An Empirical Characterization of Fair Machine Learning For Clinical Risk Prediction

Stephen R. Pfohl, BS[1], Agata Foryciarz, BS[2], Nigam H. Shah, MBBS PhD[1]
[1]Stanford Center for Biomedical Informatics Research, Stanford School of Medicine, Stanford, California; [2]Stanford University, Department of Computer Science, Stanford, California

A preprint of the full-length version of this work is available at https://arxiv.org/abs/2007.10306

Abstract

The use of machine learning to guide clinical decision making has the potential to worsen existing health disparities. Several recent works frame the problem as that of algorithmic fairness, a framework that has attracted considerable attention and criticism. The appropriateness of this framework is unclear, due to both ethical as well as technical considerations, which include trade-offs between measures of fairness and model performance that are not well-understood for clinical predictive models. To inform the ongoing debate, we conduct an empirical study to characterize the impact of penalizing violations of group fairness on an array of measures of both model performance and group fairness. We repeat the analysis across several databases, clinical outcomes, and definitions of sensitive attributes. We find that procedures that penalize differences between the distributions of predictions across groups induce nearly-universal degradation of multiple performance metrics within the groups. We also evaluate the secondary impact of these procedures and observe heterogeneity in the effect of these procedures on measures of fairness in calibration and ranking across experimental conditions.

Research Category

patient-level prediction

Introduction

Considerable attention has been devoted to reasoning about the extent to which clinical predictive models can help anticipate and mitigate harms to advance health equity, while upholding ethical standards (1–7). The role that techniques of algorithmic fairness should have in addressing this aim is actively debated (4,8–10). These algorithmic fairness methods specify a mathematical formalization of a fairness criterion representative of an ideal (such as equal error rates for male and female patients), and provide procedures for minimizing violations of the fairness criterion without unduly deteriorating model performance (11–15). Given this formalization, it is necessary to reason about the trade-off between a model's performance measures and satisfaction of fairness criteria (16–18), as long as both can be appropriately contextualized.

To inform this discussion, we conduct a large-scale empirical study characterizing the trade-offs between multiple model performance measures and algorithmic fairness definitions for predictive models of clinical outcomes. Across twenty five combinations of datasets, clinical outcomes, and definitions of sensitive attributes, we train a series of predictive models that are penalized by varying degrees for violations of several fairness criteria. We report on the observed trade off between measures of model performance and violation of fairness criteria. A schematic describing this process is in Figure 1.

Methods

We extract cohorts from three databases in the OMOP common data model: the Stanford Medicine Research Repository (STARR) (19), Optum Clinformatics Data Mart (Optum CDM), and MIMIC-OMOP. We derive a cohort of inpatient admissions from STARR and Optum CDM, returning 198,636 patients in STARR and 8,073,395 patients in Optum CDM. In both cohorts, we define binary outcome labels for length of stay greater than 7 days and 30-day readmission, and include an additional label for in-hospital mortality in STARR cohort. In MIMIC-OMOP, we match the cohort and outcome definitions defined in the MIMIC-Extract (20) project, returning admissions from 26,170 patients labeled for ICU length of stay greater than 3 and 7 days, and hospital and ICU mortality. For the purposes of evaluating measures of group fairness, we define groups of the population on the basis of a combined race and ethnicity attribute, gender, and age group constructed via discretization.

We consider fully-connected feedforward neural networks for prediction. We train a series of models with a regularized objective that penalizes violation of three threshold-free group fairness criteria: equalized odds, equal opportunity (11) (*conditional penalties*), and demographic parity (12,21) (*unconditional penalty*). To assess violation of these criteria we compare the distribution of predictions for each group and the marginal distribution over the population. We use two ways of comparing these distributions -- the maximum mean discrepancy (MMD) or the squared difference in means. We evaluate the impact of penalizing these objectives over a range of penalty weights λ to assess trade-offs between as well as among measures of performance and fairness. We examine standard model performance measures, including the area under the ROC curve, average precision, and cross entropy loss. In addition, we evaluate a novel relative calibration measure that assesses the extent to which observed outcomes differ across groups conditioned on the risk score and measures of cross-group ranking accuracy (22,23).

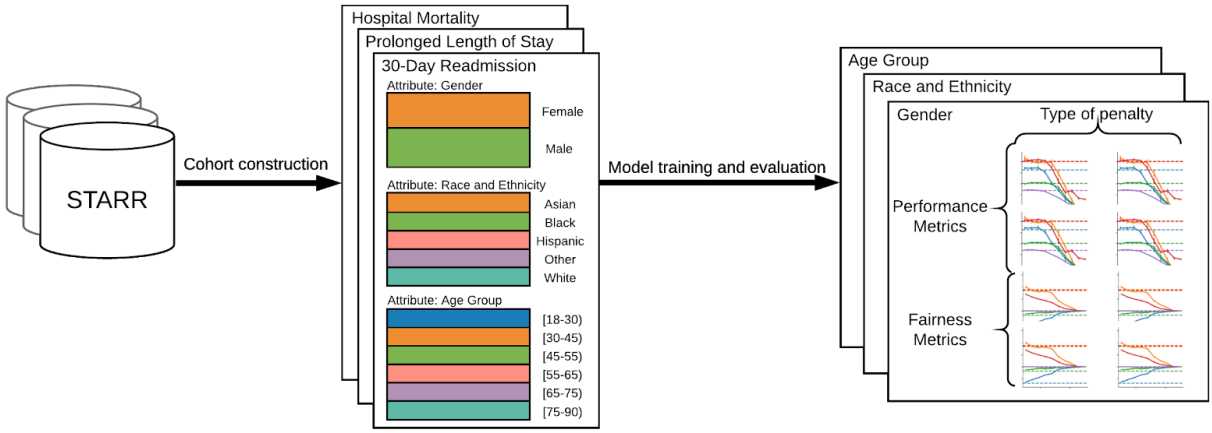

Figure 1. An overview of the experimental procedure for cohorts derived from the STARR database.

Results

The use of conditional regularization penalties universally degrades measures of model performance for all groups as well as introduces and exacerbates errors in relative calibration across groups. The effect of unconditional penalties are often similar to that of conditional penalties; however in some cases we observe improvements in model performance and calibration measures for at least one group along the trajectory of λ. The effect of algorithmic fairness procedures on cross-group ranking measures is heterogenous. The primary effect that we observe is a decline in cross-group ranking accuracy as λ increases, regardless of the type of penalty selected. In some cases, the trajectories of these measures are such that fairness is improved for one or more groups at the expense of other groups.

Conclusion

The debate on the use of algorithmic fairness techniques in healthcare has largely proceeded without empirical quantification of the effects of applying these techniques on predictive models derived from large-scale clinical data. We explicitly measure and comprehensively report on the extent of the empirical trade-offs between measures of model performance, conditional prediction parity, calibration, and ranking. Given our results, and the fundamental limitations of the algorithmic fairness framework, alternative approaches may be needed to enable proactive monitoring and auditing of the effects of intervening based on the output of clinical predictive models.

References

1. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med. 2018;169(12):866–72.

2. Goodman SN, Goel S, Cullen MR. Machine learning, health disparities, and causal reasoning. Ann Intern Med. 2018;169(12):883–5.

3. Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. Nat Med. 2020;26(1):16–7.

4. McCradden MD, Joshi S, Anderson JA, Mazwi M, Goldenberg A, Zlotnik Shaul R. Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. J Am Med Inform Assoc [Internet]. 2020 Jun 25; Available from: http://dx.doi.org/10.1093/jamia/ocaa085

5. Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care — Addressing Ethical Challenges [Internet]. Vol. 378, New England Journal of Medicine. 2018. p. 981–3. Available from: http://dx.doi.org/10.1056/nejmp1714229

6. Mccradden M, Mazwi M, Joshi S, Anderson JA. When Your Only Tool Is A Hammer: Ethical Limitations of Algorithmic Fairness Solutions in Healthcare Machine Learning. 2020;2020.

7. Ferryman K, Pitcan M. Fairness in precision medicine. Data & Society [Internet]. 2018; Available from: https://datasociety.net/wp-content/uploads/2018/02/Data.Society.Fairness.In_.Precision.Medicine.Feb2018.FINAL-2.26.18.pdf

8. Green B. The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness. :594–606.

9. Hutchinson B, Mitchell M. 50 years of test (un) fairness: Lessons for machine learning. Proceedings of the Conference on Fairness [Internet]. 2019; Available from: https://dl.acm.org/doi/abs/10.1145/3287560.3287600?casa_token=U1VcpOTF_00AAAAA:-7ubsvWtEpz3uM9DsUWZiQpNxQIpWVAqF2uA5OxMgTD5UCfaFI9cudxfF9_rNz3OJH-S3uek1fU

10. Fazelpour S, Lipton ZC. Algorithmic Fairness from a Non-ideal Perspective. 2020;(i):57–63.

11. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. Adv Neural Inf Process Syst. 2016;(Nips):3323–31.

12. Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C. Learning Fair Representations. In: International Conference on Machine Learning. 2013. p. 325–33.

13. Cotter A, Jiang H, Gupta MR, Wang S, Narayan T, You S, et al. Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals. J Mach Learn Res. 2019;20(172):1–59.

14. Agarwal A, Beygelzimer A, Dudík M, Langford J, Wallach H. A Reductions Approach to Fair

Classification [Internet]. arXiv [cs.LG]. 2018. Available from: http://arxiv.org/abs/1803.02453

15. Song J, Kalluri P, Grover A, Zhao S, Ermon S. Learning Controllable Fair Representations [Internet]. arXiv [cs.LG]. 2018. Available from: http://arxiv.org/abs/1812.04218

16. Chouldechova A. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments [Internet]. Vol. 5, Big Data. 2017. p. 153–63. Available from: http://dx.doi.org/10.1089/big.2016.0047

17. Kleinberg J, Mullainathan S, Raghavan M. Inherent Trade-Offs in the Fair Determination of Risk Scores [Internet]. arXiv [cs.LG]. 2016. Available from: http://arxiv.org/abs/1609.05807

18. Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ. On Fairness and Calibration. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems 30. Curran Associates, Inc.; 2017. p. 5680–9.

19. Datta S, Posada J, Olson G, Li W, O'Reilly C, Balraj D, et al. A new paradigm for accelerating clinical data science at Stanford Medicine [Internet]. arXiv [cs.CY]. 2020. Available from: http://arxiv.org/abs/2003.10534

20. Wang S, McDermott MBA, Chauhan G. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. Proceedings of the [Internet]. 2020; Available from: https://dl.acm.org/doi/abs/10.1145/3368555.3384469?casa_token=rcGLJo59yoEAAAAA:96ge1z5rghi_sVWhD1BPkOsg5GYHDLClCRS2Ao5Ozmfgt2W3r0iwbiWNgspEBfc6CGQpSHIEJAz2_Yc

21. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness. ITCS 2012 - Innovations in Theoretical Computer Science Conference. 2012;214–26.

22. Kallus N, Zhou A. The Fairness of Risk Scores Beyond Classification: Bipartite Ranking and the XAUC Metric. In: Wallach H, Larochelle H, Beygelzimer A, d\textquotesingle Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems 32. Curran Associates, Inc.; 2019. p. 3438–48.

23. Beutel A, Chen J, Doshi T, Qian H, Wei L. Fairness in recommendation ranking through pairwise comparisons. Proceedings of the 25th [Internet]. 2019; Available from: https://dl.acm.org/doi/abs/10.1145/3292500.3330745?casa_token=Gk0sPzJiIFIAAAAA:agBYZGoWSH3FOJr44_rEvkZHUh113rrFGsu0uoPjaJFfvykRi4bFp53RDPyKmqSzbEk0Gmy0ON0