Mitigating Racial Bias in Healthcare AI Development

By Athena Xue, Casey Nguyen, and Jodie Meng

Table of Contents

Abstract ……………………………………………………………………………...…….. 2
Introduction to Healthcare AI …………………………………………………….....….. 2
Problem ………………………………………………………………………..……….. 2
Sources of Racial Bias ….…………………………………………………..………... 2
Biases in Data Collection ……………………………………………..………... 2
Biases in Algorithmic Decisions ……………………………….……..……… 3
Developmental Recommendations ………………………………………..……… 5
Incorporating Diverse Representation ……………………………………..………5
Achieving Algorithmic Fairness ………………………….……….…………. 5
Legislative Recommendations …………………………………………………….………… 6
Conclusion …………………..……………………………………………….………… 6
References ……………………………...…………………………………….………… 9

Abstract

Physicians are guided by the principle, “first, do no harm,” but in Silicon Valley, software developers embrace a different motto, “move fast, and break things.” These contrasting philosophies clash in healthcare, where machine learning (ML) and artificial intelligence (AI) are becoming increasingly influential in the diagnosis and treatment of patients. The unintentional incorporation of bias in AI development and deployment can be severely damaging to patients’ wellbeing, impacting the quality and equity of care. Our research will review the ways bias in healthcare AI, specifically racial bias, affects patients and current regulations to prevent bias. We will investigate this information to make professional, developmental, and legislative recommendations for stakeholders in healthcare AI to mitigate bias in their work.

Introduction

AI is becoming increasingly integrated with healthcare to enhance clinical decision-making and patient outcomes. Currently, AI supports many tasks across the healthcare system, including the development of personalized treatments, patient response prediction, and the promotion of efficient telehealth practices (Johnson et al.). Several studies have suggested that AI has already met or exceeded physicians’ performance at key healthcare procedures, such as interpreting diagnostic images and analyzing symptoms of diseases. These advantages have conferred the current view of AI as a tool to drastically improve the accuracy, efficiency, and cost of healthcare delivery, such that AI is anticipated to reduce healthcare costs in the U.S. by $150 billion in 2026 (Bohr & Memarzadeh).

However, a major challenge that AI must address is the potential for racial bias to affect the development and deployment of algorithms. The usage of biased AI outputs in healthcare has a dangerous potential to perpetuate or exacerbate racial inequities by disproportionately affecting vulnerable patients. To proactively combat the consequences of racial bias in healthcare AI, we argue that interventions used by multiple stakeholders — legislators, healthcare leaders, and industry experts — are needed to fairly and safely implement AI in clinical settings.

The Problem

While AI can dramatically transform healthcare, racial bias in the algorithms can have severe consequences for patients. Biases may be statistical, wherein the dataset distribution does not reflect the population distribution and can cause algorithms to produce low-quality outputs that differ from true estimates. Biases may also be social, wherein societal inequities or power imbalances can lead to harmful outcomes for population subgroups (Norori et al.). These biases can be introduced during many different phases in the AI development cycle, including training set selection, problem analysis, and algorithmic design. Such biases have resulted in many consequences for certain patient groups, particularly Black individuals, leading to disadvantages in medical processes, lower health outcomes, and amplification of existing inequalities in the healthcare system.

Sources of Bias

Biases in Data Collection

Racial bias is first introduced in the AI development process through data collection. Disparities in the recruitment of research subjects can result in the underrepresentation of people of color within training datasets. When machine learning algorithms form predictions using biased or skewed datasets, their outputs may lead to inaccurate clinical performance. In dermatology, for example, a growing body of work on convolutional neural networks is revealing biases against Black patients, as these algorithms are mostly trained using skin lesion images from lighter-skinned patients in the United States, Australia, and Europe. In such datasets, images from Black patients comprise only 5-10% of the data (Norori et al.). When one popular online dermatology service was tested on Black patients, it was only able to diagnose Black patients with half the accuracy as white patients — a startling statistic given that Black patients have the highest melanoma mortality rates (Kamulegeya et al.). Insufficient representation of minority populations in datasets can increase misdiagnosis rates from AI, leading to the delayed treatment of patients, progression to more severe stages of disease, and ultimately higher chances of death.

Due to discriminatory healthcare practices or the biased judgment of physicians, bias against minority groups can also be embedded within training data even with perfect sampling. Prior studies have suggested that physicians' diagnostic and treatment decisions are influenced by patient race or ethnicity, as demonstrated by significant differences within physician assessments of patients and skewed tendencies to recommend procedures and prescribe medications for white, male patients (Ryn & Burke). When AI is trained against datasets reflecting such biases, such as clinical case notes or electronic health records, they can compound existing societal inequities that disadvantage minorities.

Furthermore, the practice of race correction within medicine can bias numeric estimates of patients' risk toward disease or response following clinical procedures. As reported by the *New England Journal of Medicine*, clinicians commonly employ adjustments based on race in many specialties, including cardiology, surgery, urology, obstetrics and gynecology, endocrinology, nephrology, and oncology (Vyas et al.). For example, the American Heart Association's risk assessment of in-hospital mortality for patients with acute heart failure assigns an additional three points if the patient is non-Black, despite offering no explanation for this adjustment. When physicians use this and similar estimates to guide referrals and resource allocation decisions, they can exacerbate barriers to healthcare, such as by decreasing access to cardiovascular services for Black patients (Vyas et al.). If such data collected through clinical practices are funneled into AI, the resulting outputs could compound existing inequities within healthcare systems.

Biases in Algorithmic Decisions

Racial bias is not only introduced during data collection but can also be reinforced through design decisions in machine learning algorithms. For example, two popular machine learning algorithms, the Naive Bayes regression classifier, and Logistic Regression can contribute to bias through the way they parse and classify data (Piech). Both algorithms are linearly separable such that in a basic model with two input variables or dimensions, they always try to fit a straight line that separates data instances (Figure 1). Data points are classified based on whether they fall above or below this line which is the algorithm's decision boundary. The algorithms try to classify data points on one side of the line as one prediction and everything on the other side as a different prediction.

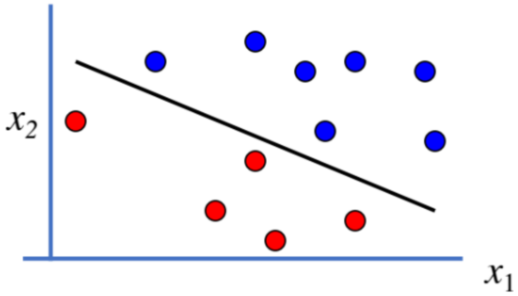

Figure 1. How an algorithm like logistic regression linearly separates and classifies data based on two input variables. Only two input variables are used for the sake of visualization.

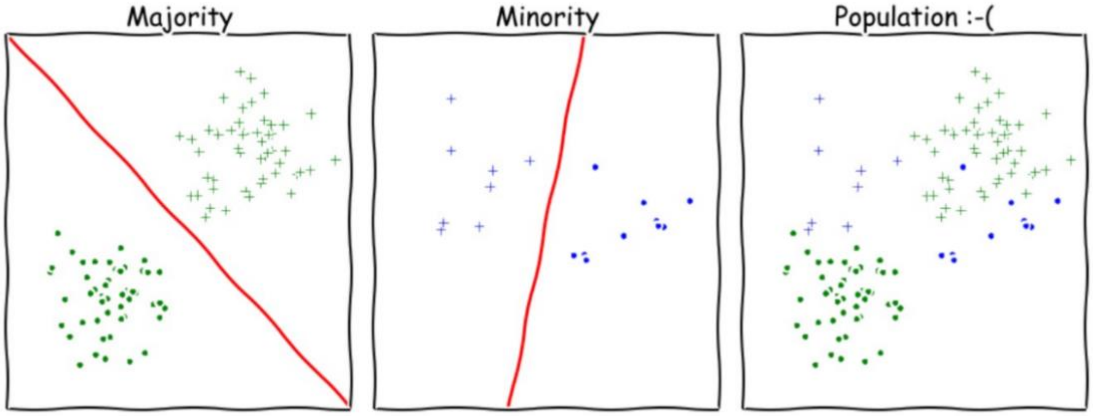

Figure 2. A visualization of how logistic regression classifies data points in a hypothetical dataset with two input variables and a minority and majority demographic.

When groups of data points can be distinctly separated, classification is most accurate. Figure 2 shows how an algorithm like logistic regression separates data in a plot with two input variables or dimensions. The rightmost plot depicts data from a population, the leftmost plot depicts data from a majority demographic in the population, and the middle plot depicts data from a minority demographic. If the algorithm was simply looking at data from the majority demographic (the leftmost plot), it would easily be able to parse the data and classify each point as either a circle or a plus sign. The algorithm would also be able to draw a clean line and easily classify data points exclusively in the minority demographic data (middle plot).

However, the reality is that datasets are often unbalanced and noisy. It is very uncommon for datasets to have perfectly equal distributions of classes or demographics (Yadav). When the algorithm parses the entire population data, it must decide how to draw a line through it. Notice how the lines for the majority and minority demographic datasets have different slopes and intercepts. When the algorithm attempts to draw a line through a population that includes both the majority and minority demographics, it fits to the majority demographic more heavily. As such, classification for minority groups may be worse, as the threshold for determining classification is established around the majority demographic. When minority groups are undersampled in databases, the predictive model will be less accurate for those groups. Even when minority groups make up a proportionate amount of the dataset to accurately represent the population, the algorithm will still attempt to fit to the majority group and perform less effectively for minorities. As these algorithms are hypersensitive to the data they are trained on, they are prone to reinforcing bias.

Developers may also unintentionally introduce racial biases into algorithms when determining target variables and proxies. Lack of understanding of the social, cultural, and geographical variables influencing characteristics within a dataset can lead to the problematic identification of targets. For example, a 2019 publication in Science revealed that an algorithm involved in a commercial risk prediction tool for 200 million U.S. patients discriminated against Black patients due to the developers' assumption that total healthcare costs accrued in a year were a proxy for the state of health (Obermeyer). As result, Black patients who generally spent less money on healthcare but carried substantially more chronic illnesses were assigned the same risk score as white patients. This example underscores the danger of failure to consider socioeconomic and environmental barriers to healthcare access, as the developers systemically mismeasured patients' health and exacerbated the ability of Black patients to treat chronic conditions like diabetes and kidney disease.

Developmental Recommendations

Incorporating Diverse Representation

The racial biases in AI can be mitigated through careful and intentional changes in the way AI is built for use in healthcare, particularly by incorporating broader representation among all people involved — patients, developers, and regulators — and in the entire AI development process (Walch). This includes data collection, design, model training, deployment, and regulation of AI systems.

Homogenous datasets are particularly harmful, especially when AI that is trained with this data is then applied to groups that were underrepresented in the original dataset. With a significant amount of bias stemming from data collection, steps must be taken to curate robust, inclusive, and well-annotated datasets that capture diversity between and within demographic groups. Many AI tools currently utilize publicly available anonymized data to train models, but creating diversified datasets necessitates actively connecting with and seeking data from underrepresented communities. This may involve aiding data-poor regions in developing technological infrastructures, such as better cloud storage and computer speed, to ensure that quality data collection can occur on a global scale (Celi et al.). With these diversified datasets, developers can then train multiple versions of algorithms on all datasets available and then combine all the models, or input all datasets into the AI and train it to learn all at once. The advantage of the latter approach is that the AI will learn to reinforce similarities between input datasets, yet still generalize to each dataset.

Furthermore, to prevent systemic errors in thinking in the AI development process, healthcare AI project teams should include clinicians and community members that can strongly advocate for constituents of healthcare datasets (Kaushal). Researchers assert that more diverse AI development teams can also lead to more representative and diversified datasets. By drawing from experiential knowledge, clinicians and community representatives can assist developers in the problem formulation step and correct biased assumptions that may otherwise lead to misdefined target variables (Kaushal). This could avoid framing problems from the perspective of majority groups and minimize implicitly biased assumptions about data.

Achieving Algorithmic Fairness

In addition to mitigating issues with data collection, biases in AI algorithm design must also be addressed. One philosophy for achieving algorithmic fairness is “Fairness through Unawareness" or procedural fairness, which focuses on ensuring that the process of developing an algorithm is fair. This philosophy in tech subscribes to the idea that if developers do not know information about protected demographics, then ML classification processes will be fair. Some software developers attempt to achieve this by excluding sensitive features such as race, gender, and age from datasets. Additionally, as a precautionary measure, proxies for sensitive features such as name and zip code can also be excluded. However, as it is difficult and unreasonable to remove all possible poxy features, especially with many input variables, algorithms can develop "awareness" of sensitive features contrary to what developers want.

A more feasible and effective approach to creating less biased algorithms is ensuring that the distribution of outcomes is fair instead of the process (Piech). By focusing on distributive fairness, developers can ensure that the distribution of outcomes is equitable according to formal definitions of fairness. One metric of fairness is parity, which involves making sure an algorithm makes a positive prediction with equal probability regardless of demographic. An example of fulfilling parity in healthcare is referring patients to a screening program with the same probability regardless of race. Another metric of fairness is calibration, which involves making sure an algorithm makes a correct prediction with equal probability regardless of demographic. An example of satisfying calibration is correctly identifying skin cancer in patients with the same level of accuracy for all demographics. Using quantitative metrics to measure the fairness of outcomes is a better approach to tracing and detecting bias compared to excluding all proxies potentially related to sensitive features.

While fairness through awareness is effective in minimizing disparities in outcomes, a flaw in this perspective is that developers need to make a qualitative decision on what fairness means in a context. Unfortunately, not all versions of fairness can be attained collectively. Depending on the purpose of an algorithm, developers will need to decide which metric of fairness to meet. It is crucial to thoughtfully choose a standard of fairness that is most relevant to the context of an algorithm because failing to do so may result in a "self-fulfilling prophecy" that creates future bias (Dwork et al.). For example, a classifier that identifies candidates for surgery may accept sufficiently many minority candidates to satisfy statistical parity and equalize selections. However, if the classifier is significantly less effective at identifying candidates in a minority group relative to the data, the minority candidates accepted may have worse health outcomes, leading to future bias.

In seeking to achieve fairness, developers may inadvertently generate bias, which is why it is essential to not only have balanced training data but also transparent reporting. One way developers can increase transparency is by using systematic checklists to investigate a model and share the results with others. An example of a checklist is a Model Card which was created by Margaret Mitchell, Timnit Gebru, and other computer scientists working in algorithmic bias and fairness. A Model Card is a comprehensive checklist that covers model details, intended use, factors, metrics, evaluation data, training data, quantitative analyses, ethical considerations, caveats, and recommendations (Mitchell et al.). Checklists like the Model Card allow developers to keep track of important considerations in the design process of an algorithm and gain external insight from researchers which may help address overlooked flaws.

Legislative Recommendations

The FDA is the primary organization responsible for the regulation of healthcare AI. However, prior studies have identified deficiencies in the ability of current FDA protocols to recognize biases and perform accurately in clinical settings. In a Nature Medicine study, Stanford researchers examined every AI device approved between 2015 to 2020, noting whether

AI evaluations met various parameters indicative of the testing comprehensiveness. Devices were focused on the chest, breast, neck, and head regions, among others. Among these devices included blood tests designed for the detection of specific cancers, blood filtration devices for children of kidney conditions, and AI-assisted imaging software (Bie). Researchers observed that almost all of the approved AI devices, including all the high-risk devices, solely underwent retrospective evaluation. These evaluations, which involve examining data on device performance before any clinical deployment, are insufficient to characterize the performance on the target population (Wu et al.).

To allow for an accurate determination of AI devices' performance in clinical settings, it is critical for AI evaluations to include prospective studies that collect data concurrently with deployment. Prospective studies will allow for the evaluation of AI on participants of diverse races and ethnicities, thereby rooting out sources of racial bias that may otherwise be near impossible to detect. Further, it is necessary to channel prospective studies to monitor devices after market introduction, as AI may continuously learn from inputs to generate new outputs. The correction of such issues before or early in deployment will prevent biases from affecting a large-scale population.

Furthermore, the Stanford study on the FDA testing protocols determined that a large majority of studies do not report multi-site testing. Of those that report this information, a significant portion only tested devices on one or two geographic sites. The lack of testing among diverse, real-world populations can lead to the failure of algorithms to perform accurately in clinical applications. For example, when the researchers tested a deep learning model trained with chest X-ray images from Stanford against data from two other hospitals, they found a 10% decrease in accuracy and a bias for white patients over black patients (Wu et al.). Therefore, in addition to conducting prospective studies, it is critical to gather data from multiple clinical sites and prioritize diversity within participant subgroups to reflect the heterogeneity of patients in the real world.

On a broader scope, governmental organizations are unprepared to address a compliance gap in which the development of technology surpasses the scope of federal and state regulatory standards. To aid the process of standardized evaluations, legislators should take a stronger role in mandating transparency in healthcare AI companies (Davis et al.). For instance, the Algorithmic Accountability Act, introduced into the Senate in February 2022, would require companies to conduct assessments and provide documentation of the high-risk AI systems that make automated decisions regarding a person's access to healthcare, financial services, employment, education, and other essential services. If approved, the bill will work to engage the Federal Trade Commission in enforcing structured guidelines for assessment and reporting and publishing aggregate reports regarding trends in automation ("The Algorithmic Accountability Act of 2022"). While the bill has previously received pushback for its extension of oversight into a number of consumer businesses, the passage of the bill will set a necessary standard for gauging risk within AI. The widespread implementation of the bill will effectively mandate companies to correct sources of racial bias, amongst others, within AI, and increase public confidence in the accuracy of high-risk technologies.

Conclusion

Racial bias in AI can compound existing inequities in the healthcare system and cause severe harm to minority populations. To combat this, we propose increased minority representation in datasets and AI development teams, building algorithmic fairness through transparency and contextual awareness, and the development of legislative measures that prospectively evaluate AI performance and mandate stricter testing protocols. While opponents may argue that subjecting AI to these regulations could hinder innovation and de-incentivize companies from developing new technologies, we argue that if AI is to be employed in high-stakes settings like healthcare, it is imperative to use it fairly and with careful scrutiny because people’s lives are on the line — patients are directly impacted by the quality of care and the tools being used. There is a lot of potential for healthcare in AI yet to be unlocked, but to do so, stakeholders in every step of the development and deployment processes must cooperate to consider and adopt measures that ensure universal safety and fairness in AI implementation.

References

Bohr, A, and K Memarzadeh. The Rise of Artificial Intelligence in Healthcare Applications. 2020, https://doi.org/10.1016/b978-0-12-818438-7.00002-2.

Bie, Ephraim. “US FDA Innovative Medical Device Approvals Rise by 25% in 2020.” *S&P Global*, 21 Jan. 2021, https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/us-fda-innovative-medical-device-approvals-rise-by-25-in-2020-62186337.

Celi, Leo Anthony, et al. “Sources of Bias in Artificial Intelligence That Perpetuate Healthcare Disparities-A Global Review.” PLOS Digital Health, Public Library of Science, 2022, https://doi.org/10.1371/journal.pdig.0000022.

Davis, Nicholas, et al. “The Anatomy of Technology Regulation.” Brookings, Brookings, 4 Mar. 2022, https://www.brookings.edu/opinions/the-anatomy-of-technology-regulation/

Dwork, Cynthia, et al. “Fairness through Awareness.” *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*, 30 Nov. 2011, https://doi.org/10.1145/2090236.2090255.

“FDA Releases Artificial Intelligence/Machine Learning Action Plan.” U.S. Food and Drug Administration, FDA, 12 Jan. 2021, https://www.fda.gov/news-events/press-announcements/fda-releases-artificial-intelligencemachine-learning-action-plan,

Johnson, Kevin, et al. Precision Medicine, AI, and the Future of Personalized Health Care. 22 Sept. 2020, https://doi.org/10.1111/cts.12884.

Kamulegeya, Louis Henry, et al. “Using Artificial Intelligence on Dermatology Conditions in Uganda: A Case for Diversity in Training Data Sets for Machine Learning.” BioRxiv, Cold Spring Harbor Laboratory, 1 Jan. 2019, https://doi.org/10.1101/826057.

Kaushal, Amit. “Health Care AI Systems Are Biased.” Scientific American, Scientific American, 17 Nov. 2020, https://www.scientificamerican.com/article/health-care-ai-systems-are-biased/.

Magnus, D., & Altman, R. (n.d.). Ai + Ethics. Ethics in Bioengineering. Stanford University.

Masters, Ken. “Artificial Intelligence in Medical Education.” Medical Teacher, U.S. National Library of Medicine, https://pubmed.ncbi.nlm.nih.gov/31007106/.

Mitchell, Margaret, et al. “Model Cards for Model Reporting.” *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 14 Jan. 2019, https://doi.org/10.1145/3287560.3287596.

Norori, Natalia, et al. “Addressing Bias in Big Data and AI for Health Care: A Call for Open Science.” *Patterns*, 8 Oct. 2021, https://doi.org/10.1016/j.patter.2021.100347.

Obermeyer, Ziad, et al. *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*. 25 Oct. 2019, https://www.science.org/doi/10.1126/science.aax2342.

Piech, Chris. “Ethics in Machine Learning.” CS109. 7 Mar. 2022, Stanford, CA, Hewlett 200.

Rice, Michelle. “The Growth of Artificial Intelligence (AI) in Healthcare.” *HRS*, Health

Recovery Solutions, 28 Apr. 2022, https://www.healthrecoverysolutions.com/blog/the-growth-of-artificial-intelligence-ai-in-healthcare.

The Algorithmic Accountability Act. 2022, https://www.wyden.senate.gov/imo/media/doc/2022-02-03%20Algorithmic%20Accountability%20Act%20of%202022%20One-pager.pdf.

Van Ryn, M, and J Burke. The Effect of Patient Race and Socio-Economic Status on Physicians' Perceptions of Patients. U.S. National Library of Medicine, 2000, https://pubmed.ncbi.nlm.nih.gov/10695979/.

Vyas, Darshali, et al. "Hidden in Plain Sight - Reconsidering the Use of Race Correction in Clinical Algorithms." *New England Journal of Medicine*, 27 Aug. 2020, https://www.nejm.org/doi/full/10.1056/NEJMms2004740.

Walch, Kathleen. "Combating Racial Bias in AI." *SearchEnterpriseAI*, TechTarget, 23 June 2021, https://www.techtarget.com/searchenterpriseai/feature/Combating-racial-bias-in-AI.

Wu, Eric, et al. "How Medical AI Devices Are Evaluated: Limitations and Recommendations from an Analysis of FDA Approvals." *Nature News*, Nature Publishing Group, 5 Apr. 2021, https://www.nature.com/articles/s41591-021-01312-x.

Yadav, Dinesh. "Weighted Logistic Regression for Imbalanced Dataset." *Medium*, Towards Data Science, 14 Apr. 2020, https://towardsdatascience.com/weighted-logistic-regression-for-imbalanced-dataset-9a5cd88e68b.