Causality in Healthcare

Estimating individual treatment effect: generalization bounds and algorithms

Sreya Francis

Brief Introduction

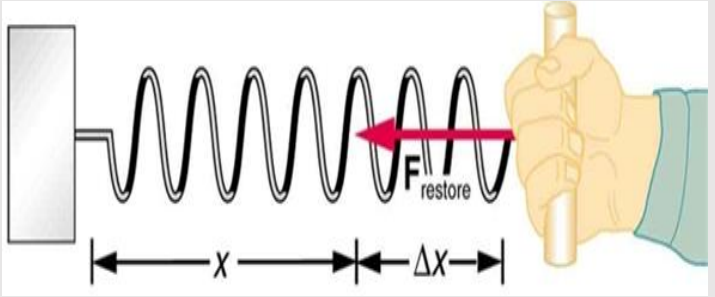
Introduction slides based on : https://dl.acm.org/citation.cfm?id=3291381

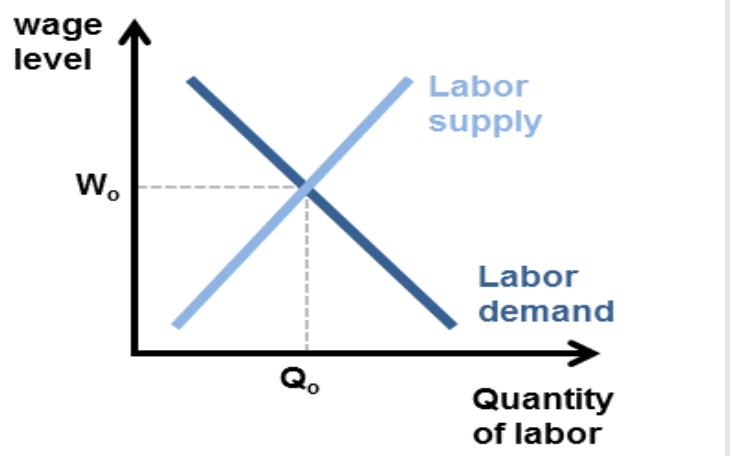
- What is causality?
- Potential Outcomes Framework
- Unobserved Confounds / Simpson’s Paradox
- Structural Causal Model Framework

Cause and Effect

- Questions of cause and effect common in biomedical and social sciences
- Such questions form the basis of almost all scientific inquiry
  - Medicine: drug trials, effect of a drug
  - Social sciences: effect of a certain policy
  - Genetics: effect of genes on disease

- **So what is causality?**
- **What does it mean to *cause* something?**





What is causality?

- A fundamental question
- Surprisingly, until very recently---maybe the last 30+ years---we have not had a mathematical language of causation. We have not had an arithmetic for representing causal relationships.

The Three Layer Causal Hierarchy

Pearl, Theoretical Impediments to Machine Learning with Seven Sparks from the Causal Revolution, arXiv:1801.04016v1. 11 Jan 2018

Level	Typical Activity	Typical Question	Examples
1. Association $P(y \mid x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y \mid do(x), z)$	Doing, Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x \mid x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

A practical definition

Definition: T causes Y iff

changing T leads to a change in Y,

keeping everything else constant.

The **causal effect** is the magnitude by which Y is changed by a unit change in T.

Called the "interventionist" interpretation of causality.

**Interventionist* definition [http://plato.stanford.edu/entries/causation-mani/]

Keeping everything else constant: Imagine a *counterfactual* world

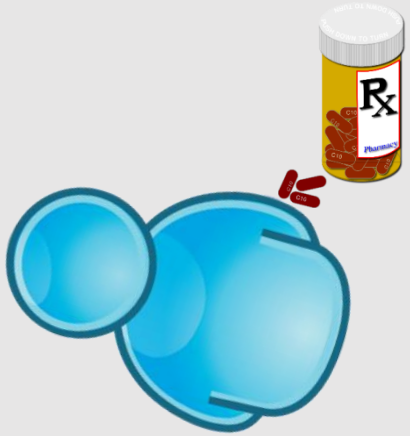
"What-if" questions

Reason about a world that does not exist.

- What if a system intervention was not done?
- What if an algorithm was changed?
- What if I gave a drug to a patient?

- What is causality?
- **Potential Outcomes Framework**
- Unobserved Confounds / Simpson's Paradox
- Structural Causal Model Framework

Potential Outcomes framework: Introduce a counterfactual quantity



$Y_{T=1}$

$Y_{T=0}$

Causal effect of treatment =

$$E[Y_{T=1} - Y_{T=0}]$$

- Potential outcomes reasons about causal effects by comparing outcome of treatment to outcome of no-treatment

Causal inference is the problem of estimating the counterfactual $Y_{t=\sim t}$

Person	T	$Y_{T=1}$	$Y_{T=0}$
P1	1	0.4	0.3
P2	0	0.8	0.6
P3	1	0.3	0.2
P4	0	0.3	0.1
P5	1	0.5	0.5
P6	0	0.6	0.5
P7	0	0.3	0.1

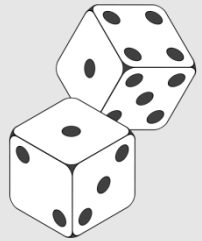
Causal effect: $E[Y_{t=1} - Y_{t=0}]$

Fundamental problem of causal inference: For any person, observe only one: either $Y_{t=1}$ or $Y_{t=0}$

Fundamental problem: counterfactual outcome is not observed

Randomized Experiments are the “gold standard”

One way to estimate counterfactual



- What is causality?
- Potential Outcomes Framework
- **Unobserved Confounds / Simpson's Paradox**
- Structural Causal Model Framework

The Simpson's paradox:

Consider success rate analysis of kidney stone treatment based on Observational Data

Kidney Stones	Treatment (A)	Treatment (B)
Success Rate	**78%**	**83%**

Which treatment do you think is better?

What if there are unobserved features that matter?

Simpson (1951)

The Simpson's paradox: Treatment B is better overall, but worse for each subgroup!

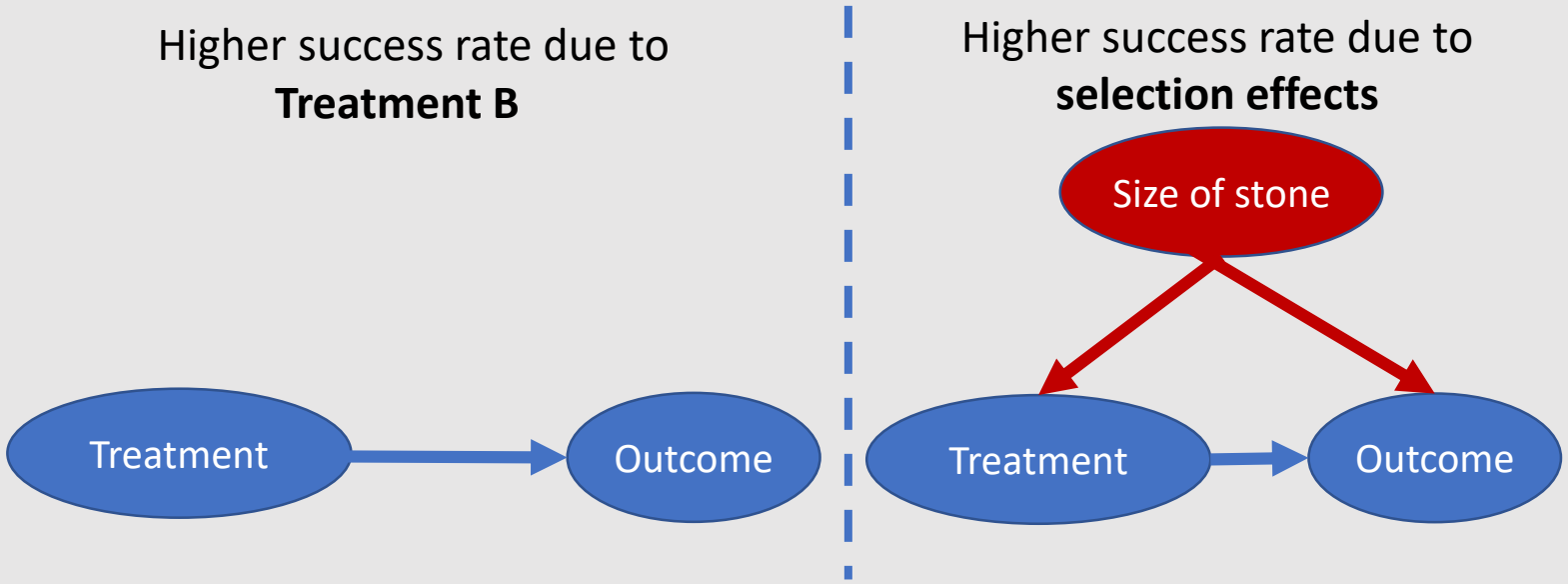
	Treatment (A)	Treatment (B)
Success Rate for small stones	93%	87%
Success Rate for large stones	73%	69%
Overall Success Rate	**78%**	**83%**

So, which is better?

Simpson (1951)

From metrics to decision-making

- **Did the change to treatment B increase success rate for the patients?**
- Answer (as usual):
- Maybe, maybe not (!)

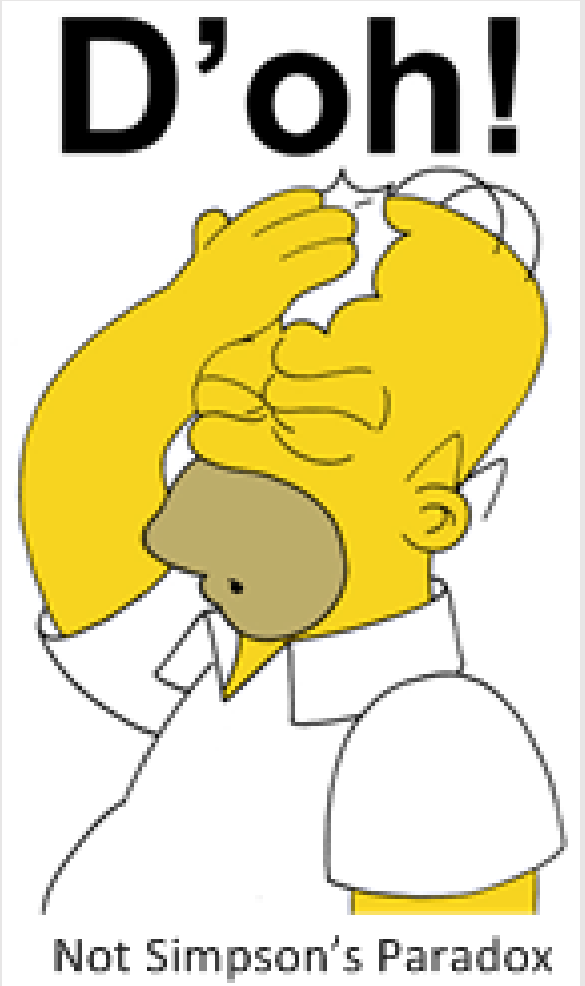


E.g., Treatment B is shown at a different time than A.

There could be other hidden causal variations.

Making sense of such data can be too complex.

Unobserved confounds are a threat to causal reasoning!



- What is causality?
- Potential Outcomes Framework
- Unobserved Confounds / Simpson’s Paradox
- **Structural Causal Model Framework**

Structural Causal Model: A framework for expressing complex causal relationships

People may have inter-related characteristics
- How are these characteristics associated with each other?

Other factors can influence the observed outcome
- How do they affect treatment and outcome?
- Which ones to include?

How to identify the causal effect in such cases?

Structural causal model and do-calculus :

- modeling the problem
- making assumptions explicit
- identifying the causal effect
- well-defined mechanisms for reasoning about causal relationships

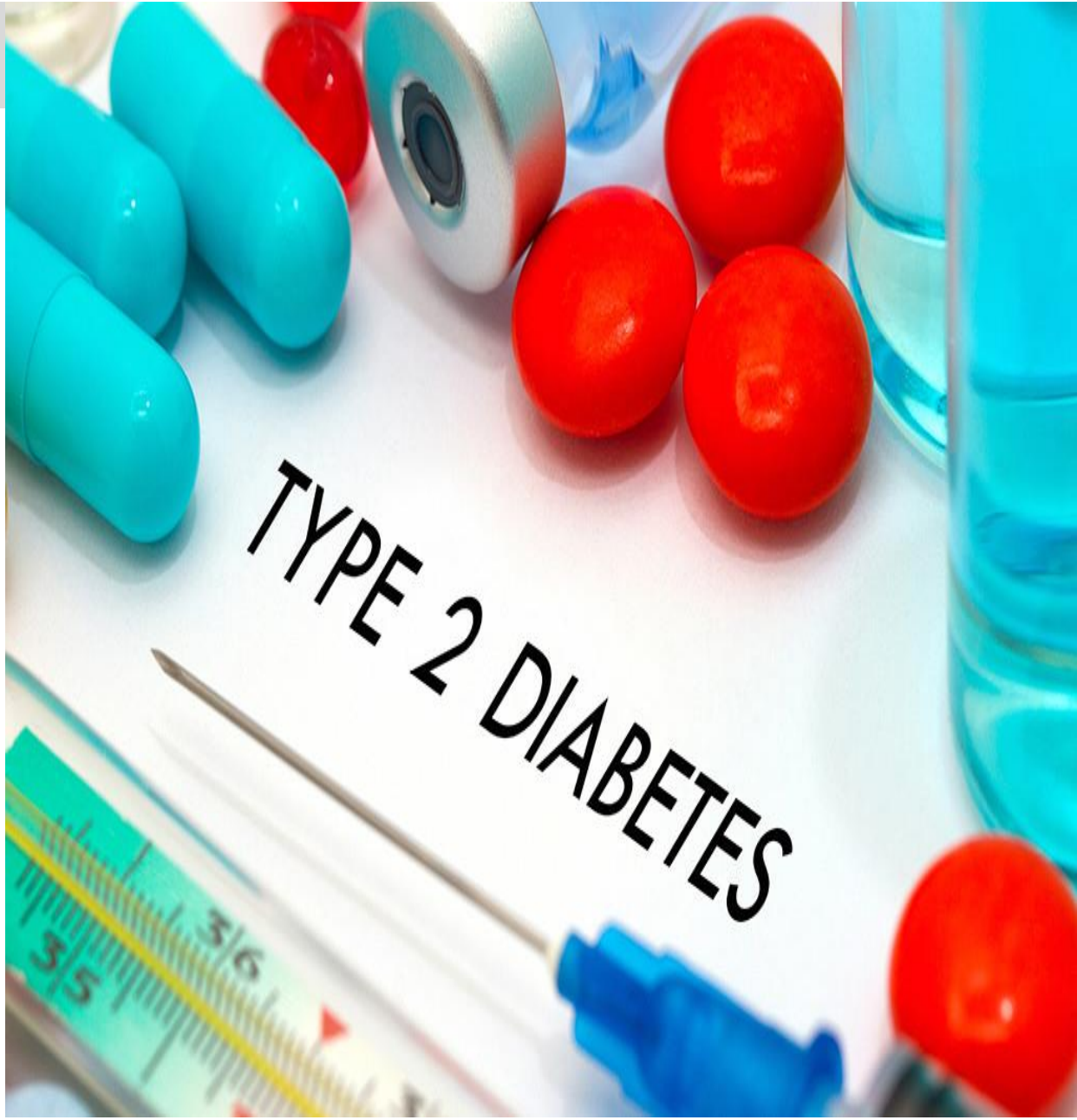
Potential outcomes-framework :

- estimating the causal effect

Potential applications of causal inference in healthcare

Predicting Diabetes Onset

- **Millions** of adults are **affected by Type 2** Diabetes
- The disease has severe symptoms and complications but is often **preventable** if risk factors are identified
- **What** is the **risk** of a person developing diabetes?
- **Who** is at risk, while we intervene on/ treat differently?



Prioritising Risk Factors for Type 2 Diabetes: Causal Inference through Genetic Approaches : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5960479/

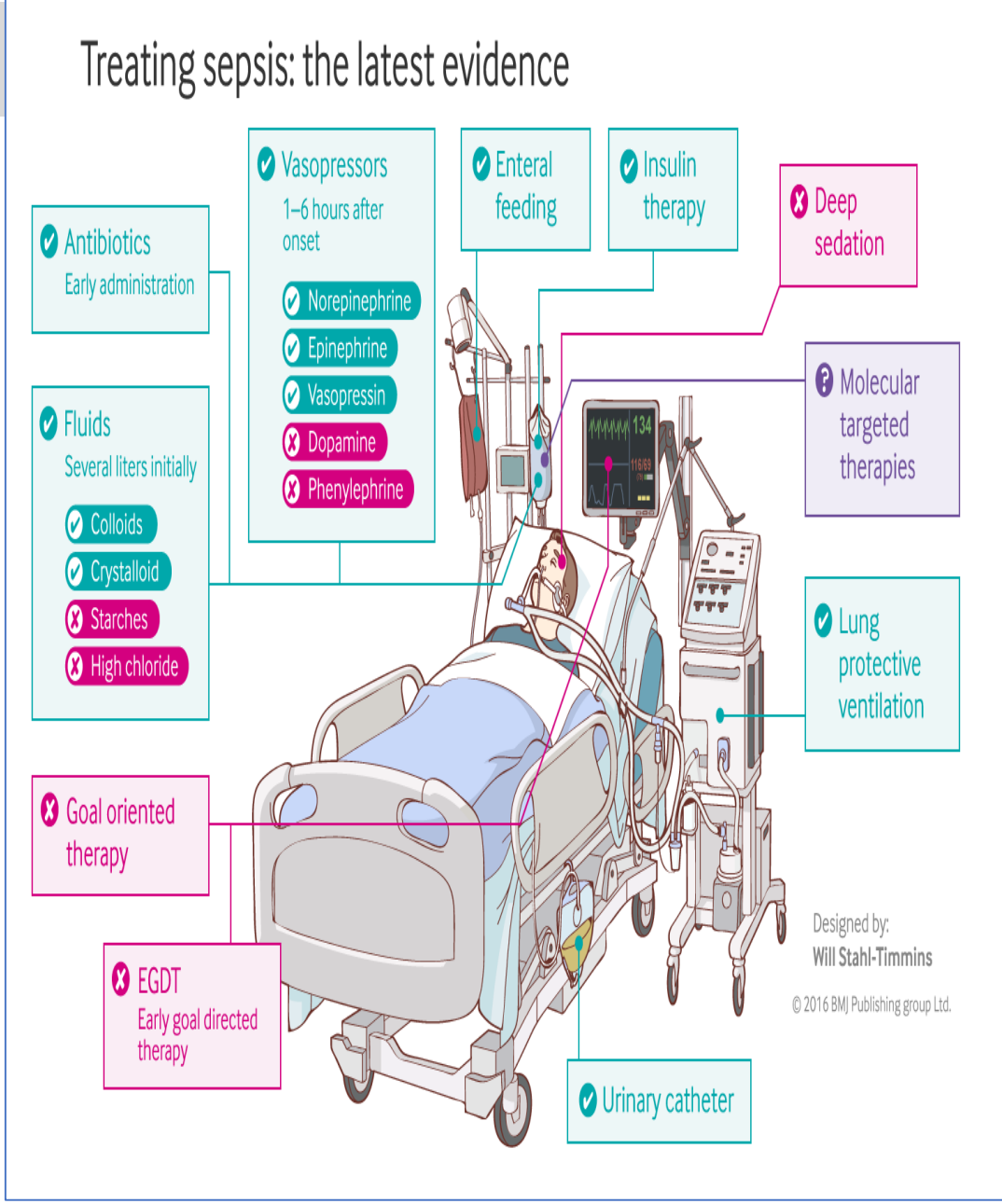
Opioid Addiction

- **Millions** of people are **addicted** to opioid medication
- **Over 10000 die** each year due to overdoses related to prescription opioids
- Try simple things like avoiding larger subscriptions that lead to higher risk
- What are the possible **drivers**?
- Who should be **prescribed what**?

Causal inference for evaluating prescription opioid abuse using trend-in-trend design : https://www.ncbi.nlm.nih.gov/pubmed/30714239

Sepsis Management

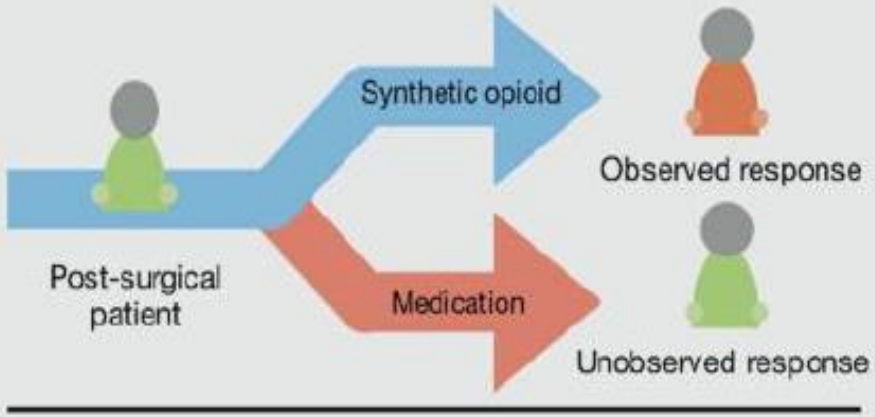
- **Sepsis –** One of the leading causes of death in ICU
- **Sepsis Management** is a continuous task of managing patients
- Treat not just the infection but make sure that all the vitals are in the right range.
- How to **maximize the long term quality** of these patients?



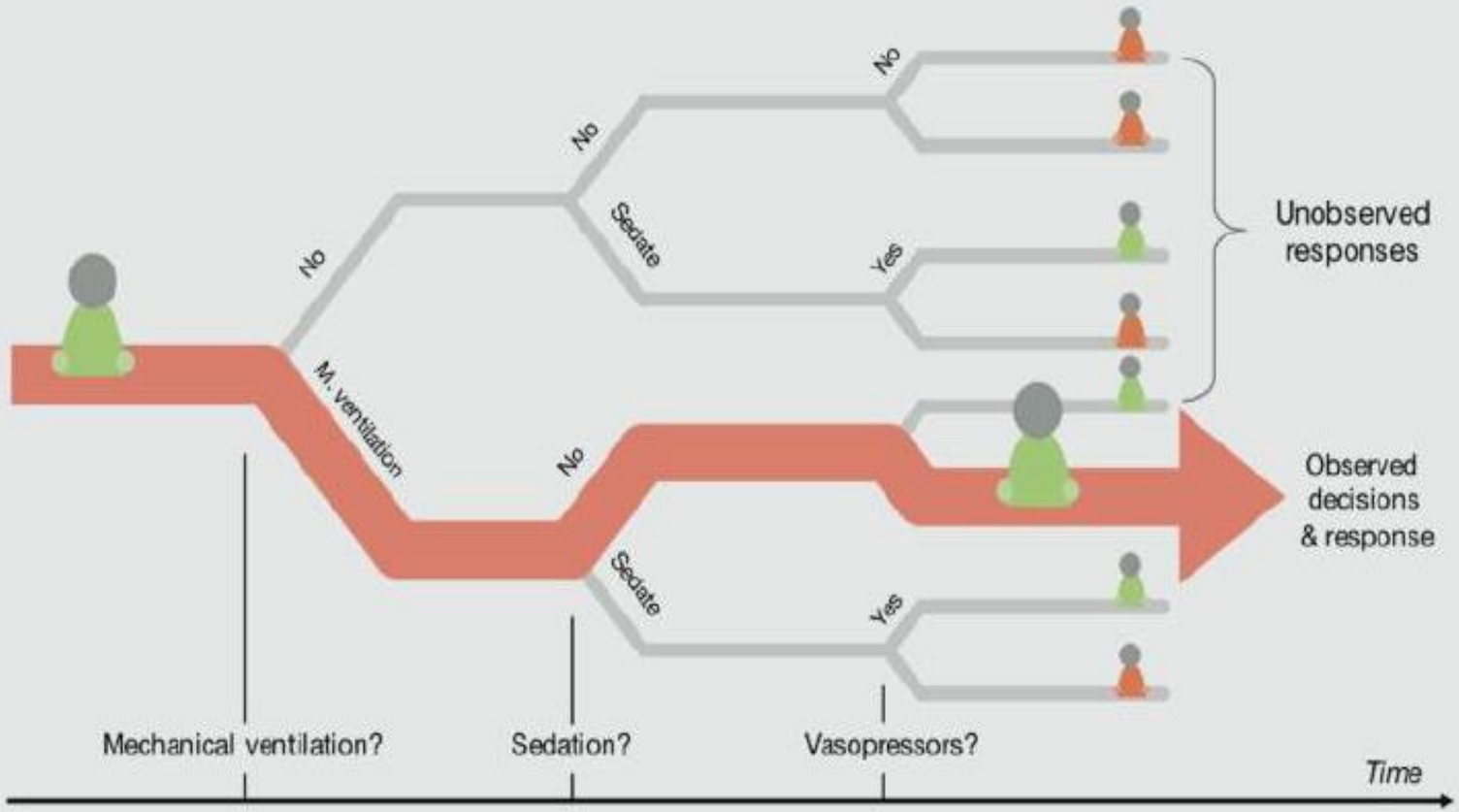
Evaluating Reinforcement Learning Algorithms in Observational Health Settings : https://arxiv.org/abs/1805.12298

Prediction and decision making in healthcare

a) Prediction: Diabetes onset



b) Treatment effect estimation: Opioid addiction

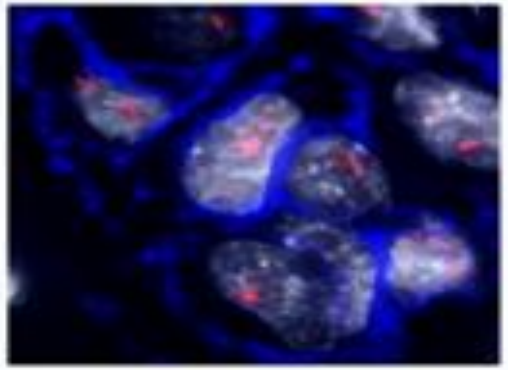


c) Sequential decision making: Sepsis management

Evaluating Reinforcement Learning Algorithms in Observational Health Settings : https://arxiv.org/abs/1805.12298

Prediction and decision making in healthcare

What treatment will work best for this patient?



Expansion pathology
(image from Andy Beck)

- People respond differently to treatment
- Goal: use data from other patients and their journeys to guide future treatment decisions
- What could go wrong if we trained to predict (past) treatment decisions?

"David" → Treatment A

"John" → Treatment B

"Juana" → Treatment A

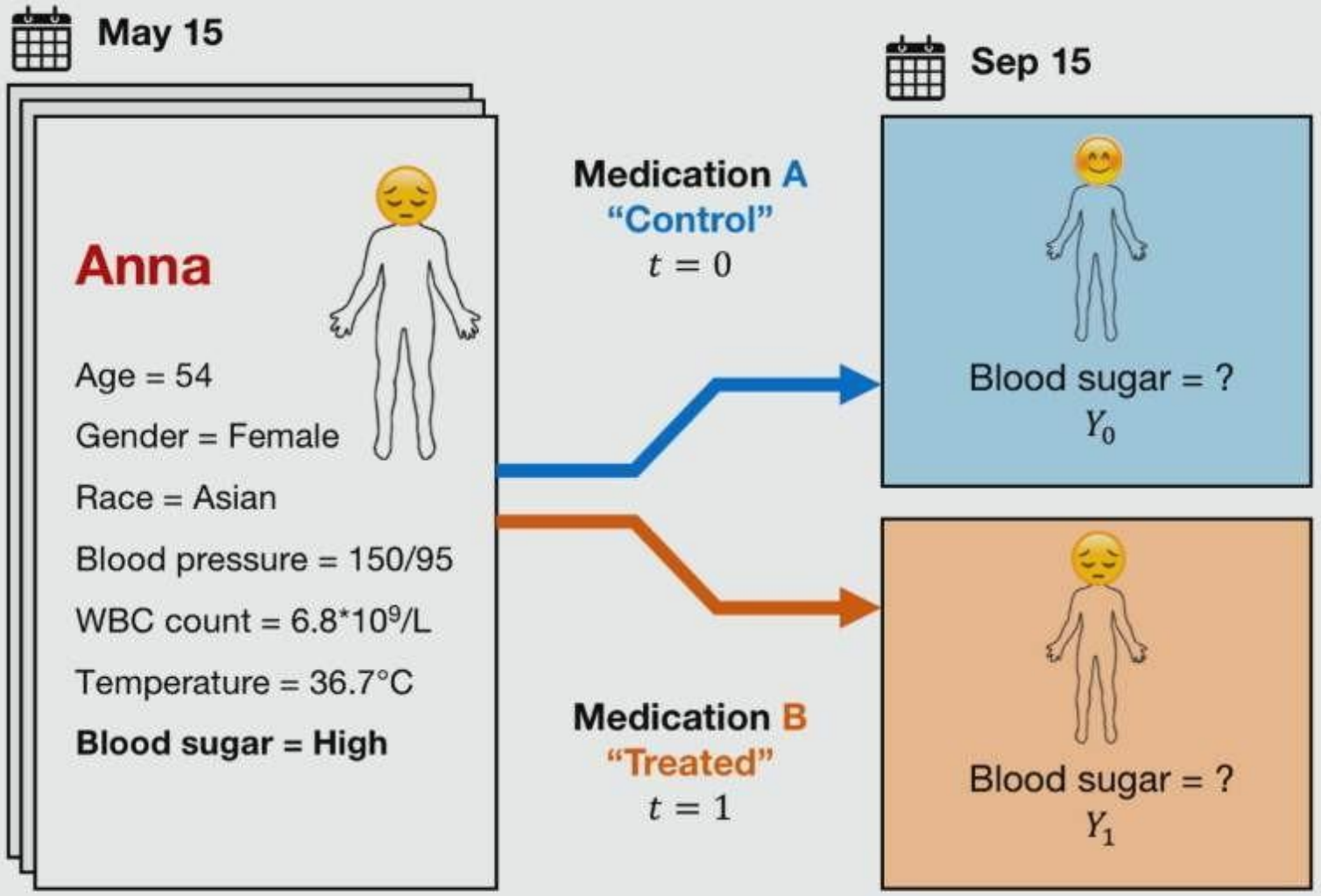
Best this can do is match current medical practice!

Predicting effects of decisions require causal reasoning

How to estimate causal effects in high dimensional data?

Slides based on : https://people.csail.mit.edu/dsontag/papers/ShalitJohanssonSontag16_slides.pdf

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

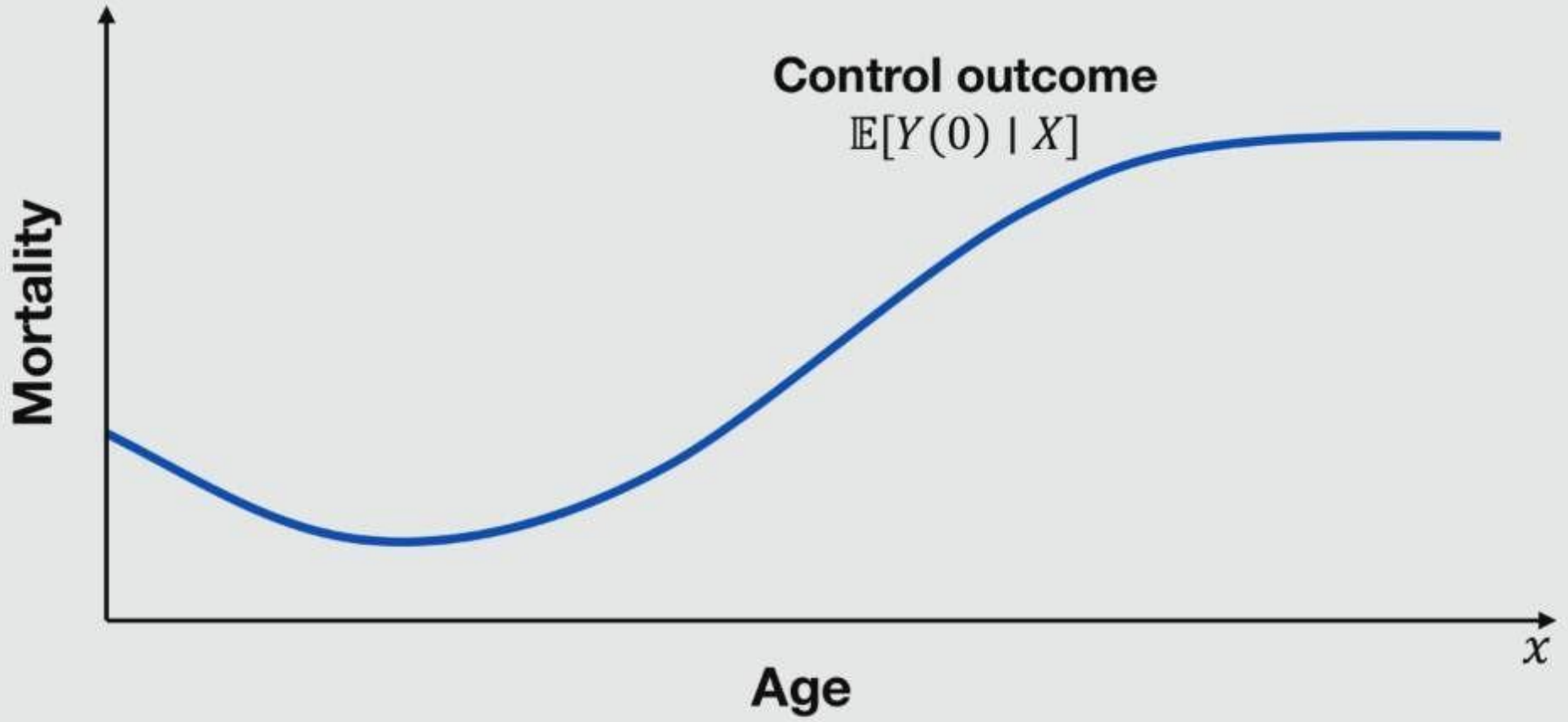
$T \in \{0, 1\}$ — **Treatment** assignments

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

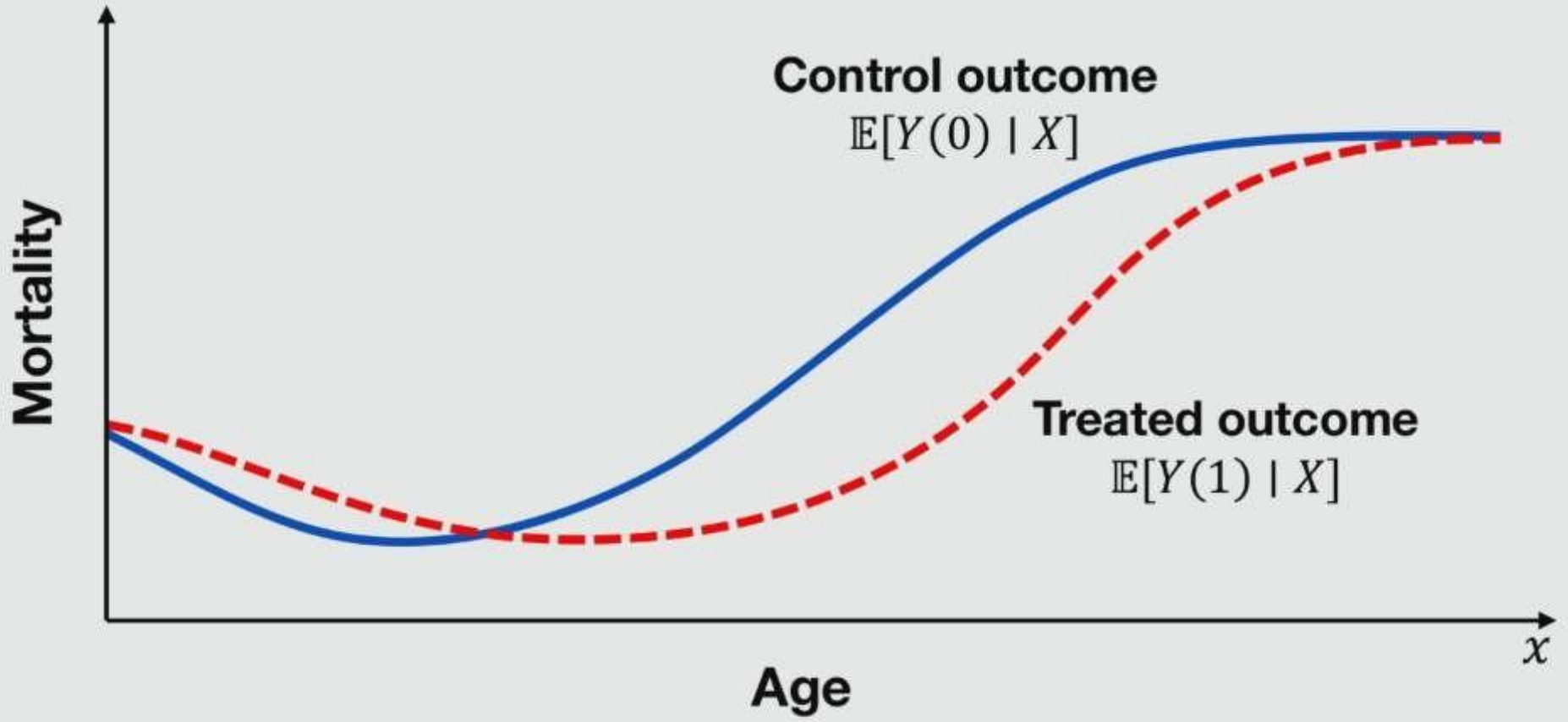
- **Goal:** Estimate counterfactual/potential outcome: $\mathbb{E}[\, Y(t) \mid X = x \,]$
- Conditional Average Treatment Effect (CATE)

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

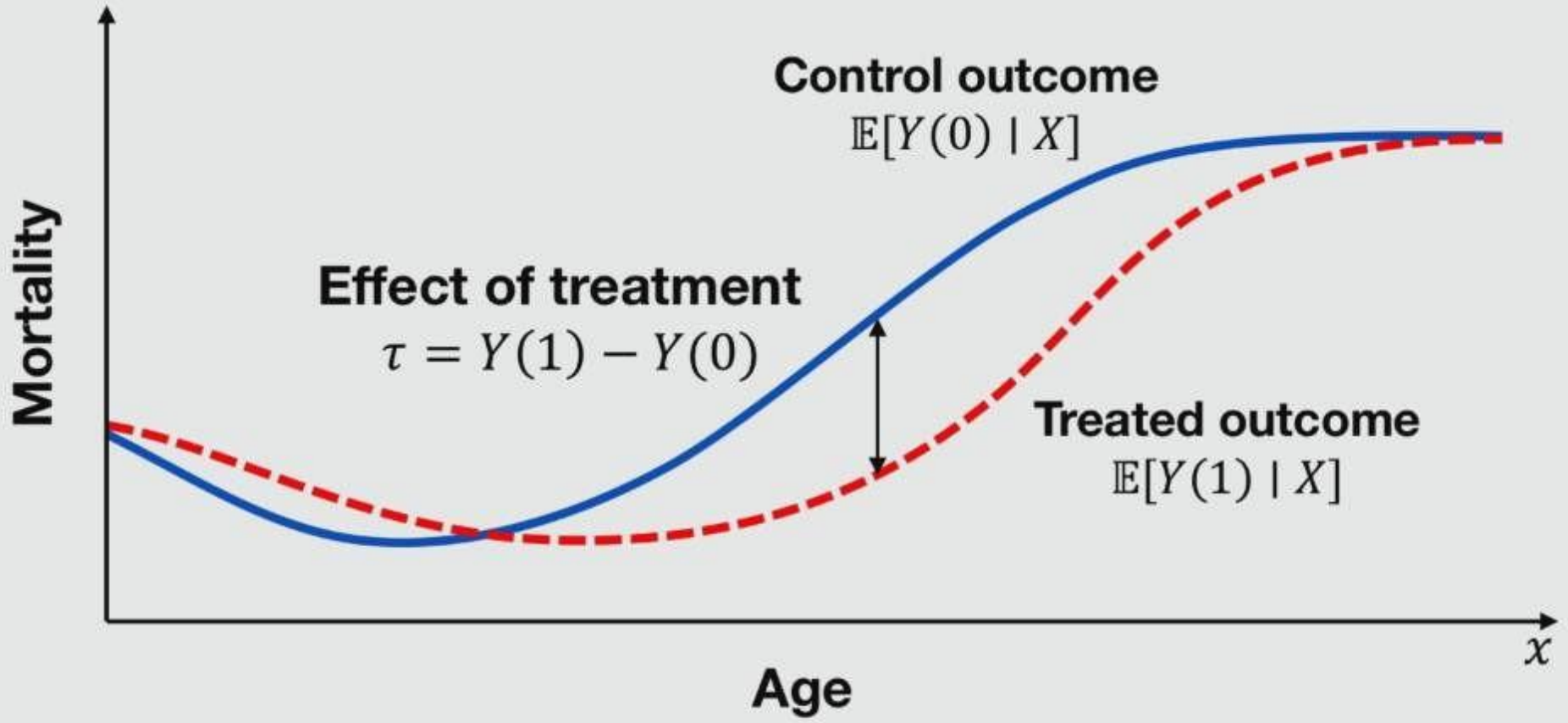
Potential outcomes and CATE



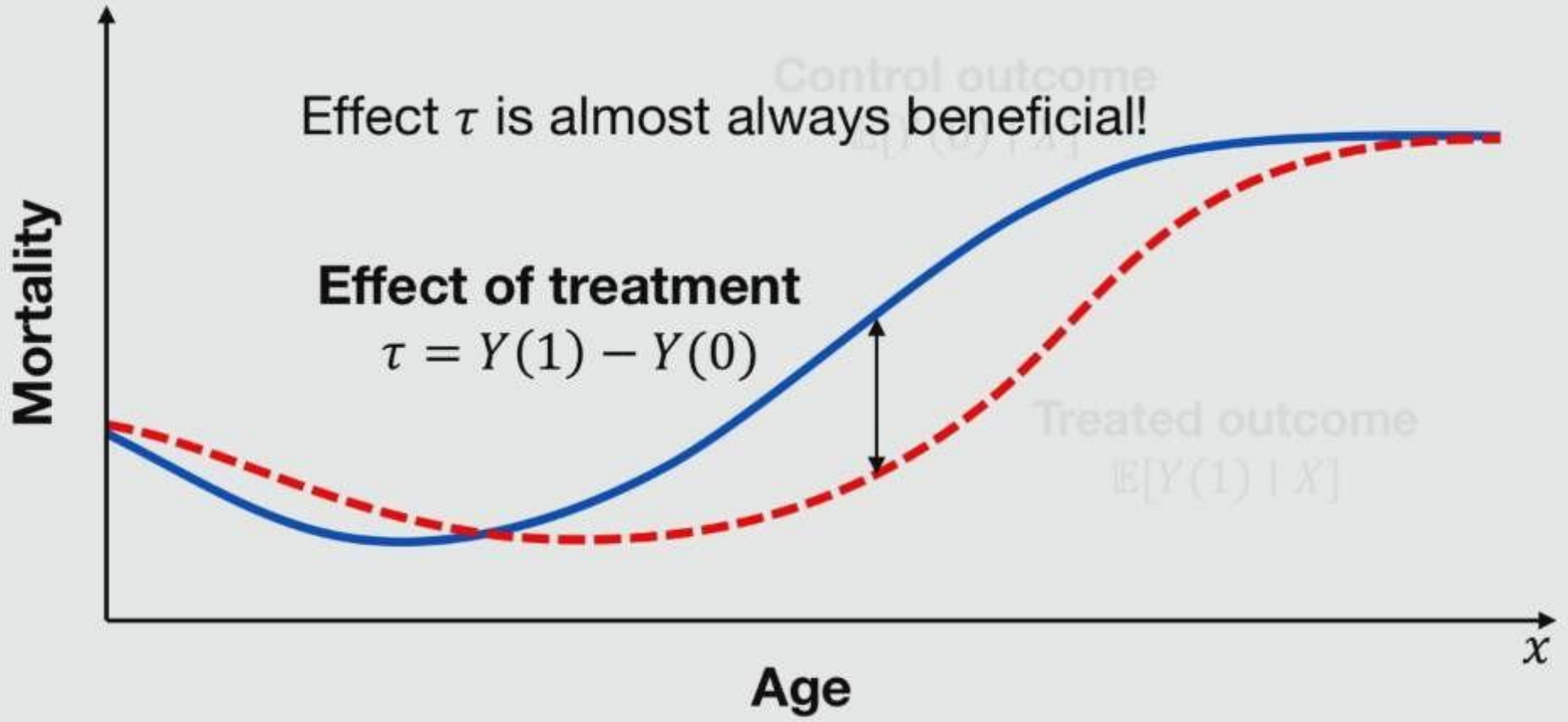
Potential outcomes and CATE



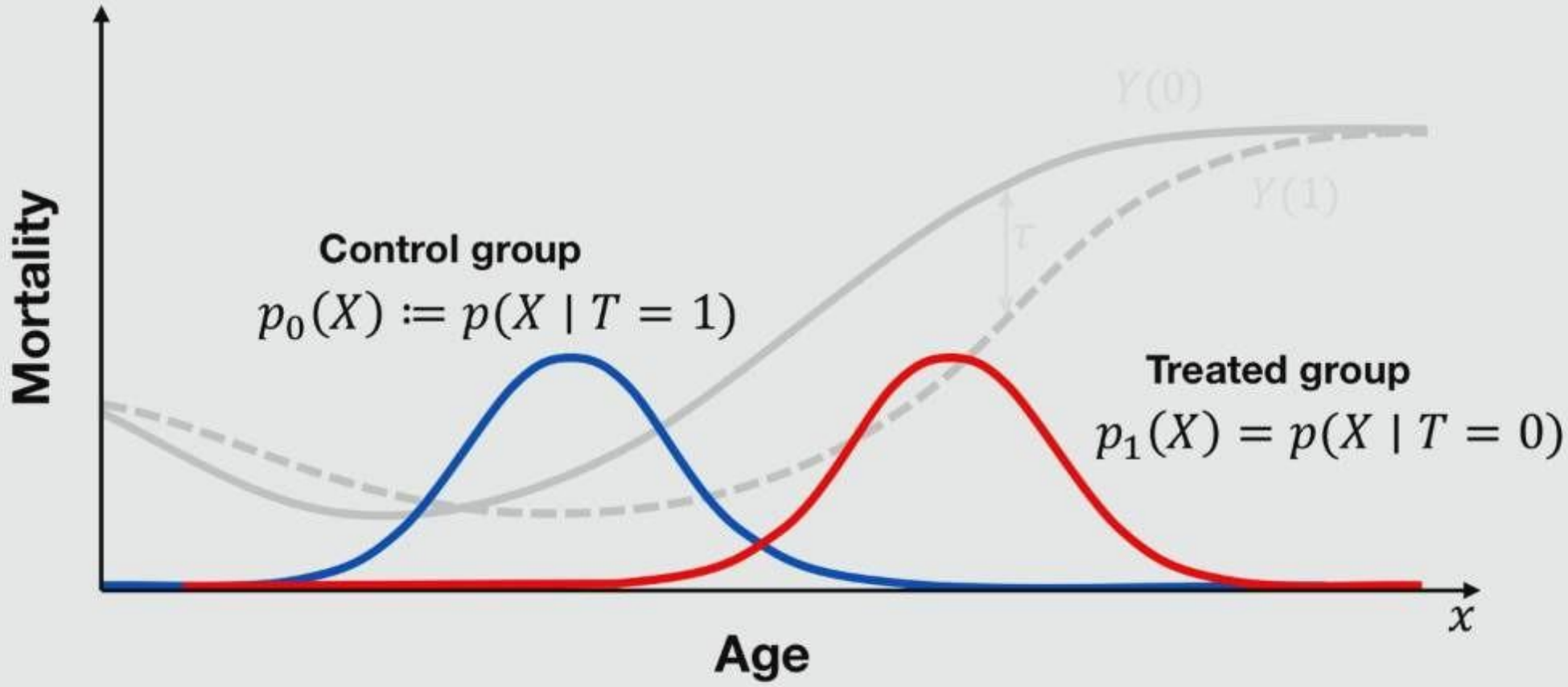
Potential outcomes and CATE



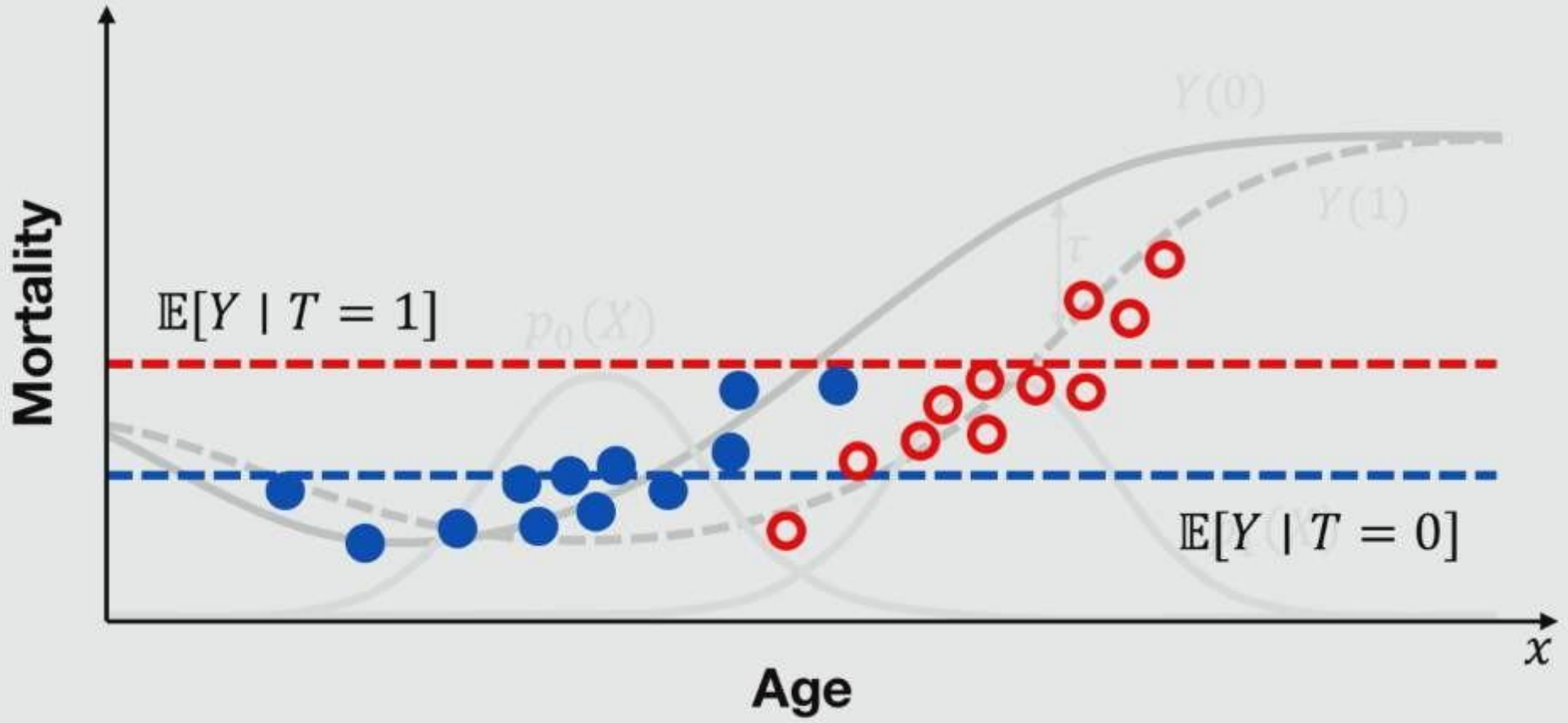
Potential outcomes and CATE



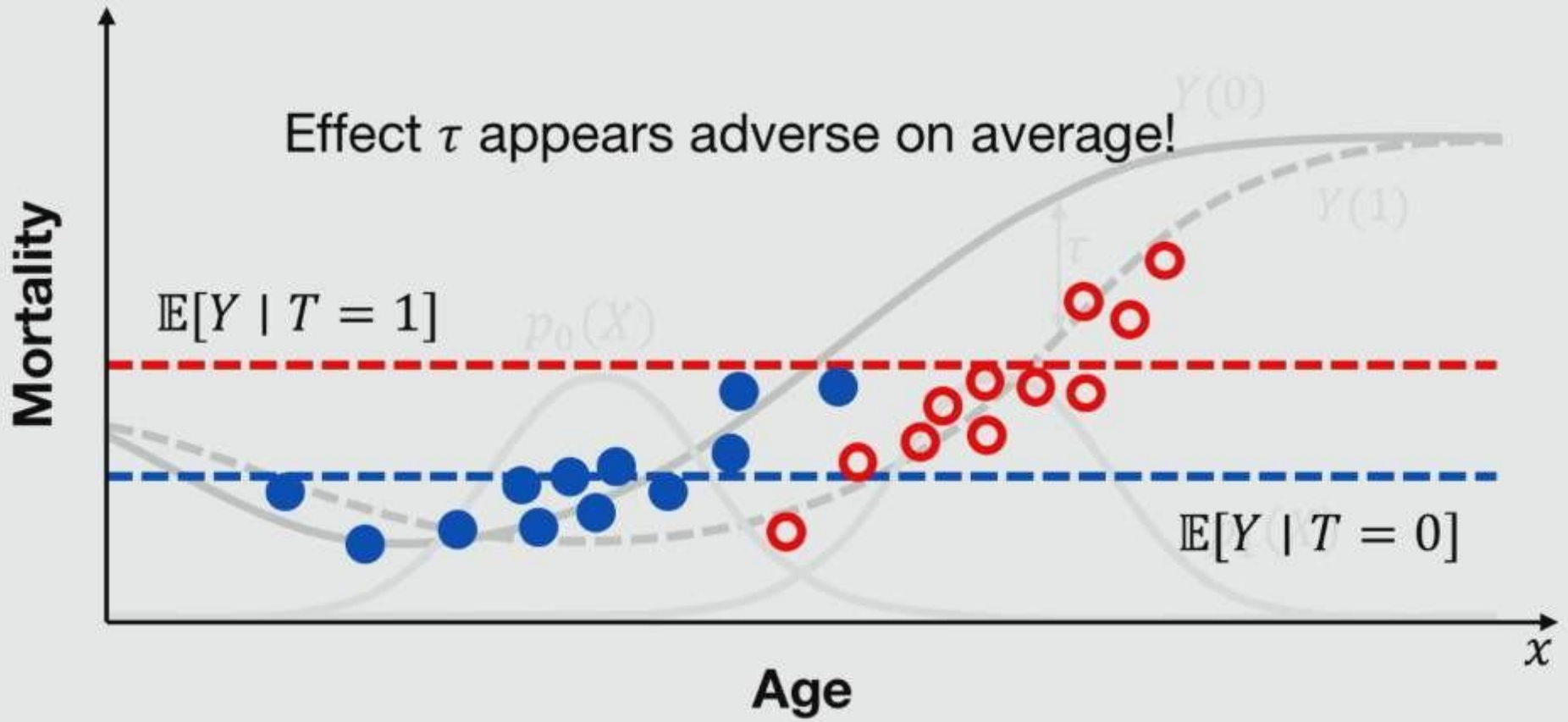
Treatment groups



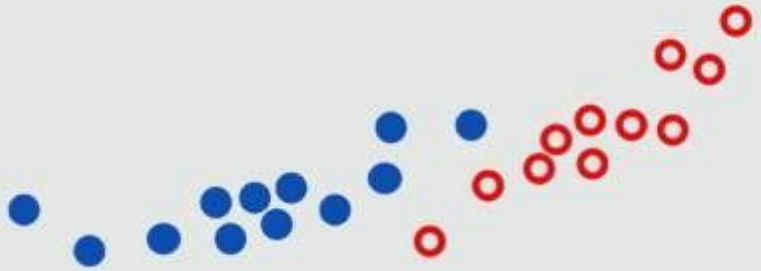
Treatment effect is confounded by age



Treatment effect is confounded by age



We have several problems



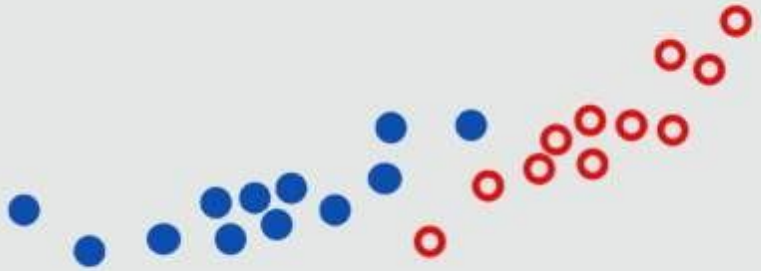
1. **Confounding:**

   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

Identifying assumptions



- **Ignorability**

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

  If we control for X, we can estimate τ

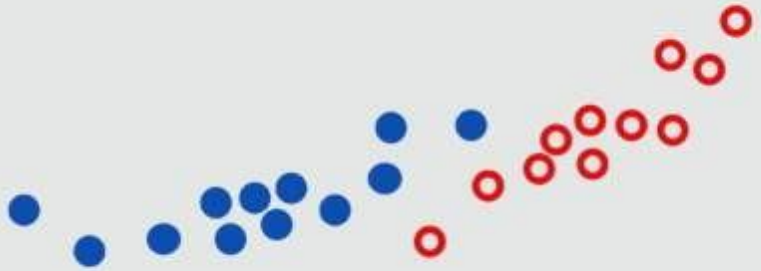
- **Common support**

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

  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

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

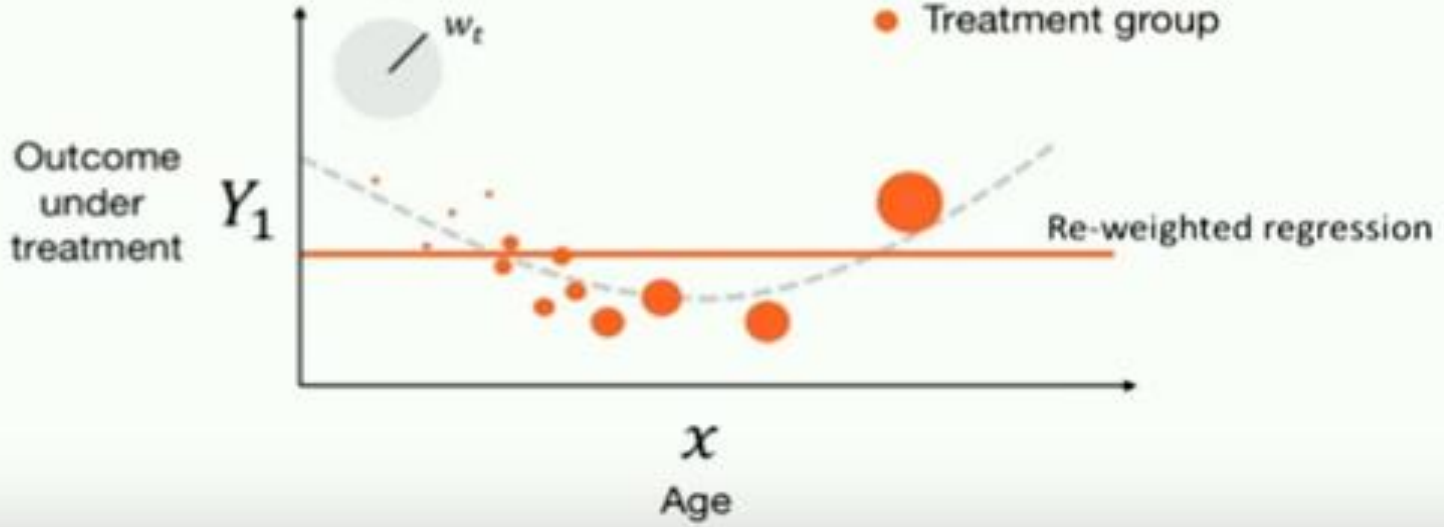
- If treatment groups are very different, the estimation error based on factual outcomes is **not representative** of the error in the counterfactual

Re-weighting

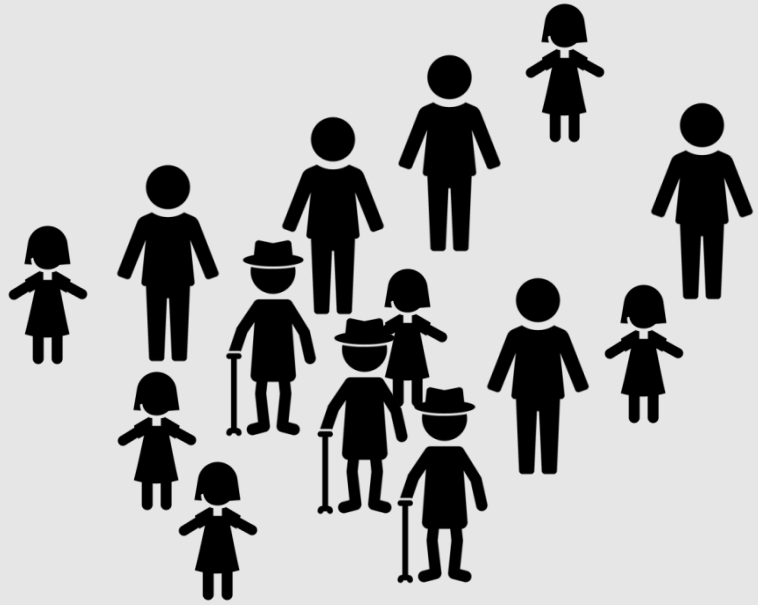
Inverse propensity re-weighting

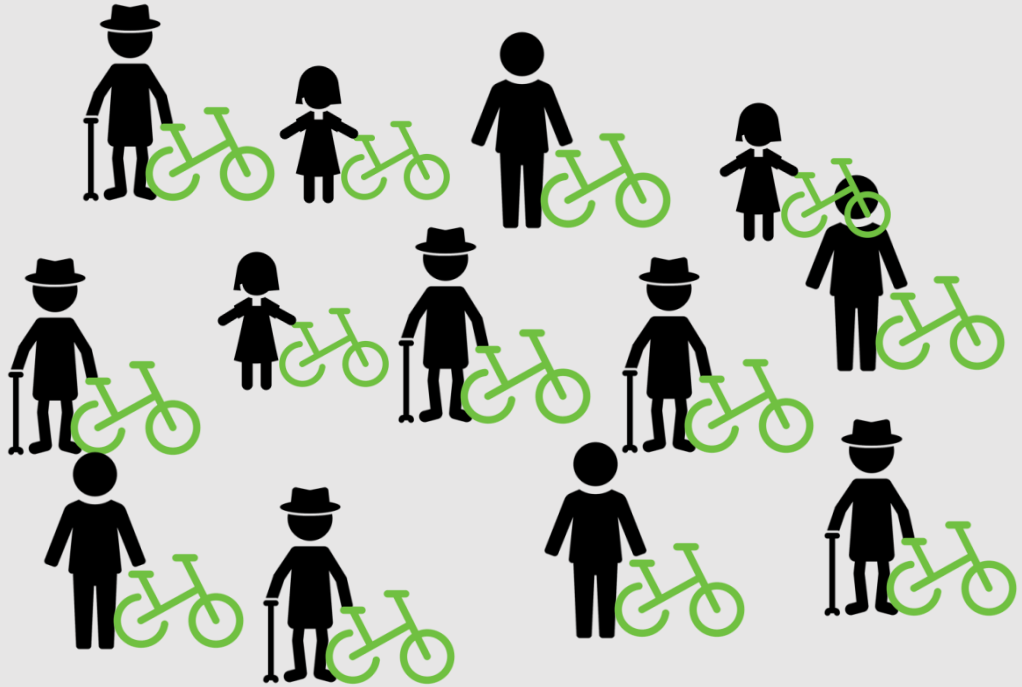
- Minimize re-weighted loss to make it representative of population

$$\min_f \mathbb{E}_{p^{t=1}(x)} \left[\frac{p(t=0 \mid x)}{p(t=1 \mid x)} \ell_f(x, 1) \right]$$



Matching





Avg blood sugar = 250

Avg blood sugar = 280

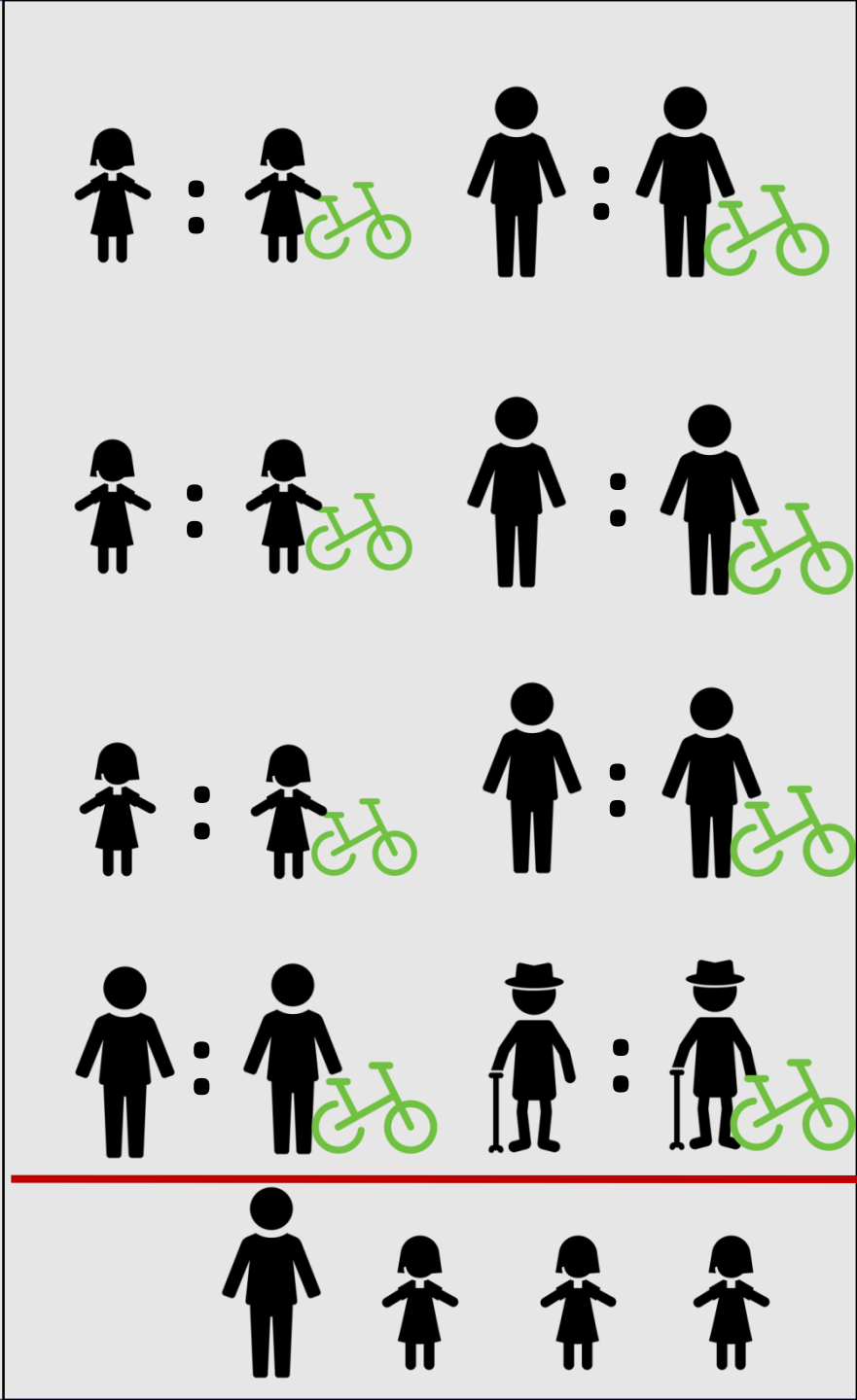
Matching

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

$$\text{Very similar} ::= Distance(X_i, X_j) < \epsilon$$

Paired individuals provide the counterfactual estimate for each other.

Average the difference in outcomes within pairs to calculate the *average-treatment-effect on the treated*



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

Mitigating the curse of dimensionality?

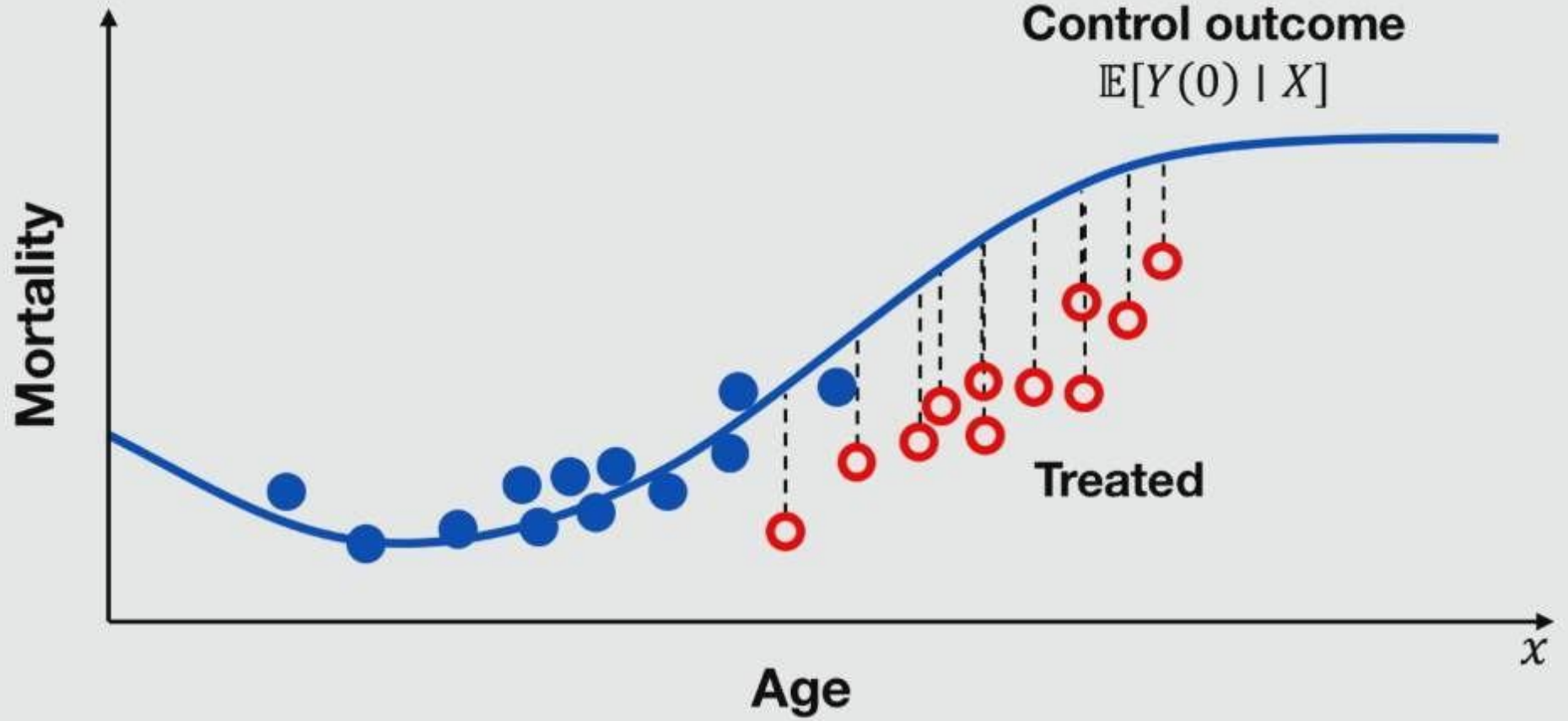
- Can we find a representation of our data $\Phi(X)$ such that
- Ignorability

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

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

How domain adaptation is related to counterfactual prediction?

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)

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

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

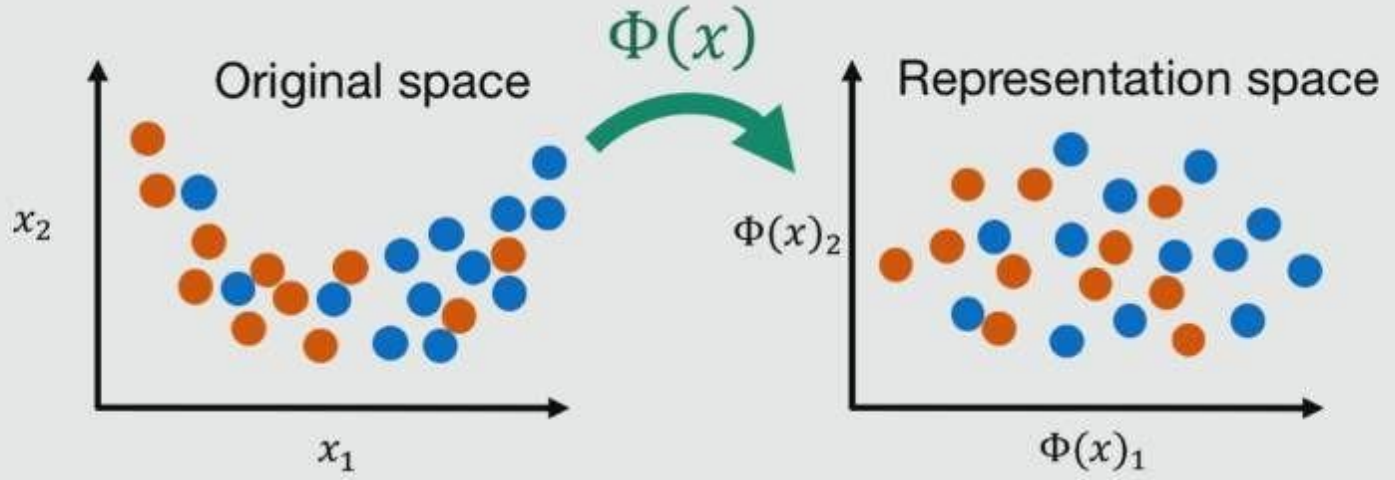
$$R_1(f_0) \leq R_0(f_0) + d_{\mathcal{H}}(p_0(X), p_1(X))$$

Counterfactual risk | **Factual risk** | **Distributional distance w.r.t. X**

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

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

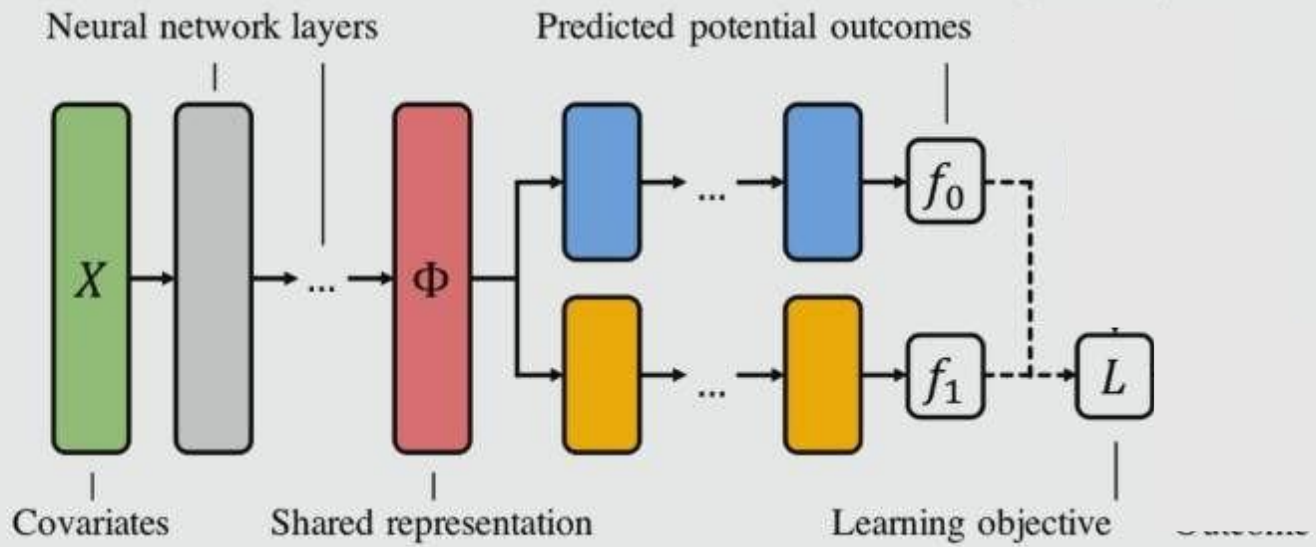
- **Approach 1:** Find a new, predictive space which exposes similarities
- $\text{minimize}_{f,\Phi} \; R_0(f_0) + d_{\mathcal{H}}\left(p_0\left(\Phi(X)\right), p_1\left(\Phi(x)\right)\right)$



[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
- $\text{minimize}_{f,\Phi} \; R_0(f_0) + d_{\mathcal{H}}\left(p_0\left(\Phi(X)\right), p_1\left(\Phi(x)\right)\right)$

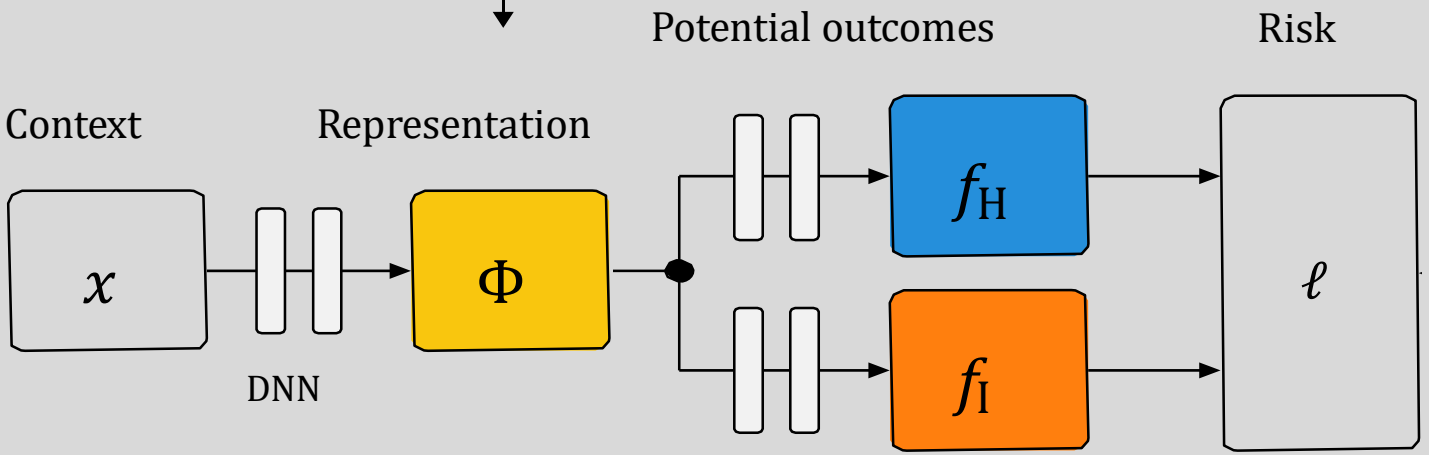


[2]Shalit, J., Sontag, *ICML* 2017

Deep learning architecture

Shared representation for shared statistical power between groups

Separate heads for different treatments to avoid washing away T



- This halved the error on a widely used causal effect benchmark!

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.	$3.8 \pm .2$	$.40 \pm .03$
BNN	$2.1 \pm .1$	$.42 \pm .03$
TARNET	$.95 \pm .02$	$.28 \pm .01$
CFR$_{\text{MMD}}$	$.78 \pm .02$	$.31 \pm .01$
CFR$_{\text{WASS}}$	$.76 \pm .02$	$.27 \pm .01$

Concatenating Φ and T — BNN

Twin-head neural net ($\alpha = 0$) — TARNET

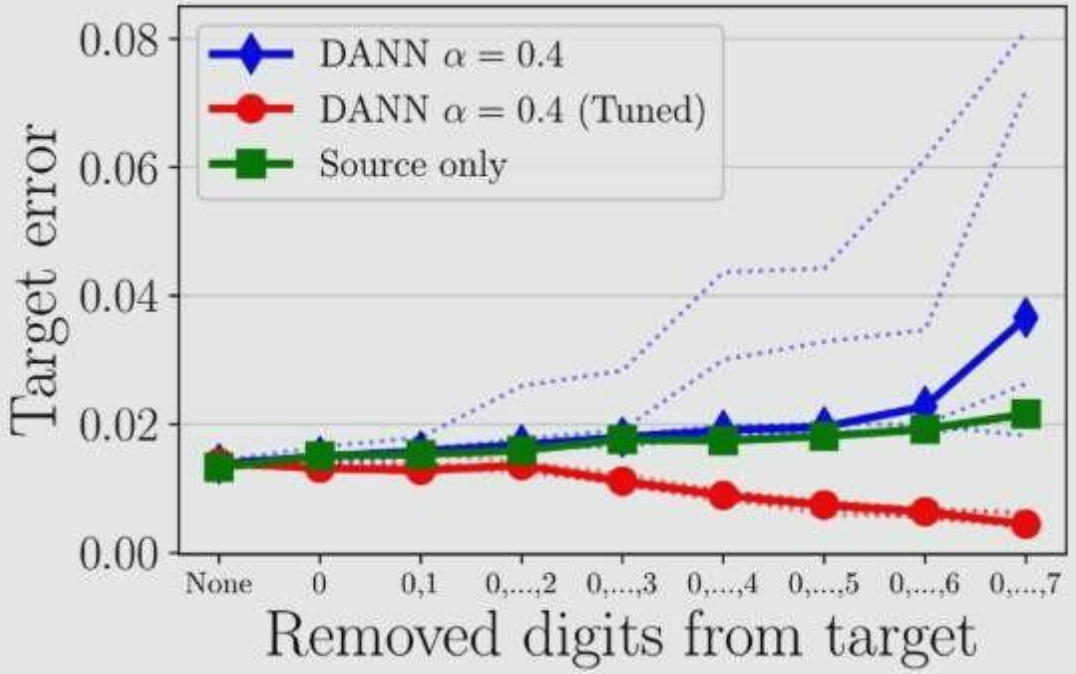
+ IPM regularization — CFR$_{\text{MMD}}$, CFR$_{\text{WASS}}$

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)**



J., Sontag, Ranganath

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

Conclusion

- **First error bound** for individual treatment effect (CATE) that holds under model misspecification
- Gives theoretical guidance for **how to change learning objective** (loss + regularization) when goal is causal inference

- Ongoing directions:
  - Generalizing to **multiple treatments**, continuous, etc.
  - Developing similar theory for sequential decision making (i.e., **off-policy RL**)
  - Algorithms for identifying **responders**
  - Causal effect variational autoencoders (Louizos et al., NIPS '17)

Thank You!