Guidance for unbiased predictive information for healthcare decision-making and equity (GUIDE) considerations when race may be a prognostic factor

Ladin, K.; Cuddeback, J.; Duru, O.K.; Goel, S.; Harvey, W.; Park, J.G.; ... ; Kent, D.M.

npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01245-y

Guidance for unbiased predictive information for healthcare decision-making and equity (GUIDE): considerations when race may be a prognostic factor

Keren Ladin[1,2], John Cuddeback[3], O. Kenrik Duru[4], Sharad Goel[5], William Harvey[6], Jinny G. Park[7], Jessica K. Paulus[8], Joyce Sackey[9], Richard Sharp[10], Ewout Steyerberg[11], Berk Ustun[12], David van Klaveren[7,13], Saul N. Weingart[6] & David M. Kent[7,14]

Clinical prediction models (CPMs) are tools that compute the risk of an outcome given a set of patient characteristics and are routinely used to inform patients, guide treatment decision-making, and resource allocation. Although much hope has been placed on CPMs to mitigate human biases, CPMs may potentially contribute to racial disparities in decision-making and resource allocation. While some policymakers, professional organizations, and scholars have called for eliminating race as a variable from CPMs, others raise concerns that excluding race may exacerbate healthcare disparities and this controversy remains unresolved. The Guidance for Unbiased predictive Information for healthcare Decision-making and Equity (GUIDE) provides expert guidelines for model developers and health system administrators on the transparent use of race in CPMs and mitigation of algorithmic bias across contexts developed through a 5-round, modified Delphi process from a diverse 14-person technical expert panel (TEP). Deliberations affirmed that race is a social construct and that the goals of prediction are distinct from those of causal inference, and emphasized: the importance of decisional context (e.g., shared decision-making versus healthcare rationing); the conflicting nature of different anti-discrimination principles (e.g., anticlassification versus antisubordination principles); and the importance of identifying and balancing trade-offs in achieving equity-related goals with race-aware versus race-unaware CPMs for conditions where racial identity is prognostically informative. The GUIDE, comprising 31 key items in the development and use of CPMs in healthcare, outlines foundational principles, distinguishes between bias and fairness, and offers guidance for examining subgroup invalidity and using race as a variable in CPMs. This GUIDE presents a living document that supports appraisal and reporting of bias in CPMs to support best practice in CPM development and use.

Predictive algorithms in healthcare (hereafter clinical prediction models, CPMs) are data-driven models that produce probabilistic predictions of patient-level outcomes. CPMs are increasingly influential in healthcare decision-making, both in doctor-patient encounters to guide treatment recommendations, and at the health system level to inform resource allocation. Despite hopes that CPMs would reduce bias in healthcare, there is growing concern that CPMs may exacerbate inequities in access to care and outcomes for structurally marginalized groups, e.g., deprioritizing African-American patients for kidney transplantation[1,2]. Much of the recent controversy surrounding CPM usage stems from the inclusion of race,

A full list of affiliations appears at the end of the paper. e-mail: David.Kent@tuftsmedicine.org

because race is socially constructed, poorly defined, and inconsistently ascertained in clinical practice[3]. As such, and owing to concerns of medicalizing race in the context of structural and individual racism in the U.S., some professional societies and scholars have argued for the removal of race from all CPMs[4–7].

However, recent studies demonstrate that omitting race from CPMs may exacerbate disparities for minoritized groups, especially those with heart disease, diabetes, prostate cancer, breast cancer, colon cancer, and other conditions that disproportionately affect racial minorities[8–16]. While 'race-unaware' models (CPMs without race, Supplementary Table 1) often predict risk sufficiently well for all patients, for conditions where the burden of disease differs by race, race-unaware models can lead to suboptimal[17], less accurate predictions especially for those in minority groups, since average outcome rates are more reflective of rates in the majority group[18]. When predictions are used to allocate limited resources to higher-risk patients, when other patients might also benefit (albeit less), and risk varies by race, omitting race from CPMs may exacerbate disparities[17,19,20]. Despite growing debate in the medical and policy communities, calls to operationalize algorithmic fairness[21], attention by the U.S. Preventative Services Task Force[22] (USPSTF) and U.S. Agency for Healthcare Research and Quality[23] (AHRQ)[24], and revisions to Section 1557 of the Affordable Care Act[25], no direct guidance clarifies how prediction modelers should approach race as a candidate variable in CPMs, nor how health systems and clinicians should consider the role of race in choosing and using CPMs, either with individual patients or at the population level.

The purpose of this Guidance for Unbiased predictive Information for healthcare Decision-making and Equity (GUIDE) is to offer a set of practical recommendations to evaluate and address algorithmic bias (here defined as differential accuracy of CPMs across racial groups) and algorithmic fairness (here defined as clinical decision-making that does not systematically favor members of one protected class over another), with special attention to potential harms that may result from including or omitting race. We approach this with a shared understanding that race is a social construct, as well as an appreciation of the profound injuries that interpersonal and structural racism cause to individual and population health. This guidance is meant to be responsive to widespread differences in health by race that are historically and structurally rooted, which have been exacerbated by racial bias embedded in the U.S. health system, and offer a starting point for the development of best practices.

Drawing upon a recently developed conceptual framework[26] (Box 1) and multidisciplinary expertise across medicine, clinical prediction, clinical trials, computer science, informatics, statistics, health disparities, ethics, policy, and law, we provide consensus-based: (1) recommendations regarding the use of race in CPMs, (2) guidance for model developers on identifying and addressing algorithmic bias (differential accuracy of CPMs by race), and (3) guidance for model developers and policy-makers on recognizing and mitigating algorithmic unfairness (differential access to care by race). Given the widespread impact of CPMs in healthcare, the GUIDE is intended to provide a first step to assist CPM developers, health system administrators, regulatory agencies and professional medical societies who share responsibility for use and implementation of CPMs.

Results

General approach and principles

The initial item list presented in the first meeting contained 7 items. In the revised list, 1 item remained unchanged, 6 were revised, and 0 items were dropped. Similarly, in subsequent meetings, 27 items were added, reviewed and iteratively revised. All 31 final items met criteria for agreement (75% or more of the TEP voting "Agree" or "Strongly Agree"). Supplementary Table 4 provides a summary of the TEP meeting votes.

Table 1 presents foundational premises, namely that race is a social construct (Item 1), and distinguishes between the goals of prediction and causal inference (Item 2; also Box 2). It identifies two common and distinct uses of CPMs (Item 3): shared decision-making (non-polar; susceptible to bias concerns) and allocation of limited resources (polar; susceptible to both bias and fairness concerns). Table 2 presents general recommendations

Box 1 | Conceptual framework

We started from a conceptual framework[26] that was also the foundation of a recent AHRQ/USPSTF report[24], which describes CPM use across health contexts involving discrete ethical considerations. 'Non-polar' contexts refer to situations in which there exists potential harms to the patient from both over- and under-prediction of risk (e.g., from over- or under-treatment) such that the patient's primary interest is to receive the most accurate prediction. For example, in shared decision-making, when balancing the benefits and harms of a given intervention, predictions are used to align decisions with patient values and preferences. An example of a non-polar context is the use of CPMs to guide decision-making regarding a prostate biopsy after an elevated antigen blood test. In contrast, 'polar' contexts describe circumstances in which the patient's primary interest is to receive a prediction that would prioritize them to receive a treatment or benefit (or to avoid a harm), as opposed to merely the most accurate prediction. Examples of polar contexts include using CPMs to prioritize patients for organ transplantation, or to determine which high-risk patients qualify for a limited resource, such as antivirals during the Covid-19 pandemic[52]. In polar contexts, there are fairness concerns beyond differential CPM accuracy (bias). That is, while the most accurate CPM can support the most efficient distribution of resources to optimize benefits (or utility) across a population, such optimization does not necessarily ensure equitable distribution of resources (distributive justice)[26]. Since different considerations apply where CPMs are either directly used to allocate scarce resources or used to align decisions with a patient's own values and preferences, separate guidelines were developed for these different contexts.

Algorithmic bias versus fairness: we use algorithmic bias to refer to statistical bias in prediction accuracy (i.e., deviation between predicted risk and observed outcome rates) across population subgroups. This is also called model miscalibration, and it is the result of model development methods, data quality, and sampling. This type of statistical bias should be distinguished from bias in the epidemiological or causal context (which implies a discrepancy between an effect estimate and some 'true' causal effect, e.g., through confounding) and bias in the sociopolitical context (which refers to discriminatory beliefs or actions). We expand on methodological guidance for avoiding and evaluating statistical bias in model development[58,63], by focusing on bias stemming from differential performance across demographic subgroups (i.e., subgroup invalidity) which has been largely overlooked in established guidance[58,63].

In contrast to bias, algorithmic fairness concerns arise largely in polar contexts where priority setting is required and where fairness criteria and priorities may conflict[75,76]. Guided by antidiscrimination principles, approaches to fairness in CPMs often appeal to either: (1) use of an input-focused approach that promotes race-unaware allocation by meticulously avoiding the inclusion of race or race proxies (this is generally aligned with the antidiscrimination principle of "anticlassification", Supplementary Table 1); or (2) use of an output-focused approach which evaluates fairness using "outcomes-focused" criteria, to ensure a fair distribution of resources by race (generally aligned with the anti-discrimination principle of "antisubordination" (Supplementary Table 1))[77].

Table 1 | Guidelines: foundational premises

Item	Statement
1	Race is a social construct Race is generally not assumed to have direct, causal effects on outcomes (except indirectly through the effects of racism on health). Yet race or ethnicity can act as a proxy for other important and often poorly measured causes of health outcomes, such as socioeconomic, environmental, cultural, genetic and other factors, and the potentially complex interactions between them. (P1)
2	Distinction between the goals of prediction and causal inference In understanding the use of race and other protected characteristics in clinical prediction models, it is important not to conflate the goals of prediction (which depend only on correlations) with those of causal inference. The use of race in prediction models does not generally support specific inferences about the mechanism of association between race and the outcome of interest (see Box 2). (P2)
3	Goals of clinical prediction a. Clinical prediction provides tailored prognoses that allow doctors and patients to weigh harms and benefits and make decisions that are consistent with a patient's own values and preferences. (P3) b. Clinical prediction models can also be used to support efficient resource allocation to maximize population-wide benefits when resources are constrained. (P4) c. In both cases, prediction models with less predictive accuracy will diminish benefits to individuals and the population (where benefit is narrowly defined by the outcomes being predicted). (P5)

P denotes premise, *R* recommendation.

Box 2 | Prediction is distinct from causal inference

We emphasize that prediction and causal inference are distinct statistical modeling tasks, with different goals and assumptions: one aimed at prognostication, the other at explanation. While there are distinct methods and procedures for these goals[78], they are often conflated in practice[79]. Coefficients within a CPM regression equation cannot be causally interpreted for many reasons (e.g., "Table 2 fallacy"[80], collider bias from cohort selection, etc.). Indeed, predictive effects of variables within a valid CPM may even have the opposite sign as the true causal effect. Simply, "risk factors" measured in observational studies may associate with health outcomes for many reasons aside from direct causation. Valid prediction only requires these associations are stable across other similarly selected population samples, not that they correspond to causal effects. Conversely, causal modeling typically requires specification of a primary exposure variable-of-interest and a set of (often unverifiable) causal assumptions based on content knowledge external to the data. Prediction is a simpler exercise, concerned only about correlations (i.e., correlation does imply prediction).

Herein, we affirm that race is a social construct and, definitionally and logically, can only cause outcomes indirectly through the health effects of racism. Nevertheless, it may be correlated with many unknown or poorly-measured variables that affect health outcomes (e.g., socioeconomic and cultural factors, genetic ancestry) and might account for differences in outcomes in groups defined by self-identified race. For these reasons (i.e., being an indirect cause of health outcomes via racism or acting as a proxy for other unknown/unmeasured causes of health outcomes), race is often empirically observed to be an important predictor of health outcomes.

The debate about including race in CPMs therefore centers around balancing the potential harms and benefits of including race, namely trade-offs related to potential social harms (Box 3) versus improvements in predictive accuracy and decision-making (Box 4).

Table 2 | Guidelines: general premises and recommendations for the inclusion of race in clinical prediction models

Item	Statement
4	There is not a universally consistent approach to conceptualizing, measuring and classifying an individual's race or ethnicity, although the 'gold standard' is typically self-report. (P6)
5	Race or ethnicity should be assessed and defined similarly for model building and application of models in practice, using standards that facilitate consistency (such as the OMB/NIH Standards for the Classification of Federal Data on Race and Ethnicity). [a](R1) Modelers should report clearly how race was obtained and defined in their sample. (R2)
6	Patients should be informed by clinicians/health systems when models including race, are used in clinical or resource allocation decisions. E.g., "This prediction makes use of demographic information, such as your age, sex and race, and clinical information, such as…" (R3)
7	Decisions supported by polar and non-polar predictions have different ethical considerations. Polar predictions most frequently arise when models are used for allocation of scarce health resources. (P7; see also P9)
8	Great caution must be exercised when attempting to adapt or use a model for a different clinical decision than the original application, or in a markedly different population. Transportability of the model must be carefully examined, both for bias (see Tables 3, 4) and for fairness (see Table 5) concerns. (R4)
9	When race is included as a candidate variable, model developers must be transparent about the reasoning and: explain the rationale, clearly outlining potential harms (Box 3) and benefits (Box 4), including references to existing models and other relevant prior literature. (R5)

P denotes premise, *R* recommendation.

[a]OMB has recently revised these standards to include a category for "Middle Eastern or North African" (MENA)[74].

Box 3 | Three potential harms to using race in clinical prediction

1. **Violates principles of anticlassification:** The anticlassification principle (Supplementary Table 1) is deeply embedded not only in antidiscrimination law but also within our culture, and captured within well-known, broadly-accepted (but not always practiced) axioms such as people should be judged by the "content of their character, not the color of their skin[81]." This foundational principle is perhaps the most salient reason for the discomfort felt when using race in clinical prediction and decision-making. Indeed, several European countries (e.g., France, Netherlands) have laws against collecting data on race or ethnicity, which impedes the study of and remedies for racial disparities (Box 5). Thus, anticlassification principles may conflict with antisubordination goals (Supplementary Table 1).
2. **Inappropriately racializes medicine:** Herein, we define 'race' as a social construct, for which there is now broad interdisciplinary consensus. However, there is a long tradition of pseudoscientific biological determinism and racial essentialism that connects race to inherited biological distinctions—explaining or justifying differences in medical outcomes, access to healthcare, risk factors, life trajectories and social status. This perspective is seen as damaging to a decent, diverse, and just society—creating a broad taboo against any use of 'race' that might be misconstrued to provide even indirect or accidental support for these racist notions. Using race in CPMs may also serve to further entrench racialization and conflict with the goal of a post-racial future.
3. **Undermines trust between clinician and patient, and in health institutions:** Effective healthcare is dependent on interpersonal trust, often between clinician and patient, which has been shown to be an important determinant of acceptance of medical recommendations, care satisfaction, and self-reported health improvement. Studies have found significantly higher levels of distrust in the medical profession in Black patients (compared to White patients)[82,83] who have experiences and expectations of racism by providers[84,85]. Though we know of no direct evidence of this, incorporation of 'race' may undermine trust not only in prediction itself but more broadly in the medical system for patients of all races, especially minoritized groups.

(Items 4–7, 9) and cautions regarding the inclusion of race in clinical predictions (Item 8), including both potential harms (Box 3) and benefits (Box 4).

Bias in predictions for non-polar decisional contexts (e.g. shared decision-decision making)

Table 3 addresses algorithmic bias (Items 10, 11), subgroup invalidity (Items 12–14; also Box 5), and "label bias" (Item 15; also Box 6), which is a concern when the meaning or measurement of the outcome might differ systematically across subgroups[27,28]. Label bias is a particular concern because this bias is not detectable with the conventional set of performance metrics that attend to model fit.

Table 4 provides recommendations for the use of race in CPMs in non-polar (shared decision-making) contexts, where predictive accuracy is the paramount modeling priority. The TEP considered how to balance anticlassification principles (which preclude use of race) and antisubordination principles (which may require use of race to prevent minoritized groups from being disadvantaged in some circumstances). Given the importance of accurate predictions to enabling patient autonomy in decision-making (Item 16), in contexts of shared decision-making, the TEP found that inclusion of race may be justified when the predictive effects are statistically robust, clinically meaningful, and go beyond other ascertainable attributes (Item 17). The precise threshold where the statistical benefits of improved calibration will outweigh anticlassification principles may differ across clinical contexts.

Fairness in predictions in polar decisional contexts (e.g. healthcare rationing)

Table 5 offers guidance to modelers and policy-makers on the use of race in CPMs in resource allocation (polar contexts) to prioritize patients for treatment (e.g., rationing scarce healthcare resources, involuntary commitment, etc.) in which the patient has an interest in receiving a prediction that will prioritize their care preferences, rather than the most accurate prediction. The TEP focused on CPMs used in rationing, since this is the most common context in healthcare where fairness concerns arise. The guidance clarifies that the development of accurate CPMs and fair decision-making are distinct domains and require different expertise (Item 27). Whether predictions are used to inform rationing choices by human decision-makers or are incorporated into formal or automated allocation models, additional analysis of distributive justice implications (i.e., beyond measures of predictive performance) is necessary to determine how best to balance conflicting fairness principles (Items 18, 19, 26). For example, using the most accurate CPM to allocate scarce resources may be appropriate to promote utility, defined as maximizing beneficial outcomes across a population, yet this would not necessarily ensure equitably distributed resources across subgroups (Item 20). In the absence of consensus on principles or unitary fairness criteria[29,30], procedural justice—stakeholder input, transparency, revisability, and an appeals process—can enhance the process for achieving fairness and consensus (Item 25)[31].

Discussion

Clinicians, health systems, professional organizations, advocacy groups, researchers, and policy-makers, are struggling to address use of race in CPMs[32–38]. While some have called for removing race from all CPMs[5–7], recent studies have underscored harms associated with race-unaware estimates, including exacerbating disparities[10–17,19,39]. In 2022, revisions to Section 1557 of the Affordable Care Act extended antidiscrimination requirements to clinical algorithms, rendering health systems and clinicians liable for "decisions made in reliance on clinical algorithms… [that rest] upon or [result] in discrimination [based on "protected traits", including race]"[40]. Yet, no standards for CPM development and validation, bias mitigation, or fairness testing exist. Most frameworks and regulations have focused on increasing transparency, without specific guidance on using protected traits such as race[3,20,29,30,41–48], and other key issues such as biased training data, model transportability, and the unique concerns of different decision-making contexts. Our GUIDE targets these gaps with a set of consensual premises and actionable recommendations.

With differing viewpoints, expertise, and backgrounds represented, starting from a shared set of premises, the TEP agreed that the use of race in CPMs implicitly encourages racialized medicine (Items 1, 2 in Table 1), and as such should be limited to cases where omitting race would harm people of color (Item 22 in Table 5)[49]. The TEP's general objection to the use of race in CPMs acknowledges the sordid history in science and medicine and the fundamental antidiscrimination principle (anticlassification), which prohibits using protected traits such as race, for the purposes of decision-making. The TEP also expressed concern that assigning a single effect estimate or coefficient in a regression model to people based on group

Box 4 | Potential benefits of including race in clinical prediction

There is broad agreement that individuals with similar outcome risks should be treated similarly regardless of race. We call this principle "equal treatment for equal risk." When race has no prognostic information independent of relevant clinical characteristics, there is no controversy, since only characteristics contributing to prognosis are included in CPMs. Controversy arises only when race is predictive of differences in outcome risk, despite clinical characteristics that appear similar.

Such is the case when there are racial disparities in outcomes. We illustrate this in prediabetic patients, where the risk of developing diabetes is ~35% higher in Black compared to White patients—also higher in Asians[10,51]. The Figure shows predicted versus observed outcomes for alternative diabetes models derived in >1 million U.S. prediabetic patients: Panel A for a "race-unaware" model and Panel B for a "race-aware" model[86]. Omitting race systematically under-estimates diabetes risk in Black patients, deprioritizing their care compared to Whites at similar risk. Including race better aligns predicted with observed risks in Black patients, supporting similar treatment for similar risk, regardless of race.

In the "shared decision-making" context, the race-aware model offers more accurate predictions across all groups, particularly minority groups—since the race-unaware model most closely reflects those in the majority. This is a general property of CPMs, since (in the absence of label bias [Box 6]) a race-aware model will generally be at least as accurate as a race-unaware model[17,19]. More generally, models restricted from using any prognostic candidate variable won't be more accurate than models considering all available information.

In this case, the race-aware model may also be disparity-reducing compared to the race-unaware model. If one were offering a lifestyle modification program to the top risk-quarter (>~10% diabetes-risk threshold), Black patients would comprise 31% of the treatment-prioritized group with a race-unaware model, and 51% with a race-aware model. The race-unaware model would prioritize lower-risk White ahead of higher-risk Black patients.

Indeed, an appealing feature of risk-based decision-making is that it can be characterized as a general-purpose, disparity-reducing algorithm. By targeting resources to patients based on risk, risk-based approaches focus resources where they're most needed—prioritizing those who are worse-off. When disparate outcomes are race-associated, leaving race out blinds predictions to these risk differences, potentially amplifying disparities. This disparity-reducing rationale has been used to justify the inclusion of race in lung and colon cancer screening CPMs[11,15].

While the causes of excess risk in some minorities may be unclear, this excess risk is no less important for decision-making than the risk associated with other variables in the model. Thus, when Black people are found to be at higher-risk than White people, despite controlling for other variables, leaving race out of risk calculations does *not* treat Black and White people equally—it systematically ignores those (unknown/unmeasured) causes of greater risk that are more common in Black than White people.

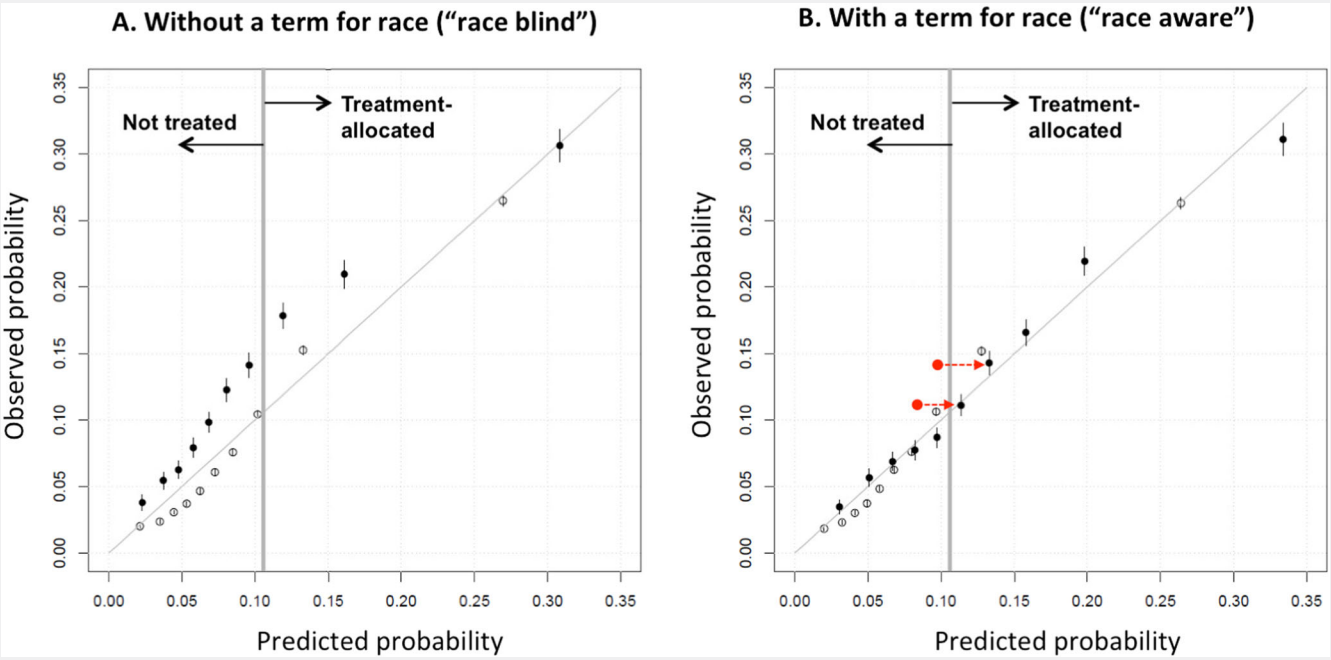

Figure. **Calibration performance of a prediction model for incident diabetes in a large nationally-representative sample of patients with prediabetes. The Figure shows the calibration plot (predicted versus observed risk) of a model for diabetes risk among patients with prediabetes, developed in a large, nationally representative sample of patients from the OptumLabs Data Warehouse (*n* = 1.1 million). Filled dots represent deciles of patients of Black race, and empty circles represent deciles of patients of White race. The solid diagonal line in each graph describes the line of identity, where predicted and observed risks are equal. Panel A describes calibration of the 10-variable model without a term for patient race ("race-blind"). Panel B describes calibration of the same risk model including a term for patient race ("race-aware"). The "race-blind" model systematically underestimates Black patient risk (predictions fall to the left of the line of identity). Including race in the model shifts predictions for Black patients rightward, aligning them on the line of identity and increases the proportion correctly identified as high risk (i.e., above a threshold of 10%). Red dots represent deciles (of Black patients) where the predicted probability fell below the threshold where treatment would be offered in a race-blind model, but exceeded the threshold in a race-aware model.**

Table 3 | Guidelines: addressing algorithmic bias/subgroup invalidity

Item	Statement
10	Samples used for prediction model derivation should represent the underlying population consistent with intended use. (R6)
11	Prediction model development should adhere to best practice guidance[57,58,63], including avoiding approaches known to increase the risk of bias in prediction. Following existing guidance is necessary (but not sufficient) to avoid algorithmic bias across racial or ethnic subgroups. (P8)
12	Model performance should not be assumed to be similar across all major demographic groups. Performance should be assessed and reported by racial or ethnic subgroup, as well as population-wide. Justification should be provided when models are not assessed or calibrated to specific subgroups. (R7)
13	When comparing performance across racial or ethnic subgroups, prevalence-insensitive measures, such as AUC and calibration, should be used to evaluate predictive validity (Box 5). (R8)
14	Best practices for model development should be designed to yield good performance across important racial or ethnic subgroups. If models are found to perform poorly on a given subgroup, modelers should explore remedies to improve performance and/or issue appropriate cautions clarifying the limitations of model applicability. (R9)
15	Careful examination is needed to explore potential "label bias" to ensure that the outcome is similarly informative across important racial or ethnic subgroups and is well suited to the decision (Box 6). (R10)

P denotes premise, *R* recommendation.

membership obscures the heterogeneity inherently experienced by individuals within that group (Item 15 in Table 3).

Nevertheless, the TEP also recognized that this anticlassification principle may conflict with principles that seek to address inequality experienced by minoritized groups (i.e., antisubordination principles). Specifically, the TEP recognized that when race has prognostic value independent of other reliably ascertainable variables, omitting race may worsen model performance and decision-making, and thus cause patient harm (Items 17 in Table 1 and 21, 26 in Table 5). The TEP identified two different classes of harms that might arise by ignoring the predictive effects of race: (1) subgroup invalidity [particularly in the (non-polar) shared decision-making context, Items 10–13 in Tables 3] and (2) exacerbation of disparities through unfair resource distribution [particularly in the rationing (polar) context, Items 21, 22 in Table 5].

For shared decision-making, subgroup invalidity may lead to misinformation and decision-making incongruent with a patient's values and preferences, especially if underrepresented groups are given less accurate estimates when models are calibrated to the overall population (Item 12 in Table 3). For example, in prostate cancer risk prediction, cancer risk is a key input to a patient's decision to obtain a biopsy, and harm can arise from both over- and under-testing. In this case, if CPMs are race-unaware[17,19,20], clinicians may inaccurately under-estimate the true risk borne by Black men and over-estimate the risk in others. Where this miscalibration is large, the inclusion of race may be evaluated to ensure that it improves predictive accuracy for every subgroup[11,15,50,51].

When prediction models are used to ration healthcare resources, particularly when there are disparities in outcomes by race, CPMs omitting race may ignore and exacerbate these disparities by treating higher-risk minoritized patients similarly to lower-risk White patients (Item 21 in Table 5). As an example, during the Covid pandemic, certain states promoted race-aware prediction models for Covid hospitalization, to prioritize high-risk patients for Covid therapies (such as Paxlovid or monoclonal antibody therapy)[52]. Including race in such a circumstance can promote both the egalitarian principle of "equal treatment for equal risk" and the utilitarian principle of "achieving the greatest overall benefit", irrespective of race[53].

Citing the nuanced trade-offs implicated in different CPMs, the TEP did not adopt a uniform recommendation to remove race from all CPMs, since it is difficult to anticipate the balance of trade-offs across all cases. When the predictive effects of race are statistically robust, of a clinically important magnitude and independent of other ascertainable variables, then judgement will be needed to weigh whether the benefits of including race outweigh anticlassification concerns (Item 17 in Table 4). Where a race-aware model is deemed beneficial, justification for its use should be provided explicitly (Item 9 in Table 2).

Reflecting a growing awareness of this issue, several race-aware equations were recently revised to remove race from their calculations, including calculators for predicting pediatric urinary tract infections[54], predicting non-progression of vaginal delivery after prior caesarian section[55] and estimating glomerular filtration rate (eGFR)[1,56]. In so doing, it is important to show model calibration of the newly estimated race-unaware model in different racial and ethnic subgroups, since overall model performance might generally be anticipated to remain robust, even when removal of race leads to substantial miscalibration in smaller subgroups. Lamentably, this has not been shown in all the above cases. More generally, the TEP recommended examining race-specific performance as good practice whenever deriving a new model (Item 9 in Table 2). This is a novel recommendation, not previously addressed in guidelines such as PROGRESS[57] or TRIPOD[58].

The TEP did not prescribe precise analytic procedures to evaluate trade-offs of race-aware versus race-unaware models, and we anticipate that methods may vary and evolve; minimum standards should be clarified in future work. We direct readers to several recent examples in which efforts have been made to carefully examine trade-offs, including examining subgroup validity and the potential effect on disparities for models used for the purposes of rationing[1,11,15,51]. In addition to not examining subgroup-specific calibration, another potential error common in the literature is assuming that directing more care toward a subgroup is generally favorable toward that group. However, since both over- and under-treatment can lead to harm in a non-polar context, the accuracy of predictions should be prioritized in this context. We acknowledge recent work demonstrating that even when race-aware prediction substantially improves statistical accuracy, it may still yield only modest clinical benefits, particularly in a non-polar (shared decision-making) context, and so race-unaware prediction may often be preferred, even in conditions where accuracy is improved with race-aware prediction[51].

The TEP agreed about the importance of identifying when spurious differences in outcomes across races might arise through label bias (Box 6), since when label bias is present the observed differences in risk may be due to bias in the data (Item 15 in Table 3). It is important to note that this does not typically require a full causal understanding of how a predictive race effect might arise, as described in Box 6.

Developing guidance to address algorithmic bias and fairness, and the use of race as a variable in CPMs requires substantial technical expertise—yet technical expertise alone is insufficient. The guidance developed here reflects the expertise, values and perspectives of this particular group and different groups may weigh differently the incommensurable harms that necessarily arise when ethical principles conflict. We note that our deliberations largely reflect a specifically American context and are unlikely to apply similarly in other countries and cultures, with different sociopolitical histories; we also do not directly address issues related to so-called "race-norming"[59]. We acknowledge that inconsistencies in operationalizing the ascertainment

Box 5 | Subgroup invalidity

Subgroup invalidity is a type of algorithmic bias that is assessed by examining the predictive performance across different groups within a population (e.g., based on gender, race, ethnicity, etc.). Performance of CPMs is typically characterized by two dimensions: discrimination (i.e., Does the model give higher predicted probabilities to individuals with the outcome versus those without the outcome?) and calibration (i.e., Does the predicted outcome rate match the observed outcome rate across subgroups at different levels of predicted risk?). Classification is typically evaluated by threshold-dependent measures related to discrimination (e.g., sensitivity, specificity, positive and negative predictive value, F1 and Matthews correlation coefficient, etc.). While these classification measures are popular for examining subgroup validity (and have also been proposed as fairness metrics), it is important to realize that all threshold-dependent measures will generally yield dissimilar scores across subgroups when the outcome rates between these subgroups differ, even in the absence of model invalidity in either group[87]. This 'prevalence-sensitivity' can be shown in simulations using prediction models that are known to have no model invalidity (i.e., they correspond exactly to the data-generating process). The Table below provides an illustration where the predictive performance of the data-generating model (i.e., a model with no model invalidity) is measured across two groups with different burdens of the same risk factors.

In contrast to threshold-dependent measures, 'valid' models will always have perfect *calibration* across subgroups, as they do in the simulated example (shown by E values of 0)[87]. For the non-polar (i.e., shared decision-making) context, good calibration also ensures 'non-harmful decision-making' using a decision-analytic framework (compared to using the best strategy for all) and the appropriate balancing of harms and benefits[88]. For these reasons, we propose that good calibration is a more appropriate and useful measure of subgroup invalidity.

Table. Simulated example of subgroup invalidity assessment of the predictive performance across different subgroups

	Group A	Group B
Predicted avg outcome rate	0.071	0.130
Observed avg outcome rate	0.071	0.130
Overall measures		
AUC	0.683	0.684
E_{avg} (calibration)	0.000	0.000
E_{90} (calibration)	0.001	0.001
At threshold 10%		
Sensitivity	0.42	0.78
Specificity	0.81	0.46
Positive predictive value	0.14	0.18
Negative predictive value	0.95	0.93
F_1 score[a]	0.22	0.29
Fowlkes-Mallows Index (FM)[b]	0.25	0.37
Net benefit[c]	0.01	0.05

[a]F_1: measures a model's accuracy (i.e., how many times a model made a correct prediction across the entire dataset).
[b]FM: determines similarity between two clusterings to measure confusion matrices (e.g., higher value indicated greater similarity between clusters).
[c]Net benefit: minimum probability of disease at which intervention may be warranted (i.e., weighted calculation of true positives—false positives).
For this simulated example, 2 million patients are randomly assigned to group A or B. Both groups are generated by the exact same model: log odds = $\alpha + \beta * x + \gamma * I\{Group = B\}$, but group A is a low prevalence group; group B is a high prevalence group. The results in the table show various measures of accuracy, testing the data-generating model—i.e., a model that is known to have no invalidity in either of the two groups. For both groups, the observed average prediction in both groups matches the predicted outcome rate; Harrell's Es (i.e., the distance between the predicted and the observed, averaged across all predictions) shows perfect calibration. The area under the receiver operator curve (AUC) is also similar across groups. Despite equal discrimination and perfect calibration, all threshold-dependent measures of model performance are unequal across the two groups, owing to their prevalence-sensitivity, and—if relied upon—may misleadingly suggest subgroup invalidity.

of the variable race may diminish its prognostic value and generalizability across settings and databases. Further, we anticipate that racial and ethnic categories will continue to evolve over time, and in particular may become progressively less distinct. Additionally, while we feel the distinction between polar and non-polar contexts is frequently useful, we note that distinguishing between these may not always be clear-cut, particularly as non-polar decisional contexts may become polar when resources are constrained. Finally, we note that the predictive effect of race in models predicting risk and the effect in models predicting life expectancy or benefits may not be congruent.[15,60] Thus, inclusion of race in risk modeling and life-expectancy modeling should be separately considered. Mindful of the above caveats, we favored an approach that underscores procedures to evaluate bias and fairness and to weigh trade-offs, rather than one that prescribes a particular outcome across all use cases, or one that prioritizes one fairness definition over others (Item 19 in Table 5).

Debates about how to best address bias and fairness and the trade-off between anticlassification and antisubordination principles have been at the forefront of many aspects of life and law in the United States; they are unlikely to be definitively settled in a position statement. Ethical dilemmas by their very nature involve conflicting terms and therefore require balancing benefits and harms, and specifics are likely to matter. We acknowledge prior work that has sought to offer formal approaches to fairness that satisfy the principle of "equal opportunity" and strive to avoid "disparate treatment"[61,62]. However, these are fundamentally causal definitions of fairness, which are challenging to satisfy in practice because causality is generally unidentifiable in observational data (without strong unverifiable assumptions), and because race might be inadvertently reconstructed

Box 6 | Label bias

Label bias (or label choice bias) arises in the presence of a mismatch between the outcome (dependent) variable as it is ascertained and the 'ideal' outcome that, in theory, should be driving decision-making, particularly when the degree of mismatch varies systematically across race or ethnicity. This can lead to disparities across groups that are spurious, due for example to differences in outcome ascertainment. A non-medical example is using arrest as a proxy for criminality, which can bias algorithms predicting recidivism if over-policing selectively affects particular communities. A medical example is the use of healthcare cost as a proxy for healthcare needs[27]. This was shown to result in predictions that systematically under-estimate need in Black versus White patients.

Label bias has unique features, making it an insidious and important potential cause of bias and unfairness. Because the bias is "baked into the data", it is difficult to uncover this bias by examining usual measures of model performance. Models that appear most accurate may propagate this bias, and including race may exacerbate the bias by improving the prediction of a spurious difference[89]. Guidance for exploring and mitigating label bias is emerging[28].

The potential for label bias in healthcare is ubiquitous. For example, under-diagnosis has been commonly found to affect disadvantaged groups[90], and over-diagnosis has been documented in the affluent[91], which can bias CPMs to under-estimate true disease risk in the under-served for diseases like cancer and diabetes[91]. Despite these issues, minoritized communities (including Black and Asian patients) are at *higher*-risk for diabetes and some cancers than White patients (e.g., colon, prostate, lung). Thus, even when the causes of a risk difference or disparity is incompletely understood, it is often implausible to attribute this difference in risk to label bias. In these examples, the plausible direction of any label bias is in the opposite direction of the disparity and there are many other potential explanations available for the observed risk differences.

Table 4 | Guidelines: premises and recommendations for the inclusion of race or ethnicity in non-polar clinical predictions

Item	Statement
16	The hallmark of a non-polar prediction is that it is used only to optimize an individual's outcomes or align a decision with a patient's own values and preferences. (P9)
17	Race or ethnicity may be included in non-polar models if (and only if) predictive effects are independent from other ascertainable attributes, statistically robust, and clinically meaningful (i.e., can alter decision-making in some patients). (R11)[a]

P denotes premise, *R* recommendation.

[a]This recommendation reflects the fact that, in some circumstances, inclusion of race predictive effects can make predictions more accurate, particularly in groups comprising a smaller proportion of the population, since omission of a race variable will yield "average" predictions more reflective of the majority population. The benefits of more accurate prediction (Box 4) may need to be balanced against other considerations, as discussed in Box 3.

through proxies even when not explicitly encoded, particularly when high-dimensional machine learning approaches are applied. Thus, we offer a pragmatic approach based on an assessment of observable outcomes that seeks to maximize benefits for the population (utilitarianism) and at the same time to reduce disparities (egalitarianism).

Future work should encourage more routine use of variables for which race may be a proxy—such as social determinants of health or genetic ancestry; better collection of more representative training data; and evolution in how health systems populate electronic health records and other healthcare databases to ensure these data consistently reflect self-reporting. We note that CMS is putting regulatory pressure behind the collection of data on social drivers of health, with quality measures that require screening in five domains: food insecurity, housing instability, transportation needs, utility difficulties, and interpersonal safety.

In conclusion, the GUIDE provides a framework for identifying, understanding and deliberating about the trade-offs inherent in these issues when developing CPMs. We present it to support those developing or implementing CPMs in their goal of providing unbiased predictions to support fair decision-making, and for the broader community to better understand these issues.

Methods

We convened a Technical Expert Panel (TEP, Supplementary Table 2) to develop guidance for considering whether and how to use race in CPMs, including technical approaches to evaluating and "debiasing" models to ensure accurate predictions across racial categories (subgroup validity), and to complement other existing guidelines, such as the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD)[58] and Prediction model Risk Of Bias ASsessment Tool (PROBAST)[63]. Our expert consensus process followed the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) Network's recommendations for guideline development[64] and used a modified Delphi process[65] with five virtual consensus meetings and asynchronous feedback after each meeting with the TEP, alongside two focus groups with patient participants (Supplementary Table 3). The project was approved by the Tufts Health Sciences Institutional Review Board. Informed consent was obtained from participants.

Initial item list generation

An initial list of candidate items was developed based on expert opinion informed by the conceptual framework[26], review of the peer-reviewed and gray literature of algorithmic bias and fairness issues in CPMs and CPM development, and a review of works citing and cited by a set of core papers[27,66–71].

Expert recruitment

We identified key stakeholder groups: health system leadership, clinicians, engineers/computer scientists, clinical prediction modelers (using both classical statistical and machine learning approaches), medical informaticians, ethicists, lawyers, health disparities scientists, methodologists, policy experts, statisticians, trialists, and patients (Supplementary Table 2).

Experts were invited based on professional expertise, authorship of key documents, reputation, and policy-making experience. TEP membership reflected disciplinary, racial, gender, age, and geographic diversity.

Delphi process and consensus meetings

We used a modified Delphi process with five 2-hour web-based video consensus meetings, held between November 2020 and November 2021. The first meeting elicited preliminary feedback on candidate items, which was used to revise items and draft more recommendations and reviewed and endorsed fairness distinctions between polar and non-

Table 5 | Guidelines: premises and recommendations for prediction models used to allocate health resources

Item	Statement
18	While accurate prediction can guide optimally efficient resource allocation, accurate prediction does not ensure (or preclude) fair decisions. (P10)
19	There is no universally accepted unitary concept of fairness, and different fairness criteria conflict. Nevertheless, justice and fairness are foundational principles of health resource allocation. (P11)
20	For prediction models used to allocate resources, model evaluation should include its potential impact on resource distribution across racial or ethnic subgroups (i.e., a "fairness assessment"). (R12)
21	As a guiding principle, algorithms should neither exacerbate nor ignore existing disparities. (P12)
22	When predictions are used in the process of allocating health resources, inclusion of race as a model variable should be determined principally by the goal of reducing disparities. (R13)
23	Fairness assessment should be done with population samples reflecting the target population, since fairness results may not generalize across different settings. (R14)
24	Fairness should be continuously audited, with corrective adjustments made to achieve predetermined (or evolving) fairness goals. (R15)
25	When predictions are used to support resource allocation and distributive justice principles conflict, procedural justice, such as stakeholder-engaged processes, offer a means of achieving fair processes for deliberation and decision-making. (P13)
26	Fairness requires prediction modelers to integrate ethical principles in developing their model, including when selecting inputs, sourcing data, and selecting and assessing outcomes. Modelers should examine whether any individuals or groups, for example by race and ethnicity, will be made worse off as a result of the algorithm's design and to identify and attempt to mitigate unintended consequences. (P14)
27	When predictions are used in allocating health resources, accurate prediction and fair decision-making are distinct processes, requiring different expertise. In general, prediction constitutes only one of several potential inputs in a decision-making process. (P15)[a]
28	When predictions are used in allocating health resources, the locus of ethical responsibility is shared between the prediction model developers and the end-user (decision-maker). (P16)
29	Modelers assume a larger share of ethical responsibility for ensuring fairness when model outputs directly allocate resources (e.g., deterioration alarms, or allocation models). (P17)
30	In general, models used for resource allocation should employ logic that is open to human scrutiny. (R16)
31	When end-users assume the responsibility for ensuring distributive fairness, at minimum, model developers should: ensure transparent models (so that predictions are driven by clinically relevant variables), ensure subgroup validity, report any other fairness evaluation, and ensure models are adaptable to local or end-user needs. (R17)

P denotes premise, *R* recommendation.

[a]For example, other inputs might include considerations related to restorative justice (identifying and prioritizing the needs of populations who have experienced past harms owing to societal and institutional discrimination), incentivizing prosocial behavior (such as rewarding living donors who subsequently need organ transplantation), or distributive considerations, such as ensuring equitable access to different population subgroups, or prioritizing the needs of the most vulnerable.

polar contexts. In the second through fifth rounds, we convened the TEP to facilitate voting, discussion, deliberation, and re-voting on specific items (Tables 1–5). Using MeetingSphere, a cloud-based collaborative platform, a TEP co-chair (JKP or KL) alongside a professional facilitator (Mark Adkins, PhD) moderated consensus-building and voting. At each meeting, TEP co-chairs (DK, JKP, KL) presented the topic with illustrative cases uniquely developed for that meeting, followed by a first round of voting. Using a 5-point scale (1-Strongly Disagree to 5-Strongly Agree, or abstain), experts were asked to rate their level of agreement with the item's importance and feasibility of assessment, and to offer comments. For each item, the vote (rating) was carried out anonymously using the MeetingSphere software, after which ratings and comments were shared with the TEP in real time.

Deliberation and discussion followed the first round of voting at each meeting. To be included, ratings on items had to have "broad agreement", or exceed the pre-specified supermajority threshold of 75% of the TEP endorsing the item as "agree" or "strongly agree" (4 or 5), excluding abstentions. A supermajority, rather than a simple majority, was required to prevent the majority from eroding the influence of minority voices, without requiring strict unanimity for all items. Items without broad agreement were always discussed and revised, and TEP members could nominate additional items to be considered, based on comments or ratings. Deliberations included refinement, clarification, and improvement of items; dissenting views were acknowledged and incorporated where possible. Revised items were then voted on a second time.

Following meetings, experts had the opportunity to refine items and revise their judgments prior to subsequent meetings where re-rating occurred. All analyses of item scores and comments were performed independently by the professional facilitator using MeetingSphere. Conflicts were resolved by consensus.

At the final TEP meeting, the TEP reviewed all items, discussed and agreed to the content and final wording of the guidelines. The final GUIDE represents points of convergence across the TEP who held diverse opinions and approaches, especially to mitigating bias in shared decision-making contexts.

The TEP divided the GUIDE into two item types: premises, which are statements that are agreed upon but require no action; and recommendations, which are statements that offer direct guidance to modelers and/or users of CPMs. The items were organized by topic, decisional context (polar/resource allocation versus non-polar/shared decision-making), stage of model development, and implications for fairness in implementation and dissemination. A glossary (Supplementary Table 1) was also produced to clarify key terms and concepts used in the GUIDE.

Patient focus groups

Patients were convened in focus groups to inform guidelines and assess acceptability. Stakeholder patient panelists were recruited from a stakeholder group provided by the Tufts Clinical and Translational Science Institute (CTSI) Community Stakeholder Engagement Core. Purposive sampling criteria included diversity of age, gender, race, ethnicity, medical history, and education (Supplementary Table 3)[72]. Each 2-hour session was conducted via Zoom. Sessions were led by experts in qualitative research (KL or JKP), following semi-structured discussion guides, using cases to illustrate conceptual aspects (e.g., CPMs, race-aware versus unaware, polar versus non-polar contexts), and elicit stakeholder feedback. Sessions were audio-recorded and professionally transcribed. Content analysis[73] was performed, and these data provided insight into the values and reasoning underlying the opinions of patient stakeholders pertaining to inclusion of race in CPMs. Patient stakeholder feedback was presented to the TEP for incorporation in the final GUIDE during the final meeting.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Received: 18 December 2023; Accepted: 31 August 2024;
Published online: 19 October 2024

References

1. Inker, L. A. et al. New creatinine- and cystatin C-based equations to estimate GFR without race. *NEJM* **385**, 1737–1749 (2021).
2. Eneanya, N. D., Yang, W. & Reese, P. P. Reconsidering the consequences of using race to estimate kidney function. *JAMA* **322**, 113–114 (2019).
3. Lu, J. H. et al. Assessment of adherence to reporting guidelines by commonly used clinical prediction models from a single vendor: a systematic review. *JAMA Netw. Open* **5**, e2227779 (2022).
4. Wright, J. L. et al. Eliminating race-based medicine. *Pediatrics* **150**, e2022057998 (2022).
5. Borrell, L. N. et al. Race and genetic ancestry in medicine—a time for reckoning with racism. *NEJM* **384**, 474–480 (2021).
6. Bonham, V. L., Callier, S. L. & Royal, C. D. Will precision medicine move us beyond race? *NEJM* **374**, 2003–2005 (2016).
7. Chokshi, D. A., Foote, M. M. K. & Morse, M. E. How to act upon racism-not race-as a risk factor. *JAMA Health Forum* **3**, e220548 (2022).
8. Ehdaie, B., Carlsson, S. & Vickers, A. Racial disparities in low-risk prostate cancer. *JAMA* **321**, 1726–1727 (2019).
9. Fletcher, S. A. et al. Geographic distribution of racial differences in prostate cancer mortality. *JAMA Netw. Open* **3**, e201839 (2020).
10. Aggarwal, R. et al. Diabetes screening by race and ethnicity in the United States: equivalent body mass index and age thresholds. *Ann. Intern Med.* **175**, 765–773 (2022).
11. Khor, S. et al. Racial and ethnic bias in risk prediction models for colorectal cancer recurrence when race and ethnicity are omitted as predictors. *JAMA Netw. Open* **6**, e2318495 (2023).
12. Gail, M. H. et al. Projecting individualized absolute invasive breast cancer risk in African American women. *J. Natl Cancer Inst.* **99**, 1782–1792 (2007).
13. Zink, A., Obermeyer, Z. & Pierson, E. Race adjustments in clinical algorithms can help correct forracial disparities in data quality. *Proc. Natl Acad. Sci. USA* **121**, e2402267121 (2024).
14. Bonner, S. N. et al. Clinical implications of removing race-corrected pulmonary function tests for African American patients requiring surgery for lung cancer. *JAMA Surg.* **158**, 1061–1068 (2023).
15. Landy, R. et al. Methods for using race and ethnicity in prediction models for lung cancer screening eligibility. *JAMA Netw. Open* **6**, e2331155 (2023).
16. Landy, R. et al. Using prediction-models to reduce persistent racial/ethnic disparities in draft 2020 USPSTF lung-cancer screening guidelines. *J. Natl Cancer Inst.* **113**, 1590–1594 (2021).
17. Manski, C. F. Patient-centered appraisal of race-free clinical risk assessment. *Health Econ.* **10**, 2109–2114 (2022).
18. Hougen, H. Y. et al. Adding a coefficient for race to the 4K score improves calibration for black men. *J. Urol.* **211**, 392–399 (2024).
19. Manski, C. F., Mullahy, J. & Venkataramani, A. S. Using measures of race to make clinical predictions: decision making, patient health, and fairness. *Proc. Natl Acad. Sci. USA* **120**, e2303370120 (2023).
20. Chohlas-Wood, A., Coots, M., Goel, S. & Nyarko, J. Designing equitable algorithms. *Nat. Comput Sci.* **3**, 601–610 (2023).
21. Wawira Gichoya, J., McCoy, L. G., Celi, L. A. & Ghassemi, M. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ Health Care Inf.* **28**, e100289 (2021).
22. Lin, J. S. et al. Addressing racism in preventive services: a methods project for the U.S. preventive services task force. (Agency for Healthcare Research and Quality, Rockville, MD, 2021).
23. Research Protocol: impact of healthcare algorithms on racial and ethnic disparities in health and healthcare. (Effective Health Care Program, Agency for Healthcare Research and Quality, Rockville, MD, Rockville, MD, 2022).
24. Evans, C. V., Johnson, E. S. & Lin, J. S. Assessing Algorithmic Bias and Fairness in Clinical Prediction Models for Preventive Services: A Health Equity Methods Project for the U.S. Preventive Services Task Force. (Agency for Healthcare Research and Quality, Portland, OR, 2023).
25. Section 1557 of the Affordable Care Act. in *42* (ed. Office for Civil Rights) (Office for Civil Rights, Washington, DC, 2022).
26. Paulus, J. K. & Kent, D. M. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ Digit. Med.* **3**, 99 (2020).
27. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
28. Obermeyer, Z. et al. Algorithmic Bias Playbook. (Center for Applied Artificial Intelligence, Chicago, IL, 2021).
29. Baeza-Yates, R. & Matthews, J. Statement on principles for responsible algorithmic systems. (ACM Technology Policy Office, Washington, DC, 2022).
30. Bates, D. W. How to regulate evolving AI health algorithms. *Nat. Med.* **29**, 26 (2023).
31. Daniels, N. Accountability for reasonableness. *BMJ* **321**, 1300–1301 (2000).
32. Majerol, M. & Hughes, D. L. CMS Innovation Center Tackles Implicit Bias. In *Health Affairs Forefront* (Health Affairs, 2022).
33. Khazanchi, R., Tsai, J., Eneanya, N. D., Han, J. & Maybank, A. Leveraging Affordable Care Act Section 1557 to address racism in clinical algorithms. in *Health Affairs Forefront* (Health Affairs, 2022).
34. Keith, K. HHS proposes revised ACA anti-discrimination rule. In *Health Affairs Forefront* (Health Affairs, 2022).
35. Shachar, C. & Gerke, S. Prevention of bias and discrimination in clinical practice algorithms. *JAMA* **329**, 283–284 (2023).
36. Ross, C. Amid the AI gold rush, a new company forms to vet models and root out weaknesses. In *STAT* (STAT News, 2023).
37. Schmidt, H., Gostin, L. O. & Williams, M. A. The Supreme Court's rulings on race neutrality threaten progress in medicine and health. *JAMA* **330**, 1033–1034 (2023).
38. Harris, E. National health care leaders will develop AI code of conduct. *JAMA* **330**, 401 (2023).
39. Goodman, K. E., Morgan, D. J. & Hoffmann, D. E. Clinical algorithms, antidiscrimination laws, and medical device regulation. *JAMA* **329**, 285–286 (2023).
40. Centers for Medicare & Medicaid Services, Office for Civil Rights, Office of the Secretary & Department of Health and Human Services. Nondiscrimination in health programs and activities. Fed. Register 87, 47824–47920 (2022).
41. Blueprint for an AI bill of rights: making automated systems work for the American people. (ed. White House Office of Science and Technology Policy) (United States Government, Washington, DC, 2022).
42. de Hond, A. A. H. et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *NPJ Digit Med.* **5**, 2 (2022).
43. Sikstrom, L. et al. Conceptualising fairness: three pillars for medical algorithms and health equity. *BMJ Health Care Inf.* **29**, e100459 (2022).
44. Vasey, B. et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* **377**, e070904 (2022).

45. Huang, J., Galal, G., Etemadi, M. & Vaidyanathan, M. Evaluation and mitigation of racial bias in clinical machine learning models: scoping review. *JMIR Med. Inf.* **10**, e36388 (2022).
46. Feng, J. et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. *NPJ Digit Med.* **5**, 66 (2022).
47. Zou, J., Gichoya, J. W., Ho, D. E. & Obermeyer, Z. Implications of predicting race variables from medical images. *Science* **381**, 149–150 (2023).
48. Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing (HTI-1) Proposed Rule. Vol. HTI-1 (ed. Office of the National Coordinator for Health Information Technology) (Office of the National Coordinator for Health Information Technology, Washington, DC, 2023).
49. Medlock, A. & Cooke, D. T. Removing structural racism in pulmonary function testing-why nothing is ever easy. *JAMA Surg.* **158**, 1069 (2023).
50. Suriyakumar, V. M., Ghassemi, M. & Ustun, B. When personalization harms performance: reconsidering the use of group attributes in prediction. *PMLR* **202**, 33209–33228 (2023).
51. Coots, M., Saghafian, S., Kent, D. & Goel, S. Reevaluating the role of race and ethnicity in diabetes screening. In *arXiv* 2306.10220 [stat.AP] (arXiv, 2023).
52. Khazanchi, R., Marcelin, J., Abdul-Mutakabbir, J. & Essien, U. Race, racism, civil rights law, and the equitable allocation of scarce COVID-19 treatments. In *Health Affairs Forefront* (Health Affairs, 2022).
53. Kent, D. M., Ladin, K. & Duru, O. K. Equal treatment for equal risk: should race be included in allocation algorithms for Covid-19 therapies? In *STAT* (STAT News, 2022).
54. Shaikh, N. et al. Reassessment of the role of race in calculating the risk for urinary tract infection: a systematic review and meta-analysis. *JAMA Pediatr.* **176**, 569–575 (2022).
55. Grobman, W. A. et al. Prediction of vaginal birth after cesarean delivery in term gestations: a calculator without race and ethnicity. *Am. J. Obstet. Gynecol.* **225**, 664.e661–664.e667 (2021).
56. Inserro, A. Flawed racial assumptions in eGFR have care implications in CKD. In *Am J Manag Care* (The American Journal of Managed Care, 2020).
57. Steyerberg, E. W. et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med.* **10**, e1001381 (2013).
58. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* **350**, g7594 (2015).
59. Diao, J. A. et al. Implications of race adjustment in lung-function equations. *NEJM* **390**, 2083–2097 (2024).
60. Zhu, J., Brenna, C. T. A., McCoy, L. G., Atkins, C. G. K. & Das, S. An ethical analysis of clinical triage protocols and decision-making frameworks: what do the principles of justice, freedom, and a disability rights approach demand of us? *BMC Med. Ethics* **23**, 11 (2022).
61. Basu, A. Use of race in clinical algorithms. *Sci. Adv.* **9**, eadd2704 (2023).
62. Loftus, J. R., Russell, C., Kusner, M. J. & Silva, R. Causal reasoning for algorithmic fairness. In *arXiv* 1805.05859v05851 [cs.AI] (arXiv, 2018).
63. Wolff, R. F. et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann. Intern Med.* **170**, 51–58 (2019).
64. Moher, D., Schulz, K. F., Simera, I. & Altman, D. G. Guidance for developers of health research reporting guidelines. *PLoS Med.* **7**, e1000217 (2010).
65. Dalkey, N. & Helmer, O. An experimental application of the delphi method to the use of experts. *Manag. Sci.* **9**, 458–467 (1963).
66. Gianfrancesco, M. A., Tamang, S., Yazdany, J. & Schmajuk, G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med.* **178**, 1544–1547 (2018).
67. Park, Y. et al. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. *JAMA Netw. Open* **4**, e213909 (2021).
68. Rajkomar, A., Hardt, M., Howell, M., Corrado, G. & Chin, M. Ensuring fairness in machine learning to advance health equity. *Ann. Intern Med.* **169**, 866–872 (2018).
69. Char, D. S., Shah, N. H. & Magnus, D. Implementing machine learning in health care - addressing ethical challenges. *NEJM* **378**, 981–983 (2018).
70. Parikh, R. B., Teeple, S. & Navathe, A. S. Addressing bias in artificial intelligence in health care. *JAMA* **322**, 2377–2378 (2019).
71. Flanagin, A., Frey, T. & Christiansen, S. L. Updated guidance on the reporting of race and ethnicity in medical and science journals. *JAMA* **326**, 621–627 (2021).
72. Patton, M. Q. *Qualitative Research & Evaluation Methods* (SAGE Publications, Thousand Oaks, CA, 2002).
73. Saldana, J. *The Coding Manual for Qualitative Researchers* (SAGE Publications, London, 2013).
74. Office of Management and Budget, Office of Information and Regulatory Affairs & Executive Office of the President. Revisions to OMB's statistical policy directive no. 15: standards for maintaining, collecting, and presenting federal data on race and ethnicity. Fed. Register 89, 22182–22196 (2024).
75. Chouldechova, A. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data* **5**, 153–163 (2017).
76. Kleinberg, J., Mullainathan, S. & Raghavan, M. Inherent trade-offs in the fair determination of risk scores. in *arXiv* 1609.05807v05802 [cs. LG] (arXiv, 2016).
77. Balkin, J. M. & Siegel, R. B. The American civil rights tradition: anticlassification or antisubordination? *Univ. Miami Law Rev.* **58**, 9–34 (2003).
78. Arnold, K. F. et al. Reflection on modern methods: generalized linear models for prognosis and intervention-theory, practice and implications for machine learning. *Int. J. Epidemiol.* **49**, 2074–2082 (2021).
79. Ramspek, C. L. et al. Prediction or causality? A scoping review of their conflation within current observational research. *Eur. J. Epidemiol.* **36**, 889–898 (2021).
80. Westreich, D. & Greenland, S. The Table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *Am. J. Epidemiol.* **177**, 292–298 (2013).
81. King, M. L., Jr. 'I Have A Dream' Speech, In Its Entirety. In *Talk of the Nation* (NPR.org, 1963).
82. Corbie-Smith, G., Thomas, S. B. & St George, D. M. Distrust, race, and research. *Arch. Intern Med.* **162**, 2458–2463 (2002).
83. Armstrong, K. et al. Differences in the patterns of health care system distrust between blacks and whites. *J. Gen. Intern Med.* **23**, 827–833 (2008).
84. Jenkins, K. A., Keddem, S., Bekele, S. B., Augustine, K. E. & Long, J. A. Perspectives on racism in health care among black veterans with chronic kidney disease. *JAMA Netw. Open* **5**, e2211900 (2022).
85. Armstrong, K. et al. Prior experiences of racial discrimination and racial differences in health care system distrust. *Med. Care* **51**, 144–150 (2013).
86. Kent, D. M. et al. An electronic health record-compatible model to predict personalized treatment effects from the diabetes prevention program: a cross-evidence synthesis approach using clinical trial and real-world data. *Mayo Clin. Proc.* **97**, 703–715 (2022).
87. Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R. & Goel, S. The measure and mismeasure of fairness. *JMLR* **24**, 1–117 (2023).
88. Van Calster, B. et al. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J. Clin. Epidemiol.* **74**, 167–176 (2016).
89. Zanger-Tishler, M., Nyarko, J. & Goel, S. Risk scores, label bias, and everything but the kitchen sink. *Sci. Adv.* **10**, eadi8411 (2024).
90. Geiger, H. J. Racial and ethnic disparities in diagnosis and treatment: a review of the evidence and a consideration of causes. in *Unequal*

Treatment: Confronting Racial and Ethnic Disparities in Health Care (eds. Smedley, B. D., Stith, A. Y. & Nelson, A. R.) 417–454 (National Academies Press (US), Washington (DC), 2003).
91. Welch, H. G. & Fisher, E. S. Income and cancer overdiagnosis—when too much care is harmful. *NEJM* **376**, 2208–2209 (2017).

Acknowledgements

The authors appreciate the research assistance of Rebecca Maunder and Hannah McGinnes, and voting facilitation from Mark Adkins, PhD. The team greatly appreciates the participation and input of patients who participated in this study, and two Technical Expert Panel members (Keith Norris, MD, PhD and Kayte Spector-Bagdady, JD, MBE) who participated in guideline development, as well as reviewed and provided comments for this work. The team also appreciates Jason Nelson, MPH for assistance with the Figure in Box 4. Research reported in this publication was funded through a "Making a Difference" and Presidential Supplement Awards from The Greenwall Foundation (PI Kent). The views presented in this publication are solely the responsibility of the authors and do not necessarily represent the views of the Greenwall Foundation. The Greenwall Foundation was not involved in the design of the study; the collection, analysis, and interpretation of the data; and the decision to approve publication of the finished manuscript.

Author contributions

D.M.K. and K.L. drafted the manuscript. D.M.K., J.K.P., and K.L. oversaw and designed the study; D.M.K., J.K.P, and K.L. obtained funding for the research. All authors (K.L., J.C., O.K.D., S.G., W.H., J.G.P., J.K.P., J.S., R.S., E.S., B.U., D.vK., S.N.W., D.M.K.) contributed to development of conclusions, and reviewed and contributed significantly to the final manuscript.

Competing interests

The authors declare the following competing interests: Dr. Duru declares no Competing Financial Interests but the following Competing Non-Financial Interests as a consultant for ExactCare Pharmacy®, research funding from the Patient Centered Outcomes Research Institute (PCORI), the Centers for Disease Control and Prevention (CDC), and the National Institutes of Health (NIH). Dr. Kent declares no Competing Financial Interests but Competing Non-Financial Interests in research funding from the Greenwall Foundation, W.L. Gore, Patient Centered Outcomes Research Institute (PCORI), and the National Institutes of Health (NIH). Dr. Ladin declares no Competing Financial Interests but Competing Non-Financial Interests in research funding from Paul Teschan Research Fund #2021-08, Dialysis Clinics Inc. (DCI), and from the Greenwall Foundation. Dr. Steyerberg declares no Competing Financial Interests but Competing Non-Financial Interests in funding from the EU Horizon program (4D Picture project, #101057332). Dr. Ustun declares no Competing Financial Interests but Competing Non-Financial Interests in research funding from the National Science Foundation IIS 2040880, the NIH Bridge2AI Center Grant U54HG012510. All other authors declare no Competing Financial or Non-Financial Interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01245-y.

Correspondence and requests for materials should be addressed to David M. Kent.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Research on Ethics, Aging and Community Health (REACH Lab), Medford, MA, USA. [2]Departments of Occupational Therapy and Community Health, Tufts University, Medford, MA, USA. [3]American Medical Group Association, Alexandria, VA, USA. [4]Department of Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA. [5]Harvard Kennedy School, Harvard University, Cambridge, MA, USA. [6]Department of Medicine, Tufts Medical Center, Boston, MA, USA. [7]Predictive Analytics and Comparative Effectiveness Center, Tufts Medical Center, Boston, MA, USA. [8]OM1, Boston, MA, USA. [9]Department of Medicine, Stanford Medicine, Stanford, CA, USA. [10]Center for Individualized Medicine Bioethics, Mayo Clinic, Rochester, MN, USA. [11]Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, Netherlands. [12]Halıcıoğlu Data Science Institute, University of California San Diego, San Diego, CA, USA. [13]Erasmus University Medical Centre, Rotterdam, Netherlands. [14]Tufts Clinical and Translational Science Institute, Tufts University, Boston, MA, USA. ✉e-mail: David.Kent@tuftsmedicine.org