Prediction-powered Generalization of Causal Inferences

Appeared at ICML 2024

Ilker Demirel, Ahmed Alaa, Anthony Philippakis, David Sontag

Causal inference objective

Some notation

Categorical treatment A

Patient features X

Potential outcome under treatment a : Y^a

Observed outcome Y

Causal inference objective

Some notation

Categorical treatment A

Patient features X

Potential outcome under treatment a : Y^a

Observed outcome Y

Have sample $\mathcal{D} = \{X_i, A_i, Y_i\}_{i=1}^{n} \sim P_{X,A,Y}$

Want to estimate

$$\mathbb{E}[Y^a]$$

Causal inference challenges

Y^a is only observed when $A = a$ $(Y = Y^a)$

$$\mathbb{E}[Y^a] \stackrel{?}{=} \mathbb{E}[Y^a \mid A = a]$$

Estimate using

$$\frac{1}{n_a} \sum_{\mathscr{D}} Y_i \times \mathbb{I}(A_i = a)$$

Causal inference challenges

Y^a is only observed when $A = a$ $(Y = Y^a)$

$$\mathbb{E}[Y^a] \stackrel{?}{=} \mathbb{E}[Y^a \mid A = a]$$

Estimate using

$$\frac{1}{n_a} \sum_{\mathcal{D}} Y_i \times \mathbb{I}(A_i = a)$$

What if the above does not hold?

Randomized controlled trials (RCT) vs. observational data

Treatment is randomized in RCTs

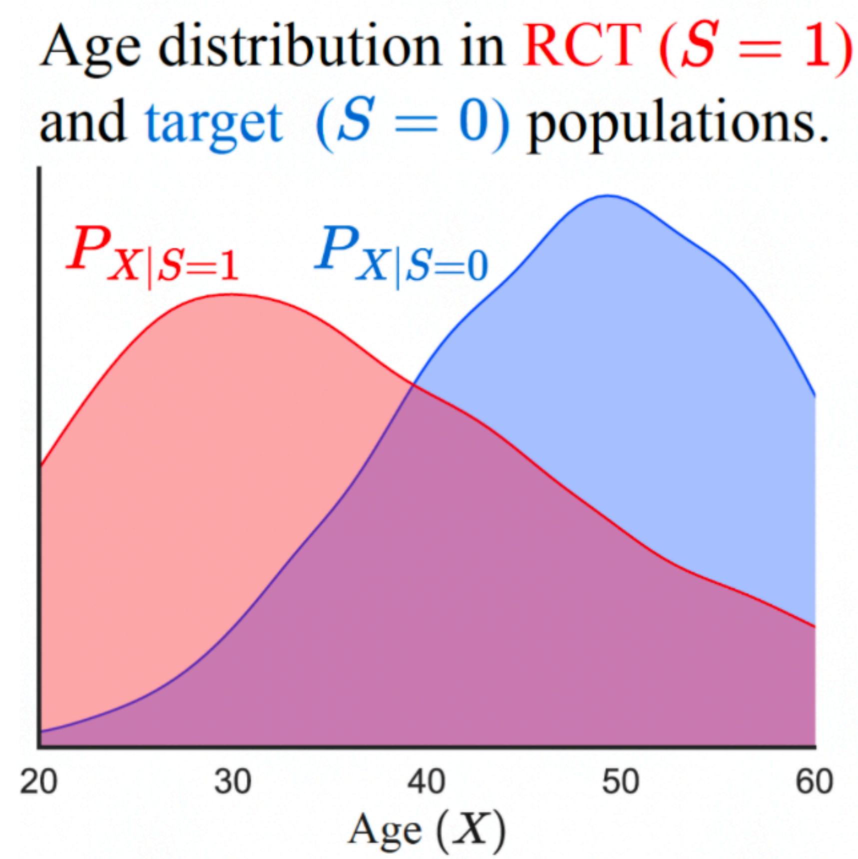
Not in the observational studies

Causal inference challenges

RCT results may not generalize

RCT-eligible super-population:

- $S = 1$ for trial participants
- $S = 0$ for non-participants



$$\mathbb{E}[Y^a \mid S = 1] \neq \mathbb{E}[Y^a \mid S = 0]$$

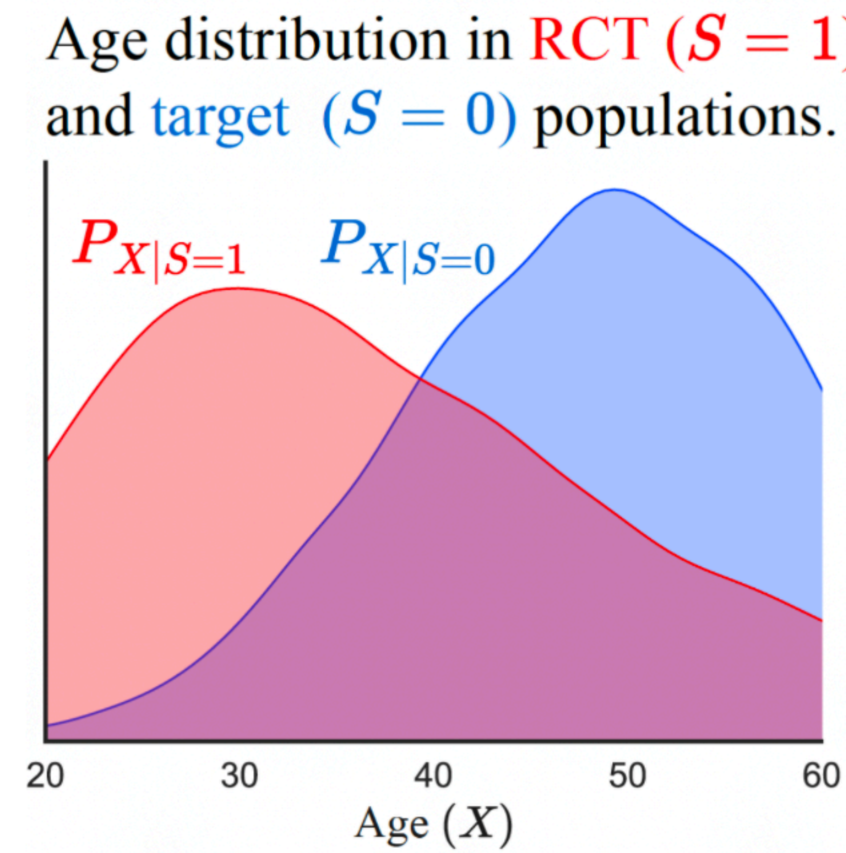
Causal inference challenges

RCT results may not generalize

RCT-eligible super-population:

$S = 1$ for trial participants

$S = 0$ for non-participants



$$\mathbb{E}[Y^a \mid S = 1] \neq \mathbb{E}[Y^a \mid S = 0]$$

Can we estimate $\mathbb{E}[Y^a \mid S = 0]$ using

$$\mathcal{D}_0 = \{X_i\}_{i=1}^N \sim P_{X|S=0}$$

$$\mathcal{D}_1 = \{X_i, A_i, Y_i\}_{i=1}^n \sim P_{X,A,Y|S=1}$$

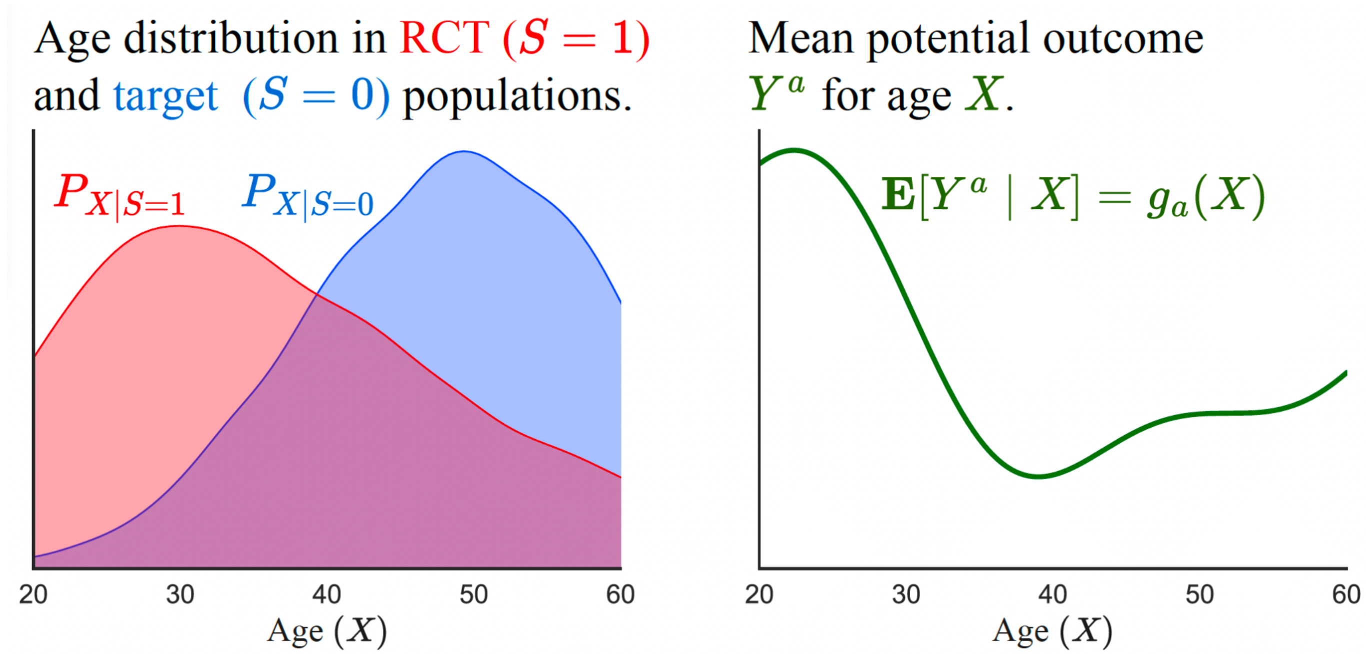
Generalizing from an RCT to a target population

$$\mathbb{E}[Y^a \mid S = 0]$$

$$= \mathbb{E}_{P_{X|S=0}}\left[\mathbb{E}[Y^a \mid X, S = 0]\right]$$

$$= \mathbb{E}_{P_{X|S=0}}\left[\mathbb{E}[Y^a \mid X, S = 1]\right]$$

$$= \mathbb{E}_{P_{X|S=0}}\left[\mathbb{E}[Y \mid X, S = 1, A = a]\right]$$



RCT cohort is **younger** than target cohort.

Outcome of interest, Y^a, is **larger** for **younger**.

Mean outcome in RCT is **larger** than in the target population.

$$\mathbf{E}[Y^a \mid S = 1] > \mathbf{E}[Y^a \mid S = 0]$$

A covariate shift problem

Generalizing from an RCT to a target population

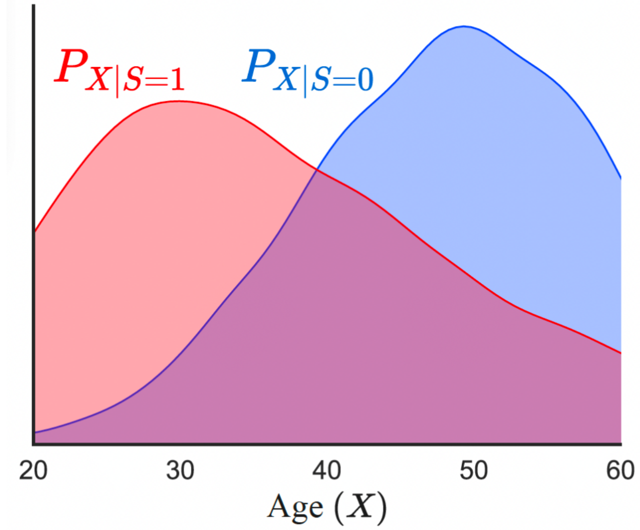
$$\mathbb{E}[Y^a \mid S=0]$$

$$= \mathbb{E}_{P_{X|S=0}}\left[\mathbb{E}[Y^a \mid X, S=0]\right]$$

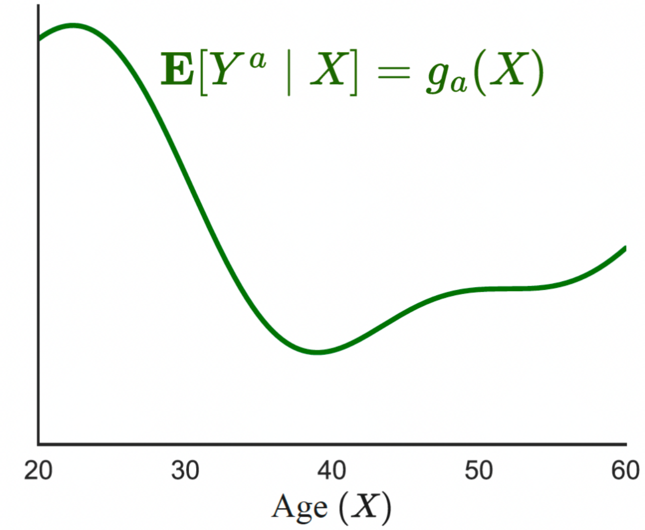
$$= \mathbb{E}_{P_{X|S=0}}\left[\mathbb{E}[Y^a \mid X, S=1]\right]$$

$$= \mathbb{E}_{P_{X|S=0}}\left[\mathbb{E}[Y \mid X, S=1, A=a]\right]$$

Age distribution in RCT ($S=1$) and target ($S=0$) populations.



Mean potential outcome Y^a for age X.



RCT cohort is **younger** than target cohort.

Outcome of interest, Y^a, is **larger** for **younger**.

Mean outcome in RCT is **larger** than in the target population.

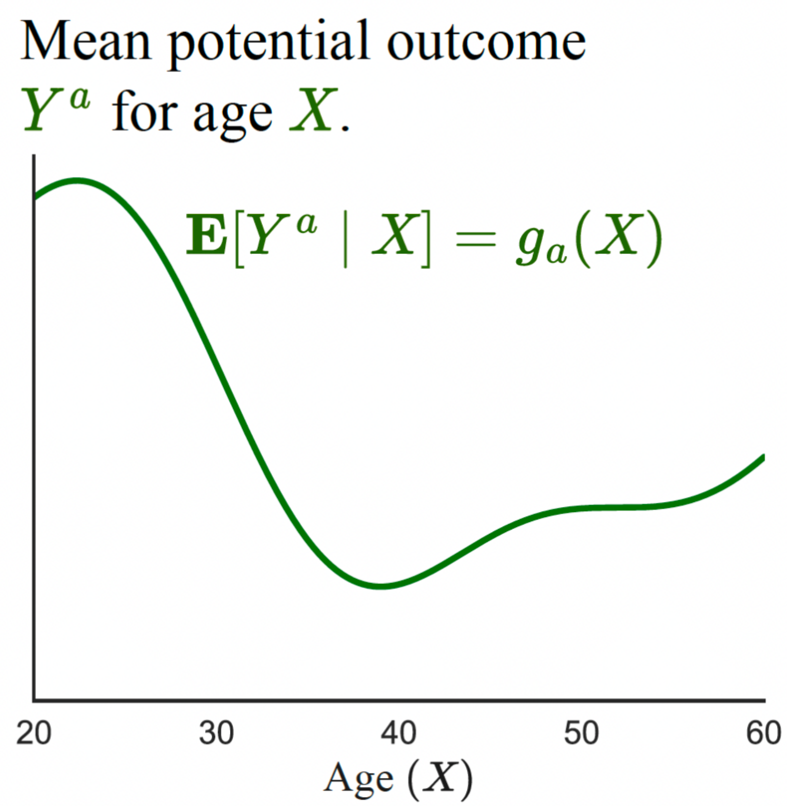
$$\mathbf{E}[Y^a \mid S=1] > \mathbf{E}[Y^a \mid S=0]$$

A covariate shift problem

$$\hat{\mu}_a^{\text{OM}} = \frac{1}{N}\sum_{\mathcal{D}_{S=0}} \hat{g}_a(X_i)$$

Fit $\hat{g}_a(X)$ using RCT data, use it in the target sample

Generalizing from an RCT to a target population



Fit $\hat{g}_a(X)$ using RCT data, use it in the target sample

$$\hat{\mu}_a^{\text{OM}} = \frac{1}{N} \sum_{\mathscr{D}_{S=0}} g_a(X_i, \hat{\theta})$$

$$\mathbf{E}[(\hat{\mu}_a^{\text{OM}} - \mu_a)^2]$$

$$\approx \mathbf{E}_{X \sim P_0} \Big[\underbrace{\mathbf{E}_{\hat{\theta} \sim \mathcal{A}(P_1)} [g_a(X; \hat{\theta})] - g_a(X)}_{=: \, \text{SB}_g(X)} \Big]^2 \tag{6}$$

$$+ \text{Var}_{\hat{\theta} \sim \mathcal{A}(P_1)} \big(\mathbf{E}_{X \sim P_0} [g_a(X; \hat{\theta})] \big). \tag{7}$$

Generalizing from an RCT to a target population

$$\mathbf{E}[(\hat{\mu}_a^{\mathrm{OM}} - \mu_a)^2]$$

$$\approx \mathbf{E}_{X \sim P_0}\Big[\underbrace{\mathbf{E}_{\hat{\theta} \sim \mathcal{A}(P_1)}[g_a(X;\hat{\theta})] - g_a(X)}_{=:\ \mathrm{SB}_g(X)}\Big]^2 \quad (6)$$

$$+ \operatorname{Var}_{\hat{\theta} \sim \mathcal{A}(P_1)}\big(\mathbf{E}_{X \sim P_0}[g_a(X;\hat{\theta})]\big). \quad (7)$$

When is the generalization MSE is large?

Small sample size of RCTs make this task statistically infeasible

Generalizing from an RCT to a target population

$$\mathbf{E}[(\hat{\mu}_a^{\mathrm{OM}} - \mu_a)^2]$$
$$\approx \mathbf{E}_{X\sim P_0}\Big[\underbrace{\mathbf{E}_{\hat{\theta}\sim\mathcal{A}(P_1)}[g_a(X;\hat{\theta})] - g_a(X)}_{=:\ \mathrm{SB}_g(X)}\Big]^2 \quad (6)$$
$$+ \mathrm{Var}_{\hat{\theta}\sim\mathcal{A}(P_1)}\big(\mathbf{E}_{X\sim P_0}[g_a(X;\hat{\theta})]\big). \quad (7)$$

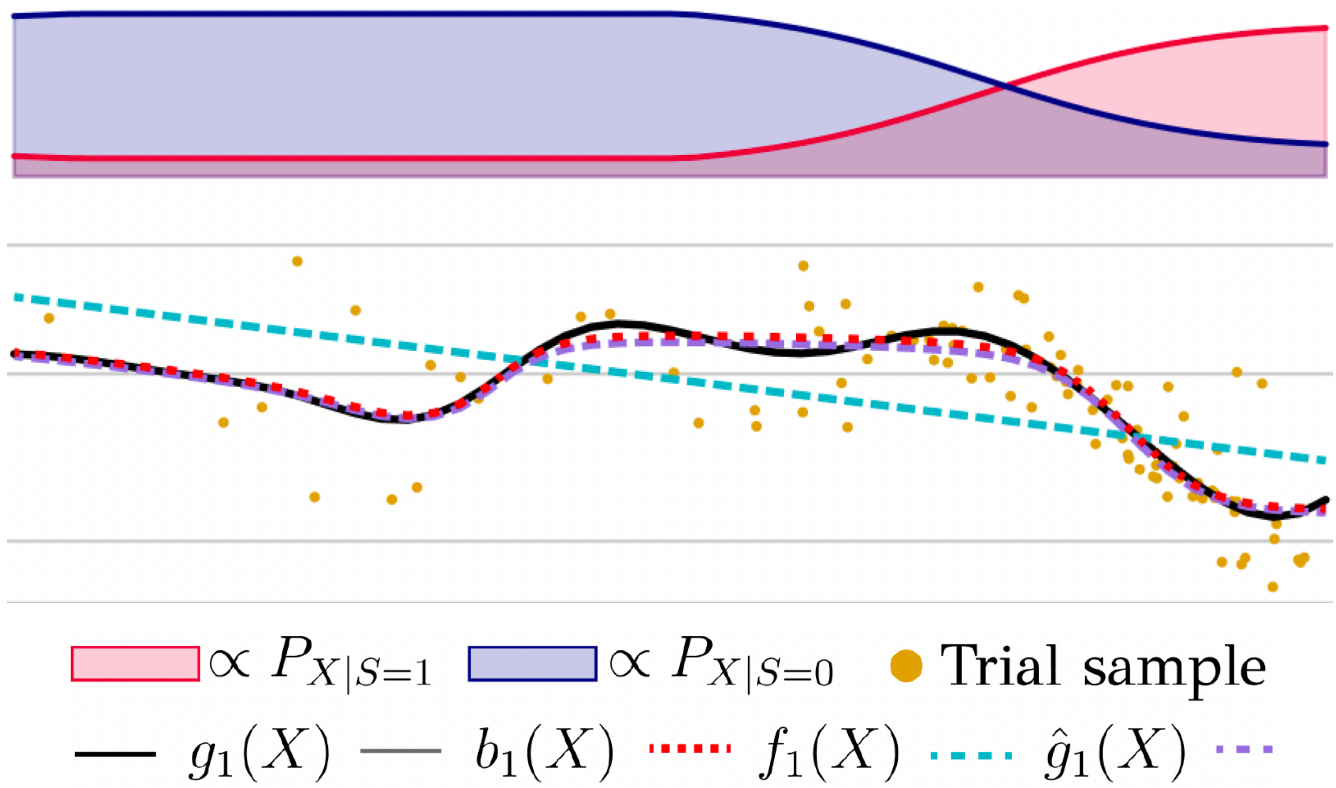
When is the generalization MSE is large?

Small sample size of RCTs make this task statistically infeasible

Consider a "complex" $g_a(X)$

Small model
- Bias (underfit)
- Hurts more due to P_X shift

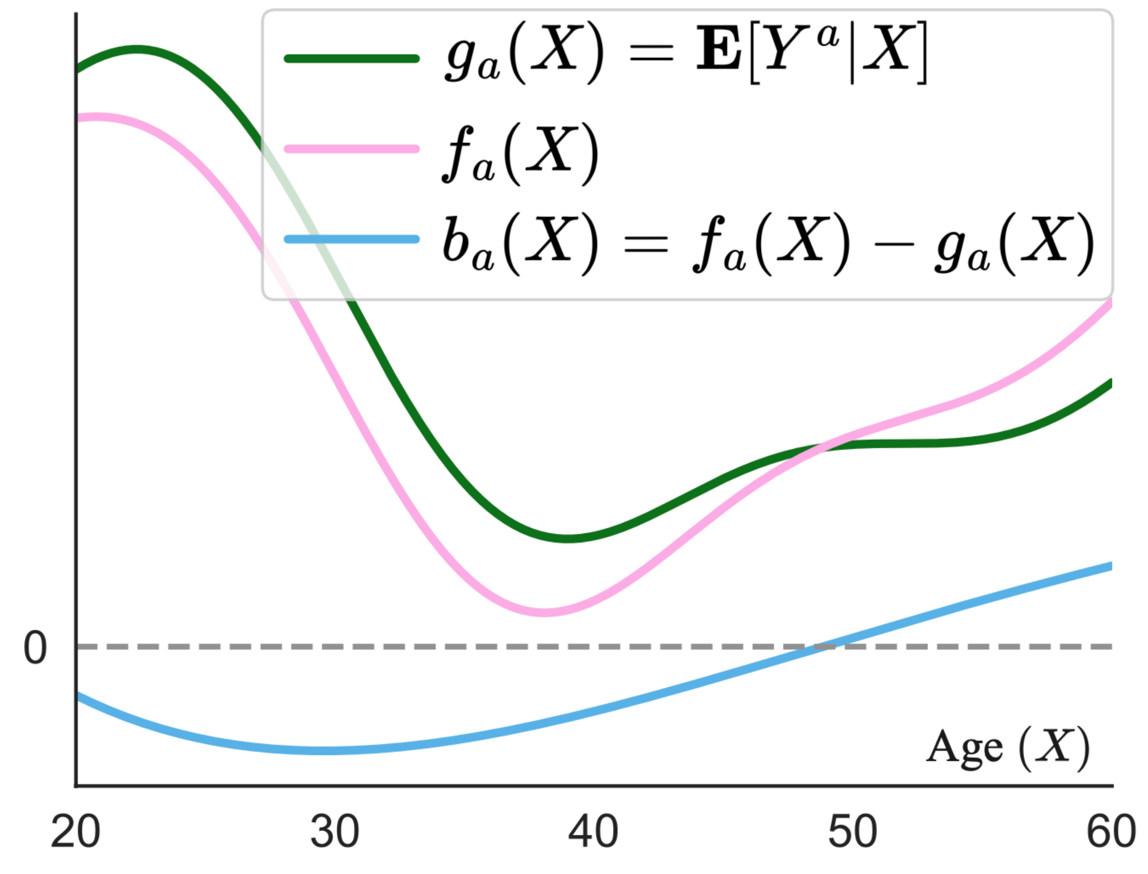
Large model
- Overfit to RCT sample
- High variance



Leveraging observational data

Large-scale & rich observational data

- Big sample size, can support complex models
- Electronic health records (EHRs), Insurance claims



Leveraging observational data

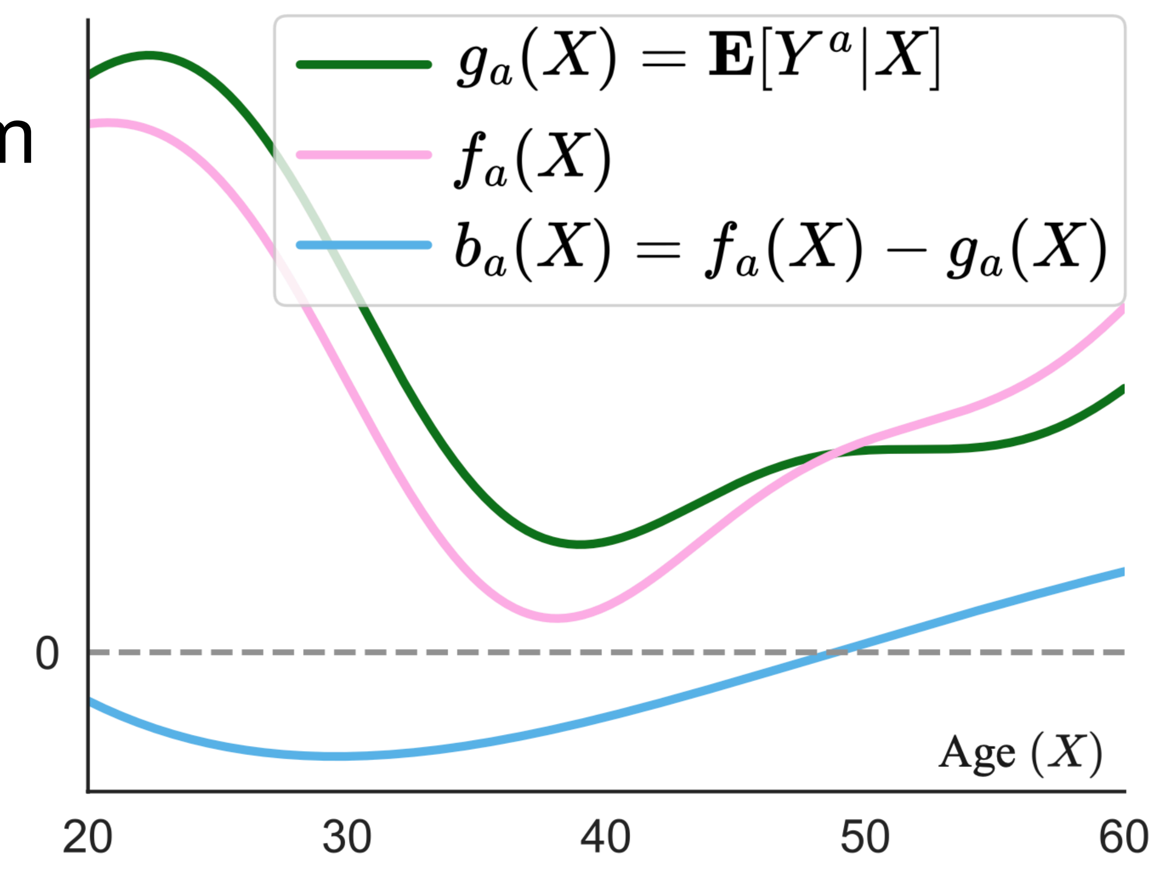
Large-scale & rich observational data

- Big sample size, can support complex models
- Electronic health records (EHRs), Insurance claims

Fit $g_a(X)$ using observational data?

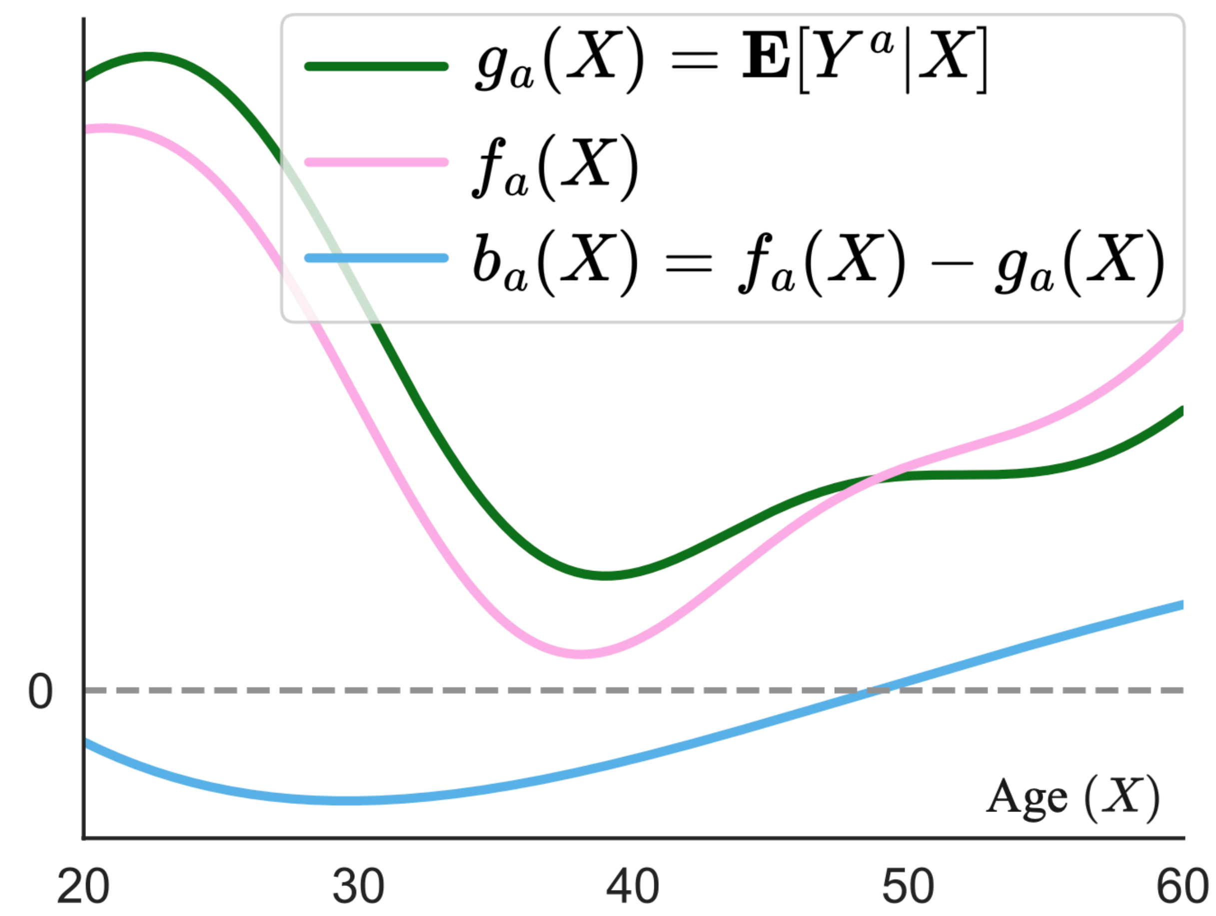
Not reliable for causal inference

- Treatment assignments are not random
- E.g., unmeasured confounding



Additive bias correction (ABC)

Observational data may still capture a lot of information



Additive bias correction (ABC)

Observational data may still capture a lot of information

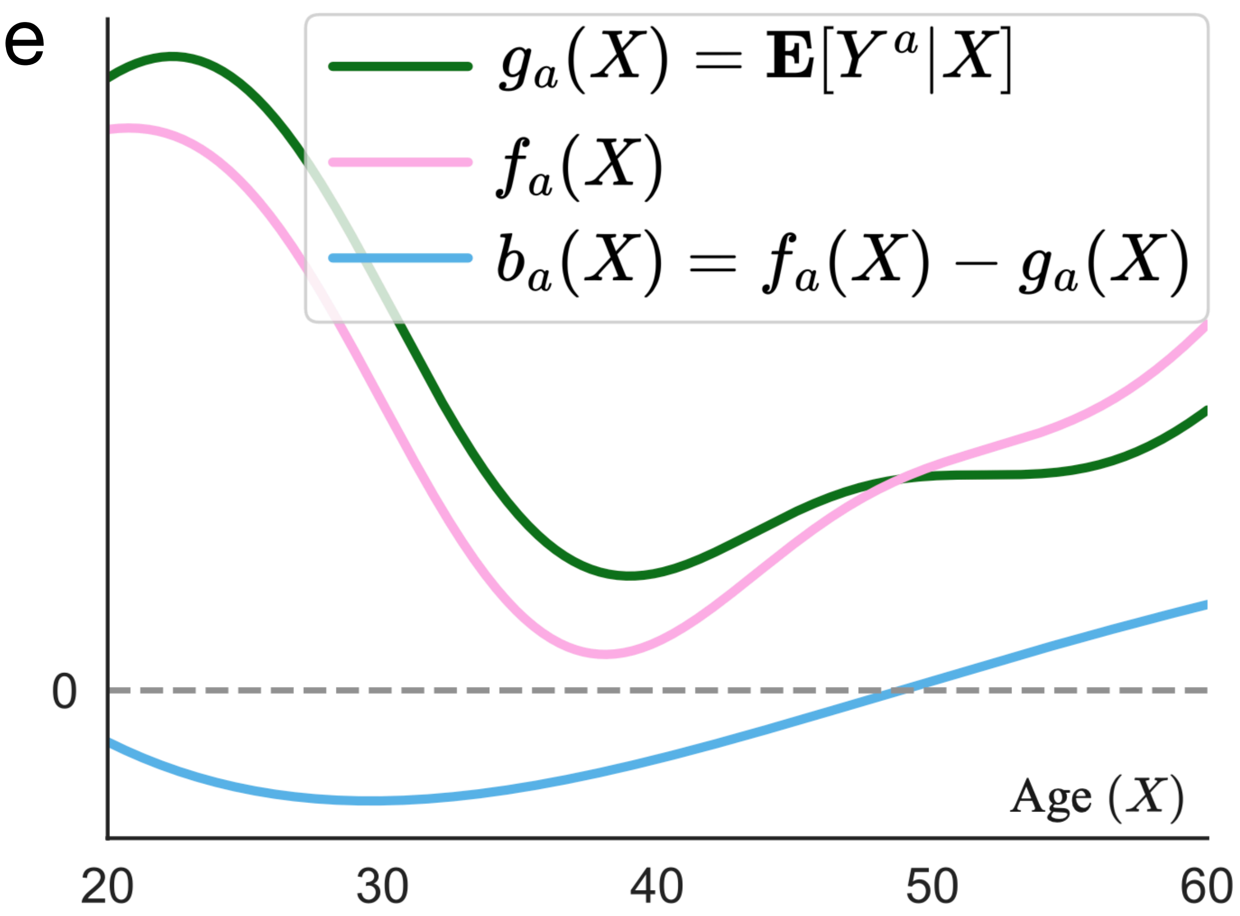
Integrating it with RCT data alleviates the

"causal reliability" and

"statistical infeasibility"

issues

$$\hat{\mu}_a^{\text{ABC}} = \frac{1}{N} \sum_{\mathscr{D}_{S=0}} f_a(X_i) - b_a(X_i, \hat{\gamma})$$



Additive bias correction (ABC)

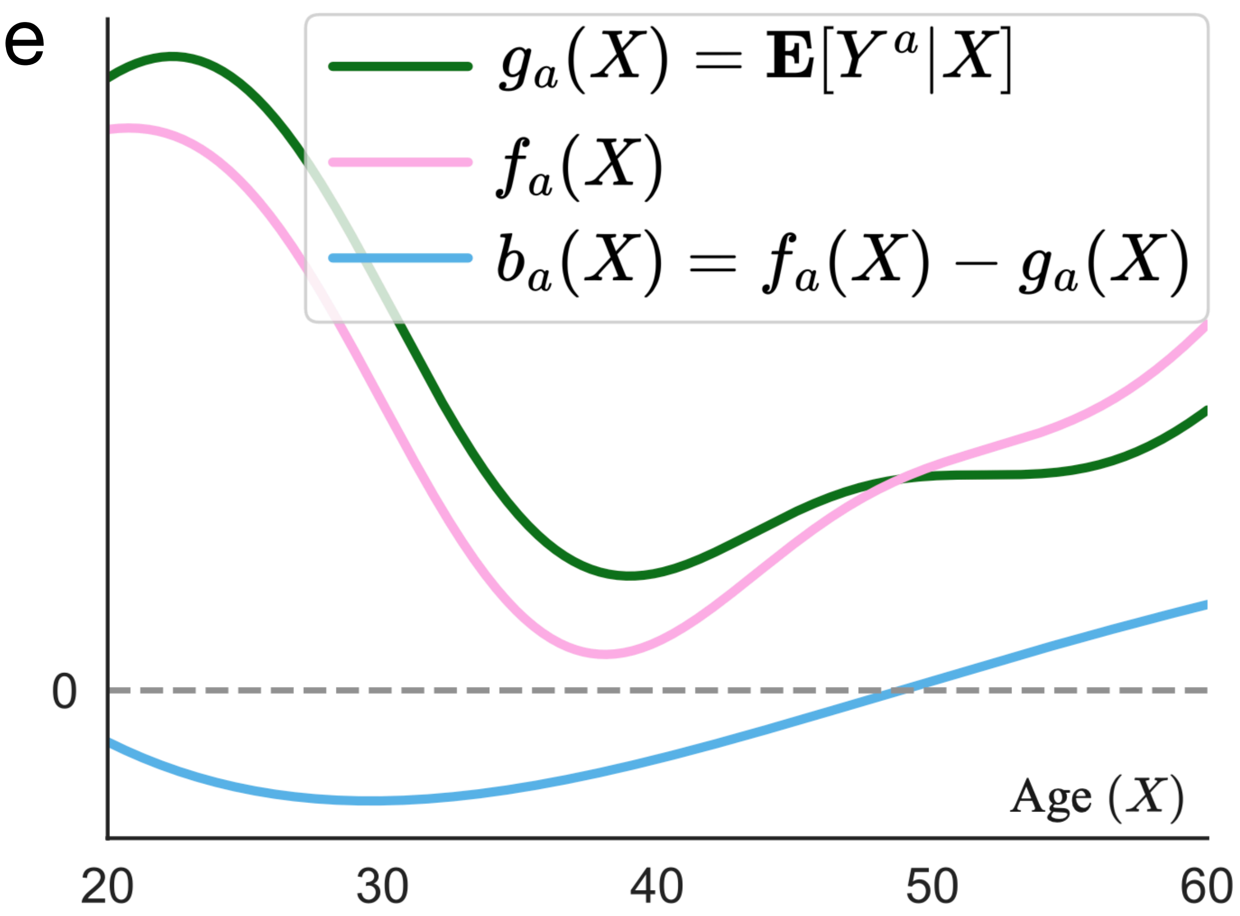
Observational data may still capture a lot of information

Integrating it with RCT data alleviates the

"causal reliability" and

"statistical infeasibility"

issues



$$\hat{\mu}_a^{\text{ABC}} = \frac{1}{N} \sum_{\mathscr{D}_{S=0}} f_a(X_i) - b_a(X_i, \hat{\gamma})$$

Instead of $g_a(X)$, learn $b_a(X)$ using the RCT data!

$$\begin{aligned} &\mathbf{E}[(\hat{\mu}_a^{\text{ABC}} - \mu_a)^2] \\ &\approx \mathbf{E}_{X\sim P_0}\left[\mathbf{E}_{\hat{\gamma}\sim\mathcal{A}(P_1)}[b_a(X;\hat{\gamma})] - b_a(X)\right]^2 \\ &\quad + \text{Var}_{\hat{\gamma}\sim\mathcal{A}(P_1)}\left(\mathbf{E}_{X\sim P_0}[b_a(X;\hat{\gamma})]\right). \end{aligned}$$

Augmented outcome modeling (AOM)

What if the bias function is harder to learn?

Imagine $f(X) \sim \mathcal{N}(0,1)$, then $b(X) = g(X) + \mathcal{N}(0,1)$

Augmented outcome modeling (AOM)

What if the bias function is harder to learn?

Imagine $f(X) \sim \mathcal{N}(0,1)$, then $b(X) = g(X) + \mathcal{N}(0,1)$

Use $f(X)$ as an additional feature to fit $g(X)$

$$g(X) = h(X, f(X))$$

If $f(X)$ is "useless," a good learning algo. would ignore it.

e.g., LASSO regression

Augmented outcome modeling (AOM)

What if the bias function is harder to learn?

Imagine $f(X) \sim \mathcal{N}(0,1)$, then $b(X) = g(X) + \mathcal{N}(0,1)$

Use $f(X)$ as an additional feature to fit $g(X)$

$g(X) = h(X, f(X))$

If $f(X)$ is "useless," a good learning algo. would ignore it.

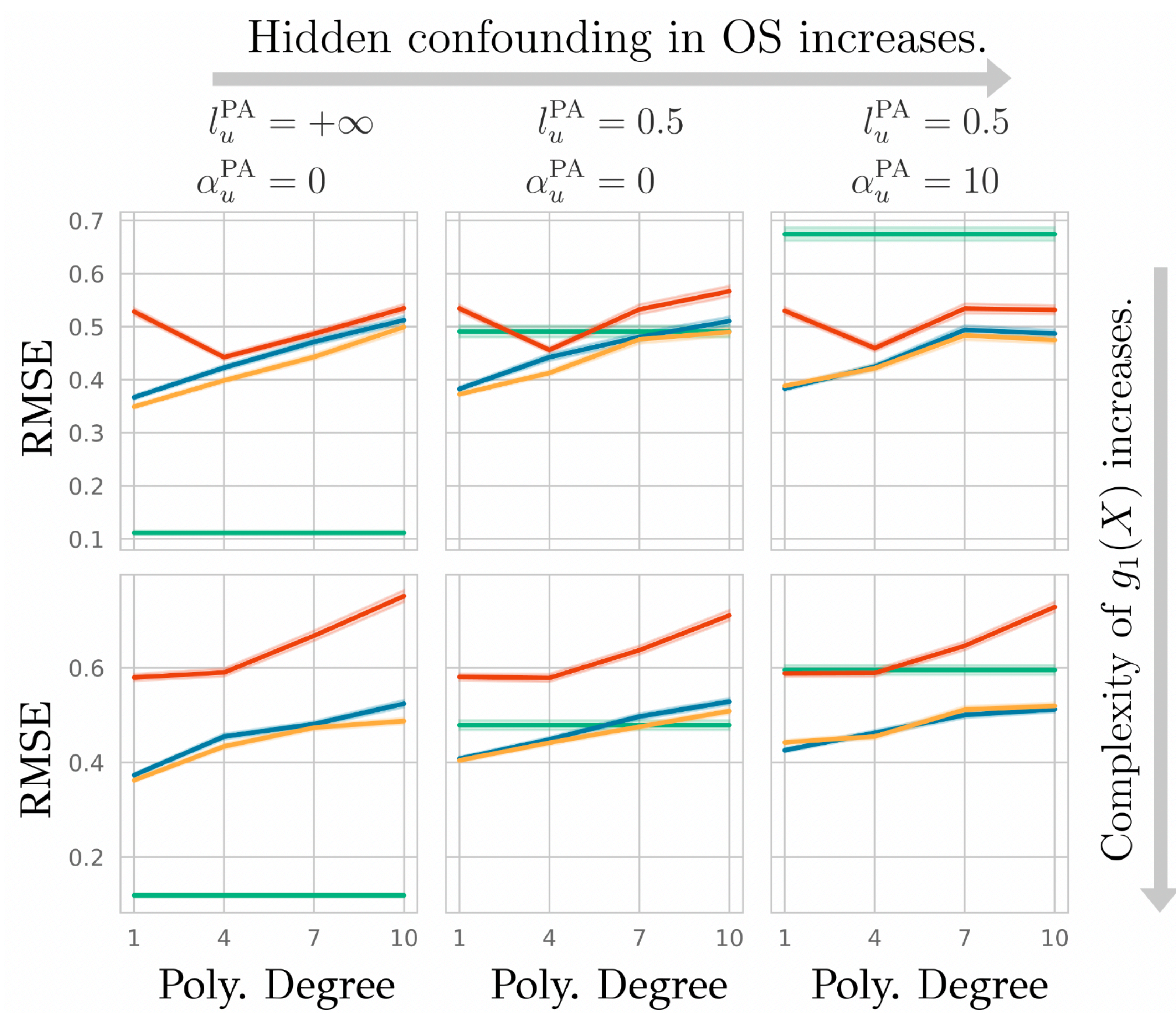
e.g., LASSO regression

If it is useful, fitting $h(X)$ maybe significantly easier than $g(X)$

e.g. $f(X) \approx g(X)$

similar to fine-tuning in deep learning

More helpful as the underlying model becomes more complex



AOM approach remains robust when the observational study is biased