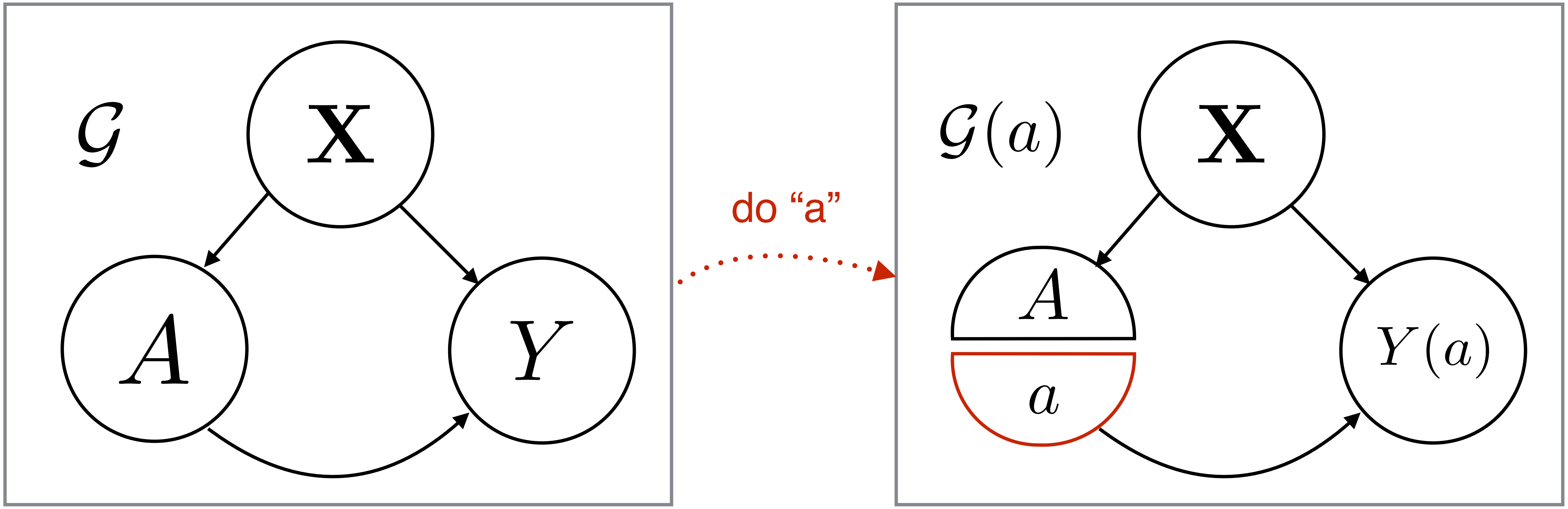
Richardson and Robins, 2014

NUC in SWIG Language

- SWIGs make NUC assumption easy to express

$$Y(a) \perp A \mid \mathbf{X} = \mathbf{x} \quad : \quad \forall a \in \mathcal{A}, \forall \mathbf{x} \in \mathcal{X}$$

- Confounders X d-separate potential outcomes from observed treatment random variable when intervening on treatment



Richardson, 2014

Richardson and Robins, 2014