When to Use Demographic Data in Healthcare Models: A Bias-Responsible Approach

Sebrina Zeleke  
The Ohio State University  
zeleke.8@osu.edu

Tanya Berger-Wolf  
The Ohio State University  
berger-wolf.1@osu.edu

Xia Ning  
The Ohio State University  
ning.104@osu.edu

Abstract—Given AI's increasing role in healthcare, it is vital to ensure that created models neither perpetuate nor introduce new biases. One of the naive approaches to mitigating bias is omitting demographic data features during model training. However, in healthcare, this method might not yield the best-performing models as these features may contain crucial care-related information. This paper explores the trade-offs between optimal performance and algorithm bias linked to using demographic data. We demonstrate the approach using a healthcare model that predicts ICU readmission risk of patients.

Index Terms—Healthcare, Fairness, Machine Learning, Artificial Intelligence, ICU Readmission Risk

I. INTRODUCTION

AI systems such as machine learning (ML) models are transforming various industries, and healthcare is no exception. In every context these systems are used, including healthcare, they raise the concern of bias against different demographic subgroups. In healthcare, ML models have been utilized for diagnosing various diseases, such as cancer [2], and most recently COVID-19 [1]. They have also been used for prediction, including patients' Intensive Care Unit (ICU) readmission risk, mortality, and ICU length of stay [4], [6]. As the use of ML in healthcare increases so does the concern to ensure that the developed models do not perpetuate existing biases or create new ones [3], [5]. In this paper, we examine the bias connected to using demographic data in ML for healthcare by evaluating the impact on model performance and fairness of including or withholding demographic data.

One naive approach to mitigate ML biases is to exclude features that might aid in identifying an individual, such as race, gender, and insurance type, from the training data, ensuring the model doesn’t explicitly use such features for predictions [7]. However, this approach may not consistently yield the best performance and can be ineffective in preventing bias as these features may provide valuable information, particularly in healthcare. Lin et al. [4] demonstrated that incorporating all demographic information enhanced predictive models performance for ICU readmission risk. In contrast, including such demographic features might introduce additional bias. To understand this trade-off between performance and bias, we developed a framework for deciding when to use demographic data as input using Lin et al.'s model for demonstration [4].

Specifically, the paper explores the models presented by Lin et al. [4] to investigate whether the increased performance after using demographic data is consistent across all patients. We systematically explored the trade-off for each demographic variable and their combinations by comparing two identical models that differ only in whether they used particular demographic information.

II. METHOD

In the work done by Lin et al., the authors used supervised machine learning models to predict ICU readmission risk using patients' clinical data. They tested several models, including the Long Short Term Memory (LSTM), Convolutional Neural Network (CNN), and a hybrid combination of the two. For input data, they tested different time series windows of the medical data, finding that the last 48 hours (L-48) before transfer/discharge data resulted in the best-performing models. Additionally, they evaluated whether adding demographic information boosted the performance of the model, finding positive results. [4]

For our analysis, we took two LSTM models with the L-48 data from the work done by Lin et al. [4] as base models where the only difference between the two is the incorporation of demographic data. The base model was the most explored and showed the third highest performance improvement with the inclusion of demographic data in the original work. We refer to the model with demographic information as \( \mathcal{D} \) and the one without it as \( \mathcal{D} \).

The original model by Lin et al. [4] utilizes True Positive Rates (TPR) for reporting results, and we adopt the same metric to examine performance and bias for two primary reasons. First, it is used to maintain consistency with the original work because it allows us to measure disparity using the originally intended metric. Second, assuming that a true positive prediction gets the benefit of extended care due to the high risk of readmission, TPR allows us to gauge the classification effectiveness of the models and assess whether the inclusion of demographic data has increased or decreased the disparity of such benefit.

To examine introduced bias resulting from the use of demographic data, we compute the TPR of model \( \mathcal{D} \) for reporting results, and we adopt the same metric to examine performance and bias for two primary reasons. First, it is used to maintain consistency with the original work because it allows us to measure disparity using the originally intended metric. Second, assuming that a true positive prediction gets the benefit of extended care due to the high risk of readmission, TPR allows us to gauge the classification effectiveness of the models and assess whether the inclusion of demographic data has increased or decreased the disparity of such benefit.

To examine introduced bias resulting from the use of demographic data, the TPR of model \( \mathcal{D} \) is computed for different demographic subgroups, and compared to the TPR of model \( \mathcal{D} \) for the same groups. The TPR for each model is derived by averaging the TPR values obtained through a 5-fold cross-validation. The difference of these TPRs between the model \( \mathcal{D} \) and \( \mathcal{D} \) is then used to measure the disparity of
Fig. 1: TPR differences of model \(WOD\) and \(WD\): 1a for gender, ethnicity, and Insurance separately; 1b, 1c, and 1d for intersectional groups (Insurance, Ethnicity), (Insurance, Gender), and (Gender, Ethnicity), respectively, where F: Female, M: Male, G: Government, M-i: Medicaid, M-r: Medicare, P: Private are different insurance groups and N: No Data, B: Black, H: Hispanic, A: Asian, W: White are different ethnicity groups.

As shown throughout this paper, depending solely on a single metric for reporting can obscure nuanced information, especially in the area of algorithmic fairness. For an increased overall performance of roughly 2 percent TPR, the figures above show the kind of benefit disparity that could be introduced. Such disparity could be attributed to inherent historical biases, systemic biases, or algorithmic biases, prompting the need for additional research to distinguish between these factors. Depending on the application, the acceptable trade-off and bias could differ, but these kinds of analyses allow us to understand such trade-offs before making decisions.

This paper presented the result of an analysis that looked to examine the trade-offs between optimal performance and algorithm bias linked to using demographic data. It is important to understand that the use of demographic information does not always increase benefits for all protected groups uniformly. This analysis is key to assessing the trade-off between performance and bias and can be used to decide whether or not to use demographic information.

**REFERENCES**


