Causal Analysis of Structural Race Bias in Syphilis Treatment: An Ethical Perspective on Algorithmic Fairness

Table of Contents

1. Introduction

Racial disparities in healthcare access and outcomes are well documented, with both historical and contemporary evidence showing how race can shape diagnosis, treatment allocation and patient prognosis. One of the most infamous examples is the Tuskegee Syphilis Study (1932–1972), in which Black men in the United States were deliberately denied effective treatment for syphilis, even after penicillin became the standard of care, in order to observe the natural course of the disease. Conducted without informed consent and rooted in racialised assumptions about Black bodies, Tuskegee stands as a profound ethical failure with lasting harm. While our study does not seek to recreate Tuskegee, it draws on its legacy to pose a pressing question: Could modern algorithmic systems, trained on biased data and shielded by the veneer of objectivity, quietly reproduce similar patterns of injustice? We investigate this possibility using a simulated cohort of syphilis patients to explore how structural race bias may influence treatment allocation and outcomes in healthcare. Crucially, this simulation is not intended to represent any specific healthcare institution or to predict real-world outcomes. Instead, it functions as a conceptual tool to explore the mechanisms through which bias, both statistical and structural, can enter algorithmic decision making. In contrast to speculative concerns about AI taking over healthcare or replacing clinicians, we focus on a more immediate and grounded ethical challenge: how seemingly neutral predictive systems may reinforce or even worsen existing disparities. Using causal inference methods, including do-calculus and inverse probability treatment weighting (IPTW), we estimate treatment effects while adjusting for both observed and unobserved confounding. To assess fairness, we evaluate both counterfactual fairness (via race-flipping interventions) and group fairness metrics, such as precision, true positive rate (TPR), false positive rate (FPR), and calibration error. By examining the intersection of race, fairness, and causality within a stylised but ethically meaningful setting, our study shows how algorithmic models, even under idealised conditions, can replicate and amplify social inequities. These findings raise critical questions about what it means to build "fair" AI in medicine, and why fairness must be understood not just as a technical constraint, but as a normative obligation grounded in historical accountability and structural reform.

2. Literature Review

A growing body of literature illustrates how algorithmic systems in healthcare often reflect and reinforce structural inequities, even when designed with neutral objectives and algorithmic design. This section reviews key studies that highlight different dimensions of fairness, bias, and accountability in AI applications.

Bias from Historical Data and Measurement Devices.

Obermeyer et al. (2019) showed that a healthcare algorithm trained to predict future healthcare costs as a proxy of clinical need underestimated care for Black patients. This was due to structural racism reflected in the training data, rather than algorithmic design, where Black patients would historically have less access to healthcare and be subject to discriminatory practices, leading the model to believe Black patients require less care than White patients. Similarly, Sjoding et al. (2020) revealed that pulse oximeters overestimated blood oxygen saturation in Black patients, leading to underdiagnosis of hypoxemia. They validated blood oxygen saturation by using arterial blood gas (ABG) analysis; a more accurate metric, requiring arterial puncture. These examples highlight how systemic racism embedded in data or devices can produce harmful outcomes, even in the absence of explicit intent.

Bias in Professional Evaluation Systems.
Kiyasseh et al. (2023) studied a surgical AI system and found it exhibited both underskilling and overskilling biases depending on surgeon attributes, hospital affiliation, and patient characteristics. They introduced an explainability framework, called TWIX (Training With EXplanations), to force the model to justify its decisions, reducing subgroup disparities while enhancing performance. This underscores that fairness and performance coexist when explicitly addressed in model design, and the potential benefits of model explainability.

Intersectionality and Diagnostic Bias.
Valentine et al. (2023) explored how race, gender, and socioeconomic status (SES) intersect in schizophrenia diagnosis. Shockingly, high-SES Black men were more likely to be diagnosed than low-SES Black men, while the opposite trend was found among White patients and was assumed to be the relationship between SES and diagnosis. The study argues for intersectional fairness frameworks that move beyond single-variable analysis, warning that ignoring such interactions can mask compounding injustice.

Accountability in AI Governance.
Novelli, Taddeo, and Floridi (2023) define accountability as a relation of answerability. This requires three conditions: recognition of authority, the possibility of interrogation, and limitation of power. They propose a structured framework around seven dimensions - including context, agent, forum, standards, and consequences - and identify four key goals of accountability: compliance, reporting, oversight, and enforcement. Importantly, they distinguish between proactive and reactive accountability, where proactive aims prevent possible harm before model deployment, a standard our project aspires towards. Conversely, reactive accountability is around redress and assigning blame after harm is inflicted upon. They argue that effective governance must go beyond vague calls for transparency and genuinely embed accountability into the sociotechnical systems in which AI operates.

3. Defining Fairness

Fairness is a notoriously contested concept in algorithmic systems. As Krasanakis (2024) notes, “it is mathematically impossible to simultaneously satisfy all conceivable definitions of fairness.” In domains like hiring and education, competing metrics - such as demographic parity, equalised odds, and calibration - can conflict, complicating the operationalisation of fairness. However, in a healthcare context, we argue that fairness can and must be grounded in both technical rigour and moral clarity, due to the nature of the domain. Healthcare is not a meritocratic domain: patients do not earn treatment through prior achievement or effort, but are entitled to care based solely on clinical need. This distinguishes it from hiring or university admissions, where a fundamental tension between meritocracy and equity often shapes fairness. In hiring, for example, a wealthy white candidate may have top qualifications due to lifelong access to elite education and professional opportunities. Meanwhile, a Black candidate from a deprived background may have slightly lower qualifications but has overcome systemic barriers to get there. While the former is more qualified and thus appears more 'employable' on paper, the latter arguably demonstrates greater resilience and potential. This creates a fairness dilemma: should decisions reward formal achievement, or account for contextual disadvantage? In healthcare, that dilemma essentially dissolves. Patients do not compete for care based on past performance; they receive it because they are unwell. As such, both narrow and broad views of fairness converge: fairness means treating clinically similar patients similarly (individual fairness; narrow view) while also ensuring that historically marginalised groups are not systematically underdiagnosed or undertreated (group fairness; broad view). This ethical clarity underpins the fairness metrics applied in our analysis, such as counterfactual and group fairness analysis. We therefore evaluate our models using three fairness criteria that can be explicitly tested in code: equalised odds (ensuring equal error rates across groups), calibration (ensuring predicted probabilities are accurate for all groups), and counterfactual fairness (ensuring decisions do not change if a sensitive attribute is altered). These metrics allow us to move beyond abstract fairness ideals and assess tangible disparities in model outputs. Crucially, fairness in healthcare cannot be disentangled from accuracy. A model that systematically underestimates risk in Black patients or misclassified symptoms in women is clinically incorrect, as well as unfair. Improving predictive accuracy for marginalised groups is, therefore, both a technical necessity and a moral obligation. This is because accuracy and fairness benefit both the patient and the healthcare provider. Fair, precise models promote allocative efficiency: patients receive the right level of care based on actual need. This means under-treated patients are less likely to return with worsened conditions, and over-treated patients do not drain unnecessary resources. In this context, fairness and accuracy are not in conflict but converge. Therefore, building equitable models should, in principle, mean building better, more cost-effective healthcare systems.

4. Dataset Overview

The dataset consists of 4000 samples, each representing an individual with simulated attributes related to syphilis infection and treatment. Key variables include race, severity of the disease, an instrumental variable, treatment status, and observed outcomes. Race categories are encoded as one-hot variables, representing five groups: White (Race_0), Black (Race_1), Hispanic (Race_2), Asian (Race_3), and Other (Race_4). The dataset is designed to study the causal impact of treatment on outcomes and analyze potential bias across racial groups.

5. Variable Generation and Derivation

The race-specific infection rates were derived from CDC data (Centers for Disease Control and Prevention [CDC], 2023). These rates were not intended to recreate real-world data, but rather to inform the simulation design with plausible epidemiological disparities. Race probabilities were therefore weighted to approximate the demographic skew observed in actual infection patterns, allowing for a more realistic stylisation of structural bias under controlled conditions:

Total Infection Rates: $\left\{R_0 = 9.1,\ R_1 = 39.7,\ R_2 = 16.9,\ R_3 = 4.4,\ R_4 = 58.2\right\}$

Total Sum: $9.1 + 39.7 + 16.9 + 4.4 + 58.2 = 128.3$

Normalized Probabilities:

$$P_0 = \frac{9.1}{128.3} \approx 0.07092$$

$$P_1 = \frac{39.7}{128.3} \approx 0.30943$$

$$P_2 = \frac{16.9}{128.3} \approx 0.13172$$

$$P_3 = \frac{4.4}{128.3} \approx 0.03429$$

$$P_4 = \frac{58.2}{128.3} \approx 0.45362$$

Severity is modeled as a continuous variable following a standard normal distribution, while the instrumental variable is binary with equal probabilities.

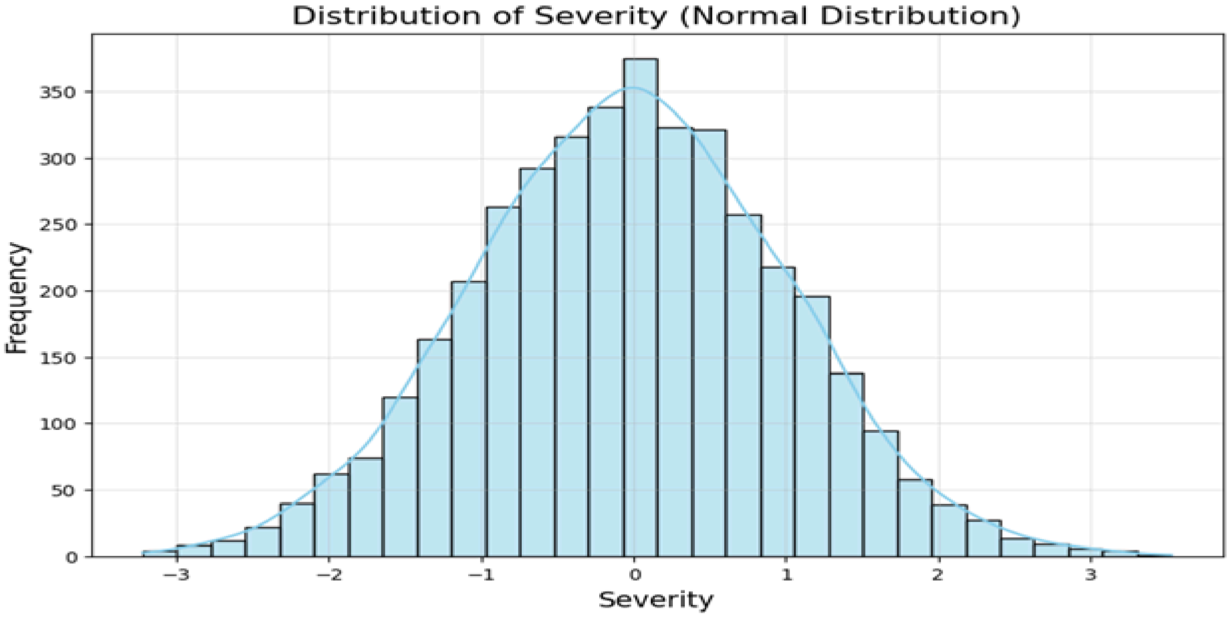


Figure 1 Distribution of Severity of the Dataset

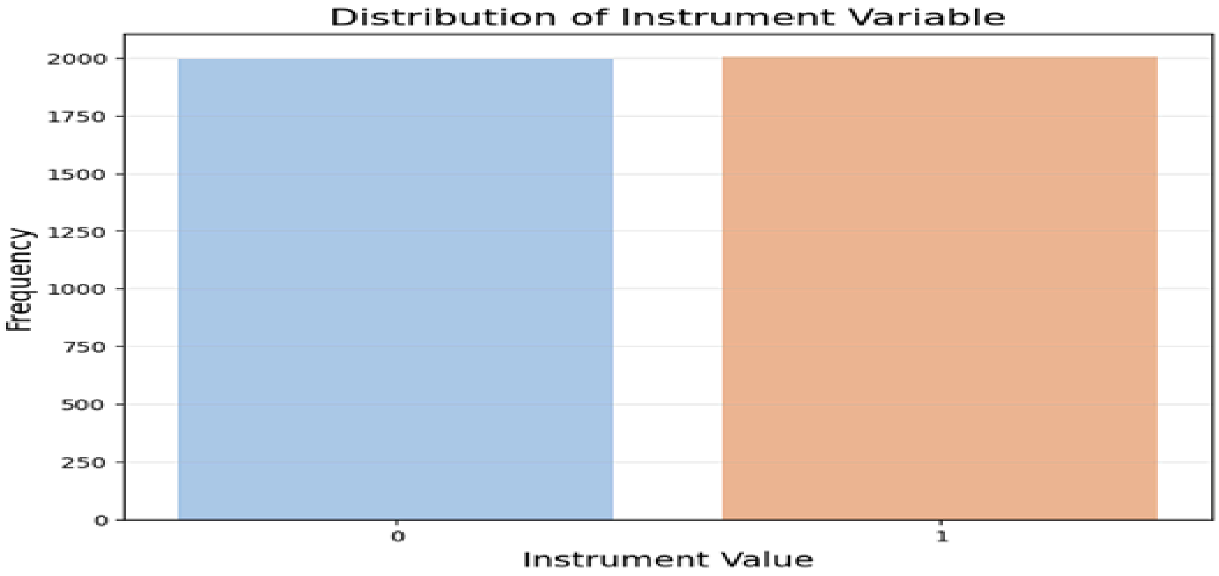


Figure 2 Distribution of Instrument Variable (Balanced)

Treatment probabilities are generated using a logistic regression model that incorporates race bias weights, severity, and the instrumental variable (Zhou, 2016).

logit_treatment = race_bias + 0.5 · severity + 0.8 · instrument

*Severity: In reality, the more severe the disease, the more likely a patient is to receive treatment. Therefore, a positive weight (0.5) is assigned.

* Instrument: The instrument variable may represent some external influences (e.g., medical insurance, policy interventions), which could have a stronger impact on the probability of receiving treatment. Hence, a larger weight (0.8) is assigned.

$$P_{Treatment} = \frac{1}{1+e^{-logit_{treatment}}}$$

$$Treatment \sim Binomial(n = 1,\ p = P(Treatment)))$$

Observed outcomes are determined by treatment status, with potential outcomes (Y_0 and Y_1) simulated using logistic functions and random noise added for realism. The potential outcome without treatment:

$$logit_{y_0} = -\ severity$$

$$Y_0 = \frac{1}{1+e^{-logit_{y_0}}}$$

The potential outcome with treatment:

$$logit_{y_1} = 1.5 - severity$$

$$Y_1 = \frac{1}{1+e^{-logit_{y_1}}}$$

The outcomes are:

$$Outcome = Y_0 \cdot (1 - T) + Y_1 \cdot T \text{ where } T \in \{0, 1\}$$

This step is a critical operation in this causal inference, where the observed outcome depends on the treatment status. It simulates the real-world impact of treatment on patient outcomes. By injecting race bias terms into the treatment probability function, the simulation explicitly models structural discrimination. While synthetic, this mirrors real-world practices where proxy variables may embed racial disparities into clinical algorithms.

6. Preprocessing and Outcome Transformation

The dataset is preprocessed to encode race as one-hot variables and ensure the validity of the observed outcome probabilities. The Outcome_prob variable is clipped to the range [0, 1], and a binary outcome (Outcome_binary) is generated using a binomial distribution. The resulting dataset is structured for causal inference analysis, enabling the examination of treatment effects and fairness metrics across racial groups. The continuous Outcome_prob represents the probability of successful treatment (ranging 0-1), simulating clinical scenarios where outcomes are measured probabilistically (e.g., cure likelihood or symptom severity reduction).

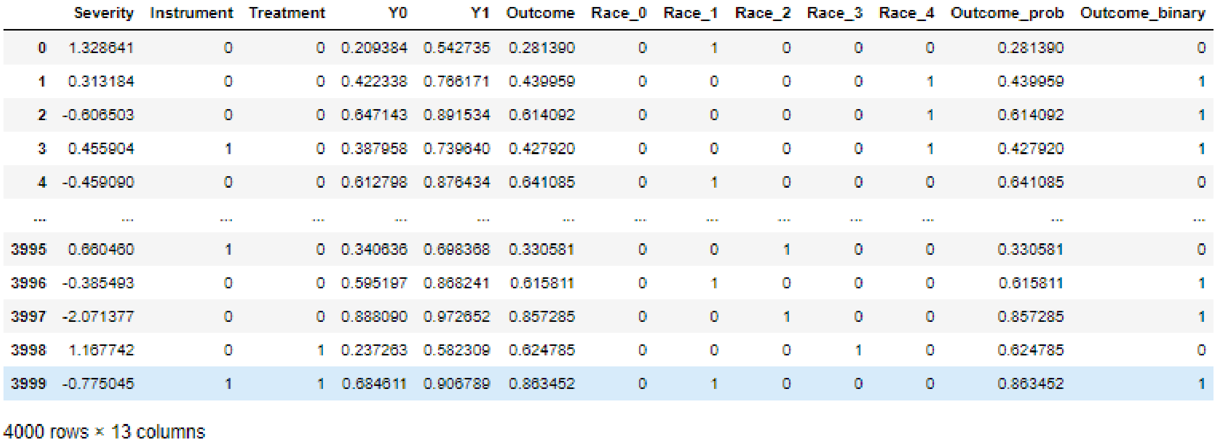
The binary Outcome_binary (1=success, 0=failure) was generated by thresholding Outcome_prob at 0.5, reflecting a clinically meaningful cutoff where >50% probability denotes treatment efficacy.

Race_0	Race_1	Race_2	Race_3	Race_4

0	1	0	0	0
0	0	0	0	1
0	0	0	0	1
0	0	0	0	1
0	1	0	0	0
…	…	…	…	…
0	0	1	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	1	0	0	0

Table 1 Race Columns Converted into Dummies

After the above process, the final dataset is created.



	Severity	Instrument	Treatment	Y0	Y1	Outcome	Race_0	Race_1	Race_2	Race_3	Race_4	Outcome_prob	Outcome_binary
0	1.328641	0	0	0.209384	0.542735	0.281390	0	1	0	0	0	0.281390	0
1	0.313184	0	0	0.422338	0.766171	0.439959	0	0	0	0	1	0.439959	1
2	-0.606503	0	0	0.647143	0.891534	0.614092	0	0	0	0	1	0.614092	1
3	0.455904	1	0	0.387958	0.739640	0.427920	0	0	0	0	1	0.427920	1
4	-0.459090	0	0	0.612798	0.876434	0.641085	0	1	0	0	0	0.641085	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
3995	0.660460	1	0	0.340636	0.698368	0.330581	0	0	1	0	0	0.330581	0
3996	-0.385493	0	0	0.595197	0.868241	0.615811	0	1	0	0	0	0.615811	1
3997	-2.071377	0	0	0.888090	0.972652	0.857285	0	0	1	0	0	0.857285	1
3998	1.167742	0	1	0.237263	0.582309	0.624785	0	0	0	1	0	0.624785	0
3999	-0.775045	1	1	0.684611	0.906789	0.863452	0	1	0	0	0	0.863452	1

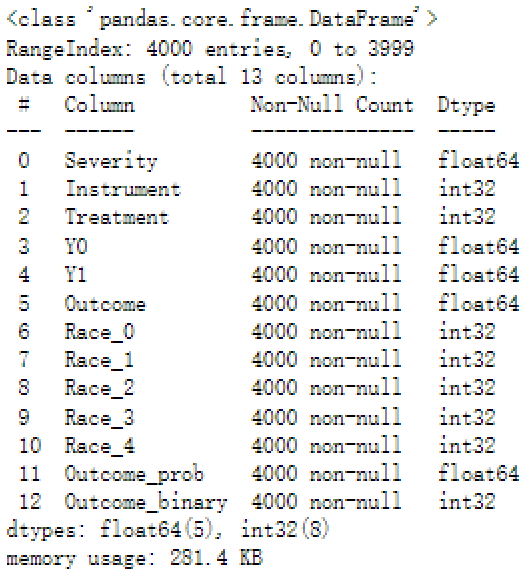
4000 rows × 13 columns

Figure 3 The Final Dataset (Snapshot of Dataframe)

7. Explore the Dataset

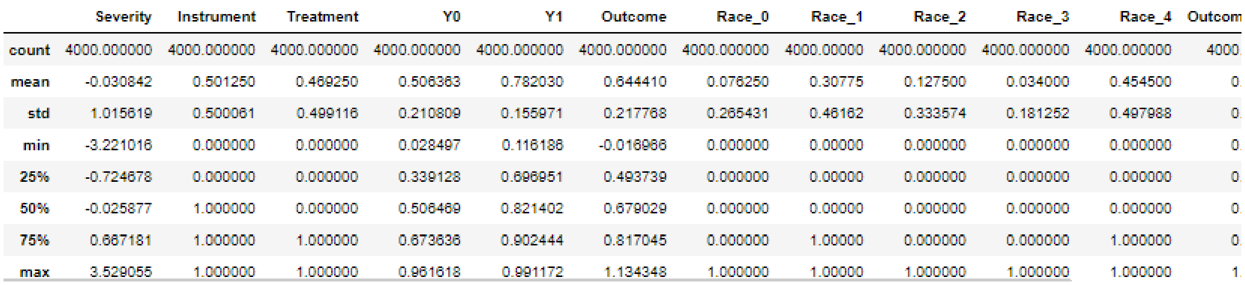
The dataset contains 4,000 entries with 13 columns, representing simulated attributes for causal inference analysis. Key variables include Severity (continuous variable indicating disease severity), Instrument (binary), Treatment (indicating whether treatment was administered), and

Outcome (the observed result). Potential outcomes (Y_0 and Y_1) represent untreated and treated states, respectively. Race is encoded using one-hot vectors across five categories: White (Race_0), Black (Race_1), Hispanic (Race_2), Asian (Race_3), and Other (Race_4). Additionally, Outcome_prob provides a clipped probability version of the observed outcome, while Outcome_binary is a binary outcome (1 = success, 0 = failure), obtained by thresholding Outcome_prob at 0.5. The dataset is well-structured with no missing values.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4000 entries, 0 to 3999
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Severity        4000 non-null   float64
 1   Instrument      4000 non-null   int32
 2   Treatment       4000 non-null   int32
 3   Y0              4000 non-null   float64
 4   Y1              4000 non-null   float64
 5   Outcome         4000 non-null   float64
 6   Race_0          4000 non-null   int32
 7   Race_1          4000 non-null   int32
 8   Race_2          4000 non-null   int32
 9   Race_3          4000 non-null   int32
 10  Race_4          4000 non-null   int32
 11  Outcome_prob    4000 non-null   float64
 12  Outcome_binary  4000 non-null   int32
dtypes: float64(5), int32(8)
memory usage: 281.4 KB
```

Figure 4 Check Missing Values



	Severity	Instrument	Treatment	Y0	Y1	Outcome	Race_0	Race_1	Race_2	Race_3	Race_4	Outcom
count	4000.000000	4000.000000	4000.000000	4000.000000	4000.000000	4000.000000	4000.000000	4000.00000	4000.000000	4000.000000	4000.000000	4000.
mean	-0.030842	0.501250	0.469250	0.506363	0.782030	0.644410	0.076250	0.30775	0.127500	0.034000	0.454500	0.
std	1.015619	0.500061	0.499116	0.210809	0.155971	0.217768	0.265431	0.46162	0.333574	0.181252	0.497988	0.
min	-3.221016	0.000000	0.000000	0.028497	0.116186	-0.016966	0.000000	0.00000	0.000000	0.000000	0.000000	0.
25%	-0.724678	0.000000	0.000000	0.339128	0.696951	0.493739	0.000000	0.00000	0.000000	0.000000	0.000000	0.
50%	-0.025877	1.000000	0.000000	0.506469	0.821402	0.679029	0.000000	0.00000	0.000000	0.000000	0.000000	0.
75%	0.667181	1.000000	1.000000	0.673636	0.902444	0.817045	0.000000	1.00000	0.000000	0.000000	1.000000	0.
max	3.529055	1.000000	1.000000	0.961618	0.991172	1.134348	1.000000	1.00000	1.000000	1.000000	1.000000	1.

Figure 5 Basic Statistics of Dataset

Severity follows an approximately standard normal distribution with a mean close to 0 and most values concentrated within the range of [-3, 3]. Y_0 (untreated potential outcome) shows a skewed distribution with values primarily between 0.3 and 0.7, reflecting the negative impact of severity. In contrast, Y_1 (treated potential outcome) is skewed towards higher values, mostly between 0.7 and 0.9, indicating the positive effect of treatment. Outcome (observed outcome) combines values from Y_0 and Y_1 with added noise, resulting in a broader distribution ranging from 0.3 to 1.0. Similarly, Outcome_prob, a clipped version of Outcome, ensures values remain within [0, 1] and aligns closely with the distribution of Outcome. These patterns validate the data generation process and highlight the treatment's impact on outcomes.

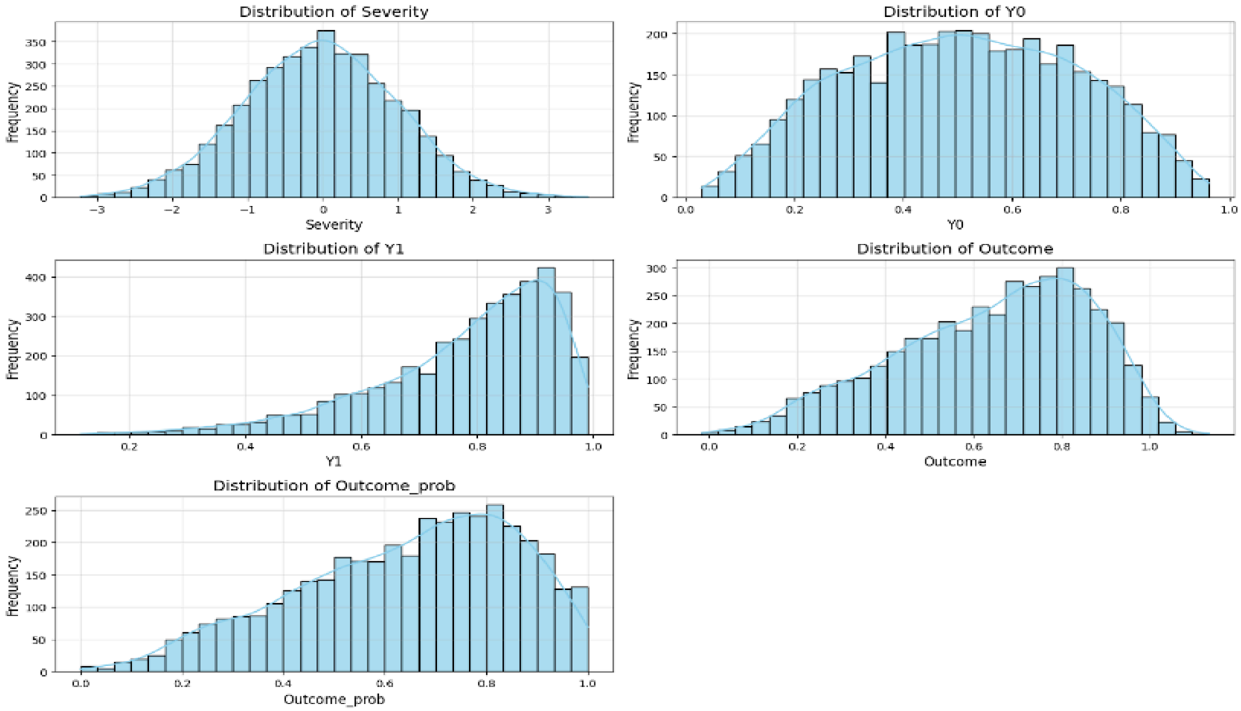


Figure 6 More Feature Distributions

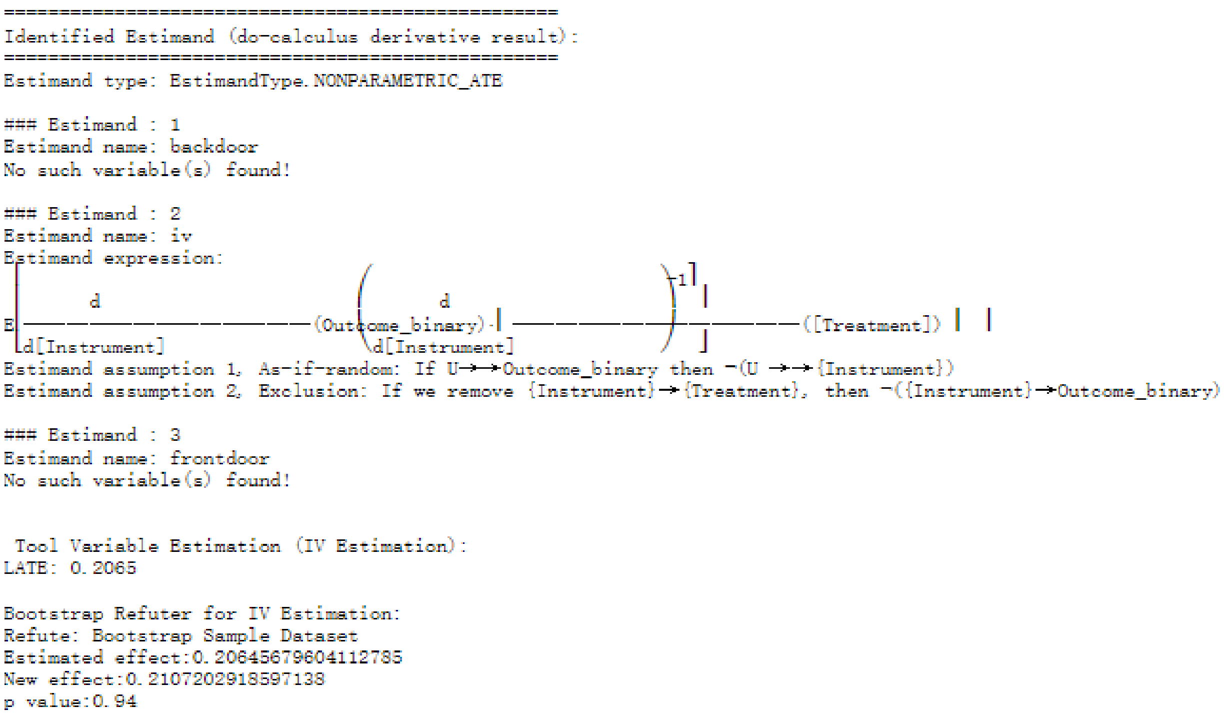
8. Causal Inference

8.1. Do-Calculus Analysis

Do-calculus is employed to identify the causal effect of treatment (Treatment) on the binary outcome (Outcome_binary) using a causal graph and an instrumental variable (Instrument). The backdoor and frontdoor criteria were found inapplicable due to the presence of unobserved confounders (U) and the absence of mediators. Consequently, the instrumental variable method was applied to derive the causal estimation, isolating the causal effect by leveraging the influence of the instrument on treatment while assuming it has no direct effect on the outcome. The derived estimation is expressed as:

$$Estimated = E[\frac{\partial outcome\_binary}{\partial Instrument} \cdot (\frac{\partial Treatment}{\partial Instrument})^{-1}]$$

The instrumental variable method estimated the Local Average Treatment Effect (LATE) as 0.2065, suggesting a moderate positive effect of treatment on patient outcomes. Bootstrap validation using 100 resampled datasets yielded an average effect of 0.2107 ($p = 0.94$), reinforcing the robustness of the estimate. Rather than highlighting statistical significance, these results underscore the reliability of IV methods in contexts with unobserved confounding - especially in healthcare, where inaccurate causal inference can amplify existing treatment disparities.



```
==================================================
Identified Estimand (do-calculus derivative result):
==================================================
Estimand type: EstimandType.NONPARAMETRIC_ATE

### Estimand : 1
Estimand name: backdoor
No such variable(s) found!

### Estimand : 2
Estimand name: iv
Estimand expression:
 ⎡                                    ⎛                          ⎞-1⎤
 ⎢       d                            ⎜       d                  ⎟  ⎥
E⎢─────────────────(Outcome_binary)⋅⎢────────────────────────([Treatment])⎥ ⎥
 ⎣d[Instrument]                       ⎝d[Instrument]             ⎠  ⎦
Estimand assumption 1, As-if-random: If U→→Outcome_binary then ¬(U →→{Instrument})
Estimand assumption 2, Exclusion: If we remove {Instrument}→{Treatment}, then ¬({Instrument}→Outcome_binary)

### Estimand : 3
Estimand name: frontdoor
No such variable(s) found!


 Tool Variable Estimation (IV Estimation):
LATE: 0.2065

Bootstrap Refuter for IV Estimation:
Refute: Bootstrap Sample Dataset
Estimated effect:0.20645679604112785
New effect:0.2107202918597138
p value:0.94
```

Our simulated causal graph further reveals that race affects both treatment allocation and health outcomes, through both direct and indirect pathways (Race → Treatment, Race → Outcome_binary). This pattern reflects structural inequities observed in real-world healthcare systems, where racial background can shape access to care and likelihood of receiving effective treatment.

To reflect more realistic confounding, we introduced a latent variable $U \sim N(0, 1)$ representing unmeasured factors like socioeconomic status or healthcare access. U simultaneously influences both treatment probability and potential outcomes Y_0 and Y_1, creating a backdoor path from Treatment to Outcome_binary. If left unadjusted, this confounding may distort estimated effects and lead to biased clinical interpretations. Ethically, failing to account for such latent confounders risks legitimizing inequities under the appearance of objective prediction.

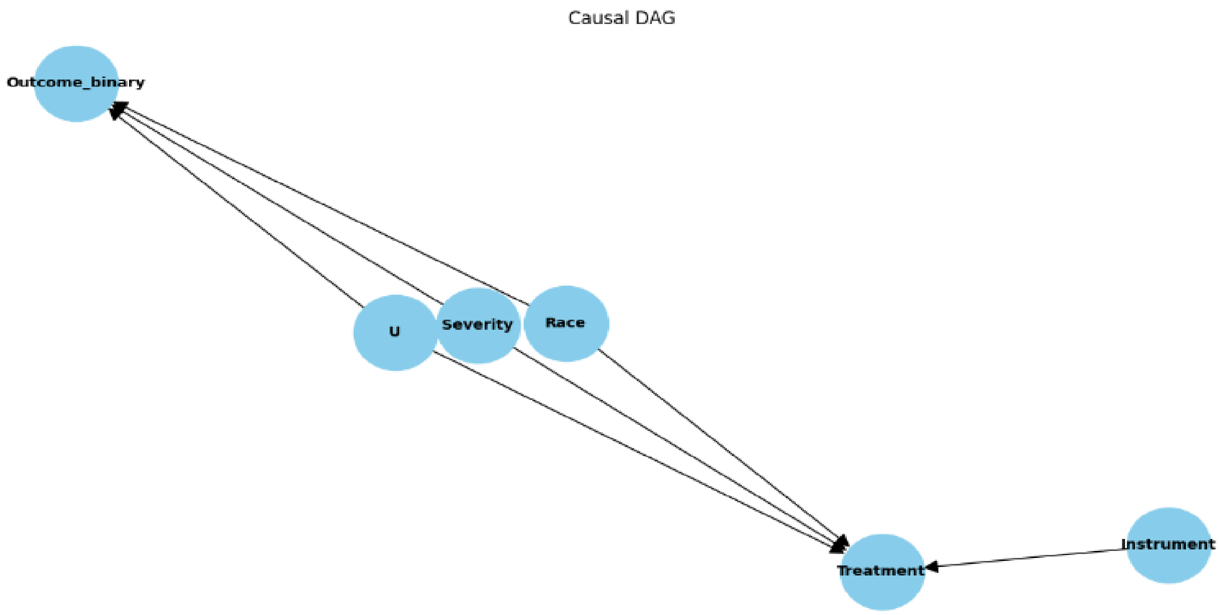


Figure 7 Causal DAG

8.2. Propensity Score Inspection

The propensity scores range from 0.2 to 0.9, with no extreme values, and are primarily concentrated within the range of 0.5 to 0.6. The mean score is 0.57, with a standard deviation of 0.12, indicating a relatively compact and symmetric distribution. The shape of the distribution is reasonable, avoiding extreme values and making it suitable for subsequent causal inference analysis. The next step is to examine the overlap in propensity score distributions between the treatment and control groups to ensure the effectiveness of matching or weighting methods.

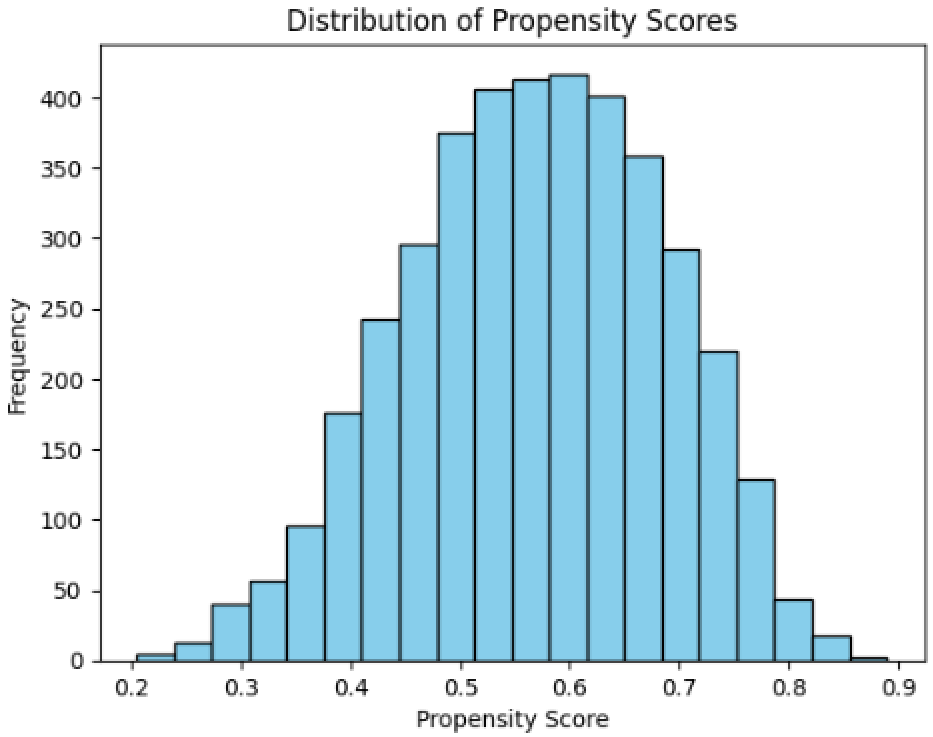


Figure 8 Propensity Scores Distribution Shows the Effectiveness of Weighting Methods

8.3. ATE Inspection

ATE represents the average causal effect of treatment on the outcome in the overall sample.

$$ATE = E[Y_1 - Y_0]$$

Where Y_1 represents the potential outcome when receiving treatment (Treated Potential Outcome). Y_0 represents potential outcome when no treatment.

The ATE value of 0.267 indicates that, on average, treatment improves patient outcomes by 26.7% points relative to the untreated baseline, under the potential outcomes framework, for syphilis patients in the dataset. This aligns with the study's objective of evaluating the causal impact of treatment while accounting for factors such as race, severity, and unobserved confounders. The moderate positive effect suggests that treatment is beneficial overall, but the analysis also highlights the need to explore how this effect varies across racial groups and other subpopulations, as structural inequities may influence both access to treatment and its effectiveness. While the average treatment effect is positive, assuming uniform benefit across all patients risks obscuring underlying disparities. Ethical practice requires disaggregating effects to ensure that no group disproportionately benefits or is left behind in treatment access.

8.4. CATE Inspection

The CATE results reveal that treatment has a positive effect across all racial groups, with values ranging from 0.2599 for White patients to 0.2733 for Asian patients. This indicates that treatment improves outcomes consistently across different racial groups, though slight differences exist. Asian patients exhibit the highest treatment effect, while White patients show the lowest. These findings suggest that while treatment is generally effective for all groups, there may be subtle disparities in its impact, potentially reflecting variations in healthcare access or other systemic factors. Even though all groups benefit from treatment, the slightly lower CATE for White patients could mean that, if resources are scarce, an algorithmic model might suggest deprioritizing them, raising concerns about equity in resource allocation

$$CATE = \frac{1}{N_{subgroup}} \sum_{i \in subgroup} (Y^{1}(i) - Y^{0}(i))$$

Where $N_{subgroup}$represents races, $Y^{1}(i)$, the potential outcome of i-th subgroup when receiving treatment.

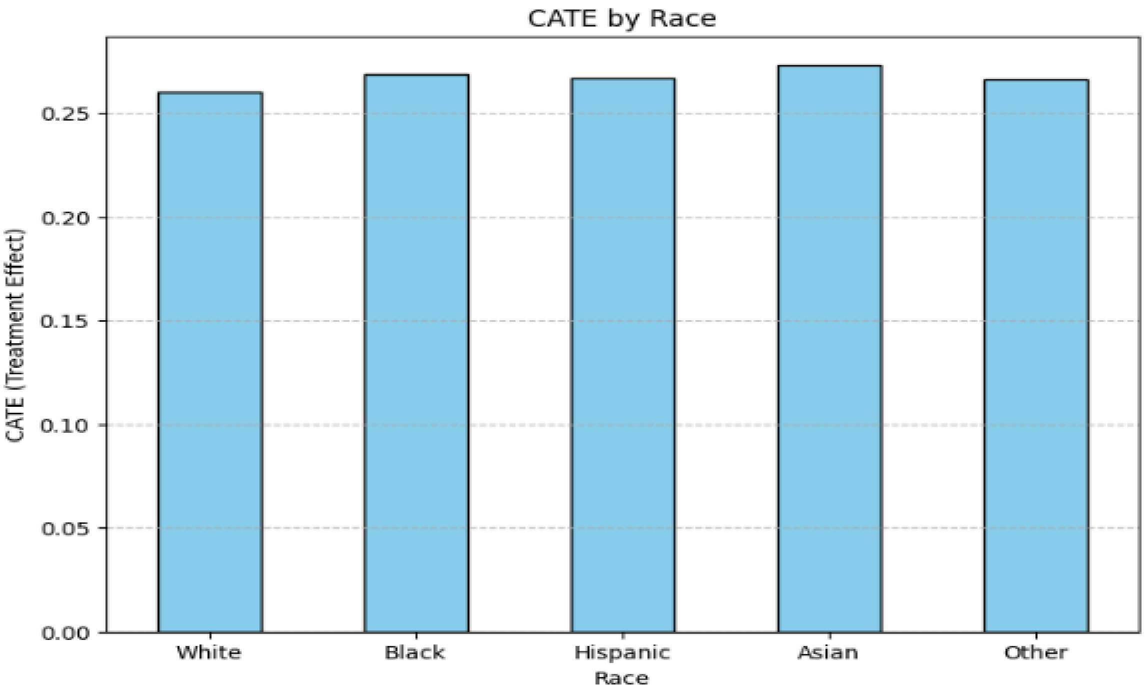


Figure 9 CATE by Race

8.5. IPTW

The Weighted Average Treatment Effect (IPTW = 0.5328) indicates a strong positive impact of treatment on patient outcomes across all racial groups. While treatment effects are relatively consistent, minor variations exist - Asians show the highest ATE (0.5419), and Hispanics the lowest (0.5255) - potentially reflecting disparities in treatment allocation or response.

Notably, the unweighted ATE underestimates this effect, likely due to covariate imbalance. By reweighting samples to account for confounding, IPTW offers a more accurate estimate of the true causal effect, improving both validity and fairness evaluation. Without proper weighting, observational estimates might lead clinicians to falsely conclude that some groups respond less to treatment. This introduces the risk of algorithmically legitimized neglect in healthcare prioritization.

$$IPTW\ Weight\ = \left\{\frac{1}{Propensity\ Score},\ if\ Treatment\ =\ 1;\ \frac{1}{1-Propensity\ Score},\ if\ Treatment\ =\ 0\right\}$$

The model used is the logistic regression model (Leskovec, Rajaraman, & Ullman, 2021, p. 407).

$P(Treatment\ =\ 1\ |\ Covariates)\ =\ \sigma(logit_{treatment})\ =\ \frac{1}{1+e^{-logit\_treatment}}$, where

$$logit_{treatment} = \beta_0 + \sum_{i=1}^{7} \beta_i \cdot X_i,\ i.e.,\ e^{-logit\_treatment}$$

$$= e^{-(0.00960988-0.216998*Race_0+0.049854*Race_1+0.101691*Race_2+0.070871*Race_3+0.00265*Race_4+0.403906*Severity+0.553076*Instrument)}$$

8.6. ITE Distribution

The Individual Treatment Effect (ITE) distribution by race is further investigated and reveals minor variations in treatment effects across racial groups. White, Black, Hispanic, and Other groups exhibit slightly negative average ITE values (-0.0009, -0.0191, -0.0168, and -0.0039, respectively),indicating that for these groups, the model predicts a limited or marginally negative average treatment benefit, which may be due to model bias, confounding, or real-world disparities. Conversely, Asian patients show a positive average ITE value (0.0161), suggesting that treatment has a modest positive effect for this group. These results highlight potential disparities in treatment effects among racial groups, where the benefits of treatment are not uniformly distributed. If clinical decisions are informed by ITE estimates, patients from groups with lower ITE, such as Black or Hispanic individuals in this model, may receive less aggressive treatment even if their clinical need is similar. This risks embedding statistical disparities into clinical judgment.

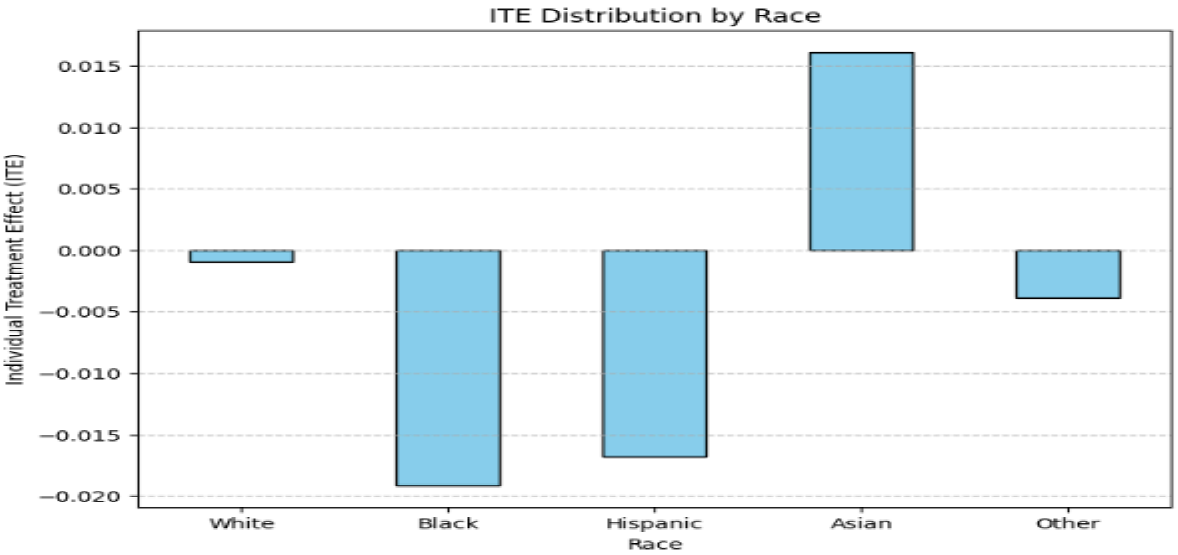


Figure 10 ITE Distribution by Race

8.7. Probability Discrepancy Analysis

Comparing interventional and observational probabilities provides insight into the causal effect of treatment. The overall interventional probability $P(Y \mid do(T = 1)) = 0.6421$ is slightly higher than the observational probability $P(Y \mid T = 1) = 0.6173$, suggesting that observational estimates may underestimate the true causal effect due to confounding. When stratified by severity bins, the contrast becomes more pronounced. Observational probabilities range from 0.8779 to 0.3457, whereas interventional estimates fall between 0.0878 and 0.0346. This reversal indicates that confounders likely inflate the observational effect across severity levels. Both curves show a declining trend, implying reduced treatment efficacy as severity increases. By adjusting for confounding, interventional estimates derived via do-calculus offer a more accurate and fair assessment of treatment impact. The decreasing treatment effect as severity increases raises ethical concerns: patients who are most ill may benefit least, yet they are often most in need. If algorithms reinforce this pattern, they could deprioritize high-risk patients, violating principles of medical justice.

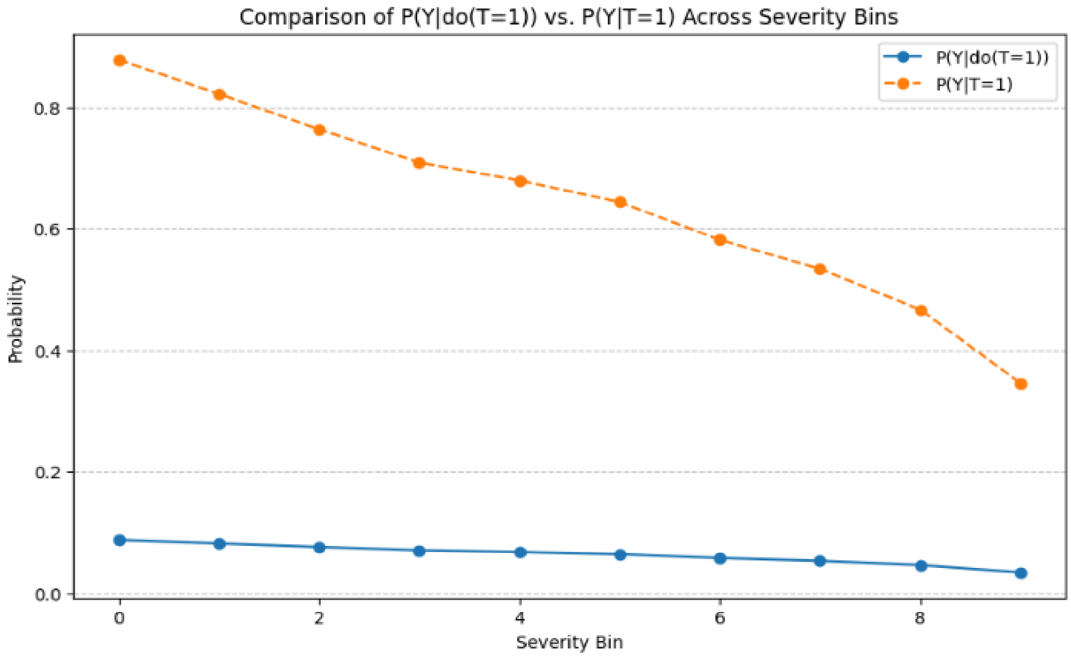


Figure 11 Comparison of Interventional and Observational Probability Across Severity Bins

9. Discussion

9.1. Counterfactual Fairness Analysis

To assess counterfactual fairness, we implemented a “race flipping” procedure that changes individuals’ race while holding all other features constant. For each original–flipped race pair, we computed the average deviation in predicted treatment probability, summarised in a fairness deviation matrix:

$$Deviation = P(Treatment = 1 \mid Original\ Race) - P(Treatment = 1 \mid Flipped\ Race)$$

$$Fairness\ Deviation\ Matrix_{i,j} = \frac{1}{N_{i,j}} \sum_{k=1}^{N_{i,j}} Deviation_k$$

where i represents the original race and j is the flipped race. $N_{i,j}$ is the number of samples which belong to the original subgroup i and flipped into j. $Deviation_k$ is the deviation of the k_{th} sample.

Results show that the model is most sensitive when flipping from "White" to "Other" (deviation = 0.0449), and least sensitive between "Black" and "Other" (deviation = 0.0048). While some deviations are small, others highlight potential bias in how the model assigns treatment likelihoods across races. These findings emphasize the ethical importance of counterfactual fairness: even minor shifts in high-stakes domains like healthcare can perpetuate systemic inequities. If predictions change solely due to race flipping, then race is influencing outcomes beyond clinical indicators. This demands accountability: developers must justify why race-sensitive predictions are acceptable, and clinicians must decide whether to override them. Future work should expand this analysis to intersectional identities (e.g., Race × Gender), following frameworks such as Valentine et al. (2023), to better capture compounding sources of unfairness.

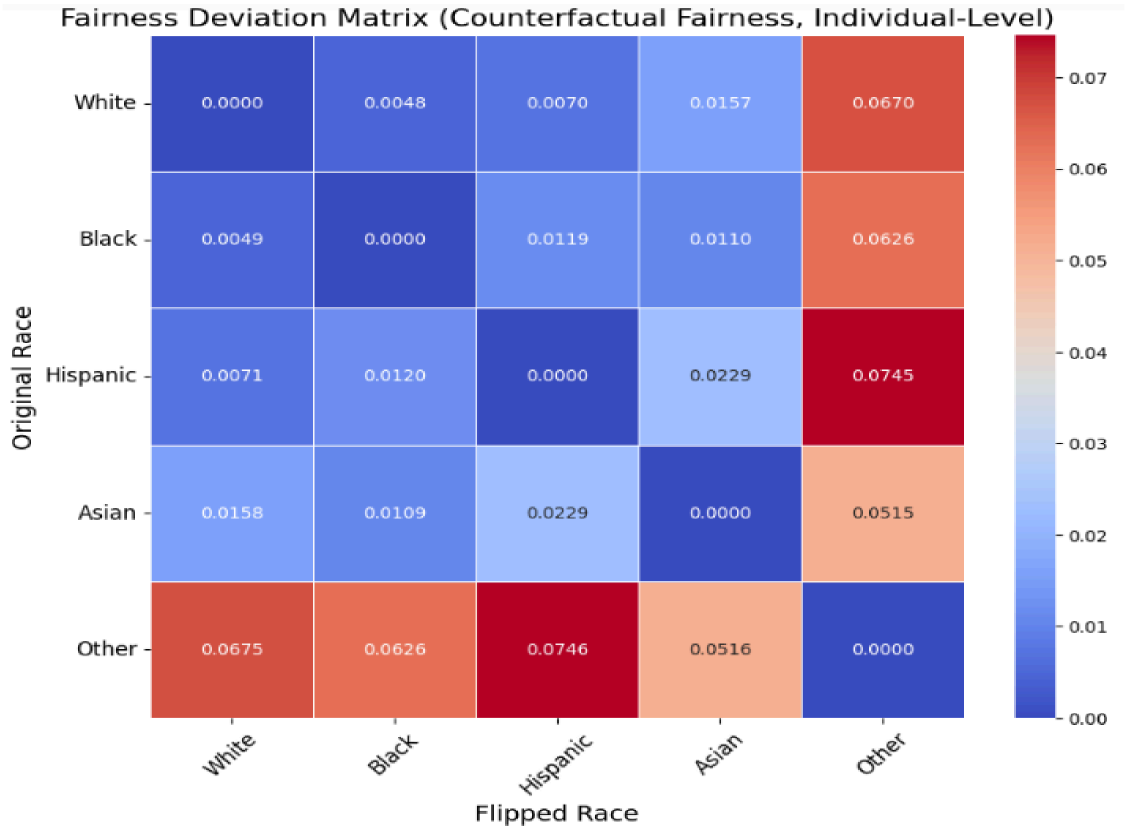


Figure 12 Counterfactual Fairness Deviation Matrix

9.2. Group Fairness Analysis

To assess group fairness, we evaluated four metrics - precision, true positive rate (TPR), false positive rate (FPR), and calibration error - across racial groups. These metrics help reveal whether the model systematically favors or disadvantages particular groups. The definitions are as follows:

$$Precision = \frac{TP}{TP+FP}$$

$$TPR = \frac{TP}{TP+FN}$$

$$FPR = \frac{FP}{FP+TN}$$

$$Calibration\ Error\ = \frac{1}{n}\sum_{i=1}^{n}\left|P_{true,\,i} - P_{pred,\,i}\right|$$

where TP represents true positive and FP false positive, $P_{true,\,i}$ represents the i-th sample's actual outcome probability. n is the number of samples.

Our results in Figure 13 show performance disparities across races. The model achieves highest precision (0.6786) and TPR (0.9048) for Asian patients, but also the highest FPR (0.6923), suggesting over-diagnosis risk. Conversely, White patients show the lowest precision (0.6071) and TPR (0.7169), indicating weaker performance. Black patients exhibit the lowest calibration error (0.0540), meaning predicted probabilities align more closely with actual outcomes for this group. These findings highlight that even when overall accuracy is high, fairness concerns remain - particularly when over-diagnosis, such as excessive treatment recommendations for certain groups, introduces its own form of harm. Ensuring equity in model performance requires not just optimizing for predictive power, but actively addressing group-level disparities. Higher accuracy for some groups, such as Asians, may misleadingly suggest fairness; however, elevated FPR implies overdiagnosis, which itself constitutes clinical harm. In fairness analysis, better metrics for one group do not always translate to better outcomes.

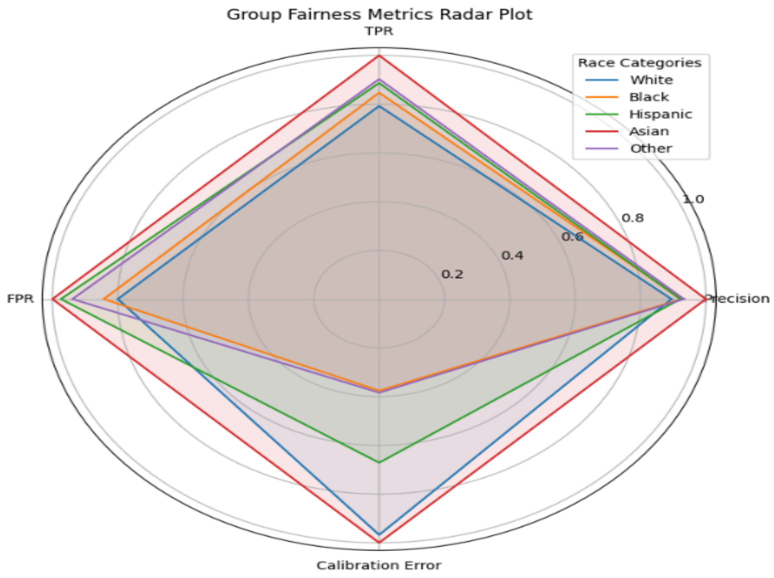


Figure 13 Group Fairness Metrics Radar Plot

The outcome-based fairness matrix in figure 14 measures prediction sensitivity to race flipping by quantifying changes in predicted outcomes across race pairs. The largest deviation occurs when flipping from "Black" to "Asian" (0.0959), suggesting the model's output is highly sensitive to this transition. Smaller deviations, such as from "Hispanic" to "Asian" (0.0689), reflect more stable predictions. These asymmetries indicate that certain race pairings exhibit higher disparity, pointing to structural biases in the model's learned associations.

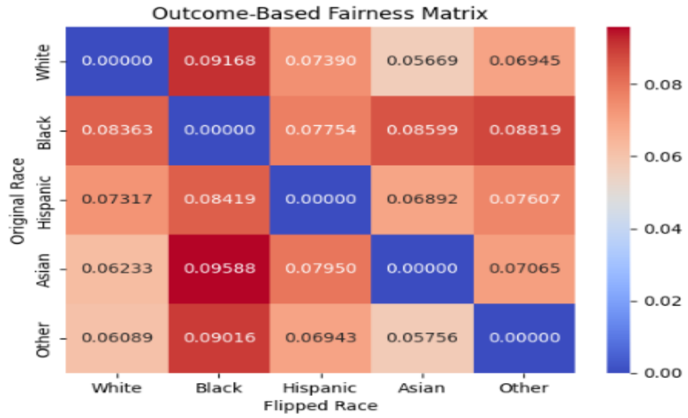


Figure 14 Outcome-Based Fairness Matrix Based on Race Flipped

The heatmap visualizations from figure 15confirm that while most deviations are small, some racial groups are consistently affected. For example, Asians exhibit the lowest mean absolute error (MAE = 0.0001) and disparity (0.01%), indicating high predictive accuracy. In contrast, the "Other" group shows the highest MAE (0.0286) and disparity (0.31%), suggesting poorer model calibration and potential underperformance.

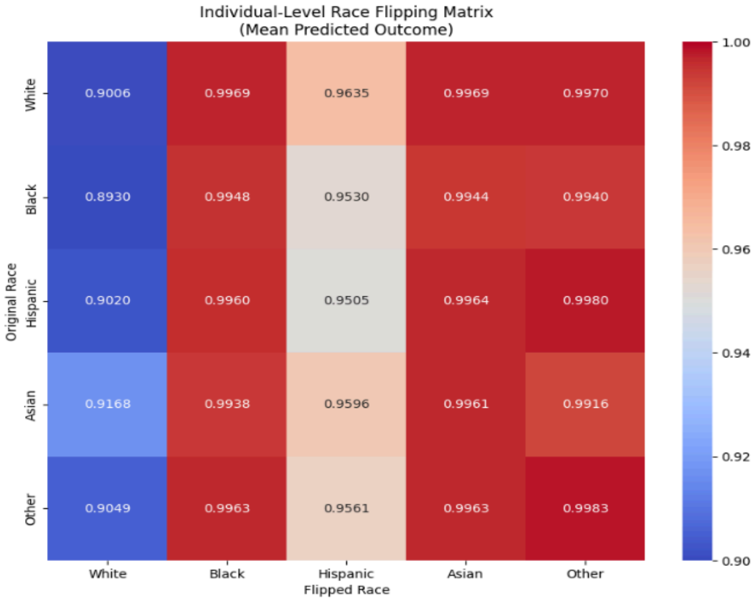
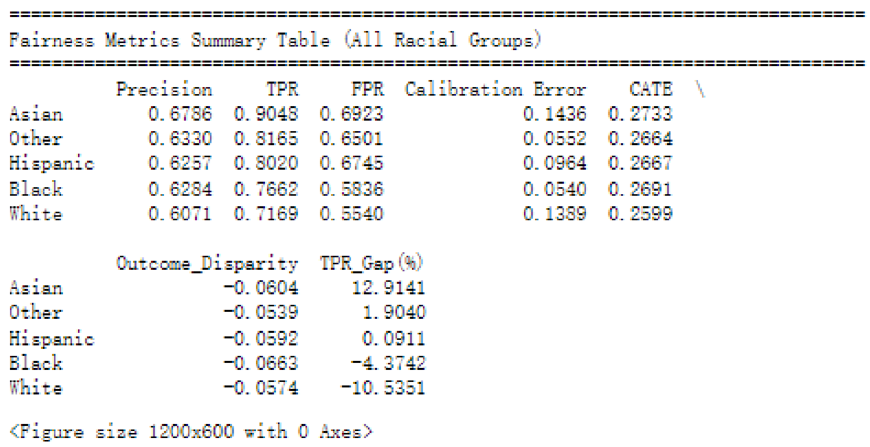


Figure 15 Individual-Level Race Flipping Matrix

At the end, we create a final summary table and aggregate key fairness metrics - precision, TPR, FPR, calibration error, and CATE - across groups, along with derived gaps (e.g., TPR gap, outcome disparity). These allow for a clearer comparison of fairness discrepancies. Ethically, even minor outcome deviations patterned by race may perpetuate unequal treatment access, reinforcing the need for bias-aware modeling and evaluation frameworks.



```
================================================================================
Fairness Metrics Summary Table (All Racial Groups)
================================================================================
          Precision     TPR     FPR  Calibration Error    CATE  \
Asian        0.6786  0.9048  0.6923             0.1436  0.2733
Other        0.6330  0.8165  0.6501             0.0552  0.2664
Hispanic     0.6257  0.8020  0.6745             0.0964  0.2667
Black        0.6284  0.7662  0.5836             0.0540  0.2691
White        0.6071  0.7169  0.5540             0.1389  0.2599

          Outcome_Disparity  TPR_Gap(%)
Asian               -0.0604     12.9141
Other               -0.0539      1.9040
Hispanic            -0.0592      0.0911
Black               -0.0663     -4.3742
White               -0.0574    -10.5351

<Figure size 1200x600 with 0 Axes>
```

Figure 16 Fairness Metrics Summary Table of All Racial Group

The results reveal significant variations in fairness metrics across racial groups. Asians exhibit the highest TPR (0.9048) and precision (0.6786), indicating the model performs most effectively for this group. However, Asians also show the highest FPR (0.6923), suggesting a higher rate of false positives. Conversely, Whites have the lowest TPR (0.7169) and precision (0.6071), highlighting weaker model performance for this group. Outcome disparity values are relatively small but vary slightly, with Blacks showing the largest disparity (-0.0663), indicating a gap between observed and predicted outcomes. The TPR gap analysis further highlights disparities, with Asians outperforming the average by 12.91%, while White patients underperform by 10.54%.. These findings underscore the need for targeted adjustments to address fairness gaps and ensure equitable model performance across all racial groups.

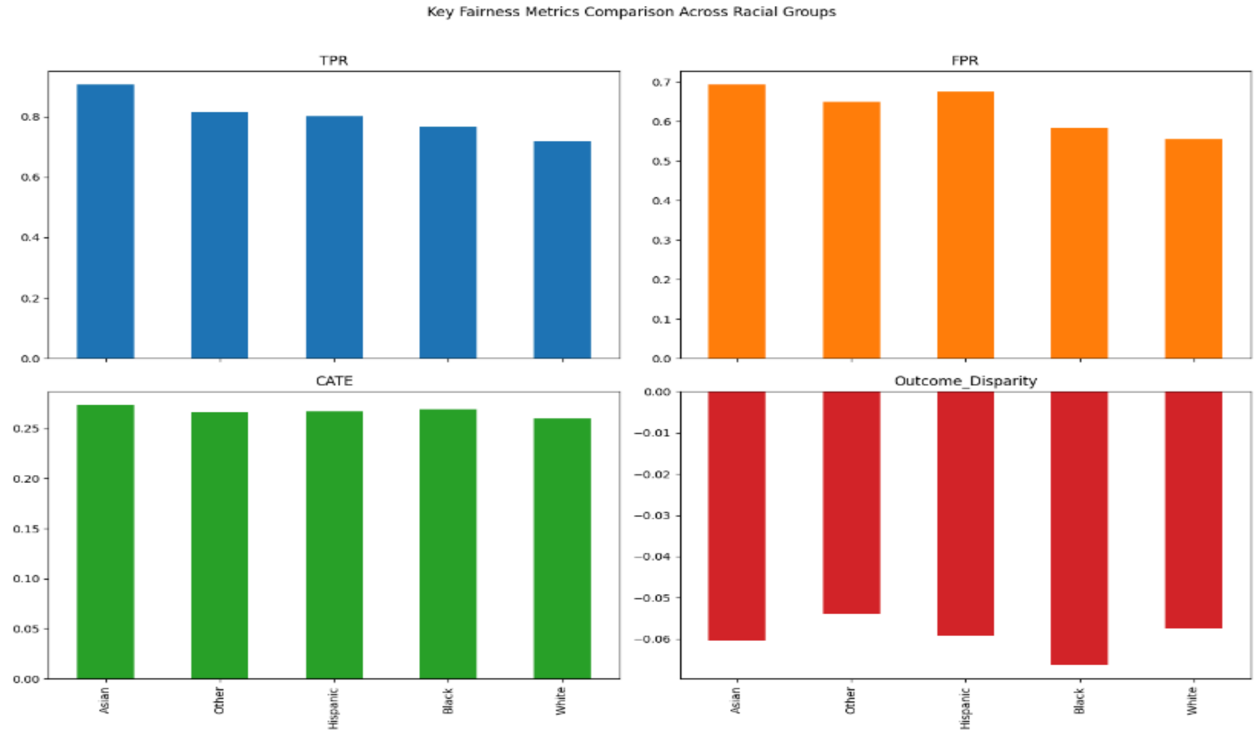


Figure 17 Comparison of Key Fairness Metrics of All Racial Groups

9.3. Intersectionality and accountability

Our analysis didn't explicitly simulate intersectional identities - such as race × gender or race × socioeconomic status - but the disparities we observed across racial groups still underscore the need for a more intersectional framework. For instance, although Asians had the highest true positive rate (TPR), they also showed the highest false positive rate (FPR), suggesting a risk of overdiagnosis. Meanwhile, Black patients exhibited the greatest outcome disparity, potentially pointing to systemic under-treatment (such as false negatives). These results may understate deeper forms of compounded disadvantage that emerge at the intersection of multiple social identities. If audits consider race and gender in isolation - for example, conducting one test for racial bias and another for gender bias - regulators may falsely label the model as “fair,” while in practice, it could still fail patients with intersecting marginalised identities. This would increase clinical risk for patients, through misdiagnosis, mistreatment, or omission of care, exposing the hospital to potential legal liability under equality and anti-discrimination law.

As Valentine et al. (2024) emphasise in their schizophrenia diagnosis study, “building fair AI tools for the healthcare space requires recognizing the intersectionality of sociodemographic factors.” Their study shows that these sensitive variables interact in complex ways to shape diagnosis outcomes - for instance, high-SES (socio-economic status) Black men faced higher schizophrenia diagnosis odds than low-SES counterparts, contradicting the common assumption that high SES is universally protective against such mental disorders. We therefore declare that fairness metrics that evaluate sensitive features - such as race, gender, SES - in isolation risk masking true harm faced by individual patients with intersecting identities, whether in healthcare or any other domain, such as recruitment and education.

Moreover, the findings reaffirm the importance of accountability in healthcare AI, especially when dealing with sensitive decisions affecting patients' treatment access. Novelli et al. 's (2023) conception of accountability as a relation of answerability provides a valuable lens through which to view the dilemmas in our study. In particular, their insistence that accountability depends on authority recognition, the ability to interrogate decisions, and limitations on arbitrary power sheds light on the shortcomings of current AI governance in healthcare.

However, assigning accountability in healthcare isn't straightforward. Developers may claim their algorithms are technically neutral, blaming biased training data or poor implementation for skewing predictions. On the other hand, hospitals - i.e. the deployers - may argue they lack the expertise to interrogate complex models, especially if the algorithm is a "black box" (completely opaque) and lacks sufficient interpretability, making it difficult for agents to interrogate. Patients, the most impacted stakeholders, are often least equipped to demand explanations, making them a weak forum in the accountability relationship. This situation mirrors the "many hands" problem described by Novelli et al., where accountability is diluted among multiple actors, none of whom feel wholly responsible. The result is an accountability vacuum, where harmful outcomes can persist without redress.

Furthermore, the fact that predictive models often produce outcomes without offering insight into their underlying reasoning exacerbates model opacity. Knowing what the model predicted is not enough, stakeholders must also understand the logic behind the model's prediction, understanding why there are racial differentials in the model's predictions. Without this insight, it becomes difficult to detect whether biased patterns, such as relying on race-linked proxies, are influencing predictions. For instance, if preliminary testing reveals that, across training data, high-risk predictions for White patients are driven by clinical variables (e.g. blood pressure, symptom history), while predictions for Black patients rely more on access-related variables (e.g. number of missed appointments), this would indicate a structural bias in the model's logic. Yet such patterns are difficult to identify in black-box systems that offer no explanation for their internal reasoning. The inability to scrutinise these mechanisms before deployment complicates the assignment of responsibility: developers may claim their models are functionally correct, while deployers may argue they lacked the tools to interpret or audit predictions. Hence, the absence of explanation hinders proactive accountability and contributes to the accountability vacuum described by Novelli et al. (2023).

The lack of precise accountability mechanisms has direct consequences for fairness. Without the ability to interrogate or rectify algorithmic outputs, existing disparities can be perpetuated by systems that claim to be objective. Novelli et al. warn of "accountability gaps," where no one is held responsible, and "accountability surpluses," where burdensome bureaucratic procedures are put in place without meaningful enforcement. Our study suggests that fairness metrics alone are insufficient to prevent such outcomes. Without robust mechanisms to tie predictions to consequences - and hold someone answerable when harm occurs - models risk entrenching the inequities they are meant to solve.

Healthcare, thus, presents a robust case for proactive accountability. Predictive tools influence treatment decisions with serious consequences for patients' lives, so accountability, compliance, oversight, reporting, and enforcement goals must be institutionally grounded. That means not only auditing models before deployment, but creating pathways for clinicians to interrogate predictions and for patients to contest outcomes. As Novelli et al. argue, accountability is not a single act but a structural relationship. Without it, fairness would remain a distant ideal.

10. Ethical and Policy Recommendations

Building fair algorithmic systems in healthcare demands structural and institutional reform, rather than merely higher technical accuracy. Based on our findings of racial disparities in simulated syphilis treatment allocation, we offer several ethical and policy recommendations grounded in both the causal results of our analysis and the literature cited prior.

First, fairness audits must move beyond static group metrics. As shown in our outcome disparity and fairness deviation matrices, models may appear fair on average while still failing patients with intersecting marginalised identities. Hence, policy frameworks should mandate intersectional auditing, integrating dimensions such as race × gender or race × SES, and ensuring that no subgroup combination's - such as low-SES Black women - systematic discrimination go unrecorded.

Second, accountability structures must be institutionalized. As Novelli et al. argue, fairness without answerability is hollow. Hospitals deploying predictive systems should establish formal *accountability chains* identifying: (i) who builds and validates models, (ii) who oversees their clinical deployment, and (iii) how patient concerns or adverse events trigger audits or redress. This includes creating designated *fairness officers* within hospital ethics boards and mandating *pre-deployment interpretability reports* for AI systems used in clinical triage or diagnosis.

Third, model explainability must support these accountability efforts. Kiyasseh et al. (2022) demonstrate with the TWIX (training with explanations) framework, where the model's justifications accompany predictions, improved transparency and performance. This would occur before deployment, but after development. Deployers (also developers) could review these explanations across demographic subgroups, exploring the model's rationale, to identify biased learning patterns. For example, preliminary explanations may show that the model's predictions for Black patients rely heavily on access-related proxies (e.g. missed appointments or zip codes), while predictions for others cite clinically relevant variables. This would signal skewed internal logic that would otherwise go undetected in models without explainability features. Furthermore, recent advances in large language models (LLMs) make installing explainability features increasingly feasible, enabling interpretable explanations to close the gap between statistical logic and clinical understanding. This would make it more difficult for deployers (i.e. clinicians) to absolve themselves of responsibility on the basis of opacity or technical complexity. As Novelli et al. (2023) argue, accountability requires the ability to interrogate decisions and limit arbitrary power, and LLMs would make it easier for non-technical agents to conduct interrogation. Embedding explainability into the pre-deployment pipeline is thus a vital component of proactive accountability, allowing issues to be identified and resolved before patients are impacted.

Fourth, data provenance and labeling must be reformed. As highlighted in *Fairness and Machine Learning*, biases often originate in flawed target definitions and proxy variables (Barocas, Hardt, & Narayanan, 2023). For example, using treatment cost as a proxy for health need - as in the case cited by Obermeyer et al. - may encode racial inequities into ostensibly neutral models. We recommend that public health institutions develop *labeling guidelines* that distinguish between structural proxies (e.g., cost, access) and clinically grounded outcomes (e.g., disease progression, biomarker change). These guidelines should be informed by affected communities, clinical ethicists, and health equity researchers.

Fifth, structural interventions must accompany algorithmic ones. Fair predictions cannot compensate for unfair systems. Following Barocas et al.’s critique of "tech fixes," we recommend that any use of AI in healthcare be paired with broader reforms - such as expanding healthcare access, supporting multilingual services, and investing in minority-serving institutions. Where algorithmic triage is used, healthcare providers should also implement *fairness-sensitive operational policies*, e.g., holding resource allocations for re-review if predicted disparities exceed certain thresholds.

Last, participatory co-design is essential. To mitigate epistemic injustice, patients from marginalized communities must be engaged throughout the model development pipeline. This includes community consultations during data collection, review boards to assess fairness criteria, and opt-out mechanisms allowing patients to reject algorithmic recommendations in favor of clinician override.

Together, these recommendations reflect a shift from fairness as a statistical property to fairness as a relational, institutional, and reparative obligation - especially in domains like healthcare, where decisions profoundly affect life outcomes. Future work should explore how these recommendations can be operationalised in regulatory standards, and how simulation-based ethical research can complement real-world EHR studies to shape safer, fairer clinical AI systems.

11. Limitations and Future Research

While this study offers a simulation-based exploration of structural race bias in syphilis treatment, several limitations must be acknowledged. First, although synthetic data allows us to isolate causal mechanisms in a controlled environment, it cannot replicate the full complexity of real-world healthcare. Social realities - such as implicit provider bias, under-resourced minority clinics, or patient mistrust rooted in historical injustices like the Tuskegee Syphilis Study - are not easily captured in code. As a result, the external validity of our fairness findings is inherently constrained. Second, while we examine treatment disparities by race, our analysis does not fully capture the sociotechnical dynamics of accountability. In real-world AI systems, fairness involves more than error rates or calibration - it also demands clarity about who designs the model, who deploys it, and who is responsible when harm occurs. These questions of "answerability," as discussed by Novelli et al. (2023), are absent in most technical pipelines and remain underexplored in our simulation. Third, our fairness analysis is limited to single-axis group comparisons, focusing only on race. In practice, algorithmic harms often emerge at the intersection of race, gender, socioeconomic status, and disability. Ignoring these compounded identities risks underestimating real-world disparities. Fourth, our evaluation is static. It does not account for how model deployment could shift provider behaviour, affect patient trust, or introduce feedback loops that reinforce bias over time. In reality, AI systems evolve within institutions, often entrenching existing power structures. Nonetheless, we argue that these limitations do not make the problem speculative. On the contrary, the patterns we simulate such as race-influenced treatment probabilities and proxy bias, mirror dynamics already seen in commercial algorithms and healthcare settings. Disparities in insurance status, language access, or incomplete patient records often serve as indirect channels through which race affects clinical decision-making, even when models exclude race as a feature. Algorithmic harm, in this sense, is not a hypothetical future but an unfolding present.

Therefore, future work should validate simulated findings using real-world electronic health record (EHR) data with rich sociodemographic labels and longitudinal outcomes. Develop accountability-aware modelling pipelines that define who is answerable and under what standards. Move beyond average treatment effects and include intersectional fairness metrics (e.g., Race × Gender). Incorporate participatory co-design, ensuring that marginalised patients can contest or influence algorithmic decisions. Embed fairness audits into institutional governance structures such as hospital ethics boards or regulatory bodies. Fairness must be treated not as a post hoc performance metric but as a structural obligation - especially in healthcare, where predictive decisions can mean the difference between care and neglect.

12. Conclusion

This study investigated how structural race bias can shape treatment decisions and outcomes in healthcare, using syphilis as a case study under controlled simulated conditions. By employing causal inference methods - such as do-calculus, inverse probability treatment weighting (IPTW), and individual treatment effect (ITE) analysis, we quantified how race can influence both treatment allocation and effectiveness. Fairness evaluations based on counterfactual flipping, group-level metrics (TPR, FPR, calibration error), and intersectional disparities further revealed that predictive models, even when trained on synthetic data, can encode unequal treatment recommendations across racial lines.

Our findings confirm that treatment is beneficial overall, as shown by consistently positive ATE and IPTW estimates. However, model performance varied significantly by race. Asian patients exhibited both the highest true positive rate and highest false positive rate, indicating potential overdiagnosis risks. Conversely, White patients had the weakest precision, and Black patients exhibited the greatest gap between observed and predicted outcomes. These patterns suggest that algorithmic bias in healthcare may not always manifest as outright denial of treatment, but rather as subtle mismatches in prediction accuracy and calibration across groups.

Importantly, counterfactual fairness analysis revealed that altering only the race attribute could significantly shift treatment probability predictions - demonstrating that the model was not truly “race-neutral.” This sensitivity shows how predictive tools can silently reproduce historical inequities under the appearance of technical objectivity. When unchallenged, such patterns risk legitimising disparities as scientific outputs rather than social constructs.

Yet some may argue that concerns raised by a synthetic dataset are speculative. We argue the opposite: this simulation reflects plausible mechanisms grounded in well-documented healthcare disparities. Historical injustices such as the Tuskegee Syphilis Study did not begin with explicit malice, but with passive, structural neglect rationalised by “clinical norms.” Today’s algorithms - if left unchecked - risk following the same trajectory, quietly embedding systemic bias into medical decisions that affect real patients. Even when race is not explicitly coded, real-world proxies such as insurance status, language barriers, or EHR completeness can carry embedded racial correlations, leading to similar disparities as those observed in our simulation.

From an ethical perspective, this study affirms that fairness in healthcare AI is not a secondary concern or post hoc audit metric - it must be a first-order design constraint. Guided by the sociotechnical framing in *Fairness and Machine Learning* and the answerability-based governance framework of Novelli et al. (2023), we advocate for proactive accountability: institutions must build in mechanisms for contestability, transparency, and recourse, rather than waiting for harm to materialise.

Future research should prioritise three directions: validating these findings using real-world electronic health record (EHR) data with rich sociodemographic attributes; expanding fairness evaluations to include intersectional identities such as race × gender or race × SES; and embedding audits into institutional accountability frameworks that empower both clinicians and patients. Only through such multidisciplinary, justice-oriented approaches can we ensure that predictive healthcare systems not only perform well, but also treat all patients equitably, transparently, and responsibly.

13. Bibliography

Asaria, M. (2024). *How AI could revolutionise NHS healthcare*. LSE Politics and Policy. https://blogs.lse.ac.uk/politicsandpolicy/how-ai-could-revolutionise-nhs-healthcare/

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. Cambridge, MA. https://fairmlbook.org

Centers for Disease Control and Prevention. (2023). *Primary and secondary syphilis - Rates of reported cases by race/Hispanic ethnicity, age group, and sex, United States*. https://www.cdc.gov/sti-statistics/data-vis/table-syph-ps-rates-caseagesex.html

Grant, C. (2022). Algorithms are making decisions about health care, which may only worsen medical racism. American Civil Liberties Union. https://www.aclu.org/news/privacy-technology/algorithms-in-health-care-may-worsen-medical-racism

Kiyasseh, D., Laca, J., Haque, T. F., Otiato, M., Miles, B. J., Wagner, C., Donoho, D. A., Trinh, Q.-D., Anandkumar, A., & Hung, A. J. (2023). Human visual explanations mitigate bias in AI-based assessment of surgeon skills. *npj Digital Medicine, 6*(1), 1–15. https://doi.org/10.1038/s41746-023-00766-2

Krasanakis, E., & Papadopoulos, S. (2024). Towards standardizing AI bias exploration. *arXiv*. https://doi.org/10.48550/arXiv.2405.19022

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2021). *Mining of massive datasets* (3rd ed., p. 407). Cambridge University Press.

Mittermaier, M., Raza, M.M. & Kvedar, J.C. Bias in AI-based models for medical applications: challenges and mitigation strategies. *npj Digit. Med.* 6, 113 (2023). https://doi.org/10.1038/s41746-023-00858-z

Novelli, C., Taddeo, M., & Floridi, L. (2023). Accountability in artificial intelligence: What it is and how it works. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00268-1

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Sjoding, M. W., Dickson, R. P., Iwashyna, T. J., Gay, S. E., & Valley, T. S. (2020). Racial bias in pulse oximetry measurement. *New England Journal of Medicine, 383*(25), 2477–2478. https://doi.org/10.1056/NEJMc2029240

Valentine, A. A., Charney, A. W., & Landi, I. (2023). Fair machine learning for healthcare requires recognizing the intersectionality of sociodemographic factors: A case study. *arXiv*. https://arxiv.org/pdf/2407.15006

Zhou, Z. (2016). *Machine learning* (pp. 58–59). Tsinghua University Press.