Fairness & Ethics In Hospital Re-Admission Prediction

Anders Hjulmand (ahju@itu.dk), Eisuke Okuda (eiok@itu.dk) & Andreas Olsen (frao@itu.dk)
IT University of Copenhagen, May 21, 2024

1 Introduction

Many facets of society are becoming increasingly reliant on big data and machine learning algorithms to aid humans in decision making processes. This is evident in a range of sectors such as employment, crime-detection, and education, where the outcome of the algorithms have a tangible impact on peoples' lives. A vital concern is therefore to implement fair and ethical algorithms that do not discriminate groups of people or individuals [1–3].

The healthcare sector is also seeing a rise in the use of algorithms, spanning from drug discovery, clinical trial assistance and disease classification [4, 5]. Ensuring fairness and accountability in healthcare algorithms is essential, as they directly affect sick patients and can have significant implications. Despite the heightened risks, existing healthcare algorithms have been shown to exhibit biases, for example in the identification of high-risk patients [6].

In this paper, we examine the fairness of an algorithm that predicts hospital readmission for diabetic patients. In particular we examine whether the algorithm exhibits discrimination based on age, race, and gender, and if so, how to mitigate the discrimination. The algorithm can help healthcare personnel identify patients likely to be readmitted after they have been discharged from the hospital. These patients may benefit from preventative measures, such as additional monitoring through follow-up phone calls to check for specific symptoms. The algorithm might detect patterns that healthcare personnel miss, and could thereby augment their judgements. However, it should only be used in conjunction with healthcare personnel expertise, with its predictions carefully weighted against experienced human judgment. The code can be found on github.

2 Data

The *Diabetes 130-US Hospitals for Years 1999-2008* dataset [7] used in this study constitutes 10 years of clinical care at 130 hospitals in the United States. It comes with 101.766 rows, where each row represents a diabetic patient's admission to the hospital. A hospital stay can last up to 14 days and may include laboratory tests and prescriptions of medicine. The dataset contains 50 variables related to the patient and their stay.

2.1 Preprocessing

One patient may have multiple admissions in the raw data. To treat each patient as an i.i.d variable we only selected the first hospital admission of each patient. To achieve more consistent entries, the dataset was filtered to only include ADMISSION_TYPE of *Emergency, Urgent* or *Elective*. Similarly, only RACE of *Caucasian, AfricanAmerican, Hispanic* and *Asian* were kept, thereby removing missing and unknown races to obtain a dichotomous variable. These filtering steps resulted in 60.571 patients.

To increase the information about each patient's admission, we transformed and aggregated some existing variables into new ones. For example, a patient's diabetic medicine prescriptions were enlisted as 23 sparse variables of 4 categories *Up, Down, No, Steady*. These were aggregated into 3 variables indicating the number of diabetic prescriptions, whether there was a change in medicine and whether there was a change in prescribed dosage. We then transformed the 23 medicine prescription variables into binary, indicating whether a specific medicine was taken at the time of admission. Other flag variables were included to indicate whether blood glucose and A1C tests were conducted during the admission. Similarly, the information on a patient's hospital admissions within the previous year was aggregated into a binary flag variable, to indicate whether they have been admitted or not.

The unique payer codes (23) was mapped to 4 binary variables of BLUE_CROSS, MEDICARE, MEDICAID, and SELF_PAY as these were the most prevalent categories disregarding *unknown*. How a patient was discharged from the hospital was aggregated into HOME, TRANSFER, UNKNOWN, and OTHER, where OTHER comprises *"Expired"* and *"Expired at home. Medicaid only, hospice."*. The age of patients were binned from intervals of 10 years into three age groups [0-30), [30-60), and [60-100). Readmission of < 30 days and > 30 days was aggregated into a single category and used as a binary target variable where $f(x_i) = \begin{cases} 1 & \text{if patient } x_i \text{ was readmitted} \\ 0 & \text{Otherwise} \end{cases}$

The categorical features DISCHARGE_DISPOSITION_ID and ADMISSION_TYPE were one-hot encoded without dropping the first category. The data was split into train (80%) and test (20%) sets. For more details on feature descriptions, see Appendix Table 1.

2.2 Protected features and Inherent biases

We consider AGE, RACE, and GENDER to be the protected features of each patient. Figure 1 shows the number of patients in each protected group (top) and their re-admission rates (bottom). There is a large difference in the number of patients across AGE and RACE. In particular, age *[60-100)* and *Caucasian* race are over-represented in the data, which illustrates a representation bias [8]. In contrast, *Asians* and *Hispanics* constitute less than 2% of the patients. This may have consequences for the algorithm's generalisability towards minority groups.

Another bias related to Figure 1 (top) is the population bias [8]. The proportion of males and females in the United States in 1999 was 49% and 51% out of which 8.1% of males were diagnosed with diabetes compared to 7.3% females [9], [10]. This distribution is not reflected in the dataset as it comprises more female than male diabetic patients, meaning that males are underrepresented compared to the true population.

Figure 1 (bottom) shows a discrepancy in the rates of re-admissions across protected groups. It is important to note that re-admission rate is not necessarily a proxy for health care needs. If a certain group has a higher re-admission rate, that does not necessarily mean that they are more sick than the other groups. It may instead be due to an omitted variable bias, since we do not know whether the patients have different access to insurances, or whether some groups are more observant of their symptoms [8].

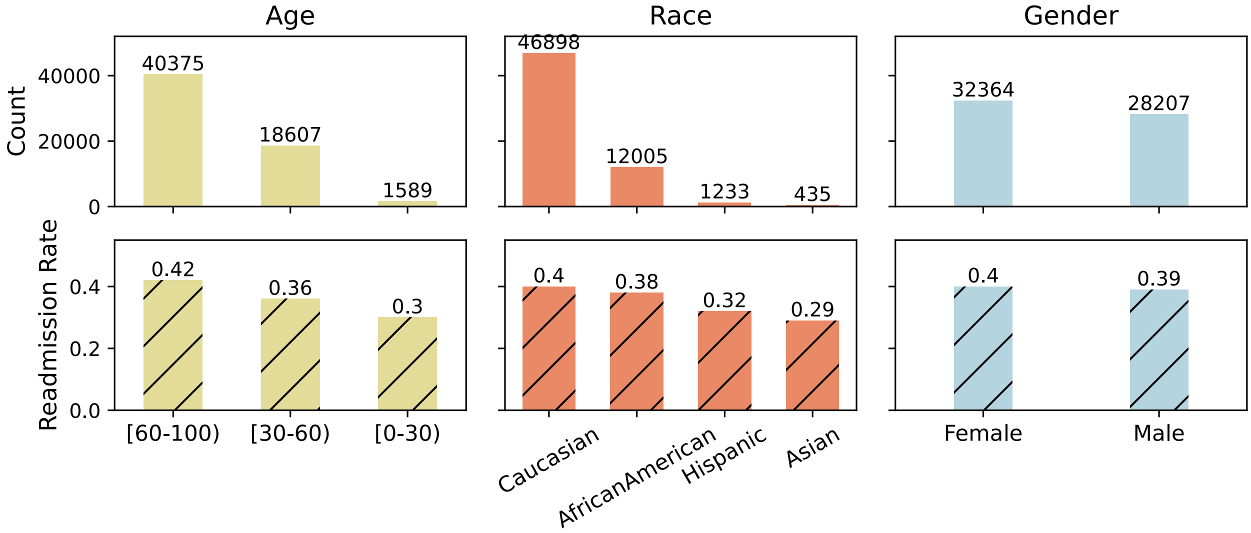

Figure 1: Number of patients (top), and readmission-rates (bottom) in the protected groups for the original data of 60.571 observations.

3 Methods

3.1 Fairness evaluation

The aim of our algorithm is to identify whether a diabetic patient is likely to be re-admitted to the hospital after being discharged. In this context, a useful notion of fairness is to address whether the algorithm's ability to predict readmissions is dependent on the protected features AGE, RACE and GENDER. Thus, we quantify our model's fairness using equalised odds as metric comprising of the true positive rate (TPR) and the false positive rate (FPR). Our goal is to ensure equalised odds i.e. equal TPRs and FPRs for all categories within the protected groups.

TPR, measures the likelihood that a patient is predicted to be re-admitted given that they are actually re-admitted. A fair algorithm should have the same TPR for all protected groups, ensuring that preventative measures such as follow-up phone calls are distributed fairly. Equal TPR for all groups is defined as: $\forall\, g \in G, \quad p(S = 1 \mid T = 1) = k$, where g is the category of the protected group (e.g. male or female), $S = 1$ indicates that the model predicts the patient to be re-readmitted, $T = 1$ indicates that the patient was actually re-admitted, and k is a constant.

FPR measures the likelihood that a patient who is not re-admitted is predicted to be re-admitted. From the patient's perspective, being falsely flagged poses little harm. In fact, additional care and attention from healthcare personnel might be beneficial even if the patient is not re-admitted. Therefore, a fair algorithm should have an equal FPR across protected groups to ensure that no category receives disproportionately excess monitoring, which is defined as: $\forall\, g \in G, \quad p(S = 0 \mid T = 1) = c$.

From the patient's perspective, an algorithm with high TPR is more important than a low FPR, as it is better to intervene unnecessarily on some patients, than to miss intervening on a patient who needed it.

3.2 Data debiasing

We use the technique described by He et al. [11] which aims to remove Pearson correlations between features and the protected features. Correlations, which otherwise can be picked up by an algorithm and result in features acting as proxy variables for the protected features. The features and protected features are expressed as vectors. The goal of the method is to construct a new representation r_j of a feature x_j which is uncorrelated with the protected feature vectors while still having high correlation to the original vector x_j. The decorrelated vector r_j is calculated by projecting the original vector x_j onto the orthonormal space of the protected vectors p: $r_j = x_j - \sum_i^p (x_j \cdot p_i) p_i$.

This constrains r_j to be strictly orthogonal to the protected feature vectors and will likely lower the model performance [12]. A parameter $\lambda \in [0, 1]$ is used to control the degree of constrain resulting in the trade-off between fairness and accuracy. Max constrain is at $\lambda = 0$ corresponding to $r'_j(\lambda) = r_j$ (high fairness) and $\lambda = 1$ produces $r'_j(\lambda) = x_j$ (low fairness): $r'_j(\lambda) = r_j + \lambda \cdot (x_j - r_j)$.

3.3 Model implementation

Ensuring fairness and transparency in healthcare algorithms is crucial due to their direct impact on patients' lives. Given this need, we opted for using a logistic regression model, making it possible to examine why a particular patient was predicted to be re-admitted. The model was trained with L2-regularization to increase robustness to the test-set. Before training, a grid-search resulted in an inverse regularization strength of 1 as the optimal hyperparameter for the L2-regularization across a 5-fold cross validation.

Before training the model, a Z-score standardisation was fitted on the training data and applied to both training and test data. In the training data there was a class imbalance where the majority class comprised of 29.277 READMITTED_FLAG=0 and the minority class had 19.179. The majority class 0 was down-sampled in the training data, such that readmissions occurred at an equal rate.

We first trained a model on the non-debiased data as a baseline model. In this data there exists Pearson correlations between features and protected groups, indicating the existence of proxy variables. Most notably, MEDICARE and NUMBER_DIAGNOSES were correlated with AGE (see figure 6). We then applied the decorrelation method on the dataset for 31 linearly spaced $\lambda \in [0, 1]$, and trained a model on each of these differently constrained representations of the data set.

We evaluated the baseline and debiased models by their equalised odds and overall accuracy. The models were compared across different values of λ to explore the fairness-accuracy trade-off. To measure the fairness within each protected group we calculated the **RMSE** for the TPRs and FPRs, thereby quantifying each category's deviation from the group mean.

4 Results

4.1 Overall model performance

The logistic regression model was unable to pick up most of the signals to classify the readmission of a patient correctly. This bad fit to the data is reflected in the overall accuracy score of 0.59 for the baseline model, barely better than a random learner.

Another indicator of this suboptimal fit is seen in Figure 2 and Appendix Figure 8 which show the TPRs for each protected category for each value of $\lambda \in [0, 1)$. From the first row of Figure 2 at $\lambda = 1$, it is clear that the max TPR of the baseline model across categories is AGE: [60-100) with a TPR of 0.56 contrary to its paired FPR of 0.41. This means that the model almost randomly predicted patients to be readmitted to the hospital. The category with the lowest performing TPR was 0.39 in the youngest age group. The protected feature AGE had the largest discrepancies among TPRs and FPRs for each category compared to RACE and GENDER, meaning that age subgroups had the most striking unequal odds with the baseline model.

4.2 Debiasing results

The aim of debiasing the data was to ensure equal TPRs and FPRs across protected groups. Figure 3 (bottom) shows the mean RMSE across each classification threshold of the ROC curve for each protected group before and after debiasing the data. The RMSE was generally smaller when using the most debiased representation of the data at $\lambda = 0$ across all protected groups. Especially AGE had a notable reduction in RMSE of more than half for the TPR, and more than a third for the FPR. The figure also shows that the TPRs and FPRs for GENDER had much smaller standard deviations compared to the other groups. This variation in TPRs and FPRs among categories is shown with

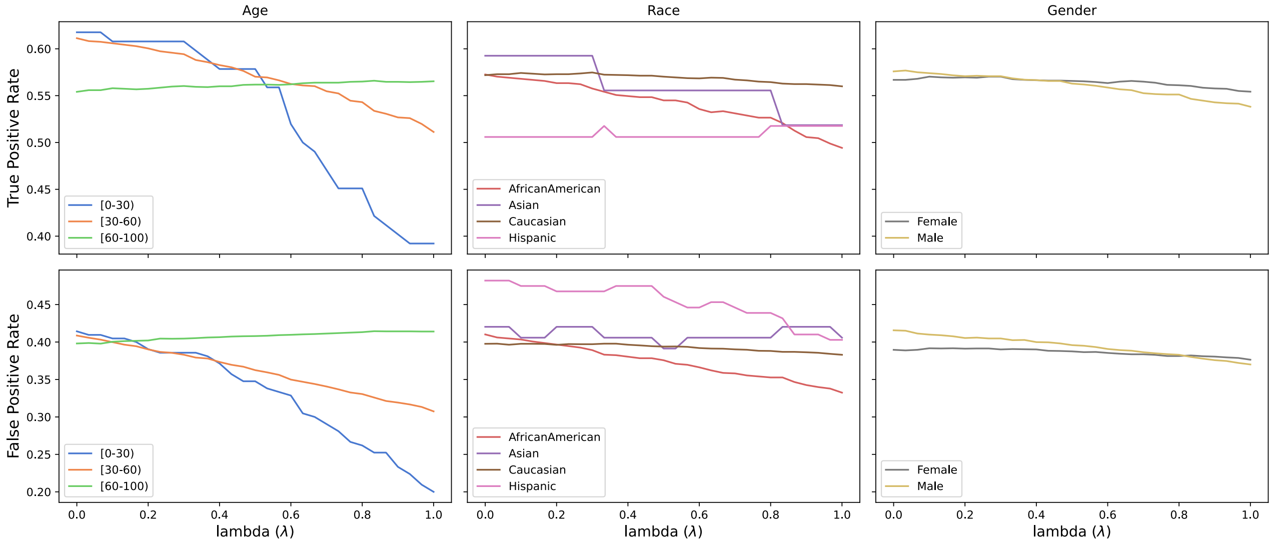

Figure 2: TPR (top) and FPR (bottom) for all protected groups with varying λ from high (0) to low (1) fairness.

higher granularity by the ROC curves in Figure 3 (top). Here it is clear that the variation among the curves for each protected group decreases after debiasing the data. In spite of this, the model still displays variance between TPRs and FPRs for some categories. This is particularly evident for patients of RACE *Asian* and *Hispanic* and AGE [0-30). Overall, Figure 3 displays success of equalizing the odds among categories when debiasing the data.

This increase in model fairness can also be seen by the pareto curve in Figure 4, that show the accuracy fairness trade-off. Fairness is calculated by the mean RMSE across each λ of each protected group for TPR and FPR. The overall accuracy only drops slightly when increasing the fairness. Notably, the most fair models can be found in the region of $\lambda \in [0, 0.5)$ and specifically we deem the best fairness accuracy trade-off to be at $\lambda = 0.33$.

4.3 Feature importance

The coefficients for all features were extracted from both the baseline and the debiased model. The change in relative effects of the different features are seen in Figure 5. It is clear, that the effects of many features have been influenced by debiasing the data. The figure shows that some effects have gained an absolute positive (green) increase whilst others have decreased (red). For example MEDICARE which had an increased negative effect and similarly NUM_LAB_PROCEDURES which had an increased positive effect. The largest difference was NUM_MEDICATIONS that grew 600% compared to the baseline estimate. The largest positive coefficient in the debiased model was PREV_YEAR_HOSPITAL which had an effect of 0.65 as predictor. This means that if a patient had been admitted within the past year, the odds of getting readmitted increased by $\exp^{0.65} = 1.91$ (See Table 2 in Appendix for a numerical summary of coefficients).

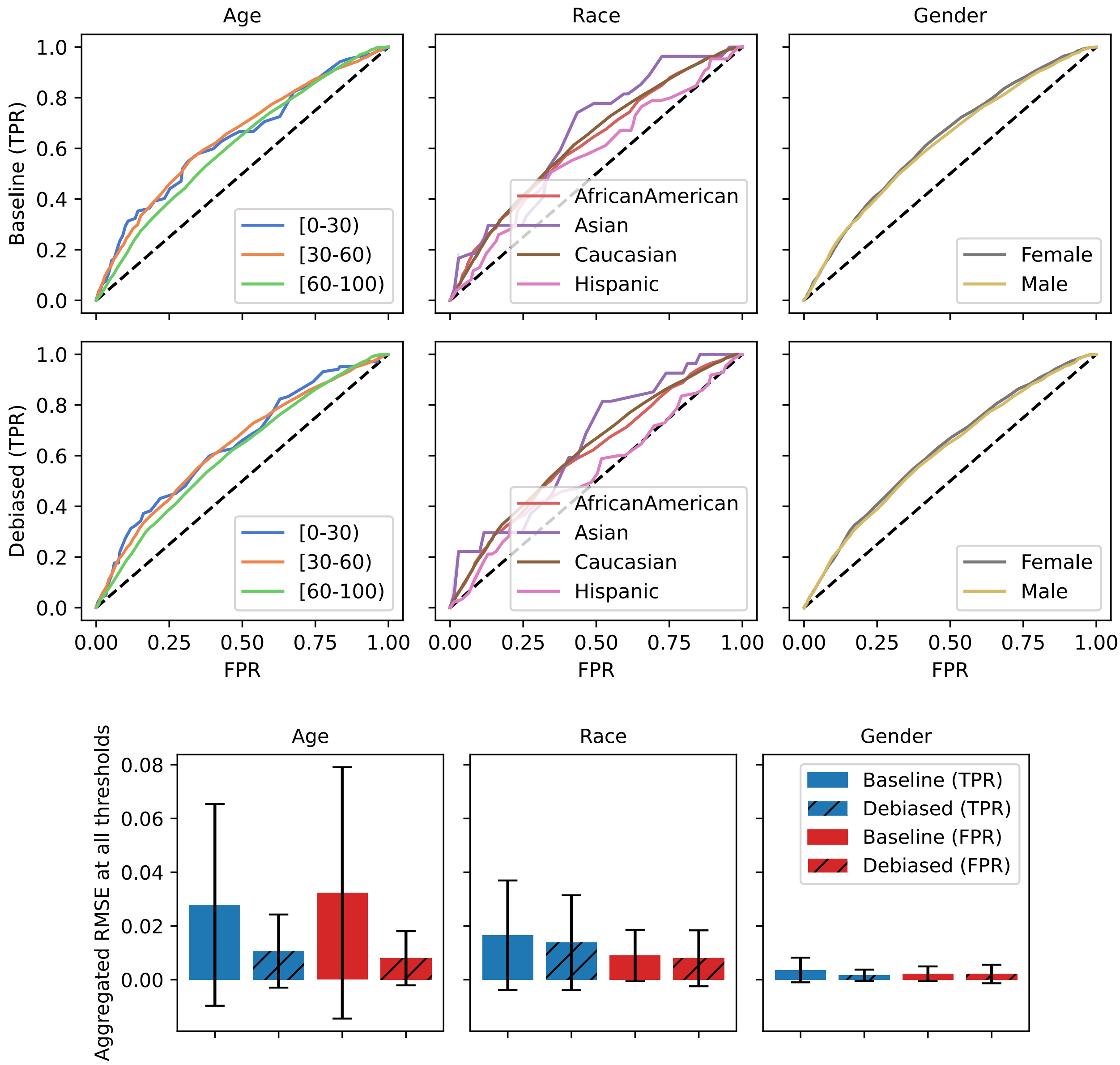

Figure 3: The top shows ROC curves for each of the protected groups before and after debiasing ($\lambda = 0$). The bottom plot shows bars equivalent to the mean RMSE of each categories' TPR and FPR calculated at each of the thresholds. Errorbars indicate the standard deviation across RMSEs. The TPR is denoted with blue and the results from the debiased data are marked with hatched bars.

5 Discussion

5.1 Fairness accuracy tradeoff

As seen in Figures 3, 4, and 8, debiasing the features using the decorrelation method was successful at mitigating most of the biases while only decreasing overall accuracy by 0.006. The decrease in accuracy was also expected to be reflected in the TPRs and FPRs, but as seen in Figure 2 this did not happen. The TPRs and FPRs after debiasing seem to have increased. This was speculated to be due to the poor fit of the model and not the data debiasing method. This is supported by the figures presented in the results.

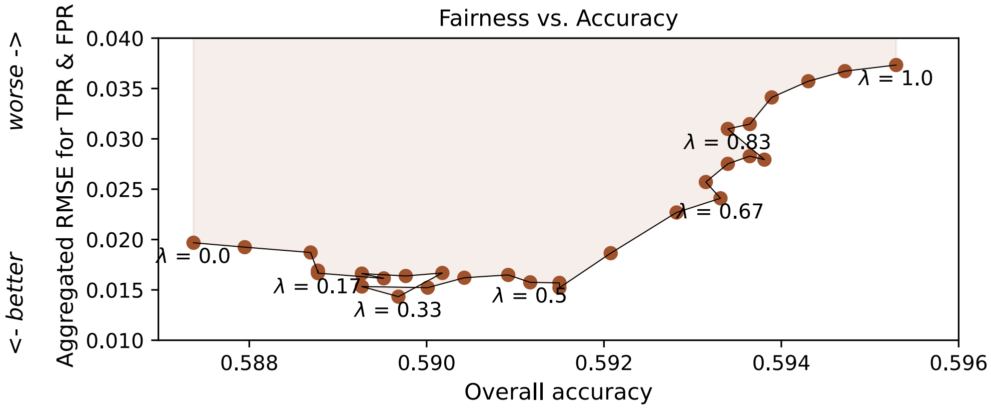

Figure 4: Overall fairness vs. overall accuracy with 31 varying λ. Accuracy is calculated as the macro average, i.e. over all protected groups. Overall fairness is calculated by first calculating the RMSE of TPR and FPR for each protected group, and then aggregating them across groups by the mean. A high value indicates that RMSE was on average high, meaning that TPR and FPR varied more across the protected groups (worse fairness). A low value indicates that the TPR and FPR of the protected groups was on average closer (better fairness).

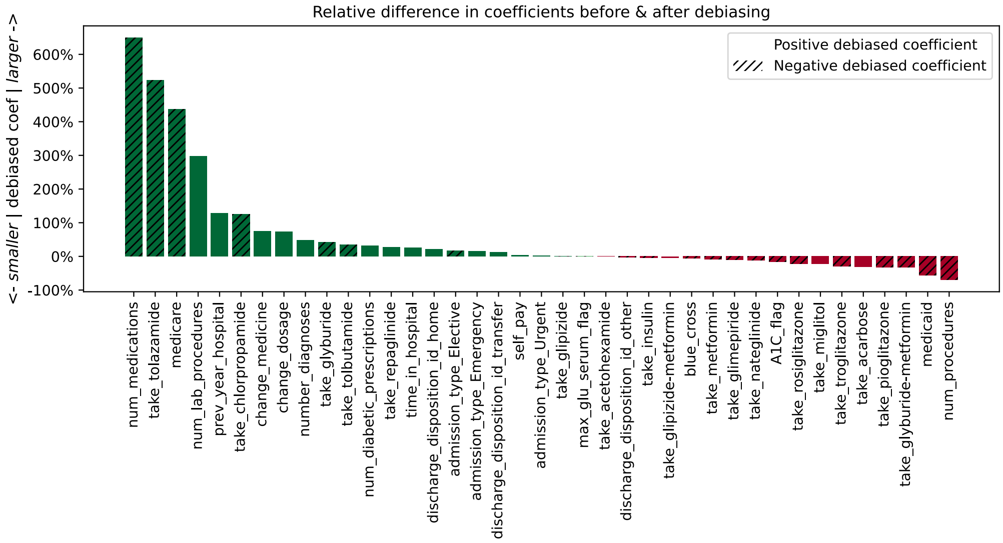

Figure 5: Relative % difference in absolute coefficient values of the logistic regression models before and after debiasing ($\lambda = 0$). Positive values indicate that the debiased coefficient was larger, whereas negative values indicate that the debiased coefficient was smaller.

5.2 Representation bias

The model although trained on the debiased data was not able to overcome the representation bias inherent in the data. This bias was speculated to cause the poor predictions of *Asian* and *Hispanic* categories as seen in Figure 3 (top). Worse, patients in age group $[0 - 30)$ were incorrectly predicted at a rate of 0.60 by the baseline model as seen in Figure 2. These three subgroups constituted less than 6% of the data and still had readmission rates of ≈ 0.30, which the model could not predict.

5.3 Limitations of data debiasing

The debiasing method was effective in removing Pearson correlations between features and protected groups (see Figure 9 in Appendix). However as seen in Figure 6, Spearman correlations still prevailed after debiasing. This could explain why most of the coefficient estimates were larger after debiasing as seen in Figure 5. We had expected that the estimated effects of proxy variables would decrease after training the model on debiased data. As an example MEDICARE had a larger estimate after debiasing, even though it was Pearson correlated with several protected groups in the unprocessed data (see Figure 6). This might be because the decorrelation method may be less effective for categorical variables as it does not remove the monotonic relationships between variables.

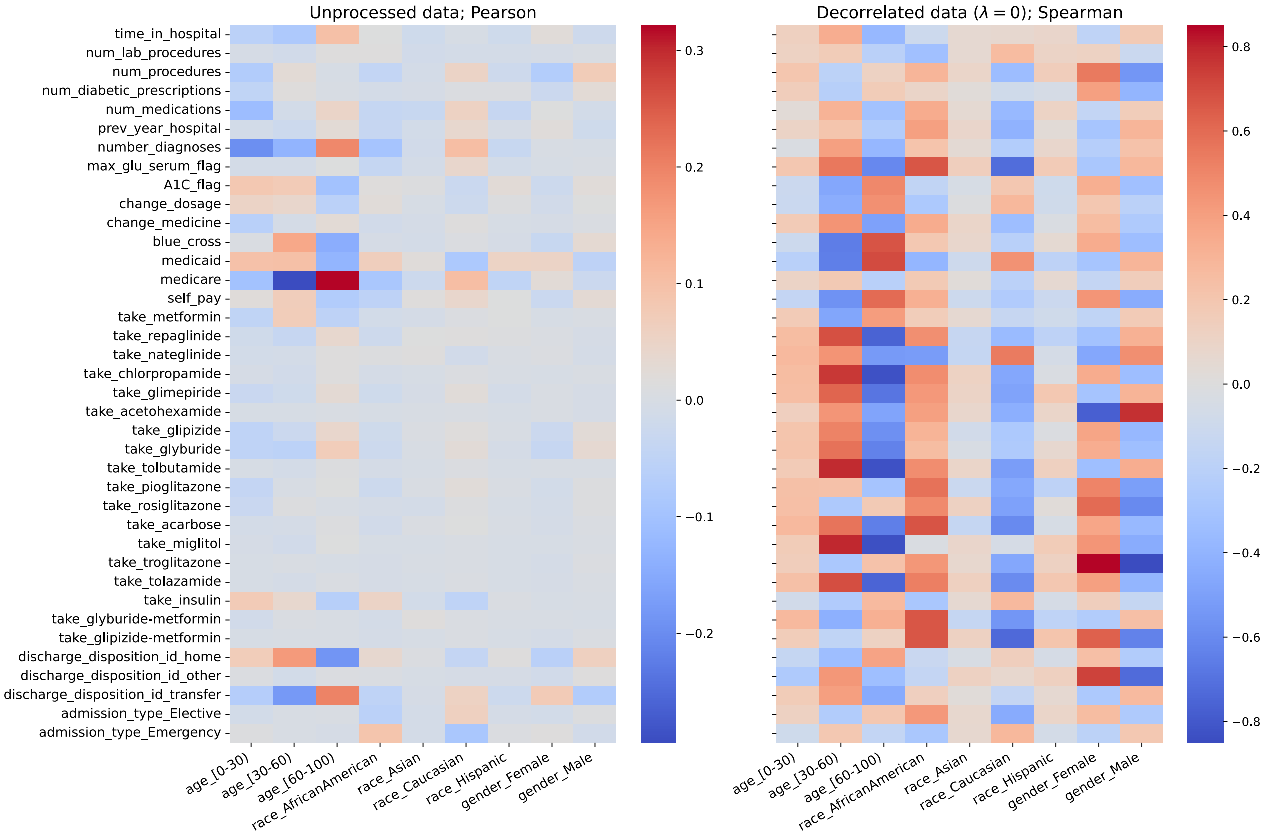

Figure 6: Correlations between protected groups (x-axis) and features (y-axis) in the downsampled data. The left subplot shows the Pearson correlations on the unprocessed data whereas the right subplot shows the Spearman correlations on the debiased data ($\lambda = 0$).

5.4 Explainability in healthcare AI

Data scientists with mechanistic knowledge of a logistic regression model may understand why a particular patient was predicted to be readmitted. However, a person unfamiliar with machine learning might not understand the decision-making process of the model. This lack of understanding can be a problem when deploying the model in a hospital. The hospital personnel, such as nurses and doctors, may not comprehend why a patient was predicted to be re-admitted by the model, especially if their judgement opposes the model decision.

This illustrates the drawbacks of using modern AI systems as opposed to the "Good Old Fashioned AI" (GOFAI). GOFAI is the idea of using computers to enhance mechanistic explanations of science, where each step of an algorithm can be explained. In a healthcare context, a GOFAI algorithm would embody the knowledge of healthcare personnel, thereby functioning like a super-intelligent doctor. In contrast, modern AI relies on discovering correlations in data to provide fast and reliable answers with little to no mechanistic explanation. Relying on correlations in data makes modern AI systems powerful, but it also means that biases inherent in the data can be learned and reproduced by the model, as seen in our baseline model. However, we also showed that decorrelating the features from the protected groups successfully mitigated most of these biases.

When using modern AI algorithms, like our logistic regression model in a hospital setting, the personnel might mistakenly view the model as a GOFAI algorithm, akin to a super-doctor whose decisions are unquestionable. This may lead them to be overly reliant on the model and disregard their own experienced knowledge. In contrast, other healthcare personnel might refuse to trust the model and hence ignore the algorithm's decision.

These scenarios demonstrate the complexity of employing a modern AI algorithm to assist healthcare personnel in decision-making processes. But if used correctly algorithms can assist by detecting patterns overlooked by human judgement, thereby supporting the mechanistic human expertise, ultimately resulting in a better quality of service to patients.

6 Conclusion

This paper examined the fairness of a logistic regression model designed to assist healthcare professionals in predicting whether diabetic patients would be re-admitted to a hospital. We compared a baseline model trained on unprocessed data to models trained on debiased data where features had been decorrelated with the protected groups AGE, RACE and GENDER. The fairness of all models were evaluated using equalized odds. We found that it was generally difficult for the models to predict patient re-admissions. In addition, we found that our baseline model exhibited discrimination towards RACE, AGE and GENDER. After training the models on a debiased representation of the data, most of the biases were mitigated, at the small cost of 0.006 in overall accuracy.

References

[1] European Parliament. Eu ai act: first regulation on artificial intelligence, jun 2023. URL https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. [Accessed on May 19, 2024].

[2] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[3] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

[4] Mohammed Yousef Shaheen. Applications of artificial intelligence (ai) in healthcare: A review. *ScienceOpen Preprints*, 2021.

[5] Andreas S Panayides, Amir Amini, Nenad D Filipovic, Ashish Sharma, Sotirios A Tsaftaris, Alistair Young, David Foran, Nhan Do, Spyretta Golemati, Tahsin Kurc, et al. Ai in medical imaging informatics: current challenges and future directions. *IEEE journal of biomedical and health informatics*, 24(7):1837–1857, 2020.

[6] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[7] Jon DeShazo John Clore, Krzysztof Cios and Beata Strack. *Diabetes 130-US Hospitals for Years 1999-2008*. UC Irvine Machine Learning Repository, 2014. URL https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008.

[8] Saxena N Lerman K Galstyan A. Mehrabi N, Morstatter F. A survey on bias and fairness in machine learning. pages 1–35, 2021.

[9] Michael Fang. Trends in the prevalence of diabetes among us adults: 1999–2016. *American journal of preventive medicine*, 55(4):497–505, 2018.

[10] USA Population 1999. https://countryeconomy.com/demography/population/usa?year=1999. Accessed: 21 May 2024.

[11] Yuzi He, Keith Burghardt, and Kristina Lerman. A geometric solution to fair representations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 279–285, 2020.

[12] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

7 Appendix

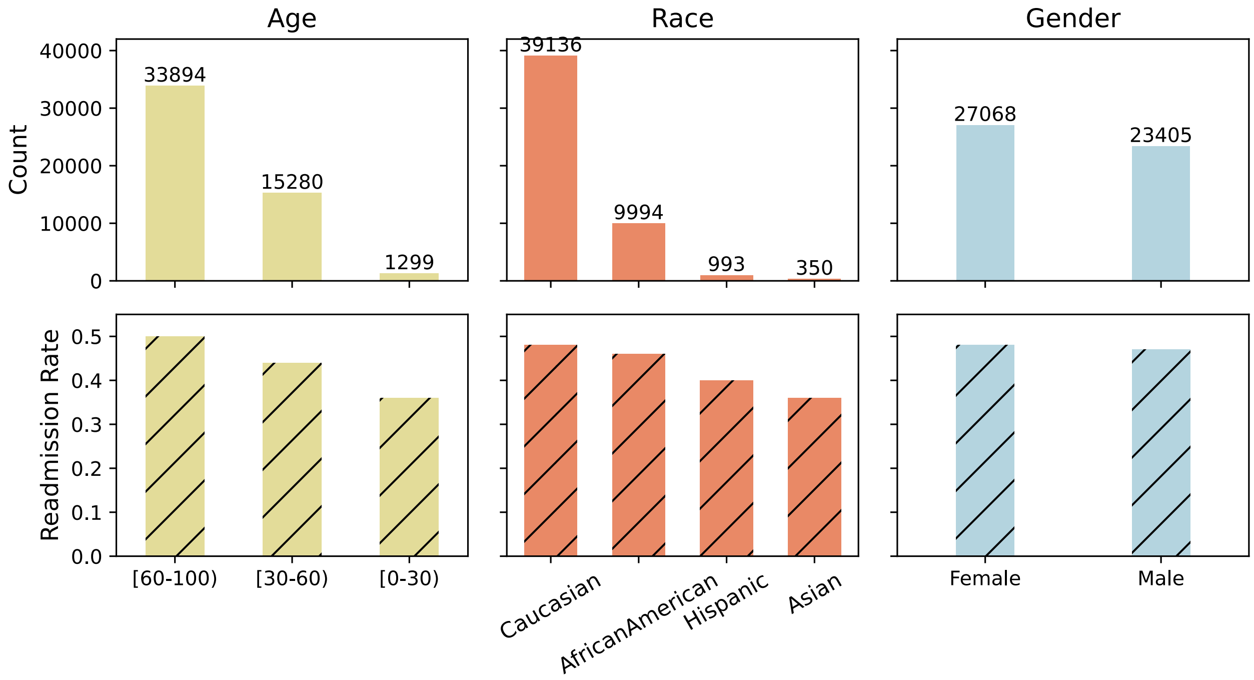

Figure 7: Number of patients (top), and readmission-rates (bottom) in the protected groups for the **downsampled data**.

Table 1: Feature Description

Feature Name	**Type**	**Value**	**Description**
time_in_hospital	numeric	1-14	Length of stay in hospital (days)
num_lab_procedures	numeric	1-132	# Lab procedures
num_procedures	numeric	0-6	# Other procedures
num_diabetic_prescriptions	numeric	0-6	# Diabetic prescriptions
num_medications	numeric	1-81	Unique medications
number_diagnoses	numeric	1-16	# Diagnoses
prev_year_hospital	binary	0-1	Any admissions in the past year
max_glu_serum_flag	binary	0-1	Glucose test performed
A1C_flag	binary	0-1	A1C test performed
change_dosage	binary	0-1	Change in diabetic prescription dosage
change_medicine	binary	0-1	New medication prescribed
blue_cross	binary	0-1	Blue Cross insurance
medicaid	binary	0-1	Medicaid insurance
medicare	binary	0-1	Medicare insurance
self_pay	binary	0-1	Self-pay
discharge_disposition_id	categorical	home, transfer, unknown, other	Type of discharge
admission_type	categorical	emergency, urgent, elective	Type of admission
take_metformin	binary	0-1	Metformin medication taken
take_repaglinide	binary	0-1	Repaglinide medication taken
take_nateglinide	binary	0-1	Nateglinide medication taken
take_chlorpropamide	binary	0-1	Chlorpropamide medication taken
take_glimepiride	binary	0-1	Glimepiride medication taken
take_acetohexamide	binary	0-1	Acetohexamide medication taken
take_glipizide	binary	0-1	Glipizide medication taken
take_glyburide	binary	0-1	Glyburide medication taken
take_tolbutamide	binary	0-1	Tolbutamide medication taken
take_pioglitazone	binary	0-1	Pioglitazone medication taken
take_rosiglitazone	binary	0-1	Rosiglitazone medication taken
take_acarbose	binary	0-1	Acarbose medication taken
take_miglitol	binary	0-1	Miglitol medication taken
take_troglitazone	binary	0-1	Troglitazone medication taken
take_tolazamide	binary	0-1	Tolazamide medication taken
take_insulin	binary	0-1	Insulin medication taken
take_glyburide-metformin	binary	0-1	Glyburide-metformin medication taken
take_glipizide-metformin	binary	0-1	Glipizide-metformin medication taken
age	categorical	[0-30), [30-60), [60-100)	
race	categorical	Caucasian, AfricanAmerican Hispanic, Asian	Race of the patient
gender	categorical	Male, Female	
patient_nbr	numeric	Unique ID	Patient number
readmitted_flag	binary	0-1	Patient readmitted or not

Table 2: Coefficients Comparison

Feature	**Coefficient**	**Decorr. Coefficient**	**Coefficient Difference**
num_medications	-0.0048	-0.0359	648.7380
take_tolazamide	-0.0138	-0.0859	523.3341
medicare	-0.0066	-0.0356	436.6757
num_lab_procedures	-0.0040	0.0159	297.4850
prev_year_hospital	0.2837	0.6496	128.9712
take_chlorpropamide	-0.0474	-0.1068	125.2019
change_medicine	-0.0586	0.1023	74.4243
change_dosage	0.1160	0.2004	72.7875
number_diagnoses	0.1789	0.2649	48.1037
take_glyburide	-0.0316	-0.0449	42.0840
take_tolbutamide	-0.1737	-0.2326	33.8790
num_diabetic_prescriptions	0.2077	0.2745	32.1791
take_repaglinide	0.0775	0.0991	27.8086
time_in_hospital	0.0567	0.0710	25.3831
discharge_disposition_id_home	0.1307	0.1580	20.9136
admission_type_Elective	-0.1230	-0.1429	16.1096
admission_type_Emergency	0.0710	0.0821	15.5438
discharge_disposition_id_transfer	0.1856	0.2095	12.8915
self_pay	0.0779	0.0812	4.2506
admission_type_Urgent	0.0593	0.0608	2.4799
take_glipizide	-0.0299	-0.0300	0.2643
max_glu_serum_flag	-0.1978	-0.1979	0.0685
take_acetohexamide	0.3270	0.3249	-0.6380
discharge_disposition_id_other	-1.6760	-1.6437	-1.9275
take_insulin	-0.1490	-0.1438	-3.4374
take_glipizide-metformin	0.6013	0.5773	-4.0010
blue_cross	-0.4307	-0.4076	-5.3701
take_metformin	-0.3031	-0.2795	-7.7800
take_glimepiride	-0.1043	-0.0938	-10.1082
take_nateglinide	-0.2999	-0.2668	-11.0162
A1C_flag	-0.1341	-0.1131	-15.6572
take_rosiglitazone	-0.0299	-0.0232	-22.2562
take_miglitol	0.5728	0.4433	-22.6111
take_troglitazone	-0.1293	-0.0924	-28.5756
take_acarbose	0.0576	0.0395	-31.3289
take_pioglitazone	-0.0424	-0.0290	-31.6439
take_glyburide-metformin	-0.0769	-0.0524	-31.8698
medicaid	-0.1248	-0.0552	-55.7721
num_procedures	-0.0477	-0.0147	-69.1264

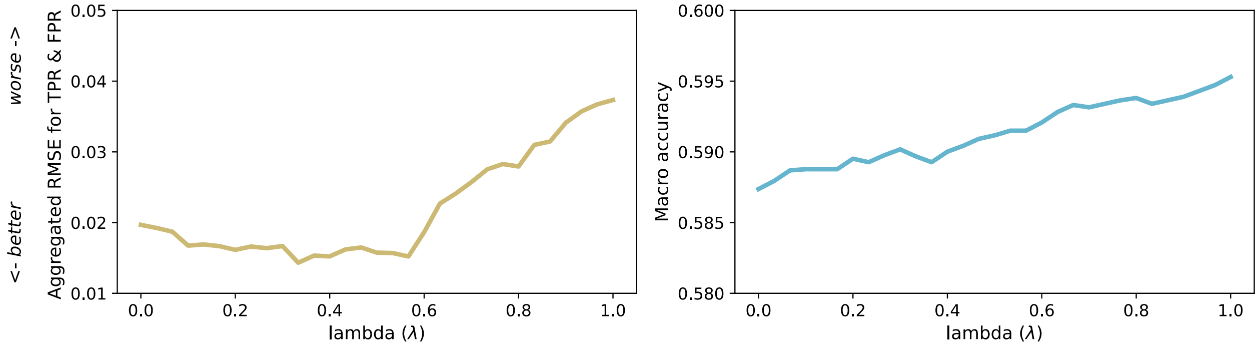

Figure 8: Overall fairness (left) and overall accuracy (right) with varying λ. Accuracy is calculated as the macro average, i.e. over all protected groups. Overall fairness is calculated by first calculating the RMSE of TPR and FPR for each protected group, and then aggregating them across groups by the mean. This is calculated for each λ. A high value means that RMSE was on average high, meaning that TPR and FPR varied more across the protected groups (worse fairness). A low value means that the TPR and FPR of the protected groups was on average closer (better fairness).

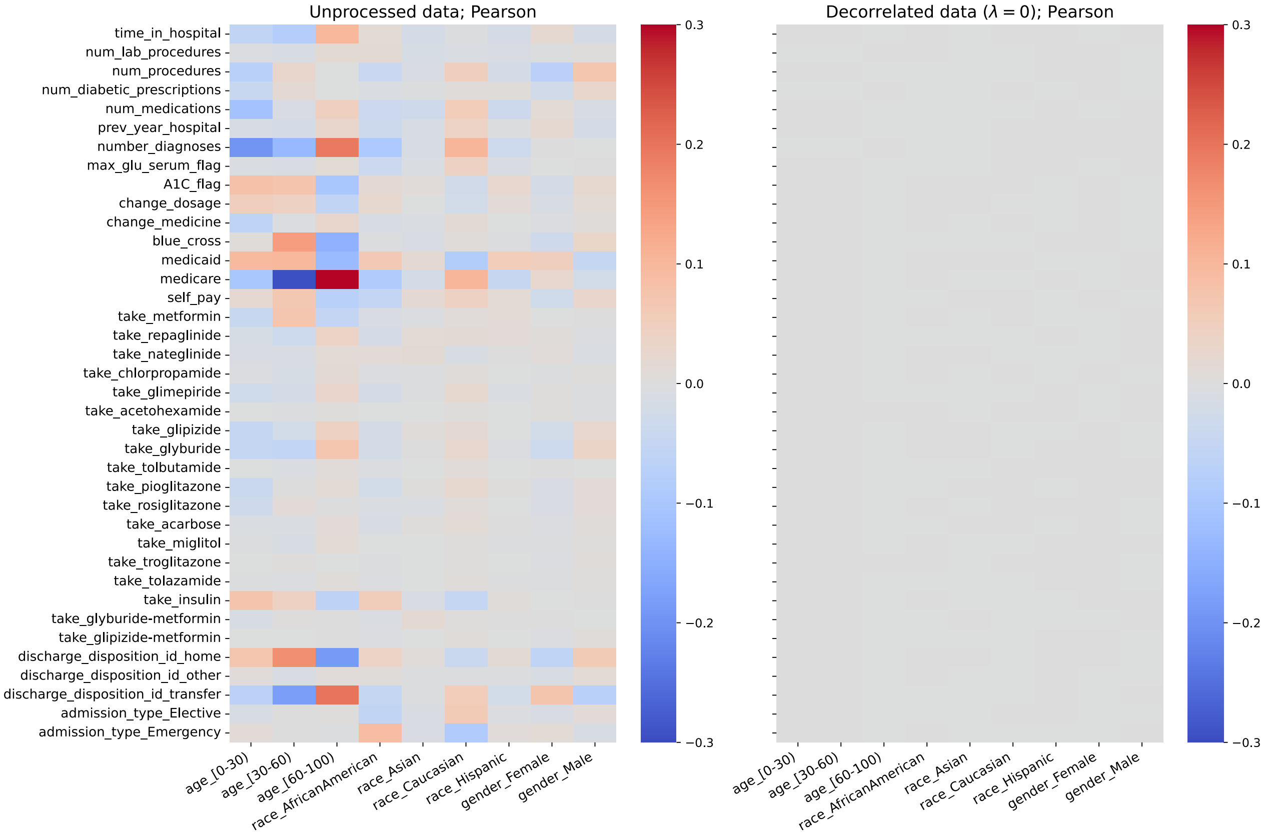

Figure 9: Correlations between protected groups (x-axis) and features (y-axis) in the downsampled data. The left subplot shows the Pearson correlations on the unprocessed data whereas the right subplot shows the correlations on the debiased data ($\lambda = 0$).