An ExplainableFair Framework for Prediction of Substance Use Disorder Treatment Completion

Mary M. Lucas*, Xiaoyang Wang*, Chia-Hsuan Chang*, Christopher C. Yang*, Jacqueline E. Braughton†, and Quyen M. Ngo†

*College of Computing and Informatics
Drexel University, Philadelphia PA, USA
Email: mml367,xw388,cc3859,chris.yang@drexel.edu

†Butler Center for Research
Hazelden Betty Ford Foundation, Minnesota, USA
Email: jbraughton,QNgo@hazeldenbettyford.org

***Abstract*—Fairness of machine learning models in healthcare has drawn increasing attention from clinicians, researchers, and even at the highest level of government. On the other hand, the importance of developing and deploying interpretable or explainable models has been demonstrated, and is essential to increasing the trustworthiness and likelihood of adoption of these models. The objective of this study was to develop and implement a framework for addressing both these issues - fairness and explainability. We propose an explainable fairness framework, first developing a model with optimized performance, and then using an in-processing approach to mitigate model biases relative to the sensitive attributes of race and sex. We then explore and visualize explanations of the model changes that lead to the fairness enhancement process through exploring the changes in importance of features. Our resulting-fairness enhanced models retain high sensitivity with improved fairness and explanations of the fairness-enhancement that may provide helpful insights for healthcare providers to guide clinical decision-making and resource allocation.**

***Index Terms*—predictive model, substance use disorder, bias, fairness, in-processing, explainability**

I. Introduction

Fairness of artificial intelligence (AI Fairness) has become increasingly important, drawing attention even from the highest levels of government [1]. Developing fair models and implementing strategies to mitigate AI biases is an active ongoing field of research. In predictive modeling for health applications, the importance of fairness goes beyond legal and ethical concerns, having significant implications for population health and the imperative to eliminate health disparities [1], [2]. Many studies have evaluated bias in predictive models and attempted to improve fairness through different preprocessing, in-processing, and post-processing approaches. Preprocessing approaches deal with the data before it goes into the model, rebalancing or reweighting it to remove disparities and imbalances that could inform the predictions and introduce bias in the model outputs. Postprocessing approaches involve adjusting the predictions after they come out of the model. In-processing approaches, on the other hand, involve introducing changes to the model training process itself, and are sensitive to the characteristics of the algorithm used for training. They are therefore less transparent and require more efforts to explain. Along with concerns regarding AI fairness, explainability has emerged as critical consideration when developing and implementing predictive models in healthcare [3]. This focus on explainable AI must extend to fairness, allowing clinicians and others who implement these models in practice to understand what the changes are from a biased model to a fair one.

A well performing but biased model presents the risk of introducing or exacerbating health disparities, leading to poorer health outcomes or higher health costs for one demographic group over another. A well performing model that is not explainable ("black box") may find low adoption by health care practitioners as they cannot observe and verify the rationale for the model's outputs or its decision-making process. Bias mitigation, the process by which a model is made more fair, may often involve trading off some performance to increase fairness [4]. A model that improves the fairness of the predictions across demographic groups but does not provide insight into how that fairness is achieved is equally problematic, as the lack of insight into the exact nature, extent, and impact of these trade-offs may present a challenge to the clinician as to how a patient with particular features is affected by this process. Pierson (2024) cautions that, in the context of AI fairness for health equity, "applying quick technical fixes... without understanding what they do or whether it's relevant" [5] may cause more harm than good.

This study aims to explore the explainability of fairness enhancement. We propose ExplainableFair, a novel framework for sequentially executing fair model training and providing explanations for enhanced fairness. Under this framework, we first develop a model with well-tuned predictive performance. We then utilize an in-processing approach on the trained model for bias mitigation. Finally, we employ feature importance

[1]https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

analysis to interpret fairness improvement, applying clarifying questions to explain why the model becomes more fair by examining how the importance of different features varies across the fairness enhancement process. We apply our framework to the task of predicting failure to complete substance use disorder (SUD) treatment.

II. Literature Review

In this section we review relevant studies related to predictive modeling in SUD, AI fairness, and explainability.

A. Predictive Modeling

Ethical concerns over the use of AI and ML in different domains have emerged over the past few years, with bias identified in several commercial AI applications, one of the most prominent being a model used in criminal recidivism risk prediction that demonstrated bias against Blacks [6], [7]. In the healthcare domain, an algorithm used to identify and help patients with complex health needs exhibited bias against Black patients, leading to reduced resources being allocated to these patients [8]. Researchers have used ML approaches in SUD treatment for various tasks, for example predicting outcomes of SUD treatment [9], [10], and specifically readmission or relapse after SUD treatment [11], [12]. While these studies individually and collectively explore multiple ML approaches and algorithms, identify important predictors, and report good performance, they do not report any bias evaluation or how the models perform for different demographic groups. Our previous work used preprocessing by resampling to improve fairness of models predicting SUD treatment completion failure [13], but did not extend into explaining the fairness enhancement beyond group distribution imbalances.

B. Bias Mitigation through In-processing

In-processing approaches for bias mitigation learn a fair model by modifying the model learning process. Learning fair representation [14]–[16] is one of two main approaches which is focused on learning a generative model that maps each data point into an arbitrary representation. The goal of the generative model is to generate a representation that reserves the ability to predict the target label while decoupling the dependence on sensitive attributes. The other main in-processing approach is regularized optimization [17]–[20], which learns a predictive model by adding a disparity regularization term with the performance loss. The common regularization terms are group-based fairness metrics, such as demographic parity, equalized opportunity, and equalized odds. Since most of in-processing bias mitigation approaches focus on reaching a fair model, they lack in reporting of feature importance change during the fairness optimization.

C. Explainable AI in Healthcare

The importance of explainability for creating trust in AI-informed decision making has been discussed in many studies. Angerschmid et. al. (2022) examine the effects of fairness and explanations through a case study, using example-based explanations and feature importance-based explanations [21]. Their study reveals that decisions accompanied by explanations result in increased trust. However, introducing fairness information has mixed results that reflects the complexity of deploying AI explanation and fairness statements. Yang, C. (2022) [3] discusses the importance of explainability in health AI and frames the process through two types of questions - explanation and clarification. The explanation questions are information-based whereas the clarification questions are instance-based, where the required information is generated as the model is executed to provide additional information regarding the predictions. Zhou, Chen, and Holzinger (2020) [22] provide an overview of the relationship between AI fairness and explanation and conclude that fairness requires "comprehensive contextual understanding" and that AI explanations can contribute to this.

III. Materials and Methods

A. The data

The data used for this study was extracted from the Hazelden Betty Ford Foundation (HBFF) electronic health record (EHR). HBFF is one of the largest nonprofit addiction treatment providers in the United States. HBFF data contain information on 20 years of patient encounters from 2000 to 2019 and is described in more detail in [11]. There are multiple variables available, including demographic information (e.g., race, ethnicity, age, legal sex), other patient related and socioeconomic variables (e.g., education level, employment status, occupation, marital status), encounter-specific variables (e.g. length of stay, primary diagnosis, discharge status), diagnosis-related variables (e.g., substance used, co-occurring mental health diagnoses), and variables encoding responses to clinical questionnaires. A distinct feature of this dataset is that unlike most structured EHR data that typically comprise objective clinical measurement variables (e.g. vital signs, laboratory measurements, medications administered, etc.), this dataset also contains multiple fields comprising answers to questionnaires administered to the patient during their stay and treatment at HBFF treatment facilities. These include the American Society of Addiction Medicine (ASAM) Criteria which measure substance use severity and are grouped into six dimensions [11]. Also included is information on the types of services the patient utilized in their treatment journey.

We included inpatient and outpatient encounters, using the discharge status variable to determine treatment completion. To reduce uncertainty in the discharge status we excluded encounters where patients had a transfer, and only included encounters with discharge "with staff aproval" (WSA), "conditional with staff approval" (CWSA), "against staff/medical advice" (ASA/AMA), or "at staff request" (ASR). Encounters with the discharge status of ASA/AMA and ASR were considered to not have successfully completed treatment. Since the aim is to build a model to predict failure to complete treatment, these encounters were labelled as positive (class label 1) and those with status WSA or CWSA were labeled

as having completed treatment (class label 0). Additional data preparation included aggregating race groups with small sample sizes and categorising race as "Caucasian" or "Not Caucasian". The final dataset comprised 10,673 encounters for 9,369 distinct patients, and is highly imbalanced both with respect to the target variable and the demographic group distributions. Table I shows the group distributions stratified by treatment completion status.

TABLE I
DISTRIBUTION OF PATIENTS BY SENSITIVE ATTRIBUTES AND CLASS LABEL

Characteristic	Negative Class (9,149)	Positive Class (1,524)
Race		
Caucasian	8,230 (90%)	1,341 (88%)
Not Caucasian	919 (10%)	183 (12%)
Sex		
Male	5,824 (64%)	1,062 (70%)
Female	3,325 (36%)	462 (30%)

To obtain more reliable and robust assessments, we divided the dataset into a training set D_{train} (90%) and a test set D_{test} (10%) ten times, with random different train-test splits on each occasion. The distribution of sensitive attributes and the target variable is similar across each combination of training and test sets.

B. ExplainableFair Framework

As delineated in Figure 1, the ExplainableFair framework is bifurcated into two distinct phases: the model training phase and the fairness explanation phase. During the model training phase, we train a Logistic Regression model $f_{\omega_{perf}}$ to maximize predictive performance. Subsequently, this trained model undergoes a fine-tuning process for addressing fairness concerns, resulting in a fair model $f_{\omega_{fair}}$ with equitable performance across demographic groups related to sensitive attributes. The fairness explanation phase then builds upon $f_{\omega_{perf}}$ and $f_{\omega_{fair}}$. Utilizing the SHapley Additive exPlanations (SHAP) values [23] obtained from the two models, we analyze the changes in feature importance attributable to the fairness optimization. This involves a comparative assessment of the most important features and a focused examination of those features that exhibit the most pronounced changes between two models. The implementation of this framework not only enhances the fairness of the resulting model but also bolsters its interpretability, particularly for clinical contexts.

C. Predictive Modeling

We select the logistic regression model as our classifier due to its robustness and interpretability. Given a patient $x \in R^m$ with the target variable y, where m is the number of features, the learning process of the logistic regression model is to find the coefficient value $\omega \in R^m$ that reduces the difference between actual value y and the predicted value $f_\omega(x) = \hat{y}$. Because of the class imbalance in our dataset, we set the classification threshold using "The Closest to (0,1) Criteria (ER)", which uses the point minimizing the Euclidean distance between the ROC curve and the (0,1) point [24], [25]. Therefore, we can find the optimal model $f_{\omega_{perf}}$ that minimizes the entropy loss in all patients X, where the coefficient ω_{perf} is identified by:

$$\text{argmin}_\omega \ \mathcal{L}^{CE}(\omega) = -E_{x\sim X}\left(y\log(\hat{y}) + (1-y)\log(1-\hat{y})\right) \tag{1}$$

D. Bias Mitigation

In this stage, we fine-tune $f_{\omega_{perf}}$ to optimize the model fairness. We select *Equalized Odds (EO)*, a comprehensive measure described by Hardt et al. [26], as the fairness criterion. EO mandates that both the True Positive Rate (TPR) and False Positive Rate (FPR) exhibit minimal disparity across different demographic groups. Based on this concept, the proposed loss function incorporates the *Equalized Odds Disparity (EOD)*, denoted as $\mathcal{L}^{EOD}$, which is the sum of the differences in TPR and FPR as presented in Equation 2:

$$\mathcal{L}^{EOD}(\omega) = \mathcal{L}_{TPR} + \mathcal{L}_{FPR}, \tag{2}$$

where the least square error is used for the calculation of TPR (FPR) difference loss term:

$$\mathcal{L}_{TPR} = (TPR_{z=0} - TPR_{z=1})^2 \tag{3}$$

$$\mathcal{L}_{FPR} = (FPR_{z=0} - FPR_{z=1})^2 \tag{4}$$

$$TPR_z = Pr(\hat{y} = 1 | z = z, y = 1) \tag{5}$$

$$FPR_z = Pr(\hat{y} = 1 | z = z, y = 0) \tag{6}$$

Note that $z \in \mathcal{Z}$ is defined as a sensitive attribute. Therefore, for example, when optimizing the pre-learned model for a race-fair model, we set $z = 1$ for Caucasian and $z = 0$ for Non-Caucasian. Similarly, to achieve a sex-fair model, we set $z = 1$ for males and $z = 0$ for females. This $\mathcal{L}^{EOD}$ loss term ensures that the TPR and FPR differences between demographic groups are minimized simultaneously. As a result, we use $\mathcal{L}^{EOD}$ to keep optimizing the coefficient values of $f_{\omega_{perf}}$ and obtain a fair model $f_{\omega_{fair}}$ after the optimization.

E. Feature Importance Analysis

We utilize a local feature attribution method, SHAP, which distributes the prediction score of a fitted model for a patient $x \in X$ to its base features R^m. The score of a base feature can be interpreted as the importance of the feature to the patient. With this method, we propose Algorithm 1 to analyze how the feature importance vary between $f_{\omega_{perf}}$ and $f_{\omega_{fair}}$. In the Algorithm, line 1 and 2 initialize two SHAP explainers using DeepSHAP [23] . Since the SHAP value is model-dependent, we initialize E_{perf} and E_{fair} for $f_{\omega_{perf}}$ and $f_{\omega_{fair}}$, respectively. We then apply the explainers to calculate the SHAP values for the test dataset (line 3), which result in two feature importance matrices $S_{perf}, S_{fair} \in R^{n\times m}$. n is the number of patients in the testing dataset, and m is the

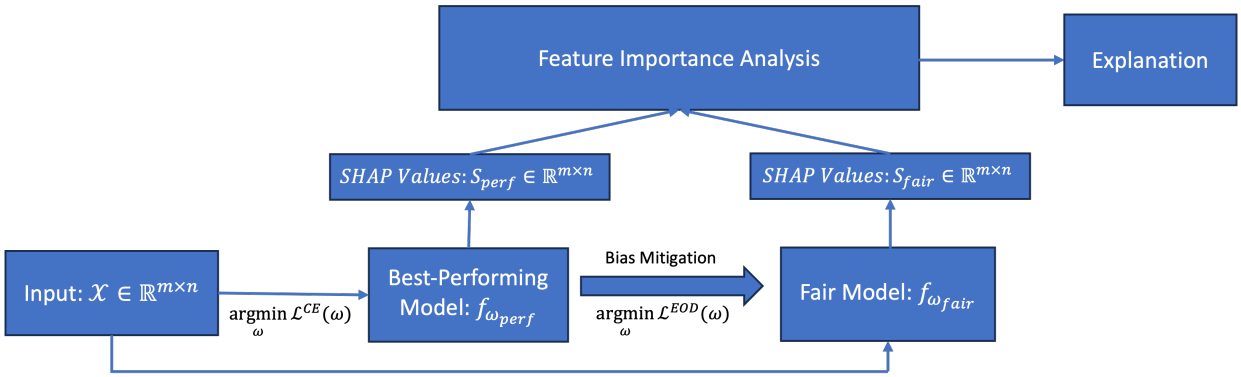

Fig. 1. The ExplainableFair framework.

Algorithm 1: Feature Importance Analysis using SHAP Values

Input : Model before Bias Mitigation $f_{\omega_{perf}}$, Fair Model after Bias Mitigation $f_{\omega_{fair}}$, Train set X_{train}, Test set X_{test}

Output: Feature importance rankings R_{perf}, R_{fair}, Most Changed features Set C

1 $E_{perf} \leftarrow shap.DeepExplainer(f_{\omega_{perf}}, X_{train})$;
2 $E_{fair} \leftarrow shap.DeepExplainer(f_{\omega_{fair}}, X_{train})$;
3 $S_{perf}, S_{fair} \leftarrow E_{perf}(X_{test}), E_{fair}(X_{test})$;
4 $\bar{S}_{perf}, \bar{S}_{fair} = mean(S_{perf}), mean(S_{fair})$;
5 $R_{perf} \leftarrow RankFeatures(\bar{S}_{perf})$;
6 $R_{fair} \leftarrow RankFeatures(\bar{S}_{fair})$;
7 $C \leftarrow CompareRankings(R_{perf}, R_{fair})$;

number of features. Each row represents a patient with m importance scores, where each corresponds to a feature.

To quantify the the extent of importance changes for each feature across the best-performing model and the fair model, we propose to apply two operations:

1) Aggregation: as shown in line 4, we aggregate feature importance scores of each feature by taking average (i.e., mean function in the algorithm) across n patients, which results in $\bar{S}_{perf}, \bar{S}_{fair} \in R^{1 \times m}$.
2) Rank transformation: SHAP values are model-dependent, so it is invalid to compare the aggregated importance scores between models. To overcome this, we sort features in $\bar{S}_{perf}$ and $\bar{S}_{fair}$ in a descending order and use the rank to represent the importance of each feature (line 5 and 6). R_{perf} and R_{fair} are the outputs, and they denote the importance ranks of features.

Consequently, we can use difference of rank between R_{perf} and R_{fair} to identify the features whose ranking changes the most. This helps us to highlight key features that may be critical for fairness, potentially providing the insights for practitioners. Moreover, to provide the robust analysis, we apply Algorithm 1 for ten different training and testing splits. Our experimental results are the aggregated results of ten splits.

TABLE II
MODEL PERFORMANCE AND FAIRNESS - RACE-FAIR OPTIMIZATION

	AUROC	Sensitivity	Specificity	EOD
Best Performing Model	0.8607	0.7948	0.8054	0.0725
Race-Fair Model	0.8585	0.8037	0.7533	0.0298

TABLE III
MODEL PERFORMANCE AND FAIRNESS - SEX-FAIR OPTIMIZATION

	AUROC	Sensitivity	Specificity	EOD
Best Performing Model	0.8607	0.7948	0.8054	0.0603
Sex-Fair Model	0.8576	0.7829	0.7545	0.0282

F. Performance and Fairness Evaluation

Predictive performance was evaluated using Area Under Receiver Operating Characteristic Curve (AUROC), sensitivity, and specificity. In the context of our study, we operationalized the concept of the 'privileged group' based on dataset representation with respect to a sensitive attribute. Specifically, the privileged group for a specific sensitive attribute is defined as the group having a larger representation within the dataset. To quantify the fairness of the model, we employed the EOD, mathematically defined as the arithmetic mean of the differences in the TPR and FPR across privileged groups and unprivileged groups, formulated as follows:

$$EOD = \frac{\mathcal{L}_{TPR} + \mathcal{L}_{FPR}}{2} \tag{7}$$

To ensure the robustness and reliability of our findings, each model training iteration was repeated ten times. We then computed and reported the average values across these iterations for key metrics, including predictive performance, fairness levels, and feature importance.

IV. RESULTS

In Tables II and III, we compare the model performance before and after fairness optimization for each sensitive attribute (race and sex).

Figures 2 through 4 present the twenty most important features, based on SHAP, for the best performing model before

fairness optimization, and for the race-fair model and sex-fair model respectively. We discuss the most important of these in the Discussion section.

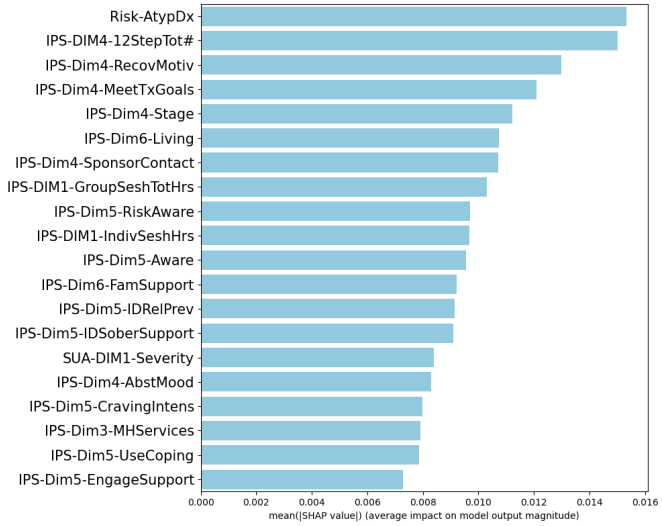

Fig. 2. Most important features before fairness optimization.

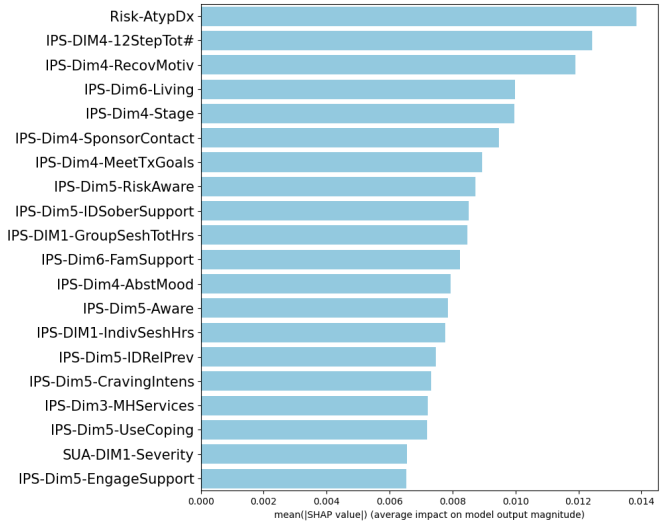

Fig. 3. Most important features after *race-fair* optimization.

In figures 5 and 6 we show the features whose importance ranking changes the most during race-fairness and sex-fairness optimization respectively. The bars to the right (in blue) indicate increase in importance ranking while those to the left (red) indicate decrease in ranking. We again expand on the most important of these, with respect to change in ranking, in the Discussion section.

V. Discussion

The overall purpose of healthcare predictive modeling is inherently pragmatic, applying data-driven results into real-world implications for clinical providers and patients. In the

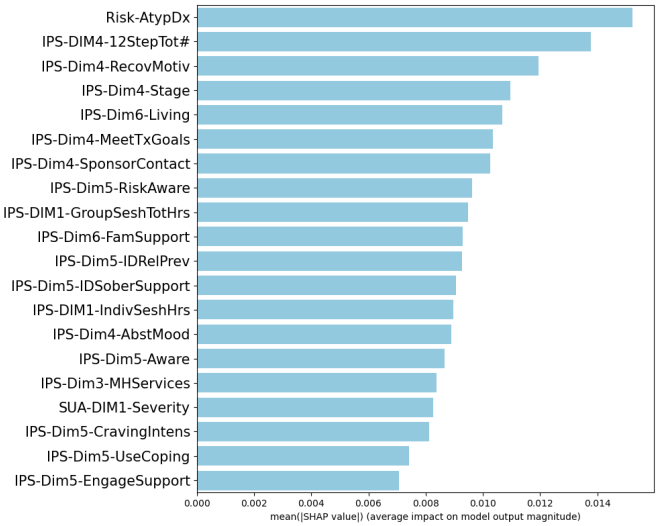

Fig. 4. Most important features after *sex-fair* optimization.

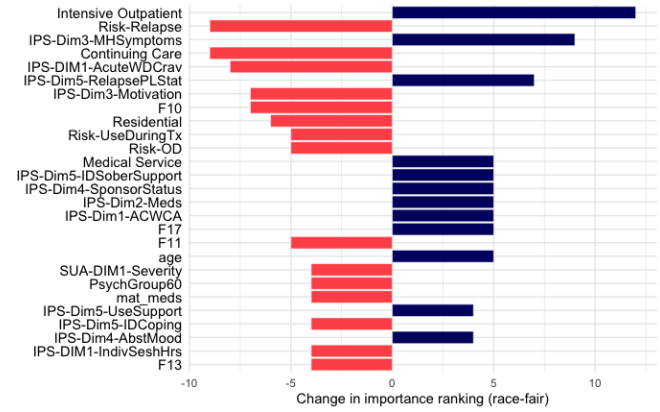

Fig. 5. Most changed features (by ranking) during race fairness optimization

field of alcohol and substance use treatment and recovery, these implications can have life-altering ramifications for patients' long-term recovery. There is a wide, and ever-growing gap in accessibility and quality of treatment for those with substance use disorders who need, seek, initiate, and attend treatment. This is particularly true for those from marginalized groups (e.g., racial minority, gender/sexual minority).

As use of machine learning and AI continues to grow in driving healthcare decisions, model features associated with optimal treatment outcomes, with their potential underlying biases, will directly impact the treatment and access that may be available to patients.

In this study, we introduced a novel framework that develops a fair model for predicting SUD treatment completion using in-processing methods, and explained the enhancement of fairness through feature importance analysis. We adapted the explainability structure of questions in [3], and crafted fairness-relevant explanation and clarification questions to guide our work. Because we are concerned with fairness across demographic groups, we framed the clarification questions as

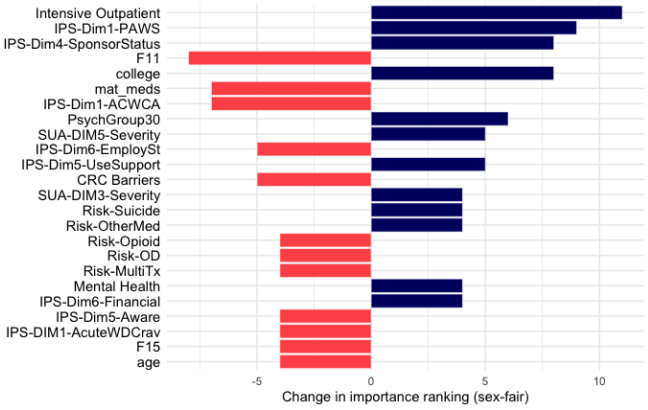

Fig. 6. Most changed features (by ranking) during sex fairness optimization

group-based rather than instance-based (Table IV).

We used logistic regression to train a best performing model, and then optimized it for fairness as described in the Methods section. Using SHAP values to determine the feature importance ranking of the different features in the base and fairness optimized models, we addressed the fairness-relevant questions to develop an understanding and explanation of the fairness enhancement process.

A. Information-based explanation questions

The results of model performance as measured by AUROC, sensitivity, and specificity, presented in Tables II and III guide us to answer the information-based questions regarding the performance of the best model before and after fairness optimization. The best performing model has an AUROC of 86.07%, sensitivity of 79.48% and specificity of 80.54%. The race-fair and sex-fair model have AUROC 85.85% and 85.76%, sensitivity 80.37% and 78.29%, and specificity 75.33% and 75.45% respectively. Thus we observe that both the best performing and the fairness optimized models have acceptable performance, above 75%, on all metrics.

We obtain insight into the most important features before and after fairness optimization from figures 2 to 4. We observe that the most important features for predicting failure to complete SUD treatment remained fairly stable across the fairness optimization process. In the best performing model, the most predictive features are documented risk of atypical discharge (*"Risk-AtypDx"*) and scores on the various dimensions of the ASAM Criteria. Of these, the most dominant are those from dimension 4 of the criteria, which assess the readiness to change, specifically participation in the 12-Step program (*"IPS-DIM4-12StepTot#"*), motivation to recover (*"IPS-Dim4-RecovMotiv"*), meeting treatment goals (*"IPS-Dim4-MeetTxGoals"*), and stage of recovery process (*"IPS-DIM4-Stage"*). These five most important features remain stable through the fairness optimization, although we do observe from figures 3 and 4 that the patient's living environment (*"IPS-Dim6-Living"*) becomes more important in the fair models (this is an element of dimension 6 of the ASAM criteria, which assesses the patients recovery environment).

Clinical Implications

The results in figure 2, before fairness optimization, with exception to risk for atypical discharge, identifies themes associated with individual readiness and motivation to change in four of the top five most important factors for treatment. In this model, clinicians would likely be instructed to place emphasis in encouraging patients, regardless of their motivation and readiness to change, to find and regularly attend a local 12-step, peer recovery support group. Previous empirical findings support that consistent involvement in community groups, such as 12-step/peer recovery support groups, is a salient factor in sustained recovery [27]–[29]. In addition, clinical treatment goals and associated interventions would likely focus on guiding patients to identify their readiness and subsequent motivation for change (i.e., recovery), and then exploring ways to build upon, maintain, or challenge these beliefs. As a result, meeting individual treatment goals would likely mean that patients have shown willingness to, and actionable change towards, developing skills that increase their confidence/motivation of sustaining change and building a solid recovery support system.

Comparison between figures 2 and 3, as well as figures 2 and 4 illustrate similarities across the top three features even after optimizing for race-fairness and sex-fairness. Yet, the optimization for both race- and sex-fairness denotes a distinct difference, raising the priority of living conditions and stage of recovery, and decreasing the overall importance of meeting treatment goals as an indicator for successful SUD treatment for non-White patients. This difference may be because marginalized groups, including women, tend to face additional barriers to meeting treatment goals [30]–[32]. Effective treatment goals are relevant and achievable; in other words, they must be relevant to the collection of symptoms/diagnosis presented and reflect consideration of individual contextual factors and resourcing. Without this consideration, treatment goals are ineffective at best, and may result in traumatization at worst. While gaining diversity, much of the extant research on substance use treatment and recovery in the United States is based on majority White, heterosexual, biological male samples. Evidence-based treatment goals are therefore, often standardized and conceptualized through the experience of the majority (i.e., in this case, White, biological male, westernized medical model) patients [33]. As a result, sole use of meeting treatment goals without meaningful changes that reflect consideration of historical, environmental, and social context that disproportionately affect marginalized communities (including women) may result in ineffective treatment and decreased patient outcomes for those populations.

In the race- and sex-fair models, clinicians would likely focus interventions to helping individuals identify their stage of recovery and ensuring individuals developed the tools necessary and connected with appropriate resources to ensure their living conditions/day-to-day responsibilities were conducive to recovery.

TABLE IV
INFORMATION-BASED AND CLARIFICATION-BASED QUESTIONS FOR EXPLAINING MODEL FAIRNESS OPTIMIZATION

Information-based explanation questions	
Best performing model	For the best performing model before fairness optimization, what is the performance for the whole population?
	What are the most important features?
Fairness-optimized model	After optimizing the fairness while retaining performance, what is the performance for the whole population?
	What are the most important features?
Group-based clarification questions	
What if ?	What are the differences between groups in the best performing model?
	What are the differences between groups in the fairness-optimized model?
	What are the fairness measures used?
	What is the improvement in fairness?
	What is the sacrifice in performance to achieve fairness?
How to be that?	Does the importance of features change after fairness optimization?
	What features become more important when the fairness is optimized based on a selected fairness criteria or measure?
	What features become less important when the fairness is optimized based on a selected fairness criteria or measure?

B. Group-based clarification questions

The "What if?" clarification questions address the overarching question "what would happen if...?", i.e. what changes occur if we optimize the best performing model for fairness. We refer again to tables II and III for the Equalized Odds Disparity (EOD) between race groups and sex groups respectively. This is the arithmetic mean of the difference in TPR and the difference in FPR between groups and is previously described in eqn 7. The EOD between race groups in the best performing model is 0.0725 while that in the race-fair model is 0.0298. In comparing male vs female groups, the EOD for the best performing model is 0.0603 while that of the sex-fair model is 0.0282. We use the EOD as the fairness measure. In looking at model performance, we observe that there was a slight drop (less than 1%) in AUROC after optimizing the model for race-fairness and for sex-fairness. The sex-fair model also had a 1.2% drop in sensitivity, while both fairness optimized models had an approximately 5% drop in specificity. We note a significant increase in fairness, with a drop in EOD of 4.27% in the race-fair model and 3.21% in the sex-fair model. Because the positive class in our prediction task is "did not complete treatment", we place a higher priority on the sensitivity of the model. This is because having a drop in TPR would be costly, resulting in missing patients who could benefit from additional interventions to help them complete treatment. Our fairness optimized models maintain high performance, close to that of the best performing model while simultaneously exhibiting less bias.

The "How to be that?" clarification questions on the other hand may be understood to address the issue of "how do we get there from here?" How do the features change to attain model fairness? We answer these questions by examining figures 5 and 6. For our logistic regression model to become race-fair the features that increase most in importance ranking include the number of days spent in intensive outpatient treatment (*"Intensive Outpatient"*), presence of mental health symptoms (*"IPS-Dim3-MHSymptoms"*), and the status of relapse prevention planning (*"IPS-Dim5-RelapsePLStat"*). The features that drop most in importance ranking include documented risk of relapse (*"Risk-Relapse"*), number of days in spent in the Continuing Care program (*"Continuing Care"*), and acute withdrawal and craving (*"IPS-DIM1-AcuteWDCrav"*). For the sex-fair model, the features that increase most in importance ranking include the number of days spent in intensive outpatient treatment (*"Intensive Outpatient"*), current post-acute withdrawal syndrome (*"IPS-Dim1-PAWS"*), status of sponsor relationship (*"IPS-Dim4-SponsorStatus"*), and having a college education (*"college"*). On the other hand, having a diagnosis related to opioid use disorder (*"F11"*), having a record of medications in the EHR medication administration table (*"mat_med"*) , and scores on Acute Withdrawal Craving Answers (*"IPS-Dim1-ACWCA"*) drop in importance ranking.

Clinical Implications

In reviewing figure 5, we see several important changes. First, the decreased importance of (a) continuing care and (b) individual motivation to recover is contrary to widely-held empirically-based beliefs that both are critical elements for successful SUD treatment outcomes. These results suggest that the type of continuing care offered may not be as helpful to

marginalized communities and may be more tailored for this sample's dominant racial group. In like manner, the race-fair model showed an increased importance of intensive outpatient treatment (IOP), which allows the ability to receive treatment services while patients remain within their communities. This suggests that successful treatment and continuing care plans must include ways for patients from marginalized groups to work on their recovery within their own community, where they are much more likely to have a felt sense of belonging, likely resulting in increased feelings of safety and having more practical, sustainable supports [29], [33], [34]. Decreased emphasis in motivation to change may also suggest that regardless of how motivated a patient is to complete treatment and recover, there may still be enduring (individual, environmental, structural, etc.) barriers that forcibly hinder an individual's ability to meet treatment goals and sustain recovery, which is often the case for marginalized groups. For example, integration of co-occurring mental health symptoms also increases in importance. Overall, statistics support that marginalized groups have higher incidences and severities of co-occurring medical and psychiatric symptoms due to long-standing barriers to care and ineffective treatment [30]–[32]. In comparison, it seems that factors that tend to be most important for successful SUD treatment for the dominant racial groups may not accurately account for structural barriers and complexity of cases that may be present for marginalized communities. Overall, the features that increased in importance in the race-fair model – IOP, mental health symptoms, sober support, sponsor status – illuminate the consideration of complexity of cases, barriers to care, vital need for community/sense of belonging during SUD treatment, and recovery support for marginalized communities.

From figure 6, we also discover several important changes. For the sex-fair model, there is an increased importance in having a college education, which is not observed in the race-fair model. Previous research supports that for patients of color, socio-economic factors (e.g. education level or income) do not appear to benefit them within the healthcare system. For example, maternal mortality for Black women in the United States is high regardless of individual income level or education status. In contrast, without intersectional factors of race and ethnicity, education status tends to be significant for female/female-identifying patients as higher education typically increases their knowledge of and access to care. In addition, the sex-fair model illustrated increased importance in post-acute withdrawal syndrome (PAWS), participation within an intensive outpatient treatment program, and obtaining a sponsor/recovery mentor. As with the race-fair model, integration of flexible treatment programs, such as an intensive outpatient treatment program, may help women get the care that they need without having to abandon day-to-day responsibilities and community support. This may be especially important for women with high parental/childcare and household management responsibilities, who may otherwise not be able to seek and receive treatment. Increased emphases in obtaining sponsors illuminates a need for women to feel like they belong and have mentors/social support within the recovery community. Due to historical stigma where women, specifically mothers, struggling with substance use are seen as amoral, and much of the treatment and recovery community being comprised of male peers, women may still struggle to find sponsors that can provide relevant and meaningful support. This suggests that clinical treatment may want to include specific interventions and resources towards identifying and solidifying a recovery community in which female-identifying patients feel a sense of belonging. While gaining more attention, PAWS is still viewed as a novel phenomenon with some evidence of gender differences. Yet, these results suggest the importance of regular assessment and integration of medical treatment for post-acute withdrawal for female/female-identifying patients. Future research should investigate the differences in PAWS prevalence among gender and race/ethnicity patient populations. As with the race-fair, the sex-fair model illustrates unique features of increased importance (e.g., college education, sponsor status, intensive outpatient program, PAWS) for ensuring successful treatment for female/female identifying patients.

Though there are salient similarities across all models, the nuanced differences illustrated in figures 5 and 6 have critical ramifications for implementing elements of successful treatment experiences across diverse patient populations.

VI. Conclusion

In this study we present ExplainableFair framework, a novel approach to developing a fair model and explaining the fairness enhancement by comparing feature importance before and after fairness optimization.

Model unfairness can come from many sources. These include biases encoded in datasets, biases resulting from the algorithm used for training, and biases caused by the features used in the model [35]. It is important to quantify and report these biases, and to use tools at our disposal to create fairer models, whether through preprocessing, in-processing, or postprocessing approaches. Equally important is that these fairness enhancements do not sacrifice the performance of the model to the detriment of its usefulness, and more importantly, that explanations are provided for the changes that occur during the fairness optimization. Examining the changes in feature importance as we optimize a model for fairness has the potential to enhance trust and willingness to apply fairness-optimized models to clinical applications. Additionally, these explanations, when presented in a clinically accessible and relevant manner may provide valuable insights to assist healthcare providers in clinical decision-making and resource allocation.

A. Strengths and Limitations

Because we begin from a model optimized for predictive performance and then apply fairness regularization with performance constraints, our approach results in a fair model with minimal performance loss. A potential limitation of this study is that we use data from only one provider (HBFF),

and therefore our findings may not be generalizable to all SUD treatment providers, facilities, or patients. However, the HBFF dataset is a large multi-year multi-site dataset, and is therefore representative of multiple different regions across the US. The sensitive attribute of race was used as entered in the EHR. The use of race in observational studies has been shown to be problematic, for example where there may be conflicts between self-reported race versus provider-perceived [36]. Additionally, aggregating all non-Caucasian patients into one category due to the smaller sample sizes may lead to our findings not being generalizable to the individual race groups.

ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation under the Grants IIS-1741306 and IIS-2235548, and by the Department of Defense under the Grant DoD W91XWH-05-1-023. This material is based upon work supported by (while serving at) the National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors would also like to thank the Hazelden Betty Ford Foundation for providing the data used in this study, and Dr. Ou Stella Liang for initial data extraction and cleaning.

REFERENCES

[1] "Health Equity in Healthy People 2030 - Healthy People 2030 | health.gov." [Online]. Available: https://health.gov/healthypeople/priority-areas/health-equity-healthy-people-2030

[2] R. K. Wadhera and I. J. Dahabreh, "The US Health Equity Crisis—An Economic Case for a Moral Imperative?" *JAMA*, vol. 329, no. 19, pp. 1647–1649, May 2023. [Online]. Available: https://doi.org/10.1001/jama.2023.4018

[3] C. C. Yang, "Explainable Artificial Intelligence for Predictive Modeling in Healthcare," *Journal of Healthcare Informatics Research*, vol. 6, no. 2, p. 228, Jun. 2022, publisher: Springer. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8832418/

[4] A. K. Menon and R. C. Williamson, "The cost of fairness in binary classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, Jan. 2018, pp. 107–118, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v81/menon18a.html

[5] E. Pierson, "Accuracy and Equity in Clinical Risk Prediction," *New England Journal of Medicine*, vol. 390, no. 2, pp. 100–102, Jan. 2024, publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMp2311050. [Online]. Available: https://doi.org/10.1056/NEJMp2311050

[6] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias." [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[7] A. Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, Jun. 2017, arXiv: 1703.00056 Publisher: Mary Ann Liebert Inc.

[8] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science (New York, N.Y.)*, vol. 366, no. 6464, pp. 447–453, Oct. 2019.

[9] L. Acion, D. Kelmansky, M. v. d. Laan, E. Sahker, D. Jones, and S. Arndt, "Use of a machine learning framework to predict substance use disorder treatment success," *PLOS ONE*, vol. 12, no. 4, p. e0175383, Apr. 2017, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0175383

[10] M. Nasir, N. S. Summerfield, A. Oztekin, M. Knight, L. K. Ackerson, and S. Carreiro, "Machine learning–based outcome prediction and novel hypotheses generation for substance use disorder treatment," *Journal of the American Medical Informatics Association : JAMIA*, vol. 28, no. 6, pp. 1216–1224, Feb. 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8661425/

[11] O. S. Liang, "Developing Clinical Prediction Models for Post-treatment Substance Use Relapse with Explainable Artificial Intelligence," Ph.D. dissertation, Drexel University, 2021, book Title: Developing Clinical Prediction Models for Post-treatment Substance Use Relapse with Explainable Artificial Intelligence.

[12] D. Morel, K. C. Yu, A. Liu-Ferrara, A. J. Caceres-Suriel, S. G. Kurtz, and Y. P. Tabak, "Predicting hospital readmission in patients with mental or substance use disorders: A machine learning approach," *International Journal of Medical Informatics*, vol. 139, p. 104136, Jul. 2020.

[13] M. M. Lucas, C.-H. Chang, and C. C. Yang, "Resampling for Mitigating Bias in Predictive Model for Substance Use Disorder Treatment Completion," in *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*. Houston, TX, USA: IEEE, Jun. 2023, pp. 709–711. [Online]. Available: https://ieeexplore.ieee.org/document/10337249/

[14] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning Fair Representations," in *Proceedings of the 30th International Conference on Machine Learning*. PMLR, May 2013, pp. 325–333, iSSN: 1938-7228. [Online]. Available: https://proceedings.mlr.press/v28/zemel13.html

[15] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating Unwanted Biases with Adversarial Learning," Jan. 2018, arXiv:1801.07593 [cs]. [Online]. Available: http://arxiv.org/abs/1801.07593

[16] T. Adel, I. Valera, Z. Ghahramani, and A. Weller, "One-network adversarial fairness," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'19/IAAI'19/EAAI'19. Honolulu, Hawaii, USA: AAAI Press, Jan. 2019, pp. 2412–2420. [Online]. Available: https://dl.acm.org/doi/10.1609/aaai.v33i01.33012412

[17] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-Aware Classifier with Prejudice Remover Regularizer," in *Machine Learning and Knowledge Discovery in Databases*. Springer, Berlin, Heidelberg, 2012, pp. 35–50. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-33486-3_3

[18] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach, "A Reductions Approach to Fair Classification," in *Proceedings of the 35th International Conference on Machine Learning*. PMLR, Jul. 2018, pp. 60–69, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v80/agarwal18a.html

[19] Y. Roh, K. Lee, S. E. Whang, and C. Suh, "FAIRBATCH: BATCH SELECTION FOR MODEL FAIRNESS," 2021.

[20] A. Shen, X. Han, T. Cohn, T. Baldwin, and L. Frermann, "Optimising Equal Opportunity Fairness in Model Training," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 4073–4084. [Online]. Available: https://aclanthology.org/2022.naacl-main.299

[21] A. Angerschmid, K. Theuermann, A. Holzinger, F. Chen, and J. Zhou, "Effects of Fairness and Explanation on Trust in Ethical AI," in *Machine Learning and Knowledge Extraction*, ser. Lecture Notes in Computer Science, A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl, Eds. Cham: Springer International Publishing, 2022, pp. 51–67.

[22] J. Zhou, F. Chen, and A. Holzinger, "Towards Explainability for AI Fairness," in *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, ser. Lecture Notes in Computer Science, A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek, Eds. Cham: Springer International Publishing, 2022, pp. 375–386. [Online]. Available: https://doi.org/10.1007/978-3-031-04083-2_18

[23] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dec. 2017, pp. 4768–4777.

[24] I. Unal, "Defining an Optimal Cut-Point Value in ROC Analysis: An Alternative Approach," *Computational and Mathematical Methods in Medicine*, vol. 2017, p. 3762651, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5470053/

[25] N. J. Perkins and E. F. Schisterman, "The Inconsistency of "Optimal" Cut-points Using Two ROC Based Criteria." *American journal of epidemiology*, vol. 163, no. 7, pp. 670–675, Apr. 2006. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1444894/

[26] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., Dec. 2016, pp. 3323–3331.

[27] S. Reif, L. Braude, D. R. Lyman, R. H. Dougherty, A. S. Daniels, S. S. Ghose, O. Salim, and M. E. Delphin-Rittmon, "Peer Recovery Support for Individuals With Substance Use Disorders: Assessing the Evidence," *Psychiatric Services*, vol. 65, no. 7, pp. 853–861, Jul. 2014, publisher: American Psychiatric Publishing. [Online]. Available: https://ps.psychiatryonline.org/doi/10.1176/appi.ps.201400047

[28] M. Leamy, V. Bird, C. L. Boutillier, J. Williams, and M. Slade, "Conceptual framework for personal recovery in mental health: systematic review and narrative synthesis," *The British Journal of Psychiatry*, vol. 199, no. 6, pp. 445–452, Dec. 2011.

[29] D. Best, J. Irving, B. Collinson, C. Andersson, and M. Edwards, "Recovery Networks and Community Connections: Identifying Connection Needs and Community Linkage Opportunities in Early Recovery Populations," *Alcoholism Treatment Quarterly*, vol. 35, no. 1, pp. 2–15, Jan. 2017, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/07347324.2016.1256718. [Online]. Available: https://doi.org/10.1080/07347324.2016.1256718

[30] C. for Behavioral Health Statistics and Quality, "Racial/Ethnic Differences in Substance Use, Substance Use Disorders, and Substance Use Treatment Utilization Among People Aged 12 or Older (2015-2019) | CBHSQ Data," 2021. [Online]. Available: https://www.samhsa.gov/data/report/racialethnic-differences-substance-use

[31] O. T. Hall, A. Jordan, J. Teater, K. Dixon-Shambley, M. E. McKiever, M. Baek, S. Garcia, K. M. Rood, and D. A. Fielin, "Experiences of racial discrimination in the medical setting and associations with medical mistrust and expectations of care among black patients seeking addiction treatment," *Journal of Substance Abuse Treatment*, vol. 133, Feb. 2022, publisher: Elsevier. [Online]. Available: https://www.jsatjournal.com/article/S0740-5472(21)00277-4/abstract

[32] E. Sahker, G. Pro, M. Sakata, and T. A. Furukawa, "Substance use improvement depends on Race/Ethnicity: Outpatient treatment disparities observed in a large US national sample," *Drug and Alcohol Dependence*, vol. 213, p. 108087, Aug. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0376871620302520

[33] E. F. Wagner and J. A. Baldwin, "Recovery in Special Emphasis Populations," *Alcohol Research : Current Reviews*, vol. 40, no. 3, p. 05, Dec. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7721369/

[34] C. K. Marino, "To belong, contribute, and hope: first stage development of a measure of social recovery," *Journal of Mental Health*, vol. 24, no. 2, pp. 68–72, Mar. 2015, publisher: Routledge _eprint: https://doi.org/10.3109/09638237.2014.954696. [Online]. Available: https://doi.org/10.3109/09638237.2014.954696

[35] D. Pessach and E. Shmueli, "A Review on Fairness in Machine Learning," *ACM Computing Surveys*, vol. 55, no. 3, pp. 1–44, Mar. 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3494672

[36] F. C. G. Polubriaginof, P. Ryan, H. Salmasian, A. W. Shapiro, A. Perotte, M. M. Safford, G. Hripcsak, S. Smith, N. P. Tatonetti, and D. K. Vawdrey, "Challenges with quality of race and ethnicity data in observational databases," *Journal of the American Medical Informatics Association: JAMIA*, vol. 26, no. 8-9, pp. 730–736, Aug. 2019.