FAMEWS: a Fairness Auditing tool for Medical Early-Warning Systems Reproduction Project

Fahad Zia

General Problem

- Searching for advancements in Early Warning Systems to predict adverse patient outcome for clinical settings
- Use cases include:
  - ICU monitoring
  - Cardiac arrest risk
  - Respiratory failure
- Biased predictions can lead to delay in proper care and unequal treatment
- Hypothesized that including demographics in training can impact outcome predictions

Paper and Approach

- FAMEWS: a Fairness Auditing Tool for Medical Early-Warning Systems (CHIL 2024)
- Used LightGBM with the ICU HiRiD dataset primarily
- Evaluated metrics including :
  - AUC
  - Demographic parity
  - Equalized odds
- Audited demographics including:
  - Age
  - Sex
  - Race
  - Admission type

Claimed Results

- Demographic factors affect:
  - Prediction accuracy
  - Model fairness
- Strong accuracy-fairness tradeoff was observed
- Bias patterns differed across demographic groups when included
- Fairness gaps varied by age and sex demographic features
- Finding were based on extensive HiRiD dataset

Reproduction

- Original HiRiD dataset is not publicly available
- Public diabetes dataset
  - ~350 patient records
  - 10 clinical features
- Demographics manually added include age, sex, and race
- Data split consisted of 70% training and 30% testing
- LightGBM (Light Gradient Boosting Machine)
  - Input of patient records with clinical features and demographics
  - Output of likelihood of adverse medical outcome

Setup and Results

- Converted to binary outcome classification
- Experiments included:
  - Baseline with no demographics
  - Demographics included
  - Ablation of age
- Retained same model and parameters across all tests
- Smaller fairness gaps compared to original research due to significantly smaller dataset
- Accuracy-fairness tradeoff was also observed

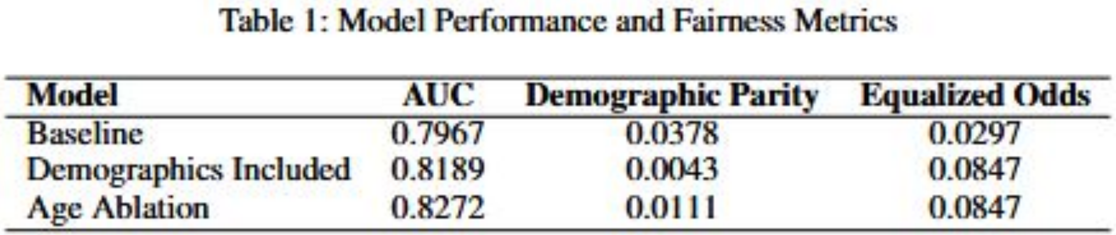

Table 1: Model Performance and Fairness Metrics

Model	AUC	Demographic Parity	Equalized Odds
Baseline	0.7967	0.0378	0.0297
Demographics Included	0.8189	0.0043	0.0847
Age Ablation	0.8272	0.0111	0.0847

Extensions/Ablations

- Alternative dataset using publicly available data
  - Tested FAMEWS framework in new environment
- Ablation of age in training
  - Led to higher accuracy when removed
  - Considering the significantly smaller dataset, manually adding demographics may have included noise in training

Thank you!

Citations

Hoche, M.; Mineeva, O.; Burger, M.; Blasimme, A.; and Ratsch, G. 2024. FAMEWS: a Fairness Auditing tool for Medical Early-Warning Systems. In Pollard, T.; Choi, E.; Singhal, P.; Hughes, M.; Sizikova, E.; Mortazavi, B.; Chen, I.; Wang, F.; Sarker, T.; McDermott, M.; and Ghassemi, M., eds., Proceedings of the fifth Conference on Health, Inference, and Learning, volume 248 of Proceedings of Machine Learning Research, 297–311. PMLR.

Y` eche, H.; Kuznetsova, R.; Zimmermann, M.; H¨ user, M.; Lyu, X.; Faltys, M.; and R¨ atsch, G. 2022. HiRID-ICU Benchmark– A Comprehensive Machine Learning Benchmark on High-resolution ICU Data. arXiv:2111.08536.