"Who[1] experiences large model decay and why[2]?"

A Hierarchical Framework for Diagnosing Heterogeneous Performance Drift

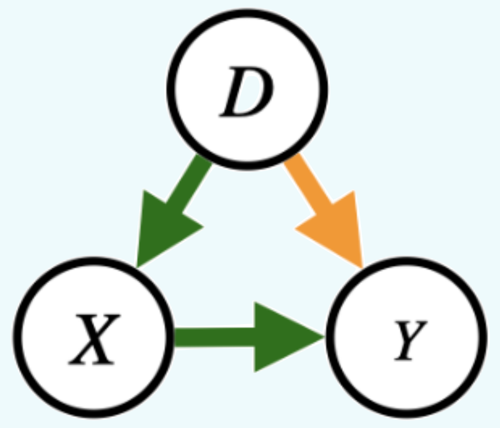
University of California San Francisco

Harvineet Singh[1], Fan Xia[1], Alexej Gossmann[2], Andrew Chuang[1], Julian Hong[1], Jean Feng[1]

[1] University of California, San Francisco, [2] Independent

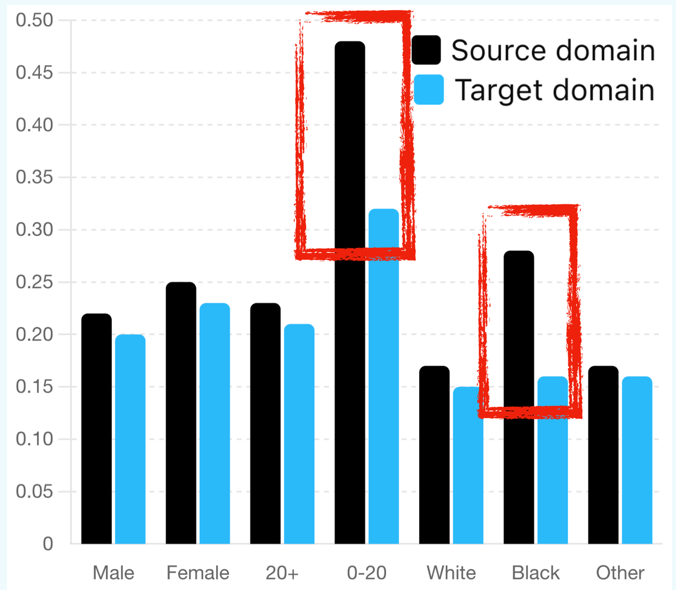
Motivation

- To understand differences in performance between a source domain ($D=0$) and target domain ($D=1$), existing methods decompose the *average* performance difference into contributions from **covariate** vs **outcome** shifts:



$$\mathbb{E}_1[\ell(Y,f(X))] - \mathbb{E}_0[\ell(Y,f(X))]$$
$$= \mathbb{E}_1[Z_0(X)] - \mathbb{E}_0[Z_0(X)]$$
$$+ \mathbb{E}_1[Z_1(X)] - \mathbb{E}_1[Z_0(X)]$$

where $Z_D(X) = \mathbb{E}_D[\ell(Y,f(X)) \mid X]$.



- However, ***performance differences can vary significantly across subgroups.***

Key contributions

- To help model developers better diagnose and mitigate large performance gaps, this work develops SHIFT, a hierarchical hypothesis testing framework that answers:
  1. ***(Who)*** Have covariate or outcome shifts led to unacceptably worse performance in any meaningfully large subgroup?
  2. ***(Why)*** If so, can these performance drops be explained by a sparse subset of variables in X?
- Unlike existing methods, SHIFT
  - Is nonparametric
  - Provides valid uncertainty quantification, even in settings with potentially limited data
  - Does not require detailed causal knowledge

SHIFT: Subgroup-scanning Hierarchical Inference Framework for performance drifT

Aggregate Covariate Shift Hypothesis

H_0: For all subgroups A with size $\geq \epsilon$, the performance drift in A due to the aggregate covariate shift is no larger than pre-specified tolerance $\tau \geq 0$, i.e.
$\mathbb{E}_1[Z_0(X) \mid X \in A] - \mathbb{E}_0[Z_0(X) \mid X \in A] \leq \tau$.

Aggregate Outcome Shift Hypothesis

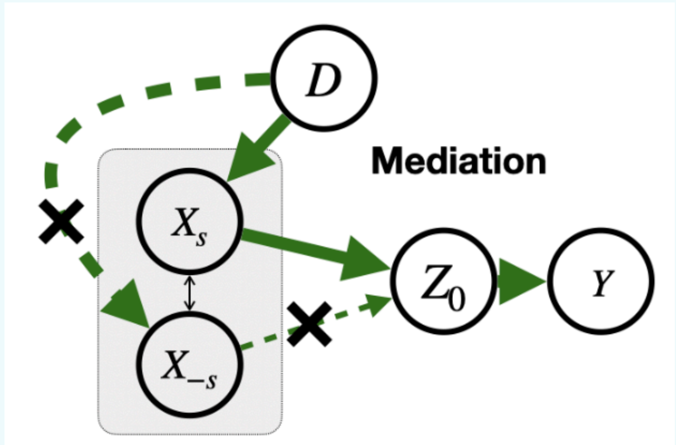
H_0: For all subgroups A with size $\geq \epsilon$, the performance drift in A due to the aggregate outcome shift is no larger than pre-specified tolerance $\tau \geq 0$, i.e.
$\mathbb{E}_1[Z_1(X) \mid X \in A] - \mathbb{E}_1[Z_0(X) \mid X \in A] \leq \tau$.

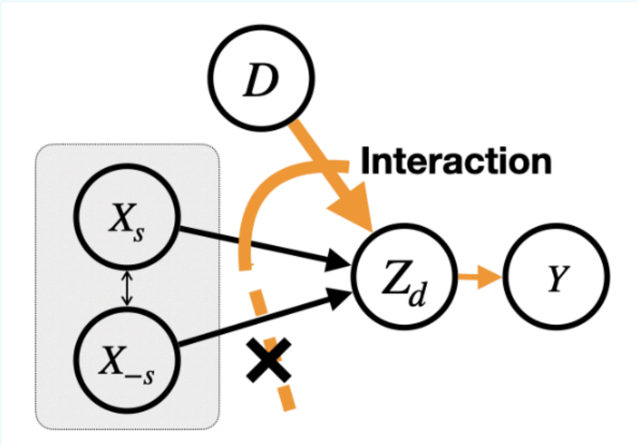
X_s-specific Covariate Shift Hypothesis

H_0: For all subgroups A with size $\geq \epsilon$, the candidate covariate shift solely with respect to variable subset X_s explains the performance change in A, i.e.
$\mathbb{E}_1[Z_0(X) \mid X \in A] - \mathbb{E}_s[Z_0(X) \mid X \in A] \leq \tau$.

X_s-specific Outcome Shift Hypothesis

H_0: For all subgroups A with size $\geq \epsilon$, the candidate outcome shift solely with respect to variable subset X_s explains the performance change in A, i.e.
$\mathbb{E}_1[Z_1(X) \mid X \in A] - \mathbb{E}_1[Z_s(X) \mid X \in A] \leq \tau$.



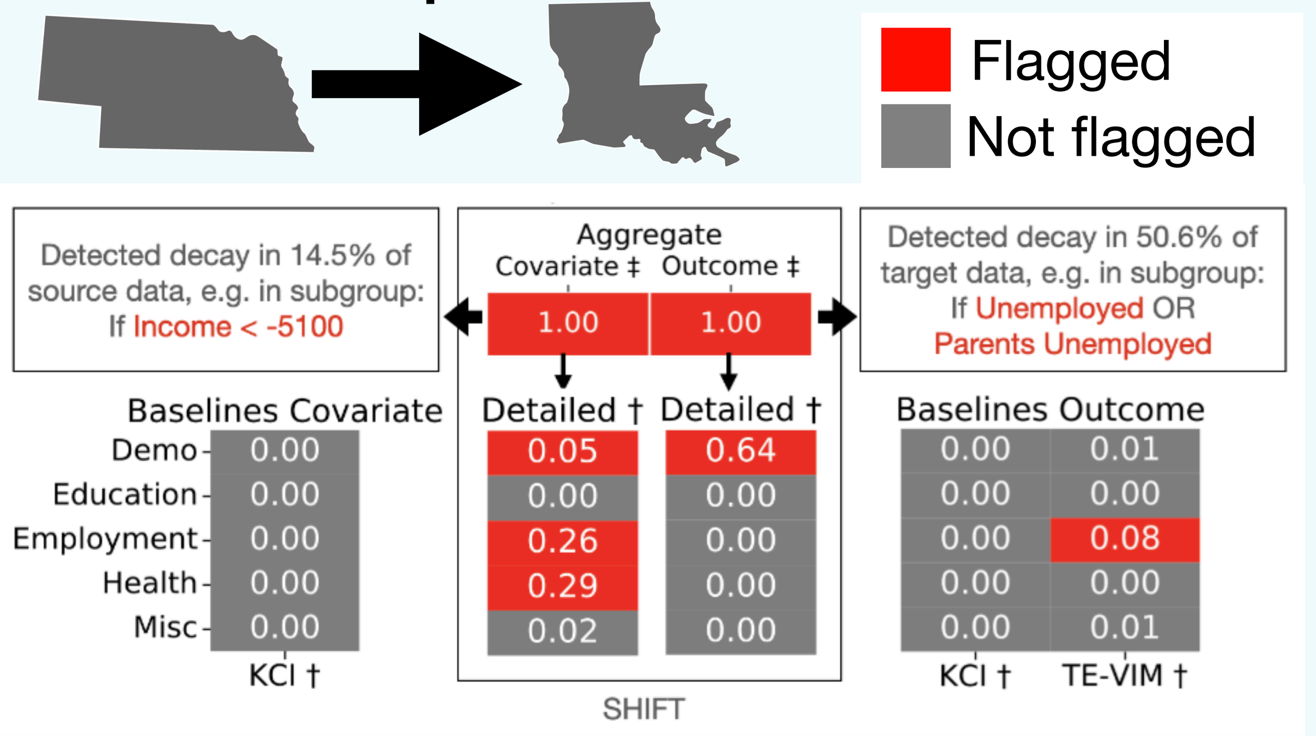


SHIFT step-by-step

Step 1. Split data into train vs test:

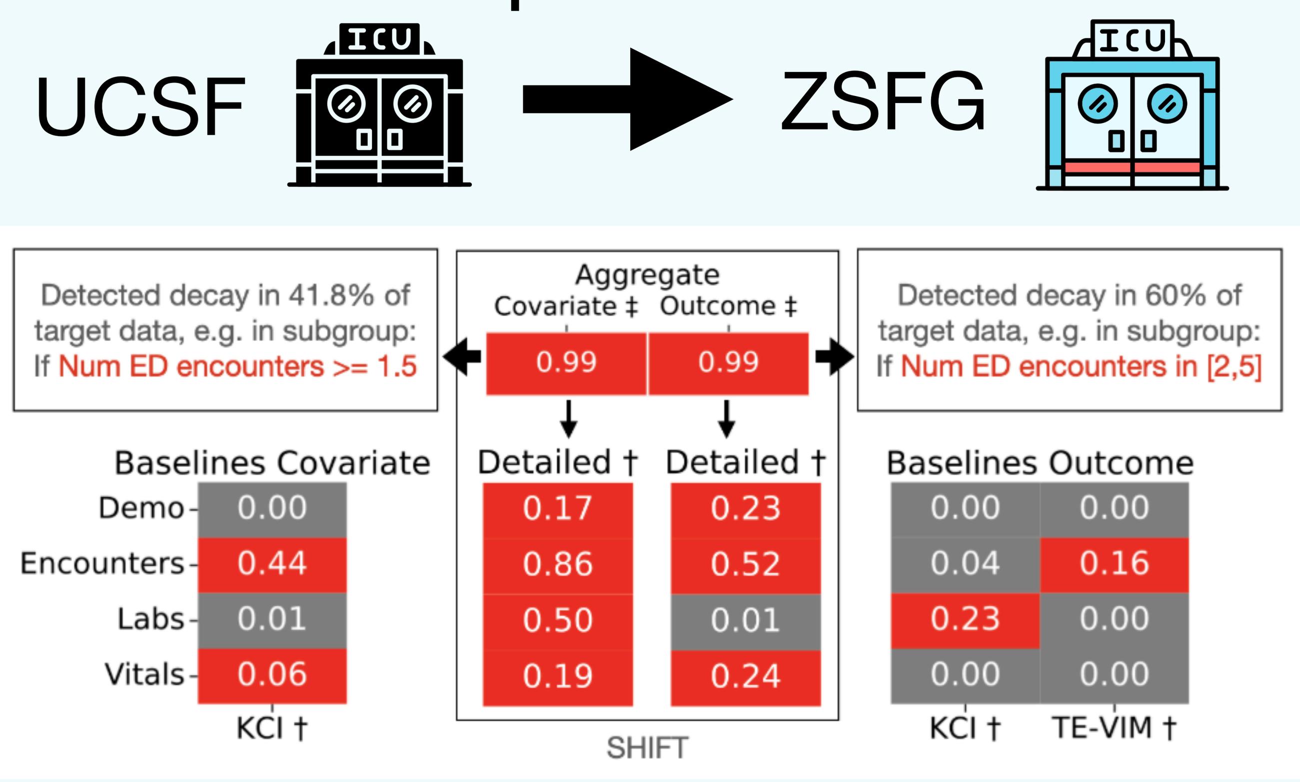
Train		Test
Step 2a. Estimate nuisance parameters, i.e. outcome models $\hat{Z}_D, \hat{Z}_s$ and density ratio models $\hat{\pi}, \hat{\pi}_s$, using ML.	**Step 2b.** Estimate candidate subgroups (i.e. $\hat{A}_{agg}, \hat{A}_s$), defined as binary functions of X, using ML.	**Step 3**. Construct test statistics (e.g. $\mathbb{E}[(\ell - Z_0(X) - \tau)1\{X \in \hat{A}\}]$) using double-debiased ML. Obtain p-values using multiplier bootstrap.

Experiment: Diagnosing an insurance prediction model



SHIFT flags aggregate tests that are rejected to indicate a subgroup has been detected and flags X_s-specific tests that are *not* rejected as potential explanations.

Experiment: Diagnosing a readmission prediction model



Paper here

This work was funded through a Patient Centered Outcomes Research Institute® (PCORI®) Award (ME-2022C125619).

The views presented in this work are solely the responsibility of the author(s) and do not necessarily represent the views of the PCORI®, its Board of Governors, or Methodology Committee.