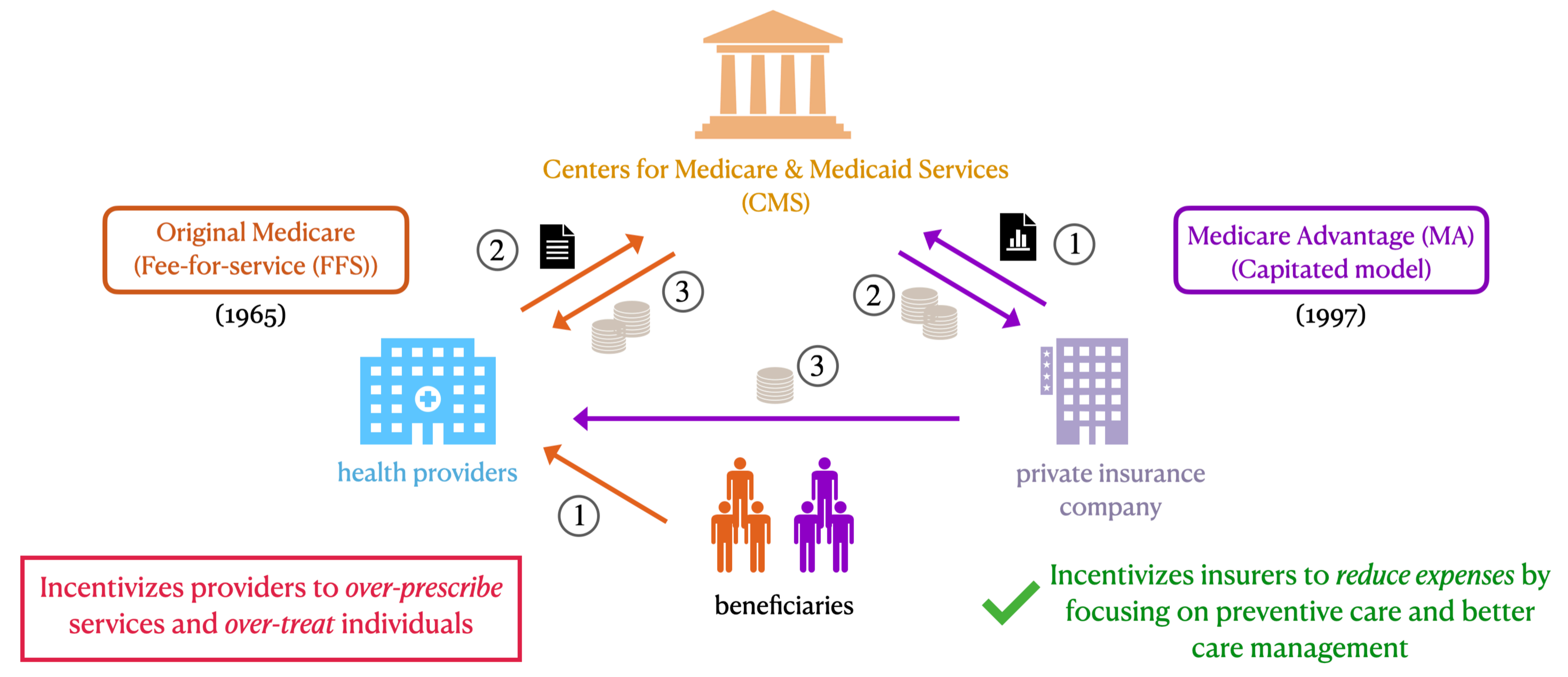
# Strategic Feature Selection

Jivat Neet Kaur[1], Pratik Patil[2], Divya Shanmugam[3], Emma Pierson[1], Michael I. Jordan[1,7], Nika Haghtalab[1], Meena Jagadeesan[4,5*], Ahmed Alaa[1*], Serena Wang[6*]

[1]UC Berkeley [2]UT Austin [3]Cornell Tech [4]UPenn [5]Stanford [6]Harvard [7]Inria, Paris

## Background on US Health insurance

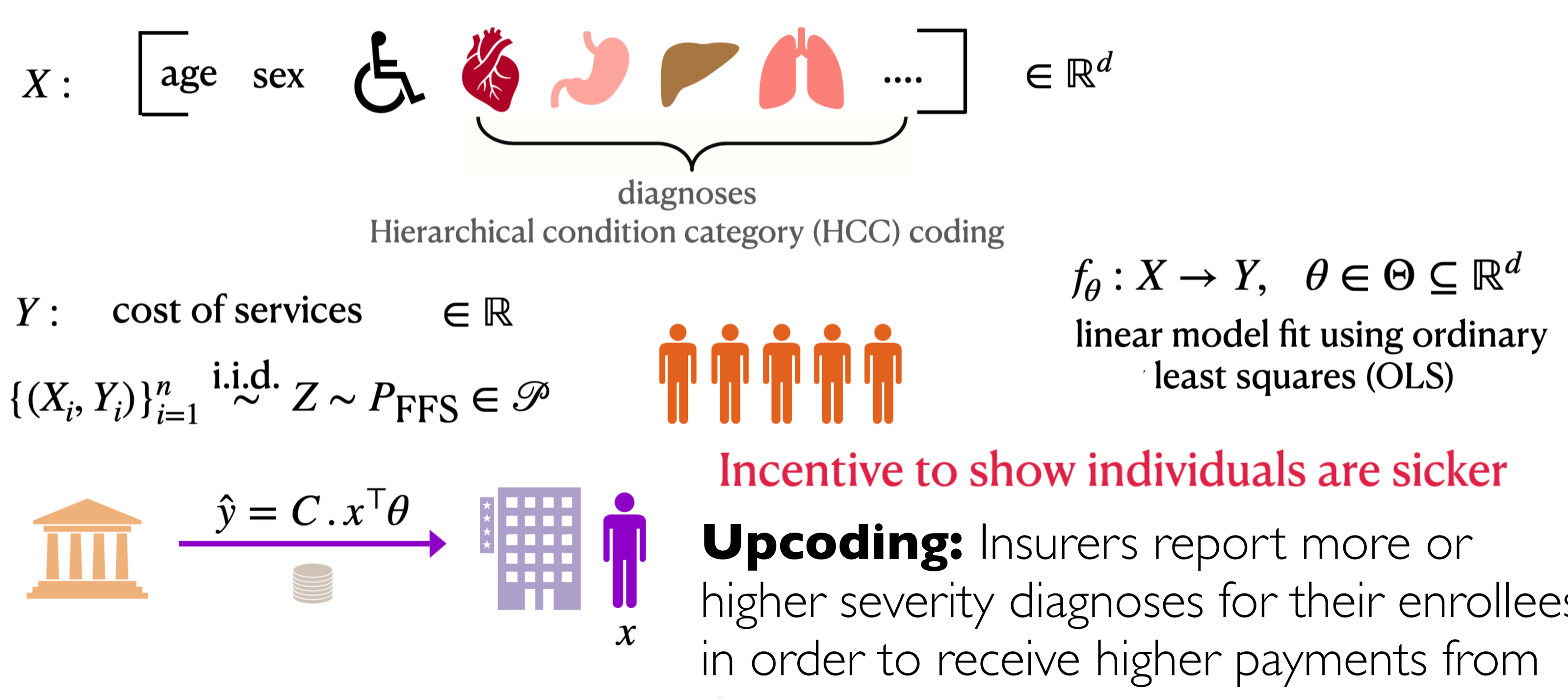




## How are payments made in MA?

### Risk adjustment model

**CMS-HCC model**




Incentive to show individuals are sicker

**Upcoding:** Insurers report more or higher severity diagnoses for their enrollees in order to receive higher payments from the government

## Problem setting and model

Interaction between a decision maker who specifies predictor $f_\theta$ and organizations that can modify the reported features of the individuals from $x$ to $x + a$ at cost $C(a) = a^\top Ha/2$, where $H \in \mathbb{R}^{d \times d} \succ 0$ is the cost matrix.

**Insurer's goal:** maximize payout while minimizing cost

Best response: $a^*(x;\theta) \in \operatorname{argmax}_{a \in \mathbb{R}^d} \ \theta^\top(x + a) - C(a)$

**Govt./CMS's goal:** predict the true outcome as accurately as possible in the presence of strategic responses

Best response: $\min_{\theta \in \mathbb{R}^d} \ \mathbb{E}[(\theta^\top(X + a^*(\theta)) - Y)^2]$

## Overview of this work

**Strategic classification** → **Coarse policy levers**

Compute *strategic optimum*, prediction rules that are optimal once manipulation taken into account.

- *Feature selection*: exclude features that are perceived to be especially manipulable
- *Regularization*: shrink coefficients on the retained features

*How should policymakers design* ***feature selection together with regularization*** *to combat strategic behavior?*

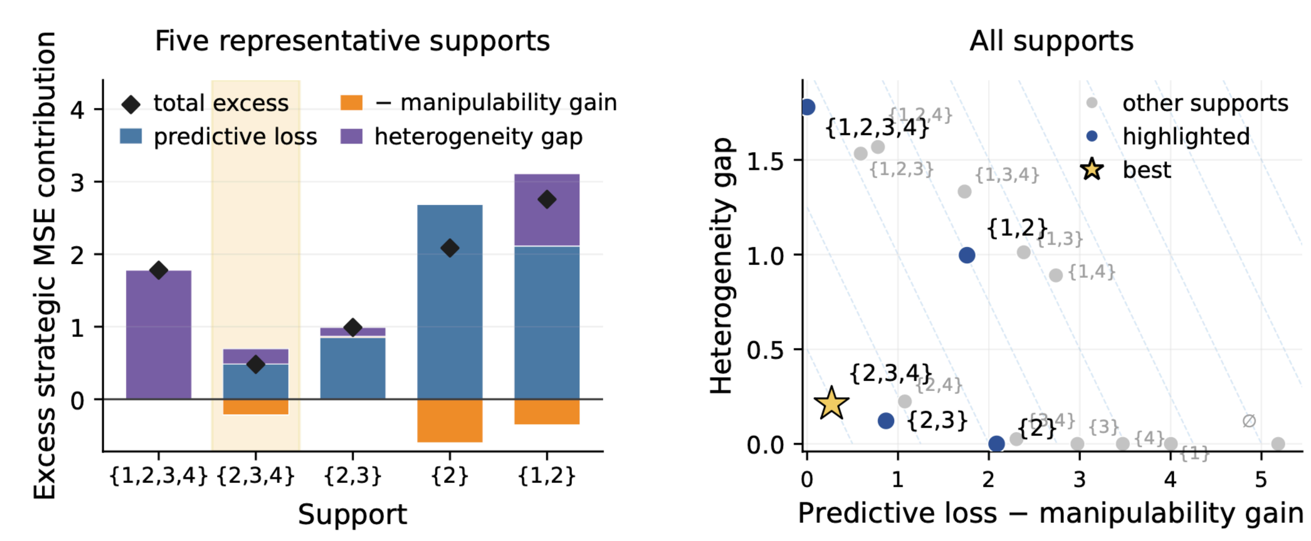
(Hardt et al., 2016; CMS, 2023)

## Theoretical results

**Main finding.** Excluding individual features based on their manipulability alone is suboptimal: features should be selected ***jointly*** based on their relative manipulability and predictability.

**Theorem 1** (Irreducible loss of coarse policy levers)**.** For any choice of feature subset and ridge regularization level, the resulting prediction error exceeds the strategic optimum by at least an irreducible gap.

**Theorem 2** (Upper bound)**.** For any feature subset, prediction error under optimally tuned ridge regularization is controlled by three terms: *manipulability gain* from excluding features, *predictive loss* from omitting them, and a *heterogeneity gap* capturing how well one regularization level accounts for heterogeneous costs.



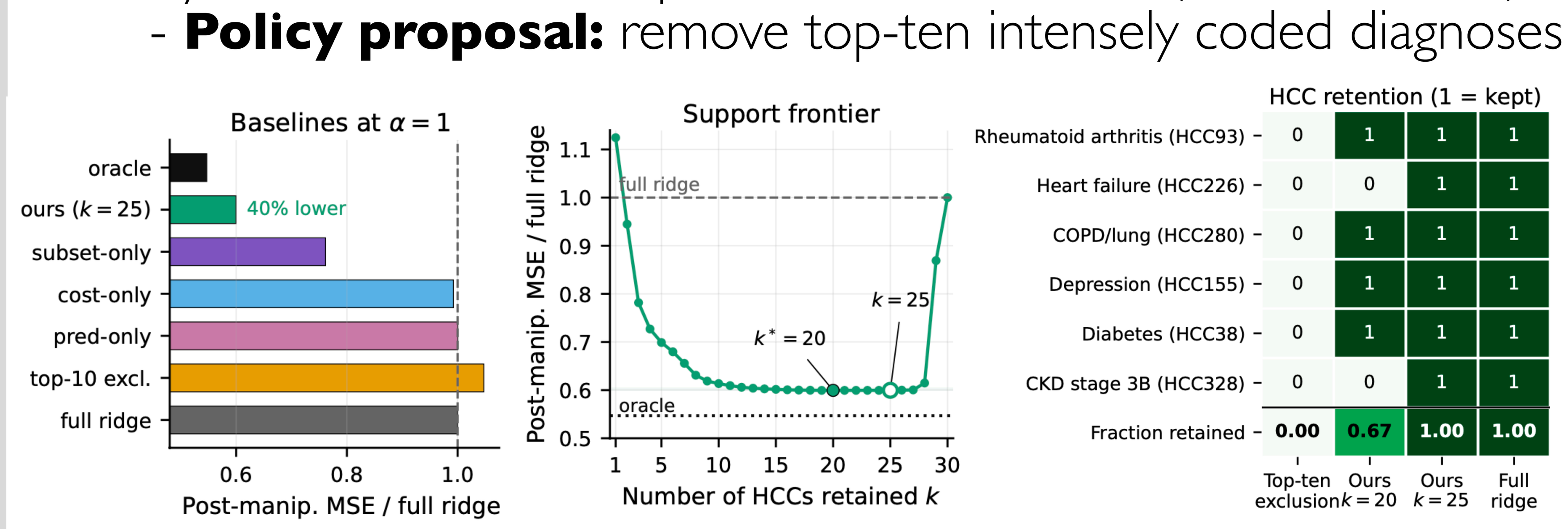

## Policy implications

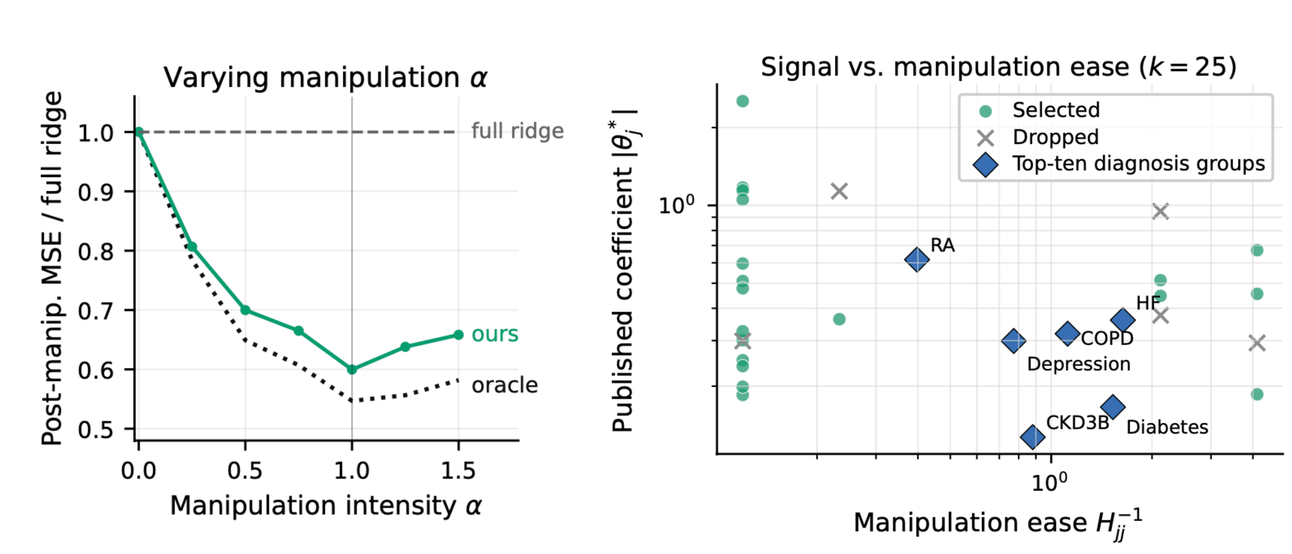
1. Feature selection and regularization must be tuned jointly.
2. Predictive manipulable groups can be retained when their manipulation costs are homogeneous.
3. Less manipulable correlated proxies can replace manipulable features.

## Case study: Feature selection in health insurance payments

Simulate realistic upcoding building on prior work to first generate diagnosis codes in the absence of coding incentives, and then modify them to match the prevalence shift in MA (Kronick et al., 2025).

- **Policy proposal:** remove top-ten intensely coded diagnoses




Low strategic error does not require aggressive exclusion.

Proposed method substantially reduces post-manipulation MSE relative to baselines and remains competitive with oracle.

Selected support is not determined by coding vulnerability alone.