01010
01010
01010
information

Special Issue Reprint

Advances in Explainable Artificial Intelligence

Edited by
Gabriele Gianini and Pierre-Edouard Portier

mdpi.com/journal/information

MDPI

information

Article

Bias Discovery in Machine Learning Models for Mental Health

Pablo Mosteiro [1,*], Jesse Kuiper [1], Judith Masthoff [1], Floortje Scheepers [2] and Marco Spruit [1,3,4]

[1] Department of Information and Computing Sciences, Utrecht University, 3584 CS Utrecht, The Netherlands; jesse94kuiper@gmail.com (J.K.); j.f.m.masthoff@uu.nl (J.M.); m.r.spruit@lumc.nl (M.S.)
[2] Afdeling Psychiatrie, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands; f.e.scheepers-2@umcutrecht.nl
[3] Department of Public Health and Primary Care, Leiden University Medical Center, 2333 ZA Leiden, The Netherlands
[4] Leiden Institute of Advanced Computer Science, Leiden University, 2311 EZ Leiden, The Netherlands
* Correspondence: p.mosteiro@uu.nl

Abstract: Fairness and bias are crucial concepts in artificial intelligence, yet they are relatively ignored in machine learning applications in clinical psychiatry. We computed fairness metrics and present bias mitigation strategies using a model trained on clinical mental health data. We collected structured data related to the admission, diagnosis, and treatment of patients in the psychiatry department of the University Medical Center Utrecht. We trained a machine learning model to predict future administrations of benzodiazepines on the basis of past data. We found that gender plays an unexpected role in the predictions—this constitutes bias. Using the AI Fairness 360 package, we implemented reweighing and discrimination-aware regularization as bias mitigation strategies, and we explored their implications for model performance. This is the first application of bias exploration and mitigation in a machine learning model trained on real clinical psychiatry data.

Keywords: fairness; bias; artificial intelligence; machine learning; psychiatry; health; mental health

Citation: Mosteiro, P.; Kuiper, J.; Masthoff, J.; Scheepers, F.; Spruit, M. Bias Discovery in Machine Learning Models for Mental Health. *Information* **2022**, *13*, 237. https://doi.org/10.3390/info13050237

Academic Editors: Gabriele Gianini and Pierre-Edouard Portier

Received: 23 March 2022
Accepted: 3 May 2022
Published: 5 May 2022

1. Introduction

For over ten years, there has been increasing interest in the psychiatry domain for using machine learning (ML) to aid psychiatrists and nurses [1]. Recently, multiple approaches have been tested for violence risk assessment (VRA) [2–4], suicidal behaviour prediction [5], and the prediction of involuntary admissions [6], among others.

Using ML for clinical psychiatry is appealing both as a time-saving instrument and as a way to provide insights to clinicians that might otherwise remain unexploited. Clinical ML models are usually trained on patient data, which includes some protected attributes, such as gender or ethnicity. We desire models to give equivalent outputs for equivalent patients that differ only in the value of a protected attribute [7]. Yet, a systematic assessment of the fairness of ML models used for clinical psychiatry is lacking in the literature.

As a case study, we focused on the task of predicting future administrations of benzodiazepines. Benzodiazepines are prescription drugs used in the treatment of, for example, anxiety and insomnia. Long-term use of benzodiazepines is associated with increased medical risks, such as cancer [8]. In addition, benzodiazepines in high doses are addictive, with complicated withdrawal [9]. From a clinical perspective, gender should not play a role in the prescription of benzodiazepines [10,11]. Yet, biases in the prescription of benzodiazepines have been explored extensively in the literature; some protected attributes that contributed to bias were prescriber gender [12], patient ethnicity [13,14], and patient gender [15], as well as interaction effects between some of these protected attributes [16,17]. There is no conclusive consensus regarding these correlations, with some studies finding no correlations between sociodemographic factors and benzodiazepines prescriptions [18].

We explored the effects of gender fairness bias on a model trained to predict the future administration of benzodiazepines to psychiatric patients based on past data, including

past doses of benzodiazepines. A possible use case of this model is to identify patients that are at risk of taking benzodiazepines for too long. We hypothesized that our model is likely to unfairly use the patient's gender in making predictions. If that is the case, then mitigation strategies must be put in place to reduce this bias. We expect that there will be a cost to predictive performance.

Our research questions are:

1. For a model trained to predict future administrations of benzodiazepines based on past data, does gender unfairly influence the decisions of the model?
2. If gender does influence the decisions of said model, how much model performance is sacrificed when applying mitigation strategies to avoid the bias?

To answer these questions, we employed a patient dataset from the University Medical Center (UMC) Utrecht and trained a model to predict future administrations of benzodiazepines. We applied the bias discovery and mitigation toolbox AI Fairness 360 [19]. Whenever we found that gender bias was present in our model, we presented an appropriate way to mitigate this bias. Our main contribution is a first implementation of a fairness evaluation and mitigation framework on real-world clinical data from the psychiatry domain. We present a way to mitigate a real and well-known bias in benzodiazepine prescriptions, without loss of performance.

In Section 2, we describe our materials and methods, including a review of previous work in the field. In Section 3, we present our results, which we discuss in Section 4. We present our conclusions in Section 5.

2. Materials and Methods

2.1. Related Work

The study of bias in machine learning has garnered attention for several years [20]. The authors in [21] outlined the dangers of selection bias. Even when researchers attempt to be unbiased, problems might arise, such as bias from an earlier work trickling down into a new model [22] or implicit bias from variables correlated with protected attributes [23,24]. The authors in [25] reviewed bias in machine learning, noting also that there is no industry standard for the definition of *fairness*. The authors in [26] evaluated bias in a machine learning model used for university admissions; they also point out the difference between *individual* and *group* fairness, as do [27]. The authors in [28,29] provided theoretical frameworks for the study of fairness. Along the same lines, refs. [30,31] provided metrics for the evaluation of fairness. The authors in [32,33] recommend methods for mitigating bias.

As for particular applications, refs. [34–36] studied race and gender bias in facial analysis systems. The authors in [37] evaluated fairness in dialogue systems, and while they did not actually evaluate ML models, ref. [38] highlighted the importance of bias mitigation in AI for education.

In the medical domain, ref. [39] pointed out the importance of bias mitigation. Indeed, ref. [40] uncovered bias in post-operative complication predictions. The authors in [41] found that disparities metrics change when transferring models across hospitals. Finally, ref. [42] explored the impact of random seeds on the fairness of classifiers using clinical data from MIMIC-III, and found that small sample sizes can also introduce bias.

No previous study on ML fairness or bias focuses on the psychiatry domain. This domain is interesting because bias seems to be present in the daily practice. We have already discussed in the introduction how bias is present in the prescription of benzodiazepines. There are also gender disparities in the prescription of zolpidem [43] and in the act of seeking psychological help [44]. The authors in [45] also found racial disparities in clinical diagnoses of mania. Furthermore, psychiatry is a domain where a large amount of data is in the form of unstructured text, which is starting to be exploited for ML solutions [46,47]. Previous work has also focused on the explainability of text-based computational support systems in the psychiatry domain [48]. It will be crucial—as these text-based models begin to be applied in the clinical practice—to ensure that they too are unbiased towards protected attributes.

2.2. Data

We employed de-identified patient data from the Electronic Health Records (EHRs) from the psychiatry department at the UMC Utrecht. Patients in the dataset were admitted to the psychiatry department between June 2011 and May 2021. The five database tables included were: admissions, patient information, medication administered, diagnoses, and violence incidents. Table 1 shows the variables present in each of the tables.

Table 1. Datasets retrieved from the psychiatry department of the UMC Utrecht, with the variables present in each dataset that are used for this study. Psychiatry is divided into four *nursing wards*. For the "medication" dataset, the "Administered" and "Not administered" variables contain, in principle, the same information; however, sometimes only one of them is filled.

Dataset	Variable	Type
Admissions	Admission ID	Identifier
	Patient ID	Identifier
	Nursing ward ID	Identifier
	Admission date	Date
	Discharge date	Date
	Admission time	Time
	Discharge time	Time
	Emergency	Boolean
	First admission	Boolean
	Gender	Man/Woman
	Age at admission	Integer
	Admission status	Ongoing/Discharged
	Duration in days	Integer
Medication	Patient ID	Identifier
	Prescription ID	Identifier
	ATC code (medication ID)	String
	Medication name	String
	Dose	Float
	Unit (for dose)	String
	Administration date	Date
	Administration time	Time
	Administered	Boolean
	Dose used	Float
	Original dose	Float
	Continuation After Suspension	Boolean
	Not administered	Boolean
Diagnoses	Patient ID	Identifier
	Diagnosis number	Identifier
	Start date	Date
	End date	Date
	Main diagnosis group	Categorical
	Level of care demand	Numeric
	Multiple problem	Boolean
	Personality disorder	Boolean
	Admission	Boolean
	Diagnosis date	Date
Aggression	Patient ID	Identifier
	Date of incident	Date
	Start time	Time
Patient	Patient ID	Identifier
	Age at start of dossier	Integer

We constructed a dataset where each data point was 14 days after the admission of a patient. We selected only completed admissions (admission status = "discharged") that lasted at least 14 days (duration in days ≥ 14). A total of 3192 admissions (i.e., data points) were included in our dataset. These were coupled with data from the other four tables mentioned above. The nursing ward ID was converted to four binary variables; some rows did not belong to any nursing ward ID (because, for example, the patient was admitted outside of psychiatry and then transferred to psychiatry); these rows have zeros for all four nursing ward ID columns.

For diagnoses, the diagnosis date was not always present in the dataset. In that case, we used the end date of the treatment trajectory. If that was also not present, we used the start date of the treatment trajectory. One of the entries in the administered medication table had no date of administration; this entry was removed. We only consider administered medication (administered = `True`). Doses of various tranquillizers were converted to an equivalent dose of diazepam, according to Table 2 [49]. (This is the normal procedure when investigating benzodiazepine use. All benzodiazepines have the same working mechanism. The only differences are the half-life and the peak time. So, when studying benzodiazepines, it is allowed to make an equivalent dose of one specific benzodiazepine).

Table 2. List of tranquillizers considered in this study, along with the multipliers used for scaling the doses of those tranquillizers to a diazepam-equivalent dose. The last column is the inverse of the centre column.

Tranquillizer	Multiplier	mg/(mg Diazepam)
Diazepam	1.0	1.00
Alprazolam	10.0	0.10
Bromazepam	1.0	1.00
Brotizolam	40.0	0.03
Chlordiazepoxide	0.5	2.00
Clobazam	0.5	2.00
Clorazepate potassium	0.75	1.33
Flunitrazepam	0.1	10
Flurazepam	0.33	3.03
Lorazepam	5.0	0.20
Lormetazepam	10.0	0.10
Midazolam	1.33	0.10
Nitrazepam	1.0	1.00
Oxazepam	0.33	3.03
Temazepam	1.0	1.00
Zolpidem	1.0	1.00
Zopiclone	1.33	0.75

For each admission, we obtained the age of the patient at the start of the dossier from the patient table. The gender is reported in the admissions table; only the gender assigned at birth is included in this dataset. We counted the number of violence incidents before admission and the number of violence incidents during the first 14 days of admission. The main diagnosis groups were converted to binary values, where 1 means that this diagnosis was present for that admission, and that it took place during the first 14 days of admission. Other binary variables derived from the diagnoses table were "Multiple problem" and "Personality disorder". For all diagnoses present for a given admission, we computed the maximum and minimum "levels of care demand", and saved them as two new variables. Matching the administered medication to the admissions by patient ID and date, we computed the total amount of diazepam-equivalent benzodiazepines administered in the first 14 days of admission, and the total administered in the remainder of the admission. The former is one of the predictor variables. The target variable is binary,

i.e., whether benzodiazepines were administered during the remainder of the admission or not.

The final dataset consists of 3192 admissions. Of these, 1724 admissions correspond to men, while 1468 correspond to women. A total of 2035 admissions had some benzodiazepines administered during the first 14 days of admission, while 1980 admissions had some benzodiazepines administered during the remainder of the admission. Table 3 shows the final list of variables included in the dataset.

Table 3. List of variables in the final dataset.

Variable	Type
Patient ID	Numeric
Emergency	Binary
First admission	Binary
Gender	Binary
Age at admission	Numeric
Duration in days	Numeric
Age at start of dossier	Numeric
Incidents during admission	Numeric
Incidents before admission	Numeric
Multiple problem	Binary
Personality disorder	Binary
Minimum level of care demand	Numeric
Maximum level of care demand	Numeric
Past diazepam-equivalent dose	Numeric
Future diazepam-equivalent dose	Numeric
Nursing ward: Clinical Affective and Psychotic Disorders	Binary
Nursing ward: Clinical Acute and Intensive Care	Binary
Nursing ward: Clinical Acute and Intensive Care Youth	Binary
Nursing ward: Clinical Diagnosis and Early Psychosis	Binary
Diagnosis: Attention Deficit Disorder	Binary
Diagnosis: Other issues that may be a cause for concern	Binary
Diagnosis: Anxiety disorders	Binary
Diagnosis: Autism spectrum disorder	Binary
Diagnosis: Bipolar Disorders	Binary
Diagnosis: Cognitive disorders	Binary
Diagnosis: Depressive Disorders	Binary
Diagnosis: Dissociative Disorders	Binary
Diagnosis: Behavioural disorders	Binary
Diagnosis: Substance-Related and Addiction Disorders	Binary
Diagnosis: Obsessive Compulsive and Related Disorders	Binary
Diagnosis: Other mental disorders	Binary
Diagnosis: Other Infant or Childhood Disorders	Binary
Diagnosis: Personality Disorders	Binary
Diagnosis: Psychiatric disorders due to a general medical condition	Binary
Diagnosis: Schizophrenia and other psychotic disorders	Binary
Diagnosis: Somatic Symptom Disorder and Related Disorders	Binary
Diagnosis: Trauma- and stressor-related disorders	Binary
Diagnosis: Nutrition and Eating Disorders	Binary

2.3. Evaluation Metrics

The performance of the model is to be evaluated by the use of the balanced accuracy (average of true positive rate and true negative rate) and the F1 score. (As seen in Section 2.2, the distribution of data points across classes is almost balanced. With that in mind, we could have used accuracy instead of balanced accuracy. However, we had decided on an evaluation procedure before looking at the data, based on previous experience in the field.

We find no reason to believe that our choice should affect the results significantly.) As for quantifying bias, we used four metrics:

- *Statistical Parity Difference:* Discussed in [26] as the difference between the correctly classified instances for the privileged and the unprivileged group. If the statistical parity difference is 0, then the privileged and unprivileged groups receive the same percentage of positive classifications. Statistical parity is an indicator for representation and therefore a group fairness metric. If the value is negative, the privileged group has an advantage.
- *Disparate Impact:* Computed as the ratio of the rate of favourable outcome for the unprivileged group to that of the privileged group [31]. This value should be close to 1 for a fair result; lower than 1 implies a benefit for the privileged group.
- *Equal Opportunity Difference:* The difference between the true positive rates between the unprivileged group and the privileged group. It evaluates the ability of the model to classify the unprivileged group compared to the privileged group. The value should be close to 0 for a fair result. If the value is negative, then the privileged group has an advantage.
- *Average Odds Difference:* The difference between false positives rates and true positive rates between the unprivileged group and privileged group. It provides insights into a possible positive biases towards a group. This value should be close to 0 for a fair result. If the value is negative, then the privileged group has an advantage.

2.4. Machine Learning Methods

We used AI Fairness 360, a package for the discovery and mitigation of bias in machine learning models. The protected attribute in our dataset is gender, while the favourable class is "man". We employ two classification algorithms implemented in ScikitLearn [50]: logistic regression and random forest (We consider these models because they are simple, widely available and widely used within and beyond the clinical field). For logistic regression, we use the "liblinear" solver. For the random forest classifier, we use 500 estimators, with min_samples_leaf equal to 25.

There are three types of bias mitigation techniques: *pre-processing*, *in-processing*, and *post-processing* [23]. Pre-processing techniques mitigate bias by removing the underlying discrimination from the dataset. In-processing techniques are modifications to the machine learning algorithms to mitigate bias during model training. Post-processing techniques seek to mitigate bias by equalizing the odds post-training. We used two methods for bias mitigation. As a *pre-processing* method, we used the reweighing technique of [32], and retrained our classifiers on the reweighed dataset. As an *in-processing* method, we added a discrimination-aware regularization term to the learning objective of the logistic regression model. This is called a *prejudice remover*. We set the fairness penalty parameter `eta` to 25, which is high enough that prejudice will be removed aggressively, while not too high, such that accuracy would be significantly compromised [33]. Both of these techniques were seamlessly implemented in AI Fairness 360. To apply *post-processing* techniques in practice, one needs a training set and a test set; once the model is trained, the test set is used to determine how outputs should be modified in order to limit bias. However, in clinical applications, datasets tend to be small, so we envision a realistic scenario in which the entire dataset is used for development, making the use of post-processing methods impossible. For this reason, we did not study these methods further. The workflow of data, models, and bias mitigation techniques is shown in Figure 1.

To estimate the uncertainty due to the choice of training data, we used 5-fold cross-validation, with patient IDs as group identifiers to avoid using the same sample for development and testing. Within each fold, we again split the development set into 62.5% training and 37.5% validation, once again with patient IDs as group identifiers, to avoid using the same sample for training and validation. We trained the model on the training set, and used the validation set to compute the optimal classification threshold, which is the threshold that maximizes the balanced accuracy on the validation set. We then retrained the model

on the entire development set, and computed the performance and fairness metrics on the test set. Finally, we computed the mean and standard deviation of all metrics across the 5 folds.

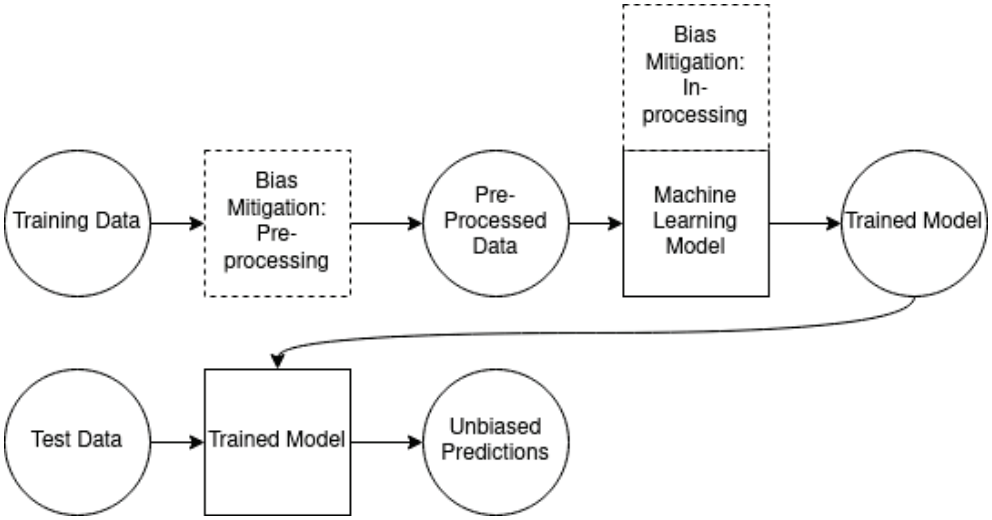

Figure 1. Workflow of data, machine learning models, and bias mitigation techniques used in this research.

The code used to generate the dataset and train the machine learning models is provided as a GitHub repository (https://github.com/PabloMosUU/FairnessForPsychiatry, accessed on 16 February 2022).

3. Results

Each of our classifiers output a continuous prediction for each test data point. We converted these to binary classifications by comparing with a classification threshold. Figures 2–7 show the trade-off between balanced accuracy and fairness metrics as a function of the classification threshold. Figures 2 and 3 show how the disparate impact error and average odds difference vary together with the balanced accuracy as a function of the classification threshold of a logistic regression model with no bias mitigation, for one of the folds of cross-validation. The corresponding plots for the random forest classifier show the same trends. The performance and fairness metrics after cross-validation are shown in Tables 4 and 5, respectively. Since we observed bias (see Section 4 for further discussion), we implemented the mitigation strategies detailed in Section 2.4. Figures 4 and 5 show the validation plots for a logistic regression classifier with reweighing for one of the folds of cross-validation; the plots for the random forest classifier show similar trends. Figures 6 and 7 show the validation plots for a logistic regression classifier with prejudice remover.

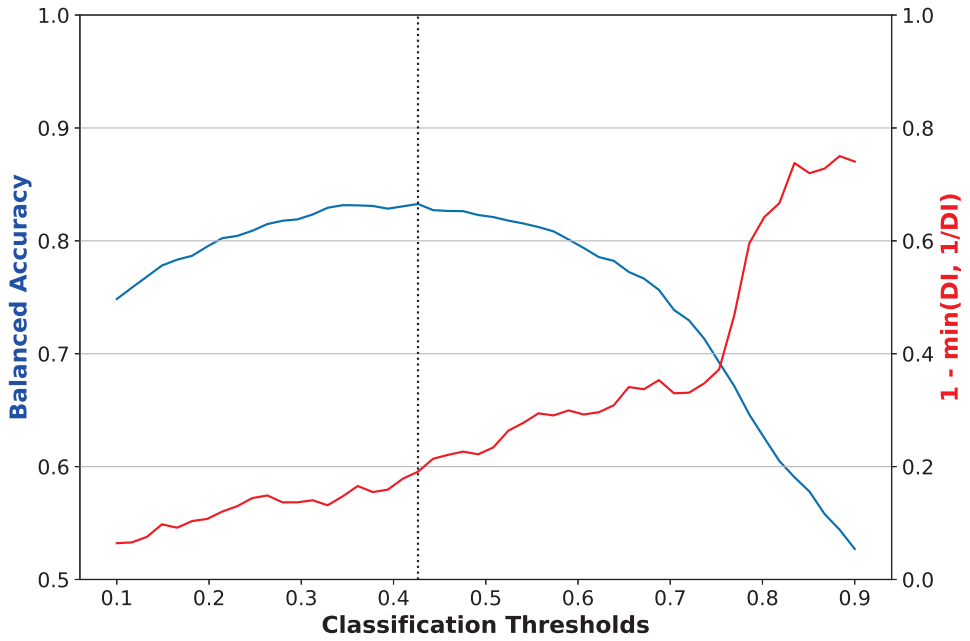

Figure 2. Balanced accuracy and disparate impact error versus classification threshold for a logistic regression classifier with no bias mitigation. The dotted vertical line is the threshold that maximizes balanced accuracy. The plot shown corresponds to one of the folds of cross-validation. Disparate impact error, equal to 1-`min`(DI, 1/DI), where DI is the disparate impact, is the difference between disparate impact and its ideal value of 1.

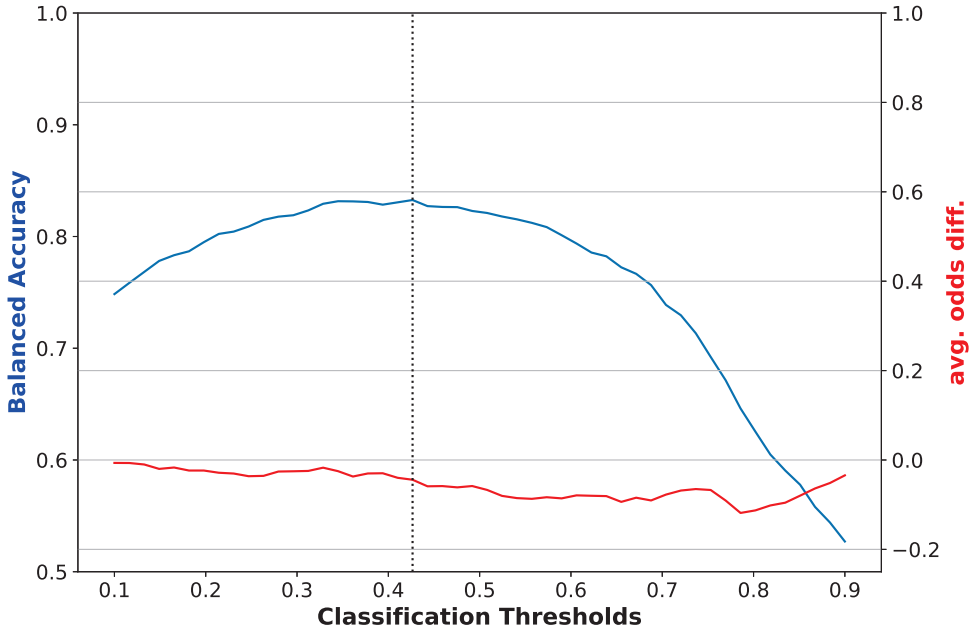

Figure 3. Balanced accuracy and average odds difference versus classification threshold for a logistic regression classifier with no bias mitigation. The dotted vertical line is the threshold that maximizes balanced accuracy. The plot shown corresponds to one of the folds of cross-validation.

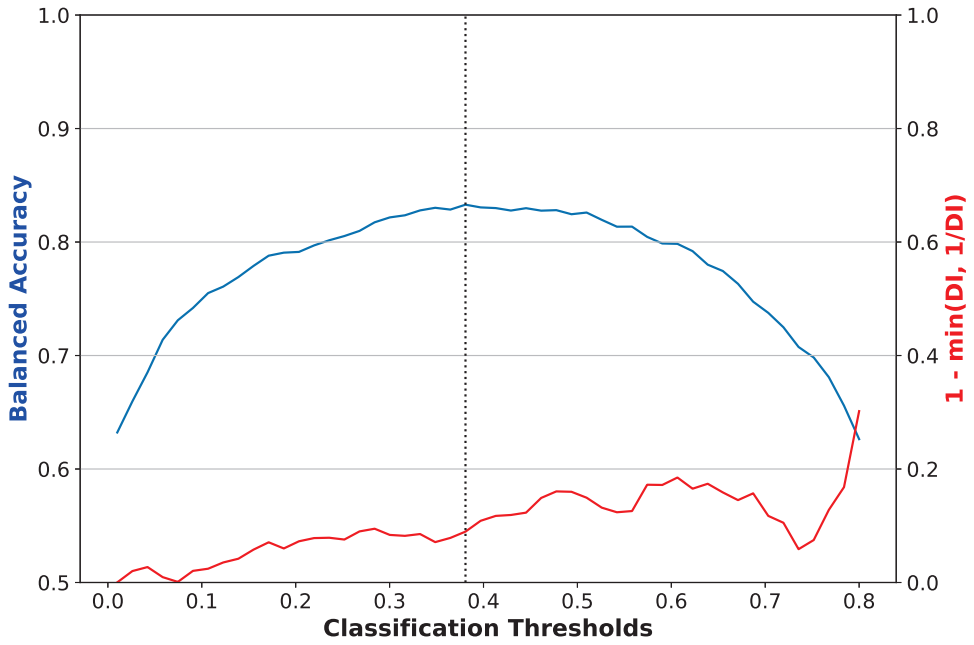

Figure 4. Balanced accuracy and disparate impact error versus classification threshold for a logistic regression classifier with reweighing. The dotted vertical line is the threshold that maximizes balanced accuracy. The plot shown corresponds to one of the folds of cross-validation. Disparate impact error, equal to 1-`min`(DI, 1/DI), where DI is the disparate impact, and the difference between disparate impact and its ideal value of 1.

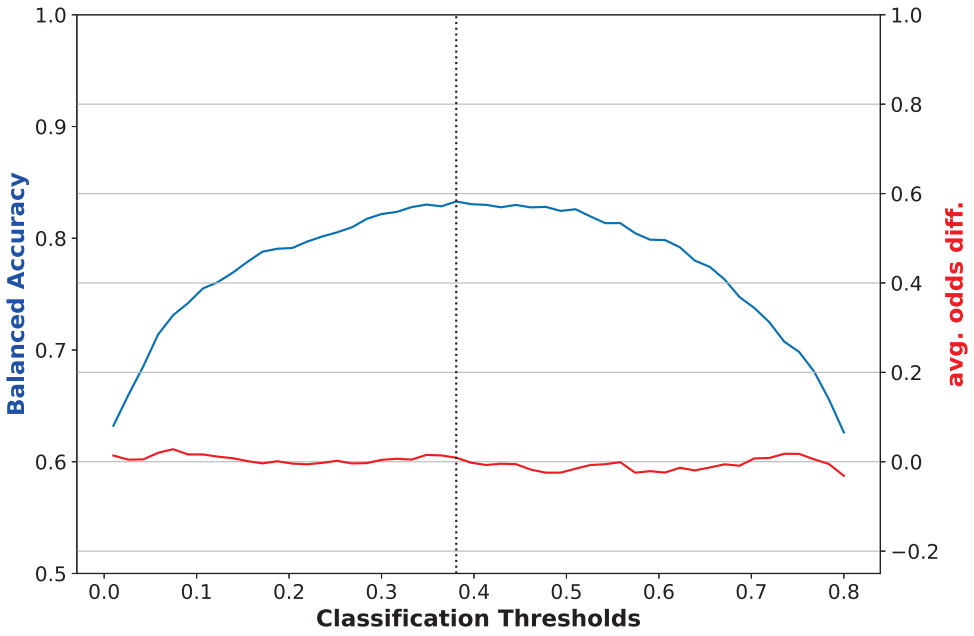

Figure 5. Balanced accuracy and average odds difference versus classification threshold for a logistic regression classifier with reweighing. The dotted vertical line is the threshold that maximizes balanced accuracy. The plot shown corresponds to one of the folds of cross-validation.

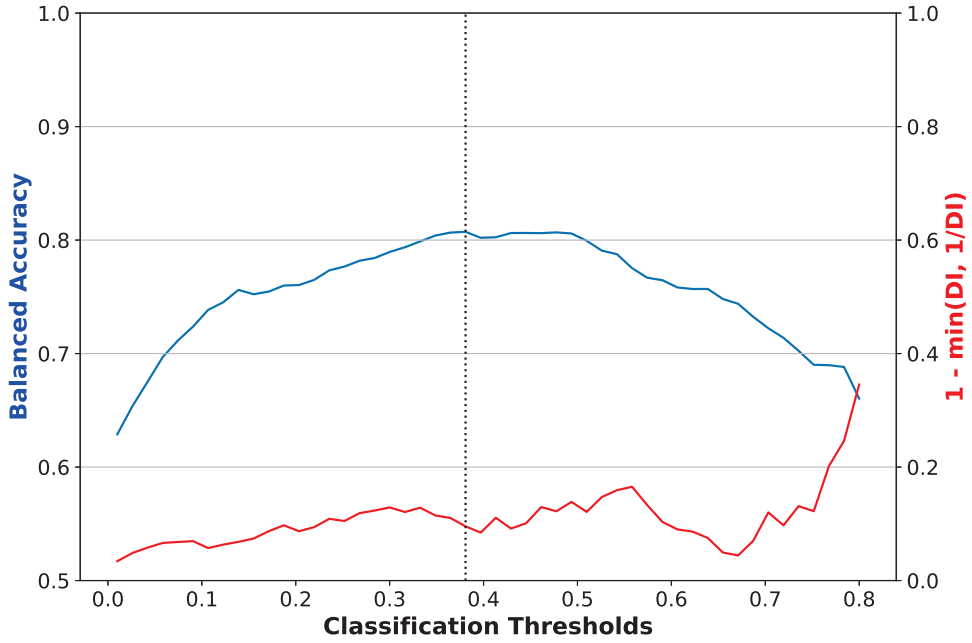

Figure 6. Balanced accuracy and disparate impact error versus classification threshold for a logistic regression classifier with prejudice remover. The dotted vertical line is the threshold that maximizes balanced accuracy. The plot shown corresponds to one of the folds of cross-validation. Disparate impact error, equal to 1-`min`(DI, 1/DI), where DI is the disparate impact, and the difference between disparate impact and its ideal value of 1.

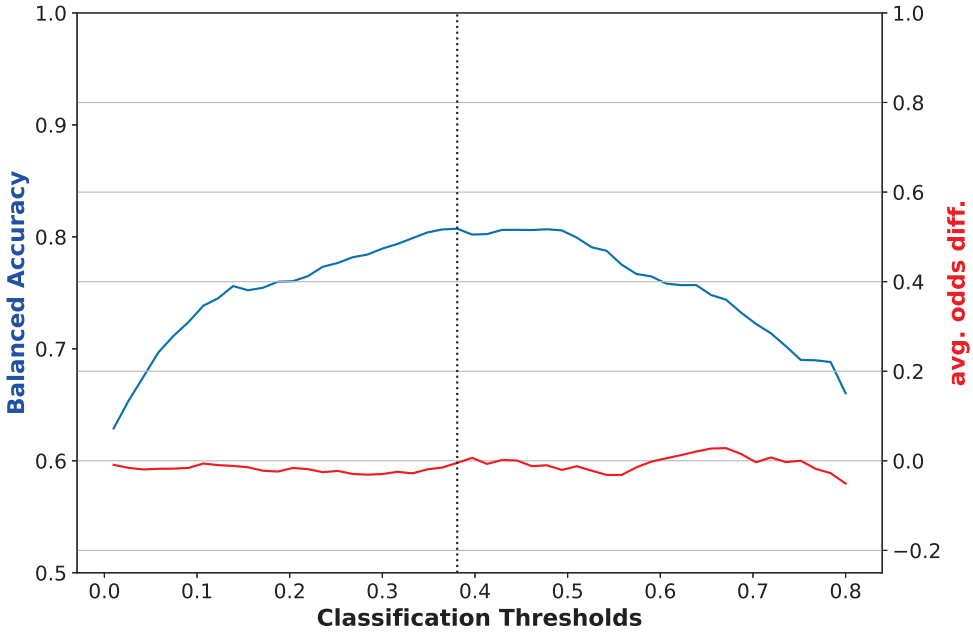

Figure 7. Balanced accuracy and average odds difference versus classification threshold for a logistic regression classifier with prejudice remover. The dotted vertical line is the threshold that maximizes balanced accuracy. The plot shown corresponds to one of the folds of cross-validation.

Table 4. Classification metrics for logistic regression (LR) and random forest (RF) classifiers including bias mitigation strategies reweighing (RW) and prejudice remover (PR). The classification metrics are balanced accuracy (Acc_{bal}) and F1 score. The errors shown are standard deviations.

Model		Performance	
Clf.	**Mit.**	**Acc_{bal}**	**F1**
LR		0.834 ± 0.015	0.843 ± 0.014
RF		0.843 ± 0.018	0.835 ± 0.020
LR	RW	0.830 ± 0.014	0.839 ± 0.011
RF	RW	0.847 ± 0.019	0.840 ± 0.020
LR	PR	0.793 ± 0.020	0.802 ± 0.029

Table 5. Fairness metrics for logistic regression (LR) and random forest (RF) classifiers including bias mitigation strategies reweighing (RW) and prejudice remover (PR). The fairness metrics are disparate impact (DI), average odds difference (AOD), statistical parity difference (SPD), and equal opportunity difference (EOD). The errors shown are standard deviations.

Model		Fairness			
Clf.	**Mit.**	**DI**	**AOD**	**SPD**	**EOD**
LR		0.793 ± 0.074	-0.046 ± 0.021	-0.110 ± 0.038	-0.038 ± 0.028
RF		0.796 ± 0.071	-0.018 ± 0.017	-0.083 ± 0.031	-0.013 ± 0.035
LR	RW	0.869 ± 0.066	-0.003 ± 0.013	-0.066 ± 0.035	0.004 ± 0.034
RF	RW	0.830 ± 0.077	-0.004 ± 0.023	-0.070 ± 0.034	0.001 ± 0.043
LR	PR	0.886 ± 0.056	-0.008 ± 0.003	-0.060 ± 0.034	-0.020 ± 0.045

4. Discussion

4.1. Analysis of Results

As reported in Table 5, all fairness metrics show results favourable to the privileged group (see Section 2.3 for a discussion of the fairness metrics we use). Reweighing improved the fairness metrics for both classifiers. The prejudice remover also improved the fairness metrics, albeit at a cost in performance. There was no big difference in performance between the logistic regression and random forest classifiers. If fairness is crucial, then the logistic regression classifier gives more options in terms of the mitigation strategies. The better mitigation strategy is the one closest to the data, for it requires less tinkering with the model, which can lead to worse explainability.

In addition, we computed, for each fold of cross-validation, the difference for each performance and fairness metric between a model with a bias mitigator and the corresponding model without bias mitigation. We then took the mean and standard deviation of those differences, and report the results for performance and fairness metrics on Tables 6 and 7, respectively. We can see that differences in performance for reweighing are mostly small, while the gains in fairness metrics are statistically significant at a 95% confidence level. Meanwhile, the prejudice remover incurs a greater cost in performance, with no apparent greater improvement to the fairness metrics.

Table 6. Classification metric differences of models with bias mitigators reweighing (RW) and prejudice remover (PR) compared to a baseline without bias mitigation, for logistic regression (LR) and random forest (RF) classifiers. The classification metrics are balanced accuracy (Acc_{bal}) and F1 score. The errors shown are standard deviations. Differences significant at 95% confidence level are shown in **bold**.

Model		Performance	
Clf.	**Mit.**	**ΔAcc_{bal}**	**ΔF1**
LR	PR	$\mathbf{-0.040 \pm 0.013}$	-0.041 ± 0.025
LR	RW	-0.003 ± 0.013	-0.005 ± 0.013
RF	RW	0.003 ± 0.002	$\mathbf{0.005 \pm 0.001}$

Table 7. Fairness metric differences of models with bias mitigators reweighing (RW) and prejudice remover (PR) compared to a baseline without bias mitigation, for logistic regression (LR) and random forest (RF) classifiers. The fairness metrics are disparate impact (DI), average odds difference (AOD), statistical parity difference (SPD) and equal opportunity difference (EOD). The errors shown are standard deviations. Differences significant at 95% confidence level are shown in **bold**.

Model		Fairness			
Clf.	**Mit.**	**ΔDI**	**ΔAOD**	**ΔSPD**	**ΔEOD**
LR	PR	$\mathbf{0.092 \pm 0.036}$	0.038 ± 0.021	$\mathbf{0.050 \pm 0.019}$	0.018 ± 0.042
LR	RW	$\mathbf{0.075 \pm 0.021}$	$\mathbf{0.043 \pm 0.017}$	$\mathbf{0.043 \pm 0.014}$	0.042 ± 0.034
RF	RW	$\mathbf{0.034 \pm 0.013}$	$\mathbf{0.014 \pm 0.006}$	$\mathbf{0.013 \pm 0.006}$	0.014 ± 0.011

4.2. Limitations

Some diagnoses did not have a diagnosis date filled out in the raw dataset. In those cases, we used the treatment end date. Some data points did not have a value for that variable either, and in those cases, we used the treatment start date. This leads to an inconsistent definition of the diagnosis date, and hence to inconsistencies in the variables related to diagnoses during the first 14 days of admission. However, we carried out the analysis again with only the diagnoses for which the diagnosis dates were present in the raw data, and the results followed the same trends.

On a similar note, we removed a few medication administrations that did not have an administering date. A better solution would have been to remove all data corresponding to those patients, albeit at the cost of having fewer data points. We carried out the analysis again in that configuration, and obtained similar results.

Finally, this work considered only the diagnoses that took place within the first 14 days of admission. It might have been interesting to also consider diagnoses that took place *before* admission. We leave this option for future work.

4.3. Future Work

The present work considered benzodiazepine prescriptions administered during the remainder of each patient's admission. To make the prediction task fairer for the computer, we could consider predicting benzodiazepines administered during a specific time window, for example, days 15–28 of an admission.

Previous work noted a possible bias between the gender of the *prescriber* and the prescriptions of benzodiazepines [16,17]. It would be interesting to look into this correlation in our dataset as well; one could train a model to predict, on the basis of patient and prescriber data, whether benzodiazepines will be prescribed. If there are correlations between the gender of the prescriber and the prescription of benzodiazepines, we could raise a warning to let the practitioner know that the model thinks there might be a bias.

Finally, there are other medications for which experts suspect there could be gender biases in the prescriptions and administrations, such as antipsychotics and antidepressives. It would be beneficial to also study those administrations using a similar pipeline as the one developed here.

As a final note, [51] warned against the use of blind applications of fairness frameworks in healthcare. Thus, the present study should be considered only as a demonstration of the importance of considering bias and mitigation in clinical psychiatry machine learning models. Further work is necessary to understand these biases on a deeper level, and what course of action should be taken.

5. Conclusions

Given our results (Section 3) and discussion thereof (Section 4.1), we can conclude that a model trained to predict future administrations of benzodiazepines based on past data is biased by the patients' genders. Perhaps surprisingly, reweighing the data (a pre-processing step) seems to mitigate this bias quite significantly, without loss of performance. The in-processing method with a prejudice remover also mitigated this bias, but at a cost to performance.

This is the first fairness evaluation of a machine learning model trained on real clinical psychiatric data. Future researchers working with such models should consider computing fairness metrics and, when necessary, adopt mitigation strategies to ensure patient treatment is not biased with respect to protected attributes.

Author Contributions: Conceptualization, F.S. and M.S.; methodology, P.M. and M.S.; software, P.M and J.K.; validation, P.M.; formal analysis, P.M.; investigation, J.K. and F.S.; resources, F.S.; data curation, P.M., J.K. and F.S.; writing—original draft preparation, P.M.; writing—review and editing, P.M.; visualization, J.K.; supervision, P.M. and J.M.; project administration, F.S. and M.S.; funding acquisition, F.S. and M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the COVIDA project, which in turn is funded by the Strategic Alliance TU/E, WUR, UU en UMC Utrecht.

Institutional Review Board Statement: The study was approved by the UMC ethics committee as part of PsyData, a team of data scientists and clinicians working at the psychiatry department of the UMC Utrecht.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated for this study cannot be shared, to protect patient privacy and comply with institutional regulations.

Acknowledgments: The core content of this study is drawn from the Master in Business Informatics thesis of Jesse Kuiper [52].

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Pestian, J.; Nasrallah, H.; Matykiewicz, P.; Bennett, A.; Leenaars, A. Suicide Note Classification Using Natural Language Processing: A Content Analysis. *Biomed. Inform. Insights* **2010**, *3*, BII.S4706. [CrossRef] [PubMed]
2. Menger, V.; Spruit, M.; van Est, R.; Nap, E.; Scheepers, F. Machine Learning Approach to Inpatient Violence Risk Assessment Using Routinely Collected Clinical Notes in Electronic Health Records. *JAMA Netw. Open* **2019**, *2*, e196709. [CrossRef] [PubMed]
3. Le, D.V.; Montgomery, J.; Kirkby, K.C.; Scanlan, J. Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting. *J. Biomed. Inform.* **2018**, *86*, 49–58. [CrossRef] [PubMed]
4. Suchting, R.; Green, C.E.; Glazier, S.M.; Lane, S.D. A data science approach to predicting patient aggressive events in a psychiatric hospital. *Psychiatry Res.* **2018**, *268*, 217–222. [CrossRef] [PubMed]
5. van Mens, K.; de Schepper, C.; Wijnen, B.; Koldijk, S.J.; Schnack, H.; de Looff, P.; Lokkerbol, J.; Wetherall, K.; Cleare, S.; O'Connor, R.C.; et al. Predicting future suicidal behaviour in young adults, with different machine learning techniques: A population-based longitudinal study. *J. Affect. Disord.* **2020**, *271*, 169–177. [CrossRef] [PubMed]
6. Kalidas, V. Siamese Fine-Tuning of BERT for Classification of Small and Imbalanced Datasets, Applied to Prediction of Involuntary Admissions in Mental Healthcare. Master's Thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 2020.

7. Delgado-Rodriguez, M.; Llorca, J. Bias. *J. Epidemiol. Community Health* **2004**, *58*, 635–641. [CrossRef]
8. Kim, H.B.; Myung, S.K.; Park, Y.C.; Park, B. Use of benzodiazepine and risk of cancer: A meta-analysis of observational studies. *Int. J. Cancer* **2017**, *140*, 513–525. [CrossRef]
9. Quaglio, G.; Pattaro, C.; Gerra, G.; Mathewson, S.; Verbanck, P.; Des Jarlais, D.C.; Lugoboni, F. High dose benzodiazepine dependence: Description of 29 patients treated with flumazenil infusion and stabilised with clonazepam. *Psychiatry Res.* **2012**, *198*, 457–462. [CrossRef]
10. Federatie Medisch Specialisten. Angststoornissen. Available online: https://richtlijnendatabase.nl/richtlijn/angststoornissen/gegeneraliseerde_angststoornis_gas/farmacotherapie_bij_gas/benzodiazepine_gegeneraliseerde_angststoornis.html (accessed on 18 November 2021).
11. Vinkers, C.H.; Tijdink, J.K.; Luykx, J.J.; Vis, R. Kiezen voor de juiste benzodiazepine. *Ned. Tijdschr. Geneeskd.* **2012**, *156*, A4900.
12. Bjorner, T.; Laerum, E. Factors associated with high prescribing of benzodiazepines and minor opiates. *Scand. J. Prim. Health Care* **2003**, *21*, 115–120. [CrossRef]
13. Peters, S.M.; Knauf, K.Q.; Derbidge, C.M.; Kimmel, R.; Vannoy, S. Demographic and clinical factors associated with benzodiazepine prescription at discharge from psychiatric inpatient treatment. *Gen. Hosp. Psychiatry* **2015**, *37*, 595–600. [CrossRef] [PubMed]
14. Cook, B.; Creedon, T.; Wang, Y.; Lu, C.; Carson, N.; Jules, P.; Lee, E.; Alegría, M. Examining racial/ethnic differences in patterns of benzodiazepine prescription and misuse. *Drug Alcohol Depend.* **2018**, *187*, 29–34. [CrossRef] [PubMed]
15. Olfson, M.; King, M.; Schoenbaum, M. Benzodiazepine Use in the United States. *JAMA Psychiatry* **2015**, *72*, 136–142. [CrossRef] [PubMed]
16. McIntyre, R.S.; Chen, V.C.H.; Lee, Y.; Lui, L.M.W.; Majeed, A.; Subramaniapillai, M.; Mansur, R.B.; Rosenblat, J.D.; Yang, Y.H.; Chen, Y.L. The influence of prescriber and patient gender on the prescription of benzodiazepines: Evidence for stereotypes and biases? *Soc. Psychiatry Psychiatr. Epidemiol.* **2021**, *56*, 1433–9285. [CrossRef] [PubMed]
17. Lui, L.M.W.; Lee, Y.; Lipsitz, O.; Rodrigues, N.B.; Gill, H.; Ma, J.; Wilkialis, L.; Tamura, J.K.; Siegel, A.; Chen-Li, D.; et al. The influence of prescriber and patient gender on the prescription of benzodiazepines: Results from the Florida Medicaid Dataset. *CNS Spectrums* **2021**, *26*, 1–5. [CrossRef] [PubMed]
18. Maric, N.P.; Latas, M.; Andric Petrovic, S.; Soldatovic, I.; Arsova, S.; Crnkovic, D.; Gugleta, D.; Ivezic, A.; Janjic, V.; Karlovic, D.; et al. Prescribing practices in Southeastern Europe—Focus on benzodiazepine prescription at discharge from nine university psychiatric hospitals. *Psychiatry Res.* **2017**, *258*, 59–65. [CrossRef]
19. Bellamy, R.K.E.; Dey, K.; Hind, M.; Hoffman, S.C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilović, A.; et al. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Res. Dev.* **2019**, *63*, 4:1–4:15. [CrossRef]
20. Baer, T. *Understand, Manage, and Prevent Algorithmic Bias*; Apress: Berkeley, CA, USA, 2019.
21. Ellenberg, J.H. Selection bias in observational and experimental studies. *Stat. Med.* **1994**, *13*, 557–567. [CrossRef]
22. Barocas, S.; Selbst, A. Big Data's Disparate Impact. *Calif. Law Rev.* **2016**, *104*, 671. [CrossRef]
23. d'Alessandro, B.; O'Neil, C.; LaGatta, T. Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification. *Big Data* **2017**, *5*, 120–134. [CrossRef]
24. Lang, W.W.; Nakamura, L.I. A Model of Redlining. *J. Urban Econ.* **1993**, *33*, 223–234. [CrossRef]
25. Chouldechova, A.; Roth, A. A Snapshot of the Frontiers of Fairness in Machine Learning. *Commun. ACM* **2020**, *63*, 82–89. [CrossRef]
26. Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; Zemel, R. Fairness through Awareness. Available online: https://arxiv.org/abs/1104.3913 (accessed on 18 November 2021).
27. Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; Dwork, C. Learning Fair Representations. In Proceedings of the 30th International Conference on Machine Learning, PMLR, Atlanta, GA, USA, 17–19 June 2013; Volume 28, pp. 325–333.
28. Joseph, M.; Kearns, M.; Morgenstern, J.; Roth, A. Fairness in Learning: Classic and Contextual Bandits. Available online: https://arxiv.org/abs/1605.07139 (accessed on 18 November 2021).
29. Friedler, S.A.; Scheidegger, C.; Venkatasubramanian, S. On the (Im)Possibility of Fairness. Available online: https://arxiv.org/abs/1609.07236 (accessed on 18 November 2021).
30. Saleiro, P.; Kuester, B.; Hinkson, L.; London, J.; Stevens, A.; Anisfeld, A.; Rodolfa, K.T.; Ghani, R. Aequitas: A Bias and Fairness Audit Toolkit. Available online: https://arxiv.org/abs/1811.05577 (accessed on 18 November 2021).
31. Feldman, M.; Friedler, S.; Moeller, J.; Scheidegger, C.; Venkatasubramanian, S. Certifying and Removing Disparate Impact. Available online: https://arxiv.org/abs/1412.3756 (accessed on 18 November 2021).
32. Kamiran, F.; Calders, T. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* **2012**, *33*, 1–33. [CrossRef]
33. Kamishima, T.; Akaho, S.; Asoh, H.; Sakuma, J. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Machine Learning and Knowledge Discovery in Databases*; Flach, P.A., De Bie, T., Cristianini, N., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 35–50.
34. Scheuerman, M.K.; Wade, K.; Lustig, C.; Brubaker, J.R. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proc. ACM Hum.-Comput. Interact.* **2020**, *4*, 1–35. [CrossRef]

35. Xu, T.; White, J.; Kalkan, S.; Gunes, H. Investigating Bias and Fairness in Facial Expression Recognition. In Proceedings of the Computer Vision—ECCV 2020 Workshops, Glasgow, UK, 23–28 August 2020; Bartoli, A., Fusiello, A., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 506–523.
36. Yucer, S.; Akcay, S.; Al-Moubayed, N.; Breckon, T.P. Exploring Racial Bias Within Face Recognition via Per-Subject Adversarially-Enabled Data Augmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Seattle, DC, USA, 14–19 June 2020; pp. 18–19.
37. Liu, H.; Dacon, J.; Fan, W.; Liu, H.; Liu, Z.; Tang, J. Does Gender Matter? Towards Fairness in Dialogue Systems. Available online: https://arxiv.org/abs/1910.10486 (accessed on 18 November 2021).
38. Kizilcec, R.F.; Lee, H. Algorithmic Fairness in Education. Available online: https://arxiv.org/abs/2007.05443 (accessed on 18 November 2021).
39. Geneviève, L.D.; Martani, A.; Shaw, D.; Elger, B.S.; Wangmo, T. Structural racism in precision medicine: Leaving no one behind. *BMC Med. Ethics* **2020**, *21*, 17. [CrossRef]
40. Tripathi, S.; Fritz, B.A.; Abdelhack, M.; Avidan, M.S.; Chen, Y.; King, C.R. (Un)Fairness in Post-Operative Complication Prediction Models. Available online: https://arxiv.org/abs/2011.02036 (accessed on 18 November 2021).
41. Singh, H.; Mhasawade, V.; Chunara, R. Generalizability Challenges of Mortality Risk Prediction Models: A Retrospective Analysis on a Multi-center Database. *medRxiv* **2021**. [CrossRef]
42. Amir, S.; van de Meent, J.W.; Wallace, B.C. On the Impact of Random Seeds on the Fairness of Clinical Classifiers. Available online: https://arxiv.org/abs/2104.06338 (accessed on 18 November 2021).
43. Jasuja, G.K.; Reisman, J.I.; Weiner, R.S.; Christopher, M.L.; Rose, A.J. Gender differences in prescribing of zolpidem in the Veterans Health Administration. *Am. J. Manag. Care* **2019**, *25*, e58–e65. [CrossRef]
44. Nam, S.K.; Chu, H.J.; Lee, M.K.; Lee, J.H.; Kim, N.; Lee, S.M. A Meta-analysis of Gender Differences in Attitudes Toward Seeking Professional Psychological Help. *J. Am. Coll. Health* **2010**, *59*, 110–116. [CrossRef]
45. Strakowski, S.M.; McElroy, S.L.; Keck, P.E.; West, S.A. Racial influence on diagnosis in psychotic mania. *J. Affect. Disord.* **1996**, *39*, 157–162. [CrossRef]
46. Rumshisky, A.; Ghassemi, M.; Naumann, T.; Szolovits, P.; Castro, V.M.; McCoy, T.H.; Perlis, R.H. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Transl. Psychiatry* **2016**, *6*, e921. [CrossRef]
47. Tang, S.X.; Kriz, R.; Cho, S.; Park, S.J.; Harowitz, J.; Gur, R.E.; Bhati, M.T.; Wolf, D.H.; Sedoc, J.; Liberman, M.Y. Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. *NPJ Schizophr.* **2021**, *7*, 25. [CrossRef] [PubMed]
48. Kaczmarek-Majer, K.; Casalino, G.; Castellano, G.; Hryniewicz, O.; Dominiak, M. Explaining smartphone-based acoustic data in bipolar disorder: Semi-supervised fuzzy clustering and relative linguistic summaries. *Inf. Sci.* **2022**, *588*, 174–195. [CrossRef]
49. Nederlands Huisartsen Genootschap. Omrekentabel Benzodiazepine naar Diazepam 2 mg Tabletten. 2014. Available online: https://www.nhg.org/sites/default/files/content/nhg_org/images/thema/omrekentabel_benzodiaz._naar_diazepam_2_mg_tab.pdf (accessed on 22 March 2022).
50. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
51. Pfohl, S.R.; Foryciarz, A.; Shah, N.H. An empirical characterization of fair machine learning for clinical risk prediction. *J. Biomed. Inform.* **2021**, *113*, 103621. [CrossRef] [PubMed]
52. Kuiper, J. Machine-Learning Based Bias Discovery in Medical Data. Master's Thesis, Utrecht University, Utrecht, The Netherlands, 2021.