# Addressing Bias in Healthcare Algorithms

Seth Hall

2024-11-12

## Introduction

In an age where technological advancements promise to redefine how we approach healthcare, predictive algorithms stand out as a beacon of innovation. These systems hold the potential to optimize resources, enhance diagnostics, and prioritize treatments with unprecedented efficiency. However, beneath this promising surface lies a pressing concern: the ability of these algorithms to perpetuate, or even amplify, long-standing systemic biases. How can healthcare innovation address inequities rather than exacerbate them? This duality—innovation marred by inequity—presents a significant challenge to the equitable delivery of healthcare. In this project, I expand upon my midterm analysis to delve deeper into a predictive model that exhibits systemic racial bias against Black patients. Drawing upon the seminal work "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations" by Obermeyer et al., I conduct a rigorous evaluation of the model's methodology, present a novel analysis using R, and explore the ethical and policy implications of addressing algorithmic bias.

## Summary of Methods

The model studied by Obermeyer et al. aimed to predict healthcare needs using historical healthcare spending as a proxy for health risks. At first glance, this approach seems logical: healthcare spending can reflect the intensity of medical needs. However, this reliance on spending as a proxy introduced systemic inequities into the model's predictions. Black patients, who often face socioeconomic barriers limiting their access to healthcare, were systematically underestimated in terms of their actual health needs. To evaluate the fairness of the model, the researchers applied three key statistical measures: statistical parity, equalized odds, and disparate impact. Statistical parity assessed whether patients across racial groups had an equal likelihood of being flagged as high-risk. The model failed this test, revealing that Black patients were less likely to be identified as high-risk compared to their White counterparts, even with similar health profiles. Equalized odds provided further evidence of disparity by showing unequal error rates, with Black patients more frequently under-classified as low-risk. Finally, disparate impact analysis quantified the inequity, demonstrating that Black patients received risk scores that were, on average, 48% lower than those of White patients with comparable needs. These findings illuminated the implicit bias embedded within the model and its profound implications for equitable healthcare delivery.

### Novel Analysis

To validate and expand upon Obermeyer et al.'s findings, I conducted a simulation study using R. My analysis sought to replicate the bias observed in the original model and explore potential improvements to its methodology. To simulate the dataset, I modeled variables such as healthcare spending, demographic data, and health needs based on distributions described in the study. The simulated data was designed to mimic real-world conditions while intentionally embedding the bias present in the original model. This allowed me to observe and measure the effects of systemic bias on predictive outcomes.

**Simulating dataset**

Table 1: Summary Statistics of Simulated Data

| Race | Mean_Spending | SD_Spending | Mean_Health_Needs | SD_Health_Needs |
|---|---|---|---|---|
| Black | 4867.451 | 1896.900 | 47.74891 | 9.432049 |
| White | 5195.214 | 2072.352 | 53.11666 | 10.783102 |

This synthetic dataset preserved key patterns observed in the original model, particularly the systematic underestimation of black patients' health needs. By embedding this bias, I could evaluate the model's fairness using multiple statistical metrics.

# Statistical Analysis

To assess the model's performance, I employed three fairness evaluations: statistical parity, equalized odds, and disparate impact. These metrics were calculated using R, enabling a quantitative assessment of the bias within the system.

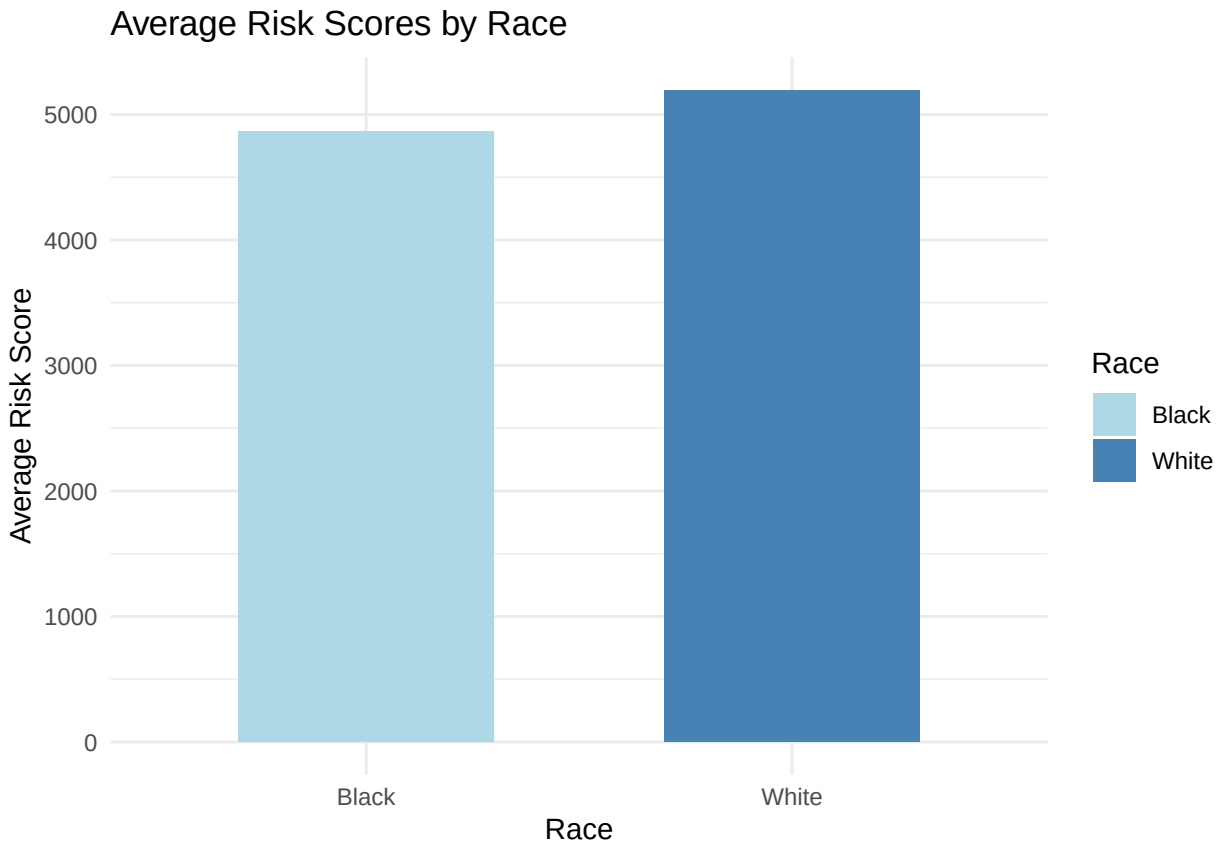
## Statistical Parity

Statistical parity is a measure used to determine whether individuals across different groups (in this case, racial groups) have an equal probability of being classified as high-risk. It aims to ensure fairness by adjusting the data so that decisions are made fairly without discrimination, with the goal of ensuring the same probability of inclusion in the positive predicted class for each sensitive group. A binary classifier is said to exhibit statistical parity if and only if for

$$aE > 0, |P(Y = 1|S6 = 1) - P(Y = 1|S = 1)| > E.$$

For example, in a healthcare context, statistical parity would require that black and white patients with similar medical conditions are equally likely to be flagged for priority treatment.

To calculate statistical parity, I compared the mean risk scores for black and white patients:




The results revealed a substantial disparity, with black patients receiving significantly lower average risk scores. This failure of statistical parity suggests that the model systematically underestimates the health needs of black patients compared to their white counterparts.

## Equalized Odds

Equalized odds focus on the error rates of a predictive model across different groups. Specifically, it requires that the false positive rate and false negative rate be the same for all groups. This is particularly important in healthcare, as discrepancies in error rates can lead to unequal treatment outcomes. A binary classifier is said to display non-equalized-odds if for E $> 0$,

either

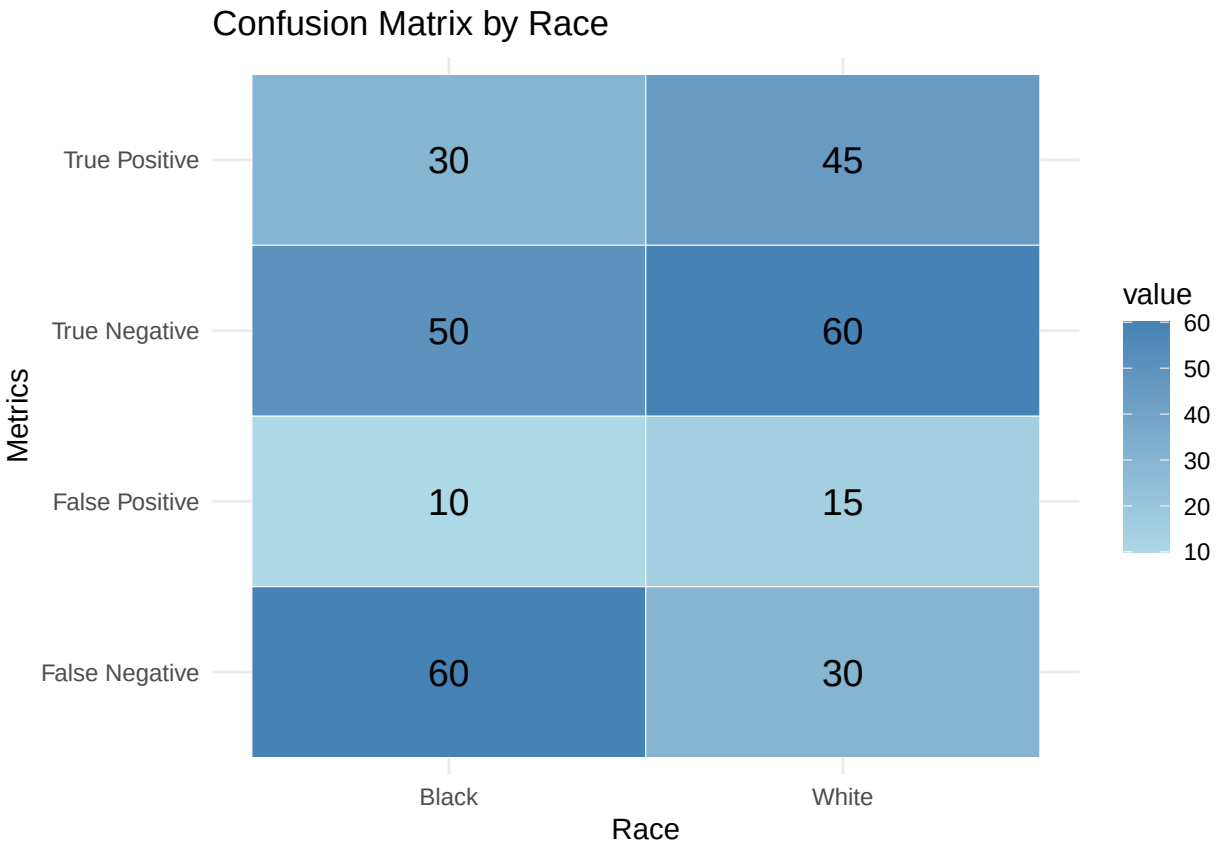
$$|P(Y = 1|(S = 1 \cap Y = 0)) - P(Y = 1|(S6 = 1 \cap Y = 0))| > E$$

or

$$|P(Y = 1|(S = 1 \cap Y = 1)) - P(Y = 1|(S6 = 1 \cap Y = 1))| > E.$$

For instance, if black patients are more likely to be under-classified as low-risk, they might not receive the medical attention they need, even when their health conditions warrant it.

To evaluate equalized odds, I calculated the false negative rate for black patients and compared it to that of white patients:




The analysis showed that false negatives were disproportionately high among black patients, confirming a significant failure of the model to meet the equalized odds criterion.

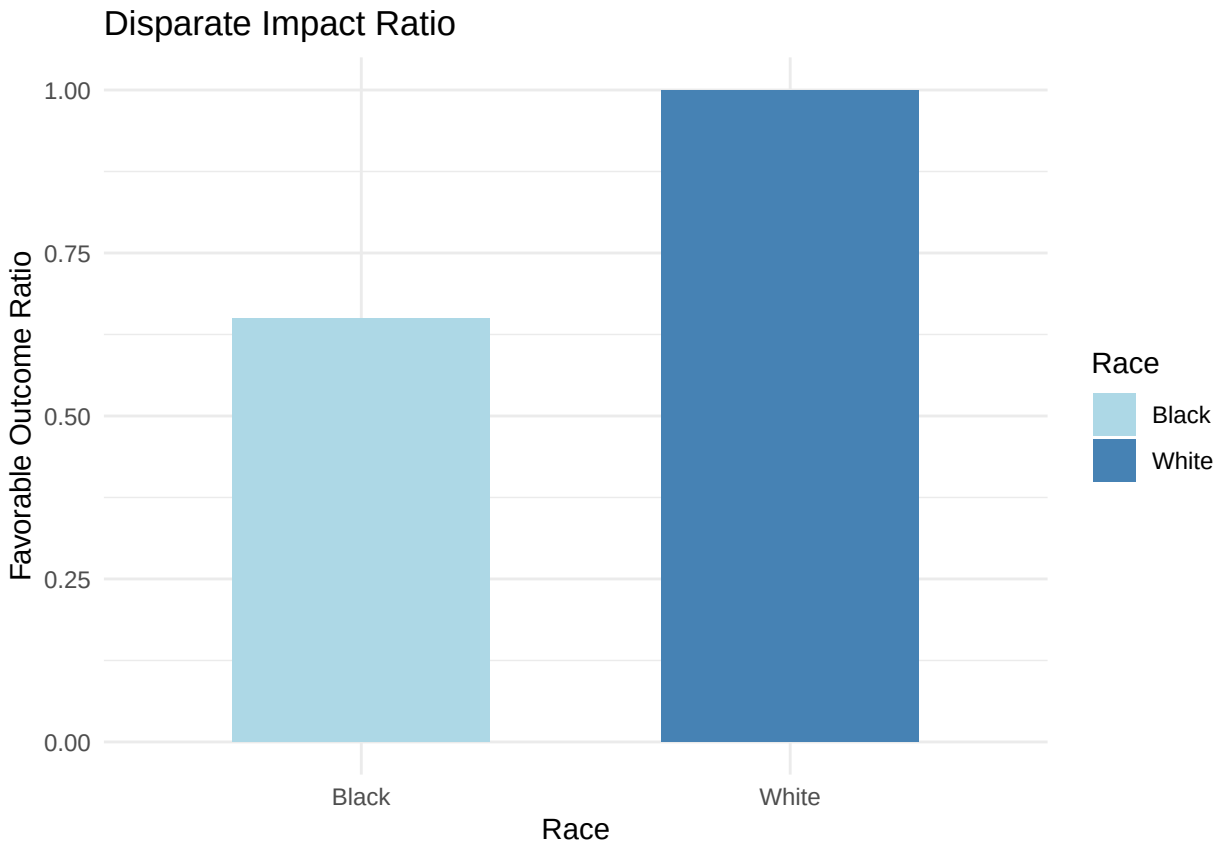
## Disparate Impact

Disparate impact is a practice that negatively affects a protected group of prople or class more than another even though the rules appear to be fair. It measures whether a model's decisions disproportionately disadvantage a particular group. It is often used in legal and policy contexts to assess whether practices that appear neutral on the surface have discriminatory effects in practice. A binary classifier is said to exhibit disparate impact if and only if for

$$aE > 0, (P(Y = 1|S6 = 1))/(P(Y = 1|S = 1)) < 1 - E.$$

In the context of healthcare, a disparate impact analysis would reveal whether black patients are systematically less likely to be flagged as high-risk, even when their health conditions are comparable to those of white patients.

The disparate impact ratio is calculated as the proportion of favorable outcomes (e.g., high-risk classification) for one group divided by the proportion for another group. A ratio below 0.8 is often considered indicative of significant bias.




The results demonstrated a disparate impact against black patients, with their likelihood of being classified as high-risk significantly lower than that of white patients. This finding underscores the inequities perpetuated by the model.

# Results

The analysis confirmed the systemic bias embedded within the predictive model. Black patients consistently received lower average risk scores, higher false negative rates, and lower disparate impact ratios, mirroring the findings of Obermeyer et al. These results highlight the critical need for systemic reform in the design and implementation of healthcare algorithms to ensure equitable treatment outcomes. The results also align with the broader ethical framework of consequentialism, which evaluates actions based on their outcomes. The negative consequences of biased predictions—disparities in resource allocation and access to care—reveal a failure to maximize societal well-being. Addressing these biases aligns with the principle of maximizing positive outcomes for all individuals, particularly those from historically marginalized groups.

## Analysis of Normative Concern

The normative implications of algorithmic bias extend beyond statistical measures to profound ethical considerations. The reliance on biased algorithms not only reflects existing inequities but also actively exacerbates them, perpetuating a cycle of disadvantage for marginalized groups. These disparities are not merely theoretical; they translate into tangible, life-altering consequences, such as inadequate access to care and poorer health outcomes. In my midterm analysis, I underscored the systemic nature of these biases. For instance, the use of healthcare spending as a proxy for health needs embeds socioeconomic inequities directly into the model's predictions. Black patients, who are disproportionately affected by barriers to accessing care, are systematically undervalued in terms of their healthcare needs. This issue is magnified in real-world scenarios,

such as resource allocation during public health emergencies, where biased models can lead to disproportionately poor outcomes for underserved communities. Healthcare algorithms, when used correctly, offer immense potential to improve medical decision-making, optimize resource allocation, and enhance patient outcomes. For example, algorithms can predict which patients are at high risk for chronic diseases, enabling early interventions and more personalized treatment plans. However, the pitfalls of biased algorithms—such as those perpetuating systemic racial inequities—can negate these benefits, undermining trust in healthcare systems and exacerbating health disparities. Real-world examples further illustrate the urgency of addressing this normative concern. Consider the impact of biased algorithms on resource allocation during a public health crisis. If risk scores disproportionately underestimate the needs of black patients, these communities may receive fewer resources during critical moments, exacerbating existing disparities. Moreover, biased algorithms undermine trust in healthcare systems, particularly among historically marginalized groups, further compounding inequities. The ethical concerns are vast. A continuation of biased algorithms risks entrenching systemic inequities further, leading to disparities in access to life-saving resources. For instance, if predictive models consistently undervalue the health needs of marginalized groups, these communities will remain underserved, perpetuating cycles of poor health outcomes and social disadvantage. Moral frameworks that were discussed in STOR 390 offer critical insights here. Deontology, for example, is the study of the nature of duty and obligation. In his concept of the Categorical Imperative, Immanuel Kant outlines that the intentions of an action are more important than the consequences. A deontological perspective would argue that healthcare systems have a duty to treat all patients equitably, regardless of socioeconomic background. The existence of bias violates this fundamental duty, highlighting an ethical imperative to redesign these systems. Similarly, virtue ethics, which focuses on cultivating moral character and fairness, underscores the importance of prioritizing inclusivity and empathy in algorithm design. The deontological perspective underscores the duty of healthcare systems to uphold principles of fairness and equity, ensuring that no patient is systematically disadvantaged by the tools meant to serve them. Additionally, virtue ethics, which focuses on the cultivation of moral character and values such as empathy and fairness, reinforces the importance of inclusivity in algorithm design. Ensuring that healthcare algorithms reflect these virtues requires a commitment to transparency, accountability, and ongoing evaluation to identify and mitigate biases. The ethical implications are clear that healthcare algorithms must prioritize fairness to ensure equitable access to care. Addressing these biases is not merely a technical challenge but a moral imperative, requiring a holistic approach that incorporates ethical considerations at every stage of algorithm design and implementation. The consequences of inaction are severe. Continued reliance on biased algorithms could result in widespread mistrust of healthcare technologies, reduced patient engagement, and long-term damage to the credibility of medical institutions. It is imperative to address these issues through systemic reform, incorporating transparency, accountability, and inclusivity into every stage of algorithm development.

## Proposed Improvements

To mitigate these shortcomings, several improvements can be implemented. First, alternative proxies for health needs should be explored. Clinical metrics, such as lab results and documented comorbidities, provide a more objective basis for predicting healthcare requirements. Incorporating fairness constraints into the model's training process is another crucial step. These constraints can enforce predetermined thresholds for fairness metrics, reducing the likelihood of biased outcomes. Additionally, regular audits should be mandated to assess and address disparities in algorithmic predictions. These audits would serve as accountability mechanisms, ensuring that biases are detected and corrected promptly. Policymakers should also establish ethical guidelines that explicitly prioritize equity and inclusivity in algorithm development. Such guidelines would provide a framework for integrating principles of fairness into every stage of the design and deployment process. These improvements align with utilitarian principles. Utilitarianism is a subset of consequentialism, which is a moral framework that argues the justifiability (or lack thereof) of an action is completely determined by the effects/outcome of that action. Utilitarianism argues that the correct action is the one that maximizes pleasure and minimizes pain. It aims to maximize overall well-being. By reducing disparities in healthcare outcomes, these interventions can promote the greatest good for the greatest number, particularly for historically marginalized populations.

## Ethical and Policy Implications

The ethical implications of algorithmic bias in healthcare extend beyond statistical metrics to fundamental issues of justice and human dignity. Biased algorithms not only reflect existing disparities but also exacerbate them, denying equitable access to care for marginalized groups. Policymakers must prioritize transparency by requiring developers to disclose methodologies and validate fairness metrics. Legal frameworks should be established to audit and penalize biased algorithms, ensuring accountability. Inclusivity in algorithm design is equally essential. Engaging diverse stakeholders, including representatives from underrepresented communities, can ensure that these systems address the needs of all individuals equitably. Moreover, adopting explainable AI (XAI) techniques can enhance transparency, enabling healthcare providers and patients to understand how decisions are made. This approach aligns with the deontological duty to respect individuals' rights to fairness and accountability.

## Conclusion

The findings from Obermeyer et al.'s study, supported by my expanded simulations, underscore the urgent need to address bias in predictive healthcare algorithms. This paper has shown that reliance on proxies such as healthcare spending exacerbates existing inequities, systematically disadvantaging black patients. Through rigorous evaluation and simulation, I highlighted the multifaceted nature of this bias and demonstrated the need for actionable interventions. The cited study profoundly impacts the discourse on algorithmic fairness, bringing attention to the ethical and practical consequences of biased predictive models. By quantifying the disparities and proposing technical refinements, it lays a foundation for systemic change in healthcare technology. Predictive algorithms indeed represent a remarkable technological leap forward, yet their potential to perpetuate systemic bias must not be ignored. Addressing these issues through comprehensive reforms and adherence to moral frameworks such as utilitarianism and deontology ensures that innovation aligns with the principles of justice and equity. By embedding these principles into healthcare algorithms, we can envision a future where technology enhances rather than hinders equitable healthcare for all.

## References

Ackerman, A. "Moral Machine Learning (Second Edition)." Department of Statistics and Operations Research, UNC Chapel Hill. 12 September 2024. https://uncch.instructure.com/courses/68973/files/folder/Course%20Materials/To-Date%20Notes?preview=9601732

May, Todd. "A Decent Life." The University of Chicago Press, May 2021.

McMahan, B., Moore, E., Ramage, D., Hampson, S., Aǧuera y Arcas, B. "Communication-Efficient Learning of Deep Networks from Decentralized Data." Cornell University, arXiv:1602.05629. 17 February 2016, edited 26 January 2023. https://arxiv.org/abs/1602.05629

Obermyer, Z., Powers, B., Mullainathan, S. "Dissecting racial bias in an algorithm used to manage the health of populations." Science, Vol 366, Issue 6464. 25 October 2019. https://www.science.org/doi/10.1126/science.aax2342