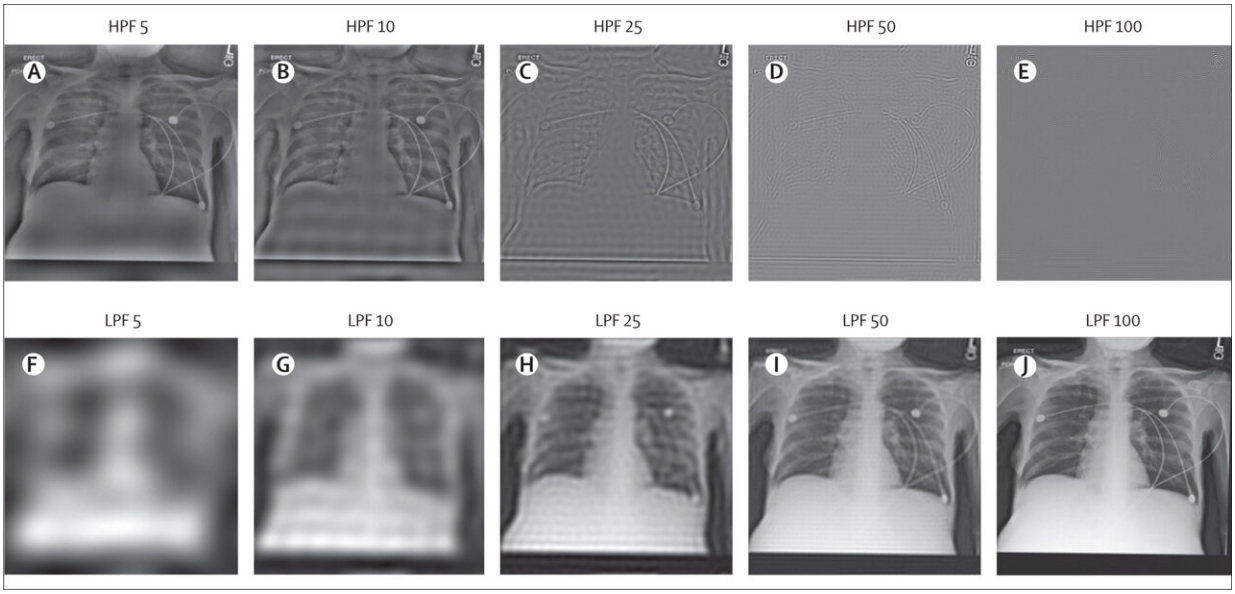
Artificial intelligence predicts patients' race from their medical images

May 20 2022, by Rachel Gordon



Samples of the images after low-pass filters and high-pass filters in MXR dataset. HPF=high-pass filtering. LPF=low-pass filtering. MXR=MIMIC-CXR dataset. Credit: *The Lancet Digital Health* (2022). DOI: 10.1016/S2589-7500(22)00063-2

The miseducation of algorithms is a critical problem; when artificial intelligence mirrors unconscious thoughts, racism, and biases of the humans who generated these algorithms, it can lead to serious harm. Computer programs, for example, have wrongly flagged Black defendants as twice as likely to reoffend as someone who's white. When

an AI used cost as a proxy for health needs, it falsely named Black patients as healthier than equally sick white ones, as less money was spent on them. Even AI used to write a play relied on using harmful stereotypes for casting.

Removing sensitive features from the data seems like a viable tweak. But what happens when it's not enough?

Examples of bias in natural language processing are boundless—but MIT scientists have investigated another important, largely underexplored modality: medical images. Using both private and public datasets, the team found that AI can accurately predict self-reported race of patients from medical images alone. Using imaging data of chest X-rays, limb X-rays, chest CT scans, and mammograms, the team trained a deep learning model to identify race as white, Black, or Asian—even though the images themselves contained no explicit mention of the patient's race. This is a feat even the most seasoned physicians cannot do, and it's not clear how the model was able to do this.

In an attempt to tease out and make sense of the enigmatic "how" of it all, the researchers ran a slew of experiments. To investigate possible mechanisms of race detection, they looked at variables like differences in anatomy, bone density, resolution of images—and many more, and the models still prevailed with high ability to detect race from chest X-rays. "These results were initially confusing, because the members of our research team could not come anywhere close to identifying a good proxy for this task," says paper co-author Marzyeh Ghassemi, an assistant professor in the MIT Department of Electrical Engineering and Computer Science and the Institute for Medical Engineering and Science (IMES), who is an affiliate of the Computer Science and Artificial Intelligence Laboratory (CSAIL) and of the MIT Jameel Clinic. "Even when you filter medical images past where the images are recognizable as medical images at all, deep models maintain a very high performance.

That is concerning because superhuman capacities are generally much more difficult to control, regulate, and prevent from harming people."

In a clinical setting, algorithms can help tell us whether a patient is a candidate for chemotherapy, dictate the triage of patients, or decide if a movement to the ICU is necessary. "We think that the algorithms are only looking at vital signs or laboratory tests, but it's possible they're also looking at your race, ethnicity, sex, whether you're incarcerated or not—even if all of that information is hidden," says paper co-author Leo Anthony Celi, principal research scientist in IMES at MIT and associate professor of medicine at Harvard Medical School. "Just because you have representation of different groups in your algorithms, that doesn't guarantee it won't perpetuate or magnify existing disparities and inequities. Feeding the algorithms with more data with representation is not a panacea. This paper should make us pause and truly reconsider whether we are ready to bring AI to the bedside."

The study, "AI recognition of patient race in medical imaging: a modeling study," was published in *The Lancet Digital Health* on May 11. Celi and Ghassemi wrote the paper alongside 20 other authors in four countries.

To set up the tests, the scientists first showed that the models were able to predict race across multiple imaging modalities, various datasets, and diverse clinical tasks, as well as across a range of academic centers and patient populations in the United States. They used three large chest X-ray datasets, and tested the model on an unseen subset of the dataset used to train the model and a completely different one. Next, they trained the racial identity detection models for non-chest X-ray images from multiple body locations, including digital radiography, mammography, lateral cervical spine radiographs, and chest CTs to see whether the model's performance was limited to chest X-rays.

The team covered many bases in an attempt to explain the model's behavior: differences in physical characteristics between different racial groups (body habitus, breast density), disease distribution (previous studies have shown that Black patients have a higher incidence for health issues like cardiac disease), location-specific or tissue specific differences, effects of societal bias and environmental stress, the ability of deep learning systems to detect race when multiple demographic and patient factors were combined, and if specific image regions contributed to recognizing race.

What emerged was truly staggering: The ability of the models to predict race from diagnostic labels alone was much lower than the chest X-ray image-based models.

For example, the bone density test used images where the thicker part of the bone appeared white, and the thinner part appeared more gray or translucent. Scientists assumed that since Black people generally have higher bone mineral density, the color differences helped the AI models to detect race. To cut that off, they clipped the images with a filter, so the model couldn't color differences. It turned out that cutting off the color supply didn't faze the model—it still could accurately predict races. (The "Area Under the Curve" value, meaning the measure of the accuracy of a quantitative diagnostic test, was 0.94–0.96). As such, the learned features of the model appeared to rely on all regions of the image, meaning that controlling this type of algorithmic behavior presents a messy, challenging problem.

The scientists acknowledge limited availability of racial identity labels, which caused them to focus on Asian, Black, and white populations, and that their ground truth was a self-reported detail. Other forthcoming work will include potentially looking at isolating different signals before image reconstruction, because, as with bone density experiments, they couldn't account for residual bone tissue that was on the images.

Notably, other work by Ghassemi and Celi led by MIT student Hammaad Adam has found that models can also identify patient self-reported race from clinical notes even when those notes are stripped of explicit indicators of race. Just as in this work, human experts are not able to accurately predict patient race from the same redacted clinical notes.

"We need to bring social scientists into the picture. Domain experts, which are usually the clinicians, public health practitioners, computer scientists, and engineers are not enough. Health care is a social-cultural problem just as much as it's a medical problem. We need another group of experts to weigh in and to provide input and feedback on how we design, develop, deploy, and evaluate these algorithms," says Celi. "We need to also ask the data scientists, before any exploration of the data, are there disparities? Which patient groups are marginalized? What are the drivers of those disparities? Is it access to care? Is it from the subjectivity of the care providers? If we don't understand that, we won't have a chance of being able to identify the unintended consequences of the algorithms, and there's no way we'll be able to safeguard the algorithms from perpetuating biases."

"The fact that algorithms 'see' race, as the authors convincingly document, can be dangerous. But an important and related fact is that, when used carefully, algorithms can also work to counter bias," says Ziad Obermeyer, associate professor at the University of California at Berkeley, whose research focuses on AI applied to health. "In our own work, led by computer scientist Emma Pierson at Cornell, we show that algorithms that learn from patients' pain experiences can find new sources of knee pain in X-rays that disproportionately affect Black patients—and are disproportionately missed by radiologists. So just like any tool, algorithms can be a force for evil or a force for good—which one depends on us, and the choices we make when we build algorithms."

More information: Judy Wawira Gichoya et al, AI recognition of patient race in medical imaging: a modelling study, *The Lancet Digital Health* (2022). DOI: 10.1016/S2589-7500(22)00063-2

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology