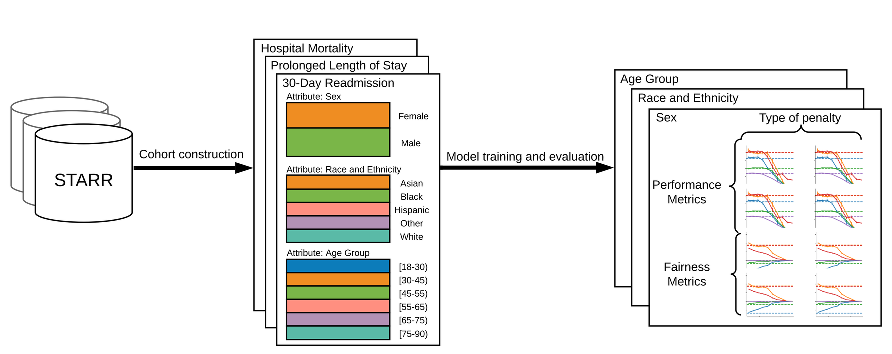
An Empirical Characterization of Fair Machine Learning for Clinical Risk Prediction

***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 study to characterize the impact of imposing group fairness on measures of model performance and fairness
- We find that group fairness penalties
  - Degrade model performance
  - 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 fairness criteria
  - *Upstream biases* due to misguided problem formulation or measurement error
  - *Downstream biases* defined in terms of disparate impact of an intervention

METHODS



- **Apply regularized learning objectives for *conditional prediction party***
- **Evaluate**
  - Conditional prediction parity
  - Relative calibration error
  - Cross-group ranking (xAUC)
  - Standard performance measures (AUROC, AP, etc)

1. **There is heterogeneity in trade-offs among measures of algorithmic fairness and model performance for patient-level prediction**
2. **We encourage researchers to step outside of the algorithmic fairness frame and engage critically with the broader sociotechnical context of machine learning in healthcare**

Preprint: tinyurl.com/fair-models

COHORTS

- **Databases**
  - STARR (Stanford)
  - Optum CDM
  - MIMIC-III (MIMIC-OMOP)
- **Target cohorts**
  - **STARR (198,644) / Optum (8,074,571)**
    - Hospital admissions lasting at least 24 hours
    - Index date at admission
    - 18+ at admission
  - **MIMIC-III (26,170)**
    - 24 hours after hospital admission associated with first ICU stay if 6 hours < ICU LOS < 12 days
- **Outcome cohorts**
  - **STARR/Optum**
    - Hospital mortality
    - LOS >= 7 days
    - 30-day Readmission
  - **MIMIC-III**
    - ICU LOS > 3/7 days
    - ICU/Hospital Mortality

RESULTS