Informatics in Medicine Unlocked 54 (2025) 101627

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Learning unbiased risk prediction based algorithms in healthcare: A case study with primary care patients

Vibhuti Gupta [a,*,1], Julian Broughton [b,1], Ange Rukundo [a], Lubna J. Pinky [b]

[a] Department of Computer Science and Data Science, School of Applied Computational Sciences, Meharry Medical College, 1005 Dr DB Todd Jr Blvd, Nashville, 37208, TN, USA

[b] Department of Biomedical Data Science, School of Applied Computational Sciences, Meharry Medical College, 1005 Dr DB Todd Jr Blvd, Nashville, 37208, TN, USA

ARTICLE INFO

Dataset link: https://gitlab.com/labsysmed/dissecting-bias

Keywords:
Artificial intelligence
Trustworthy AI
Bias
Fairness

ABSTRACT

The proliferation of Artificial Intelligence (AI) has revolutionized the healthcare domain with technological advancements in conventional diagnosis and treatment methods. These advancements lead to faster disease detection, and management and provide personalized healthcare solutions. However, most of the clinical AI methods developed and deployed in hospitals have algorithmic and data-driven biases due to insufficient representation of specific race, gender, and age group which leads to misdiagnosis, disparities, and unfair outcomes. Thus, it is crucial to thoroughly examine these biases and develop computational methods that can mitigate biases effectively. This paper critically analyzes this problem by exploring different types of data and algorithmic biases during both pre-processing and post-processing phases to uncover additional, previously unexplored biases in a widely used real-world healthcare dataset of primary care patients. Additionally, effective strategies are proposed to address gender, race, and age biases, ensuring that risk prediction outcomes are equitable and impartial. Through experiments with various machine learning algorithms leveraging the Fairlearn tool, we have identified biases in the dataset, compared the impact of these biases on the prediction performance, and proposed effective strategies to mitigate these biases. Our results demonstrate clear evidence of racial, gender-based, and age-related biases in the healthcare dataset used to guide resource allocation for patients and have profound impact on the prediction performance which leads to unfair outcomes. Thus, it is crucial to implement mechanisms to detect and address unintended biases to ensure a safe, reliable, and trustworthy AI system in healthcare.

1. Introduction

The widespread use of Artificial Intelligence (AI) technologies in data-driven decision-making systems has become increasingly popular because of their remarkable predictive capabilities. These systems have made significant advancements across various sectors, specially in the field of medicine. While these AI-based systems are effective in making important life-changing decisions, it is of the utmost importance to ensure that these decisions do not reflect discriminatory behavior towards certain individuals or groups. Recent findings of AI applications in medical field indicate that AI can lead to both biased, and erroneous decisions with complete lack of transparency in sensitive safety-critical scenarios, while potentially enhancing already existing biases against marginalized groups and exacerbating inequities [1]. This has raised concerns among clinicians, policymakers and patients, and hence resulted in a decline in AI's trustworthiness and commercialization of these systems despite it is predictive power [2]. Therefore, it has become essential to make these systems safe, reliable, and trustworthy in order to utilize the effectiveness of AI technologies in full capacity. Recently, several requirements related to transparency, ethics and legal issues, such as explainability, accountability, reliability, data privacy, and fairness, have been proposed in this direction to make these systems trustworthy. In addition, researchers in the field have developed a wide variety of fairness-enhanced classifiers and fairness matrices in traditional machine learning and deep learning setting to address these issues. Nevertheless very few such techniques have been translated into the real-world practice of data-driven decisions [3].

Existing clinical AI methods are often biased to specific ethnic groups or subpopulations in their predictions or mirror human biases in decision making. These biases can be categorized into data-driven bias and algorithmic bias. Data-driven biases [4] occur when the data do not

* Corresponding author.
E-mail address: vgupta@mmc.edu (V. Gupta).
[1] Vibhuti Gupta and Julian Broughton as equal first authors.

https://doi.org/10.1016/j.imu.2025.101627
Received 6 October 2024; Received in revised form 13 February 2025; Accepted 14 February 2025
Available online 26 February 2025

reflect the true distribution of population including enough samples for a specific ethnicity/race [5], data quality considerations (e.g., data errors and omissions in the original data entry process, non-standardized data, lack of metadata, inaccurate data annotation), which affects data training and biased predictions. However, algorithmic bias [6] occurs when algorithms have unfair outcomes due to their training on the data reflecting inherit bias that exist in the history of our world (i.e., societal prejudices, power imbalances), class imbalance, non-inclusion of some variables such as age, sex, socioeconomic status, social determinants of health factors (SDOH) etc.

Although there are some works [7–13] done in identifying and mitigating bias in healthcare datasets, the thorough assessment of fairness and biases in these data and models is still lacking. These biases can be devastating for an AI-based system if gone unchecked. Therefore, it is essential to thoroughly examine these issues and create computational methods that can analyze and address biases effectively. Due to the greater need to develop trustworthy AI systems and minimize harm due to existing biases in the dataset and algorithms, it is crucial for guaranteeing that the algorithms perform consistently and accurately across different patient populations. Furthermore, understanding these biases can help pinpoint their sources and develop strategies to address them.

The objective of this paper is to explore different types of biases in the dataset and algorithms, identify the potential features that are the sources of bias in the outcomes, propose solutions to mitigate these biases, and finally compare the effect of biased and unbiased data and algorithms on the predictive performance. The comparison of biases during pre-processing, in-processing, and post-processing is crucial to get a complete picture of this problem in a complex healthcare domain. Thus, we critically analyze this problem in this paper with detailed description of the data and algorithmic biases and demonstrate that using an open-source healthcare dataset [14] and applying a widely used algorithmic fairness tool (Fairlearn) developed by Microsoft [15]. Our major contributions in this paper can be summarized as follows:

1. Revisiting a widely used real-world healthcare dataset of primary care patients and thoroughly examining various types of data and algorithmic biases during both pre-processing and post-processing phases to uncover additional, previously unexplored biases.
2. Proposing effective strategies to mitigate gender, racial, and age biases, ensuring that the risk-prediction results are fair and unbiased.
3. Leveraging an open-source algorithmic fairness tool (Fairlearn) to identify and mitigate biases in the dataset and evaluate the impact of these biases on risk prediction through various evaluation metrics.
4. Providing valuable practical insights that can benefit computer scientists and healthcare professionals working in the area of AI in healthcare.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 describes the dataset followed by data description and analysis models in Sections 4 and 5. Section 6 describes the results and Section 7 concludes the paper.

2. Related work

2.1. AI fairness tools to mitigate bias

There are several AI fairness tools exist in the research community. These tools are used to identify, measure, and mitigate bias in the datasets. These tools also provide features to compare the impact of these biases on the predictive outcomes. Below are some of the major tools used in the industry and research community:

- **IBM's AI Fairness 360 Toolkit**: It is a comprehensive Python open source toolkit focusing on technical solutions through fairness metrics and algorithms to mitigate bias in datasets at the pre-processing and model training stages [16].
- **Aequitas** is an open-source Python library, designed to enable developers to test ML models for a list of bias and fairness metrics in relation to multiple population sub-groups [17].
- **Google's What-If Tool**: This is an open-source application that let users to explore a models' performance on a dataset, including examining several preset definitions of fairness constraints such as, equality of opportunity. This tool is interesting as users can explore as well as visualize counterfactuals against different subsets of the same input data [18].
- **Microsoft's fairlean.py**: This is Python package developed by Microsoft that contains numerous algorithms to mitigate "unfairness" in supervised machine learning [15].

2.2. Bias identification and mitigation in healthcare

Recently, significant efforts have been made to identify and mitigate bias in healthcare datasets [7–13]. There are many studies associated with the classical Framingham risk score [19] which is developed based on the data obtained from the Framingham Heart Study [7]. The Framingham risk score is a risk prediction algorithm developed to estimate an individual's 10-year cardiovascular risk, which was further refined in 2008 by incorporating additional cardiovascular conditions. Careful examination of the Framingham risk score revealed that it is biased towards certain races and ethnicities [20]. Predicting the risk of cardiovascular events in nonwhite populations using data from the Framingham Heart Study [7] leads to biased results, causing both overestimation and underestimation of associated risks [8]. This study is similar to ours as it is associated with the bias examination of risk-prediction algorithms.

Dermatology algorithms are found biased towards fair-skinned patients which leads to biased results as shown in [9]. Age and sex differences are reported in chronic conditions such as diabetes, cardiovascular disorders, neurological diseases [10], cancer [11], mental health disorders [12], and autoimmunity [13] for algorithmic bias.

There are some works done in mitigating inherent biases in the healthcare datasets [21–26]. Pre-processing mitigation approaches including reweighing, resampling, and blinding are used in [21–24] to mitigate racial, label, and systematic biases in the healthcare datasets. In-processing and post-processing approaches such as dynamic reweighing and transformation are used in [25,26] to mitigate systematic and racial biases.

Obermeyer's study [14] is a classical study examining racial bias in risk prediction algorithms. This study discovered disparities in an automated screening algorithm implemented across multiple healthcare centers and utilized in health insurance plans. The algorithm used to predict risk underestimated the risk for Black patients compared to White patients. Our work in this paper revisits the dataset used in [14] however the goal differs from [14]. Authors in [14] focused on identifying the label bias in the risk-prediction algorithm and demonstrated it with their analysis results. However, we examine biases in the data and algorithms through detailed exploratory and statistical analysis, highlight disparities in the risk prediction model outcomes, and address these biases using various techniques.

One of the major challenges in the modern application of artificial intelligence (AI) in healthcare is the presence of bias in the datasets used to train these models. These datasets often underrepresented minority groups, including women[27], people of color [28], and individuals from lower socioeconomic backgrounds [29]. Such skewed representations can result in AI models that produce inaccurate predictions and perpetuate disparities in healthcare outcomes for these populations. Although it is widely acknowledged that AI models can

inherit biases from their source datasets, there remains a lack of robust methods for quantifying these biases. Our review revealed that while many studies discuss the presence of bias, they frequently fail to report tools or metrics used to measure it [30]. This absence of quantitative evidence makes it more difficult to assess, address, and ultimately mitigate these biases [31]. Moreover, both academic and industry stakeholders often require demonstrable evidence to invest in efforts to improve fairness in AI systems. The ability to quantify bias not only substantiates its existence but also underscores the urgency of implementing mitigation strategies. Several studies also emphasize the need for consistent standards to evaluate the performance of AI algorithms and the diversity of datasets.

In addition to the scarcity of quantitative methodologies, there is limited exploration of how multiple forms of bias intersect within a single AI model or dataset. Most studies tend to focus on isolated biases, such as racial or gender bias, without addressing the co-occurrence of multiple biases – such as race, age, and gender – within a single dataset [27,30,31]. Addressing these intersectional biases is critical for understanding their compounded effects on vulnerable populations. A comprehensive approach to bias mitigation is necessary to ensure that AI models used in healthcare applications do not exacerbate existing inequities, but instead promote equitable and accurate outcomes for all individuals.

3. Materials and methods

This section provides a comprehensive overview of the study design. Initially, we will present the dataset used and the preprocessing procedures to prepare the data. Subsequently, the bias identification and mitigation techniques will be briefly discussed, followed by the fairness evaluation metrics used in the paper.

3.1. Dataset and preprocessing

This paper utilizes the dataset from Obermeyer paper [14], a well-known study on the identification of racial bias in the risk-prediction algorithms used to identify the primary care patients requiring special care programs. Since the data used in the study [14] are protected health information, the authors provided a synthetic version of dataset with similar characteristics. The dataset consists of primary care patients enrolled in risk-based contracts from 2013 to 2015. This dataset is of 48,784 patients with 160 variables. The key attributes included are demographic variables, algorithm generated risk scores, cost variables, medication variables, and various health metrics.

The variables were grouped into categories as shown below.

- **Variables at time** t: A vector of "outcome" for a given calendar year (t): cost, health, program enrollment, and the commercial risk score.
- **Demographic variables**: This consists of race, age, and gender of the patients.
- **Comorbidity variables at time** $t-1$: A vector of indicators for specific chronic comorbidities (illnesses) that were active in the previous year, and their sum.
- **Cost variables at time** $t-1$: Costs claimed from the patients' insurance payer, rounded to the nearest $100 and broken down by type of cost, over the previous year.
- **Biomarker/medication variables at time** $t-1$: A set of indicators capturing normal or abnormal values (or missingness) of biomarkers or relevant medications, over the previous year.

There are a total of 9 demographic variables, 34 comorbidity variables, 13 cost variables, 94 biomarker/medication variables, and 10 variables at time t. We pre-processed the data after extraction which includes data cleaning, column renaming, and sorting, of the original dataset to streamline the analysis process. Each demographic analysis involved preprocessing datasets to ensure data completeness, with imputation for missing values.

3.2. Types of biases in the machine learning workflow

We have employed various bias identification and mitigation techniques in this paper to evaluate the effectiveness of identifying and mitigating bias in the given dataset. The bias identification can be categorized into pre-processing, in-processing, and post-processing biases inherent in the data and models during the data analysis. This paper only focuses on the pre-processing and post-processing bias identification and mitigation.

3.2.1. Pre-processing bias

This is a type of bias that arises usually earlier in the machine learning process usually during the pre-processing stage. The pre-processing bias typically stems from how the data is collected and its intended usage. The lack of carefulness and attentiveness to details when collecting data and going through pre-processing methods can lead to devastating outcomes for protected attributes such as race, age, and gender and have devastating outcomes for specific groups. Identifying biases at the pre-processing stage requires detailed data description and visualization of variables during the exploratory data analysis to determine if any variable affects the machine learning model outcome due to skewness or errors.

3.2.2. Post-processing bias

These are the type of biases that arises usually during post-processing stages of the machine learning process. The post-processing biases arise after model training is complete and act on the results. The post-processing biases come under the category of algorithmic bias which in a modern-day processes can be a challenging task as it requires interpretability of models and a collaboration between the AI-designers and end users. Identifying these biases requires a thorough check of the models and their outcomes. If the pre-processing biases are not mitigated, they propagate to the models which leads to algorithmic bias.

3.3. Bias identification and mitigation techniques

We have employed several bias identification and mitigation techniques to identify and mitigate the biases in the pre-processing and post-processing stages of ML workflow. Some of the techniques are described in detail below:

3.3.1. Overall baseline model

We aimed to develop an overall baseline model which assumes that there is no bias in the data and models. To achieve this, we initially trained the model on the complete training dataset regardless of race, age, and gender and tested it to determine the effectiveness of the model. This provides us a baseline model to compare and help in identifying the inherent algorithmic bias.

3.3.2. Group-specific models

This approach considers training the individual models for each protected group (i.e., race, age, gender) and testing it within the group and on other groups to identify the fairness in the models.

3.3.3. Data balancing and adjustment factor

This approach is used to mitigate the inherent biases in the dataset. The data balancing approach balance the number of instances of majority and minority protected groups to mitigate the bias and improve the performance on the minority groups. Moreover, we also applied an adjustment factor for the protected groups to reduce the biases in the dataset.

In this paper, we have used FairLearn [15] framework, a Python library designed for identifying and mitigating bias in machine learning model building. This library provides us the tools to quantify potential biases and apply strategies to mitigate them. The process involved

several key stages of detection of bias, assessment of its impact, and the implementation of bias mitigation strategies. When detecting bias with FairLearn, we used the library's *Metric Frame* function in our model's prediction.

3.4. Machine learning models

The paper collectively analyze disparities in predictive modeling outcomes across gender, age, and race using machine learning algorithms and tailored experimental setups to detect biases and evaluate fairness. The primary algorithms implemented include Linear Regression for its simplicity and interpretability, Support Vector Regression (SVR) for modeling non-linear relationships, and the Gradient Boosting Regressor, which iteratively refines predictions by minimizing errors. These algorithms were chosen for their robustness in handling structured data and their ability to highlight performance variations across demographic groups, including gender, age, and race. The details of the machine learning models used in the paper are as follows:

- **Support Vector Regression (SVR)**: SVR predicts continuous values by fitting a hyperplane that minimizes error within a tolerance margin (epsilon). It balances complexity and accuracy by penalizing points outside this margin. Using kernel tricks, SVR can handle both linear and non-linear relationships. While it performs well on small or complex datasets and is robust to outliers, it can be computationally expensive and sensitive to hyperparameter tuning.
- **Linear Regression**: Linear regression models the relationship between independent variables and a continuous dependent variable by fitting a straight line. It minimizes the sum of squared errors, making it simple, interpretable, and computationally efficient. However, it assumes linearity, independence, and normality, which limits its performance with non-linear or noisy data. It is widely used for straightforward predictive tasks like forecasting or trend analysis.
- **Gradient Boosting Regression**: Gradient Boosting Regression builds an ensemble of decision trees sequentially, with each tree correcting the errors of the previous one to minimize a loss function. It handles linear and non-linear relationships, excels at complex problems, and achieves high accuracy with proper hyperparameter tuning. However, it is computationally intensive and prone to overfitting without regularization, making it ideal for tabular data and high-stakes predictions.

4. Experiments and results

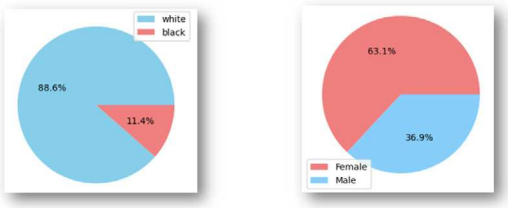
This section provides the experimental results to evaluate the effectiveness of our approach in identifying and mitigating biases in the dataset.

4.1. Experimental setup

We applied our approach to the healthcare dataset (i.e., details are in dataset section). We have used Python Jupyter notebooks for our analysis that would serve to house all of the processes that we use on the dataset. We have done experiments in Fairlearn framework [15], a Python library designed for understanding and mitigating bias in machine learning model building, to evaluate the effectiveness of our approach. We did our experiments in a system with Intel Core -i7-8550U CPU 2 GHz processor, 16 GB RAM 8 cores and 1TB of Hard disk with Windows 10 OS. We used Python 3.10, Anaconda Navigator 2.2.0., and Fairlearn version 0.10 for the experiments.

4.2. Bias identification in the pre-processing stage

This section describes the visualization approach to identify biases in the given dataset with critical analysis and descriptive statistics.



(a) Race Distribution (b) Gender distribution

Fig. 1. Race and gender distribution.

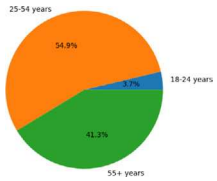


Fig. 2. Age distribution.

4.2.1. Race distribution

The distribution shown in Fig. 1 (a) reveals a notable disparity in the dataset, with a significantly larger proportion of individuals identified as white (43,202 or 88.6%) compared to those identified as black (5,582 or 11.4%). This distribution highlights a significant overrepresentation of white individuals compared to their black counterparts in the dataset. Such imbalances in sample sizes across racial categories could potentially influence the analysis and interpretation of outcomes derived from the dataset.

4.2.2. Gender distribution

The distribution shown in Fig. 1 (b) indicates a higher representation of individuals identified as Female in the dataset compared to those identified as Male. Specifically, there are 30,763 individuals identified as Female, comprising 63.1% of the dataset, whereas there are 18,021 individuals identified as Male, constituting 36.9% of the dataset. This shows a disparity in the population considered in the dataset for the healthcare risk scores and costs.

4.2.3. Age distribution

Fig. 2 shows the age distribution of individuals in the dataset. We have five different age groups in the dataset (i.e., 18–24, 25–34, 35–44, 45–54, 55–64, 65–74, and 75+). We have combined the age groups into three categories i.e., 18–24, 25–54, 55+. Fig. 2 shows the percentage of people in the age group of 18–24, 25–54, and 55+ as 3.7%, 54.9%, and 44.3% respectively. The distribution indicates that the dataset represent a population that is predominantly composed of 25–54 age group and 55+ years adults, with a smaller representation of 18–54 years age group adults.

This distribution could potentially impact any analysis or conclusions drawn from the dataset, as it may not be representative of a broader population with a more balanced age distribution.

4.2.4. Risk score and costs by race

The mean and median risk scores for individuals of the black race (5.37 and 3.02, respectively) surpass those of the white race (4.27 and 2.87, respectively), suggesting that, on average, individuals from the black community tend to achieve higher risk scores compared to their white counterparts. Moreover, the standard deviation of the risk scores among individuals of the black race (7.98) notably exceeds that among individuals of the white race (5.10), indicating greater variability in scores within the black community compared to the white community. Similarly, the mean and median costs for individuals of the black race

surpass those of the white race, implying that, on average, expenses associated with individuals from the black community are higher than those from the white community. Additionally, the standard deviation of costs among individuals of the black race (25,068.99) notably exceeds that among individuals of the white race (16,849.42), signifying greater variability in costs within the black community compared to the white community.

In summary, individuals from the black race tend to exhibit higher risk scores and incur higher costs on average compared to individuals from the white race for healthcare expenses. Furthermore, there is greater variability in both risk scores and costs among individuals from the black race compared to those from the white race, as evidenced by the higher standard deviations.

4.2.5. Risk score and cost by gender

The relatively high standard deviation for both genders suggests a widespread or variability in scores around the mean, indicating significant diversity within each gender group. Females exhibit a higher mean cost compared to males, with females averaging approximately $8,004, while males average $7,071. This disparity in mean cost may reflect variations in utilization or expenditure patterns between genders within the context under study.

Moreover, the high standard deviation for both genders in terms of cost indicates considerable variability or dispersion in the costs incurred within each gender group. Although there are differences in mean scores and costs between genders, these variances are not dramatic. However, the notable difference in mean costs may warrant further investigation into potential contributing factors. Exploring factors influencing variations in costs between genders and investigating outliers impacting the distribution of scores and costs could be potential areas for further analysis.

4.2.6. Joint distribution of "race" and "gender" analysis

Among the dataset's demographics, there are 3686 black females and 1896 black males, while 27,077 white females and 16,125 white males are accounted for. In the black racial category, females outnumber males, with 3686 females compared to 1896 males. Similarly, within the white racial category, there is a notable female majority, with 27,077 females and 16,125 males. These findings reveal a gender imbalance within both racial categories. For instance, in the black racial category, there are approximately two females for every male, and in the white racial category, there are roughly 1.7 females for every male. These gender imbalances within racial categories may point to underlying socio-cultural or demographic influences affecting healthcare utilization, program enrollment, or dataset participation. To ensure equitable representation and access to resources and services, further investigation into the factors driving these gender disparities, such as healthcare accessibility, participation rates, or societal norms, is needed.

4.2.7. Joint distribution of "race", "age", and "gender" analysis

A combined analysis of all the demographic or protected attributes (i.e., race, age, gender) indicates that the most common combination in the dataset consists of white females in the age group of 25–54 years. The percentage of individuals matching this combination is around 30%. This suggests a potential bias in the dataset with respect to the individual as well as the combined demographics. This also suggest a potential sampling bias towards white females of 25–54 years age group in the data collection process.

Overall, these pre-processing results suggest that there is a race, age, and gender bias in the dataset which indicates that we need to consider these biases at an early stage and might need methods to mitigate these biases initially. Sampling/selection bias is a common issue in most of the dataset due to under-representation of a particular demographic subgroup. These biases lead to class imbalance problem which will produce incorrect predictions. If identified early, there is a potential to improve the outcomes. To alleviate this problem resampling methods [32], and data augmentation techniques [33,34] are used. We can use data augmentation techniques such as cGAN [35] to generate synthetic data for the minority attribute, reduce bias and develop trustworthy models that are generalizable.

Table 1
Racial bias identification results.

Model type	LR	SVR	GBR
Overall Baseline Model	11.7	18.0	**9.5**
Black (tested on Black)	11.7	18.0	**9.5**
White(tested on White)	10.8	15.8	**8.9**
Black (tested on White)	21.2	41.5	**19.3**
White(tested on Black)	12.3	15.7	**9.8**
Overall (tested on Black)	19.2	38.1	**12.0**
Overall(tested on White)	10.2	14.5	**7.3**

4.3. Bias identification in the post-processing stage

Fairness assessment is conducted using Fairlearn. We have done experiments to identify racial, gender, and age group biases in the dataset. One of the key steps in detecting bias involve identifying the sensitive attribute and target variable. The sensitive attributes are age, race, and gender and the target variable is the risk score in the dataset. Preprocessing tasks include handling missing values and categorical variable encoding to optimize the set for modeling. Experimental scenarios for the bias identification include: 1) Overall baseline model; 2) Group-specific models (i.e., training and testing on specific subgroups of sensitive attributes). All of our experiments use a train-test split of 80–20 where 80% data is used for training and 20% for testing. The goal of the predictive model in the experiments is to predict the risk scores of primary care patients and compare the outcomes using various evaluation metrics. The evaluation metric used in the paper is Mean squared error (MSE) values. To ensure consistency in the results, the dataset is shuffled for 5 times, and average MSE values are reported for each experimental scenario.

4.3.1. Models results for racial bias identification

Table 1 summarizes the racial bias identification results. We built three models including linear regression (LR), support vector regression (SVR), and gradient boosting regression (GBR) for the experiments for seven scenarios as shown in the Model Type column of Table 1. To examine the racial bias of each model, we compare the model trained on a mixed racial dataset (as shown in the overall baseline model) with models trained exclusively on Black group data and tested with Black testing data, as well as models trained on White group data and tested with White testing data. Moreover, the models trained on black and white groups are also tested on their counterparts interchangeably. Finally, models trained on overall training data (unbiased) are also tested on the individual subgroups of race.

As shown in Table 1, First of all, when the model is trained on all the training data (unbiased) regardless of race, the average MSE values are 11.7, 18.0, and 9.5. These values are comparable to those obtained when the model is trained exclusively on Black group data and tested on Black group data. However, when the model is trained on the white-group data and tested on the white-group, the MSE values decreased to 10.8, 15.8, and 8.9 respectively. When the model is trained only on the black-group data and tested on white, there is a significant variation in the MSE values and among the highest for all the other scenarios, however, when the model is trained on white-group and tested on black-group, it reduces significantly. We see the similar pattern when the models are trained on overall unbiased data and tested on the Black-group data with the higher MSE values of 19.2, 38.1, and 12.0 respectively. Through these results, we confirmed that the model trained on the white-group and tested on white-group has

Table 2
Gender bias identification results.

Model type	LR	SVR	GBR
Overall Baseline Model	11.7	18.0	**9.5**
Female (tested on Female)	10.9	16.3	**9.3**
Male (tested on Male)	13.9	22.6	**11.7**
Female (tested on Male)	11.7	16.8	**10.2**
Male (tested on Female)	14.3	23.4	**12.6**
Overall (tested on Female)	10.1	14.8	**7.3**
Overall (tested on Male)	13.0	20.8	**8.3**

Table 3
Age bias identification results.

Model type	LR	SVR	GBR
Overall Baseline Model	11.7	18.0	**9.5**
Lower (tested on Lower)	8.8	12.8	**6.8**
Upper (tested on Upper)	15.5	21.1	**13.1**
Lower (tested on Upper)	9.9	15.5	**9.4**
Upper (tested on Lower)	22.6	31.0	**16.8**
Overall (tested on Lower)	8.5	12.3	**5.8**
Overall (tested on Upper)	14.8	23.6	**10.3**

Table 4
Intersectional bias identification results.

Race	Gender	Age-group	LR	SVR	GBR
Black	Female	Lower	21.6	37.3	16.2
		Upper	20.2	32.0	13.8
	Male	Lower	15.9	20.0	16.0
		Upper	27.3	59.9	24.2
White	Female	Lower	5.8	6.8	5.2
		Upper	14.4	22.0	12.6
	Male	Lower	7.8	9.5	6.7
		Upper	19.0	30.0	13.3

the best performance among all the other cases. For individual models, all models perform best for white race and worse for the black race.

Overall the analysis of these results revealed significant disparities in the error rates between racial groups, indicating that there is a potential bias in the model's prediction. Also, gradient boosting model performs best in all the scenarios. Fairlean's facilitation of these disparities offered a clear measure of model fairness and laid the groundwork for our mitigation efforts. However the results suggest that the overall unbiased model is not a good indicator to identify biases in the dataset. The results are consistent with the existing work [14].

4.3.2. Models results for gender bias identification

Table 2 summarizes the gender bias identification results. The models are similar to the racial bias identification results. To assess gender bias in each model, we compare the model trained with data from both genders (as shown in the overall baseline model) with models trained exclusively on male-group data and tested on male testing data, as well as models trained on female-group data and tested on female testing data. Moreover, the models trained on male and female groups are also tested on their counterparts interchangeably. Finally, models trained on overall training data (unbiased) are also tested on the individual subgroups of gender.

As shown in Table 2, First of all, when the model is trained on all the training data (unbiased) regardless of gender, the average MSE values are 11.7, 18.0, and 9.5 respectively. For the model trained on the female-group data and tested on the female-group, the MSE values decreased to 10.9, 16.3, and 9.3 respectively. However, when the model is trained on male-group data and tested on the male-group, the MSE values increased to 13.9, 22.6, and 11.7 respectively. When the model is trained only on the female-group data and tested on male, there is a significant decrease in the MSE values, however, when the model is trained on male-group and tested on female-group, it increases significantly. We see the similar pattern when the models are trained on overall unbiased data and tested on the female-group with the lower MSE values of 10.1, 14.8, and 7.3 respectively. Through these results, we confirmed that the model trained on the overall data and tested on the female-group has the best performance among all the other cases. For individual models, all models perform best for females and worse for the male.

Overall the analysis of these results revealed significant disparities in the error rates between the gender groups, indicating that there is a potential bias in the model's prediction. Also, gradient boosting model again outperforms in all the scenarios. However, the results suggest that the overall unbiased model is not a good indicator to identify biases in the dataset. This suggests another demographic bias in the dataset based on gender.

4.3.3. Models results for age bias identification

For the age-related bias, we have five different age groups in the dataset (i.e., 18–24, 25–34, 35–44, 45–54, 55–64, 65–74, and 75+). We combined the groups to create a new variable with two groups(i.e., lower and upper age groups). The lower comprises of patients with age groups between 18–54 including 18 and 54 years people, however, the upper group consists of all people with age greater than or equal to 55 years. Table 3 summarizes the age bias identification results. The models are similar to the racial and gender bias identification results. o examine age bias in each model, we compare the model trained with data from both age groups (as shown in the overall baseline model) with models trained exclusively on data from the lower age group and tested on lower age group data, as well as models trained on data from the upper age group and tested on upper age group data. Moreover, the models trained on lower and upper groups are also tested on their counterparts interchangeably. Finally, models trained on overall training data (unbiased) are also tested on the individual subgroups of age.

As shown in Table 3, First of all, when the model is trained on all the training data (unbiased) regardless of age, the average MSE values are 11.7, 18.0, and 9.5 respectively. For the model trained on the lower-group data and tested on the lower-group, the MSE values decreased to 8.8, 12.8, and 6.8 respectively. However, when the model is trained on upper-group data and tested on the upper-group, the MSE values are increased to 15.5, 21.1, and 13.1 respectively. When the model is trained only on the lower-group data and tested on upper group, there is a significant decrease in the MSE values, however, when the model is trained on upper-group and tested on lower-group, it increases significantly. We see the similar pattern when the models are trained on overall unbiased data and tested on the lower-group MSE values of 8.5, 12.3, and 5.8 respectively. Through these results, we confirm that the model trained on the overall data and tested on lower-group has the best performance among all the other cases. For individual models, all models perform best for lower group and worse for the upper group.

Overall the analysis of these results revealed significant disparities in the error rates between the age groups, indicating that there is a potential bias in the model's prediction. Also, gradient boosting model again outperforms in all the scenarios. However, the results suggest that the overall unbiased model is not a good indicator to identify biases in the dataset. This suggests another demographic bias in the dataset based on age. In summary, the bias identification results suggest that there are significant racial, gender, and age bias in the dataset.

4.3.4. Models results for intersectional bias identification

We have also analyzed the intersectional bias in the dataset by considering all the sensitive variables (i.e., race, gender, age) together. Table 4 summarizes the results of the intersectional bias identification. On comparing the black female with the White female at same age groups, the error rate of black female of lower age group (18–54 years) is much higher than their white counterpart. Black female of lower age

Table 5
Oversampling bias mitigation results.

Model type	Avg. MSE values for GBR		
	Updated MSE	Initial MSE	% diff
Black (tested on White)	18.5	19.3	4.1
White (tested on Black)	9.9	9.8	−1.0
Male (tested on Female)	11.2	12.6	11.1
Female (tested on Male)	9.7	10.2	4.9
Lower (tested on Upper)	9.1	9.4	2.9
Upper (tested on Lower)	15.6	16.8	7.1

Table 6
Undersampling bias mitigation results.

Model type	Avg. MSE values for GBR		
	Updated MSE	Initial MSE	% diff
Black (tested on White)	20.3	19.3	−5.1
White (tested on Black)	9.9	9.8	−1.0
Male (tested on Female)	12.0	12.6	4.8
Female (tested on Male)	9.8	10.2	3.9
Lower (tested on Upper)	8.4	9.4	10.6
Upper (tested on Lower)	16.0	16.8	4.8

group are incorrectly predicted for their risk scores at the rate of almost 3.72 times that of white female of same age group. Similarly, black women of upper age group (55+ years) also have higher error rate as compared to black women. However, on comparing the black males of both lower and upper age groups with their white counterparts, the error rates are still higher. This signifies that black females are unfairly treated by this model. On comparing the black females with white males, a similar pattern of high error rate is observed. However, there is a significantly higher error rates in the risk score prediction for black males as compared to white females and males. This signifies that black males are also unfairly treated by the model. For the individual models results, all the models have similar patterns of higher error rates for black males and females of lower and higher age groups as compared to their white counterparts. Overall the results of intersectional bias suggests that Black males and females are an intersectional group that is being unfairly harmed.

4.4. Bias mitigation in pre-processing stage

We have used oversampling and undersampling techniques for bias mitigation in pre-processing stage. The oversampling technique oversamples the minority group data to balance it with the majority group however the undersampling approach reduces the majority group to balance it with the minority group. Table V summarizes the oversampling bias mitigation results. We run the experiments with the same setting as in the bias identification however we report the results only for the GBR model as it is the best performing model. As shown in Table 5, the MSE values after racial bias mitigation are 4.1% lower compared to the values before mitigation when the model is trained exclusively on Black group data and tested on White group data. However, when the model is trained on White group data and tested on the Black group, the error rates increases. Similarly, when the model is trained exclusively on male group data and tested on female group data, there is a notable change in the error values, with a decrease of up to 11.1%. Finally, there is a slight improvement in the MSE values when the model is trained on lower-group data and tested on upper-group data. However it reduces significantly when the model is trained on upper-group data and tested on lower-group data.

Table 6 summarizes the undersampling bias mitigation results. As shown in Table VI, the error rates increase after the racial bias mitigation. Thus, undersampling does not work well in the racial bias mitigation. However, it works for the gender and age bias mitigation. When the model is trained exclusively on male group data and tested on female group data, there is a slight reduction in the error values with 4.8%. However the reduction is 3.9% when the model is trained on female group data and tested on male group. Finally, there is a notable reduction of 10.6% in the MSE values when the model is trained on lower-group data and tested on upper-group data. However there is a slight reduction when the model is trained on upper-group data and tested on lower-group data.

Through these results, we confirm that the mitigation strategies applied at the pre-processing stage help in building a fair model.

4.5. Bias mitigation in post-processing stage

Based on the above results, there is a significant racial, gender, age and intersection bias in the dataset. We have used adjustment factor technique for bias mitigation in the post-processing stage. Conceptualizing a post-processing mitigation strategy tailored towards regression tasks and the detected bias, adjustments are made specifically for the 'black' group, in an effort to align its performance more closely with that of the 'white' group. These adjustments aimed to mitigate the observed bias and enhance the model's fairness of prediction, regardless of racial group.

We have developed an adjustment factor dictionary that includes factors assigned to both racial groups. These factors are determined to align the MSE values of the Black and White groups more closely with the MSE value of the overall baseline model, thereby reducing the performance disparity between the biased and unbiased models. An adjustment factor of 1.5 is applied to the White group to increase the MSE values, as they are lower than the overall MSE. This results in MSE values of 12.74, 13.3, and 10.5 for the linear regression, SVR, and gradient boosting models, respectively. An adjustment factor of −0.33 is applied to the Black group to decrease the MSE value, as it is higher than the overall MSE. This results in MSE values of 20.79, 21.5, and 19.5 for the linear regression, SVR, and gradient boosting models, respectively. These results indicate that while we attempt to address the racial bias, this mitigation strategy does not fully eliminate it. Further analysis is needed to definitively identify the most effective mitigation strategy for these biases, which will be a focus of our future work in this paper.

5. Discussion and conclusions

This paper aimed to perform a detailed analysis of different types of data and algorithmic biases and their impact on the outcomes using a case study of a widely used real-world healthcare dataset. We have identified pre-processing and post-processing bias in the dataset which leads to data and algorithmic bias. Through the experimental results and evaluation, we can conclude that the dataset has racial, gender, age, and intersectional bias. These biases not only lead to unfair outcomes from the model but also introduce harm to specific demographic groups (including historically marginalized groups (e.g., based on gender, race, age). Based on the experimental results, we can infer that checking for bias is necessary at every stage of the Machine learning workflow starting from the data collection to the model building and evaluation phases. Finally, we can deduce that racial biases lead to health disparity that affect certain group of people in receiving access to healthcare facilities and treatment for various serious health conditions. The gender and age bias affect certain groups disproportionately. Based on intersectional bias results, black male and females in the age group of 18–54 and 55+ years are unfairly harmed and have higher error rates in the risk score prediction as compared to their white counterparts. This signifies that black groups are categorized in the incorrect risk category due to which they get deprived of special government healthcare programs.

Our results demonstrate clear evidence of racial, gender-based, and age-related biases in the healthcare dataset used to guide resource allocation for patients. These biases have the potential to lead to

higher rates of misclassification, disproportionately affecting marginalized communities. Such classification errors risk perpetuating existing healthcare inequities by allocating fewer resources to underrepresented groups while reinforcing systemic disparities. To address the adverse consequences of these biases, we underscore the need for continued research into sophisticated mitigation techniques. While existing toolkits such as Fairlearn and IBM's AI Fairness 360 provide valuable technical solutions, these tools alone are insufficient. A critical, often overlooked component is the integration of interdisciplinary expertise in clinical practice, data science, and ethics to oversee the development and implementation of these tools. This integration ensures that mitigation strategies are not only effective but also aligned with ethical principles and real-world clinical contexts. Furthermore, there is a pressing need for robust policies and regulatory frameworks to govern the use of AI in healthcare. These policies should establish clear standards for fairness, accountability, and transparency in AI systems. Such measures are crucial to prevent the deployment of biased algorithms and ensure that AI-powered decisions benefit all patients equitably, regardless of their demographic background.

A potential limitation of this study is that the study is conducted on one healthcare-related dataset. This limitation is due to the unavailability of healthcare datasets with characteristics to demonstrate various types of biases and privacy issues. With more access, we could conduct this experiment across a wide range of datasets and see how bias arises in multiple instances. One way to address this limitation is by enabling private sector entities to test their datasets using these or similar fairness metrics and evaluate the results they obtain. We will explore additional metrics to detect biases in the dataset and experiment with other mitigation techniques in the future extension of this work. Further validation on different types of dataset will also be a part of our future work. There is also potential to create more comprehensive mitigation strategies by developing and testing mitigation techniques that are more sophisticated to address the detected biases. This includes exploring more algorithmic adjustments or fairness-aware model training approaches. Interdisciplinary collaboration can also be integrated into the bias mitigation process to provide deeper insight into the ethical implications of various bias mitigation methods. This would ensure that adverse effects are directed towards the intended consumers of these models during testing. Addressing these areas of work can advance the understanding of bias mitigation in AI models used in healthcare, contributing to more equitable healthcare practices.

In conclusion, it is imperative to develop AI systems that are safe, reliable, and trustworthy, while also implementing mechanisms for continuous monitoring and evaluation to detect and address unintended biases. We advocate for collaboration across sectors, bringing together researchers, policymakers, clinicians, and ethicists to design and validate mitigation strategies. This collaborative effort will create a robust system of checks and balances, ensuring that AI technologies do not exacerbate existing disparities but rather contribute to a more equitable healthcare landscape.

CRediT authorship contribution statement

Vibhuti Gupta: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Julian Broughton:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Ange Rukundo:** Writing – original draft, Visualization, Validation, Methodology, Formal analysis. **Lubna J. Pinky:** Writing – review & editing, Validation, Methodology, Investigation, Conceptualization.

Ethical statement

This study adheres to the highest ethical standards. An open-access dataset was utilized, and validation was performed through computer simulations. As no human subjects were involved at any stage of the research, the study poses no risk to patient safety or privacy. This approach supports the advancement of medical research while upholding ethical integrity and ensuring the protection of human subjects.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used GPT-4o in order to edit the sentences. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Funding

This work was supported in part by the: National Science Foundation, United States [grant number 2334391], RCMI Program in Health Disparities Research [grant number 3U54MD007586-37S3], and by the, AIM-AHEAD Coordinating Center, award number OTA-21-017, and was, in part, funded by the National Institutes of Health, United States Agreement No. 1OT2OD032581.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Vibhuti Gupta reports financial support was provided by Meharry Medical College. Vibhuti Gupta reports a relationship with Meharry Medical College that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to acknowledge mainly the School of Applied Computational Sciences at Meharry Medical College including their department of computer science and data science for providing time and support for accomplishing this work.

Data availability

The paper uses the publicly available data provided in the paper [14]. The link to access the data is at https://gitlab.com/labsysmed/dissecting-bias.

References

[1] Nasir S, Khan RA, Bai S. Ethical framework for harnessing the power of AI in healthcare and beyond. IEEE Access 2024;12:31014–35. http://dx.doi.org/10.1109/ACCESS.2024.3369912.

[2] Ueda D, Kakinuma T, Fujita S, Kamagata K, Fushimi Y, Ito R, et al. Fairness of artificial intelligence in healthcare: review and recommendations. Jpn J Radiol 2024;42(1):3–15. http://dx.doi.org/10.1007/s11604-023-01474-3.

[3] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv (CSUR) 2021;54(6):1–35. http://dx.doi.org/10.1145/3457607.

[4] Li F, Ruijs N, Lu Y. Ethics & AI: A systematic review on ethical concerns and related strategies for designing with AI in healthcare. Ai 2022;4(1):28–53. http://dx.doi.org/10.3390/ai4010003.

[5] Zhang J, Zhang Z-m. Ethics and governance of trustworthy medical artificial intelligence. BMC Med Inform Decis Mak 2023;23(1):7. http://dx.doi.org/10.1186/s12911-023-02103-9.

[6] Akter S, McCarthy G, Sajib S, Michael K, Dwivedi YK, D'Ambra J, et al. Algorithmic bias in data-driven innovation in the age of AI. IJIM 2021;60:102387. http://dx.doi.org/10.1016/j.ijinfomgt.2021.102387.

[7] Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: the Framingham study. AJC 1976;38(1):46–51. http://dx.doi.org/10.1016/0002-9149(76)90061-8.

[8] Gijsberts CM, Groenewegen KA, Hoefer IE, Eijkemans MJ, Asselbergs FW, Anderson TJ, et al. Race/ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events. PLoS One 2015;10(7):e0132321. http://dx.doi.org/10.1371/journal.pone.0132321.

[9] Adamson AS, Smith A. Machine learning and health care disparities in dermatology. JAMA Dermatol 2018;154(11):1247–8. http://dx.doi.org/10.1001/jamadermatol.2018.2348.

[10] Ferretti MT, Iulita MF, Cavedo E, Chiesa PA, Schumacher Dimech A, Santuccione Chadha A, et al. Sex differences in Alzheimer disease—the gateway to precision medicine. Nat Rev Neurol 2018;14(8):457–69. http://dx.doi.org/10.1038/s41582-018-0032-9.

[11] Kim H-I, Lim H, Moon A. Sex differences in cancer: epidemiology, genetics and therapy. Biomol Ther 2018;26(4):335. http://dx.doi.org/10.4062/biomolther.2018.103.

[12] Kuehner C. Why is depression more common among women than among men? Lancet Psychiatry 2017;4(2):146–58. http://dx.doi.org/10.1016/s2215-0366(16)30263-2.

[13] Natri H, Garcia AR, Buetow KH, Trumble BC, Wilson MA. The pregnancy pickle: evolved immune compensation due to pregnancy underlies sex differences in human diseases. TiG 2019;35(7):478–88. http://dx.doi.org/10.1016/j.tig.2019.04.008.

[14] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Sci 2019;366(6464):447–53. http://dx.doi.org/10.1126/science.aax2342.

[15] Weerts H, Dudík M, Edgar R, Jalali A, Lutz R, Madaio M. Fairlearn: Assessing and improving fairness of AI systems. J Mach Learn Res 2023;24:1–8.

[16] Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al. AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Dev 2019;63(4/5):1–4.

[17] Saleiro P, Kuester B, Hinkson L, London J, Stevens A, Anisfeld A, Rodolfa KT, Ghani R. Aequitas: A bias and fairness audit toolkit. 2018, http://dx.doi.org/10.48550/arXiv.1811.05577, arXiv preprint arXiv:1811.05577.

[18] Wexler J, Pushkarna M, Bolukbasi T, Wattenberg M, Viégas F, Wilson J. The what-if tool: Interactive probing of machine learning models. IEEE Trans Vis Comput Graphics 2019;26(1):56–65. http://dx.doi.org/10.1109/TVCG.2019.2934619.

[19] Hemann BA, Bimson WF, Taylor AJ. The Framingham risk score: an appraisal of its benefits and limitations. Am Heart J 2007;5(2):91–6. http://dx.doi.org/10.1111/j.1541-9215.2007.06350.x.

[20] Brindle PM, McConnachie A, Upton MN, Hart CL, Smith GD, Watt GC. The accuracy of the Framingham risk-score in different socioeconomic groups: a prospective study. Br J Gen Pract 2005;55(520):838–45.

[21] Allen A, Mataraso S, Siefkas A, Burdick H, Braden G, Dellinger RP, et al. A racially unbiased, machine learning approach to prediction of mortality: algorithm development study. JPHS 2020;6(4):e22400. http://dx.doi.org/10.2196/22400.

[22] Karlsson I, Boström H. Handling sparsity with random forests when predicting adverse drug events from electronic health records. In: 2014 ieee international conference on healthcare informatics. IEEE; 2014, p. 17–22. http://dx.doi.org/10.1109/ICHI.2014.10.

[23] Li F, Wu P, Ong HH, Peterson JF, Wei W-Q, Zhao J. Evaluating and mitigating bias in machine learning models for cardiovascular disease prediction. J Biomed Inf 2023;138:104294. http://dx.doi.org/10.1016/j.jbi.2023.104294.

[24] Wolk DM, Lanyado A, Tice AM, Shermohammed M, Kinar Y, Goren A, et al. Prediction of influenza complications: development and validation of a machine learning prediction model to improve and expand the identification of vaccine-hesitant patients at risk of severe influenza complications. J Clin Med 2022;11(15):4342. http://dx.doi.org/10.3390/jcm11154342.

[25] Li C, Jiang X, Zhang K. A transformer-based deep learning approach for fairly predicting post-liver transplant risk factors. J Biomed Inf 2024;149:104545. http://dx.doi.org/10.1016/j.jbi.2023.104545.

[26] Cui S, Pan W, Zhang C, Wang F. Bipartite ranking fairness through a model agnostic ordering adjustment. IEEE PAMI 2023. http://dx.doi.org/10.1109/TPAMI.2023.3290949.

[27] Kamulegeya L, Bwanika J, Okello M, Rusoke D, Nassiwa F, Lubega W, et al. Using artificial intelligence on dermatology conditions in Uganda: A case for diversity in training data sets for machine learning. Afri Heal Sci 2023;23(2):753–63. http://dx.doi.org/10.4314/ahs.v23i2.86.

[28] Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JA, et al. Risk of bias of prognostic models developed using machine learning: a systematic review in oncology. Diagn Progn Res 2022;6(1):13. http://dx.doi.org/10.1186/s41512-022-00126-w.

[29] Du M, Yang F, Zou N, Hu X. Fairness in deep learning: A computational perspective. IEEE Intell Syst 2020;36(4):25–34. http://dx.doi.org/10.1109/MIS.2020.3000681.

[30] Yang J, Soltan AA, Eyre DW, Clifton DA. Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. Nat Mach Intell 2023;5(8):884–94. http://dx.doi.org/10.1038/s42256-023-00697-3.

[31] Puyol-Antón E, Ruijsink B, Mariscal Harana J, Piechnik SK, Neubauer S, Petersen SE, et al. Fairness in cardiac magnetic resonance imaging: assessing sex and racial bias in deep learning-based segmentation. Front Cardiovasc Med 2022;9:859310. http://dx.doi.org/10.3389/fcvm.2022.859310.

[32] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. JAIR 2002;16:321–57. http://dx.doi.org/10.1613/jair.953.

[33] Um TT, Pfister FM, Pichler D, Endo S, Lang M, Hirche S, et al. Data augmentation of wearable sensor data for Parkinson's disease monitoring using convolutional neural networks. In: Proceedings of the 19th ACM international conference on multimodal interaction. 2017, p. 216–20. http://dx.doi.org/10.1145/3136755.3136817.

[34] Iwana BK, Uchida S. Time series data augmentation for neural networks by time warping with a discriminative teacher. In: 2020 25th international conference on pattern recognition. IEEE; 2021, p. 3558–65. http://dx.doi.org/10.1109/ICPR48806.2021.9412812.

[35] Chen W-H, Cho P-C. A GAN-based data augmentation approach for sensor-based human activity recognition. IJCCE 2021;10(4):75–84. http://dx.doi.org/10.17706/ijcce.2021.10.4.75-84.