Brief Reviews

Misguided Artificial Intelligence: How Racial Bias is Built Into Clinical Models

Atin Jindal, MD[1]

[1] Division of Hospital Medicine, The Miriam Hospital, Lifespan Health System, Warren Alpert Brown School of Medicine, Providence, RI

Keywords: artificial intelligence, racial bias, racism, bias, machine learning, health equity

https://doi.org/10.56305/001c.38021

Vol. 2, Issue 1, 2023

Artificial Intelligence is being used today to solve a myriad of problems. While there is significant promise that AI can help us address many healthcare issues, there is also concern that health inequities can be exacerbated. This article looks specifically at predictive models in regards to racial bias. Each phase of the model building process including raw data collection and processing, data labelling, and implementation of the model can be subject to racial bias. This article aims to explore some of the ways in which this occurs.

INTRODUCTION

Race is a social construct that can be defined in a few ways but generally involves skin color, language, and phenotypic features as major criteria.[1] Perspectives on race include racial identity, which refers to self-identification, race reflection, which refers to how one thinks others identify them, and observed race which is how others actually identify them. Race can also overlap with ethnicity, such as with "Hispanic" or "Jewish" groups.[1,2]

Racism also has several perspectives, but it is generally understood to include not just implicit bias or personal discrimination but also the structure of rules and practices which encourages and fosters such discrimination.[1] One definition of *structural racism* is the "totality of ways in which societies foster racial discrimination through mutually reinforcing systems of housing, education, employment, earnings, benefits, credit, media, health care, and criminal justice".[3] Measuring racism is challenging. Not only can it be subjective, but it can also be politically charged and based on data that is difficult to collect. Nonetheless, it is crucial to have a measurable outcome in order to monitor change and set goals. For this reason, many approaches to measuring racism have emerged, including the Perceived Discrimination Scale.[3]

Artificial Intelligence (AI) can be defined as "a field of science and engineering concerned with the computational understanding of what is commonly called intelligent behavior, and with the creation of artifacts that exhibit such behavior".[4] It is a quickly growing field used broadly today in and outside the clinical context. One overarching theme of AI is to affect behavior change.[5] Some of the areas that AI encompasses are machine learning, deep learning, neural networks, computer vision, natural language processing (NLP), skin imaging, and extensive dataset analysis. Machine learning (ML) is a method of data analysis that builds and improves predictive models automatically. ML is used in the clinical setting to diagnose conditions, predict outcomes, and guide clinical decision-making. NLP is a branch of AI that focuses on making sense of text or voice data using ML as well as other methodologies. It has become particularly prominent in medical AI because a large proportion of clinical data is unstructured, and NLP can help structure, understand, and thus analyze this data.

There is promise that AI can help address issues of health inequities and socioeconomic determinants of health. However, there is also a concern and significant research that shows that AI models can exacerbate racial bias. The social costs of inaccurate predictions are substantial, especially as racial divides in the country deepen.[1] People tend to evaluate AI as relatively autonomous, similar to other humans.[5] This increases blame when AI is wrong to levels similar to that of humans. People tend to blame programmers and companies rather than the algorithms involved.[5] This article aims to examine some of the ways in which racial bias gets included in AI and touch upon what can be done to mitigate these biases. **Figure 1** outlines the steps involved in developing AI-based systems and the ways bias may be introduced.

RAW DATA CAN BE RACIALLY BIASED

Machine learning starts with raw data harvested from an ever-growing collection of data sources. Electronic health records, administrative health records, data warehouses, social media data,[6] as well as population health data are collected and stored with various entities.[7] If the raw data available for training and validation is biased, the analytical results will be biased. Raw data can be racially biased in an underrepresentation of minority groups during data collection, favoring racially biased data types, and in racial discrepancy with health information exchange (HIE).

During the COVID-19 pandemic, there was a significant disparity of data collection in minority groups for a few major reasons; African Americans were less likely to be offered COVID-19 testing, less likely to have testing clinics

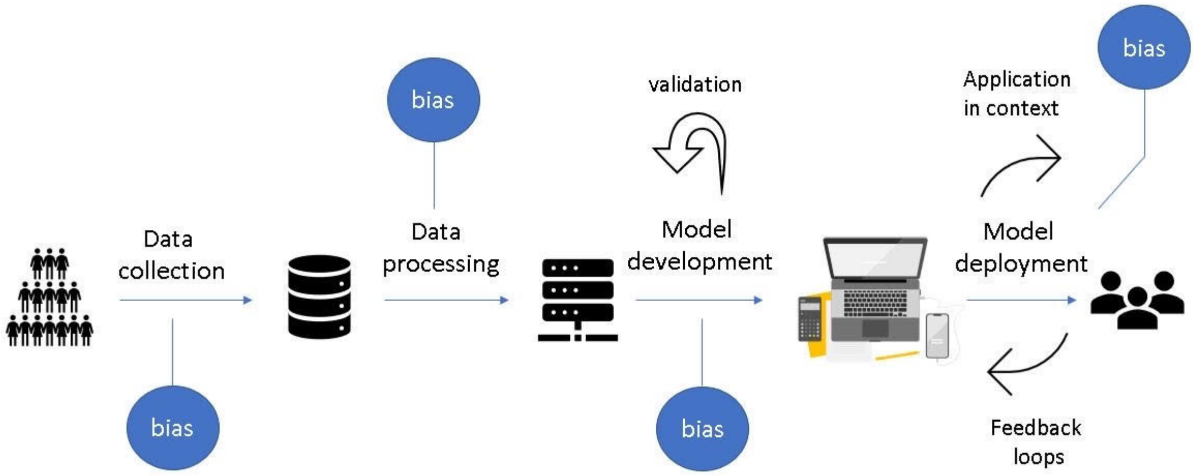

Figure 1. Diagram outlining the steps involved in developing AI-based systems for medical applications and the ways bias may be introduced.

available in their neighborhood, and less likely to be able to follow public health guidelines with their living situation.[8] Randomized control trials often have strict inclusion and exclusion criteria which can select for a racially biased study sample. Social media data is inherently biased towards people with consistent internet and mobile platform access or who speak English.[6,7] Population databases can underrepresent low- and middle-income individuals due to resource constraints and undocumented immigrants due to safety concerns.

Furthermore, HIE is not robust. Disadvantaged or minority groups can have fewer options for providers leading to a greater need for information exchange between health systems.[8] When HIE fails, this can lead to a disproportionately large amount of missing data for minority groups. Data collection and information exchange discrepancies lead to an underrepresentation of data for these groups. When this biased data is used to build models, it limits the generalizability of the model results but not in a way that is easy to identify. In many ways, an inaccurate score might be more harmful than no score.

Machine learning models also tend to favor discrete, structured, physiological data due to the reliability and abundance of this data.[9] Excluding socioeconomic information such as income and family status, which are correlated to race, can exacerbate bias by making all data sources look equal. For example, disadvantaged patients may be more likely to present later in their disease, have fewer data collection points (less primacy care visits), or have less ability to follow up with recommended care. This can racially bias models to be more inaccurate in diagnosis or prognosis for these patients.

DATA EXTRACTION, ORGANIZATION, AND MODEL DEVELOPMENT CAN BE RACIALLY BIASED

Raw data is prepared for predictive analytics by extracting, cleaning, organizing, and labeling. Data labeling involves adding meaningful labels to raw data to give context. Labeled data can then be used in model development. In NLP, for example, word embeddings are used to label language. Word embeddings are the algorithmic representation of words such that similar word meanings have similar representations. These embeddings have been shown to be biased, such as how “man” is represented similar to “doctor” and “woman” is represented similar to “nurse”.[10] It is not difficult to imagine how embeddings may similarly be racially biased. Language is particularly tricky as it has been used throughout colonization to exclude groups of people and justify social hierarchy and violence. Stereotypes and biases are thus inherent in daily language. In this way, NLP is intricately tied to race. For example, language flagged as offensive can lack interpretation based on racial context.[1] NLP implementations can thus build racial bias into their models. In addition, data extractors, cleaners, and organizers themselves can be biased. Anti-racist activists can assign offensive labels differently than other workers, for example. Politically motivated projects can organize differently based on race as well.

After data preparation, manual aspects of model development usually involve decisions on tuning parameters (components used to optimize the model’s accuracy), feature selection (choosing which variables are considered predictive in the model), and performance metrics. Without careful attention and a clear understanding of how these variables affect racial disparities, it can be challenging to account for racially biased models. Blindly choosing features such as race and ethnicity in the model can embed inequities in model results. Even automatic feature selection can be racially biased due to the misleading nature of p-values.[11]

One classic example is using statistical models to derive equations to estimate kidney function using an estimated glomerular filtration rate (eGFR).[12] These models and thus equations use race as a coefficient to predict eGFR. Authors referenced literature that supported Black individuals having higher blood creatinine levels due to increased muscle mass, though that literature was not scientifically robust and has yet to be replicated. Further, several studies have shown that the inclusion of the race or ethnicity coefficient does not improve accuracy among other ethnic or racial groups or Black people outside the USA.[12] Despite these studies, these eGFR equations continue to be prevalent in the USA and ultimately classify Black people as having increased kidney function compared to the gold standard. This can disproportionately disqualify Black people from transplant evaluation and have downstream effects on prognostication and appropriate medical care. This example also highlights how predictive algorithms in medicine can quickly be widely adopted without robust external validation and how difficult it can be to change already implemented models.

IMPLEMENTATION OF AI RESULTS CAN BE RACIALLY BIASED

Implementing AI can involve clinical decision support, predictions of patient care and prognosis, analysis of public health data, and distribution of health resources, to name just a few. To look at an example from the legal world, ML is used to inform decisions on every level of law enforcement, including punishment, bond amounts, and rehabilitation as an option. Predictive policing as a concept aims to pre-emptively intervene in people more likely to commit a crime. The ProPublica study looked at arrests using an ML scoring system and found that these scores were highly unreliable. They also found that black defendants were almost twice as likely to get labeled as future criminals even when criminal history, age, and gender were controlled for.[13] Of particular interest is that the scoring system studied never asks for the race as raw data. This would suggest that despite excluding race as a discrete variable, there is significant racial disparity in predictions. Racially correlated variables may be responsible for this result.

Predictive algorithms like this could be used in the medical world to make decisions on policing communities during pandemics, such as enforcing quarantine or mask rules. If the models we use are racially biased, this could exacerbate the current racial divide in the US.[14] Similar research shows poor accuracy or racial bias for AI in population health,[3,6,15,16] dermatology,[15] heart failure,[16] opioid use,[17] kidney function,[18] speech recognition,[19] gender classification,[20] and many others.[21] There are very few validations of AI in general.[22–24] This continues to be a highly needed area of research today. Externally validated, independent, robust studies aggregated in systematic reviews and clinical practice guidelines are practically non-existent today. Unlike areas of medicine where excellent quality of evidence is available, there is a high risk of widespread adoption before rigorous testing with AI. Further, trainers and end users could use biased and unvalidated model results to create, confirm, or exaggerate personal biases, leading to a perpetuating cycle.

Bias in prognostic predictions, for example, in labeling admitted patients for risk of readmission or decompensation, can lead to discrepancies in disposition or level of care based on those biases. Population-based interventions have also implemented ML, such as vaccine prevalence monitoring, digital contact tracing, and combating anti-vaccine misinformation.[6] Social media data is often mined to identify hotspots of communicable disease,[6] and racial bias in these identification algorithms can misallocate resources. Diagnostic AI, such as with skin recognition, has also been shown to be biased.[15] It is easy to imagine how an algorithm focused on clinical decision support for a dermatological disease might worsen health disparity. For example, if the prediction algorithm is trained with a predominantly Caucasian population in diagnosing skin cancer, it could lead to poor accuracy in Black or Brown populations.[15] In this case, skin cancer could be under-identified and undertreated in dark-skinned populations, worsening already existing medical discrepancies.

IMPROVING RACIAL BIAS

The solution to these problems is not straightforward. It is challenging to simply ignore race because variables otherwise well-correlated to race may still encode racial disparity in the prediction.[25] Even if one endeavors to ignore race in a more sophisticated way and somehow erases any correlation to race in the data, this ignores the issue. In this way, current racial inequity is reproduced. Attempting to inject fairness by rebalancing data can introduce even more issues of inaccuracy or create equality when it does not exist. Another barrier is that there is little incentive to improve models for racial reasons, especially in the private sector.[25]

Nonetheless, improvements can be made. Reducing the underrepresentation of racial minorities can be improved by directly involving these groups in data participation. Partnering with racially diverse organizations like Black in AI, Data for Black Lives, and the Algorithmic Justice League can help as well.[1] Reducing barriers to testing can improve data collection as well.

Improving data transparency, increasing external validation by independent sources before widespread implementation, and strengthening data governance can significantly increase trust and reduce model bias.[8,25] Improving model selection can help, too, as some models are fairer than others. For example, linear support vector machine models produce fairer predictions without compromising accuracy.[6] Alongside process improvements in data collection, labeling, and modeling, comprehensive interdisciplinary racial training for designers and users in each model creation step is needed to address these issues.[26]

CONCLUSION

AI can be racially biased in many ways. Every step of the machine learning process, including raw data collection and processing, data labeling, model building, validation, and implementation, is prone to racial bias. It is essential to acknowledge that even erasing any correlation to race in raw or processed data will reproduce racial inequity. It can be daunting to challenge and improve this bias as significant improvement requires interdisciplinary and extensive training for builders, trainers, and end users. Nevertheless, several steps can be taken in data collection, model building, and the implementation process to reduce racial bias. By addressing these issues now, we can mitigate an exacerbation of racial divides due to AI and maybe even start to bridge some of these gaps and improve racial equity.

DISCLOSURES/CONFLICTS OF INTEREST

The author has no conflicts of interest to disclose.

CORRESPONDING AUTHOR

Atin Jindal, MD
Assistant Professor of Medicine, Clinical Educator
The Warren Alpert Medical School of Brown University
Academic Hospitalist, Physician Informatics Liaison
Division of Hospital Medicine, The Miriam Hospital
Providence, RI, 02906
Email: AJindal@lifespan.org

Submitted: July 20, 2022 EST, Accepted: August 16, 2022 EST

REFERENCES

1. Field A, Blodgett SL, Waseem Z, Tsvetkov Y. A survey of race, racism, and anti-racism in NLP. *arXiv*. Published online June 21, 2021.

2. Crawford K. Artificial intelligence's white guy problem. *The New York Times*. June 25, 2016.

3. Adkins-Jackson PB, Chantarat T, Bailey ZD, Ponce NA. Measuring structural racism: a guide for epidemiologists and other health researchers. *American journal of epidemiology*. 2022;191(4):539-547.

4. Shapiro SC. Artificial intelligence. In: Shapiro SC, ed. *Encyclopedia of Artificial Intelligence*. Vol 1. 2nd ed. Wiley.

5. Hong JW, Williams D. Racism, responsibility and autonomy in HCI: Testing perceptions of an AI agent. *Computers in Human Behavior*. 2019;100:79-84. doi:10.1016/j.chb.2019.06.012

6. Lwowski B, Rios A. The risk of racial bias while tracking influenza-related content on social media using machine learning. *Journal of the American Medical Informatics Association*. 2021;28(4):839-849. doi:10.1093/jamia/ocaa326

7. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical Machine Learning in Healthcare. *Annu Rev Biomed Data Sci*. 2021;4(1):123-144. doi:10.1146/annurev-biodatasci-092820-114757

8. Geneviève LD, Martani A, Shaw D, Elger BS, Wangmo T. Structural racism in precision medicine: leaving no one behind. *BMC Medical Ethics*. 2020;21(1):1-3.

9. Sveen WN, Dewan M, Dexheimer JW. The risk of coding racism into pediatric sepsis care: the necessity of anti-racism in machine learning. *The Journal of Pediatrics*. Published online April 22, 2022.

10. Garg N, Schiebinger L, Jurafsky D, Zou J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc Natl Acad Sci USA*. 2018;115(16):E3635-3644. doi:10.1073/pnas.1720347115

11. Thompson B. Stepwise regression and stepwise discriminant analysis need not apply here: a guidelines editorial. *Educ Psychol Meas*. 1995;55(4):525-534. doi:10.1177/0013164495055004001

12. Eneanya ND, Boulware LE, Tsai J, et al. Health inequities and the inappropriate use of race in nephrology. *Nat Rev Nephrol*. 2022;18(2):84-94. doi:10.1038/s41581-021-00501-8

13. Hong JW, Williams D. Racism, responsibility and autonomy in HCI: Testing perceptions of an AI agent. *Computers in Human Behavior*. 2019;100:79-84. doi:10.1016/j.chb.2019.06.012

14. Shachar C, Gerke S, Adashi EY. AI surveillance during pandemics: ethical implementation imperatives. *Hastings Center Report*. 2020;50(3):18-21.

15. Butt S, Butt H, Gnanappiragasam D. Unintentional consequences of artificial intelligence (AI) in dermatology for patients with skin of colour. *Clinical and Experimental Dermatology*. Published online May 10, 2021.

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

17. Johnson AE, Brewer LC, Echols MR, Mazimba S, Shah RU, Breathett K. Utilizing Artificial Intelligence to Enhance Health Equity Among Patients with Heart Failure. *Heart Failure Clinics*. 2022;18(2):259-273.

18. Thompson HM, Sharma B, Bhalla S, et al. Bias and fairness assessment of a natural language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across racial subgroups. *Journal of the American Medical Informatics Association*. 2021;28(11):2393-2403.

19. Braun L, Wentz A, Baker R, Richardson E, Tsai J. Racialized algorithms for kidney function: Erasing social experience. *Social Science & Medicine*. 2021;268:113548. doi:10.1016/j.socscimed.2020.113548

20. Koenecke A, Nam A, Lake E, et al. Racial disparities in automated speech recognition. *Proc Natl Acad Sci USA*. 2020;117(14):7684-7689. doi:10.1073/pnas.1915768117

21. Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: *Conference on Fairness, Accountability and Transparency*. PMLR; 2018:77-91.

22. Fefegha A. Racial bias and gender bias examples in AI systems. Accessed August 8, 2019. https://medium.com/thoughts-and-reflections/racial-bias-and-gender-bias-examples-in-ai-systems-7211e4c166a1;

23. Obermeyer Z, Topol EJ. Artificial intelligence, bias, and patients' perspectives. *The Lancet*. 2021;397(10289):2038. doi:10.1016/s0140-6736(21)01152-1

24. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: Addressing ethical challenges. *PLoS Med*. 2018;15(11):e1002689. doi:10.1371/journal.pmed.1002689

25. Fountain JE. The moon, the ghetto and artificial intelligence: Reducing systemic racism in computational algorithms. *Government Information Quarterly*. 2022;39(2):101645. doi:10.1016/j.giq.2021.101645

26. Owens K, Walker A. Those designing healthcare algorithms must become actively anti-racist. *Nature medicine*. 2020;26(9):1327-1328.