Stanford University
Human-Centered
Artificial Intelligence

DECEMBER 2021

SPECIAL
HEALTHCARE
SERIES

Risks of AI Race Detection in the Medical System

Matthew Lungren

ARTIFICIAL INTELLIGENCE (AI) IS BEING DEPLOYED FOR A RANGE OF TASKS ACROSS THE MEDICAL SYSTEM, from patient face-scanning to early-stage cancer detection. The U.S. Food and Drug Administration (FDA) and other regulatory bodies around the world are in the process of vetting a range of such algorithms for use. Many in the field hope these AI systems can lower the costs of care, increase the accuracy of medical diagnostics, and boost hospital efficiency, among other benefits.

At the same time, however, AI systems drawing conclusions about demographic information could seriously exacerbate disparities in the medical system—and this is especially true with race. Left unexamined and unchecked, algorithms that both accurately and inaccurately make assessments of patients' racial identity could possibly worsen long-standing inequities across the quality and cost of—and access to—care.

Extensive research has already documented how facial and image recognition systems are often more accurate at recognizing lighter-skinned faces than darker-skinned ones. In practice, this has led to facial recognition systems that wrongly identify one Black person as another, or algorithms that do not even recognize darker skin tones. On the flip side, there has been much discussion about what kind of harm could be inflicted when AI systems classify race remarkably well—accurate recognition tools could be used to harm people of color as well.

KEY TAKEAWAYS

- Algorithms that guess a patient's race, without medical professionals even knowing it, may exacerbate already serious health and patient care disparities between racial groups.
- Technical "de-biasing" techniques often discussed for other algorithms, like distorting inputs (e.g., altering images), may have little effectiveness with medical imaging AI.
- This research was only made possible due to the efforts of several universities and hospitals to make open medical data a public good, allowing our researchers to explore important research questions without conflicts with commercial interests.
- Future research on AI medical imaging regulation and approval should include audits explicitly focused on evaluating an algorithm's assessment on data that includes racial identity, sex, and age.

A groundbreaking series of findings was recently reported by a large international AI research consortium led by Dr. Judy Gichoya, an assistant professor at Emory University, in *Reading Race: AI Recognizes Patient's Racial Identity In Medical Images*. This work explores how well AI models, of the kind already deployed in the medical field, can be trained to predict a patient's race. The investigator team, including researchers from Stanford Center for Artificial Intelligence in Medical & Imaging (AIMI), worked together to apply multiple, commonly deployed machine learning (ML) models to large, publicly and privately available datasets of medical images. These databases included everything from chest and limb X-rays to CT scans of the lungs to mammogram screenings.

Human experts cannot determine a patient's race on these medical imaging examinations, and so, until our study, it was never seriously investigated as it was not thought possible. To our surprise, we found that AI models can very reliably predict self-reported race from medical images across multiple imaging modalities, datasets, and clinical tasks. Even when we altered characteristics like age, tissue density, and body habitus (physique), the models' accuracy held true. In and of itself, this may be concerning, as this attribute could be exploited to reproduce or exacerbate racial inequalities in medicine. But the greater risk is that AI systems will trivially learn to predict a patient's race, without a medical professional even realizing it and reinforce disparate outcomes. Since medical professionals often do not have access to patient race data when performing routine tasks (like a clinical radiologist reviewing a medical image), they would not be able to notice if an algorithm was routinely making bad or harmful decisions based on patient race.

Far more than a medical professional issue, these findings matter for users, developers, and regulators overseeing AI technologies.

> *The purpose was to study not just how a single ML model would handle one patient's race data, but to broadly examine how a range of algorithms could predict a patient's race in a range of scenarios with multiple types of data.*

Introduction

To understand how ML's use in medicine could strengthen or exacerbate inequalities, it is important to understand how algorithms in medicine process race information. Indeed, race and racial identity are not easily quantified and studied in healthcare; many also wrongly conflate race and racial identity with biological concepts like genetic ancestry. In our paper, we define racial identity as "a social, political, and legal construct that relates to the interaction between external perceptions (i.e. 'how do others see me?') and self-identification."

Previous research has found that ML in medicine can predict or make assessments about a patient's demographic information in potentially harmful ways, yet little in the field has focused explicitly on the ability of algorithms to predict race. With that in mind, the team performed three core groups of experiments for the paper, focused on this issue of medical AI and patient race. Each experiment used combinations of ML models, publicly and privately available medical image datasets, and race data self-reported by individuals. The purpose was to study not just how a single ML model would handle one patient's race data, but to broadly examine how a range of algorithms could predict a patient's race in a range of scenarios with multiple types of data.

First, the team focused on quantifiying how well ML models could be trained to predict race from medical images, including in new environments and with different types of imagery. Using three large datasets, we trained the algorithms to detect race and tested their results against self-reported race data corresponding with the image datasets. Second, we tested those models with altered medical images to evaluate whether there were other influences on the algorithms' race predictions. Breast density in mammogram scans, disease labels, and bone density information were just some of the factors we tested against. And third, we examined how factors like medical image quality might influence an algorithm's ability to predict a patient's race.

> *Blurring the images, adding "noise" to them, and scaling their resolutions up and down likewise had no substantial effect on their ability to identify patients' race.*

Research Outcome

After conducting three core groups of experiments, we found that ML algorithms in medicine can be trained, relatively easily, to accurately predict a patient's race. In the first experiment (detecting race in radiology imaging), the deep learning models all had high accuracy in predicting whether a patient's self-reported race was Black, Asian, or white. Performing different combinations of tests, like switching from chest X-rays to mammogram images, did not change this outcome.

In the second experiment, manipulating the medical images did not substantially change the algorithms' ability to predict race. In mammography, tissue density and age did not account for the majority of the ML model's performance. With X-rays, the algorithms predicted race less accurately with disease labels—which were tested as a possible proxy for a patient's race—than it did with the original images. Testing against bone density, age, and sex produced similar findings; the ML models were not predominantly relying on those factors to make a race prediction, but instead were derived from the medical image itself.

Finally, in the third experiment, distorting the resolution and quality of the medical images did not affect the ML models' race-prediction accuracy. Blurring the images, adding "noise" to them, and scaling their resolutions up and down likewise had no substantial effect on their ability to identify patients' race. This included specific testing to intentionally disrupt the model's ability to key in on specific machine intelligible approaches to high and low level features of the images. Testing all this across institutions and populations was especially important given that racial disparities in healthcare may have correlated with imaging equipment or protocols as a proxy for the race of the patient in the image.

The ML models used in the paper's experiments are generalizable to many clinical environments, meaning they may be deployed in a range of different, real medical settings for a range of tasks. Moreover, one of the key takeaways from this large multidisciplinary, multi-institutional effort was that this research was only made possible due to the efforts of several universities and hospitals—notably Stanford AIMI and Beth Israel Deaconess-Massachusetts Institute of Technology—to make open medical data a public good, allowing researchers involved in this project to explore important research questions without conflicts with commercial interests.

All told, there are some limits to the study. We relied on self-reported race for our predictions, and when it comes to racial discrimination, the "vector of harm," as we put it, is the social and cultural construct of racial identity and not genetic ancestry. We also focused on imaging modalities that use ionizing radiation, and there is future work to be done with ultrasound, magnetic resonance, and other kinds of imaging. But going forward, developers, regulators, buyers, and users of this technology should exercise great caution around, and put much thought into, its use. And future medical imaging research should have audits explicitly focused on racial identity, sex, and age—to ensure that in the process of trying to use AI to help advance medicine, it does not systematically worsen racial inequalities.

> *Future medical imaging research should have audits explicitly focused on racial identity, sex, and age—to ensure that in the process of trying to use AI to help advance medicine, it does not systematically worsen racial inequalities.*

Policy Discussion

Racial disparities in the medical system are widespread, pronounced, and often deadly, intertwined with other inequities across such lines as sex and class. As ML algorithms are increasingly deployed in medical environments for a range of tasks, some policymakers at the FDA and elsewhere are studying these algorithms and their effects. What these efforts must understand is just how trivially a medical AI system can be trained to accurately predict a patient's race—perhaps without a medical professional even knowing it.

Much of the discussion around ML, discrimination, and inequality focuses on purported technical solutions to reduce "bias" in algorithms, such as by modifying training data and diversifying training data sources to reduce accuracy disparities between groups. However, our research suggests that such approaches may fail in medical imaging, simply because of the difficulty in isolating race from images. Indeed, in our multiple attempts to change algorithm inputs (e.g., looking at different mammogram tissue densities, selecting different-aged patients) or modify the images altogether (e.g., adding noise, lowering image quality), the ML algorithms could still predict patients' race with high accuracy. Medical professionals still need high-quality images to perform their jobs; even if reducing the quality of an image would mitigate an algorithm's ability to predict race, a human radiologist would still need a higher-quality image to make accurate diagnoses.

The nascent regulatory environment for medical AI has not yet produced robust safeguards against AI models that unexpectedly learn to identify race. Presently, many deployed AI systems in medicine could be learning about patients' race without medical professionals, or perhaps even the original developers of those systems, realizing it. Algorithms could make decisions that factor in race unbeknownst to users, exacerbating discrimination and inequality in medicine—for instance, Black Americans' historical, unequal access to quality healthcare. This could even happen in cases where race is poorly correlated with good outcomes, like pathology detection, but where ML systems will use race to make decisions anyway.

One of the more concerning aspects of the research is that we are not sure why or how these AI models can identify race, despite performing more than a dozen experiments (and 20 more since the publication of the research) and speaking with experts around the world. We cannot fully address what we do not understand, but at the very least, our research calls for careful consideration in the deployment of medical AI systems in patient care and better infrastructure for monitoring and evaluating the impact of AI in healthcare for medical imaging.

Racial disparities in the medical system are widespread, pronounced, and often deadly, intertwined with other inequities across such lines as sex and class.

Regulators and lawmakers should weigh this new research when evaluating the benefits and costs that come with deploying specific AI tools throughout the healthcare system—tools that could inadvertently perpetuate biases inherent in the data—by enacting requirements for explicit testing and monitoring of model development and performance on demographic subgroups.

The original article, **"Reading Race: AI Recognizes Patient's Racial Identity in Medical Images,"** can be accessed at: https://arxiv.org/ftp/arxiv/papers/2107/2107.10356.pdf

Matthew Lungren is an associate professor of radiology at Stanford University and co-director of the Stanford Center for Artificial Intelligence in Medicine and Imaging.

Stanford University's Institute on Human-Centered Artificial Intelligence (HAI), applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu.**

Stanford University
Human-Centered
Artificial Intelligence

Stanford HAI: Cordura Hall, 210 Panama Street, Stanford, CA 94305-4101

T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**