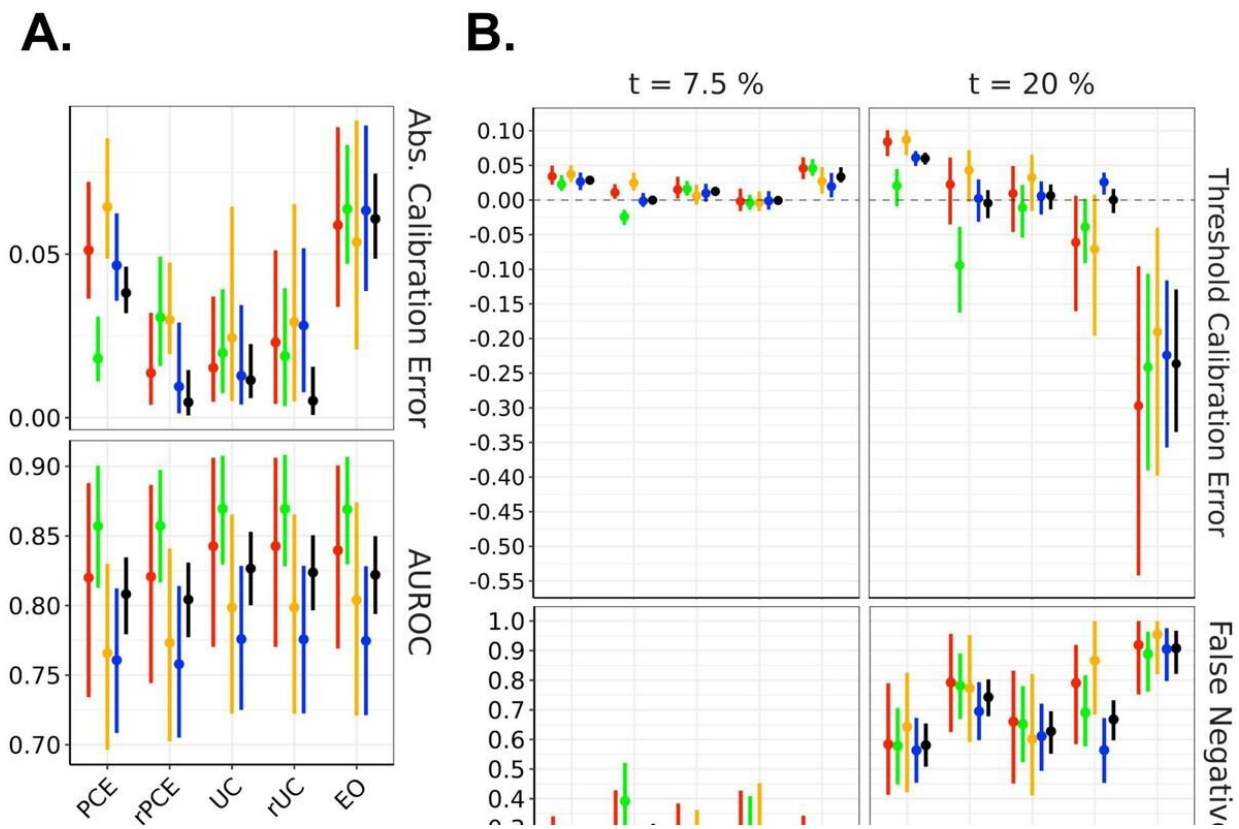
Ensuring the fairness of algorithms that predict patient disease risk

August 1 2022, by Adam Hadhazy



Model performance across evaluation metrics, stratified by demographic group, evaluated on the test set. The left panel showsAUROC and absolute calibration error. The right panel shows false negative rates, false positive rates and threshold calibration error at two therapeutic thresholds (7.5% and 20%). EO, equalized odds; PCEs, original pooled cohort equations; rPCE, revised PCEs; rUC, recalibrated model; UC, unconstrained model. Credit: *BMJ Health & Care Informatics* (2022). DOI: 10.1136/bmjhci-2021-100460

"To treat or not to treat?" is the question continually faced by clinicians. To help with their decision making, some turn to disease risk prediction models. These models forecast which patients are more or less likely to develop disease and thus could benefit from treatment, based on demographic factors and medical data.

With the growth of these tools across the medical field and especially in this area of clinical guidance, researchers at Stanford and elsewhere are grappling with how to ensure the fairness of the models' underlying algorithms. Bias has emerged as a significant problem when models are not developed using data reflecting diverse populations.

In a new study, Stanford researchers examined important clinical guidelines for cardiovascular health that advise the use of a risk calculator to guide prescription decisions for Black women, white women, Black men, and white men. The researchers looked at two ways that have been proposed for improving the fairness of the calculator's algorithms. One approach, known as group recalibration, re-adjusts the risk model for each subgroup of patients to better match frequency of observed outcomes. The second approach, called equalized odds, seeks to ensure that error rates are similar for all groups. The researchers found that the recalibration approach overall produced the better match with the guidelines' recommendations.

The findings underscore the importance of building algorithms that take into account the full context relevant to the populations they serve.

"While machine learning has a lot of promise in medical settings and other social contexts, there is the potential for these technologies to worsen existing health inequities," says Agata Foryciarz, a Stanford Ph.D. student in computer science and lead author of the study published in *BMJ Health & Care Informatics*. "Our results suggest that evaluating disease risk prediction models for fairness can make their use more

responsible."

In addition to Foryciarz, the researchers include senior author Nigam Shah, Chief Data Scientist for Stanford Health Care and a Stanford HAI faculty member; Google Research Scientist Stephen Pfohl, and Google Health Clinical Specialist Birju Patel.

Prudent prevention

The clinical guidelines evaluated in the study are for the primary prevention of atherosclerotic cardiovascular disease. This condition is caused by fats, cholesterol, and other substances building up as so-called plaques on the walls of arteries. The sticky plaques block blood flow and potentially lead to adverse outcomes including strokes and kidney failure.

The guidelines, put out by the American College of Cardiology and the American Heart Association, provide recommendations for when patients should start medications called statins—drugs that reduce the levels of certain cholesterol that lead to arterial buildup.

The atherosclerotic cardiovascular disease guidelines take into account medical measures including blood pressure, cholesterol levels, diabetes diagnoses, smoking status, and hypertension treatment, along with the demographics of sex, age, and race. Based on these data, the guidelines suggest the use of a calculator that then estimates patients' overall risk of developing cardiovascular disease within 10 years. Patients identified as being at intermediate or high risk of disease are advised to initiate statin treatment. For patients who are instead at borderline or low risk of disease, statin therapy could be unnecessary or unwanted given potential medication side effects.

"If you as a patient are perceived to be higher risk than you actually are,

you can be put on a statin that you don't need," says Foryciarz. "Then on the other hand, if you're predicted to be low risk but you really should be on a statin, doctors might fail to put preventive measures in place that could have prevented heart disease later on."

Clinical practice guidelines are increasingly recommending physicians use clinical risk predictions models for various conditions and patient populations. The proliferation of medical-decision support calculators—for instance on phones and other electronics used in clinical settings—means such apps are often right at hand.

"Clinicians are likely to encounter and use more and more of these algorithm-based decision-support tools, so it's important that designers try to ensure the tools are as fair and accurate as possible," says Foryciarz.

Refining risk assessment

For their study, Foryciarz and colleagues used a cohort of more than 25,000 patients age 40–79 collected across several large datasets. The researchers compared the patients' actual incidence of atherosclerosis with the predictions made by risk models. As part of these experiments, the researchers built models using the two approaches of group recalibration and equalized odds and then compared the estimates the model's calculators generated with those generated by a simple model calculator with no fairness adjustment.

Recalibrating separately for each of the four subgroups involved running the model for a subset of each subgroup and obtaining a risk score of the actual percentage of patients who developed disease, and then adjusting the underlying model for the broader subgroup. This approach did successfully boost the model's desired compatibility with the guidelines for those patients at low levels of risk. On the other hand, differences in

the error rates between the subgroups overall did emerge, especially at the high-risk end.

The equalized odds approach, in contrast, required building a new predictive model that was constrained to yield equalized error rates across populations. In practice, this approach achieves similar false-positive and false-negative rates across populations. A false positive refers to a patient who was identified as high risk and would be started on a statin, but who did not develop atherosclerotic cardiovascular disease, while a false negative refers to a patient identified as low risk, but who did develop atherosclerotic cardiovascular disease and would likely have benefited from taking a statin.

Going with this equalized odds approach ultimately skewed the decision threshold levels for the various subgroups. Compared with the group recalibration approach, using the calculator built with equalized odds in mind would have led to more under- and over-prescribing of statins and would fail to potentially prevent some of the adverse outcomes.

The gain in accuracy with group recalibration does require additional time and effort to adjust the original model versus leaving the model as-is, though this would be a small price to pay for improved clinical outcomes. An additional caveat is that dividing a population into subgroups does increase the chances of creating too small a sample size to as effectively assess risks within the subgroup, while also lessening the ability to extend the model's predictions to other subgroups.

Overall, algorithm designers and clinicians alike should keep in mind which fairness metrics to use for evaluation and which, if any, to use for model adjustment. They should also understand how a model or calculator is going to be used in practice and how erroneous predictions could lead to clinical decisions that can generate adverse health outcomes down the line. Awareness of potential bias and further

development of fairness approaches for algorithms can improve outcomes for all, Foryciarz notes.

"While it's not always easy to identify which of possibly many subgroups to focus on, considering some subgroups is better than not considering any," Foryciarz says. "Developing algorithms to serve a diverse population means that the algorithms themselves have to be developed with that diversity in mind."

More information: Agata Foryciarz et al, Evaluating algorithmic fairness in the presence of clinical guidelines: the case of atherosclerotic cardiovascular disease risk estimation, *BMJ Health & Care Informatics* (2022). DOI: 10.1136/bmjhci-2021-100460

Provided by Stanford University