Regularizing towards Causal Invariance: Linear Models with Proxies

Mike Oberst
MIT

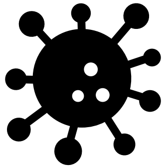
Nikolaj Thams
Univ. of Copenhagen

Jonas Peters
Univ. of Copenhagen

David Sontag
MIT

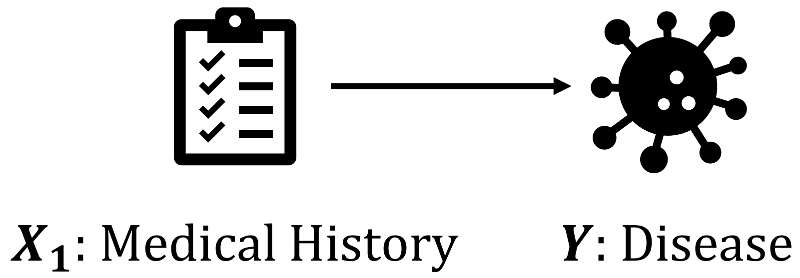
@MichaelOberst

Motivation: Robustness to Dataset Shift

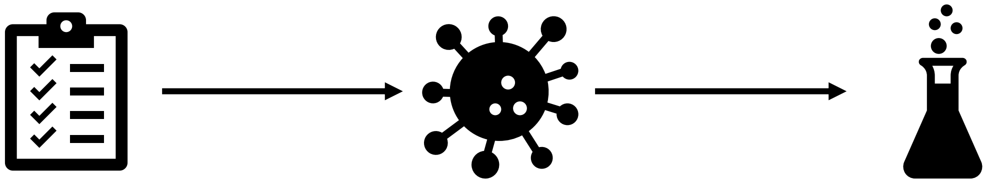


$\boldsymbol{Y}$: Disease

Motivation: Robustness to Dataset Shift

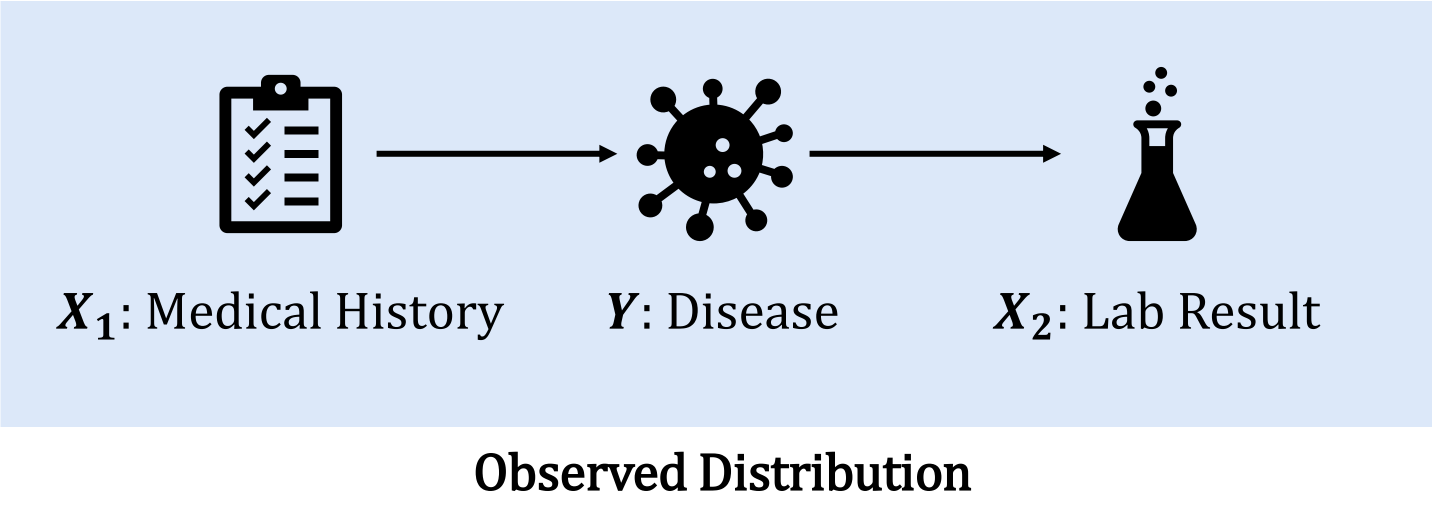


Motivation: Robustness to Dataset Shift

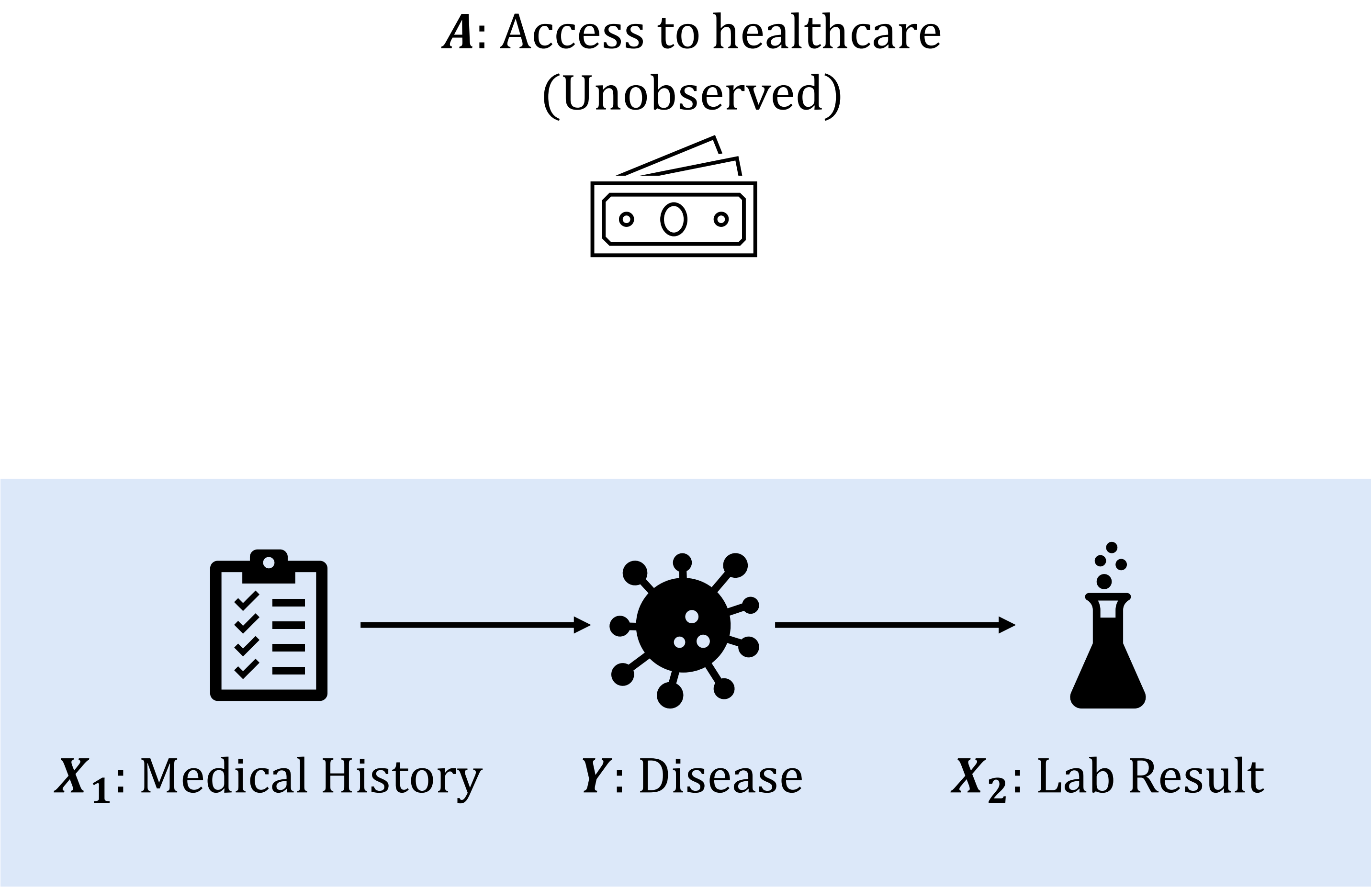


X_1: Medical History Y: Disease X_2: Lab Result

Motivation: Robustness to Dataset Shift

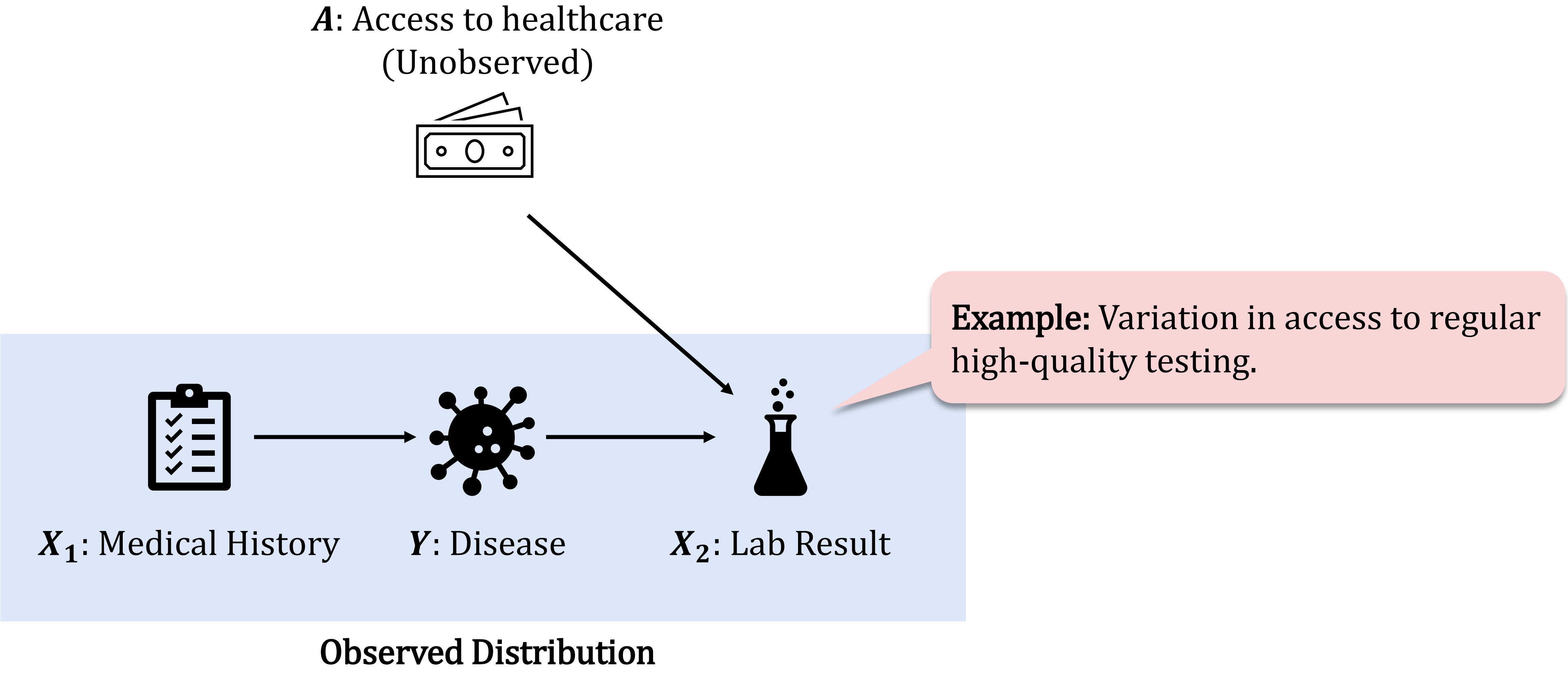


Motivation: Robustness to Dataset Shift

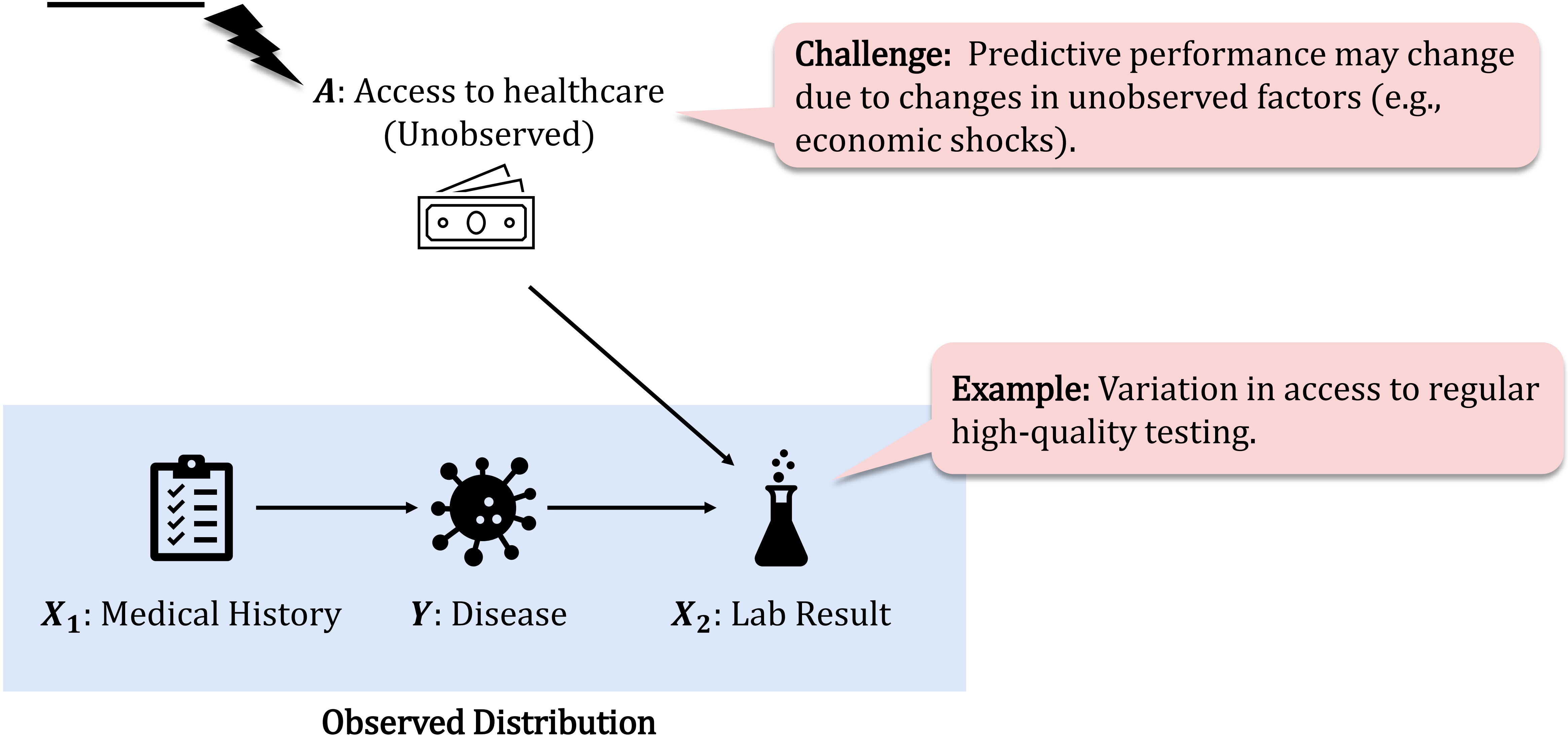


Observed Distribution

Motivation: Robustness to Dataset Shift

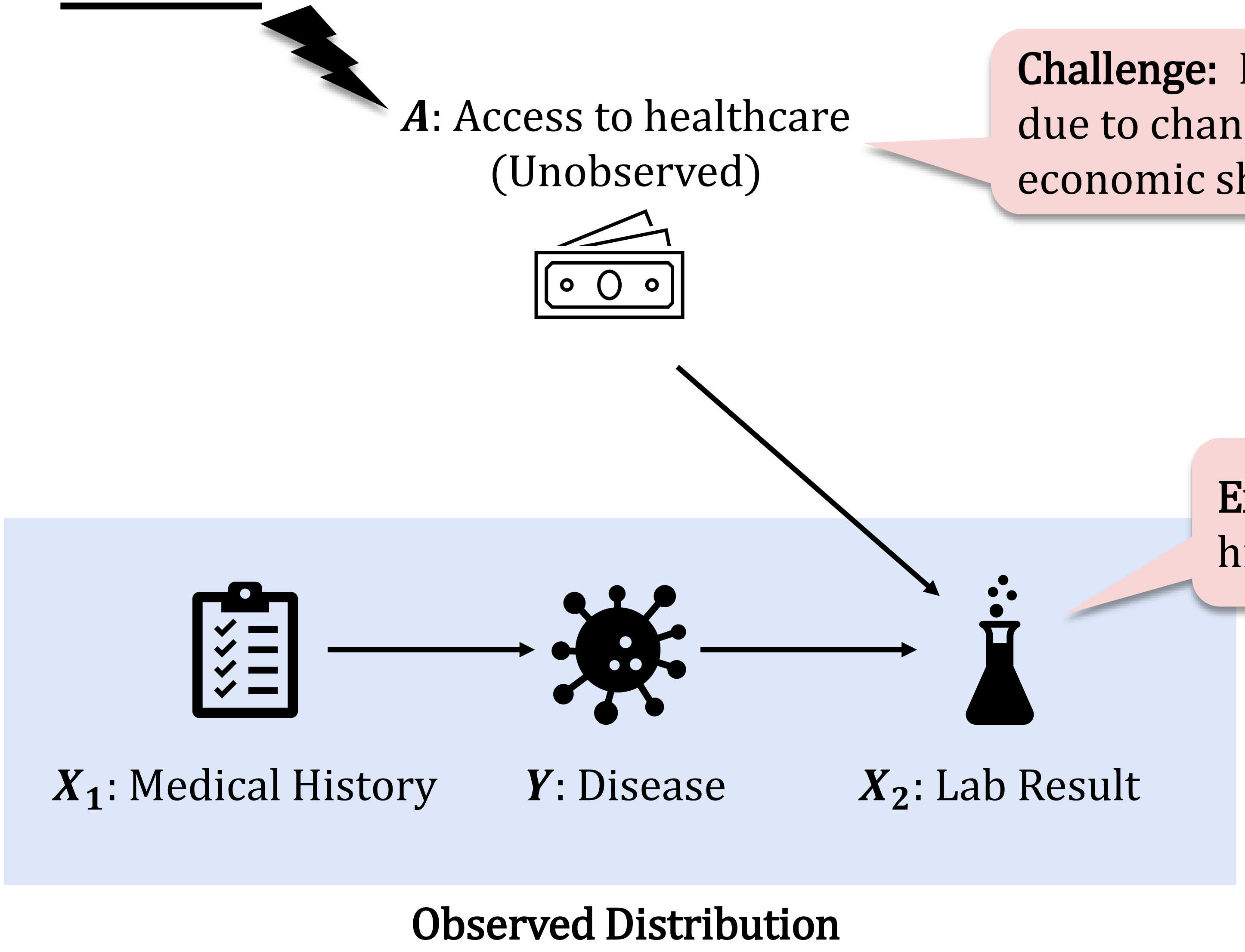


Motivation: Robustness to Dataset Shift



Motivation: Robustness to Dataset Shift

Challenge: Predictive performance may change due to changes in unobserved factors (e.g., economic shocks).



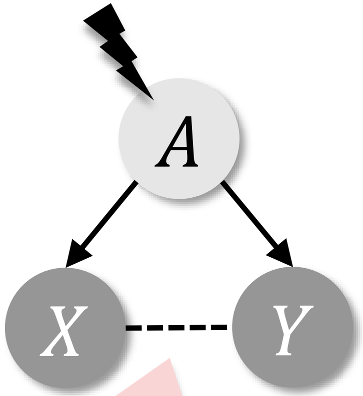
Example: Variation in access to regular high-quality testing.

1. Protect against shift in unobserved factors?
2. Balance between accuracy & robustness?

Motivation: Robustness to Dataset Shift

Minimize worst-case loss over a set of interventions

Interventions on A change the distribution of P(X, Y)



$$\min \sup_{\nu \in C} E_{do(A:=\nu)}[(Y - \gamma^\top X)^2]$$

Unknown causal graph between X, Y

Assumptions

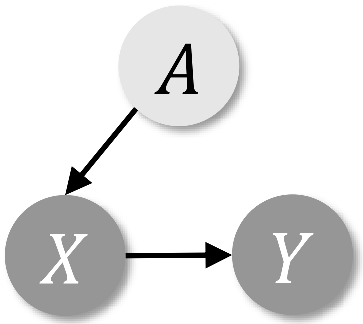
Linear structural causal model (SCM) over all observed and unobserved variables, and one or more **noisy proxies of A**

Assumptions

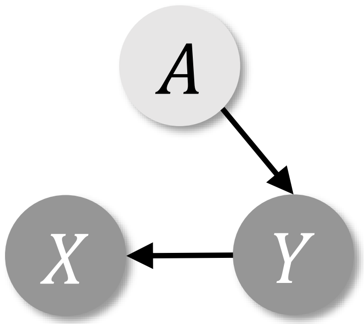
Linear structural causal model (SCM) over all observed and unobserved variables, and one or more **noisy proxies of A**

Any causal graph over X, Y, H is permitted, but A is an "anchor" with no causal parents.

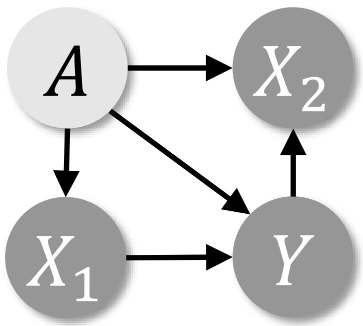
$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} := B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + M_A A + \epsilon$$



Covariate Shift



Label Shift



...and more

Assumptions

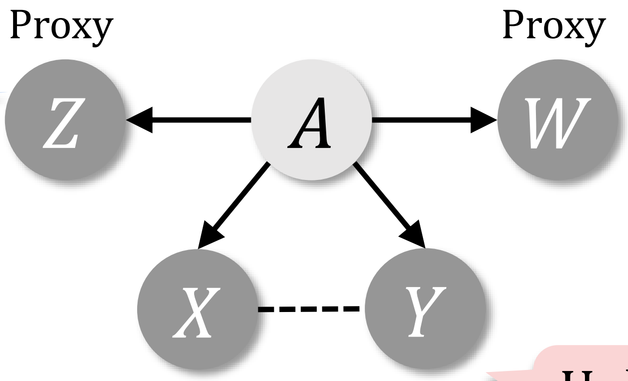
Linear structural causal model (SCM) over all observed and unobserved variables, and one or more **noisy proxies of A**

Any causal graph over X, Y, H is permitted, but A is an "anchor" with no causal parents.

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} := B \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + M_A A + \epsilon$$

Proxies are linear functions of A with independent additive noise.

Example: Self-reported data on income, distance to closest clinic, etc.



Unknown causal graph between X, Y

Motivation: Robustness to Dataset Shift

Minimize worst-case loss over a set of interventions

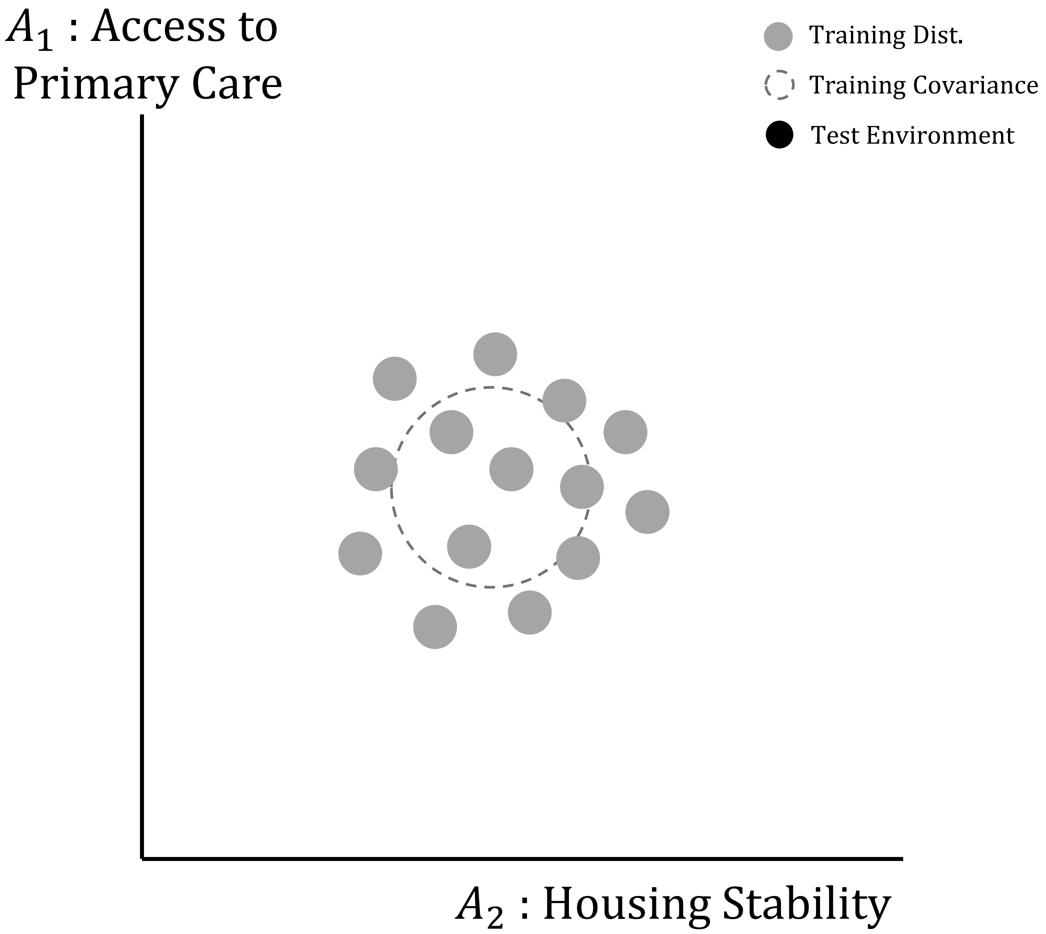
Interventions on A change the distribution of P(X, Y)

$$\min \sup_{\nu \in C} E_{do(A:=\nu)}[(Y - \gamma^\top X)^2]$$

Using prior knowledge:

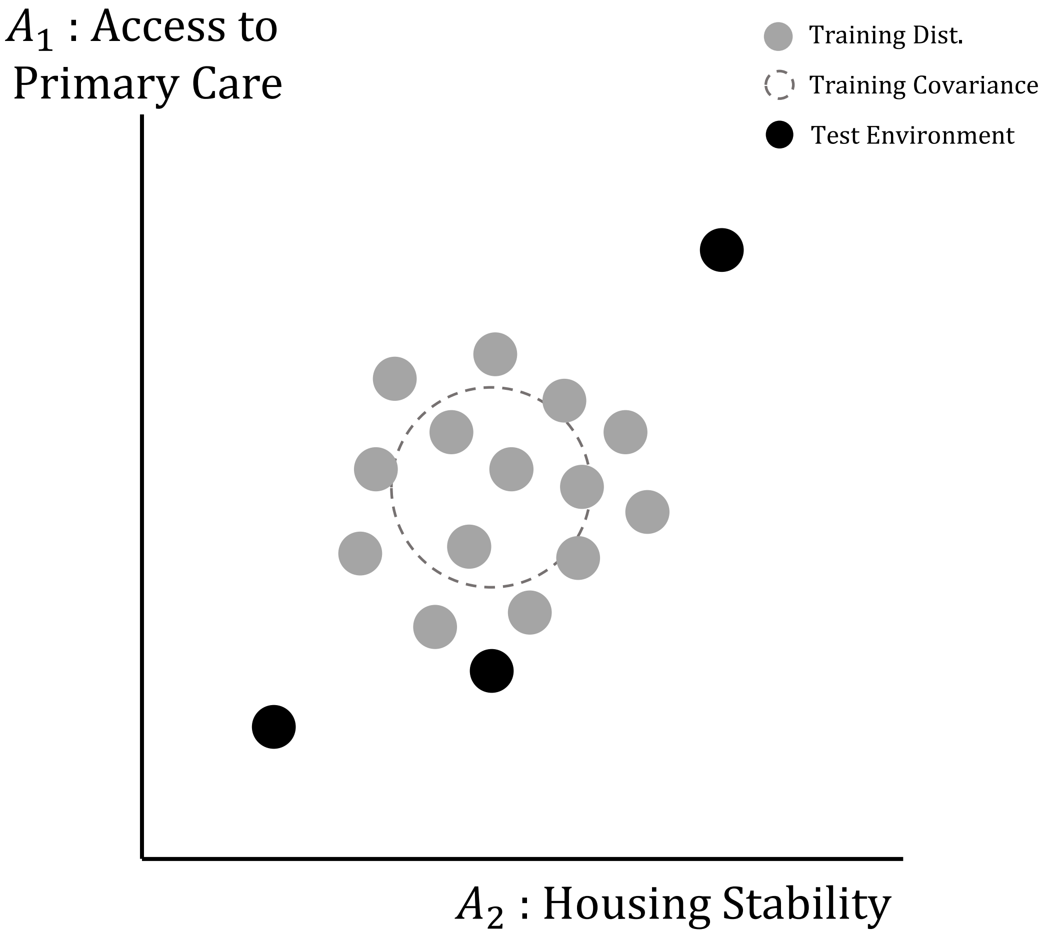
1. Specify **relevant factors of variation** A via proxies.
2. Specify **plausible shifts** via robustness set C.

Robustness to bounded interventions on A



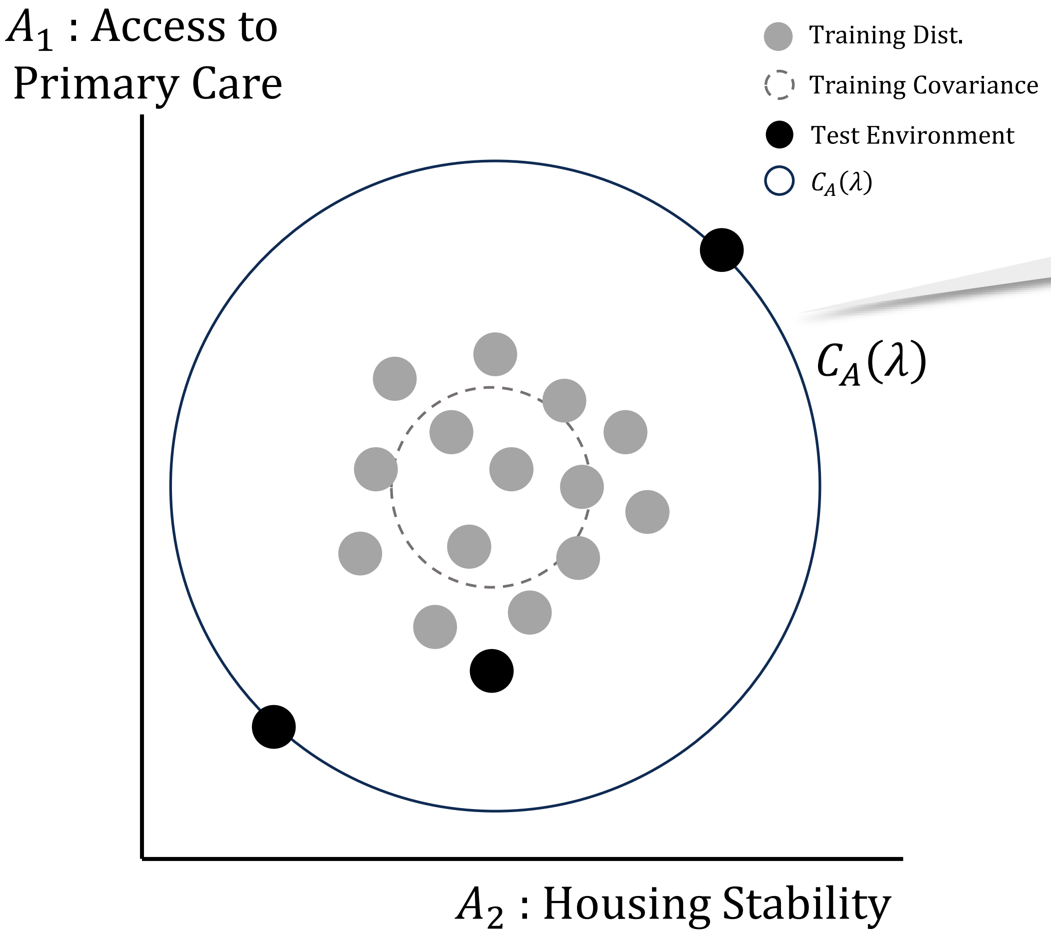
X, Y not shown here, just the dimensions of A

Robustness to bounded interventions on A



X, Y not shown here, just the dimensions of A

Robustness to bounded interventions on A



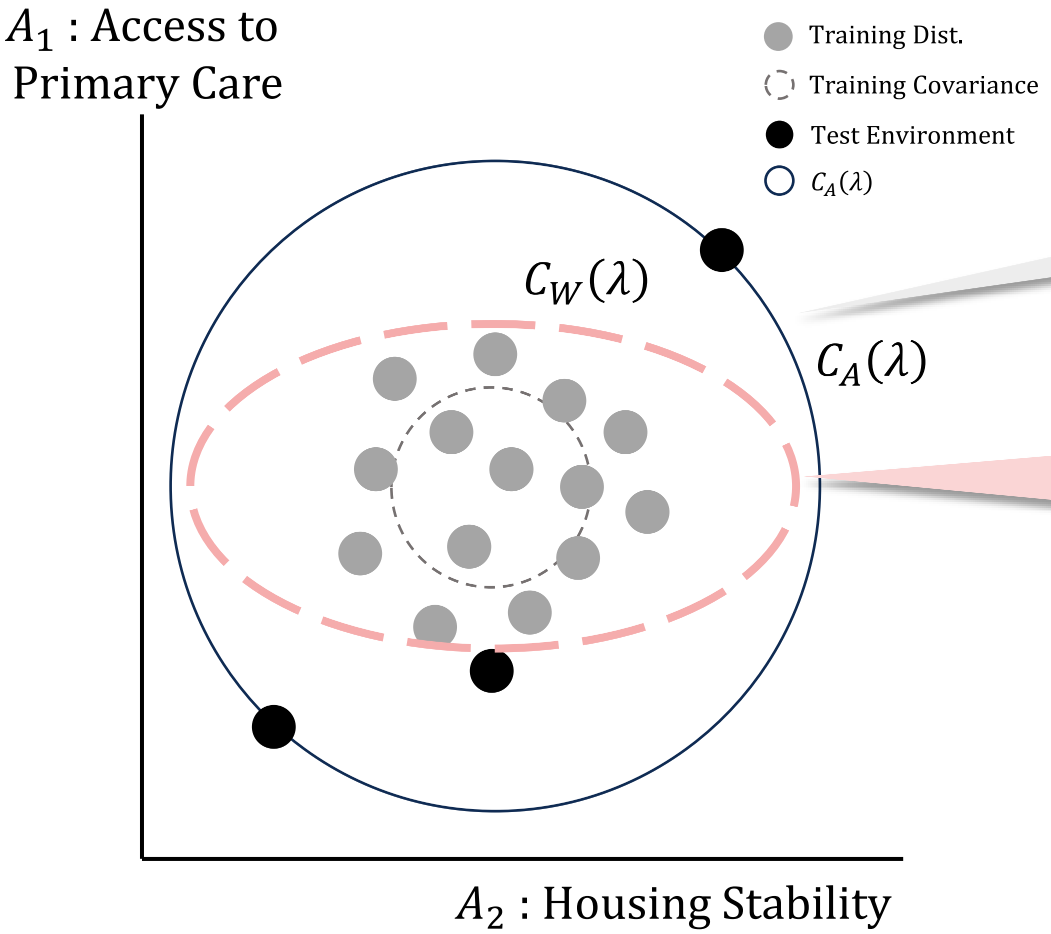
Anchor Regression [1] assumes that A is observed, and optimizes a worst-case loss over bounded interventions

$$\sup_{\nu \in C_A(\lambda)} E_{do(A := \nu)}[(Y - \gamma^\top X)^2]$$

X, Y not shown here, just the dimensions of A

[1] Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. Anchor regression: Heterogeneous data meet causality. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 83(2):215–246, 2021.

Robustness to bounded interventions on A



Anchor Regression [1] assumes that A is observed, and optimizes a worst-case loss over bounded interventions

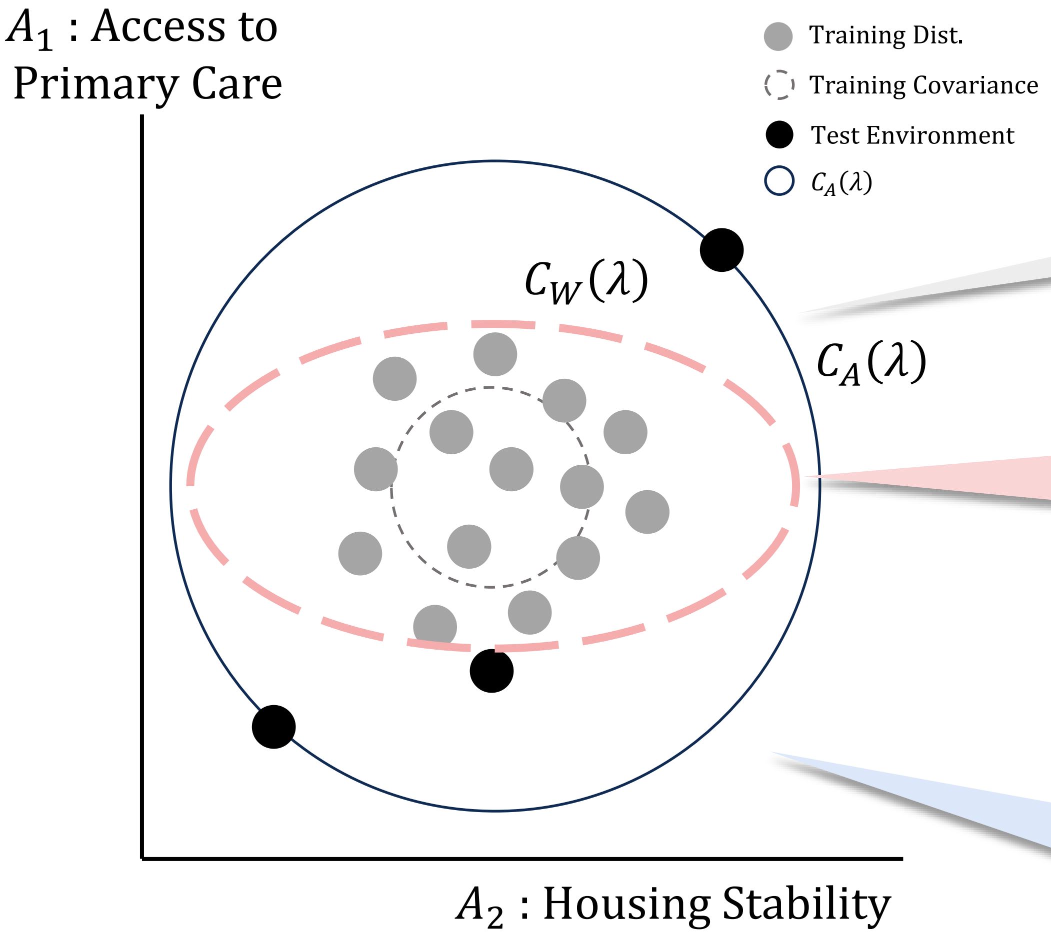
$$\sup_{\nu \in C_A(\lambda)} E_{do(A := \nu)}[(Y - \gamma^\top X)^2]$$

Theorem 1 (Informal)
Given a single noisy proxy W of A, the robustness set is provably reduced, and this reduction is not identifiable.

X, Y not shown here, just the dimensions of A

[1] Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. Anchor regression: Heterogeneous data meet causality. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 83(2):215–246, 2021.

Robustness to bounded interventions on A



Anchor Regression [1] assumes that A is observed, and optimizes a worst-case loss over bounded interventions

$$\sup_{\nu \in C_A(\lambda)} E_{do(A:=\nu)}[(Y - \gamma^\top X)^2]$$

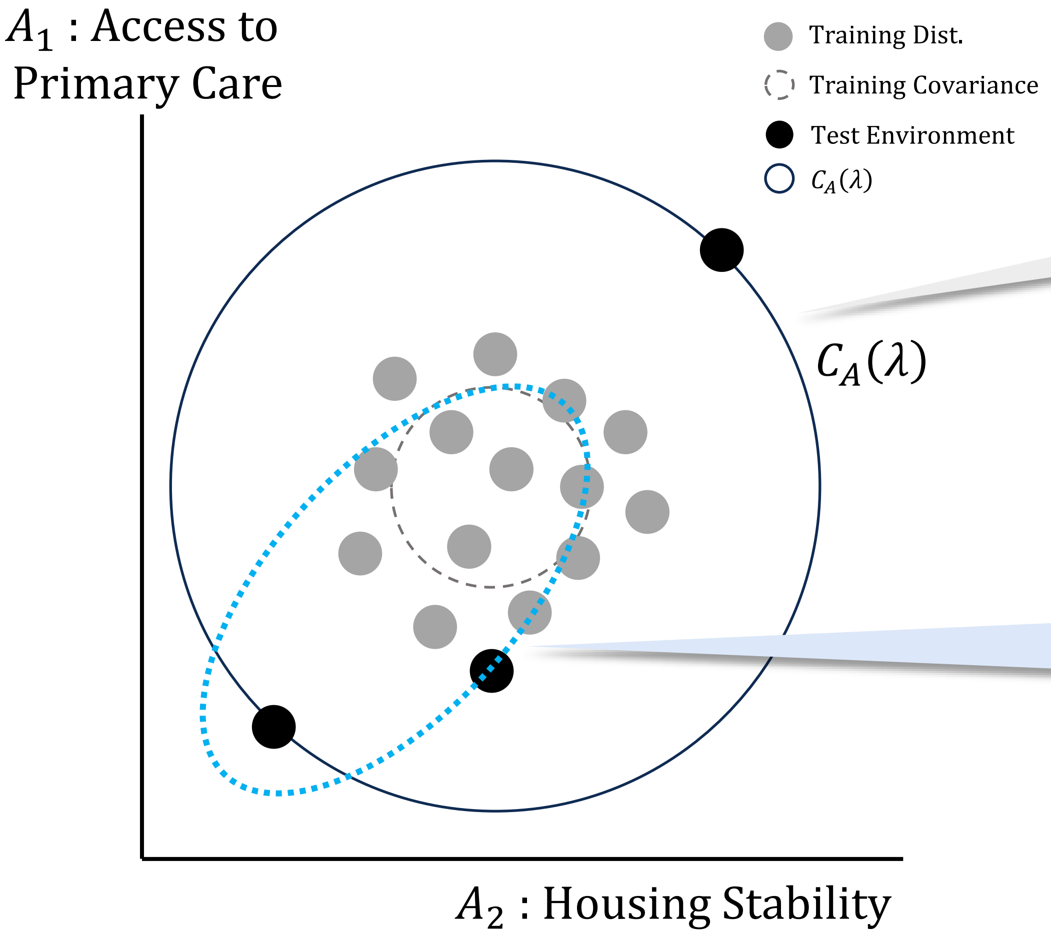
Theorem 1 (Informal)
Given a single noisy proxy W of A, the robustness set is provably reduced, and this reduction is not identifiable.

Theorem 2 (Informal)
Given two noisy proxies of A, one can recover the original robustness set, using a modified objective

X, Y not shown here, just the dimensions of A

[1] Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. Anchor regression: Heterogeneous data meet causality. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 83(2):215–246, 2021.

Targeting the robustness set with prior knowledge



Anchor Regression [1] assumes that A is observed, and optimizes a worst-case loss over bounded interventions

$$\sup_{\nu \in C_A(\lambda)} E_{do(A := \nu)}[(Y - \gamma^\top X)^2]$$

Theorem 3 (Informal)
We generalize to a larger class of robustness sets, and prove that the objective is identified with <u>two proxies</u>

X, Y not shown here, just the dimensions of A

[1] Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. Anchor regression: Heterogeneous data meet causality. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 83(2):215–246, 2021.

Conclusion

Incorporate prior knowledge about future shifts, instead of seeking invariance to arbitrary changes

1. Specify **relevant factors of variation** A via proxies
2. Specify **plausible shifts** via targeted robustness sets