Clinical decision-making and algorithmic inequality

Robert Challen ,[1,2] Leon Danon[1,2]

[1]Engineering Mathematics, University of Bristol, Bristol, UK
[2]Bristol Vaccine Centre, Bristol Medical School, University of Bristol, Bristol, UK

Correspondence to
Dr Robert Challen, Engineering Mathematics, University of Bristol, Bristol, BS8 1QU, UK; rob.challen@bristol.ac.uk

Accepted 15 May 2023
Published Online First 8 June 2023

To cite: Challen R, Danon L. *BMJ Qual Saf* 2023;**32**:495–497.

Decision support algorithms based on historical data will make recommendations that are influenced by past inequality. Detailed historical health data contain patterns that identify demographic factors, such as race,[1] socioeconomic status or religion. These factors are linked to societal disadvantage and hence are indirectly correlated with unequal health outcomes. Machine learning or statistical models trained on such data will be able to identify these patterns and associate unequal outcomes with these disadvantaged groups even if the demographics are not explicitly recorded in the data.[1 2] If the indirect associations later influence a decision support algorithm, it is possible to unknowingly create further disadvantage and reinforce social inequality.[2] The reinforcement of social inequality is at highest risk when the behaviour of an algorithm is not transparent, embedded in a 'black box' and used to influence decisions in the fields of health, education, employment or justice.[3]

In both machine learning and statistical models, inequality can be the result of inadequate observational data or model design. Inadequate data might be the result of practicalities in collecting data from disadvantaged groups,[3] for example, women of childbearing age are under-represented in drug trials and 96% of UK Biobank participants are of European ancestry[4] leading to under-representation of disadvantaged groups within the data, which even the most sophisticated models cannot correct. It may also be due to structural societal issues, such as delayed presentation to healthcare observed in lower income groups, particularly in countries without national health services, which lead to poorer health outcomes for disadvantaged groups. During model development covariates like race, gender or socioeconomic status are a poor proxy for a range of associated cultural and societal risks such as language barriers, diet, exercise, sunlight exposure, poor housing or family history,[5] about which there are typically limited data available. Although methods for identifying and correcting for complex causal relationships exist, they are entirely dependent on the existence and availability of informative data, which, we argue, remain lacking due to historical and current structural inequalities. Statistical or machine learning models developed on routinely available data cannot differentiate between these nuances, if the data are not available, and will collapse multifactorial drivers of historical poor outcomes onto non-specific factors like ethnicity. Suppose, for example, that poor outcomes in an ethnically distinct immigrant population were due to exposure to endemic disease, or some other transient socioeconomic risks, that are not well described in the data. Models trained on such historical data will not be representative of the changing risks faced by future generations, an example of 'temporal drift'.[6] Decision support algorithms based on such models may then propagate the kind of 'race-based' decisions predicated on historical inequality described by Vyas *et al*.[7]

BIAS IN PREDICTIONS

In this *issue of BMJ Quality & Safety*, Teeple and colleagues[8] quantify algorithmic bias in a retrospective observational model validation of 'Palliative Connect'—a risk model developed in 2016 in the University of Pennsylvania and used at the point of an acute care admission to predict risk of death within 6 months. 'Palliative Connect' was used to identify patients who would benefit from palliative care consultation, defined by a threshold of ≥30% mortality risk. The model uses 35 factors including age and gender, emergency presentation, clinical

comorbidities from the Elixhauser score[9] and a panel of laboratory values,[10] but by design excluded sociodemographic factors, such as ethnicity, educational or employment status.

The mortality prediction among 41 327 patients admitted in 2017 was evaluated in different marginalised subpopulations, including race, Hispanic ethnicity, health insurance status, household income and education level, none of which were predictors in 'Palliative Connect'. Teeple and colleagues[8] found that false negatives (ie, mortality risk underestimated) were more common in patients with younger age, lower income, black race or Hispanic ethnicity and lower education (eg, false negative rate was 4.9% higher for black vs non-Hispanic white patients). Conversely, false positives (ie, mortality risk overestimated) were more common in older, white/Asian and male patients with insurance and hence these patients were more likely to be considered for palliative care referral (eg, false positive rate lower by 6.0% in blacks vs non-Hispanic whites). Racial disparity in the decision support algorithm performance appeared to be driven by a correlation between race and a subset of the factors in the model, particularly age, emergency admission, hypertension and gender.

If the decision support algorithm's result were blindly followed in clinical practice, false negative predictions would represent missed opportunities to provide palliative care to a patient and support to their family. False positive predictions would represent unnecessary cost and intervention. The balance of these impacts is needed to quantify the effect of inequality of this decision support algorithm, but is not explicitly described in the study.

Other performance measures presented in the paper include commonly used scores such as the balanced accuracy, Brier score, c-statistic (area under the receiver operating characteristic curve) and the Integrated Calibration Index in accordance with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis reporting guidelines. These do not capture the impact of an algorithm in real-world practice, even though the algorithm had better balanced accuracy for marginalised groups, as these measures do not account for the differential harms of false negative and false positive predictions. The 'insensitivity to impact' of traditional performance measures[6] means that the balance of harm and benefit of two hypothetical predictive models with identical overall performance scores may be very different for a clinical application that results in significant unnecessary intervention (eg, surgery) versus one that results in harmful undertreatment. As Teeple and colleagues[8] nicely prove here, it is not sufficient to determine that an algorithm is equivalently accurate in a disadvantaged subgroup to rule out inequality resulting from its use.

The possible negative implications of race in clinical prediction algorithms are well described by Vyas *et al*[7] and influence a range of clinical decisions resulting in reduced admission rates for heart disease, undertreatment of chronic kidney disease and many other potential harms associated with the explicit use of race as a factor in algorithms. An analysis of the UK QRISK cardiovascular prediction algorithm found it an underpredicted risk in people of South Asian ancestry, potentially leading to undertreatment and a failure to prevent heart disease.[11] The reasons for this have yet to be established. A recent assessment of a stroke risk algorithm[12] determined that risk ranking was 'significantly weaker for Black participants than for White participants' with potential for both undertreatment of high-risk and overtreatment of low-risk individuals. The implication is that important predictors for this subgroup were not included in the final models, either because the models have been optimised for the majority group, or as the result of inadequate data in the subgroup, or both. Obermeyer *et al*[13] assessed the behaviour of a 'black box' commercial risk score which guides targeted healthcare interventions to high-risk patients and found that 'at a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses'. They found this resulted from the inclusion of healthcare spend in the model, which was lower in black people for a myriad of socioeconomic reasons. Using the risk score as a predictive algorithm would result in less intervention being offered to those most in need and persistence of the inequality. Issues with the explicit use of race in models lead Vyas *et al*[7] to recommend that three conditions are met before it is included in algorithms: first, that there is no evidence of confounding between race and more specific factors; second, that there is a plausible causative mechanism; and third, that including race will relieve rather than exacerbate health inequalities. Teeple and colleagues[8] demonstrate that these considerations are important even if race (or indicators of other disadvantaged groups) is not directly represented in clinical algorithms as other factors included in the model may still be correlated with race. Advancing health equity involves identifying inadequate care due to socioeconomic factors as a patient safety issue and developing performance measures that highlight inequality.[2 14] Teeple and colleagues[8] have identified how relevant these goals are in the context of clinical decision algorithms.

IMPLICATIONS IN CLINICAL DECISION-MAKING

Looking to the future, identifying and correcting for under-representation of disadvantaged groups in prospective research cohorts seems like an obvious priority, but one fraught with the practical difficulties recruiting in communities with low levels of trust and engagement in medical research, or safely involving women of childbearing age.[3 5] It will not address the

realities of working with observational studies using historical data, in which significant inequality of care may be recorded and which are not necessarily representative of the standard of care we wish to deliver in the future. Regardless of the data sources used, researchers developing clinical models which include socioeconomic factors, such as race, as covariates must consider whether they are adjusting for the effect of race on the outcome, or for the effect of historical inadequacy of care associated with race.[7] Given the nature of socioeconomic covariates as proxies for issues such as diet, exercise and housing as described above, any significant association of these factors in a model warrants more detailed investigation. Opaque models that perform inscrutable predictions must be explicitly tested for these biases using metrics that take into account differential harms of misclassification.[2 6]

From a clinical perspective, the use of any decision-making algorithm must be combined with well-informed scepticism. Clinicians must be supplied with the tools to interpret a clinical algorithm in the context of the patient in front of them. Studies such as those described here[8 12 13] are the first step to this goal; however, it is unrealistic to expect clinicians to keep abreast of the limitations of all the algorithms they use for all the different patients they are applied to. We previously argued that a key risk in clinical algorithms is prediction drift over time[6] and a one-off model validation study cannot hope to track changes in the effects of sociodemographic inequality. This validation needs to be embedded into the deployment of these algorithms, in such a way that a clinician can immediately see both how confident and how reliable a prediction is for the purposes of making a decision in the context of similar patients in the same clinical environment.

Contributors RC and LD discussed the concept for this article. RC created the initial draft. RC and LD reviewed and edited the final draft.

Funding RC and LD were supported by UKRI through the JUNIPER consortium (grant number MR/V038613/1). LD was further supported by MRC (grant number MC/PC/19067), EPSRC (EP/V051555/1) and the European Centre for Disease Control on unrelated topics.

Competing interests RC and LD are affiliated with the Bristol Vaccine Centre which receives funding from Pfizer, UKRI and UKHSA for unrelated projects.

Patient consent for publication Not applicable.

Provenance and peer review Commissioned; internally peer reviewed.

ORCID iD

Robert Challen http://orcid.org/0000-0002-5504-7768

REFERENCES

1 Gichoya JW, Banerjee I, Bhimireddy AR, *et al*. AI recognition of patient race in medical imaging: a Modelling study. *Lancet Digit Health* 2022;4:e406–14.
2 Rajkomar A, Hardt M, Howell MD, *et al*. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018;169:866–72.
3 Geneviève LD, Martani A, Shaw D, *et al*. Structural racism in precision medicine: leaving no one behind. *BMC Med Ethics* 2020;21:17.
4 Yang G, Mishra M, Perera MA. Multi-Omics studies in historically excluded populations: the road to equity. *Clin Pharmacol Ther* 2023;113:541–56.
5 Adigbli G. Race, science and (Im)Precision medicine. *Nat Med* 2020;26:1675–6.
6 Challen R, Denny J, Pitt M, *et al*. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf* 2019;28:231–7.
7 Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight — reconsidering the use of race correction in clinical Algorithms. *N Engl J Med* 2020;383:874–82.
8 Teeple S, Chivers C, Linn KA, *et al*. Evaluating equity in the predictive performance of an electronic health record-based 6-month mortality risk model to trigger palliative care consultation: a retrospective model validation analysis. *BMJ Qual Saf* 2023;32:503–16.
9 Elixhauser A, Steiner C, Harris DR, *et al*. Comorbidity measures for use with administrative data. *Med Care* 1998;36:8–27.
10 Courtright KR, Chivers C, Becker M, *et al*. Electronic health record mortality prediction model for targeted palliative care among hospitalized medical patients: a pilot quasi-experimental study. *J Gen Intern Med* 2019;34:1841–7.
11 Tillin T, Hughes AD, Whincup P, *et al*. Ethnicity and prediction of cardiovascular disease: performance of Qrisk2 and Framingham scores in a UK Tri-ethnic prospective cohort study (SABRE—Southall and Brent Revisited). *Heart* 2014;100:60–7.
12 Hong C, Pencina MJ, Wojdyla DM, *et al*. Predictive accuracy of stroke risk prediction models across black and white race, sex, and age groups. *JAMA* 2023;329:306–17.
13 Obermeyer Z, Powers B, Vogeli C, *et al*. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366:447–53.
14 Chin MH. Advancing health equity in patient safety: a reckoning, challenge and opportunity. *BMJ Qual Saf* 2021;30:356–61.