scientific reports

Check for updates

OPEN

Fairness and bias correction in machine learning for depression prediction across four study populations

Vien Ngoc Dang[1]✉, Anna Cascarano[1], Rosa H. Mulder[2,3], Charlotte Cecil[2,4,5], Maria A. Zuluaga[6], Jerónimo Hernández-González[7] & Karim Lekadir[1,8]

A significant level of stigma and inequality exists in mental healthcare, especially in under-served populations. Inequalities are reflected in the data collected for scientific purposes. When not properly accounted for, machine learning (ML) models learned from data can reinforce these structural inequalities or biases. Here, we present a systematic study of bias in ML models designed to predict depression in four different case studies covering different countries and populations. We find that standard ML approaches regularly present biased behaviors. We also show that mitigation techniques, both standard and our own post-hoc method, can be effective in reducing the level of unfair bias. There is no one best ML model for depression prediction that provides equality of outcomes. This emphasizes the importance of analyzing fairness during model selection and transparent reporting about the impact of debiasing interventions. Finally, we also identify positive habits and open challenges that practitioners could follow to enhance fairness in their models.

Keywords Machine learning for depression prediction, Algorithmic fairness, Bias mitigation, Novel post-hoc method, Psychiatric healthcare equity

Depression is a leading cause of disability worldwide, a major risk factor for the global burden of disease, and can even lead to suicide[1,2]. Taking into account that the global prevalence of depression increased by 25% during the COVID-19 outbreak[3], being able to identify those individuals at risk would be of great value in order to enable the application of personalized preventive measures. To this end, it is necessary to characterize the factors leading up to the development of depression. Research to date points to the importance of both genetic and environmental factors (along with their interactions) in the etiology of depression[4,5]. Furthermore, environmental factors have been shown to co-occur, exerting cumulative effects on depression risk. The totality of these environmental influences is often referred to as the exposome and includes environmental and lifestyle factors, as well as traumatic life events[6]. Exposome data does not only provide an alternative picture, it is also relatively inexpensive and easy to acquire, typically through questionnaires[7]. Motivated by the successful application of machine learning (ML) in different contexts in the domain of medicine, there has been a spike in the use of ML for the detection, diagnosis, and treatment of depression[8–10]. Specifically, supervised ML methods are commonly used to learn predictive models from historical data, which are then applied to predict the possible development of the illness in new cases and patients.

There have been historical concerns about the fairness of automatic decision making systems[11] and, with the growing adoption of machine learning in health care applications, these concerns have also extended to ML models' potential unfair bias[8]. It has been shown[12] that ML models can amplify unfair behaviors masked

[1]Departament de Matemàtiques i Informàtica, Facultat de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain. [2]Department of Child and Adolescent Psychiatry/Psychology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [3]The Generation R Study Group, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [4]Department of Epidemiology, Erasmus MC, Rotterdam,University Medical Center Rotterdam, Rotterdam, The Netherlands. [5]Molecular Epidemiology, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands. [6]Data Science Department, EURECOM, Biot, France. [7]Departament d'Informàtica, Matemàtica Aplicada i Estadística, Universitat de Girona, Girona, Spain. [8]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. ✉email: dangn@ub.edu

in past practice, that is, in the data used for model learning. The term "algorithmic bias" refers to differences in the predictive power of models when applied to different subgroups of the population. Such differences are particularly worrying when the subgroups are determined according to some protected attribute, such as ethnicity, sex, or age. The subgroups that are unfairly impacted or benefited by the bias of the ML model are known as the unprivileged and the privileged groups, respectively. Although this undesirable behavior of ML models is nowadays well known, assessing the bias of ML models (and trying to mitigate it) is not a common practice in healthcare applications. Chen et al.[8] examine an ML algorithm on psychiatric notes to predict 30-day psychiatric readmission regarding sex, ethnicity, and insurance type without addressing algorithmic bias. Park et al.[9] reduce bias for clinical prediction models of postpartum depression only associated with one protected attribute - binarized ethnicity (Black/White individuals). In mental health, several unintentional discriminative behaviors have been detected, which could potentially be reproduced by ML models if they are reflected in the training data. Specifically, a lack of representation of the patient subgroups has been reported; for example, some ethnic groups do not use mental health services as much as others due to cultural stigma surrounding mental illness[13,14]. Prior research shows that the prevalence or incidence of depression differs across sex subgroups; women are about twice as likely as men to develop depression during their lifetime[15]. In addition, manifest discrepancies in relevant factors such as lifestyle or dietary habits between subgroups have also been reported[16]. These, and possibly other factors, can be rooted in the data which is used to learn ML models for depression prediction.

In this paper, we present a systematic analysis of algorithmic bias in ML models designed to predict the presence or absence of depression from environmental and lifestyle data, using four public datasets: LONGSCAN[17], FUUS[18], NHANES[19], and the UK Biobank (UKB)[20]. We study unfair bias on protected attributes including demographic factors (sex, ethnicity, nationality), socioeconomic status (age, income, academic qualifications), and co-morbidities (cardiovascular disease (CVD), diabetes), and evaluate the interplay of model accuracy and fairness. For this study, protected attributes were agreed based on standard choices in the related literature[21]. We analyze the ability of different bias mitigation techniques to reduce the discrimination level of the models learned for our four case studies. The mitigation effect is measured as the performance difference, in terms of fairness metrics, before and after performing bias mitigation. We have found unfair biases in the behavior of the models learned with standard ML techniques regarding several protected attributes in all the case studies. We also found, however, that mitigation techniques are effective in reducing discrimination levels. Our results suggest that bias monitoring is pertinent in the evaluation of ML-based predictive models in mental health and current mitigation techniques provide a powerful toolset to mitigate unfair algorithmic bias.

Methods

In this section, we introduce our case studies and perform an initial descriptive analysis of the available data ("Datasets"). We describe the type of predictive models tested and the five strategies we considered to mitigate bias ("Bias mitigation approaches"). We also explain both standard ML and fairness evaluation metrics used in our experiments ("Results").

Datasets

This study uses four public datasets: LONGSCAN, FUUS, NHANES, and UK Biobank. We select these datasets to cover a spectrum of sizes, from small to large, and to showcase diverse methodologies for diagnosing depression. We aim to assess potential biases across different scales and diagnostic methods. These differences allow us to investigate how the size of the dataset (both in terms of the number of samples and input variables) influences bias in both plain ML models and those combined with debiasing techniques. All the participants and/or their carers provided written informed consent in LONGSCAN, NHANES, and UKB studies. This was not required in the case of FUUS according to the laws that regulate "non-interventional clinical research" in France[18]. While LONGSCAN and FUUS are datasets of late adolescents, the subjects in the NHANES and UKB datasets are primarily between the ages of 40 and 80. Supplementary Table S1 describes the protected attributes considered for each dataset, as they differ between datasets. The LONGSCAN, FUUS, NHANES, and UKB datasets have relatively equal proportions of male and female subjects. A higher prevalence of depression in women versus men is evident in the LONGSCAN, NHANES, and UKB datasets. No sex effect is found among college freshmen in the FUUS dataset, which is consistent with previous studies[22–24]. The distribution of other protected attributes is highly skewed.

Participants and features

Participants in LONGSCAN were from five regions in the United States (the South, East, Midwest, Northwest, and Southwest), with different selection criteria, representing varying levels of risk or exposure to maltreatment during the period spanning from 1991 to 2012. The LONGSCAN interview and questionnaire data were collected when target children were 4, 6, 8, 12, 14, 16, and 18 years of age. Out of 1,354 total participants, we kept the 67.3% of children who completed an interview at 18 years-old which included depression outcomes. This left us with 911 samples for our study. Among these 911 individuals, there were 363 cases with depression at the age of 18, and 548 controls. The study design is depicted in Supplementary Fig. S1: data from three different stages (early childhood, late childhood, teen) were collected to predict depression at the age of 18. Up to 23 descriptive variables were considered, grouped as demographic variables, lifestyles variables, and adverse exposures variables, both time-invariant and repeatedly measured along these stages (see Supplementary Table S2). Data is available under request; its use for this study was approved by the National Data Archive on Child Abuse and Neglect (NDACAN). Participants in FUUS were undergraduate students who underwent a compulsory medical visit at the university medical service in Nice (France) between September 2012 and June 2013. Among the 4184 total participants, there are 528 cases with depression and 3656 controls. A total of 62 biomedical

and demographic features were used, including binary, ordinal and continuous variables (see Supplementary Table S3). Participants in NHANES provided data between 2005 and 2018 and were selected by random sampling of the American population. Among the 36,259 total participants, there were 3168 cases with depression and 33,091 controls. A total of 86 features were used in our study, including demographic data, socioeconomic status, medical history, lifestyle characteristics, and prescription medications. However, to prevent label leakage, we specifically excluded participants' feelings and expressions, specific lifestyle characteristics such as physical activity, diet, and sleep habits, as well as depression-specific medications from the set of descriptive features (see Supplementary Table S4). FUUS and NHANES datasets are publicly available. Participants in UKB were enrolled from 22 United Kingdom's centers from 2006 to 2010. Among the 461,033 participants who did not initially have depression, 18,112 cases (3.93%) developed depression. Up to 143 descriptive variables were considered, including demographic data, socioeconomic status, medical history, lifestyle characteristics, early life factors, and traumatic events (see Supplementary Table S5). Data is available under request; its use for this study was approved by UKB, under the project title "Association between Early-Life-Stress and Psycho-Cardio-Metabolic Multi-Morbidity: The EarlyCause H2020 Project" (application number 65769).

Building the ground truth: depression outcome

We acquired ground truth label information –whether a participant has depression or not– using dataset-specific information. In LONGSCAN, depression was assessed using a self-reported questionnaire at age 18, which includes a specific question regarding having depression or not. In NHANES, the Patient Health Questionnaire-9 (PHQ-9) was used. This screening questionnaire consists of 9 items (scored 0–3) and has a specificity and sensitivity of 88% for major depressive disorder (MDD) at a threshold score of 10 or more[25]. Therefore, we chose the threshold at PHQ-9 score 10. In FUUS, the depression outcome was evaluated in a two-stage process. If the result of an initial four-item screening questionnaire indicated possible presence of MDD (at least two of the four symptoms present), the participants were assessed by a medical provider for the full Diagnostic and Statistical Manual of Mental Disorders Fourth Edition (DSM IV) criteria[10]. In UKB, the depression outcome was defined as an occurrence of a depressive episode (ICD10 code F32 and F33) after the date of assessment, which was drawn from hospital inpatient diagnoses or self-reported conditions. Note that this label information might be noisy due to the use of questionnaires. Addressing noisy labels, which could have a relative impact on the results of this study, is a complex challenge in ML problems that falls beyond the scope of our work. In the rest of the paper, we use the presence and absence of depression as the positive and negative class, respectively.

Model evaluation

Two popular types of ML models were learned from the data presented above: (i) logistic regression (LR)[26], a linear classifier, and (ii) extreme gradient boosting (XGB)[27], a boosting ensemble method. In this paper, we do not focus on the algorithmic aspects of the ML methods considered, but rather on their clinical application and the fairness assessment of their predictions. We use k-fold cross validation ($k = 5$ for UKB, $k = 10$ for the rest) for performance evaluation and nested cross-validation for hyper-parameters tuning (see Supplementary Table S6).

Performance metrics

To assess predictive performance, we considered two standard ML performance metrics, namely, the area under the receiver operating characteristic curve (AUC-ROC), which measures the area under the curve formed by points of true positive rates versus false positive rates across all the possible thresholds, and the balanced accuracy (BAcc), which is the arithmetic mean of sensitivity and specificity. BAcc is particularly appropriate for class-unbalanced datasets, which is the case of our four study populations. It ensures equitable representation in performance assessment, that is, it ranks algorithms according to their ability to accurately detect positive and negative examples. In this way, it aligns with the objective of the task, which is the prediction of presence/absence of a mental health issue. When discussing the fairness-accuracy trade-off, we report on empirical accuracy measured by BAcc because, in practice, classification is performed at a fixed threshold[9], making it a more pertinent measure in such scenarios. Complete experimental results in terms of AUC-ROC are provided in the Supplementary material.

Fairness metrics

There are different concepts of fairness: group fairness, individual fairness, or a combination of both[29]. In this study, we focus on group fairness, meaning that the model should perform similarly for all subgroups according to a certain statistical metric. Firstly, we consider the *equal opportunity* criterion, which states that a binary classifier is fair if its true-positive rates (TPR) are equal across groups. The Equal Opportunity Difference (EOD) metric, defined as the maximum difference in the TPR between any two subgroups defined by the protected attribute (i.e., a value of 0 indicates complete fairness), meets the aforementioned criterion. EOD is an attainable and practical fairness metric which mandates equal TPRs across the demographic subgroups. Let $D = (X, Y, C)$ be the dataset, with the protected variable X, regular descriptive variables Y, and the binary class variable C. Predictions provided by an ML model are denoted $\hat{C}$. Let us define Ω_C as the set of class labels. In this study, we only consider binary classification task, i.e., $|\Omega_C| = 2$, namely $C = \{0, 1\}$. Let us also define Ω_X as the set of possible values of variable X. For example, for the protected attribute "sex", $\Omega_X = \{\text{male, female}\}$. A subgroup of the population is formally defined as all the samples in dataset D with the same value $x \in \Omega_X$ assigned to the protected attribute X. We define the TPR of a specific subgroup $x \in \Omega_X$ as

$$TPR_x(\hat{C}) := \mathbb{E}_Y[\hat{C}|C = 1, X = x]$$

and then, EOD can be described as:

$$\text{EOD} = \min_{x \in \Omega_X} \text{TPR}_x - \max_{x \in \Omega_X} \text{TPR}_x$$

This study is contextualized in a project focused on developing efficient screening and risk prediction tools for depression in primary care settings. At this point, underdiagnosis has more severe implications than misdiagnosis. When a misdiagnosis occurs, patients still receive clinical care, and clinicians can draw upon additional symptoms and data sources to correct the error. On the other hand, underdiagnosis may lead to individuals not receiving the necessary treatment and support, exacerbating their mental health condition. Our clinicians support the idea that the focus on fairness should be on ensuring that individuals at risk of depression are equally identified across groups, so they are fairly provided with the necessary care and support. This objective aligns with the use of equal opportunity criterion, which is also considered by previous studies in ML for mental health[9,30]. In the absence of an expert-informed opinion, the *equalized odds* criterion[31] serves as an alternative fairness objective which offers a stricter standard than the *equal opportunity* criterion. It asserts that a classifier achieves fairness when both its TPR and FPR are consistent across groups. Exact equality, $TPR_{x_0}(\hat{C}) = TPR_{x_1}(\hat{C})$ and $FPR_{x_0}(\hat{C}) = FPR_{x_1}(\hat{C})$, for all $x_0, x_1 \in \Omega_X$, is often hard to force in practice. Alternatively, in this paper, we measure *equalized odds* criterion by means of the Average Odds Difference (AOD) metric, which assesses the average discrepancy between the TPR and FPR across subgroups. Formally, if the FPR of a specific subgroup $x \in \Omega_X$ is defined as:

$$FPR_x(\hat{C}) := \mathbb{E}_Y[\hat{C}|C = 0, X = x]$$

then, AOD can be described as:

$$\text{AOD} = \frac{1}{2}\left[\min_{x \in \Omega_X} (\text{FPR}_x + \text{TPR}_x) - \max_{x \in \Omega_X} (\text{FPR}_x + \text{TPR}_x)\right]$$

Complete experimental results based on group-specific FPRs and AOD are provided in the Supplementary material.

Bias mitigation approaches

Many techniques have been proposed over the last few years to address algorithmic fairness[32]. However, there is a significant shortfall in addressing fairness and bias concerns when ML is applied to the field of psychiatry. Only a handful of studies have adopted methods to counteract bias. For instance, reweighing (RW) bias-mitigation technique[33] was used to minimize bias when forecasting future benzodiazepine administrations[30]. Likewise, others applied Suppression (SUP)[34] and RW approaches to reduce bias in the prediction of postpartum depression[9]. In healthcare and other fields, the Disparate Impact Remover (DIR) method[35] has shown its ability to effectively mitigate bias while maintaining a satisfactory level of predictive performance. Furthermore, the Calibrated Equalized Odds Post-processing (CPP) method[36] was proposed in the healthcare context to predict whether an individual will have a heart condition. Among all the available bias-mitigation techniques, we consider four standard methods in this study: SUP, RW, DIR, and CPP. Moreover, we propose a novel post-hoc disparity mitigation named Population Sensitivity-Guided Threshold Adjustment (PSTA).

Suppression (SUP)

Protected attributes are directly removed from the training dataset under the assumption that access to this information is the main cause of bias. First, the protected attribute is removed from the dataset. Then, a new ML model is learned from this new version of the dataset.

Reweighing (RW)

This method weighs the samples in each (group, label) combination differently to make the protected attribute and outcome statistically independent of each other before model learning. The weight of a sample is directly proportional to the frequency of its label in the whole population and inversely proportional to the frequency of its label in its subgroup. The model is learned from this new training dataset with weighted samples.

Disparate impact remover (DIR)

Given a dataset $D = (X, Y, C)$, with protected attribute X, remaining attributes Y, and class variable or outcome C, this repair process attempts to remove bias in the remaining features Y. New values are assigned to all the cases and variables in Y. The new values ensure that all the groups follow the same distribution over every variable, making adjustments based on percentiles and quantile functions. The predictive model is learned from the new training dataset. In the three previous bias-mitigation techniques, we act on the training data. Once a newly prepared version of the training dataset is obtained, an ML model is learned from it, with the effectiveness of bias mitigation evaluated based on this model's outputs. The following two techniques perform differently. The model is learned from the original data. Then, the outputs of the model are modified under certain criteria to reduce disparities. The success of the mitigation approach is then evaluated using these modified outputs.

Calibrated equalized odds post-processing (CPP)

The method calibrates the predicted probability so that the false-positive rate or the false-negative rate of privileged and unprivileged groups are, on average, equal. It modifies the score outputs of the model for the different subgroups so that the output labels meet an equalized odds objective. In our clinical application, we focus on

identifying individuals at risk of depression, and the goal is equitable outcomes; therefore, we consider recall more important than precision, which leads us to set a cost-constraint objective: the equal false-negative rates between the subgroups.

Population sensitivity-guided threshold adjustment (PSTA)

This approach is our own bias mitigation technique, and it is inspired by the observation that when there are prevalence rate discrepancies between groups in the training dataset ML models usually primarily associate the positive class with the characteristics of the subgroup of highest prevalence. Supplementary Fig. S2a,b illustrate the predicted probability distributions generated by an ML model with data from UKB with different prevalences for females and males suffering from depression. For both positive and negative samples, the Mann-Whitney U test[37] indicates significant differences between male and female predicted probabilities (positive: $U = 21316361.0$, $p < 0.001$; negative: $U = 1297359183.0$, $p < 0.001$). The difference arises from the prevalence rate discrepancies in the training dataset, leading the model to inherently assign lower probabilities of suffering from depression to individuals from the male subgroup, which has a lower prevalence rate. Applying the default single decision threshold at 0.5 for both groups results in a lower TPR and a higher FPR for the male group, exacerbating outcome disparities. This example model would benefit from a tailored threshold adaptation that considers the disparities in the distribution of probabilities between different subgroups, ultimately improving both fairness and accuracy. The use of threshold adaptation to address bias and unfairness in ML models is not entirely new, as evidenced in previous works, such as[38,39]. However, the procedure for obtaining the optimal threshold differentiates our proposal from existing approaches. Our method is guided by the fairness criterion, *Equal Opportunity*, which is applicable to categorical protected features within a binary classification framework, i.e., it is not limited to binary protected features only. The main objective is to enhance sensitivity for unprivileged groups while maintaining an acceptable FPR, which refers to a level that is not significantly higher than the overall population's FPR, ensuring a balance between minimizing incorrect positive predictions and effectively detecting true positive cases. It determines subgroup-specific thresholds for unprivileged groups to ensure that their sensitivity aligns with that of the overall population on training data, encompassing the sensitivity of all subgroups. For the remaining groups, the standard threshold at 0.5 is employed. This conventional 0.5 threshold can be adjusted as a hyperparameter to better suit specific context requirements. The whole procedure is detailed in Algorithm 1. To illustrate the type of correction that PSTA enforces, Supplementary Fig. S2c shows that, after learning the optimal threshold with PSTA on the training set for the unprivileged group (male, in this case), the FPR of the unprivileged group reaches an acceptable value in the test set. The PSTA method focuses on enhancing performance for the most disadvantaged groups to achieve fairness. This approach contrasts with other methods like CPP which, aiming for uniform FN and FP error rates across subgroups, adjusts predicted probabilities for all the groups.

In this study, we use three different pre-processing mitigation techniques (SUP, RW, and DIR), which operate over training data, and two different post-processing mitigation techniques (CPP and PSTA), which operate over the model's predictions. See a graphical description of how they operate in Supplementary Fig. S3. Generally, pre-processing and post-processing techniques are model-agnostic. There exists a third type of bias mitigation technique, called in-processing methods, which operate during classifier construction, leading to a significant reliance on the specific type of model in use. Given our study's focus on broadly applicable and adaptable mitigation techniques, we concentrated on pre- and post-processing methods, as their adaptability aligns better with our research objective. An in-depth comparison of bias mitigation techniques is beyond the scope of this study. Our aim is to demonstrate that, with a diverse toolkit, ML practitioners can expect to build a model with increased fairness without a significant loss of predictive performance. In this study, we focus on bias mitigation concerning a single protected attribute. However, we study the unintended consequences of mitigating bias for a given protected attribute on fairness regarding the untreated protected attributes and provide preliminary results in the Supplementary material. Future work will require intersectional fairness, that is, addressing bias regarding several protected attributes at the same time, as individuals may belong to multiple unprivileged groups. Note that not all the techniques may be able to deal with several protected attributes at once (an effective workaround is defining subgroups as the Cartesian product of the values of the different protected attributes); and even the fairness metrics need to be adapted for intersectional analysis.

Input: A training set $D = \{(y_i, x_i, c_i)\}_{i=1}^{n}$, where y_i are the values of the descriptive variables, x_i is the value of the protected attribute and c_i the class for individual i, and $\hat{P} = \{\hat{p}_i\}_{i=1}^{n}$ where $\hat{p}_i$ is the vector of probabilities predicted for individual i and each class label.

Result: A vector of threshold values $\boldsymbol{\theta}$, one per group $\boldsymbol{\theta} = (\theta_x)_{x \in \Omega_X}$

Initialize threshold vector to $\theta_x = 0.5$ for all the groups, $x \in \Omega_X$
$sensitivity_{\text{overall}}$:= Calculate sensitivity in the overall training set with current θ, i.e., all decision thresholds at 0.5
for each *unprivileged* group, u in Ω_X **do**
    Initialize minimum difference $minDiff = \infty$
    Initialize optimal threshold for group $\hat{\theta}_u = 0.5$
    for each possible threshold value $\theta' \in [0, 1]$ **do**
        Calculate the sensitivity of the unprivileged group u, $sensitivity_u$
        $diff := \text{abs}(sensitivity_u - sensitivity_{overall})$
        if $diff < minDiff$ **then**
            $minDiff = diff$
            $\hat{\theta}_u = \theta'$
        end if
    end for
    Adopt threshold $\theta_u = \hat{\theta}_u$
end for
Use the vector of thresholds $\boldsymbol{\theta}$ to predict the class of new data instances

Algorithm 1. Population Sensitivity-Guided Threshold Adjustment (PSTA).

Results

In this section, we study the behavior of the different ML models before and after bias mitigation techniques are applied. Firstly, ML models are learned from data and their predictive performance and unfair bias are quantified ("Initial model performance and bias assessment"). Secondly, bias mitigation techniques are applied to these models and to assess their efficiency, we evaluate the interplay between predictive performance and fairness in the adapted models ("Model performance after bias mitigation"). We perform this analysis on four datasets for the prediction of depression: LONGSCAN, FUUS, NHANES, and UKB.

Initial model performance and bias assessment

For the sake of simplicity, in this manuscript we only report results with LR models. Results with XGB models, qualitatively similar to those of LR, are available in the Supplement. Regarding predictive performance, LR models achieve lower prediction accuracy with LONGSCAN and FUUS datasets, with BAcc of 0.621 (95% CI 0.577–0.664) and 0.615 (95% CI 0.598–0.632), respectively. The predictive performance of LR models is considerably better with NHANES and UKB datasets, with BAcc of 0.719 (95% CI 0.711–0.727) and 0.729 (95% CI 0.725–0.734), respectively. Aligned with these results, AUC-ROC measurements are available in Supplementary Table S9. These results support that the predictive ability of ML models increases with the amount of training data, as NHANES and UKB are larger. Figure 1 displays the TPR of the base and debiased LR models by dataset, protected attribute and subgroup, providing a comprehensive comparison of fairness performance before and after the application of bias mitigation techniques. In each plot, the horizontal shift or difference in performance for different subgroups reveals the bias. In each frame, the upper plot shows results before applying a bias mitigation technique, and the lower plot shows results after mitigation. Similar figures for FPR measurements are available in the Supplementary material. We use Tukey's range test[28] for pairwise comparisons across multiple groups to assess statistical significance in true positive rates and false positive rates among these groups, and to map the 95% confidence intervals. Let us analyze the bias assessment by (type of) protected attribute: demographic factors, socioeconomic status, and co-morbidities.

Protected attributes: demographic factors (sex, ethnicity, nationality)

There are consistent sex differences across datasets, as shown in the top row of Fig. 1, with higher TPRs for female subjects. The differences in TPRs between sexes are only statistically significant at the 95% confidence level for the LONGSCAN, NHANES, and UKB datasets. Note that there are no sex differences in rates of depression among subjects in the FUUS dataset (see Supplementary Table S1). Interestingly, the mean difference between sexes in the UKB and NHANES datasets (0.1121, 0.167, $p < 0.001$) is less than in the LONGSCAN dataset (0.4103, $p < 0.001$). Sex, the top importance feature in LR and XGB for the LONGSCAN dataset, has a much lower rank in the feature importance ranking for the NHANES and UKB datasets (see the 'feature importance rankings' in the Supplement), which means depression outcome is less sensitive to sex bias in the NHANES and UKB datasets with a large sample size and features compared to the LONGSCAN dataset. We highlight the benefit obtained from considering a larger number of risk factors in the predictive model and adequate sample size to reduce bias. This evidence is in line with[12], where enhanced data collection is pointed out as a means of lessening discrimination without sacrificing accuracy. The second row of Fig. 1 shows differences in TPRs between racial groups, which were not statistically significant, with black subjects and "other/multiracial" subjects having the lowest true-positive rates for LONGSCAN and NHANES datasets, respectively; except for the case between black subjects and white subjects in the LONGSCAN dataset, their TPRs have non-overlapping confidence intervals,

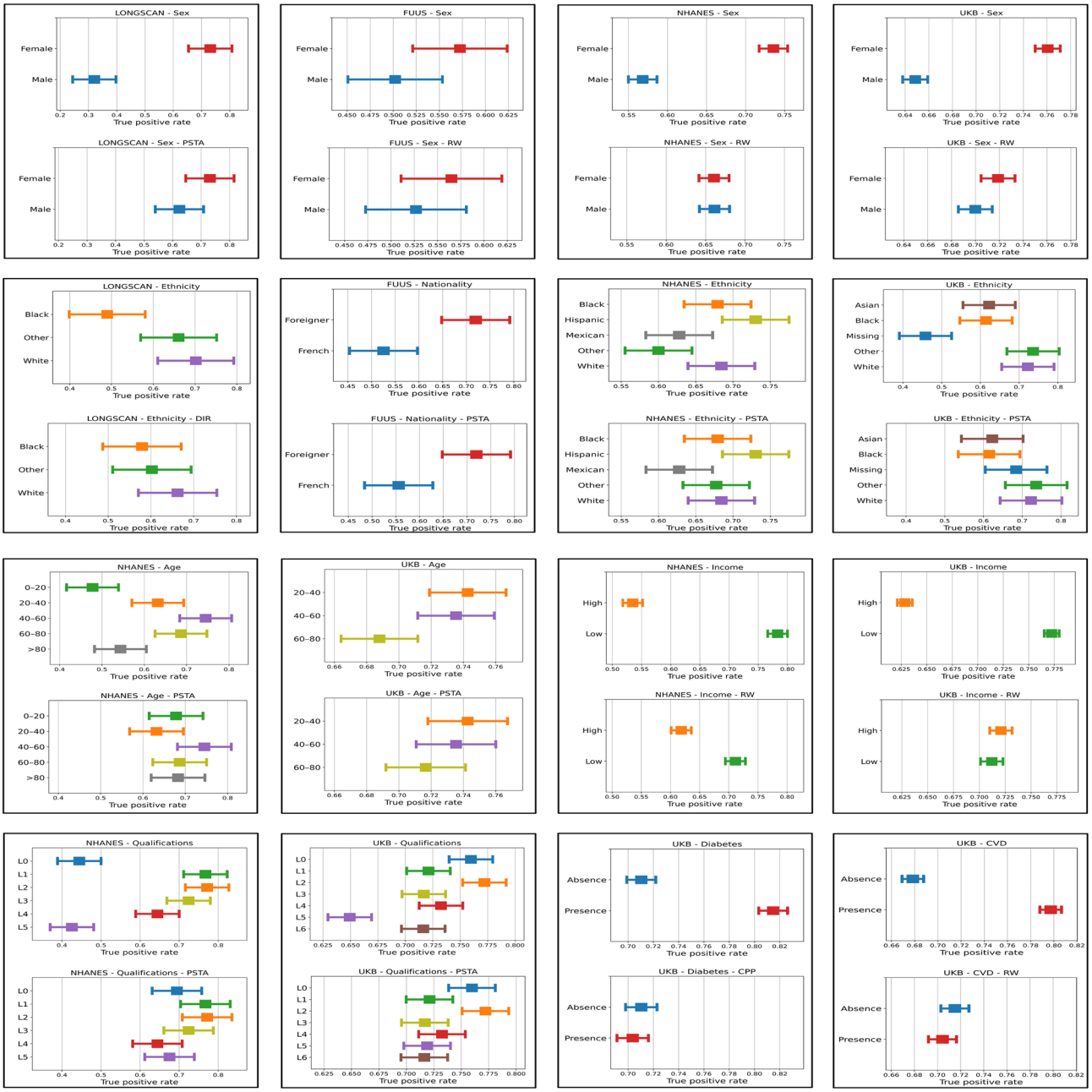

Figure 1. Comparative analysis of group-specific TPRs for LR classifiers: baseline and bias mitigation outcomes. Each plot represents a Dataset-Protected attribute pair, with paired rows displaying base classifiers and debiased classifiers, reporting the best results among the tested bias mitigation algorithms. Points represent mean TRP and error bars indicate 95% confidence intervals over k-fold cross validation. Note that the base classifiers show regularly biased behaviors due to a lack of representation, varying rates of depression across groups, unequal distribution of features between groups, or a combination of any of these characteristics in the four different study populations.

indicating a significant difference. Interestingly, the UKB subjects in the "do not know/prefer not to answer" ('Missing') group have the lowest TPR, compared with others. As shown in Fig. 1, TPRs for nationality have non-overlapping confidence intervals. Specifically, foreign subjects have a higher rate than French subjects. We note that foreigners have a higher observed depression rate in the training set. Supplementary Fig. S5 displays similar patterns regarding FPR, although the differences between racial groups in the UKB dataset are more limited.

Protected attributes: socioeconomic status (age, income, academic qualifications)
We find that the TPRs do not differ much across age groups with many overlapping intervals. However, subjects under 20 years of age have the lowest TPR. This may be partially due to the fact that small subset sizes (see Supplementary Table S1) may not reflect accurate depression rates amongst the subpopulation of adolescents and

young adults, whose symptoms of depression and other mental illnesses have increased significantly over the past decades[40]. Differences in TPRs in the NHANES dataset are also observed between qualification groups. As seen in Fig. 1, subjects in the "Level 5" group have the lowest TPR, compared with others in both the NHANES and UKB datasets. Subjects in the "Level 0" (Refused/Don't know/Missing) group in the NHANES dataset are also on the unprivileged side. As shown in Fig. 1, TPRs for income have non-overlapping confidence intervals. Specifically, low-income subjects have a much higher rate than high-income subjects. We note that low-income subjects have a higher baseline observed event rate. Supplementary Fig. S5 displays similar patterns regarding FPR, although the differences between age groups in the NHANES dataset are more limited.

Protected attributes: comorbidities (CVD, diabetes)
As shown in Fig. 1, TPRs for CVD and diabetes all have non-overlapping confidence intervals. Specifically, individuals experiencing CVD/diabetes have a much higher rate than subjects without CVD/diabetes. These inequitable outcomes support that CVD and diabetes should be considered as important comorbidities of depression. Similar behaviors are observed regarding FPR in Supplementary Fig. S5.

Model performance after bias mitigation

To assess bias mitigation, we analyze changes in fairness metrics between the previously discussed base models and the new classifiers obtained after bias mitigation. Moreover, given the clinical application, most people would not find it fair to reduce discriminatory outcomes if it identifies fewer actual positives overall. There exists an open discussion in the related literature[41,42] regarding the actual existence of the so-called fairness-accuracy trade-off when bias mitigation is implemented, meaning that predictive performance is reduced if one tries to make the model fairer. A trade-off of accuracy for fairness is often undesirable in healthcare. Thus, we have two objectives: increasing accuracy and decreasing discrimination. We report the results using 2D points combining two metrics: (i) a fairness metric, measured by EOD and (ii) a standard ML performance metric, measured by BAcc, on the test set, to determine whether our models for depression prediction are experiencing a fairness-accuracy trade-off. A model can be considered as fair if EOD is between − 0.1 and 0.1, its ideal value is 0 and for BAcc, the larger the better. Figure 2 presents, for each dataset and protected attribute, the 2D points in the

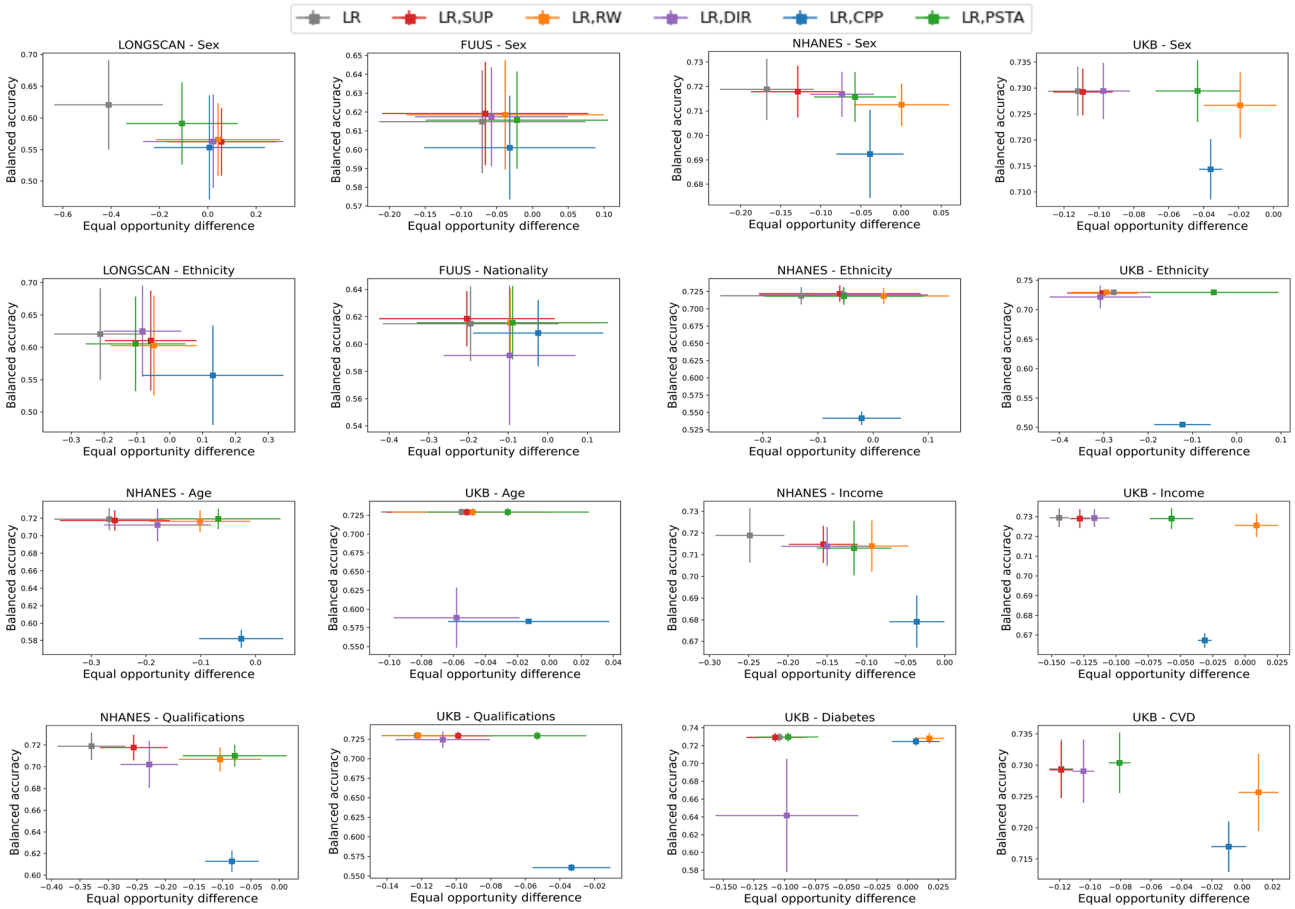

Figure 2. Fairness-accuracy performance in terms of EOD vs. BAcc of the base model and the new classifiers after applying five bias mitigation algorithms to LR classifiers. Points represent mean BAcc-TRP and error bars indicate the standard deviation over k-fold cross validation. Each plot shows the results per subgroup for a Dataset-Protected attribute pair.

fairness-accuracy space achieved by each bias mitigation technique. The limited BAcc variation together with larger EOD variations observed in the assessment of different techniques in some subplots (e.g., CPP or PSTA applied to UKB-Ethnicity) can be explained by the relatively small sample size of the subgroups (see Supplementary Table S1). BAcc is less sensitive to even larger changes in group-specific TRPs when one or more subgroups are small. The base model, without bias mitigation, is also shown (gray point and cross). Absolute and relative changes in performance and fairness metrics behind these plots are provided in detail in Supplementary Table S7. The implemented bias methods help to improve fairness, for all protected attribute perspectives. This is a desirable mitigation result. Note that while Fig. 2 summarizes the overall impact of bias mitigation techniques, Fig. 1 showcases group-specific TPRs for LR classifiers before and after applying the best-performing bias-mitigation techniques for each protected attribute, uncovering the fine-grained performance enhancements leading to fairness improvement. This allows for a detailed evaluation of their effectiveness across different study populations. It is noteworthy that in general debiasing through RW, DIR, and PSTA substantially improves fairness without compromising model accuracy. SUP only partially achieves fairness between groups. This method removes the protected attribute from the training dataset, as they are considered biased features. This result suggests that bias is not only contained in those features but elsewhere. DIR not only excludes these attributes but also adjusts non-protected features that are highly correlated to them. In any case, this repair tool is a good baseline to investigate whether this bias comes from non-protected features. On the other hand, the CPP technique preserves precision when calibrating recall, which results in BAcc reduction, according to most of the protected attributes, except for the diabetes attribute. In this study, the PSTA method is considered as an alternative to post-processing methods like CPP that, aiming for group fairness, may inadvertently compromise performance for all subgroups. We compare the five considered mitigation techniques across a total of 32 scenarios (all the combinations of model, dataset and protected attribute in this study), using the weighted harmonic mean of predictive performance and fairness metrics (BAcc and EOD, resp.) to assess the dual fairness-accuracy objective:

$$\Delta = \frac{(1+\beta^2)\cdot \mathrm{BAcc}\cdot(1-|\mathrm{EOD}|)}{\beta^2\cdot \mathrm{BAcc}+(1-|\mathrm{EOD}|)}$$

with $\beta = 0.5$, indicating a preference for performance over fairness. This is particularly pertinent in healthcare, where compromising accuracy for fairness can have ethical implications, making widespread fairness adjustments less justifiable if they lead to reduced detection of health outcomes. In this comparison, SUP does not outperform other methods in any case, CPP is superior in 1 case (3.125%), DIR in 1 case (3.125%), RW in 12 cases (37.5%), and PSTA in 18 cases (56.25%). In comparison with CPP, the other post-processing technique used in this study, PSTA seems to stand out. A future study of PSTA should find (empirical) evidence that characterizes the scenarios where this technique performs competitively in comparison with other post-processing bias mitigation techniques. Furthermore, it is important to highlight that, when applying bias-mitigation methods to the LONGSCAN dataset with the "sex" protected attribute, a significant decrease in BAcc is observed across all methods. This observation serves as evidence that the model may struggle to effectively learn the underlying structure of the data due to the relatively small dataset. Consequently, this emphasizes the need to have sufficient samples and relevant predictors when utilizing ML algorithms for risk prediction tasks in order to develop both fair and accurate models. Note that TPRs for the best mitigation technique (RW) regarding the "income" protected attribute on models learned from the NHANES dataset still have non-overlapping confidence intervals, although the mean difference between groups was considerably reduced from 0.2485 ($p < 0.001$) to 0.0929 ($p < 0.001$). Additionally, results according to group-specific FPR (see Supplementary Figs. S5 and S6) and AOD metrics (see Supplementary Table S7), demonstrate a consistency with the findings based on EOD, reinforcing the validity of our debiasing approach across different fairness measures.

Our analysis on the impact of bias mitigation on untreated protected attributes analysis (see Supplementary Table S12) shows that there is not a uniform trend toward worsening fairness on untreated attributes when the debiasing method mitigates bias for a single protected attribute, although this undesirable behavior happens frequently and deserves attention. SUP seems more inclined to worsening, while post-processing methods (CPP and PSTA) are more robust, as evidenced by both EOD and AOD metrics. Our analysis also suggests that when fairness is enhanced for one attribute, it may inversely affect another if the attributes are negatively correlated. More details about this analysis are provided in the Supplementary material (see Supplementary Figs. S8–S10).

Discussion

ML algorithms have achieved state-of-the-art performance in many clinical tasks. However, when applying them in these life-or-death-stakes applications, it must be understood that they can induce biases against unprivileged subgroups and precautionary actions need to be taken in different deployment stages[43]. Here, we leverage our empirical study on four datasets to analyze the need of bias mitigation techniques, as well as their effectiveness, when using ML models to predict mental health issues such as depression.

Bias found when following the standard ML approach

Our results indicate that models learned following the standard ML approach show regularly unfair biased behaviors (see Fig. 1). We find that, for the classification problem, unequal distribution of classes between groups in the training dataset can lead the predictive model to learn that one group has a higher probability of being part of one class or another. (Evidence of this behavior can be observed in Supplementary Table S1). Therefore, ML models trained on the unbalanced dataset of a trial population, even if the sample in clinical trials is representative of the patient population, provide potential inequitable outcomes if deployed without fairness analysis.

This evidence encourages ML healthcare practitioners not only to report the model performance on the overall population regardless of the subject membership to subpopulations but also to audit and address algorithmic bias.

Bias can be mitigated

Our results show that the bias mitigation techniques improve fairness compared to the no-intervention base models. All the techniques considered enhance results in terms of the difference in TPRs, in differing proportions. However, the techniques exhibit differences regarding the effect on the accuracy of the classifiers. Those learned in combination with the RW, DIR, and PSTA mitigation techniques tend to preserve predictive performance, whereas other techniques (SUP, CPP) usually compromise the level of accuracy (see Fig. 2). Our findings point out that the RW technique combined with the use of a larger number of risk factors diminishes the impact of the protected attributes (and other possible proxy attributes) on the outcome of the model, which leads to reduced algorithmic bias in NHANES and UKB datasets. This solution is appropriate in our case study as it performs well when it is integrated into predictive model learning while preserving the distribution and values of the original training data, unlike DIR and SUP. In addition, we find that our proposed post-hoc disparity mitigation method (PSTA) tends to mitigate bias while preserving predictive performance. In this method, the distinct treatment of subpopulations aims to address subgroup-specific disparities. Conventional one-size-fits-all methods might inadvertently accentuate biases, especially for underrepresented groups. Therefore, our approach ensures fairness by recognizing and addressing these unique subgroup challenges. However, as other post-doc technical solutions for imposing fairness, it might be difficult to obtain the optimal threshold for subgroups with small populations underrepresented in the training set. The choice of the method must depend on the specific domain of application and desired outcomes. In the larger effort to create a fair system overall, methods like RW and DIR not only increase the TPR for the unprivileged groups but also reduce the privileged group's FPR, resulting in fewer false positives. In contrast, PSTA increases the TPR of the unprivileged groups, accepting a higher FPR to reduce the TPR gap without intervening in the privileged groups. In primary care settings, where early detection and prevention are key, PSTA may be more appropriate, as it ensures that more people are identified for further evaluation and potential intervention, maximizing the overall benefit. Conversely, in secondary care settings focused on diagnosing and treating specific conditions, techniques such as RW and DIR may be more suitable, as they balance TPR and FPR across demographic groups, ensuring effective and fair resource allocation. We also highlight the importance of prioritizing adequate data collection before employing any debiasing technique to improve fairness in predictive models. By doing so, we aim to encourage communities to open health data, further contributing to the development of more equitable ML solutions. When mitigating bias against one protected attribute, further investigations are required to fully understand the potential inadvertent introduction or exacerbation of bias in other, untreated attributes. When required, future studies will have to address fairness in several protected attributes at once.

Fairness-accuracy, a real trade-off?

In fairness literature, the existence of a trade-off between fairness and accuracy is a common assumption; that is, that fairness cannot be improved without sacrificing predictive performance[33]. A few studies also have pointed out that this trade-off may necessitate the application of more complex methods[44,45]. Based on our empirical observations, we find that the fairness-accuracy trade-off for the datasets examined in this paper can be lessened if a set of bias mitigation techniques is considered, and not just one. Standard techniques RW, DIR, and CPP, or our proposed method PSTA can reduce bias while preserving predictive performance for specific (dataset, protected attribute) pairs. This was particularly evident for the modest-sized to large datasets (FUUS, NHANES, UKB), though the dynamics shifted a bit for the smaller dataset (LONGSCAN). Thus, this trade-off is not consistently observed in our case studies in depression prediction without requiring complex ML methods. In line with[46], this evidence encourages the ML community to intentionally propose frameworks that maximize both predictive performance enhancement and bias reduction, aiming to lessen the trade-off. There is probably no golden ticket, as there is no single best ML model for all prediction problems providing equality of outcomes naturally. ML practitioners need to figure out which combination in terms of type of classifier and bias mitigation algorithm is appropriate for the case at hand so that it produces the best results in terms of both accuracy and fairness.

In conclusion, we conducted an empirical study on four exposome datasets to show the ability of bias mitigation techniques to increase fairness of machine learning models obtained to predict depression from environmental and lifestyle data. In addition, our main effort in this work has been directed toward providing empirical evidence to encourage clinical decision-makers to carefully evaluate a proposed framework in terms of both its accuracy and fairness prior to deployment. Experimental results support the idea that it is possible to improve algorithmic fairness regarding a single protected attribute without sacrificing predictive performance. We consider that our promising results could enable a wider use of ML techniques in mental healthcare. This should inevitably go hand-in-hand with the assessment of possible biases in the models and the appropriate mitigation techniques if required. In the future, we expect to simultaneously examine the effects of having multiple protected attributes, such as ethnicity, sex, socio-economic status, geographical location, or comorbidities (type 1 and type 2 diabetes, or cardiovascular disease), as well as broadening to include protective factors data, such as intelligence, temperament, cognitive appraisal, and support from a significant person, all of which may counteract the negative effects of risk factors for depression, along with environmental and lifestyle data. In addition, we aim to integrate genetic and biological data, which are robust risk factors for depression, into our research.

Data availability

Code for data processing and analysis is available at https://github.com/ngoc-vien-dang/FairML-Depression. Detailed dataset sources and accessibility conditions are also provided in the repository's README.

Received: 26 October 2023; Accepted: 28 March 2024
Published online: 03 April 2024

References

1. Friedrich, M. J. Depression is the leading cause of disability around the world. *JAMA* **317**, 1517–1517 (2017).
2. Bachmann, S. Epidemiology of suicide and the psychiatric perspective. *Int. J. Environ. Res. Public Health* **15**, 1425 (2018).
3. Bueno-Notivol, J. *et al.* Prevalence of depression during the covid-19 outbreak: A meta-analysis of community-based studies. *Int. J. Clin. Health Psychol.* **21**, 100196 (2021).
4. Anttila, V. *et al.* Analysis of shared heritability in common disorders of the brain. *Science* **360**, eaap8757 (2018).
5. Geschwind, D. H. & Flint, J. Genetics and genomics of psychiatric disease. *Science* **349**, 1489–1494 (2015).
6. Harald, R. *et al.* An exposome perspective: Early-life events and immune development in a changing world. *J. Allergy Clin. Immunol.* **140**, 24–40 (2017).
7. Olesen, J. *et al.* The economic cost of brain disorders in Europe. *Eur. J. Neurol.* **19**, 155–162 (2012).
8. Chen, I. Y., Szolovits, P. & Ghassemi, M. Can AI help reduce disparities in general medical and mental health care?. *AMA J. Ethics* **21**, E167-179 (2019).
9. Park, Y. *et al.* Comparison of methods to reduce bias from clinical prediction models of postpartum depression. *JAMA Netw. Open* **4**, 213909-e213909 (2021).
10. Nemesure, M. *et al.* Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Sci. Rep.* **11**, 1980 (2021).
11. Hutchinson, B. & Mitchell, M. 50 years of test (un)fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability and Transparency*, 49–58 (2019).
12. Chen, I. Y., Johansson, F. D. & Sontag, D. Why is my classifier discriminatory? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 3543–3554 (2018).
13. Corrigan, P. W., Druss, B. G. & Perlick, D. A. The impact of mental illness stigma on seeking and participating in mental health care. *Psychol. Sci. Public Interest* **15**, 37–70 (2014).
14. Wong, E. C., Collins, R. L., Cerully, J., Seelam, R. & Roth, B. Differences in mental illness stigma and discrimination among Californians experiencing mental health challenges. RAND Corporation research report no. RR-1441-CMHSA (2017). https://doi.org/10.7249/RR1441.
15. Albert, R. P. Why is depression more prevalent in women?. *J. Psychiatry Neurosci.* **40**, 219–221 (2015).
16. Lubin, F., Lusky, A., Chetrit, A. & Dankner, R. Lifestyle and ethnicity play a role in all-cause mortality. *J. Nutr.* **133**, 1180–1185 (2003).
17. Runyan, D. *et al.* Longitudinal studies on child abuse and neglect (LONGSCAN) ages 0-18, version 1.4 dataset. National Data Archive on Child Abuse and Neglect (2014).
18. Tran, A. *et al.* Health assessment of French university students and risk factors associated with mental health disorders. *PLoS ONE* **12**, e0188187 (2017).
19. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.
20. Sudlow, C. *et al.* UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
21. Mhasawade, V., Zhao, Y. & Chunara, R. Machine learning and algorithmic fairness in public and population health. *Nat. Mach. Intell.* **3**, 659–666 (2021).
22. Feng, Q., Zhang, Q., Du, Y., Ye, Y. & He, Q. Associations of physical activity, screen time with depression, anxiety and sleep quality among Chinese college freshmen. *PLoS ONE* **9**, e100914 (2014).
23. Bayram, N. & Bilgel, N. The prevalence and socio-demographic correlations of depression, anxiety and stress among a group of university students. *Soc. Psychiat. Epidemiol.* **43**, 667–672 (2008).
24. Ovuga, E., Boardman, J. & Wasserman, D. Undergraduate student mental health at Makerere University, Uganda. *World Psychiatry* **5**, 51–52 (2006).
25. Kroenke, K., Spitzer, R. L. & Williams, J. B. W. The PHQ-9: Validity of a brief depression severity measure. *J. Gen. Intern. Med.* **16**, 606–613 (2001).
26. Yu, H., Huang, F. & Lin, C. Dual coordinate descent methods for logistic regression and maximum entropy models. *Mach. Learn.* **85**, 41–75 (2011).
27. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (2016).
28. Tukey, J. W. Comparing individual means in the analysis of variance. *Biometrics* **5**, 99–114 (1949).
29. Bellamy, R. K. E. *et al.* AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint (2018). https://arxiv.org/abs/1810.01943.
30. Mosteiro, P., Kuiper, J., Masthoff, J., Scheepers, F. & Spruit, M. Bias discovery in machine learning models for mental health. *Information* **13**, 237 (2022).
31. Hardt, M., Price, E. & Srebro, N. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 3323–3331 (2016).
32. Xu, J. *et al.* Algorithmic fairness in computational medicine. *EBioMedicine* **84**, 104250 (2022).
33. Calders, T., Kamiran, F. & Pechenizkiy, M. Building classifiers with independency constraints. In *IEEE International Conference on Data Mining Workshops*, 13–18 (2009).
34. Verma, S. & Rubin, J. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, 1–7 (2018).
35. Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C. & Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268 (2015).
36. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. & Weinberger, K. Q. On fairness and calibration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (2017).
37. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **1947**, 50–60 (1947).
38. Rodolfa, K. T. *et al.* Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 142–153 (2020).
39. Jang, T., Shi, P. & Wang, X. Group-aware threshold adaptation for fair classification. *In Proceedings of the AAAI Conference on Artificial Intelligence* **36**, 6988–6995 (2022).

40. Twenge, J. M., Cooper, A. B., Joiner, T., Duffy, M. & Binau, S. Age, period, and cohort trends in mood disorder indicators and suicide-related outcomes in a nationally representative dataset, 2005–2017. *J. Abnorm. Psychol.* **128**, 185–199 (2019).
41. Calders, T., Kamiran, F. & Pechenizkiy, M. Building classifiers with independency constraints. In *IEEE International Conference on Data Mining Workshops*, 13–18 (2009).
42. Menon, A. K. & Williamson, R. C. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 107–118 (2018).
43. de Hond, A., Leeuwenberg, A., Hooft, L. *et al.* Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: A scoping review. *NPJ Digit. Med.* **5**, 2 (2022).
44. Friedler, S. A. *et al.* A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 329–338 (2019).
45. Zafar, M. B., Valera, I., Rodriguez, M. G. & Gummadi, K. P. Fairness constraints: mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 962–970 (2017).
46. Rodolfa, K. T., Lamba, H. & Ghani, R. Empirical observation of negligible fairness-accuracy trade-offs in machine learning for public policy. *Nat. Mach. Intell.* **3**, 896–904 (2021).

Acknowledgements

VND, RHM, CC, JHG, and KL have received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 848158, EarlyCause.

Author contributions

Vien Ngoc Dang: methodology, investigation, data curation, software, visualization, writing- original draft, review and editing. Anna Cascarano, Rosa H. Mulder, Charlotte Cecil, and Maria A. Zuluaga: writing - review and editing. Jerónimo Hernández-González: methodology, conceptualization, writing - original draft, review and editing, co-supervision. Karim Lekadir: conceptualization, resources, writing - review, supervision.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-58427-7.

Correspondence and requests for materials should be addressed to V.N.D.