Fairness Regularized Risk Assessment Models: Balancing Risk Prediction and Racial and Ethnic Equality

Rhys Hester (Clemson University)
Greg Ridgeway (University of Pennsylvania)
Ryan Labrecque (RTI)

Overview

- Actuarial risk assessment tools
  - Aim to provide objective measures of risk, but…
  - Generate concerns over racial bias
- Typical process
  - Fit models to maximize prediction accuracy…
  - Then assess racial fairness
- Fairness regularized models
  - Simultaneously optimize predictive performance and minimize racial differences
  - Logistic regression fit with a “lack of fairness penalty” added to the negative Bernoulli log-likelihood

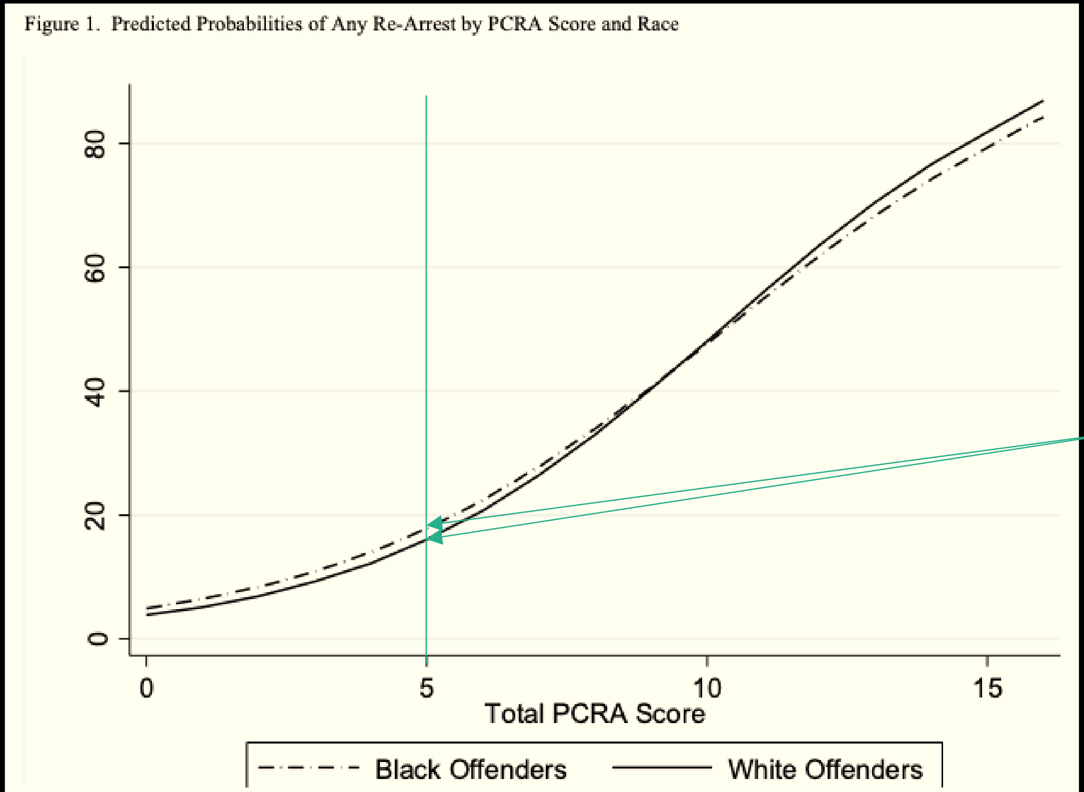
Background

- Actuarial risk assessment is increasingly prevalent in the justice system

- Several widely publicized critiques
  - Attorney General Eric Holder's Comments (NACDL Speech, 2014)
  - Propublica/COMPAS Controversy (e.g., Angwin et al. 2016; cf. Flores et al. 2016)
  - *Weapons of Math Destruction* (Cathy O'Neil 2016)

- Multiple, conflicting definitions of "fairness" (Chouldechova 2017; Berk et al. 2021)
  - Mathematical proofs that all common fairness measures cannot be satisfied simultaneously

Calibration Is One Way of Defining Fairness in Risk Assessment



Skeem & Lowenkamp (2016)

At any PCRA score, rearrest probabilities are similar

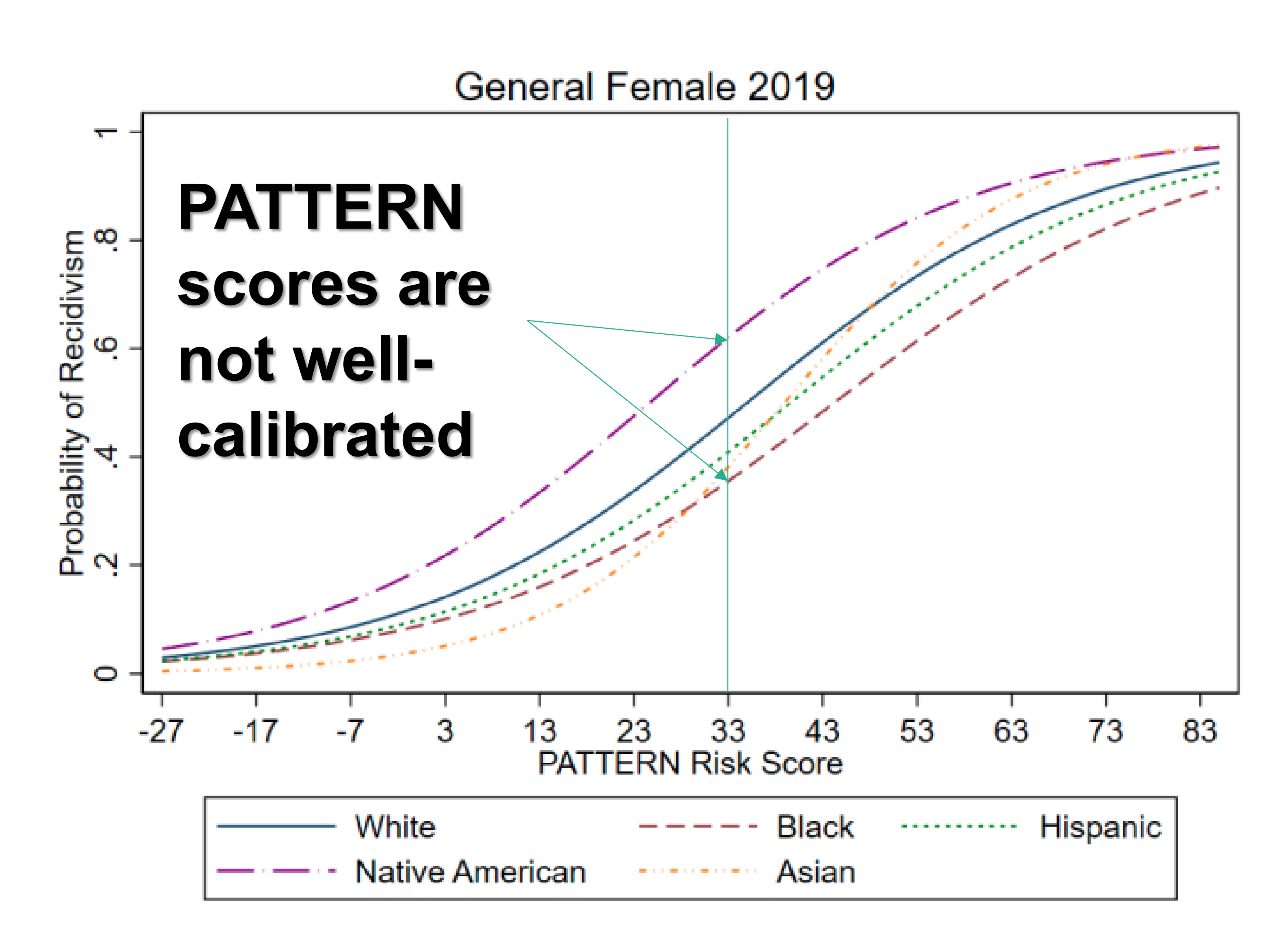
A score *S* is well-calibrated if

$$P(Y = 1|S = s, R = \text{black}) = P(Y = 1|S = s, R = \text{white})$$

PATTERN Risk Tool Is Accurate…

- High overall accuracy relative to instruments used in criminal justice
  - Female group AUCs range from 0.73 - 0.86
  - 0.76 and 0.78 for White and Black women

…But PATTERN Risk Scores Do Not Seem Fair



Source: https://www.ojp.gov/pdffiles1/nij/309264.pdf

Measure Lack-of-Calibration with F-statistic

- Lack-of-calibration penalty
  - Compute score-and-sum predictions as

$$\hat{f}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots$$

$$\log \frac{P(y_i=1)}{1-P(y_i=1)} = \alpha_0 + \alpha_1 ns_1(\hat{f}_i) + \alpha_2 ns_2(\hat{f}_i) + \alpha_3 ns_3(\hat{f}_i) + \alpha_4 ns_4(\hat{f}_i) + \alpha_5 \text{black}_i + \alpha_6 \text{black}_i ns_1(\hat{f}_i) + \alpha_7 \text{black}_i ns_2(\hat{f}_i) + \alpha_8 \text{black}_i ns_3(\hat{f}_i) + \alpha_9 \text{black}_i ns_4(\hat{f}_i)$$

Natural splines allowing non-linear relationship between score and log odds

Main effect for race

- Measure calibration with F-statistic testing

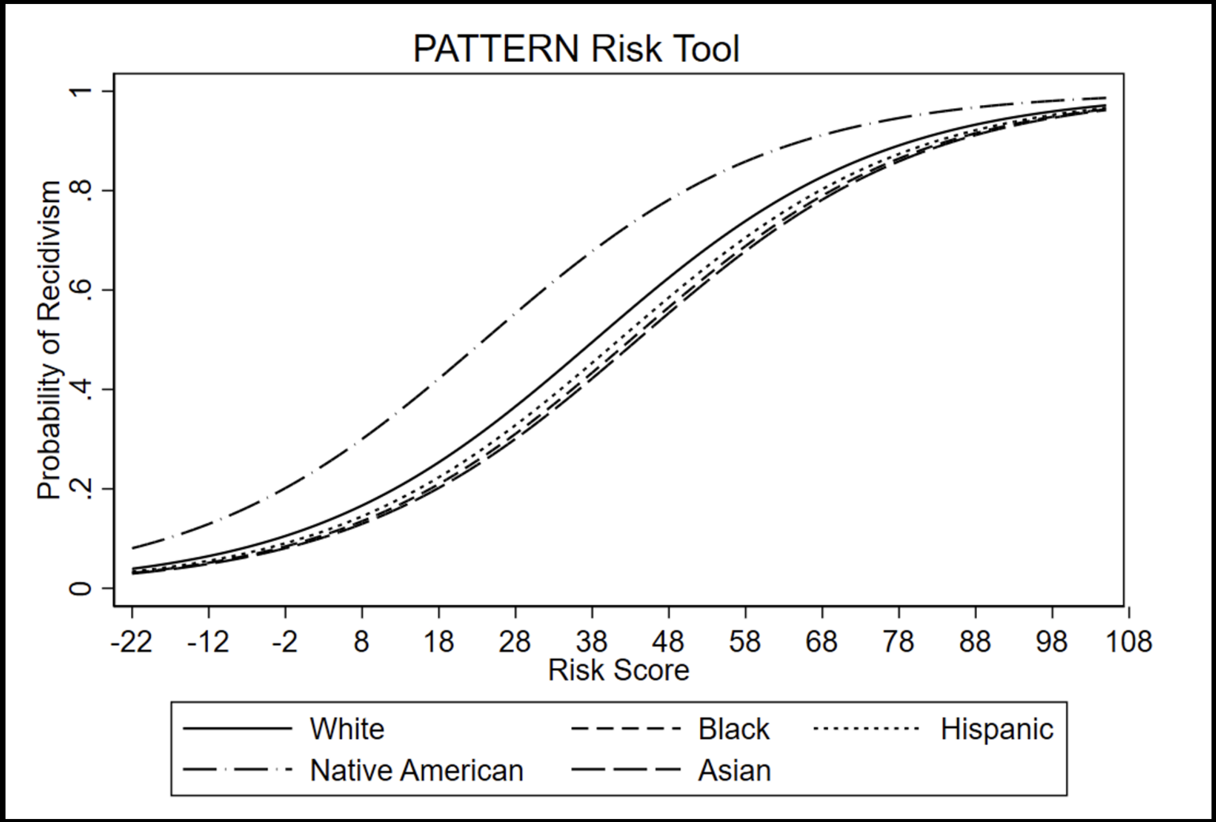
$$\alpha_5 = \alpha_6 = \alpha_7 = \alpha_8 = \alpha_9 = 0$$

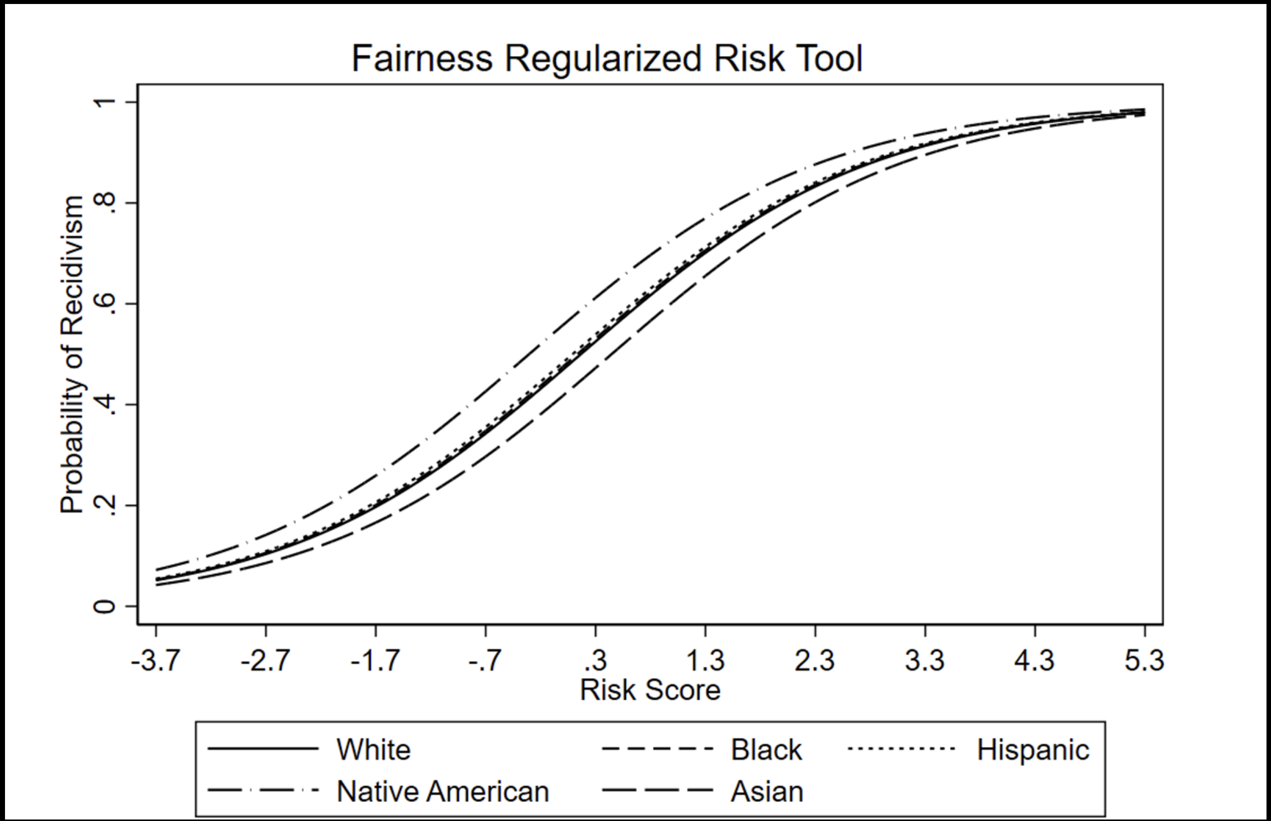
Capture difference in calibration curves across race

Minimize Deviance with Unfairness Penalty

- Finds β to minimize
$$\ell(\beta) = -2\sum_{i=1}^{n} y_i\beta'\mathbf{x}_i - \log(1+\exp(\beta'\mathbf{x}_i)) + \lambda F(\beta)$$
  - No differences in calibration by race group, $F \approx 0$
  - Increasing λ focuses optimization focuses on equal calibration
- May create scores that fail to incentive constructive rehabilitation
  - For example, more serious criminal history predicts lower recidivism risk
  - Additional constraints on β
    - Risk must increase with more serious criminal history
    - Risk must decrease with more participation in rehabilitation programming

Fairness Regularization Improves Within Race Calibration





Improving Calibration Slightly Reduces Predictive Performance (AUC)

	PATTERN	FR
White	0.80	0.79
Black	0.75	0.74
Hispanic	0.77	0.75
Native American	0.70	0.70
Asian	0.84	0.84
Overall	0.78	0.77

Conclusion

- When unconstrained, risk assessments
  - Are not calibrated within groups
  - May encode undesirable incentives
- Fairness regularization improves within group calibration
  - Optimization can also enforce desired incentives
- Improving fairness comes with a price: reduced predictive performance
  - Forcing perfectly calibration reduces the model to predict the baseline rearrest rate for everyone (perfectly fair, but no risk assessment)

Fairness Regularized Risk Assessment Models: Balancing Risk Prediction and Racial and Ethnic Equality

Rhys Hester (Clemson University)
Greg Ridgeway (University of Pennsylvania)
Ryan Labrecque (RTI)

ASC Nov 2025