Fair Clinical Risk Prediction: Trade-offs and Alternatives

Stephen R. Pfohl, Katherine Heller, Nigam H. Shah
spfohl@stanford.edu, kheller@google.com, nigam@stanford.edu

An Empirical Characterization of Fair Machine Learning for Clinical Risk Prediction

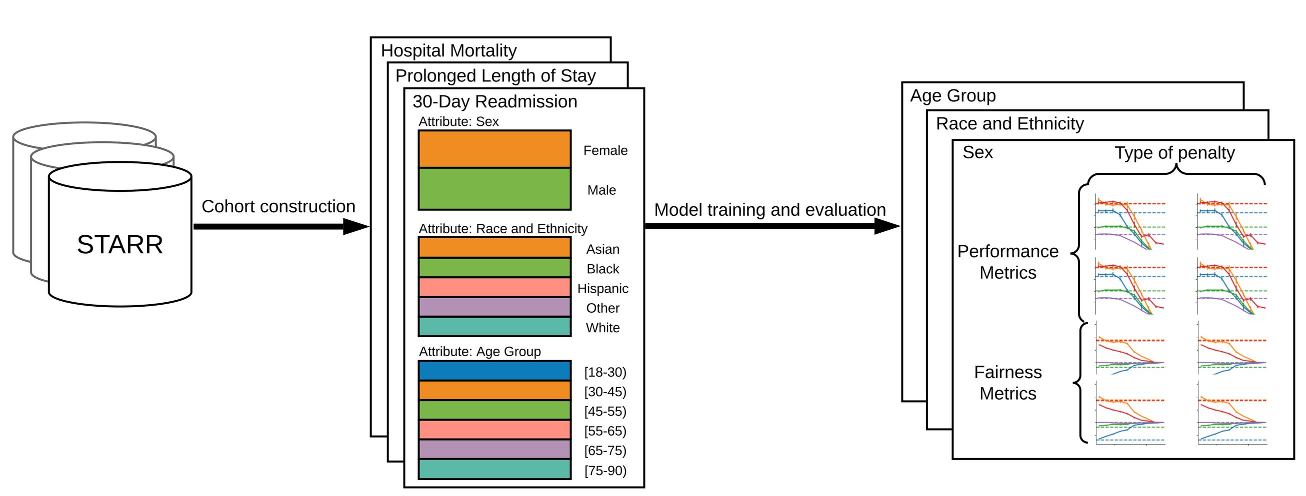
Stephen R. Pfohl, Agata Foryciarz, Nigam H. Shah

Key points

- The effects of imposing fairness constraints on clinical predictive models are not well understood
- We conduct a large-scale empirical study to characterize the impact of imposing group fairness on measures of model performance and fairness
- We find that group fairness penalties generally
  - Degrade model performance for all groups
  - Introduce *relative calibration errors* that occurs across groups -- independent of changes in absolute calibration error
- Algorithmic fairness is incapable of auditing or correcting for *causal quantities* not captured by observational criteria
  - *Upstream biases* due to the interaction of structural inequities with errors in problem formulation and measurement
  - *Downstream biases* defined in terms of disparate impact of a model-guided intervention
- We encourage researchers to step outside of the algorithmic fairness frame and engage critically with the broader sociotechnical context of machine learning in healthcare

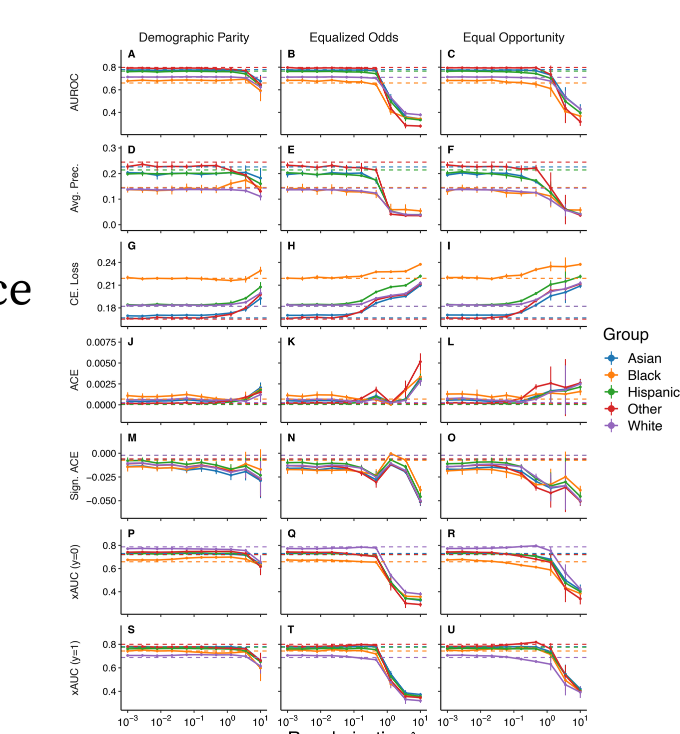
Methods

- Apply regularized learning objectives to impose conditional prediction parity
- Evaluate
  - Conditional prediction parity
  - Relative calibration error
  - Cross group ranking performance (xAUC)
  - Standard performance measures (AUROC, AP, etc)
- Repeat in a grid of 25 experimental conditions across datasets (STARR, Optum Clinformatics Data Mart, MIMIC-III), clinical outcomes, and sensitive attributes

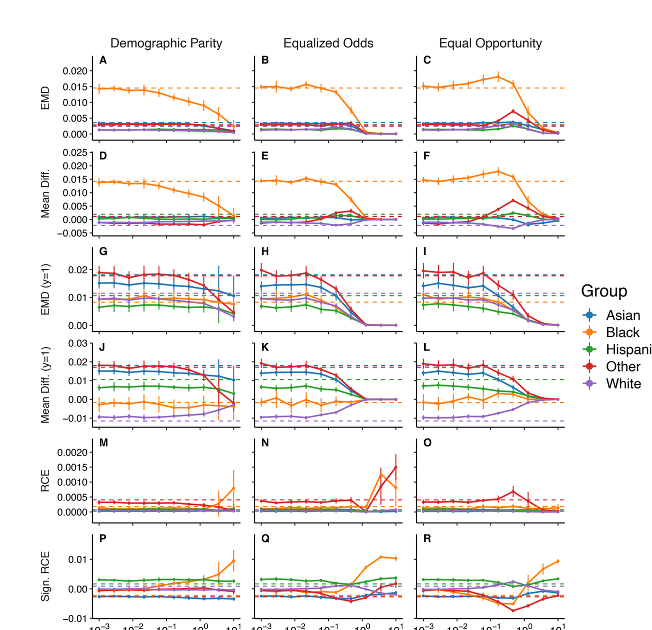


Results

Performance Metrics



Fairness Metrics



Alternative Algorithmic Approaches for Reliable and Fair Clinical Risk Prediction

Key points

- If group-level model performance is a suitable proxy for benefit, then the algorithmic fairness approaches that we study generally introduce harm
  - Always critically evaluate this assumption in the context of the assumptions underlying problem formulation, measurement, and intended use of the intervention
- Increasing the effective size and diversity of datasets via pooling across siloes may improve model performance for underrepresented groups without the trade-offs of algorithmic fairness objectives
- Key barriers to pooling data across siloes
  - Ethical and legal necessity of respecting privacy constraints
  - Distribution shift and heterogeneity limit transfer across siloes
- **Hypothesis**: We may improve group-level model performance while achieving notions of algorithmic fairness by composing
  - Approaches to addressing privacy constraints in learning across siloes, such as federated and differentially private learning
  - Approaches to learning robust and transferable models
  - Approaches to imposing algorithmic fairness
- Invariance provides a common framework for fairness and distribution shift
  - We have a common algorithmic toolbox for these problems
  - Empirical characterization of trade-offs among fairness criteria informs our empirical understanding of invariance as a tool for addressing distribution shift
- **Proposal and on-going work:** Assessing the above hypothesis in several settings:
  - Learning robust and transferable ASCVD risk scores in multi-center EHR and large national claims databases without data sharing, partitioning by state, zip code, and care site
  - Benchmarking with mortality and length of stay outcomes in the eICU Collaborative Research DB

Invariance as a Common Framework

Algorithmic Fairness		Distribution Shift	
Definition(s)	Name(s)	Definition(s)	Class of approach
$f(X) \perp A$	Demographic Parity	$Z \perp E$	Domain Adaptation
$f(X) \perp A \mid Y$	Equalized Odds	$Z \perp E \mid Y$	Conditional Domain Adaptation
$Y \perp A \mid f(X)$ $\mathbb{E}[Y \mid f(X), A] = f(X)$	Sufficiency Calibration	$Y \perp E \mid Z$	Invariant Risk Minimization