Is this model reliable for everyone? Testing for strong calibration

Jean Feng[1], Alexej Gossmann[2], Romain Pirracchio[1], Nicholas Petrick[2], Gene Pennello[2], Berkman Sahiner[2]

[1]University of California, San Francisco, [2]U.S. Food and Drug Administration

Introduction

- A strongly calibrated model is one that is reliable for *all* subgroups.
- Methods for identifying subgroups for which a model is poorly calibrated are often low-powered due to:
  - Correction for multiple testing after searching over a large number of potential subgroups
  - Little remaining signal if a highly flexible model was fit (e.g. via machine learning)
- *An omnibus test for the existence of a poorly calibrated subgroup* is more feasible in settings with limited data.
- Although newer GOF tests can be adapted to test for strong calibration, they lack power in settings with small subgroups or low signal-to-noise ratios.

Test of strong calibration

- Let $\hat{p}$ be a risk prediction algorithm and p_0 be the true risk.
- Null hypothesis: The prevalence of the subgroup where the true and predicted risk differ by more than δ is no larger than $\epsilon \geq 0$, i.e.

$$H_0 : \Pr(|\hat{p}(X) - p_0(X)| > \delta) \leq \epsilon$$

Funding: This work was supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) as part of a financial assistance award Center of Excellence in Regulatory Science and Innovation grant to University of California, San Francisco (UCSF) and Stanford University, U01FD005978. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by FDA/HHS, or the U.S. Government.

Omnibus test for subgroups = Test for changepoints

Intuition: If we order test observations by their predicted residuals, there should be a drop in the association between observed and predicted residuals if a poorly calibrated subgroup exists.

Advantages of changepoint tests:

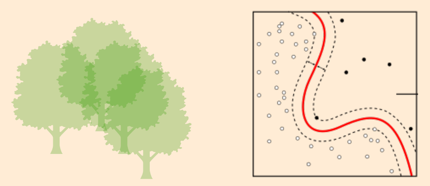
- Avoids specifying subgroup size
- Good for detecting small subgroups
- Nested subgroups
- Leverages predicted residuals

The Adaptive Score CUSUM test

Train data

1. Train an ensemble of residual models.

$\{\hat{g}_k : k = 1, \cdots, K\}$



Test data

2. For each residual model, order observations by predicted residuals.

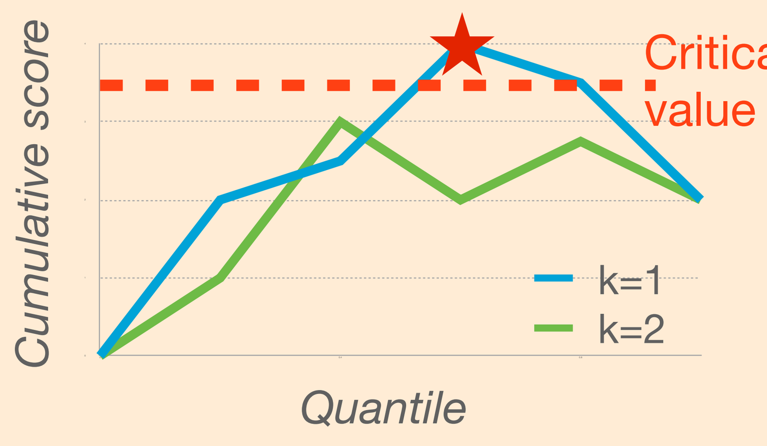
Predicted residuals from $\hat{g}_k$

Largest ⟵⟶ *Smallest*

$\hat{\epsilon}_{(1),k}$	$\hat{\epsilon}_{(2),k}$	$\hat{\epsilon}_{(3),k}$	$\hat{\epsilon}_{(4),k}$
$\epsilon_{(1)}$	$\epsilon_{(2)}$	$\epsilon_{(3)}$	$\epsilon_{(4)}$

Observed residuals

3. Test for structural change using an ensemble of score-based CUSUM tests.



4. Employ cross-validation

An ensemble of score-based CUSUM tests

- Consider the simple case of a one-sided test with $\epsilon = 0$.
- For each residual model $\hat{g}_k$, we define working models for structural change:

$$\text{logit}(p_{k,\gamma}(y=1|x)) = \text{logit}(\hat{p}_\delta(y=1|x)) + \theta \underbrace{\hat{g}_k(x) 1\{\hat{g}_k(x) > \gamma\}}_{\text{Subgroup detector}} \qquad \forall \gamma \geq 0$$

Maximum allowable risk if $\hat{p}(x)$ is calibrated

- Under the null, the expected score is non-positive for all models for structural change:

$$H_0 : \max_{k=1,\cdots,K} \sup_{\gamma \geq 0} \mathbb{E}\left[\left(Y - \hat{p}_\delta(Y|X)\right) \hat{g}_k(X) 1\{\hat{g}_k(X) > \gamma\}\right] \leq 0$$

- Ensemble CUSUM test statistic:

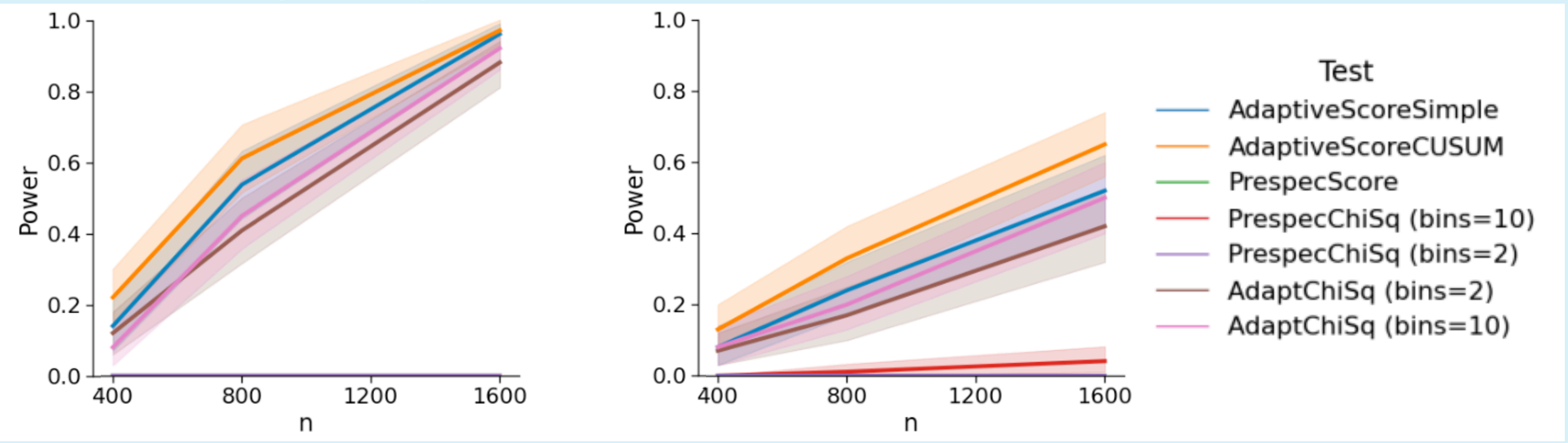
$$\hat{C}_n = \max_{k=1,\cdots,K} \sup_{\gamma \geq 0} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{p}_\delta(Y_i|X_i)\right) \hat{g}_k(X_i) 1\{\hat{g}_k(X_i) \geq \gamma\}$$

Simulations

Data: $X \in \mathbb{R}^{10}$, binary Y with logit

$$(0.6x_0 + 0.4x_2 + 0.2x_3)1\{\max(x_1, -x_2) \geq -2\} + 0.2x_1 1\{\max(x_1, -x_2) < -2\}$$

Tests: Residual models fit using RFs and kernel logistic regression. 4-fold CV.

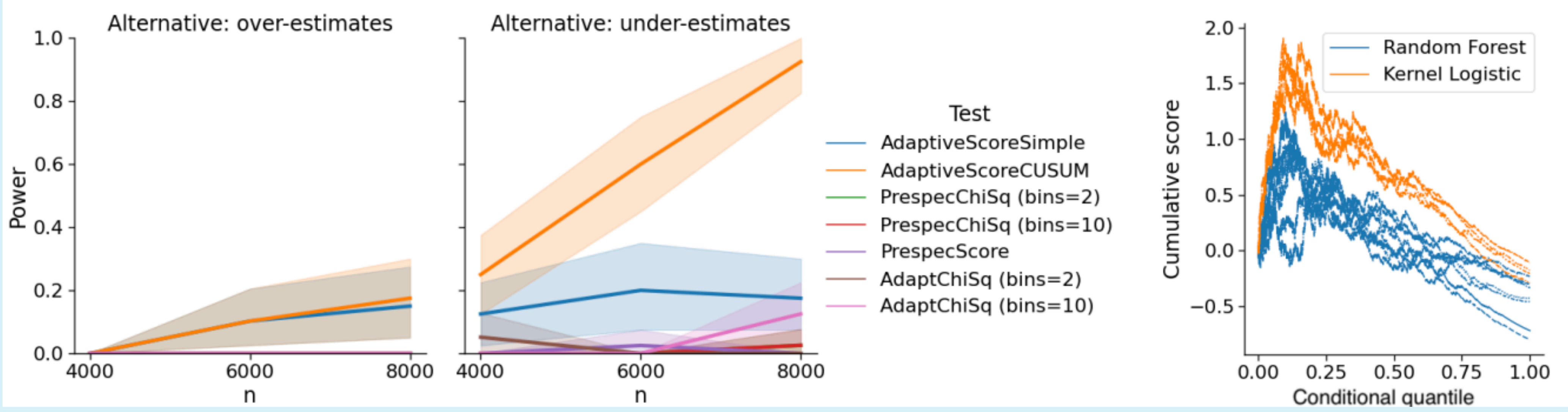


Auditing a mortality prediction model

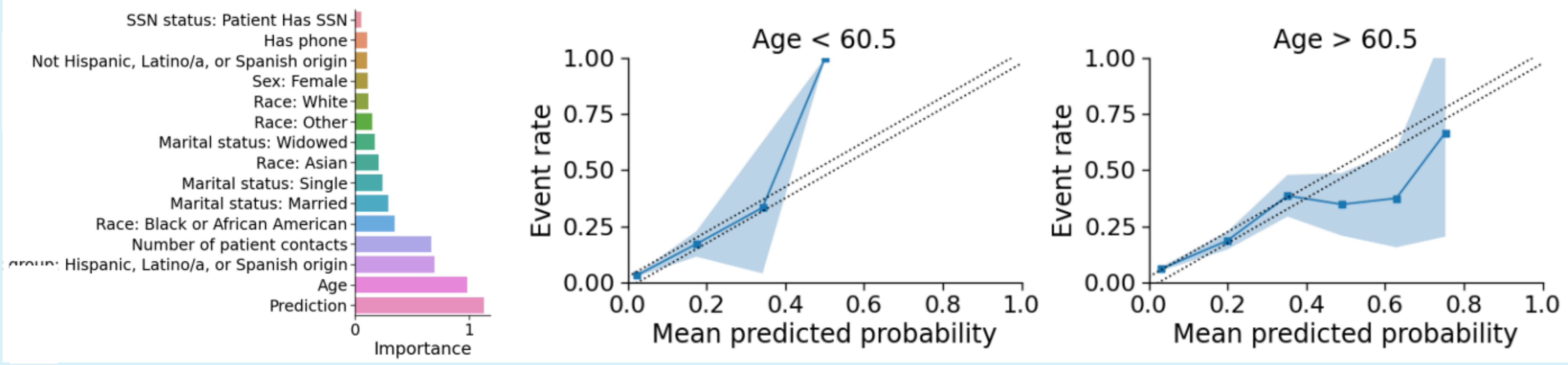
Mortality model: RF trained on data from 250,000 patients from the Zuckerberg San Francisco General Hospital. Input features include demographic variables and diagnosis codes.

Tests: Separately detect under- and over-estimation of the true risks

Test results: Power (left) and CUSUM plot (right)



Detected subgroup: variable importance (left) and calibration plots (right)



Paper at https://arxiv.org/abs/2307.15247