An Approach to Identify and Mitigate Algorithmic Performance Biases in Sepsis Prediction Models

Adam Kotter[1][0000-0002-5071-0399]

[1] University of Utah, Salt Lake City UT 84108, USA
adam.kotter@utah.edu

Abstract. While clinical machine-learning (ML) methods offer tremendous opportunities to advance healthcare, existing disparities in health outcomes may be exacerbated if care is not taken in the development of ML-based clinical decision support tools. In spite of this, comparatively little effort has been made to systematically mitigate or even identify performance disparities in clinical ML models. This project aims to develop a systematic approach for identifying and mitigating performance disparities in ML models used for clinical prediction tasks so that disparities in health outcomes may be reduced.

Keywords: Healthcare disparities, Predictive analytics in healthcare, Clinical decision support systems (CDSSs).

1 The Importance of Addressing Algorithmic Bias

Machine learning (ML) and other forms of artificial intelligence (AI) stand to greatly improve the practice of medicine by uncovering new insights, supporting clinical decision making, and automating tasks [1]. However, if care is not taken in the development and implementation of ML-based clinical decision support (CDS) systems, disparities in healthcare outcomes will likely be exacerbated and vulnerable populations may be harmed [2-5]. Identifying and mitigating performance disparities is an essential part of addressing algorithmic bias.

ML models acquire the biases inherent in the data on which they are trained. Healthcare utilization, practice, and outcomes vary across population groups, so potential disparities often reflect as biases in the data used to train clinical algorithms [1]. These biases can cause harm. A recent study found that AI can "propagate harmful, inaccurate, race-based" clinical information [6]. Other studies found that a broad range of clinical predictive ML algorithms trained on commonly used datasets significantly underperformed for females and for racial/ethnic minorities by as much as 0.3 AUROC and 20% recall [7,8]. Systematic underperformance of clinical ML algorithms may lead to unidentified illness, overuse of risky interventions, and even clinician dismissal of important alerts about patients in minority groups [5]. The American Medical Association (AMA) has laid out proactive identification and mitigation of bias in AI algorithms as a key principle to mitigate patient and physician risk [9]. Thus, addressing

algorithmic bias is critical to ensure equitable implementation of AI in healthcare and improve outcomes among patients from historically marginalized groups.

There have been few efforts to address algorithmic bias despite its importance [10]. Even studies that do address algorithmic performance disparities often suffer from systematic weaknesses. Most studies that identify disparities do so in a post-hoc fashion with models that were not designed with disparity mitigation in mind from the ground up [7,8,11,12]. Meta-analyses of disparity-mitigation or -identification studies are not possible due to heterogenous methodology [10]. Systematic methods are needed that include bias detection and mitigation at every step of ML-based CDS development.

2 What is Already Known?

Different forms of bias can affect the training data of ML models [5]. These training data biases can be broadly categorized into two types: a lack of representation (data poverty)[3] and a lack of informativeness. Lack of representation can be described as a lack of available data and may lead ML models to miss patterns in underrepresented groups, such as racial minorities. Lack of informativeness can be described as having data that is missing important information and can result from issues including 1) missing data about social determinants of health (SDoH) and 2) unclear or inaccurate data. Missing SDoH data can lead to model blindness about conditions affecting patient outcomes, such as living in a food desert. Unclear or inaccurate data may lead to inaccurate models. For example, patients who don't share a language with healthcare providers may appear to have erratic vital signs as a result of nervousness, and identifying melanoma on dark skin may be more difficult than on light skin [5]. Addressing these sources of bias in model development is essential to avoid disparities mediated by ML-based CDS.

3 Research Questions and Goal

The goal of my research is to develop a generalizable approach for detecting and mitigating performance disparities in ML-based prediction models. This research will answer two important questions for equitable use of ML in CDS: 1) can deliberate detection during model development be used to assess algorithmic performance disparities and their potential causes (Aim 1); and 2) can rebalancing representation in datasets and inclusion of SDoH data mitigate performance disparities during model development (Aim 2).

4 Planned Approach and Methods

The approach for this research plan can be divided into two conceptual steps: disparity analysis (Aim 1) and mitigation (Aim 2). In Aim 1, I will identify performance disparities and their potential sources in a sepsis-prediction ML model. In Aim 2, I will redevelop the ML model using disparity mitigation approaches targeted at the identified

potential sources of disparities. These two steps may provide the foundation for a future framework for ground-up bias mitigation: train low-effort preliminary models to identify potential biases, then implement bias-mitigation strategies targeted at the identified potential biases throughout the development of the final models.

The data used for this research will be obtained from MIMIC-III, a publicly available deidentified dataset of patients admitted to critical care units at Beth Israel Deaconess Medical Center in Boston, Massachusetts. MIMIC-III has been widely used for ICU sepsis prediction research and serves as an excellent benchmark. Future work beyond this planned research will test the generalizability of the findings of my research to other datasets.

MIMIC-III contains information about various socioeconomic variables. I will use five of these SDoH features as model inputs and sampling categories at different stages of my research: race/ethnicity (combined into one feature in MIMIC-III), language, age, sex, and insurance status. Performance disparities will be evaluated according to just two of these demographic categories: race/ethnicity and language.

4.1 Establish Baseline

I do not have any theoretical or experimental rationale to suspect that one type of ML model will perform better than others, so I will select an ML model for this research project based on benchmarking evaluation of a panel of commonly used model architectures. I will use established methods of hyperparameter optimization and cross validation to ensure that each model is performing optimally. The model with the highest AUROC will be selected for further experiments and hypothesis testing.

For each subgroup of each of the two evaluated demographic categories, I will compare the AUROCs and positive predictive values (PPVs) at fixed sensitivity of the selected model to check for disparities. The Friedman chi-square test will be used to assess differences at a significance level corrected for multiple comparisons. For significant differences, I will perform the Nemenyi post-hoc test on the subgroup performances for the given metric and demographic category to assess pairwise disparities between each subgroup. For each significant pairwise disparity, I will assess possible causes (Aim 1) and identify mitigation strategies (Aim 2).

4.2 Aim 1 - Assess Disparities and their Potential Sources in Sepsis Prediction Algorithm Performance.

I will assess potential sources of the performance disparities identified in benchmark experiments. This aim focuses on the two main types of training data biases: lack of representation and lack of informativeness.

Lack of Representation. I will assess the extent to which data imbalance contributes to model performance disparities by testing if artificially equalizing representation reduces disparities. For each of the five SDoH features, I will retrain the model on new datasets produced by using the Synthetic Minority Oversampling Technique (SMOTE)

to generate representative synthetic data for the underrepresented classes of the SDoH feature. For example, a dataset with oversampled insurance status would have equal representation of patients with private insurance, Medicare, Medicaid, and no insurance.

I will test the retrained models on testing sets that have not been resampled and conduct comparisons for the pairwise disparities in question. The Wilcoxon signed-rank test will be used to test the significance of the differences between the baseline disparities and the disparities in the retrained models. Because the only change between the baseline model and the retrained models is the use of oversampling, if a significant pairwise performance disparity is significantly reduced, then I will conclude that the original disparity is at least partially associated with lack of representation in the resampled SDoH feature to the degree that the disparity was reduced. If instead a disparity is significantly exacerbated by minority oversampling, I will conclude that the original disparity was mitigated by lack of representation, which may indicate that a minority class of that SDoH feature enjoys some privilege that results in improved model predictive performance for that minority class. I will use these conclusions to decide which SDoH features will be oversampled in Aim 2.

Lack of Informativeness. I will assess the extent to which lack of informativeness contributes to model performance disparities by testing if addition of SDoH information reduces disparities. For each of the five SDoH features, I will retrain and test the models on new datasets generated by adding the given SDoH as a model input feature. I will perform statistical tests and draw conclusions as above while noting in my conclusions that the added SDoH features may be proxies for a more informative correlated variable or variables. I will use these conclusions to decide which SDoH features will be included as model features in Aim 2.

4.3 Aim 2 - Mitigate Identified Sepsis Prediction Model Performance Disparities.

I will create methods to develop predictive models with disparity mitigation targeted at the potential disparity causes identified in Aim 1. This aim focuses on three main steps: (1) mitigate lack of representation disparities using targeted rebalancing of training data, (2) mitigate lack of informativeness disparities using targeted addition of SDoH features, and (3) mitigate overall disparities using a combination of targeted minority oversampling and targeted SDoH feature addition.

Lack of Representation. I will mitigate the baseline model performance disparities associated with lack of representation by using SMOTE to simultaneously oversample all SDoH features identified as contributing to the respective performance disparities. For each pairwise disparity identified as associated with lack of representation in Aim 1, I will retrain the model on new datasets produced by using simultaneous SMOTE oversampling on each underrepresented class of each SDoH feature identified as contributing to the disparity through lack of representation. This may require grouping of

the datasets into intersectional subclasses to oversample each group simultaneously. I will test the retrained models and perform statistical analyses as in Aim 1. I will also compare overall model performances between the baseline and the retrained model to see if overall model performance changed with the resampling. If the overall model performance significantly increased, I will conclude that data rebalancing is beneficial even without considering disparity reduction. If the overall model performance is significantly reduced, I will conclude that data rebalancing induces a tradeoff between disparity reduction and overall model performance.

Lack of Informativeness. I will mitigate the baseline model performance disparities caused by lack of informativeness using targeted inclusion of SDoH features identified in Aim 1 as associated with disparities. For each pairwise disparity with significant reduction from inclusion of SDoH features in Aim 1, I will retrain the model on data with inclusion of each SDoH feature identified as contributing to the disparity as model features. This may result in different models being trained on different features in different contexts. Statistical analyses will be performed as in Aim 1 to determine if the baseline disparities were reduced. I will also compare overall performance between the baseline and retrained model to check for tradeoffs as above.

Combination Method. I will simultaneously mitigate the baseline model performance disparities associated with lack of representation and associated with lack of informativeness using targeted SMOTE rebalancing and SDoH feature inclusion. I will create new datasets with simultaneous oversampling and inclusion of SDoH model features as above. I will train the model on these datasets and test them on datasets without any resampling. I will perform statistical analyses and compare performance as above.

5 Expected Results

My aim with this research is to produce an approach to systematically identify and mitigate bias from the ground up in ML tools for clinical prediction. I expect to develop a robust, generalizable approach for detecting and mitigating performance disparities during model development, which will enable equitable ML models in CDS and further research into healthcare disparity mitigation. Consistent application of this approach to develop ML tools for predicting clinically relevant events would reduce performance disparities in these ML tools.

6 Questions

Which courses, if any, would I need to take to carry out this work? What courses would be recommended but not necessary to enhance this work?

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Dorr, DA., Adams, L., Embí, P.: Harnessing the Promise of Artificial Intelligence Responsibly. JAMA. **329**(16):1347-1348. doi: 10.1001/jama.2023.2771.
2. Vyas, DA., Eisenstein, LG., Jones, DS.: Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. New England Journal of Medicine 2020 Vol. 383 Issue 9 Pages 874-882, DOI: 10.1056/NEJMms2004740
3. Ibrahim H., Liu X., Zariffa N., et al.: Health data poverty: an assailable barrier to equitable digital health care. The Lancet Digital Health, doi.org/10.1016/S2589-7500(20)30317-4
4. Matheny, M., Israni, ST., Ahmed M., et al.: Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. NAM Special Publication. Washington, DC: National Academy of Medicine.
5. Rajkomar, A., Hardt, M., Howell, MD., et al.: Ensuring Fairness in Machine Learning to Advance Health Equity. Ann Intern Med. 169(12):866-872. doi: 10.7326/M18-1990.
6. Omiye, JA., Lester, JC., Spichak, S,. et al.: Large language models propagate race-based medicine. npj Digit. Med. 6, 195 (2023). https://doi.org/10.1038/s41746-023-00939-z
7. Wang, H., Li, Y., Naidech, A., Luo, Y.: Comparison between machine learning methods for mortality prediction for sepsis patients with different social determinants. BMC Med Inform Decis Mak. 22(Suppl 2):156. doi: 10.1186/s12911-022-01871-0.
8. Straw, I., Wu, H.: Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction. BMJ Health Care Inform. 2022 Apr;29(1):e100457. doi: 10.1136/bmjhci-2021-100457.
9. Mills, RJ.: AMA issues new principles for AI development, deployment & use. https://www.ama-assn.org/press-center/press-releases/ama-issues-new-principles-ai-development-deployment-use. Nov 28, 2023.
10. Taber, P., Armin, JS., Orozco, G., et al.: Artificial Intelligence and Cancer Control: Toward Prioritizing Justice, Equity, Diversity, and Inclusion (JEDI) in Emerging Decision Support Technologies. Curr Oncol Rep. 25(5):387-424. doi: 10.1007/s11912-023-01376-7.
11. Reese, TJ., Schlechter, CR., Potter, LN., et al.: Evaluation of Revised US Preventive Services Task Force Lung Cancer Screening Guideline Among Women and Racial/Ethnic Minority Populations. JAMA Netw Open. 4(1):e2033769. doi: 10.1001/jamanetworkopen.2020.33769.
12. Economou-Zavlanos, NJ., Bessias, S., Cary, MP., et al.: Translating ethical and quality principles for the effective, safe and fair development, deployment and use of artificial intelligence technologies in healthcare. J Am Med Inform Assoc 2023, doi: 10.1093/jamia/ocad221