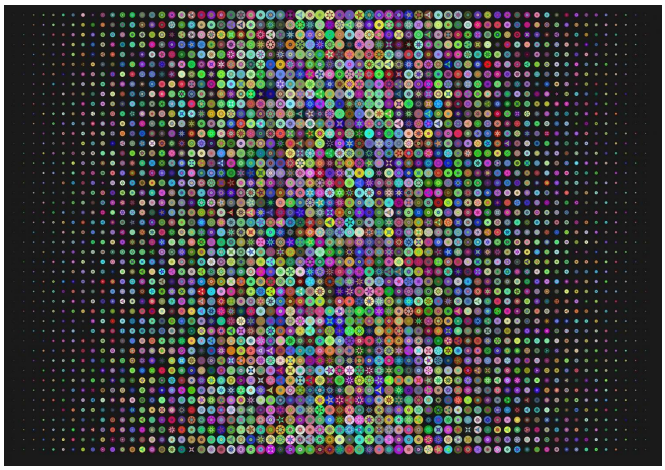
When algorithmic fairness fixes fail: The case for keeping humans in the loop

6 November 2020, by Katharine Miller



Credit: CC0 Public Domain

Attempts to fix clinical prediction algorithms to make them fair also make them less accurate.

As healthcare systems increasingly rely on predictive algorithms to make decisions about patient care, they are bumping up against issues of fairness.

For example, a hospital might use its electronic healthcare records to predict which patients are at risk of cardiovascular disease, diabetes or depression and then offer high-risk patients special attention. But women, Black people, and other ethnic or racial minority groups might have a history of being misdiagnosed or untreated for these problems. That means a predictive model trained on historic data could reproduce historical mistreatment or have a much higher error rate for these subgroups than it does for white male patients. And when the hospital uses that algorithm to decide who should receive special care, that can make matters worse.

Some researchers have been hoping to address model fairness issues algorithmically – by recalibrating the model for different groups or developing ways to reduce systematic differences in the rate and distribution of errors across groups.

But Nigam Shah, associate professor of medicine (biomedical informatics) and of biomedical data science at Stanford University and an affiliated faculty member of the Stanford Institute for Human-Centered Artificial Intelligence (HAI), and graduate students Stephen Pfohl and Agata Foryciarz wondered whether algorithmic fixes were really the answer.

In a recent paper, the team found that the various methods that have been proposed to address algorithmic fairness indeed make algorithms fairer, but they can also make them perform more poorly. "You might actually make the algorithm worse for everybody," Shah says.

The upshot, Shah says, is that when institutions are dealing with issues of fairness in prediction algorithms for clinical outcomes, applying an algorithmic fix is one of three options that should be on the table. The second is to keep a human in the loop to make sure subgroups are treated fairly; and the third is to ditch the algorithm altogether. Knowing which option is most appropriate will require a good understanding of the broader context in which the perceived unfairness arises, he says.

To that end, computer scientists trying to develop fair prediction algorithms for use in the clinic need to connect with stakeholders (clinicians, patients and community members), Pfohl says. "Careful problem formulation, grounded in the values of the population you are trying to help, is fundamental and crucial."

Algorithmic Fairness Approaches' Limited Usefulness

To assess the various approaches that have been

proposed for fixing unfair predictive models, Shah and Pfohl started by training a machine-learning algorithm to predict a handful of health outcomes for thousands of patients in three large datasets. For example, they used 10-plus years of Stanford's electronic health records data to predict hospital mortality, prolonged stays in the hospital and 30-day readmissions. First, they broke the datasets up by age, ethnicity, gender and race. Then, using several different definitions of fairness, they applied related algorithmic fairness fixes to the outcome predictions. "What we get in the end is a big matrix of how different notions of fairness and model performance covary for each subgroup," Pfohl says.

In most cases, the original trained model produced unfair results: Predictions were better calibrated for some racial and ethnic groups than for others, or yielded different numbers of false positives and negatives, for example.

When various algorithmic fairness methods were applied to the model, they actually worked: The distributions of predictions matched up better or the error rates became more similar across groups. But the imposition of fairness came at a cost to model performance: Predictions were less reliable. Moreover, Pfohl says, the various approaches to fairness are in conflict with one another. "If you satisfy one notion of fairness, you won't satisfy another notion of fairness and vice versa – and different notions can be reasonable in different settings."

Despite these problems, it is possible that algorithmic fairness fixes will work in some contexts, Pfohl says. If developers, with input from appropriate stakeholders, put in the hard work to understand what notion of fairness or equity is most relevant to a particular setting, they might be able to balance the tradeoffs between fairness and performance for a narrowly tailored prediction algorithm. "But it's not a general-purpose solution," he says. "Our tech solutions are narrow in scope, and it's important to always remember that."

An Alternative: Focus on Fair Treatment with a Human in the Loop

To Shah, the problem of algorithmic fairness is most concerning when it leads to unfair treatment in the clinic. A recent paper by Ziad Oberyemer received a lot of attention for exactly this reason, Shah says. There, a healthcare provider had used a cost predictive algorithm to decide which patients should be referred to a special high-risk care management program. The algorithm was one that used historic healthcare costs to predict future healthcare costs (and did so in an unbiased way). But when the healthcare provider used future healthcare cost projections as a proxy for healthcare need, the impact of that usage led to unfair treatment: Black patients had to be a lot sicker than white patients before they received the extra care.

This is what people care the most about, Shah says. "If you and I are treated differently by a government or health agency because of an algorithm, we will get upset."

But, Shah says, people tend to blame the algorithm itself. "An often-unstated assumption is that if we fix the systematic error in the estimate of an outcome [using algorithmic fairness approaches], that will in turn fix the error in benefit assignment," he says. "That might be wishful thinking."

Indeed, even if an algorithm is fair for one purpose, or has been patched with an algorithmic fix, clinicians will still need to be aware of a model's limitations so that it's not deployed inappropriately.

Having humans in the loop matters when it comes to making sure predictive algorithms are used fairly in the clinic, Shah says. He points to a widely used algorithm called the pooled cohort equations that predicts a person's risk of having an adverse cardiovascular event in the next 10 years. The algorithm is known to overestimate risk for East Asians, Shah says. As a result, clinicians often prescribe statins for East Asian patients at a different cutoff than the typical cutoff of a 7.5% 10-year risk.

"Algorithms don't live in a vacuum," Shah says. "They are built to enable decisions." There are some situations where fairness may lie in having two different cutoff values for two different subgroups, he says. "And we are perfectly fine

doing that."

Finally, if an algorithmic fix doesn't work, health providers should consider abandoning the algorithm altogether. "That is a perfectly viable option in my view," Shah says. "There are some situations where we should not be using machine learning, period. It's just not worth it."

Pfohl agrees: "I would argue that if you're in a setting where making a prediction doesn't allow you to help people better, then you have to question the use of machine learning, period. You have to step back and solve a different problem or not solve the problem at all."

More information: Pfohl et al., An Empirical Characterization of Fair Machine Learning For Clinical Risk Prediction. arXiv:2007.10306 [stat.ML]. arxiv.org/abs/2007.10306

Provided by Stanford University