Causal Effects and Overlap in High-Dimensional or Sequential Data

Fredrik D. Johansson
IMES & CSAIL, MIT

w. David Sontag, Uri Shalit, Rajesh Ranganath, Nathan Kallus


Diabetes onset

Predicting Type 2 diabetes onset

Millions of US adults are affected by Type 2 diabetes

The disease has severe symptoms and complications but is often preventable if risk factors are identified

Who is at risk of developing Type 2 diabetes?


Opioid addiction

Drivers of opioid addiction

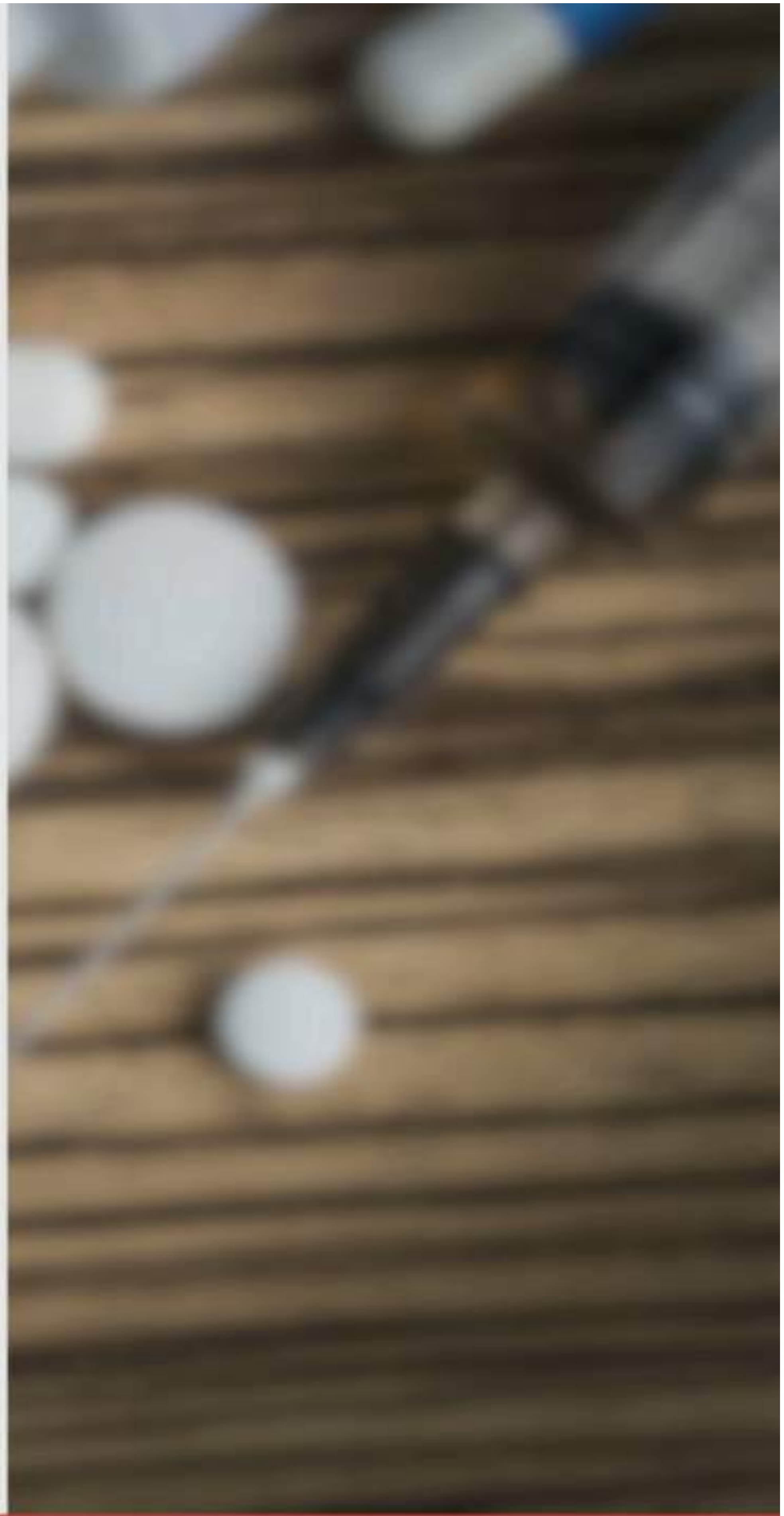
Millions of patients are addicted to opioid medications

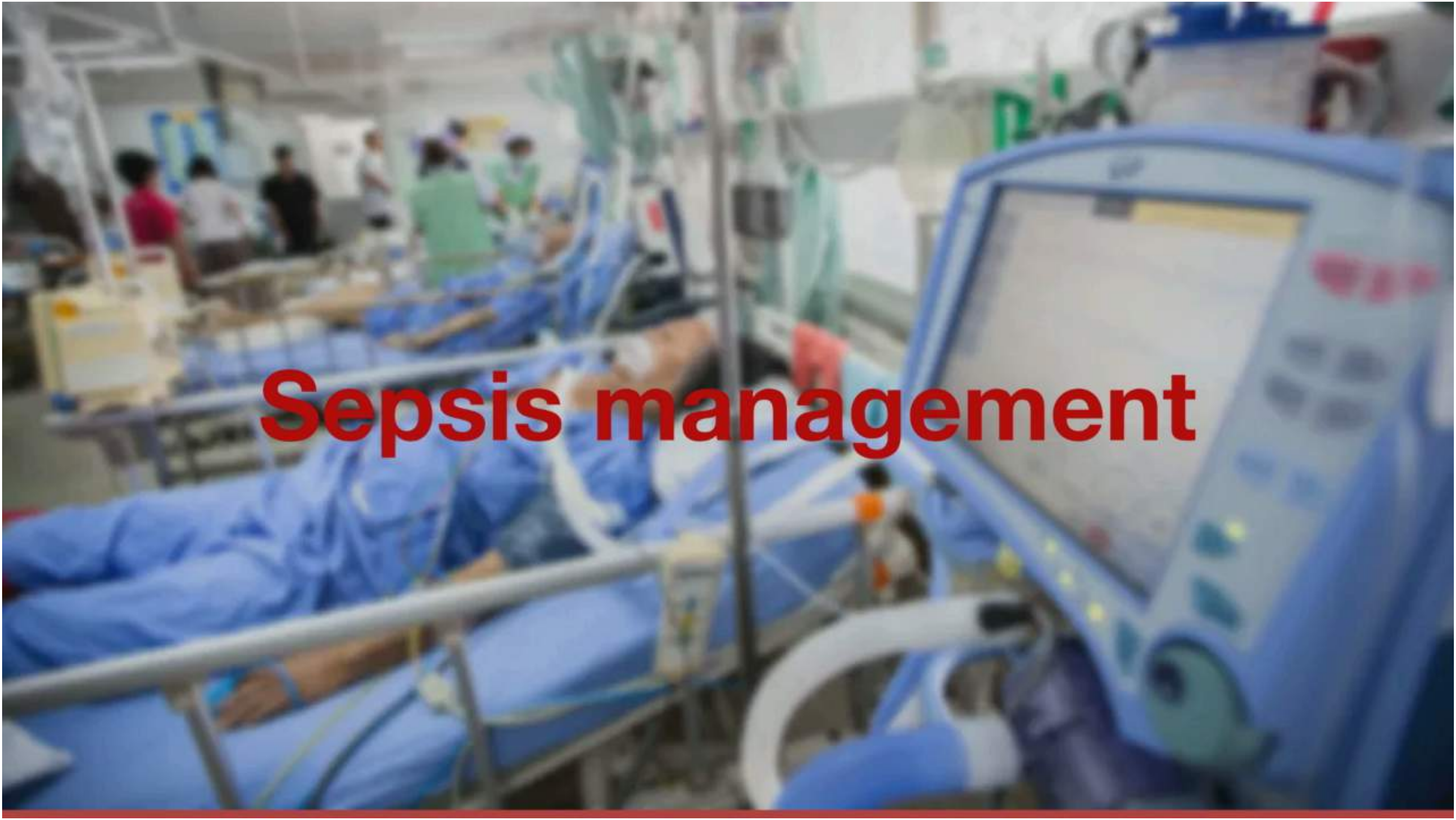
10 000 people die each year from overdoses

Larger subscriptions are associated with higher risk

What else **drives addiction**?

Who should be prescribed what?



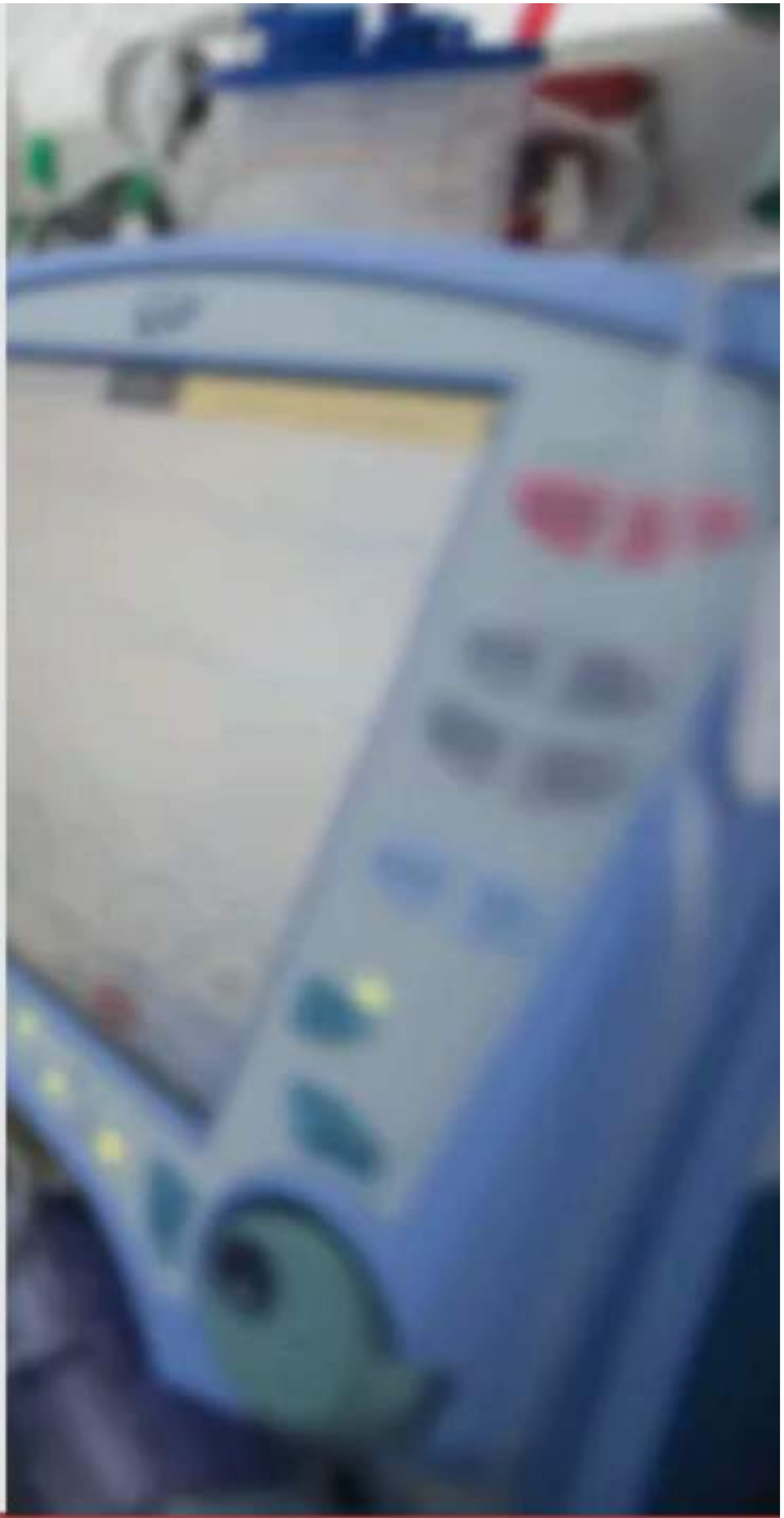

Sepsis management

Sepsis management

Sepsis is one of the **leading causes of death** in the ICU, and is a complication of an infection

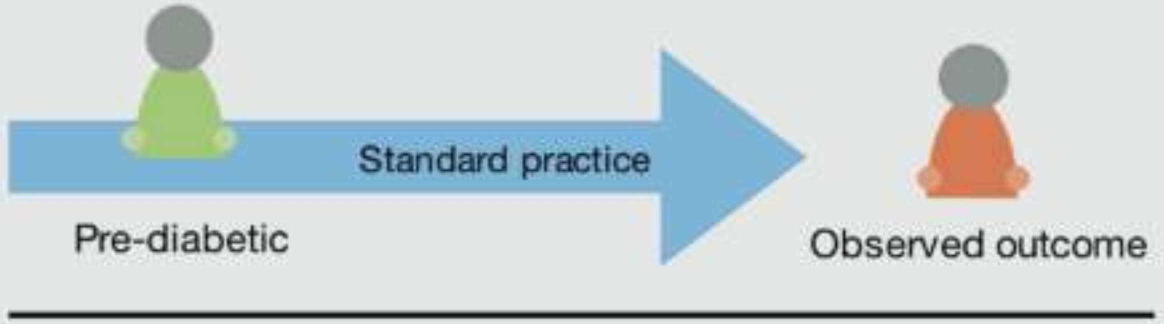
Aside from staving off infection, treatment involves **managing vitals** like blood pressure, heart rate, oxygen intake

How should we manage patients to maximize **long-term quality of life**?



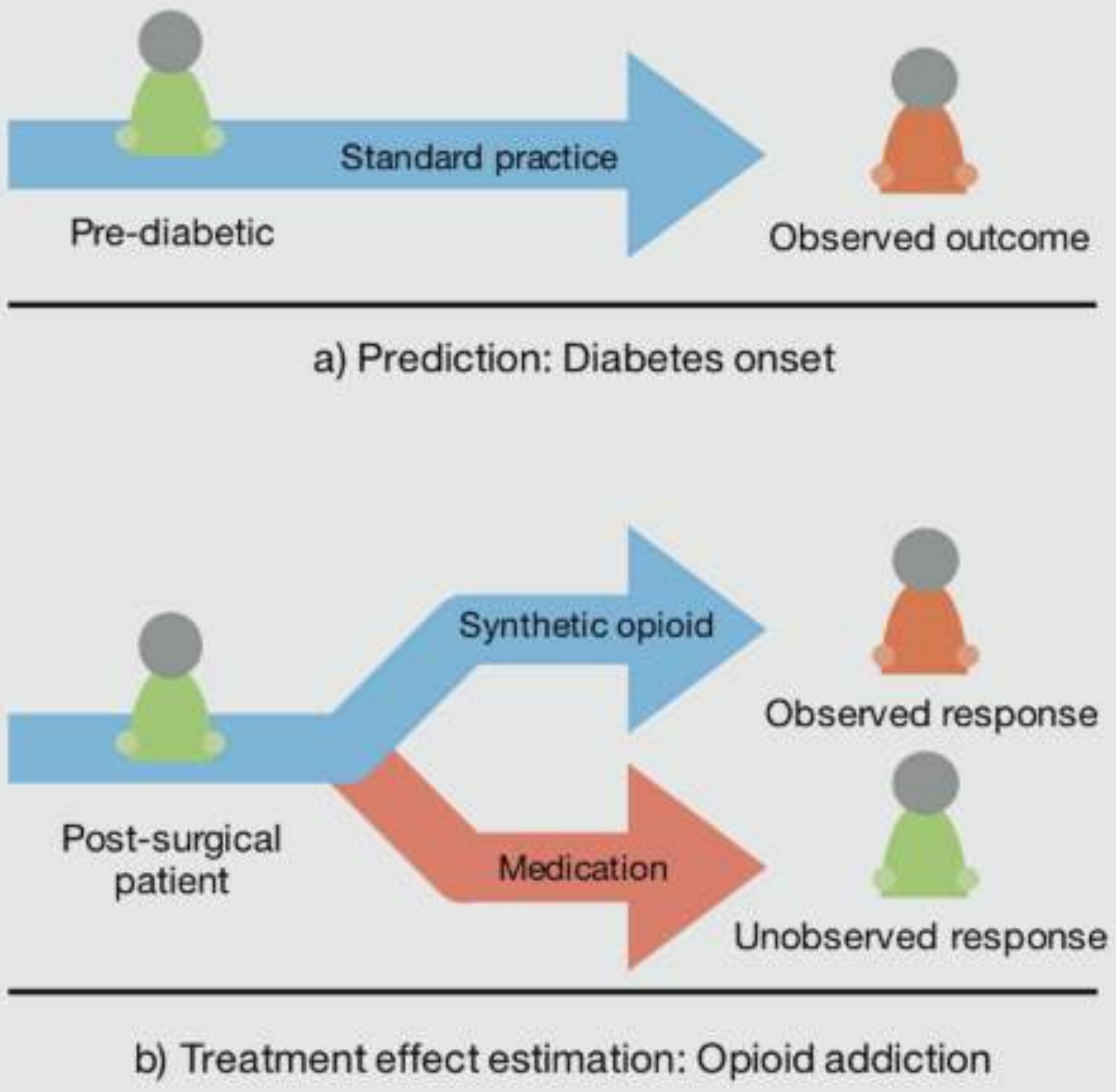
Prediction & decision making in healthcare

Prediction & decision making in healthcare

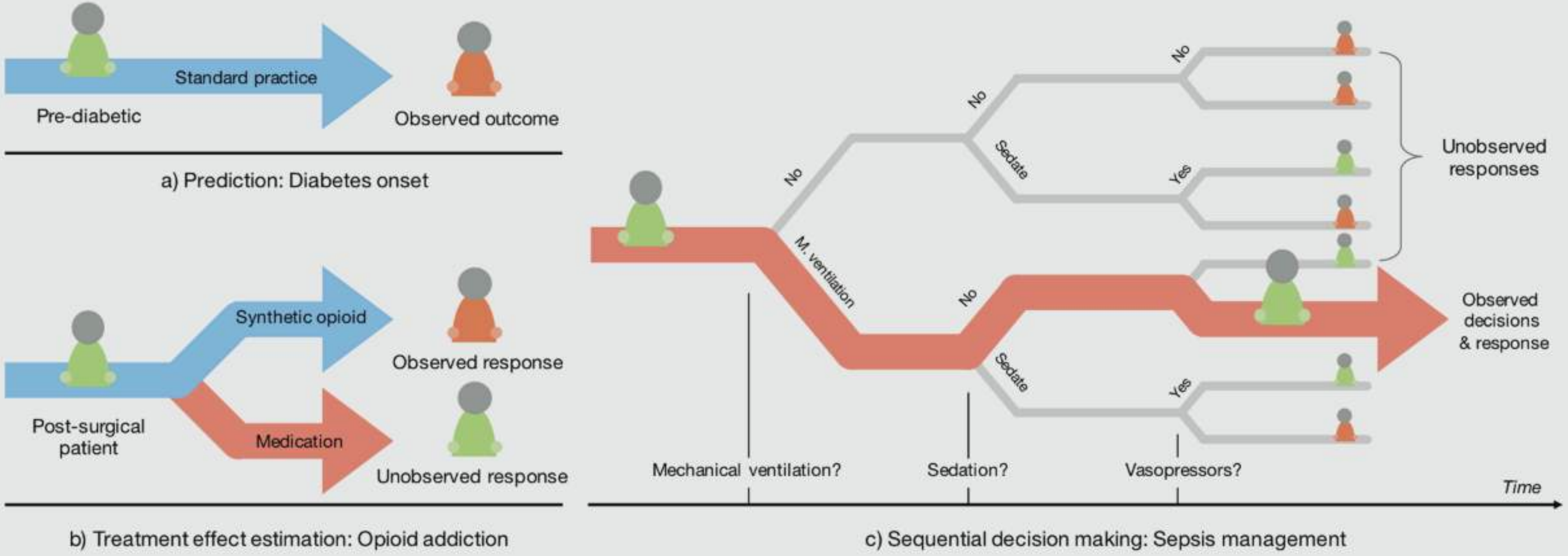


a) Prediction: Diabetes onset

Prediction & decision making in healthcare



Prediction & decision making in healthcare



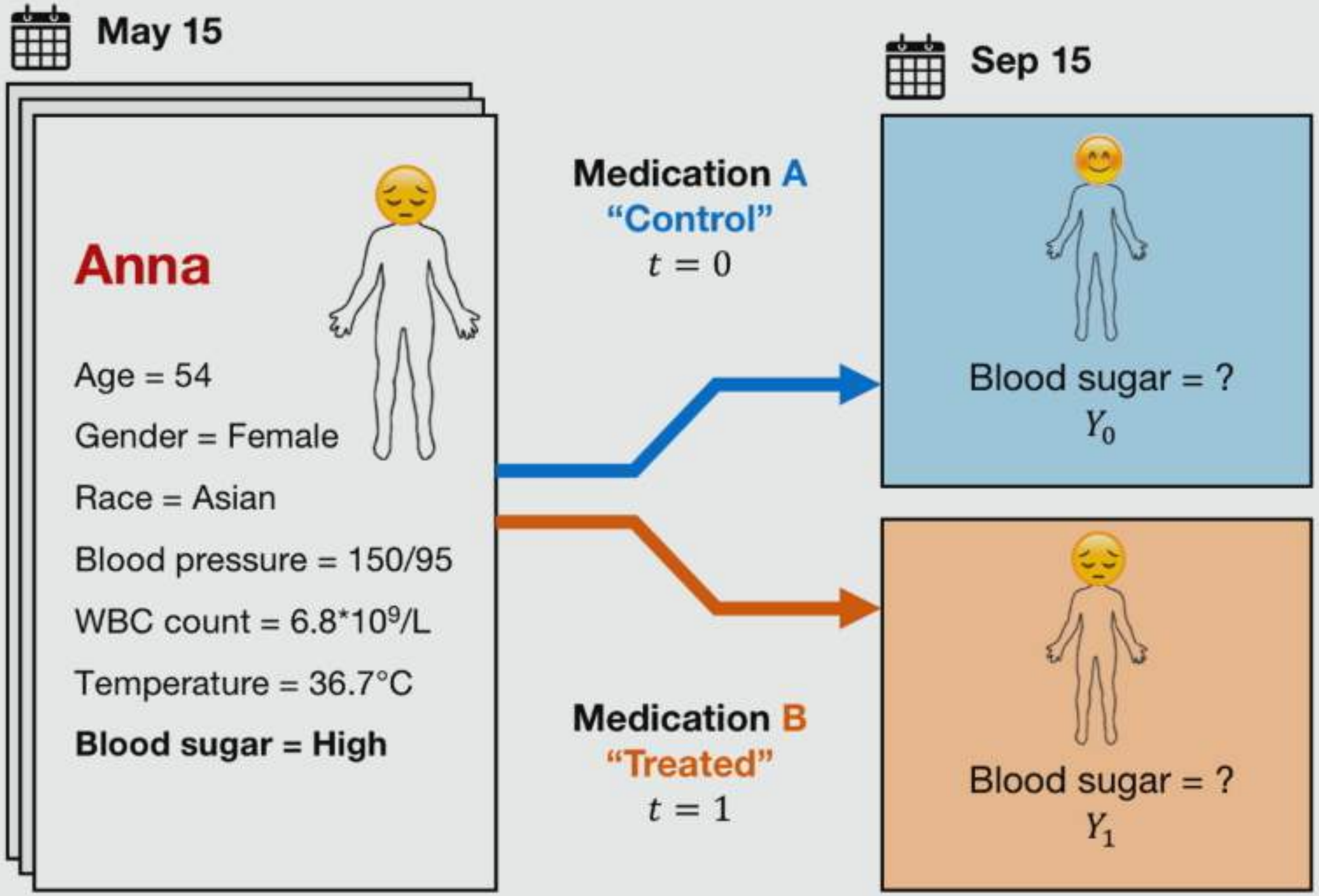
Predicting effects of decisions
requires **causal** reasoning

Outline

1. Estimating causal effects in **high-dimensional** data
2. What I learned from studying **domain adaptation**
3. How should we evaluate and new **policies**?

1. Estimating causal effects in **high-dimensional** data

Potential outcomes of medication



Often, we can perform an experiment
(e.g. randomized controlled trial).

Can we learn from historical **observational** data?

Observational datasets

- Observe medical records

Patient	Age	Blood pressure	Treatment	Blood sugar
Anna	54	150/95	A	High
Calvin	52	140/80	A	Low
John	48	135/70	B	Low
Peter	60	150/80	B	High

Observational datasets

- Unobserved **counterfactual** outcomes

Patient	Age	Blood pressure	Blood sugar **(A)**	Blood sugar **(B)**
Anna	54	150/95	High	**?**
Calvin	52	140/80	Low	**?**
John	48	135/70	**?**	Low
Peter	60	150/80	**?**	High

Observational datasets

Missing **not at random!**

- Unobserved **counterfactual** outcomes

Patient	Age	Blood pressure	Blood sugar **(A)**	Blood sugar **(B)**
Anna	54	150/95	High	**?**
Calvin	52	140/80	Low	**?**
John	48	135/70	**?**	Low
Peter	60	150/80	**?**	High

Predicting outcomes of interventions

$X \in \mathbb{R}^k$ — **Covariate** representation of units in k dimensions

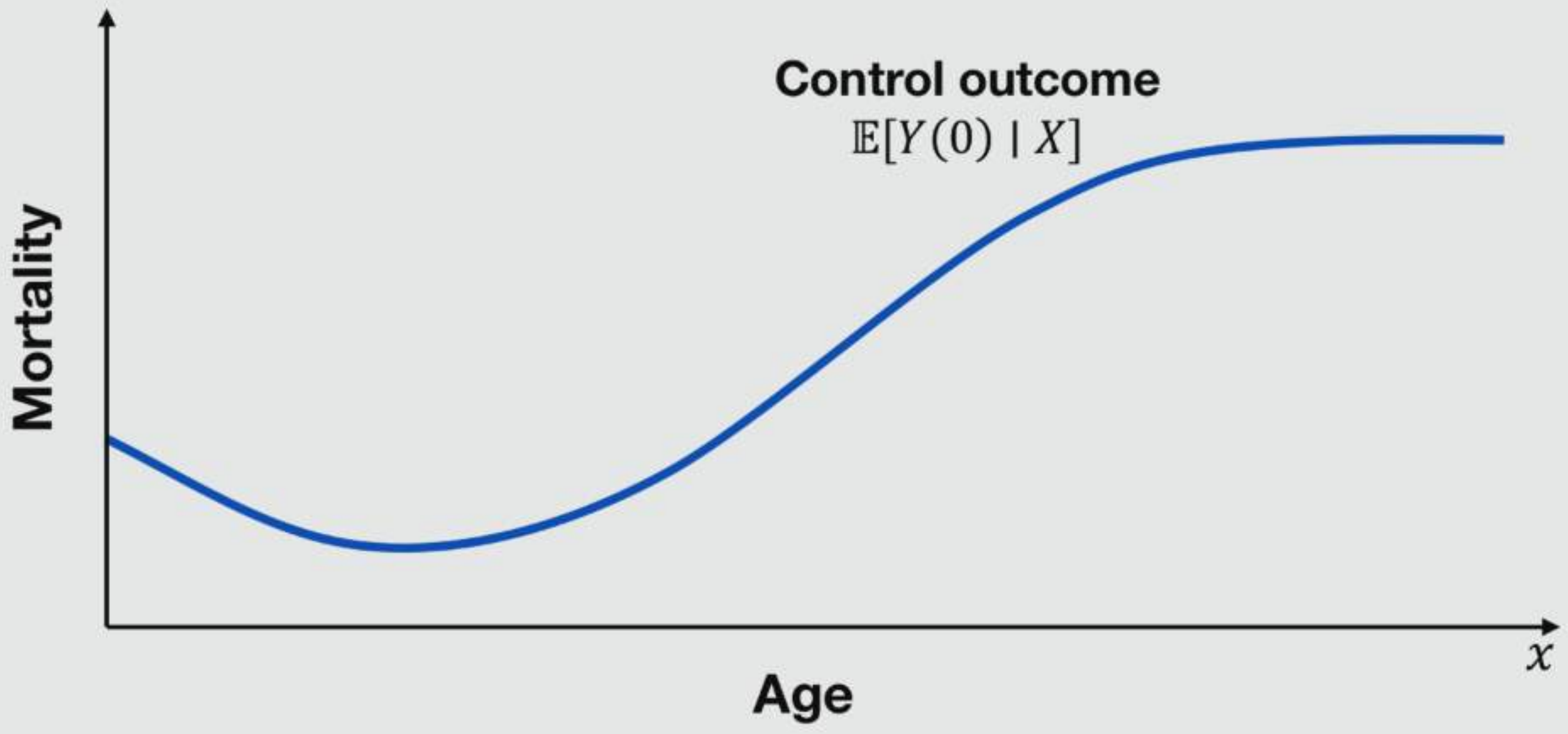
$T \in \{0, 1\}$ — **Treatment** assignments

$Y(0), Y(1)$ — **Potential outcomes** under $T = 0, 1$, respectively

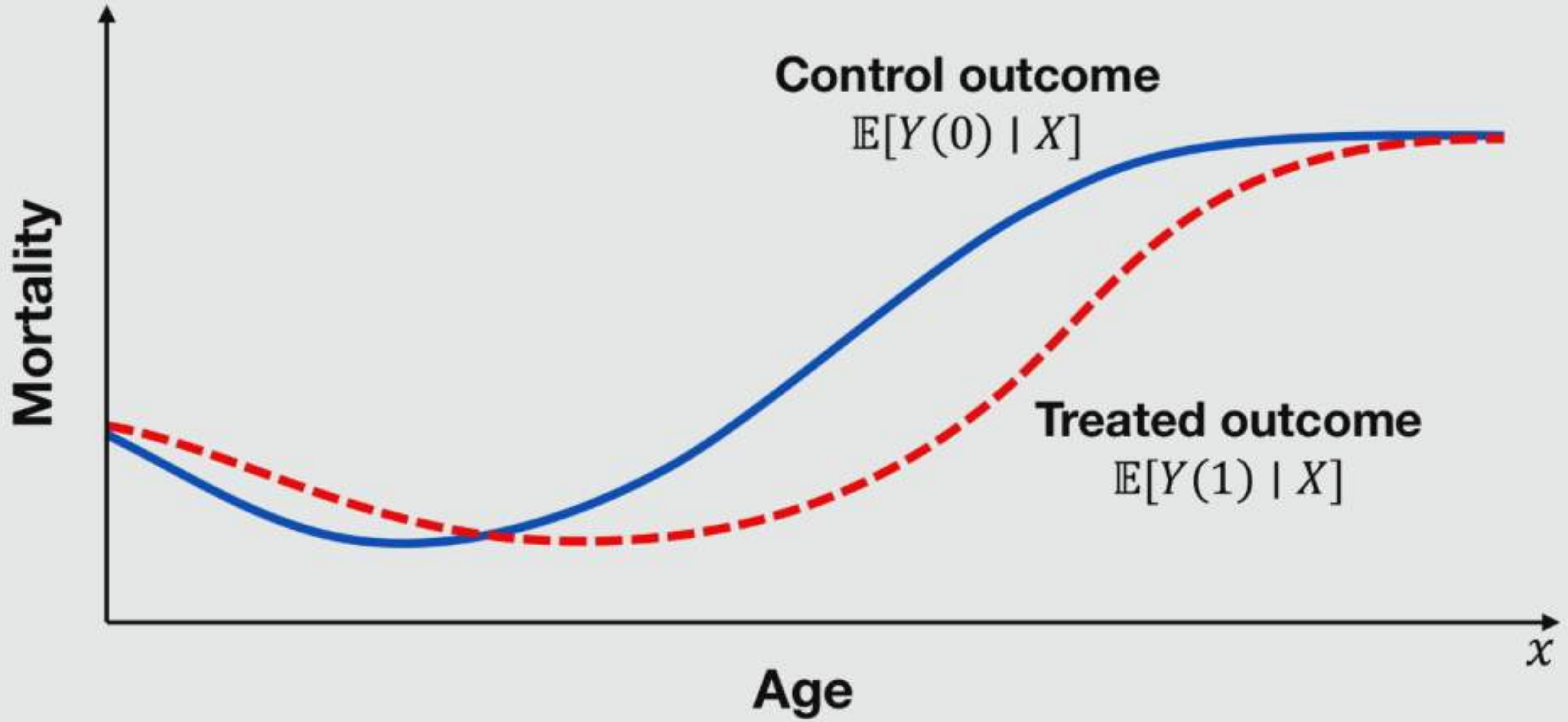
- **Goal:** Estimate counterfactual/potential outcome: $\mathbb{E}[\,Y(t) \mid X = x\,]$
- Conditional Average Treatment Effect (CATE)

$$\tau(x) = \mathbb{E}[\,Y(1) - Y(0) \mid X = x\,]$$

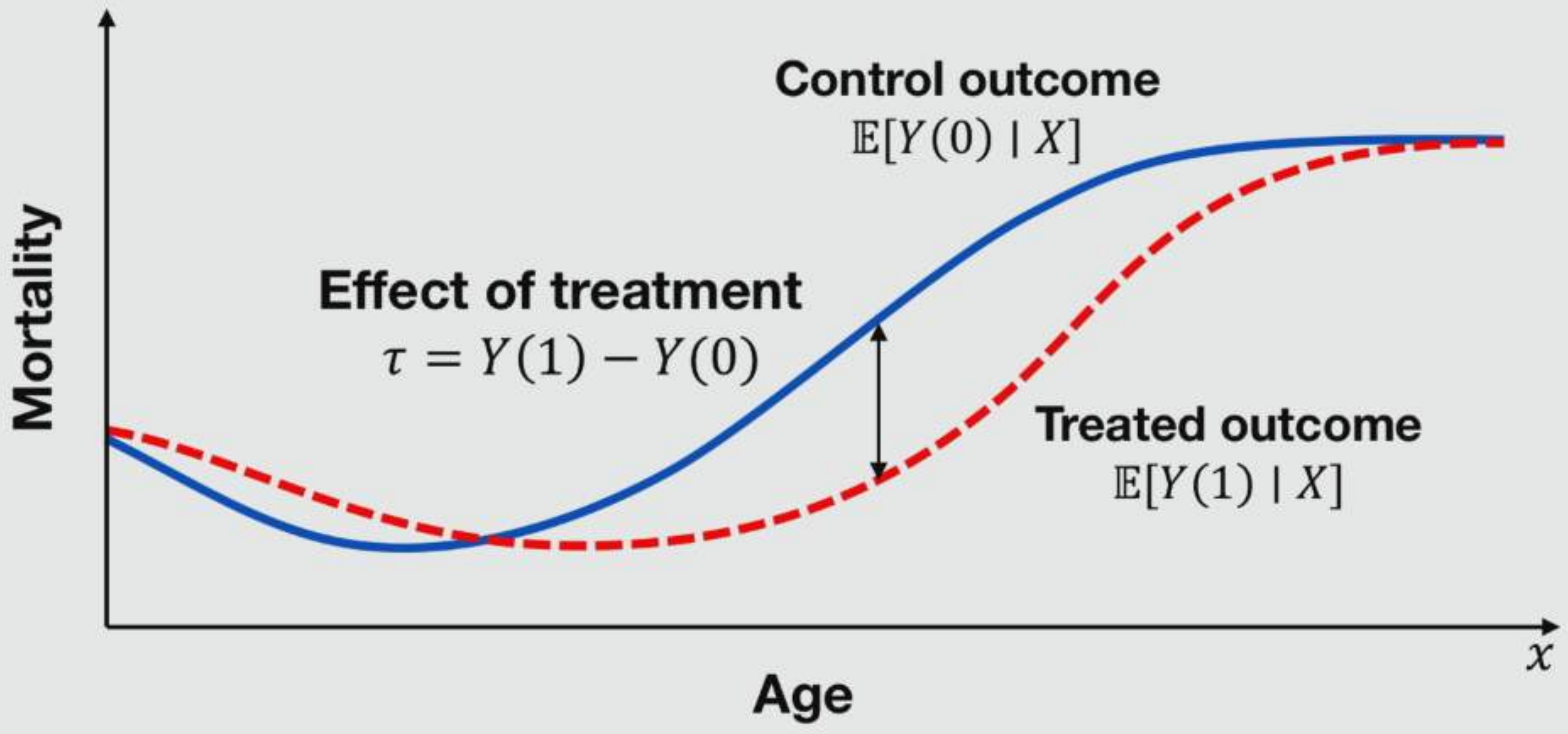
Potential outcomes and CATE



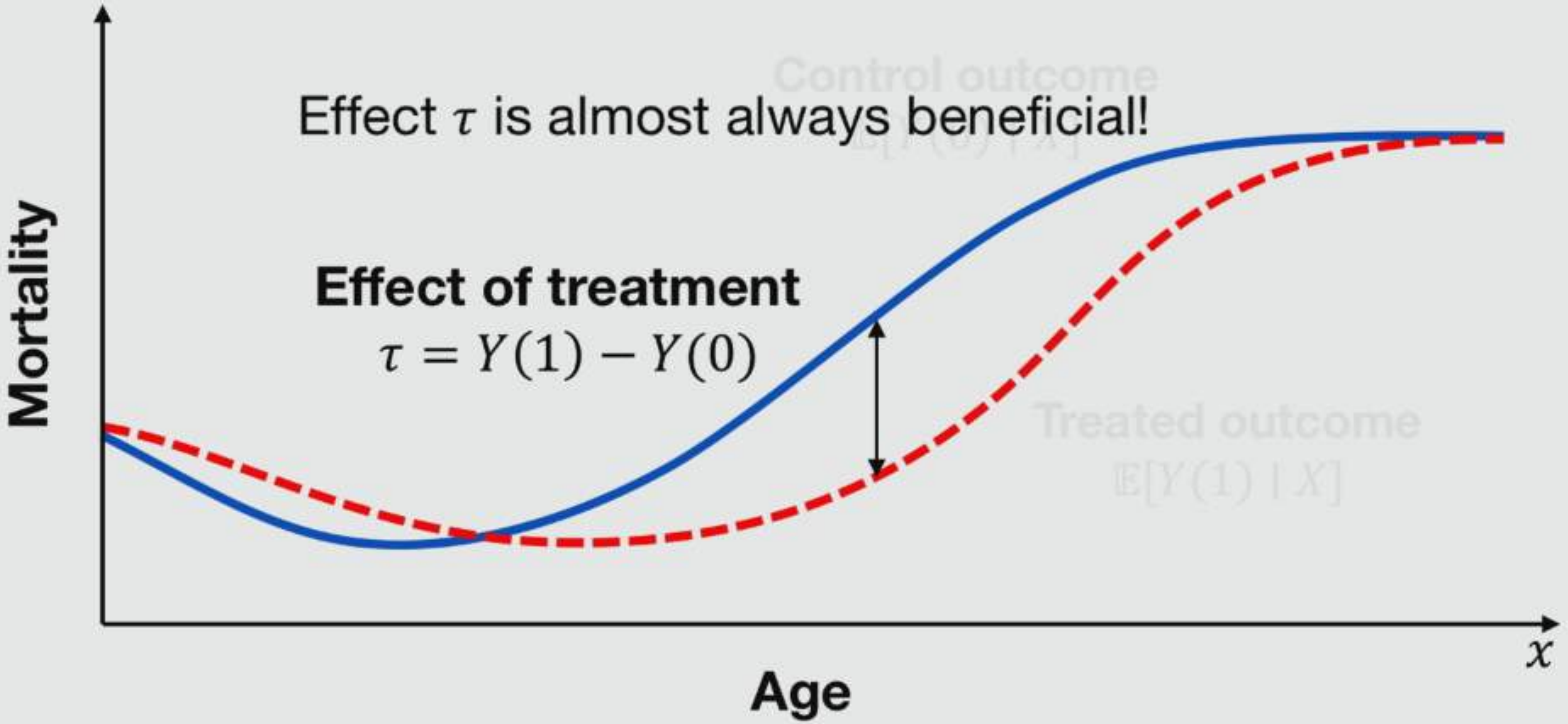
Potential outcomes and CATE



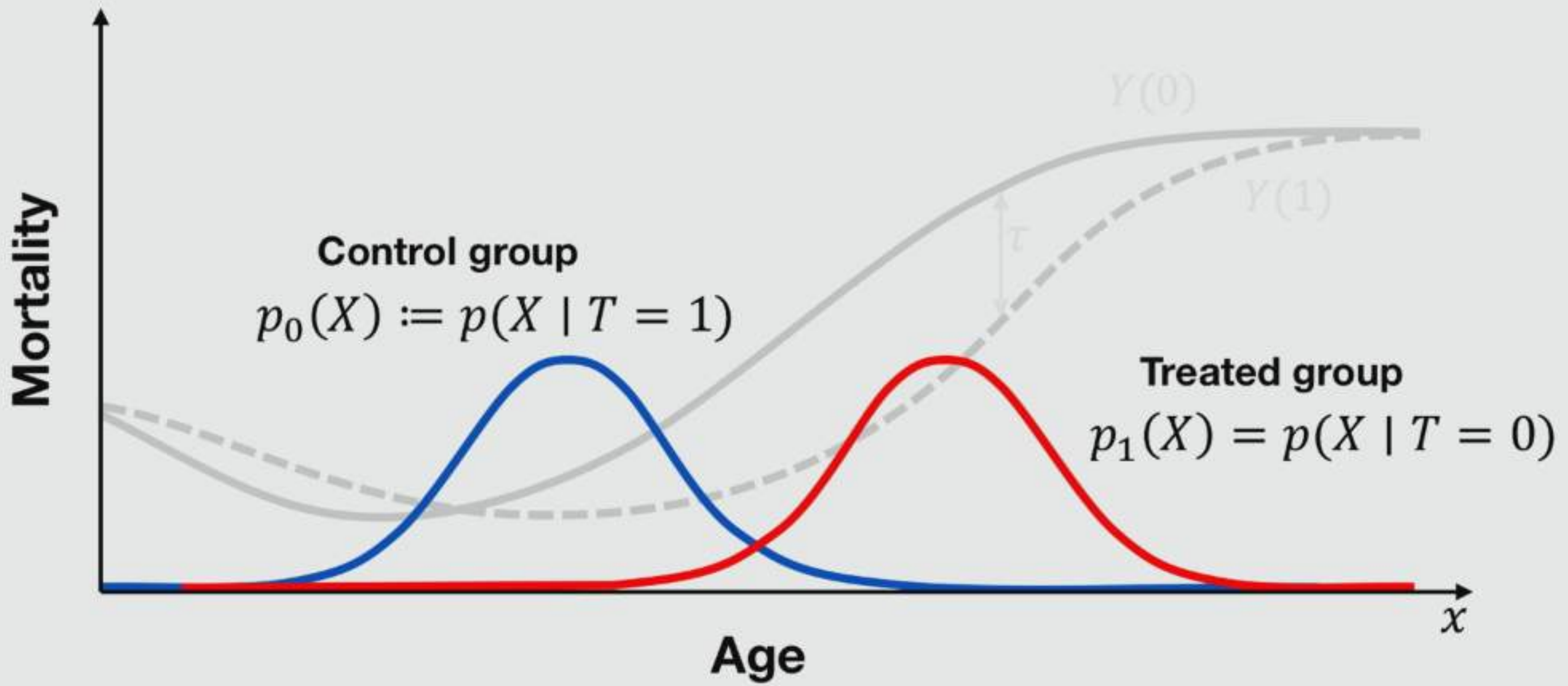
Potential outcomes and CATE



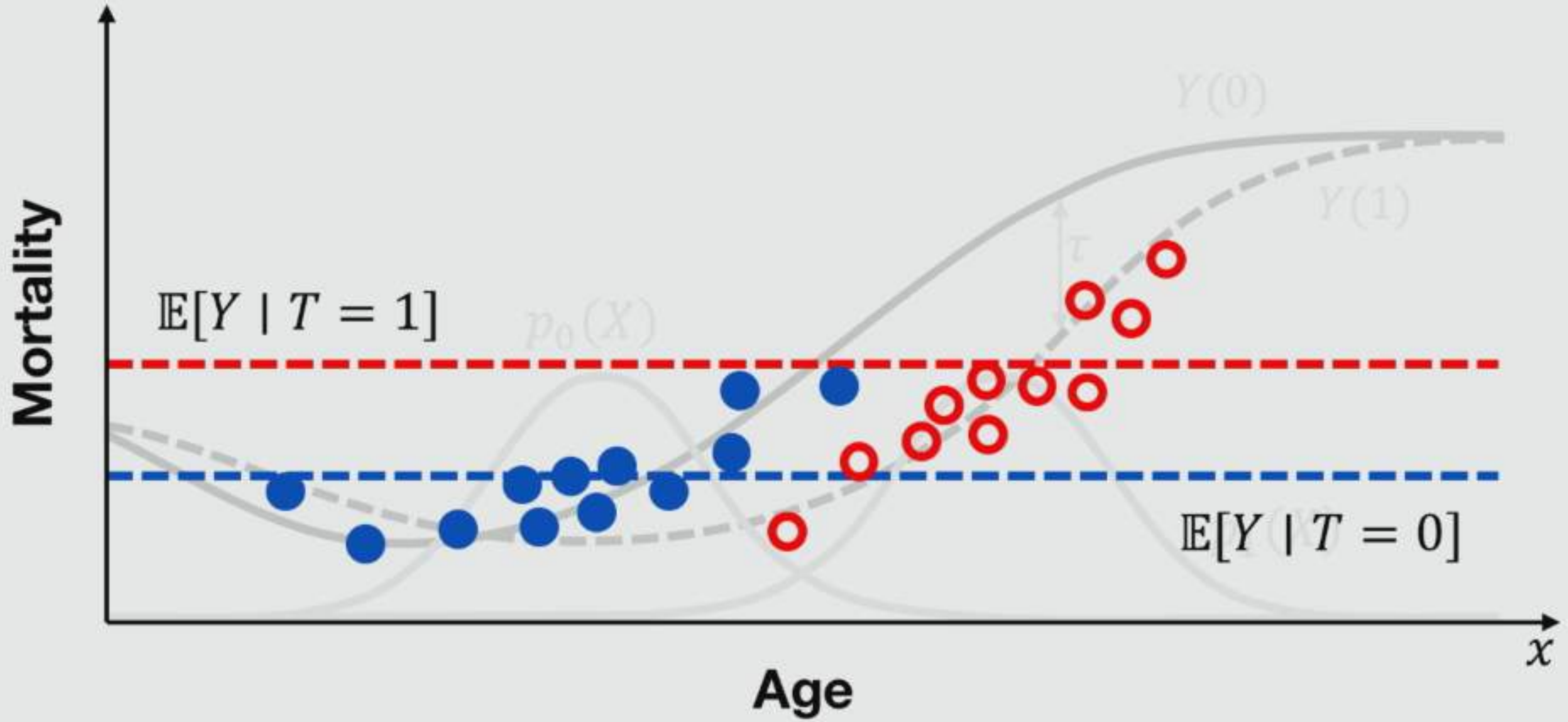
Potential outcomes and CATE



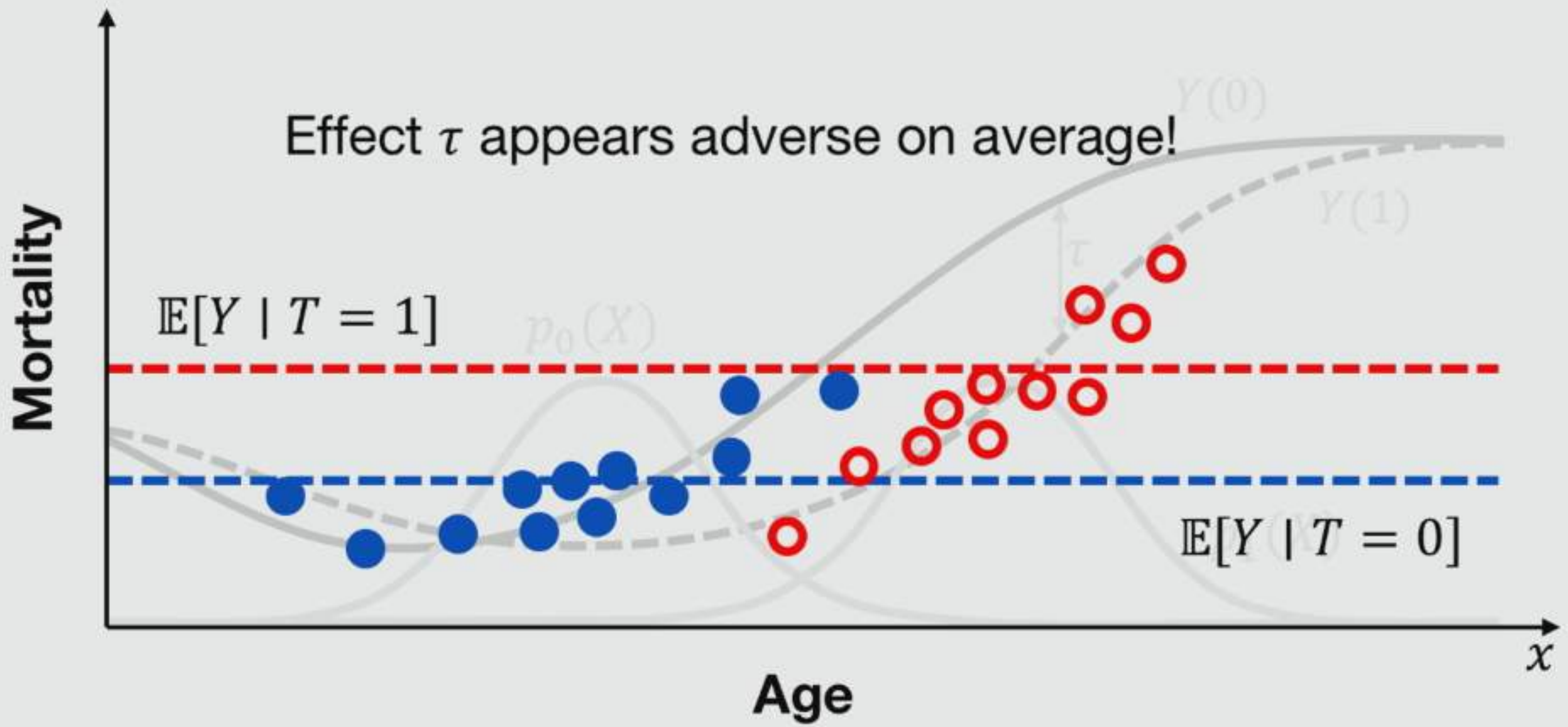
Treatment groups



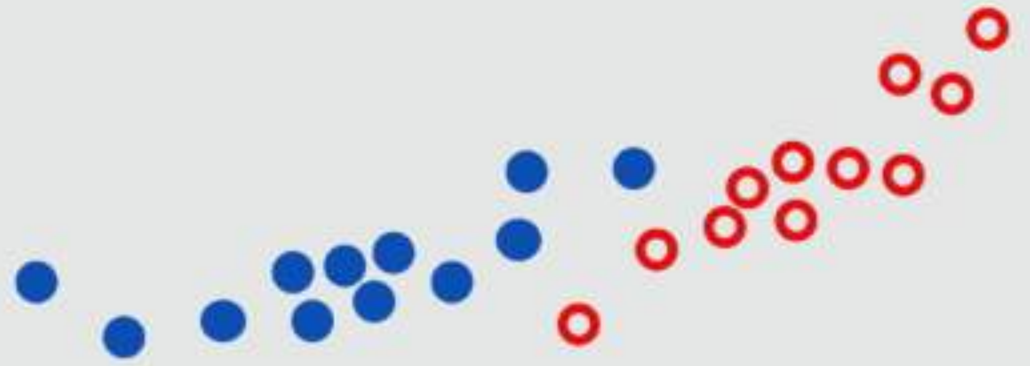
Treatment effect is confounded by age



Treatment effect is confounded by age



We have several problems



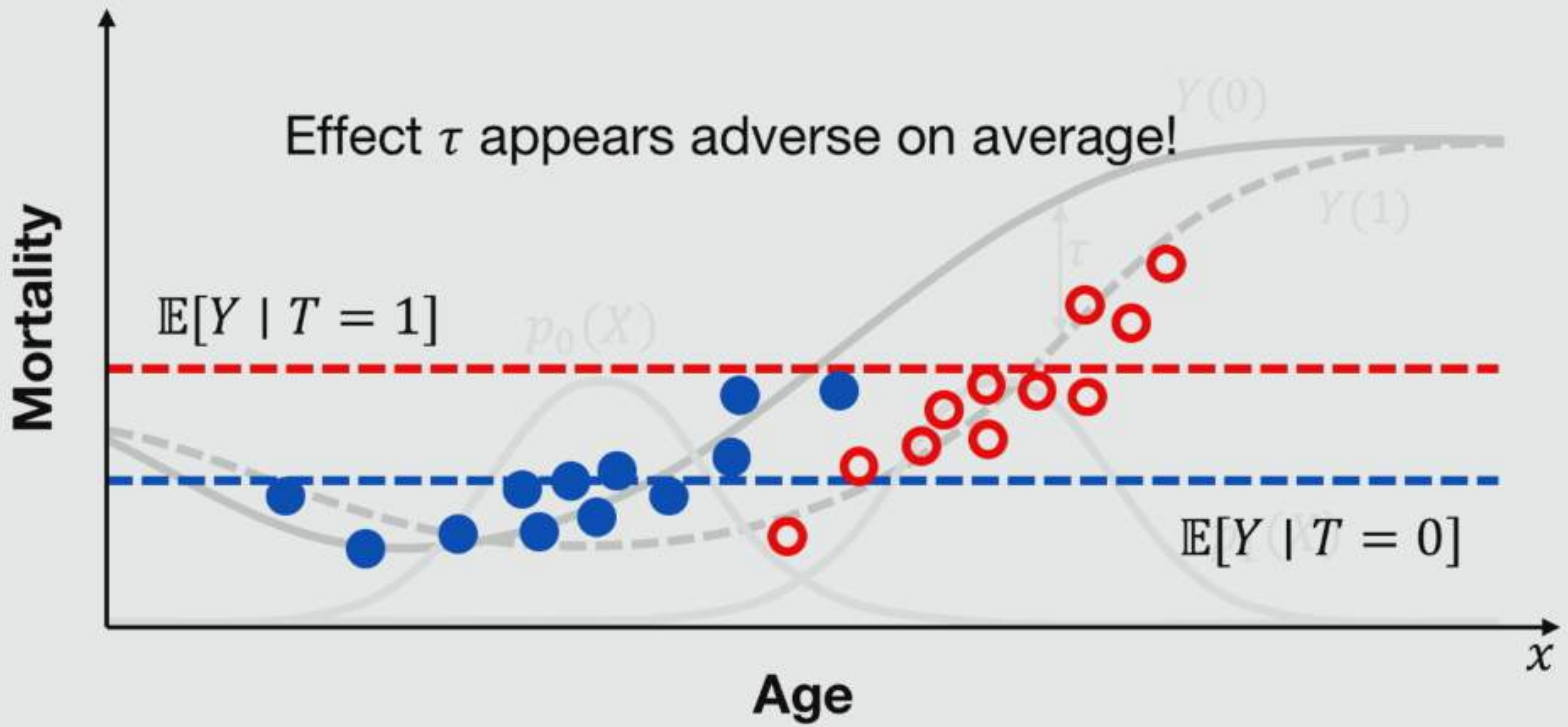
1. **Confounding:**

   Both the treatment groups and treatment effect vary with age. Naïve estimates are wrong

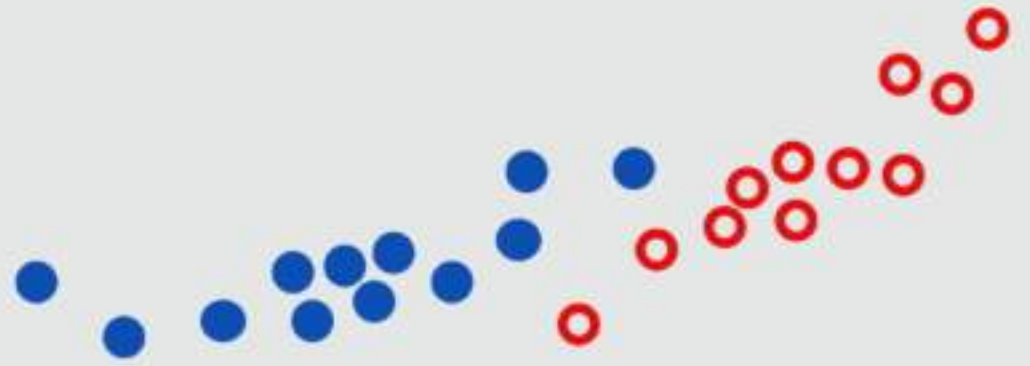
2. **Overlap:**

   We know very little about older patients off treatment

Treatment effect is confounded by age



We have several problems



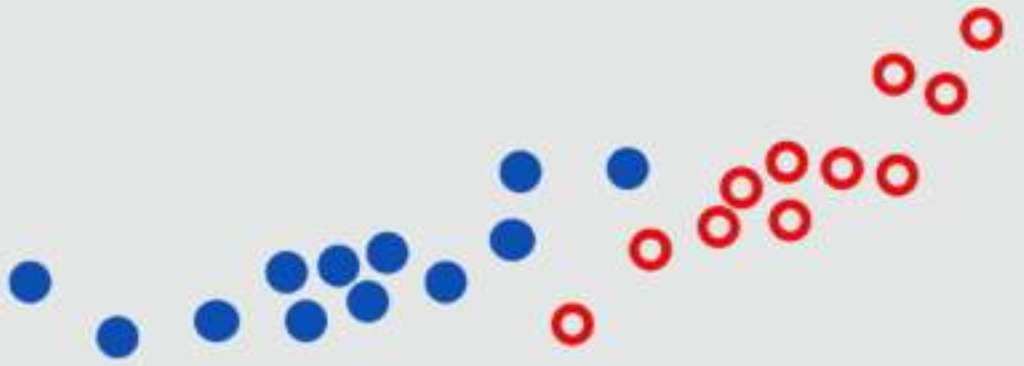
1. **Confounding:**

   Both the treatment groups and treatment effect vary with age. Naïve estimates are wrong

2. **Overlap:**

   We know very little about older patients off treatment

Identifying assumptions



- **Ignorability**

  $Y(0), Y(1) \perp T \mid X$

  If we control for X, we can estimate τ

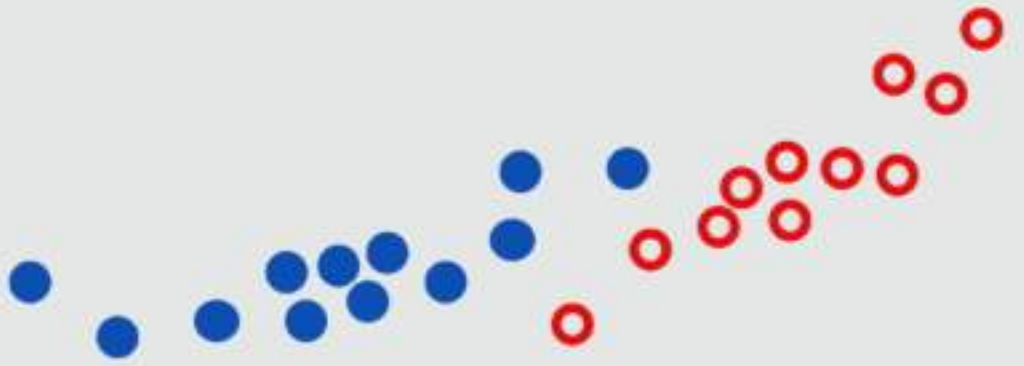
- **Common support**

  $\forall x: \; 0 < p(T = 1 \mid X = x) < 1$

  Treatment groups overlap everywhere

Essentially: Assume we don't have the problems I mentioned…

Identifying assumptions



- **Ignorability**

  $Y(0), Y(1) \perp T \mid X$ — If we control for X, we can estimate τ

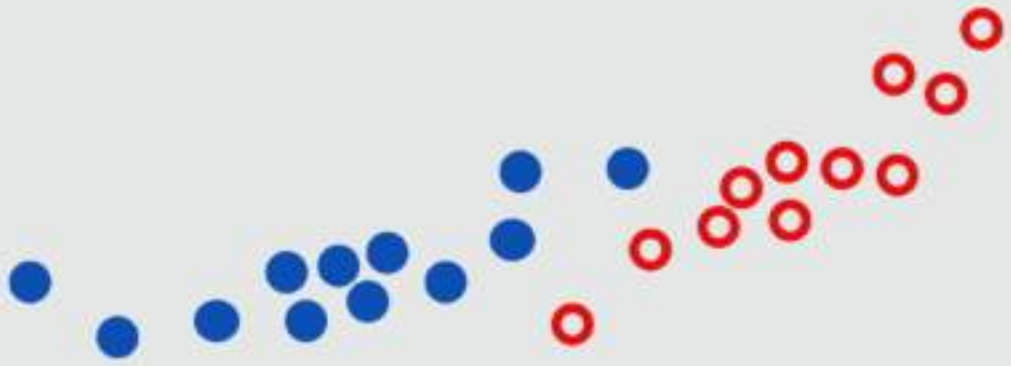
- **Common support**

  $\forall x: \; 0 < p(T = 1 \mid X = x) < 1$ — Treatment groups overlap everywhere

- Consistency

  $Y = TY(1) + (1 - T)Y(0)$ — If we assign treatment, we observe treated

Identifying assumptions



- **Ignorability**

  $Y(0), Y(1) \perp T \mid X$

- **Common support**

  $\forall x: \; 0 < p(T = 1 \mid X = x) < 1$

Focus in this talk

- Consistency

  $Y = TY(1) + (1 - T)Y(0)$

The remaining problem—observed confounding

- We observe only **factual** outcomes
- Roughly speaking

$$\mathbb{E}[Y(1) \mid X = x, T = 1] \quad \text{and} \quad \mathbb{E}[Y(0) \mid X = x, T = 0]$$

- We need both outcomes for **everyone**

$$\mathbb{E}[Y(1) \mid X = x] \quad \text{and} \quad \mathbb{E}[Y(0) \mid X = x]!$$

- How do we get there?

Classical solutions

- **Regression**

  Fit functions to predict outcomes of interventions

- **Re-weighting**

  Adjust for treatment group bias

  by emphasizing representative samples

- **Matching**

  Impute counterfactual outcomes by pairing up similar subjects

All of these rely on overlap!

Regression estimators

- Under ignorability with respect to X,

$$\mathbb{E}[\,Y(t) \mid X, T = t\,] = \mathbb{E}[Y \mid X, T = t]$$

- Regression is often used to estimate outcomes under different treatments separately

- If treatment groups are very different, the estimation error based on factual outcomes is **not representative** of the error in the counterfactual

Two conflicting observations

- **Blessing*** of high dimensionality
  - Less likely to have left out confounding variables
- **Curse** of high dimensionality
  - Less overlap between treatment groups
  - High-variance estimates
  - More likely to introduce selection bias, M-bias etc

* Adjusting for more potential confounders does not always lead to less bias (see e.g. M-bias, Z-bias)

Mitigating the curse of dimensionality?

- Can we find a representation of our data $\Phi(X)$ such that

- Ignorability

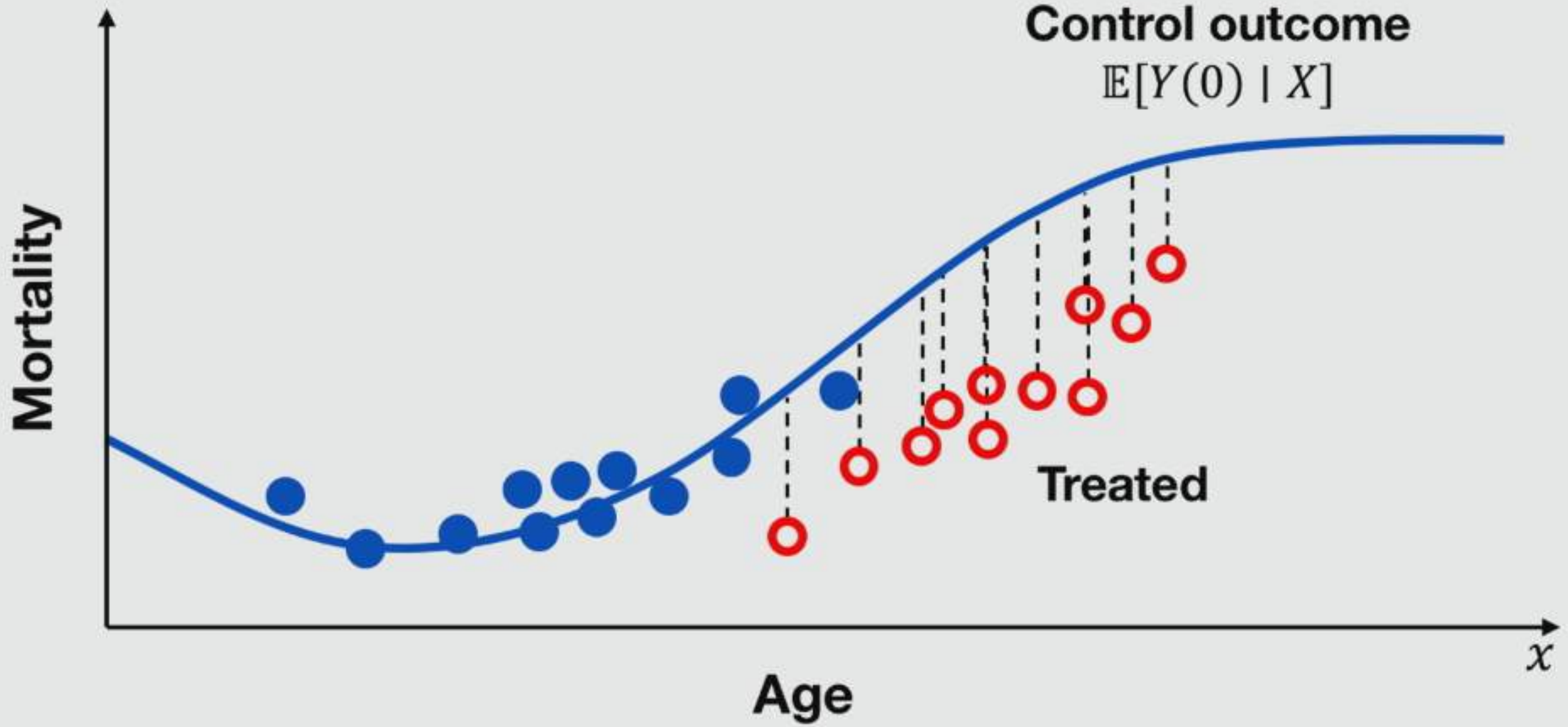
  $Y(0), Y(1) \perp T \mid \boldsymbol{\Phi}(\boldsymbol{X})$

- Common support

  $\forall z : \boldsymbol{\epsilon} < p(T = 1 \mid \boldsymbol{\Phi}(\boldsymbol{X}) = \boldsymbol{z}) < \mathbf{1} - \boldsymbol{\epsilon}$

2. What I learned from (and about) **domain adaptation**

Consider counterfactual for the treated



Counterfactual prediction & domain adaptation[1]

- Domain adaptation: Learn from **source** domain, predict in **target**

		Counterfactual prediction	Domain adaptation
Data	$(x, y) \sim p_0(X, Y(0))$	Factual control	Labeled source
	$x \sim p_1(X)$	Treated	Unlabeled target
Goal	$Y(0)$ for $x' \sim p_1(x)$	Counterfactual	Target labels
Assum.	$Y(0) \perp T \mid X$	Ignorability	Covariate shift

[1] J, Shalit, Sontag, *ICML,* 2016

Domain adaptation without overlap[1]

SVHN | Syn Signs

Source (Training)

Target (Test)

MNIST | GTSRB

Target accuracy: ~ 83% | Target accuracy: ~ 93%

[1]Ganin et al, *JMLR*, 2015

Risk minimization

(Machine learning view)

$$\hat{f} := \arg\min_{f \in \mathcal{H}} R(f) \approx Y$$

Risk minimization

- Find hypothesis f_0 that minimizes the **counterfactual risk** $R_1(f_0)$

The risk in predicting the control outcome for the treated

$$R_1(f_0) = \mathbb{E}\Big[\underbrace{\ell\big(f_0(X), Y(0)\big) \mid T = 1}_{\text{Unobserved}}\Big]$$

- for e.g. the squared loss, $\ell(y, y') = (y - y')^2$
- Use importance weights? $R_1(f_0) = R_0^{\boldsymbol{w}}(f_0) \approx \frac{1}{n}\sum_{i=1}^{n} \frac{p_1(x_i)}{p_0(x_i)} \ell(f_0(x_i), y_i)$

Risk minimization

- Find hypothesis f_0 that minimizes the **counterfactual risk** $R_1(f_0)$
  The risk in predicting the control outcome for the treated

$$R_1(f_0) = \mathbb{E}\left[\, \ell\big(f_0(X), Y(0)\big) \mid T = 1 \,\right]$$

- for some loss function ℓ such as the squared loss, $\ell(y, y') = (y - y')^2$

No overlap in high dimensions!

We can't do importance weighting!

Domain adaptation bounds

- Take inspiration from domain adaptation[1,2]—bound the risk!

- **Under ignorability** w.r.t. X, the following bound holds for any f_0

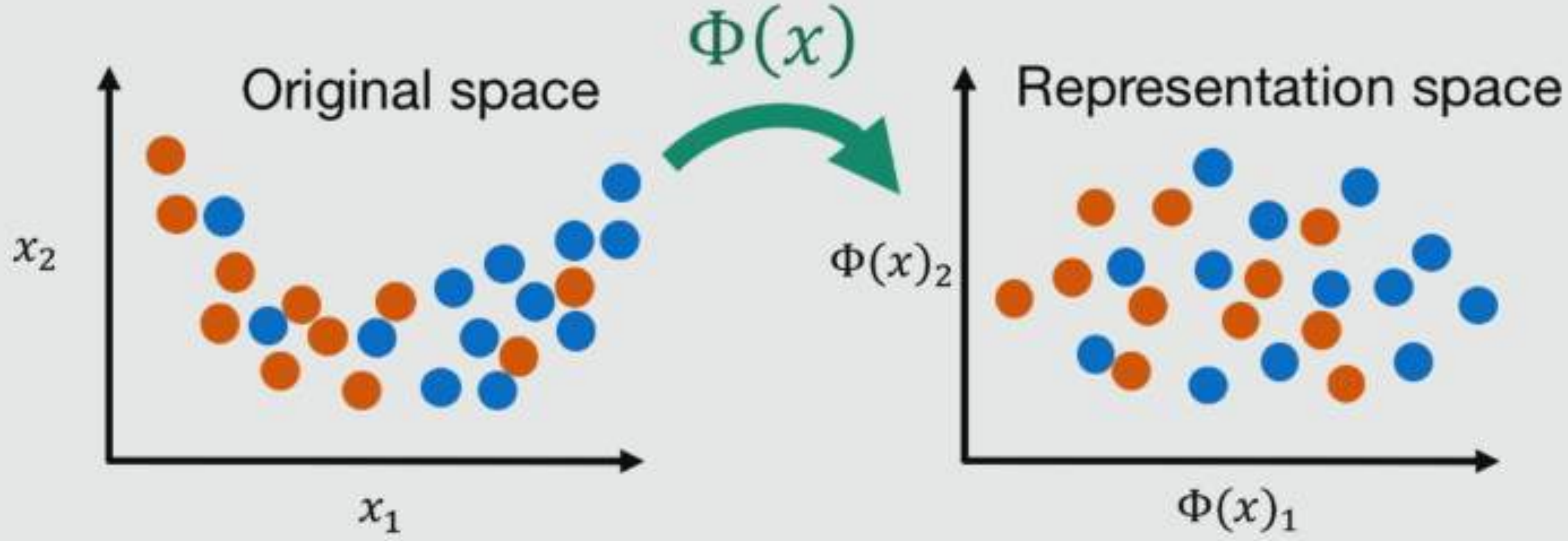
$$\boxed{\underbrace{R_1(f_0)}_{\text{Counterfactual risk}} \leq \underbrace{R_0(f_0)}_{\text{Factual risk}} + \underbrace{d_{\mathcal{H}}(p_0(X), p_1(X))}_{\text{Distributional distance w.r.t. } X}}$$

- The distance $d_{\mathcal{H}}(p, q) \coloneqq \sup_{g \in \mathcal{H}} \left| \mathbb{E}_p[g] - \mathbb{E}_q[g] \right|$ such that $\ell \in \mathcal{H}$

[1]Ben-David et al., 2008, [2]J., Shalit, Sontag, *ICML* 2016

Learn representations to minimize $d(p_0, p_1)$

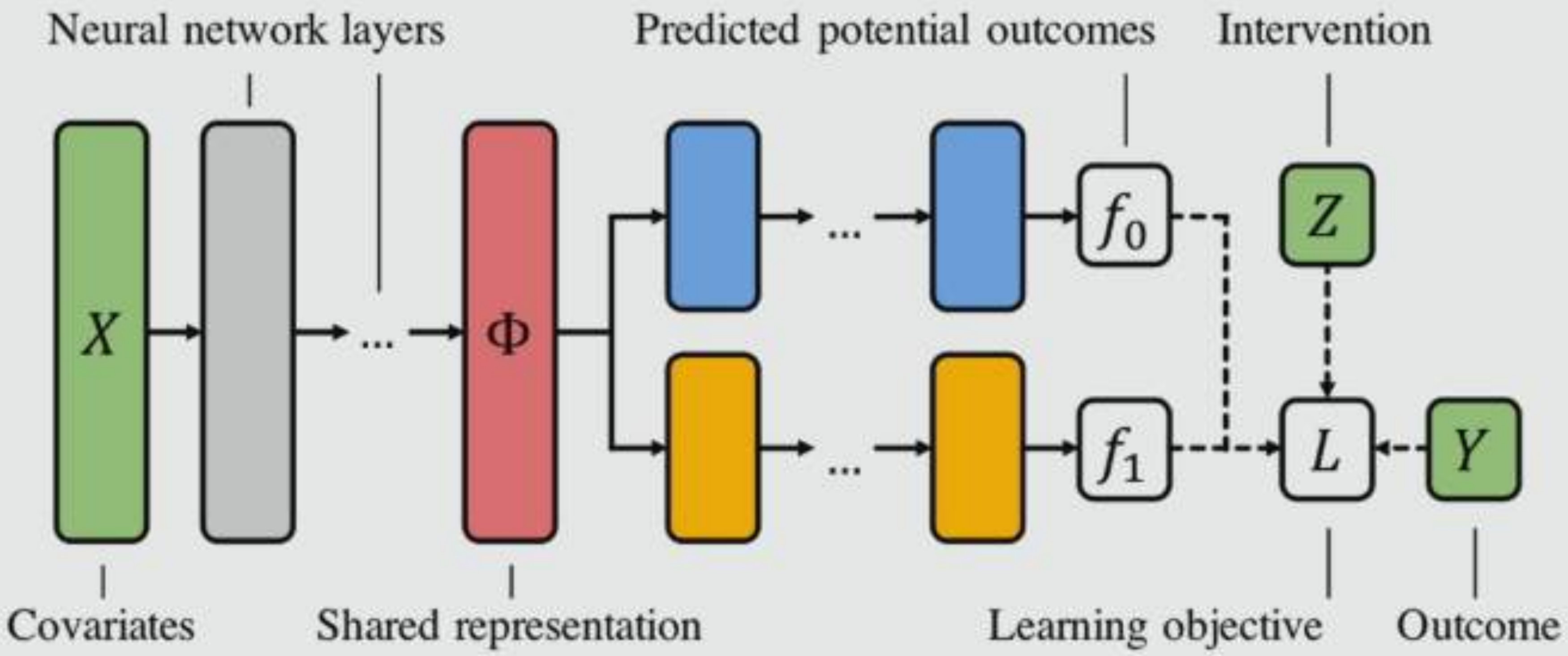
- **Approach 1:** Find a new, predictive space which exposes similarities
- $\text{minimize}_{f,\Phi}\ \ R_0(f_0) + d_{\mathcal{H}}\left(p_0(\Phi(X)), p_1(\Phi(x))\right)$



[2] Shalit, J., Sontag, *ICML* 2017

Learn representations to minimize $d(p_0, p_1)$

- **Approach 1:** Find a new, predictive space which exposes similarities
- $\text{minimize}_{f,\Phi} \;\; R_0(f_0) \; + \; d_{\mathcal{H}}\Big(p_0\big(\Phi(X)\big), p_1\big(\Phi(x)\big)\Big)$



---

[2]Shalit, J., Sontag, *ICML* 2017

Learn representations to minimize $d(p_0, p_1)$

- Worked well in practice
- Results on the **IHDP** benchmark
- Semi-synthetic dataset

		Error in conditional effect	Error in average effect
		IHDP	
		$\sqrt{\epsilon_{\text{CATE}}}$	ϵ_{ATE}
	OLS/LR$_1$	$5.8 \pm .3$	$.94 \pm .06$
	OLS/LR$_2$	$2.5 \pm .1$	$.31 \pm .02$
	BLR	$5.8 \pm .3$	$.93 \pm .05$
	k-NN	$4.1 \pm .2$	$.79 \pm .05$
	TMLE	†	†
	BART	$2.3 \pm .1$	$.34 \pm .02$
	R.For.	$6.6 \pm .3$	$.96 \pm .06$
	C.For.	$3.8 \pm .2$	$.40 \pm .03$
Concatenating Φ and T	BNN	$2.1 \pm .1$	$.42 \pm .03$
Twin-head neural net ($\alpha = 0$)	TARNet	$.95 \pm .02$	$.28 \pm .01$
+ IPM regularization	CFR$_{\text{MMD}}$	$.78 \pm .02$	$.31 \pm .01$
	CFR$_{\text{WASS}}$	$.76 \pm .02$	$.27 \pm .01$
+ Learned re-weighting	RCFR	$.67 \pm .05$	—

What can go wrong?

- The objective is **not an upper bound** for all Φ!*

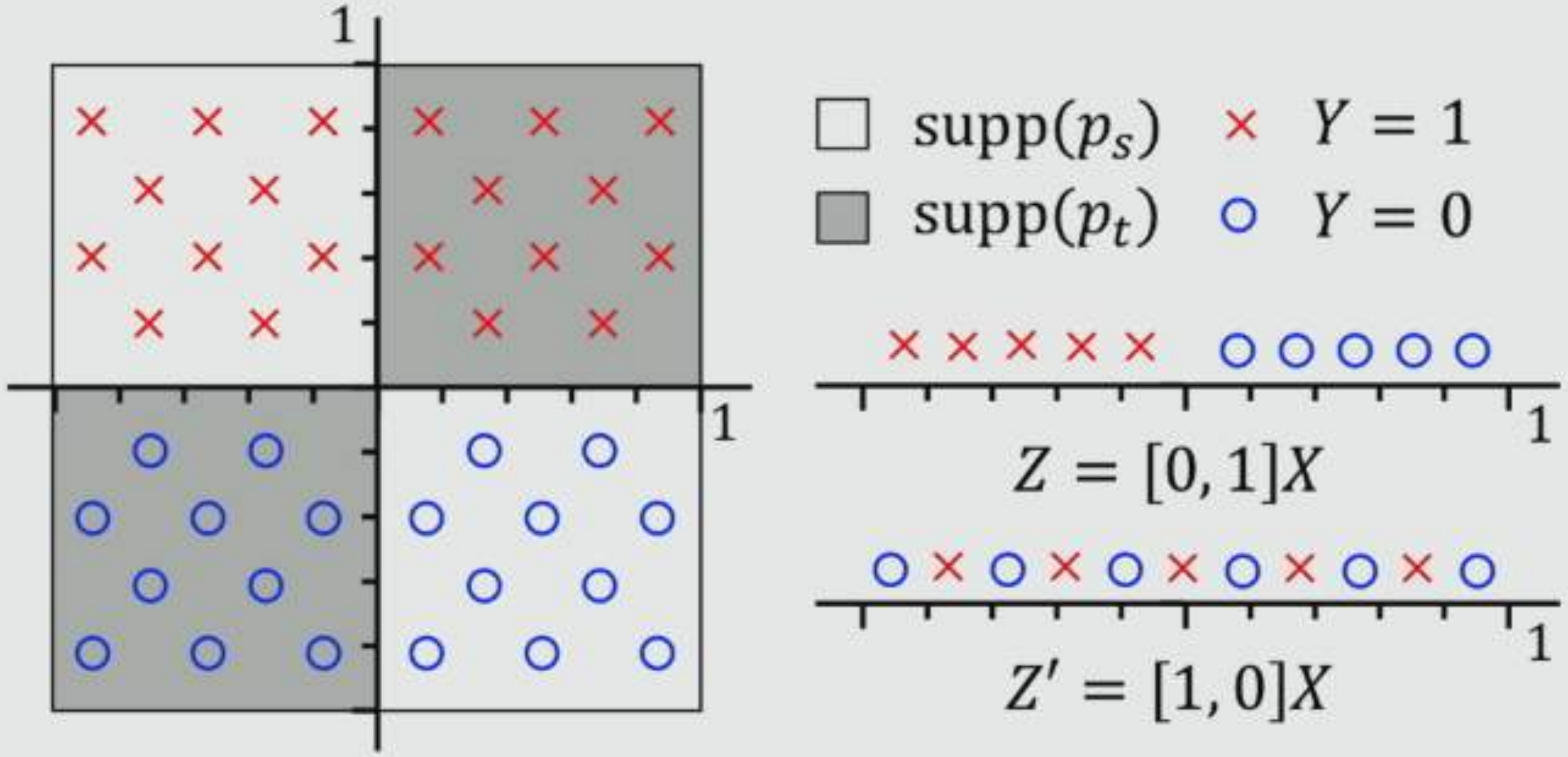
$$\exists\Phi\colon\ R_1\big(f_0(\Phi(X))\big) \nleq R_0(f_0(\Phi(X))) + d_{\mathcal{H}}\Big(p_0\big(\Phi(X)\big), p_1\big(\Phi(X)\big)\Big)$$

- A) **Ignorability need not hold** w.r.t. $Z = \Phi(\mathrm{X})$!

- B) We may lose information that is more useful for the counterfactual than the factual!

*In Shalit, J., Sontag, *ICML* 2017, we assume that Φ is invertible

Failure case: variable selection



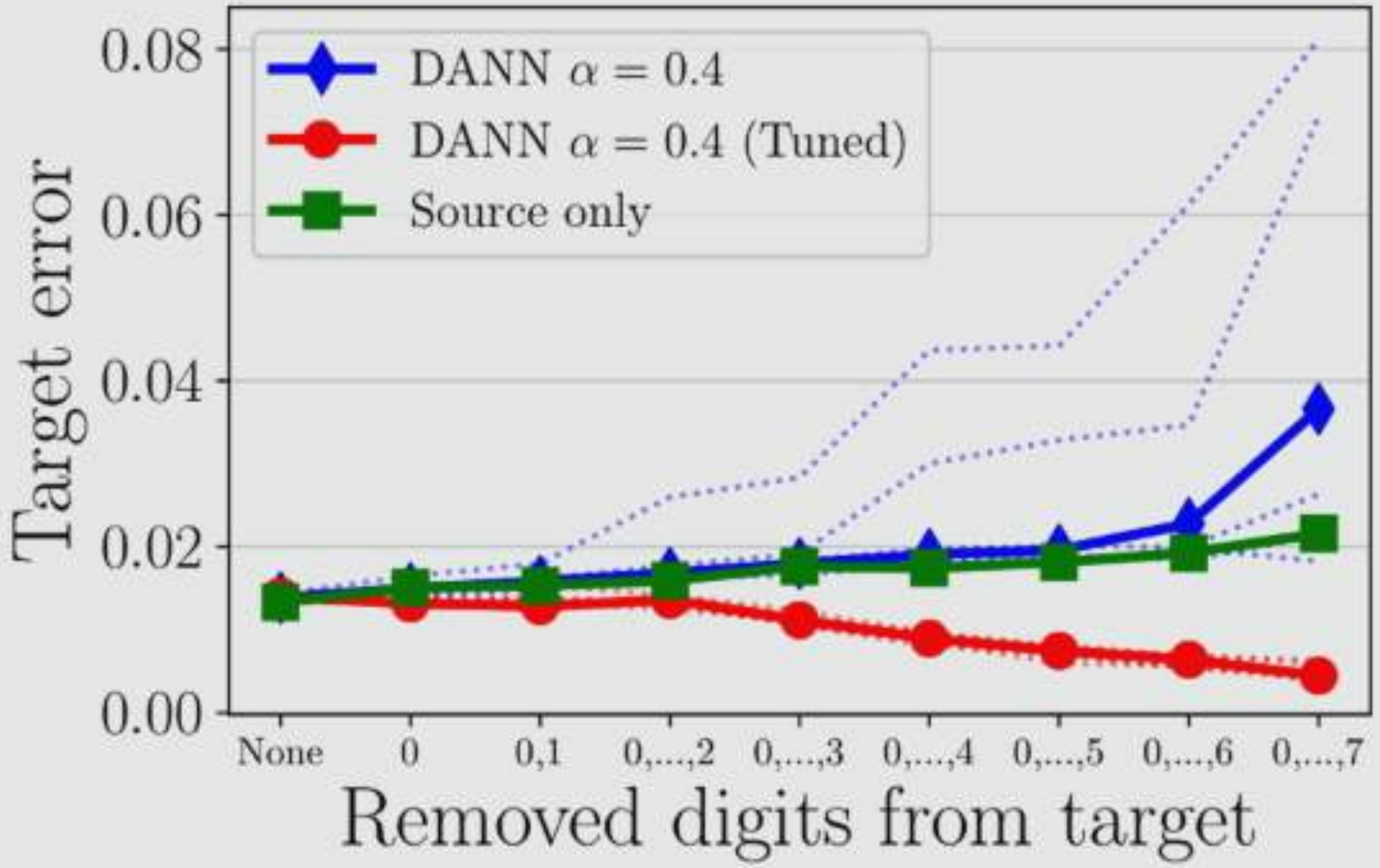
- Projection onto x_1 admits $R_0(f_0(x_1)) + d_{\mathcal{H}}(p_0(x_1), p_1(x_1)) = 0$
- But ignorability does not hold: $Y(0) \not\perp T \mid X_1$

Failure case: variable selection

- Consider predicting the **effect of a drug T** vs no treatment
- Now, assume that **T induces an allergic reaction** in some patients
- The allergy indicator **will not be predictive** of the treated outcome, as treated allergic patients **will be rare** in data (if this is known)
- Selecting variables based on overlap and prediction will remove the allergy indicator!

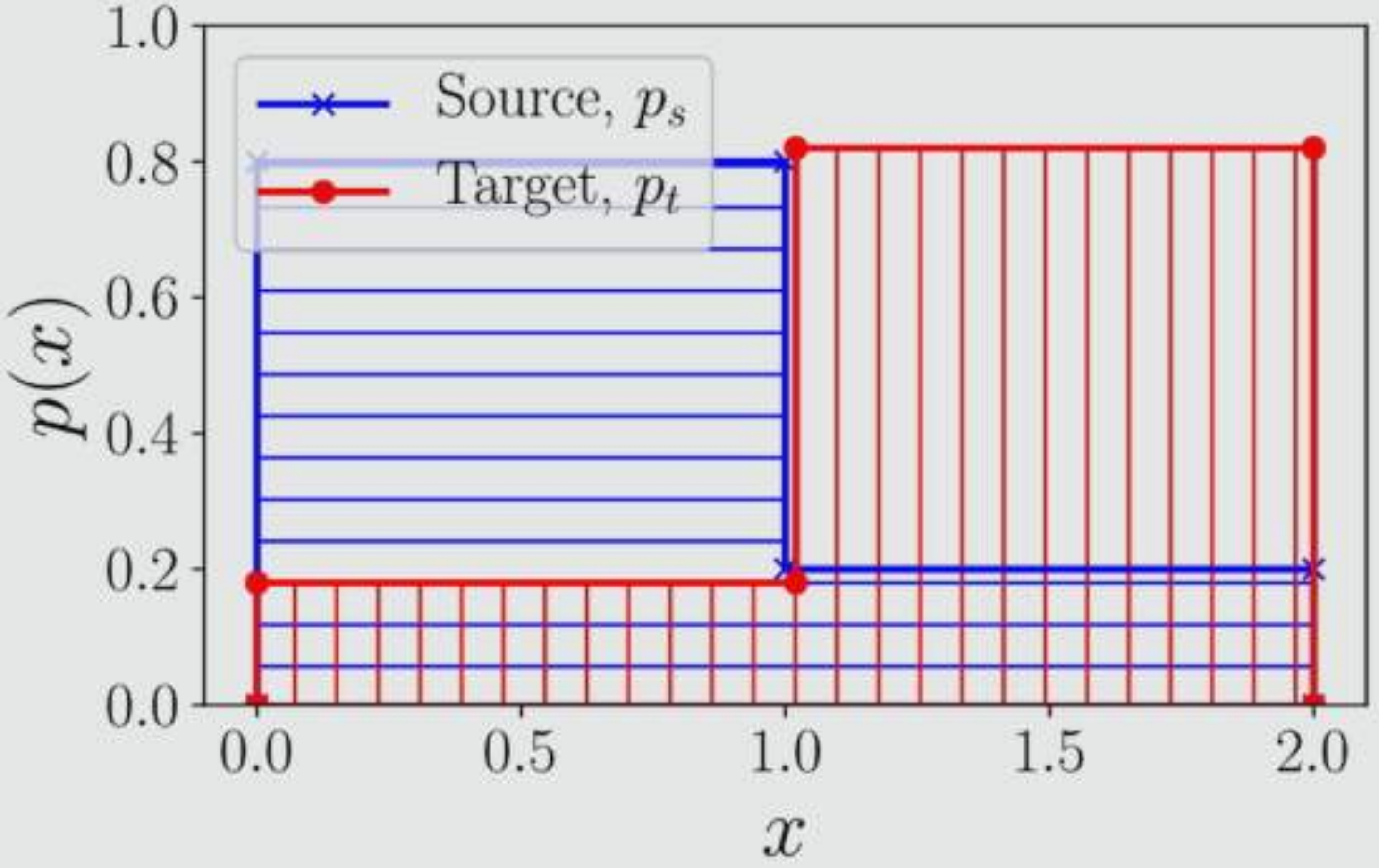
Distance metrics matter

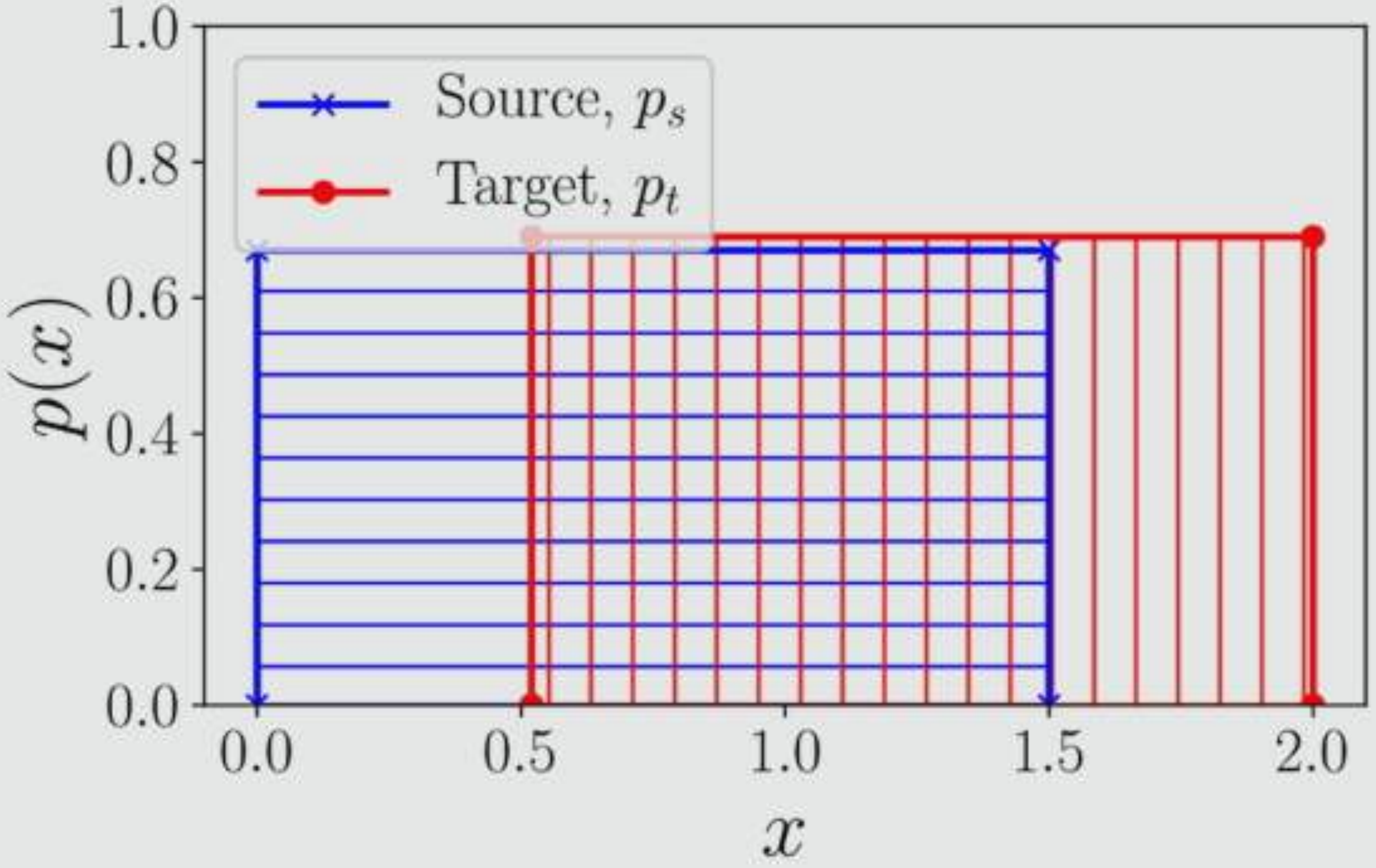
- Source: **MNIST**, Target: **MNIST (with digits removed)**



What is going on?

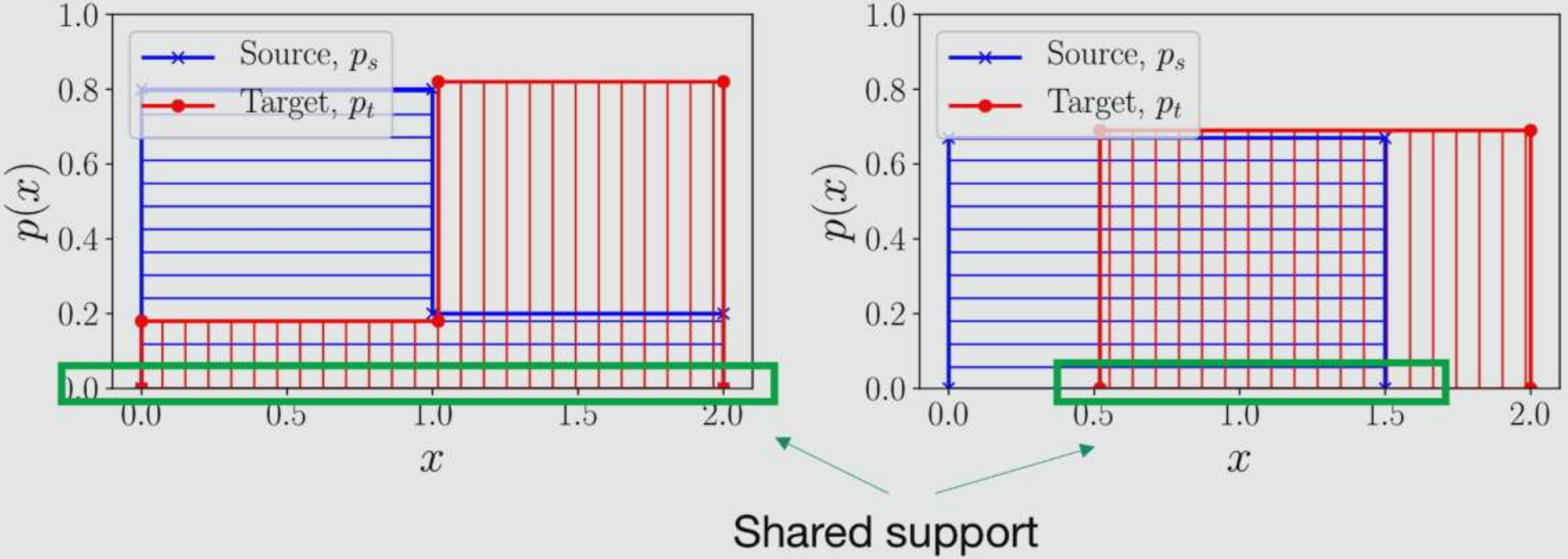
- Distance (MMD) greater in left than in right example





What is going on?

- Can we instead look at the support?



New representation learning bound

- Acknowledges information loss from representation
- Holds for any $\mathrm{Z} \coloneqq \phi(X)$

$$\underbrace{\underbrace{R_1(f_0(Z))}_{\text{Counterfactual risk}} \quad \leq \quad \underbrace{R_0^w\big(f_0(Z)\big)}_{\text{Weighted factual risk}} \quad + \quad \underbrace{d_{\mathrm{supp}}\big(p_0(Z), p_1(Z)\big)}_{\text{Lack of shared support}}}_{\text{Observable}} \quad + \quad \underbrace{\underbrace{\eta_\phi^\ell(f_0)}_{\text{Excess information loss}}}_{\text{Unobservable}}$$

[1]J., Ranganath, Sontag. *In preparation*.

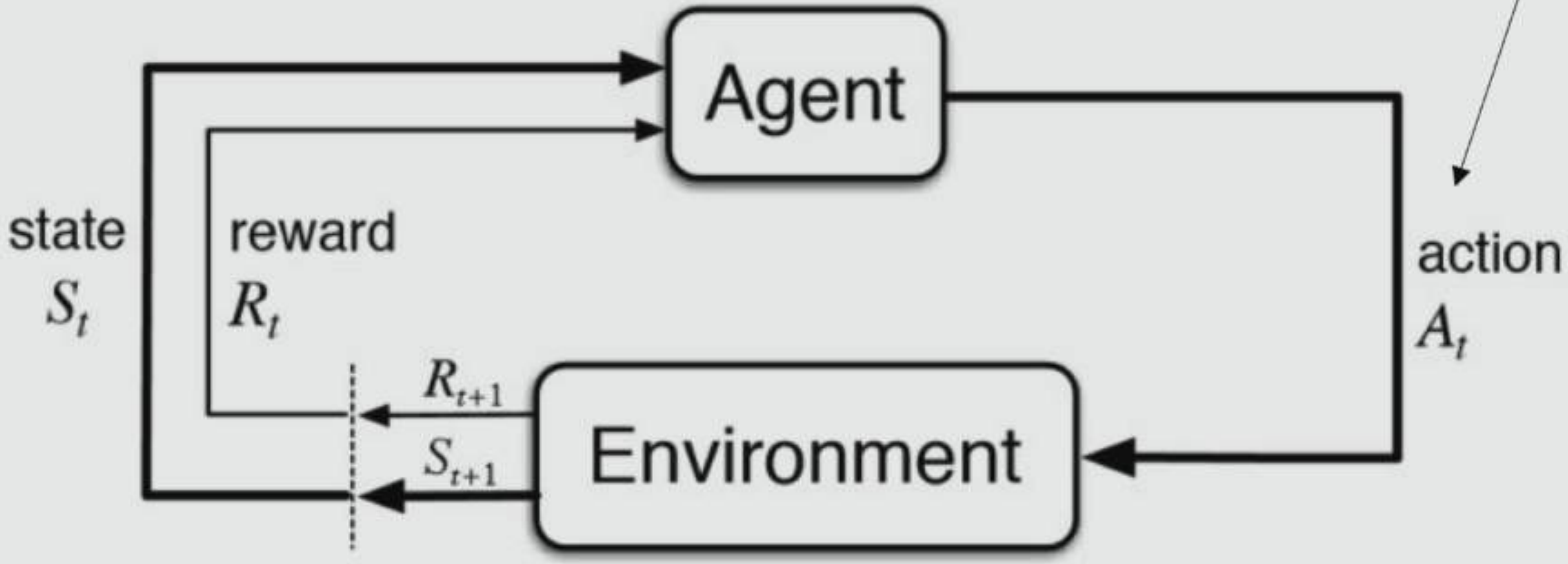
Takeaways

- **Domain adaptation** can inspire but are not magic
  - Same old problems from causal inference remain…
- Low-dimensional **representations** can help with regression, weighting
- New assumptions needed for consistent estimation

Reinforcement learning

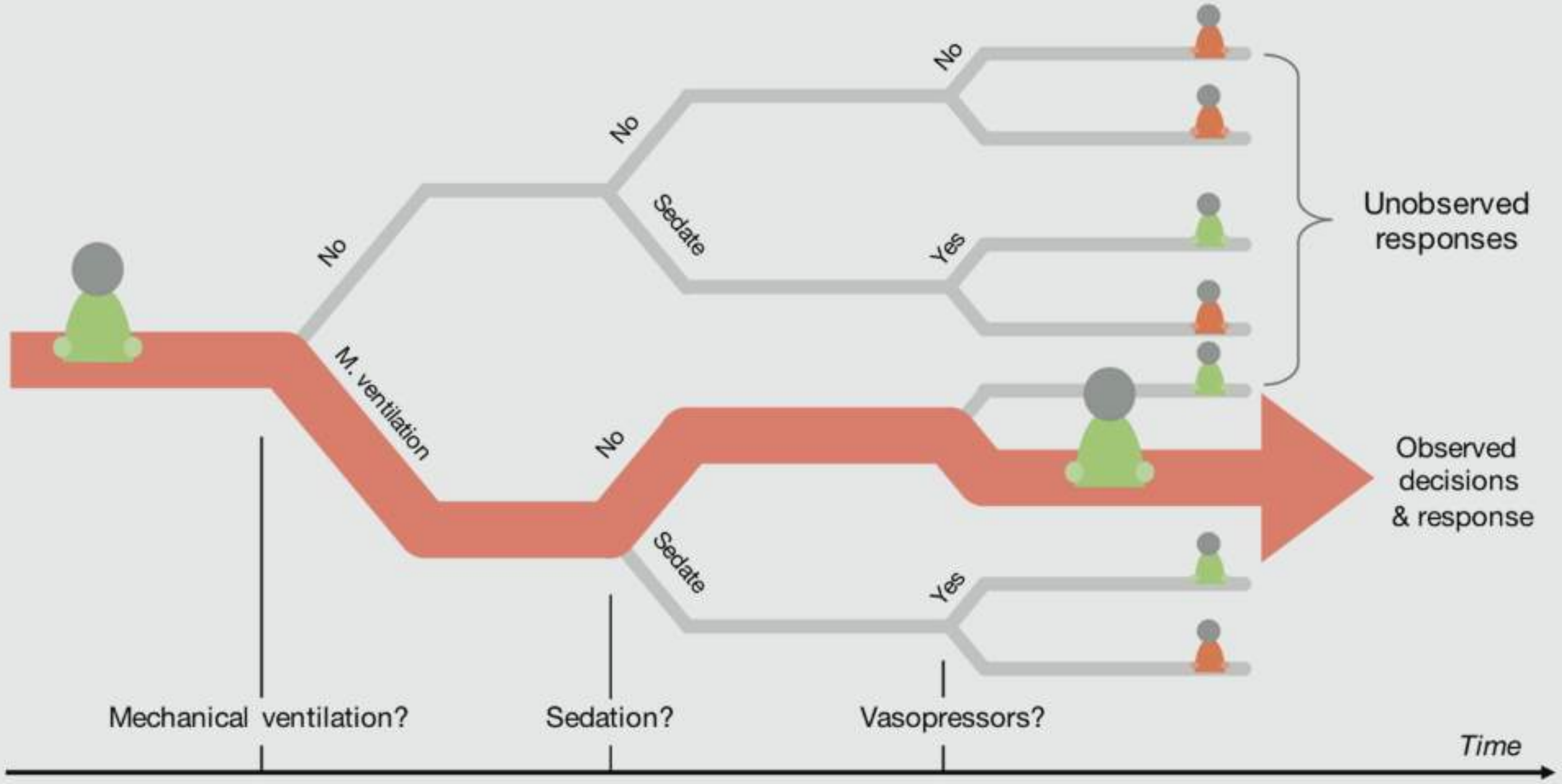
Policy $p(A_t \mid S_t)$

- Find a policy that maximizes future reward



Reinforcement learning (RL) in healthcare

Reinforcement learning (RL) in healthcare



c) Sequential decision making: Sepsis management

New representation learning bound

- Acknowledges information loss from representation
- Holds for any $Z := \phi(X)$

$$\underbrace{\underset{\text{Counterfactual risk}}{R_1(f_0(Z))} \leq \underset{\text{Weighted factual risk}}{R_0^w\big(f_0(Z)\big)} + \underset{\text{Lack of shared support}}{d_{\text{supp}}\big(p_0(Z), p_1(Z)\big)}}_{\text{Observable}} + \underbrace{\underset{\text{Excess information loss}}{\eta_\phi^\ell(f_0)}}_{\text{Unobservable}}$$

[1]J., Ranganath, Sontag. *In preparation*.

Domain adaptation bounds

- Take inspiration from domain adaptation[1,2]—bound the risk!
- **Under ignorability** w.r.t. X, the following bound holds for any f_0

$R_1(f_0)$	$\leq$	$R_0(f_0)$	$+$	$d_{\mathcal{H}}(p_0(X), p_1(X))$
Counterfactual risk		**Factual risk**		**Distributional distance w.r.t.** X

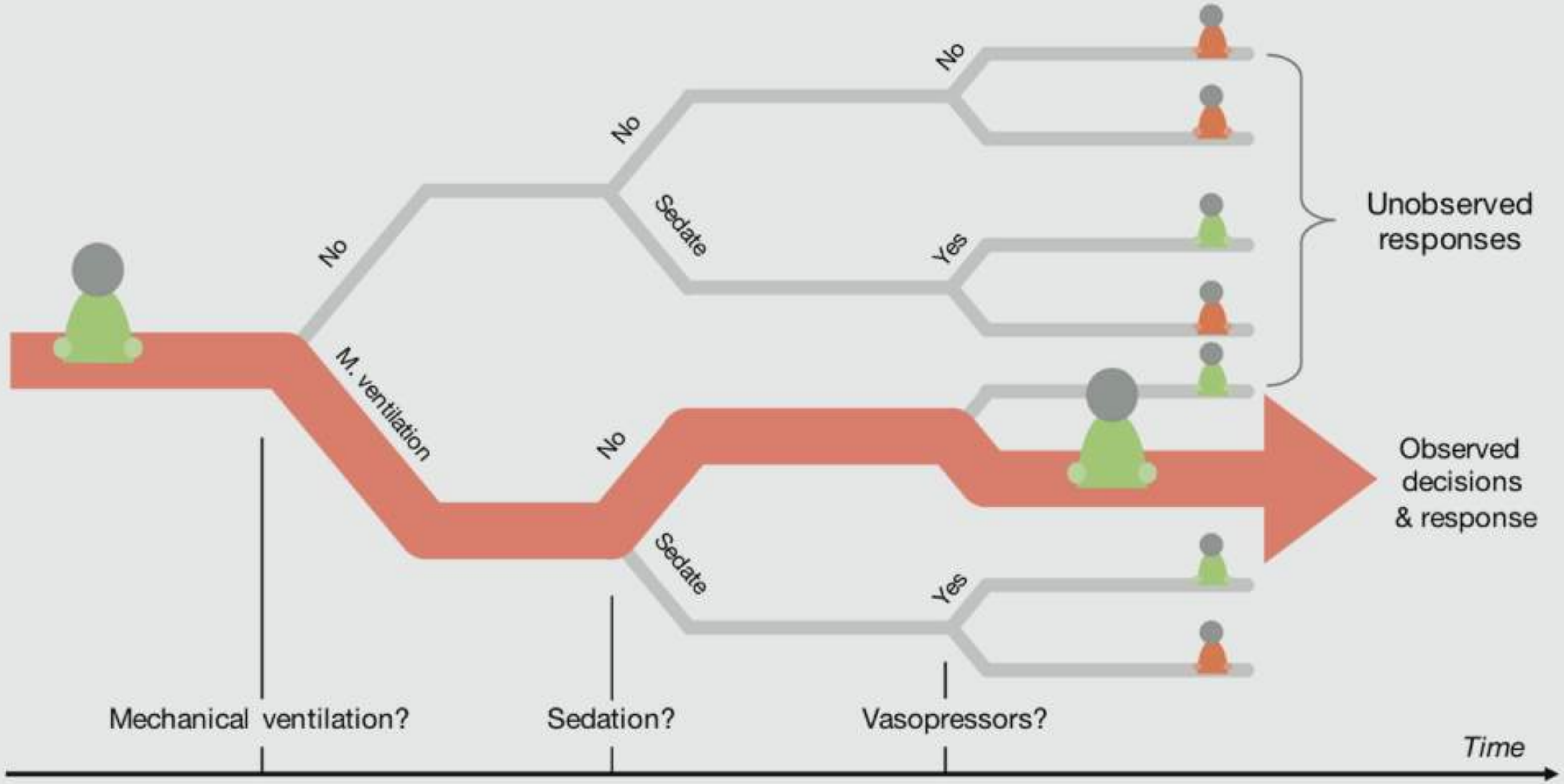
- The distance $d_{\mathcal{H}}(p, q) \coloneqq \sup_{g \in \mathcal{H}} \left| \mathbb{E}_p[g] - \mathbb{E}_q[g] \right|$ such that $\ell \in \mathcal{H}$

[1]Ben-David et al., 2008, [2]J., Shalit, Sontag, *ICML* 2016

Takeaways

- **Domain adaptation** can inspire but are not magic
  - Same old problems from causal inference remain…
- Low-dimensional **representations** can help with regression, weighting
- New assumptions needed for consistent estimation

Reinforcement learning (RL) in healthcare



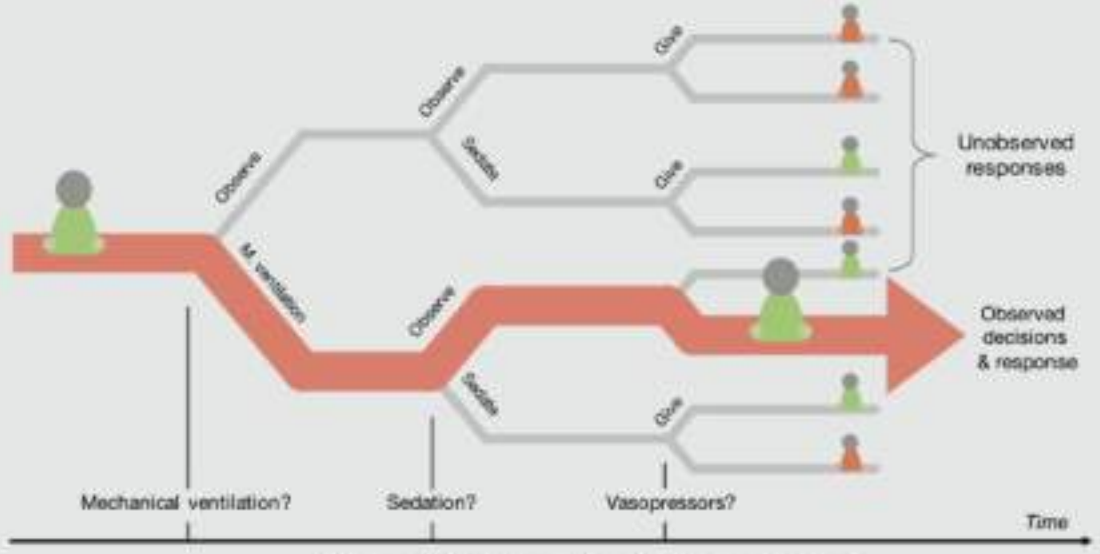
c) Sequential decision making: Sepsis management

Reinforcement learning (RL) in healthcare

- A string of work has applied **off-policy RL** to healthcare problems

- A popular application is **sepsis management**

- How should we evaluate learned policies?



c) Sequential decision making: Sepsis management

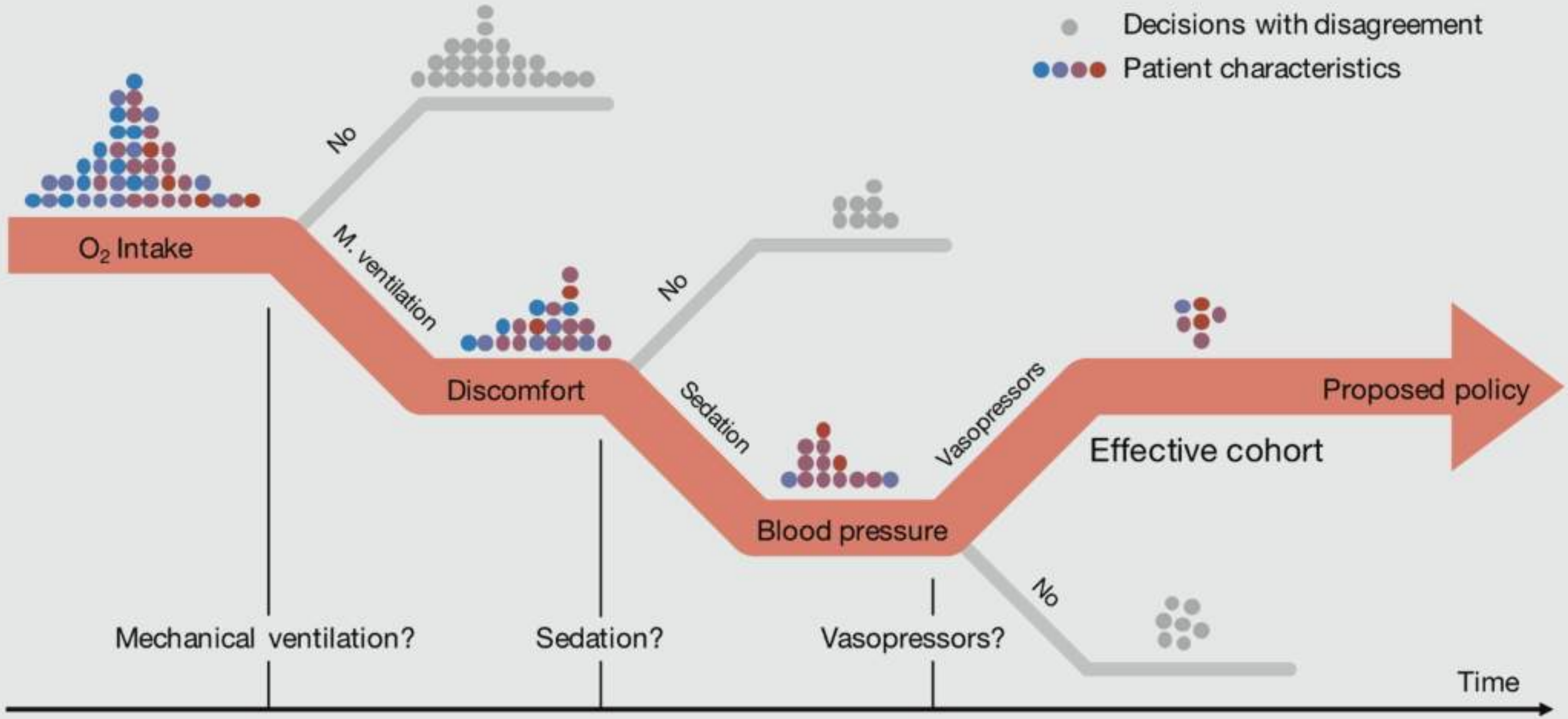
Evaluating RL policies

- In RL, we care about the long-term reward after following a policy

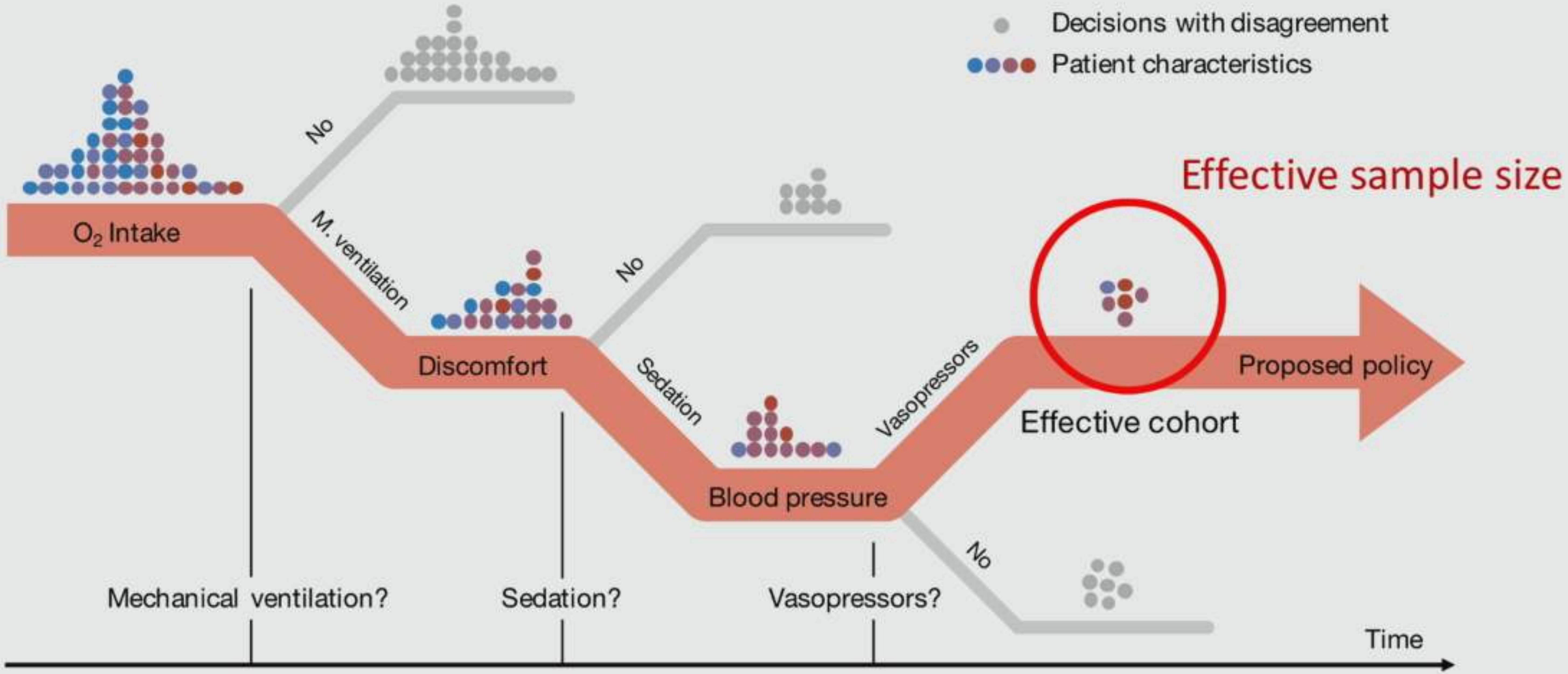
$$V(\pi) = \mathbb{E}_{a_t \sim \pi}\left[\sum_{t=1}^{T} R_t\right]$$

- **Overlap** now refers to the states in which the observed policy and proposed policy agree

Support for proposed decisions



Support for proposed decisions

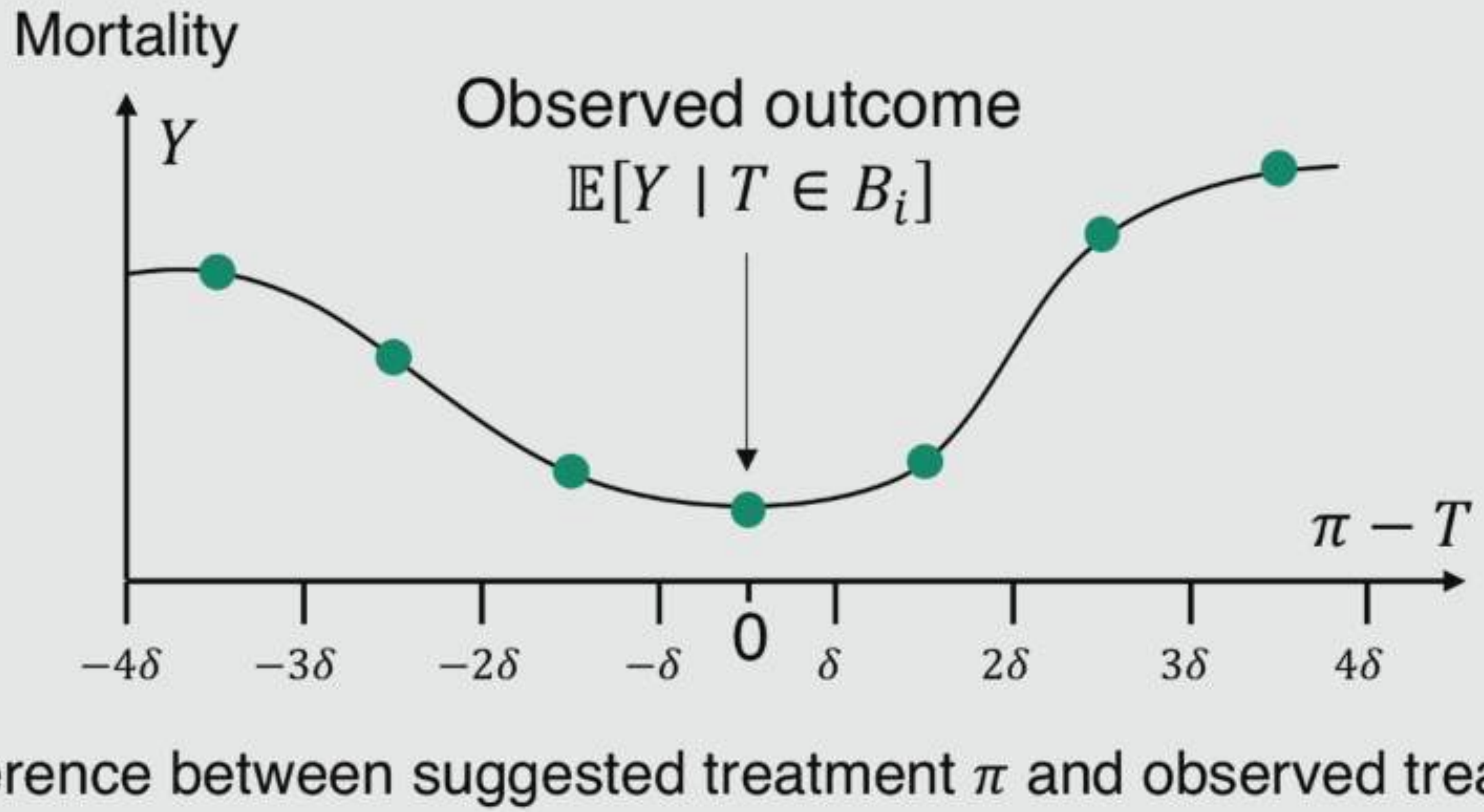


Evaluating off-policy RL

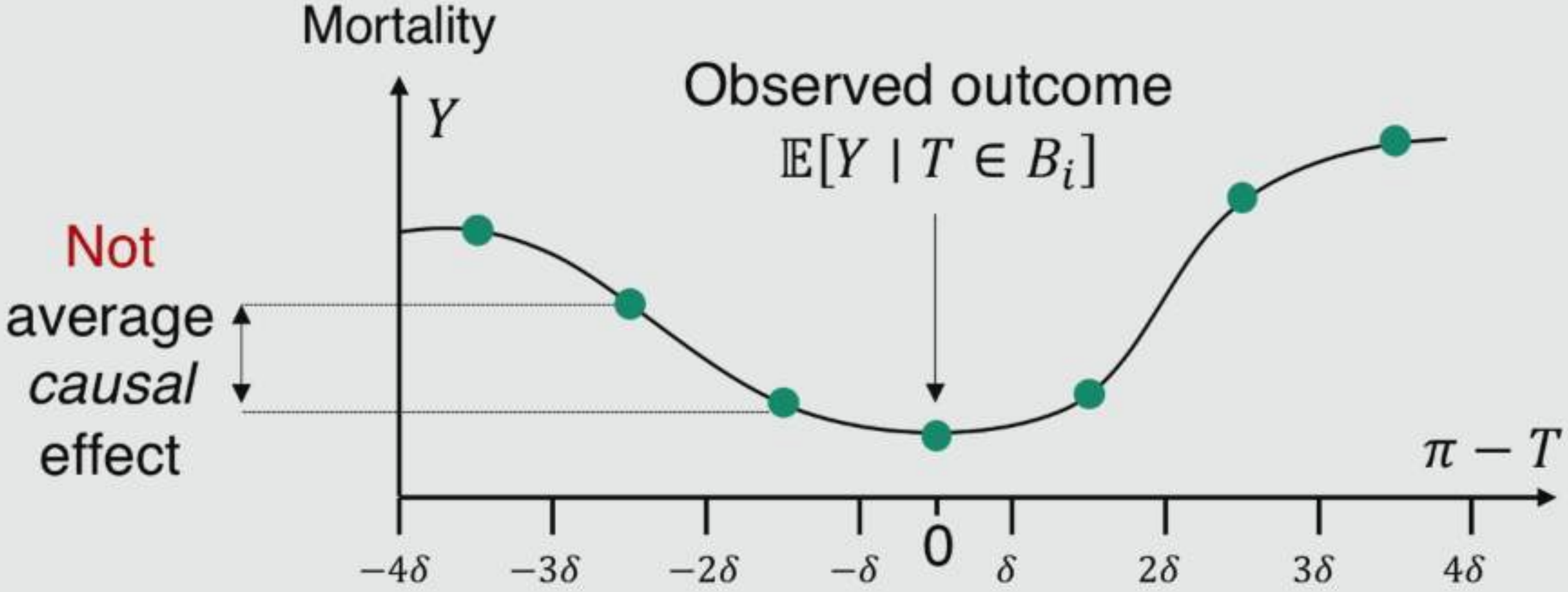
- Importance sampling has incredibly high variance due to small effective sample size
- Regression (model-based) evaluation is often biased
- Time-varying confounding is unverifiable
- Researchers have tried heuristic methods instead…

Heuristic alternative

- Authors have considered the following heuristic (Published at UAI/MLHC, …)

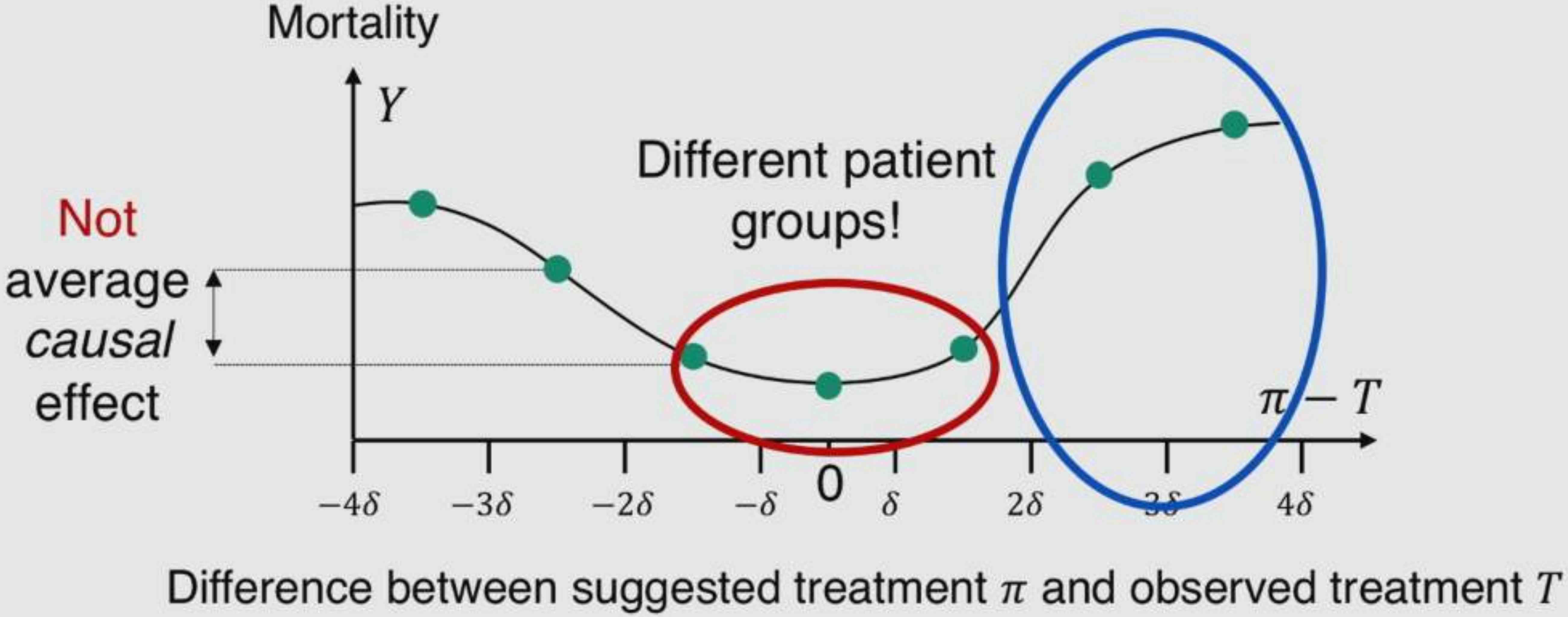


Heuristic alternative



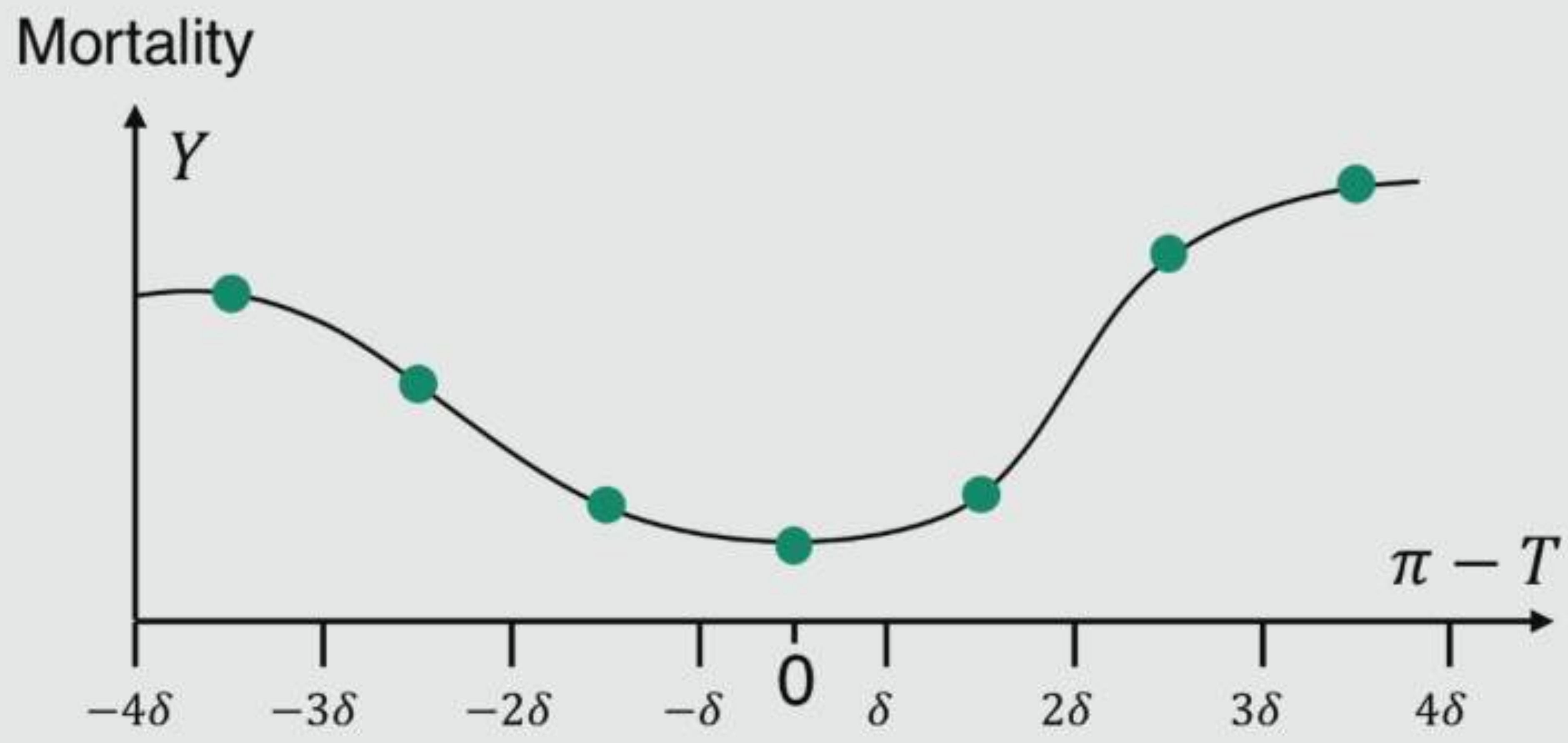
Difference between suggested treatment π and observed treatment T

Heuristic alternative



Heuristic alternative

- Often, the same figure appears for a **random** policy



Difference between suggested treatment π and observed treatment T

Takeaways

- Off-policy RL is **only harder** than estimating the effect of one decision
- The overlap problem is aggravated, and ignorability harder to verify
- Heuristic evaluation may lead to **dangerous** policies
- Check out https://arxiv.org/abs/1805.12298 for more

[1]Gottesman, J. et al. *In submission*, 2018. [2]Gottesman, J. et al. https://arxiv.org/abs/1805.12298, 2018


David Sontag


Rajesh Ranganath


Uri Shalit


Nathan Kallus

Fredrik D. Johansson
fredrikj@mit.edu