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

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
5600 Fishers Lane
Rockville, MD 20857
www.ahrq.gov

Prepared by:
Kaiser Permanente Evidence-based Practice Center
Kaiser Permanente Center for Health Research
Portland, OR

Investigators:
Corinne V. Evans, MPP
Eric S. Johnson, PhD
Jennifer S. Lin, MD, MCR

AHRQ Publication No. 23-05308-EF-1
August 2023

Acknowledgments
We thank David Kent, MD, CM, MSC, and Keren Ladin, PhD, MSc, from Tufts Medical Center and the following individuals from the ECRI-University of Pennsylvania EPC for their contributions to this report: Shazia Siddique, MD, MSHP; Brian Leas, MS; Nikhil Mull, MD; Michael Harhay, PhD; Gary Weissman, MD, MSHP; Harald Schmidt, PhD; and Jaya Aysola, MD, MS.

Contents

Introduction

Clinical prediction models are tools that combine multiple predictors to estimate the risk or probability that a specific disease or condition is present (diagnostic) or that a specific clinical outcome will occur in the future (prognostic).[1] Several prognostic clinical prediction models are invoked by the United States Preventive Services Task Force (USPSTF) to inform decision making for the initiation of preventive services, such as statin and aspirin use to prevent cardiovascular disease,[2-4] screening for breast cancer and chemoprevention,[5,6] and screening for osteoporosis.[7] All of these examples are "race-aware" clinical prediction models, meaning that race is included as a predictor or stratifying characteristic.[8-11] To date, there is no consensus regarding if, when, and how best to use race and ethnicity as a predictor in clinical prediction models.[12] A number of recent publications have highlighted examples of race-aware clinical prediction models with potential harm of diverting care away from populations experiencing health inequities.[13,14] A 2018 analysis showed that race was included in just 3% of cardiovascular-related clinical prediction models,[15] yet the most commonly used clinical prediction model in current U.S. practice for primary cardiovascular disease prevention, the Pooled Cohort Equations (PCE), is stratified by race.[8] The inclusion of race in clinical prediction models is motivated by the potential to improve predictive accuracy when prognostic differences exist between racial groups, acknowledging that the mechanisms for these prognostic differences are complex, multifactorial, and not necessarily biologic.[16] Reasons for excluding race in clinical prediction models are likewise compelling. These include a desire to avoid racial profiling such as examples from other clearly objectionable contexts such as law enforcement, wanting to avoid elevating race from a poorly defined social construct to biologic predictor, and avoiding using race as a poor proxy for biological or other risk factors.[16]

The past several years have witnessed an explosion of interdisciplinary interest in evaluating the consequences of including race and ethnicity in clinical prediction models and questions about whether health inequities are reinforced or exacerbated by the use of "race-aware" clinical prediction models.[13,14] Concerns about whether prediction models—also referred to as algorithms—outside of the healthcare context are reinforcing inequities have been investigated more universally in the artificial intelligence field—including machine learning—as such tools have a growing reach in many aspects of modern life.[17] Machine learning is a method where models iteratively learn from data, identify data patterns, and automate model building;[18] thus, this method selects predictors and may identify predictors that are proxies for race and ethnicity. In contrast, most regression models typically seen in clinical prediction models include race and ethnicity because of investigator intention. Given the inclusion of several "race-aware" clinical prediction models in the USPSTF portfolio, these concerns are directly relevant to several topics. As a continuation of methods work around health equity for the USPSTF, we conducted this project to address gaps in the literature for evaluating algorithmic bias and fairness in clinical prediction models as they relate to race and ethnicity. We acknowledge that there are health inequities in addition to those specific to race and ethnicity; however, this is the focus of the present work.

We have intentionally used the term "race-aware" clinical prediction model rather than the term "race-based medicine." "Race-based medicine" has been defined in various ways, such as "the system by which research characterizing race as an essential, biological variable, translates into clinical practice, leading to inequitable care."[14] In a 2022 Policy Statement, the American Academy of Pediatrics (AAP) describes the term "race-based" medicine as meaning "the misuse of race as a corrective or risk-adjusting variable in clinical algorithms or practice guidelines."[19] We acknowledge the varied examples of misuse of race and ethnicity in prediction models. We chose a more agnostic term for this specific work because our goal was to investigate upstream factors to interrogate the rationale, mechanisms, and implications of inclusion of race and ethnicity variables in clinical prediction models. We further

wanted to allow for the possibility that race and ethnicity were not being included as stand-ins for biology and that there may be possible circumstances where "race-aware" models may be responsive in directing resources to communities experiencing inequities. We agree with the AAP that "the effects of racism require consideration in clinical decision-making tools in ways that are evidence informed and not inappropriately conflated with the limiting phenotype of race categorization."[19]

Methods

Our overall aim was to develop guidance that the Evidence-based Practice Centers (EPCs) and USPSTF could use to evaluate algorithmic bias and fairness considerations for topics and recommendations involving race-aware clinical prediction models. To this end, we began our work by identifying and synthesizing foundational literature of the main concepts addressed in this guidance. We did not use a systematic literature review approach. Instead, we started our scan of the literature using references from a presentation given to the USPSTF Health Equity Workgroup in Fall 2021, explored the reference lists of those citations, and supplemented with other articles suggested by experts as well as recently published editorials.

For algorithmic bias, our goal was to develop and pilot an extended set of signaling questions to identify algorithmic bias and applicability concerns in the context of race-aware models for clinical preventive services that can be used alongside an existing critical appraisal tool, such as the Prediction model Risk Of Bias ASessment Tool (PROBAST).[20,21] Using PROBAST as the foundation, we then added 11 equity-based signaling questions focused on the inclusion of race and ethnicity–specific data. In designing these additional signaling questions, we envisioned that the PROBAST tool would be applied in its entirety by a systematic review team that included expertise in clinical prediction models. We then piloted the tool on four clinical prediction models. We additionally incorporated feedback from the ECRI-Penn EPC, which is conducting a related systematic review for AHRQ's Effective Health Care Program titled *Impact of Healthcare Algorithms on Racial Disparities in Health and Healthcare*.[22] Additionally, the signaling questions were reviewed by nine experts as part of an eDelphi process. All items had an agreement rate of greater than 70%. Using feedback from this process, one item was removed, the wording of four items were changed, and the rationale text was modified for six items.

For fairness, our goal was to develop a set of questions to guide decision makers through qualitative evaluation of potential threats to fairness when considering the implementation of a particular algorithm. We envisioned this step being completed by guideline makers after risk of bias and algorithmic bias are assessed by the systematic review team. Because of the complexity of implementation and clinical context, a potentially limited set of validated clinical prediction model alternatives, and the possibility for the same clinical prediction model to be recommended for different treatment decisions (e.g., aspirin and statin use for cardiovascular disease prevention), we avoided a rigid approach and intended to create a flexible discussion guide to help articulate the limitations of a model. The discussion guide was created by synthesizing and extending upon the limited amount of fairness criteria available to decision makers that did not take the form of mathematical criteria.[23-25]

Framework and Definitions

Background

Bias and fairness are concepts addressed in a host of disciplines beyond healthcare to include ethics, law, insurance, finance, computer science and machine learning, and psychometrics, just to name a few. The multidisciplinary character of these terms and growing use of algorithms in varied applications has led to a "proliferation of terminology, rediscovery and simultaneous discovery, conflicts between

disciplinary perspectives"[26] and what other scholars have described as "wildly inconsistent motivations, terminology, and notation, presenting a serious challenge for cataloging and comparing definitions."[17]

For the purposes of this work, we began with a framework grounded in the work of Paulus and Kent[12] that delineates between algorithmic bias and fairness. **Figure 1** illustrates key conceptual differences between algorithmic bias and fairness as they relate to this work. Simply stated, *algorithmic bias* refers narrowly to attributes intrinsic to a model that may result in differential model performance in different groups, whereas *fairness* refers more broadly to downstream outcomes and whether algorithmic decisions create discriminatory or unjust impacts in different populations. Definitions of fairness in the published literature are quite complex, with no single agreed upon definition.[17] Thus, we selected an expansive definition for fairness that emphasizes downstream outcomes. This definition is from *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms*, developed by the consortium for Fairness, Accountability, and Transparency in Machine Learning (FAT/ML).[23]

Algorithmic bias

In the development to implementation pathway of a clinical prediction model, there are several points at which health equity concerns can be introduced. There are upstream considerations related to model development and data attributes itself that could result in systematically worse predictive performance in specific racial and ethnic groups. We refer to this as algorithmic bias, where issues related to model design, data, and sampling may disproportionately affect prediction model performance in specific racial or ethnic groups.[12] Because model performance is quantified using a number of different measures, assessments of algorithmic bias may differ for each measure with respect to population characteristics (i.e., case mix) and varying condition prevalence or incidence. For example, discrimination can be higher in samples with greater variation in the explanatory variable.[27] While differences in discrimination measures are expected in different groups,[12] accurate models will always have good calibration across groups. This differential predictive performance has the potential to cause under or overuse of preventive services in specific populations. Technical solutions in model development may be available to mitigate some differential performance issues. An empirically validated instrument, PROBAST, is available to assess the risk of bias and applicability of clinical prediction models.[20,21,28] The Quality Assessment of Prognostic Accuracy Studies (QUAPAS) tool is a related instrument published during the course of our work that addresses prognostic tests more generally (prognostic studies involving single biomarkers, multimarker scores, imaging, or other methods).[29] PROBAST, on the other hand, is directly applicable to prediction models and their specific development methods.

In systematic reviews of clinical prediction models, risk of bias is assessed to evaluate whether best practices were used in model design, conduct, and analysis. PROBAST does not explicitly address risk of bias from including social risk factors such as race or ethnicity. Risk of bias is a broader concept than algorithmic bias, which is specific to calibration and other performance measures and can be measured directly. Standards for addressing this type of bias in the context of machine learning are beginning to emerge[30-32] but are potentially less relevant to the type of clinical prediction models currently used in primary care and recommended by clinical practice guidelines.

Fairness

Beyond the narrower and technical aspects of development that are intrinsic to a clinical prediction model, there are broader downstream considerations for how model implementation could contribute to health inequities even if it is at minimal risk for algorithmic bias. We refer to this broader and normative concept as fairness, which addresses whether algorithmic decisions create discriminatory or unjust impacts in different populations. In the healthcare context, an example of unfairness is the

allocation of intensive disease management resources away from a sociodemographic group experiencing health inequities. While we have carried forward the term fairness,[12,23] we acknowledge that there are limitations to this term. It has been noted that the term "fairness" does not address the interplay between power relations in society and outcomes and instead has a limited emphasis on outcomes.[33]

In addition to establishing whether algorithmic decisions are creating or reinforcing inequities, it is important to consider multiple mechanisms of how a clinical prediction model might contribute to inequities and how to mitigate this risk. These considerations could include: the availability of alternative decision-making criteria in usual care, communication of risk, issues around decision thresholds, and limitations of the model that do not appear as performance issues. In the context of imperfect clinical prediction models that may not be readily modified, deliberate and upfront consideration of model implementation and contextual factors may help to articulate and address limitations. To mitigate the potentially harmful outcomes of a clinical prediction model, the model itself could be adjusted, or its implementation modified. It may be that the latter is more feasible. Further, depending on the availability of alternative models or decision-making criteria in usual care for a clinical scenario, it would be valuable for transparency to justify the selection of a race-aware vs. race-unaware clinical prediction model.

As noted previously, there is a rich and multidisciplinary history to various definitions and mathematical criteria for fairness; however, none of these are adequate for our purposes. Two primary limitations of these criteria are well described by Paulus and Kent[12] and include the mutual incompatibility of fairness criteria as well as the greater relevance of fairness in specific decision-making contexts. Both are discussed further below. To these reasons, we further add that rigid criteria may not be acceptable in cases when an imperfect algorithm may be the best available option and a guideline body serves as a crucial intermediary in guiding implementation in a complex clinical context.

Mutual incompatibility of fairness criteria. There are three broad fairness criteria that have their origins in the education testing and psychometrics literature from the 1960s and 1970s, and much of this literature formalizes fairness quantitatively.[26,34] These criteria have been given different names and a multitude of variations on these criteria have been proposed.[26] Brief definitions are provided below as they relate to fairness considerations centering on race and ethnicity.

- Independence (demographic parity, statistical parity, group fairness, or disparate impact): the algorithm's score is independent of race
- Separation (equalized odds, conditional procedure accuracy, or avoiding disparate mistreatment): the algorithm's score is independent of race given true health state
- Sufficiency (calibration within groups or conditional use accuracy): the true health state is independent of race given the algorithm's score

It has been shown mathematically, however, that these criteria are mutually incompatible.[17,26] Specifically, when outcome rates (e.g., disease incidence) differ in two groups, consistent calibration and error rates are mutually conflicting.[12] Thus, it can be difficult to decide which definition of fairness to use when a decision maker is implementing an algorithm.[35] Scholars have noted that certain definitions of fairness can increase discrimination[35] and that there may be a tradeoff between fairness and accuracy.[36,37]

Fairness concerns may not be equally applicable in all decision-making contexts. Paulus and Kent suggest that fairness concerns are most salient in the context of competing interests, or what they term "polar" decisions, where one pole of a prediction is associated with a clear benefit or harm.[12] In the medical context, this could apply to the allocation of scarce resources such as expensive medications or specialized services,[38-40] where it is in a patient's interest to be scored high and thus receive the medication or service. For example, who will receive an available donor organ[12] or be able to participate in a resource-intensive care-management program.[38] In the preventive services context, where patients and providers have a shared goal of accurate prediction to balance the benefit and harm of an inexpensive clinical action (e.g., aspirin), decisions are likely less "polar" in nature.[12] Our three examples of clinical prediction models recommended by the USPSTF, including cardiovascular disease risk prediction to guide preventive therapies, breast cancer risk prediction, and osteoporosis risk prediction, are more non-polar than polar.

The consideration of polarity, however, may not be straightforward. While some decisions are clearly more polar than others, the perspective of the individual and the relative scarcity or cost of the service are relevant. While clinical prediction algorithms such as the Framingham Risk Score and FRAX may have initially been used to target use of once expensive therapies such as statins and bisphosphonates to the highest-risk individuals while drugs were on patent, in the context of much cheaper off-patent drugs, we consider these to be non-polar decisions. Additionally, to some individuals, "more" may be thought of as better because of patient preferences, even when a model may suggest net harm for a particular decision threshold. Harms could include cost, adverse effects, or downstream consequences such as delays in access to other medical treatments. Further, some applications of a clinical prediction model may suggest a narrow and more polar decision space of "receive" vs. "do not receive" a service, when in fact a larger decision space may be available.[17] For example, receipt of one intervention instead of another or immediate vs. delayed receipt of an intervention. Regardless of polarity, if a clinical prediction model consistently allocates preventive services away from groups experiencing the greatest burden of disease or disparate access to care, this would not be fair.

Mitigating algorithmic bias and unfairness

The first step in mitigating algorithmic bias and fairness concerns is to determine whether the decision is shared decision making or polar, where the model is used for rationing of scarce services (**Figure 2**). In non-polar decisions, statistical solutions may be available to reduce algorithmic bias. On the other hand, assessing fairness is more complex. As previously noted, various quantitative criteria can be evaluated, but a more comprehensive view looks not just at metrics of predictive accuracy but at direct harm and downstream health outcomes, requiring different types of evidence extrinsic to a model or validation, to evaluate. These study designs could include comparative effectiveness trials of an algorithm vs. usual care or microsimulation modeling. Paulus and Kent emphasize the concept of algorithmic bias in non-polar decisions where the reduction of bias (improvement in predictive accuracy in each group) promotes accurate estimates of net benefit. In contrast, fairness is central to polar decisions where there is the greatest risk of allocation harm because of scarcity.[26] Given the complexity of fairness, it follows that addressing fairness concerns in models applied for rationing can involve a wider range of solutions and methodologic tools. These could include the use of causal models that avoid race or race proxies, the application of different decision thresholds to different groups, or systematic reclassification of individuals to equalize allocation between groups.[12]

An imperfect clinical prediction model may be the best available option. In determining whether to recommend a particular clinical prediction model, the essential question is "compared with what?" in usual care and its evidence for bias and fairness. We acknowledge that decision makers may conclude that among various alternatives (e.g., using another clinical prediction model that is not race-aware, or not using multivariate risk prediction), the best option is an imperfect model. In a situation where the clinical prediction model is not readily modifiable, decision makers should be armed with tools to help articulate its limitations, with possible mitigation strategies for implementation. Such guidance for decision makers is nascent. The recent reassessment of the inclusion of race in glomerular filtration rate (GFR) estimation by the National Kidney Foundation and American Society of Nephrology is the most robust example to date.[24,25] A task force convened by these groups selected and used six attributes to evaluate model alternatives: assay availability and standardization; implementation; population diversity in equation development; performance compared with measured GFR; consequences to clinical care, population tracking, and research; and patient centeredness.

Translating evidence to decision making. Clinical prediction models are not proscriptive in a way that algorithms in other contexts may be; for example, credit scores providing automated decisions around lending. Instead, in the clinical context, clinician judgement or competing decision criteria influence when and how an algorithm supports decision making. As such, guideline developers are a crucial intermediary between the availability of a clinical prediction model and its implementation. While evidence reviewers (e.g., EPCs conducting systematic reviews) should appraise a clinical prediction model's risk of bias and algorithmic bias, guideline bodies should consider threats to fairness in their deliberations and recommendations.

Guidance to decision makers concerning fairness is also available from the machine learning community;[23] however, is nascent. Guiding questions include: 1) Are there particular groups which may be advantaged or disadvantaged, in the context in which you are deploying, by the algorithm/system you are building? 2) What is the potential damaging effect of uncertainty/errors to different groups?

Algorithmic Bias

Development of health equity signaling questions for critical appraisal

We developed and piloted an extended set of signaling questions to the PROBAST critical appraisal tool to enable systematic reviewers to identify algorithmic bias specific to race and ethnicity in clinical prediction models. Algorithmic bias refers to issues related to model design, data, and sampling that may disproportionately affect prediction model performance in specific populations, such as those classified by race or ethnicity.[12] Algorithmic bias is a distinct concept from overall risk of bias of a clinical prediction model, which refers broadly to whether a clinical prediction model's results are flawed, and this risk of bias can come about from a number of reasons (e.g., selection bias, measurement error, or model optimism). When applying PROBAST and the health equity signaling questions, the reviewer obtains information on risk of bias broadly and algorithmic bias. If reporting on a model's development methodology is sparse, as is most often the case for older models, then there may be inadequate information to determine the overall risk of bias. Further, if validation studies do not report model performance by race and ethnicity, then algorithmic bias cannot be determined. When prediction model performance is robustly reported by race and ethnicity, with the use of observed to expected ratios, for example, then the magnitude and direction of miscalibration can be assessed numerically.

Our work focuses on clinical prediction models that include race and ethnicity as a predictor or stratifying variable; however, the exclusion of race and ethnicity from a model does not ensure lack of

bias or fairness. Most of the signaling questions apply to prediction models not including race and ethnicity as a predictor, and evaluation of the signaling questions in these cases is currently underway in a systematic review funded by AHRQ's Effective Health Care Program.

Table 1 presents the health-equity specific signaling questions with accompanying rationale and considerations. The organization of the questions follows the PROBAST structure, which consists of four domains: participants, predictors, outcome, and analysis.

Piloting of health equity signaling questions

We piloted the tool on three clinical prediction models used in preventive services and mentioned in USPSTF recommendations: the PCE, the Breast Cancer Risk Assessment Tool (BCRAT), and FRAX®. We additionally piloted QFracture, which is not addressed by the USPSTF, so that the signaling questions could be tested on a model published after the establishment of modern reporting requirements.[1]

General Findings

Our general findings were that the completion of the full PROBAST tool with extension took a few hours and that methodological expertise in clinical prediction models was needed, particularly for the analysis domain. If a limited number of prediction models were being evaluated in a review, it would be feasible to complete the extended PROBAST assessment for all model development papers. The approach to evaluating external validation studies for a particular model should be carefully considered. A systematic review approach is ideal but could be highly time intensive. Other approaches could be to apply a subset of questions to all validation papers or to establish a priori selection criteria for which validation studies would be appraised; for example, the largest, most applicable, and most recent external validation conducted in the United States.

PROBAST is a high standard even without the addition of health equity signaling questions, particularly for models published prior to modern reporting guidelines, such as TRIPOD.[1] This finding for PROBAST has been documented elsewhere: a 2021 validation study reported that 96% (98/102) of evaluated models were at high risk of bias according to PROBAST.[28] Important signaling questions with no information in our pilots were: 1) consistency of categorization of racial and ethnic groups where the handling of ethnicity was sometimes unclear and no model addressed multiracial populations, 2) differential missingness of predictor data by race and ethnicity with problematic complete case analyses, and 3) differential followup by race and ethnicity. Important methodological limitations related to health equity questions and beyond were: 1) differential life expectancy without use of a competing risks model, 2) overfitting and optimism were not addressed, 3) lack of confidence intervals for predicted and observed events limited precision of calibration comparisons, and 4) external validation studies using routine care databases were problematic due to differing outcome ascertainment methods than prospective studies with a protocol and event adjudication.

Pooled Cohort Equation Detailed Findings

Here we provide additional detail on the pilot for the PCE[8] to illustrate the challenges in assessing the risk of bias and algorithmic bias and explore model limitations with respect to race and ethnicity. We further describe opportunities for further refinements to directly address potential risks of bias and algorithmic bias as they relate to the inclusion of race. The PCE was developed prior to modern reporting guidelines for multivariable prediction models;[1] therefore, several items could not be assessed because no information was provided: the proportion of individuals in the development data set with missing data, the potential for differential followup, and model optimism.

Assessment of the PCE using new signaling questions also identified issues that could contribute to algorithmic bias. Because the PCE was not developed with a competing risks model,[47] and Black

Americans suffer a higher age-specific all-cause mortality, the predicted 10-year probabilities of a cardiovascular event from the PCE's Cox model may be overestimated—an overestimation that would be worse than in White Americans. Further, smaller effective sample sizes (numbers of events for Black individuals) likely led to model overfitting in equations for this population. Further, the use of multiple imputation to handle missing data would be preferable for reducing selection bias; instead, the PCE excluded participants with missing predictors. If there are differences in the number of participants with missing data by race, a further selected and less representative sample would be used. While not required by PROBAST, the lack of confidence intervals for expected to observed events precludes firm conclusions about how calibration compares in Black and White individuals.

A key clinical question is whether the PCE is differentially miscalibrated in Black Americans.[12] Our findings were inconclusive. Inconsistency in who is experiencing worse model performance across validation cohorts, and especially the lack of confidence intervals for calibration data in validation datasets, limits precise conclusions. However, design features of the PCE are consistent with increased overprediction, such as overfitting in the Black population as well as no use of a competing risk model in the context of known life expectancy differences. Additionally, the lack of specific PCE equations for Hispanic/Latino, Asian, and Native American populations raise critical questions about the populations to whom the PCE is applicable. The lack of equations in these populations arose from limited longitudinal outcome data from which to derive equations; this represents a foundational evidence gap. Developing equations in the context of very limited data would have resulted in poor model performance with subsequent algorithmic bias. The lack of robust longitudinal cohort data in these populations raises deeper equity questions about in whom research is funded.

Fairness

We developed a discussion guide for considering fairness and health equity when making recommendations involving clinical prediction models. **Table 2** presents a set of critical questions to assist guideline developers in identifying, and potentially mitigating, fairness concerns for recommended clinical prediction models. These questions are envisioned as a discussion guide for deliberations specific to topics involving risk assessment and are designed to be considered after risk of bias and algorithmic bias have been assessed in a more technical and less normative sense during the evidence synthesis phase of a topic. These questions are not intended as a formula to determine whether a tool, when implemented, is unjust, but instead to help articulate potential concerns. The applicability and fairness of a clinical prediction model require judgement in relation to a specific clinical decision defined by the relevant population, intervention, outcomes, setting, and other factors. Guideline developers may choose to explicitly address items from the question prompts in this discussion guide in recommendation statements or guideline documentation. Consideration of any one model's fairness should be in the context of a larger body of evidence for a topic, which may include multiple risk prediction models, each with their own specific limitations or known fairness concerns regarding usual care without an algorithm.

Conclusions

The pilot testing of PROBAST health equity signaling questions suggests that all or nearly all clinical prediction models relevant to the USPSTF's portfolio would be scored as high risk of bias, with many items being unscoreable for models published before TRIPOD reporting guidelines. Rather than using this as grounds for not recommending such models, our conclusions are that this tool can help to articulate limitations of available models. Ideally, the best available clinical prediction model can be contrasted with usual care, which is often not formally evaluated for risk of bias, but may carry the

potential for implicit bias, lack of transparency, incompleteness, or concerns about reproducibility. The critical appraisal process could further help to identify specific areas of future research related to risk prediction. Further, our recommendation that performance measures be reported in specific populations may result in the increasing inclusion of race and ethnicity as a predictor variable.

Our discussion guide to assess fairness is grounded in a similar approach—in the context of imperfect models, it is important to be able to comprehensively consider up front possible fairness concerns for a model as they relate to clinical context and implementation. As stated before, to mitigate the potential negative effects of a race-aware clinical prediction model (or a race-unaware model with known bias), the model itself could be adjusted, or its implementation modified.

This work is meant to foster discussion and consideration on criteria for assessing bias when race and ethnicity are included in clinical prediction models, with a specific focus on the context of recommended models in preventive services. We acknowledge that this is a rapidly evolving field, with concurrent ongoing work from other scholars, including a parallel review funded by AHRQ's Effective Health Care Program.

Further discussion among experts should address the following questions:

- Is improvement in calibration alone sufficient rationale for the inclusion of race in a prediction model? If the inclusion of race and ethnicity as a predictor variable improves predictive accuracy for a specific group, does that potential downstream benefit outweigh possible negative consequences of including it?
- When is it not appropriate to include race and ethnicity as a predictor? Are there situations where it would be exceptionally clear at the outset that its inclusion could be harmful?

References

1. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). *Ann Intern Med*. 2015;162(10):735-736. doi:10.7326/L15-5093-2
2. US Preventive Services Task Force. Statin use for the primary prevention of cardiovascular disease in adults: US Preventive Services Task Force recommendation statement. *JAMA*. 2022;328(8):746-753. doi:10.1001/jama.2022.13044
3. US Preventive Services Task Force. Statin use for the primary prevention of cardiovascular disease in adults: US Preventive Services Task Force recommendation statement. *JAMA*. 2016;316(19):1997-2007. doi:10.1001/jama.2016.15450
4. US Preventive Services Task Force. Aspirin use to prevent cardiovascular disease: US Preventive Services Task Force recommendation statement. *JAMA*. 2022;327(16):1577-1584. doi:10.1001/jama.2022.4983
5. U.S. Preventive Services Task Force. Screening for breast cancer: US Preventive Services Task Force recommendation statement. *Ann Intern Med*. 2016;164(4):279-296. doi:10.7326/M15-2886
6. US Preventive Services Task Force. Medication use to reduce risk of breast cancer: US Preventive Services Task Force recommendation statement. *JAMA*. 2019;322(9):857-867. doi:10.1001/jama.2019.11885
7. US Preventive Services Task Force. Screening for osteoporosis to prevent fractures: US Preventive Services Task Force recommendation statement. *JAMA*. 2018;319(24):2521-2531. doi:10.1001/jama.2018.7498
8. Goff DC Jr, Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation*. 2014;129(25 Suppl 2):S49-S73. doi:10.1161/01.cir.0000437741.48606.98
9. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst*. 1989;81(24):1879-1886. doi:10.1093/jnci/81.24.1879
10. Kanis JA, Johansson H, Oden A, Dawson-Hughes B, Melton LJ 3rd, McCloskey EV. The effects of a FRAX revision for the USA. *Osteoporos Int*. 2010;21(1):35-40. doi:10.1007/s00198-009-1033-8
11. Kanis JA on behalf of the World Health Organization Scientific Group. Assessment of Osteoporosis at the Primary Health-Care Level. Technical Report. 2007. Accessed August 16, 2023. https://www.sheffield.ac.uk/FRAX/pdfs/WHO_Technical_Report.pdf
12. Paulus JK, Kent DM. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ Digit Med*. 2020;3:99. doi:10.1038/s41746-020-0304-9
13. 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(9):874-882. doi:10.1056/NEJMms2004740
14. Cerdena JP, Plaisime MV, Tsai J. From race-based to race-conscious medicine: how anti-racist uprisings call us to act. *Lancet*. 2020;396(10257):1125-1128. doi:10.1016/S0140-6736(20)32076-6
15. Paulus JK, Wessler BS, Lundquist CM, Kent DM. Effects of race are rarely included in clinical prediction models for cardiovascular disease. *J Gen Intern Med*. 2018;33(9):1429-1430. doi:10.1007/s11606-018-4475-x
16. Paulus JK, Kent DM. Race and ethnicity: a part of the equation for personalized clinical decision making? *Circ Cardiovasc Qual Outcomes*. 2017;10(7). doi:10.1161/CIRCOUTCOMES.117.003823

17. Mitchell S, Potash E, Barocas S, D'Amour A, Lum K. Algorithmic fairness: choices, assumptions, and definitions annual review of statistics and its application. *Annu Rev Stat Its Application*. 2021;8(1):141-163.
18. Andaur Navarro CL, Damen JA, Takada T, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. *BMC Med Res Methodol*. 2022;22(1):12. doi:10.1186/s12874-021-01469-6
19. Wright JL, Davis WS, Joseph MM, et al. Eliminating race-based medicine. *Pediatrics*. 2022;150(1). doi:10.1542/peds.2022-057998
20. Wolff RF, Moons KG, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51-58. doi:10.7326/M18-1376
21. Moons KG, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med*. 2019;170(1):W1-W33. doi:10.7326/M18-1377
22. Agency for Healthcare Research and Quality. Research Protocol: Impact of Healthcare Algorithms on Racial and Ethnic Disparities in Health and Healthcare. 2022. Accessed August 16, 2023. https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/protocol
23. Diakopoulos N, Friedler S, Arenas M, et al. Principles for Accountable Algorithms and a Social Impact Statement for Algorithms. https://www.fatml.org/resources/principles-for-accountable-algorithms
24. Delgado C, Baweja M, Burrows NR, et al. Reassessing the inclusion of race in diagnosing kidney diseases: an interim report from the NKF-ASN task force. *J Am Soc Nephrol*. 2021;32(6):1305-1317. doi:10.1681/ASN.2021010039
25. Delgado C, Baweja M, Crews DC, et al. A unifying approach for GFR estimation: recommendations of the NKF-ASN task force on reassessing the inclusion of race in diagnosing kidney disease. *Am J Kidney Dis*. 2022;79(2):268-288e1. doi:10.1053/j.ajkd.2021.08.003
26. Barocas S, Hardt M, Narayanan A. Fairness and Machine Learning. Limitations and Opportunities. 2021. Accessed August 16, 2023. https://fairmlbook.org/
27. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Med Res Methodol*. 2012;12:82. doi:10.1186/1471-2288-12-82
28. Venema E, Wessler BS, Paulus JK, et al. Large-scale validation of the Prediction model Risk Of Bias ASsessment Tool (PROBAST) using a short form: high risk of bias models show poorer discrimination. *J Clin Epidemiol*. 2021;138:32-39. doi:10.1016/j.jclinepi.2021.06.017
29. Lee J, Mulder F, Leeflang M, Wolff R, Whiting P, Bossuyt PM. QUAPAS: an adaptation of the QUADAS-2 tool to assess prognostic accuracy studies. *Ann Intern Med*. 2022;175(7):1010-1018. doi:10.7326/M22-0276
30. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA*. 2019;322(24):2377-2378. doi:10.1001/jama.2019.18058
31. U.S. Food and Drug Administration, Health Canada, and U.K. Medicines and Healthcare Products Regulatory Agency. Good Machine Learning Practice for Medical Device Development: Guiding Principles. 2021. Accessed August 16, 2023. https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles
32. Schwartz R, Down L, Jonas A, Tabassi E. A Proposal for Identifying and Managing Bias in Artificial Intelligence. Draft NIST Special Publication 1270. 2021. Accessed August 16, 2023. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270-draft.pdf

33. American Medical Association and Association of American Medical Colleges. Advancing Health Equity: Guide on Language, Narrative and Concepts. 2021. Accessed August 16, 2023. https://www.ama-assn.org/about/ama-center-health-equity/advancing-health-equity-guide-language-narrative-and-concepts-0
34. Barocas S, Hardt M. Fairness in Machine Learning. Accessed August 16, 2023. https://vimeo.com/248490141
35. Kusner M, Loftus J, Russell C, Silva R. Counterfactual Fairness. Ppresented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017; Long Beach, CA, USA. Accessed August 16, 2023. https://arxiv.org/pdf/1703.06856.pdf
36. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness. ITCS '12: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. 2012:214-226. doi:10.1145/2090236.2090255
37. Barocas S, Selbst AD. Big data's disparate impact. *California Law Review*. 2016.
38. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453. doi:10.1126/science.aax2342
39. Eberly LA, Richterman A, Beckett AG, et al. Identification of racial inequities in access to specialized inpatient heart failure care at an academic medical center. *Circ Heart Fail*. 2019;12(11):e006214. doi:10.1161/CIRCHEARTFAILURE.119.006214
40. Peterson PN, Rumsfeld JS, Liang L, et al. A validated risk score for in-hospital mortality in patients with heart failure from the American Heart Association get with the guidelines program. *Circ Cardiovasc Qual Outcomes*. 2010;3(1):25-32. doi:10.1161/CIRCOUTCOMES.109.854877
41. Lett E, Asabor E, Beltran S, Michelle Cannon A, Arah OA. Conceptualizing, contextualizing, and operationalizing race in quantitative health sciences research. *Ann Fam Med*. 2022;20(2):157-163. doi:10.1370/afm.2792
42. Institute of Medicine. *Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement*. National Academies Press; 2009.
43. ASHG Denounces attempts to link genetics and racial supremacy. *Am J Hum Genet*. 2018;103(5):636. doi:10.1016/j.ajhg.2018.10.011
44. Obermeyer Z, Nissan R, Stern M, Eaneff S, Bembeneck EJ, Mullainathan S. *Algorithmic Bias Playbook*. 2021.
45. Whittle J, Conigliaro J, Good CB, Lofgren RP. Racial differences in the use of invasive cardiovascular procedures in the Department of Veterans Affairs medical system. *N Engl J Med*. 1993;329(9):621-627. doi:10.1056/NEJM199308263290907
46. Steyerberg EW. *Clinical Prediction Models. A Practical Approach to Development, Validation, and Updating.* 2nd ed. Springer; 2019.
47. Wolbers M, Koller MT, Witteman JC, Steyerberg EW. Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology*. 2009;20(4):555-561. doi:10.1097/EDE.0b013e3181a39056
48. Emanuel E, Schmidt H, Steinmetz A. *Rationing and Resource Allocation in Healthcare. Essential Readings*. Oxford University Press; 2018.
49. Wynants L, van Smeden M, McLernon DJ, et al. Three myths about risk thresholds for prediction models. *BMC Med*. 2019;17(1):192. doi:10.1186/s12916-019-1425-3

Figure 1. Framework for consideration of algorithmic bias and fairness in race-aware clinical prediction models

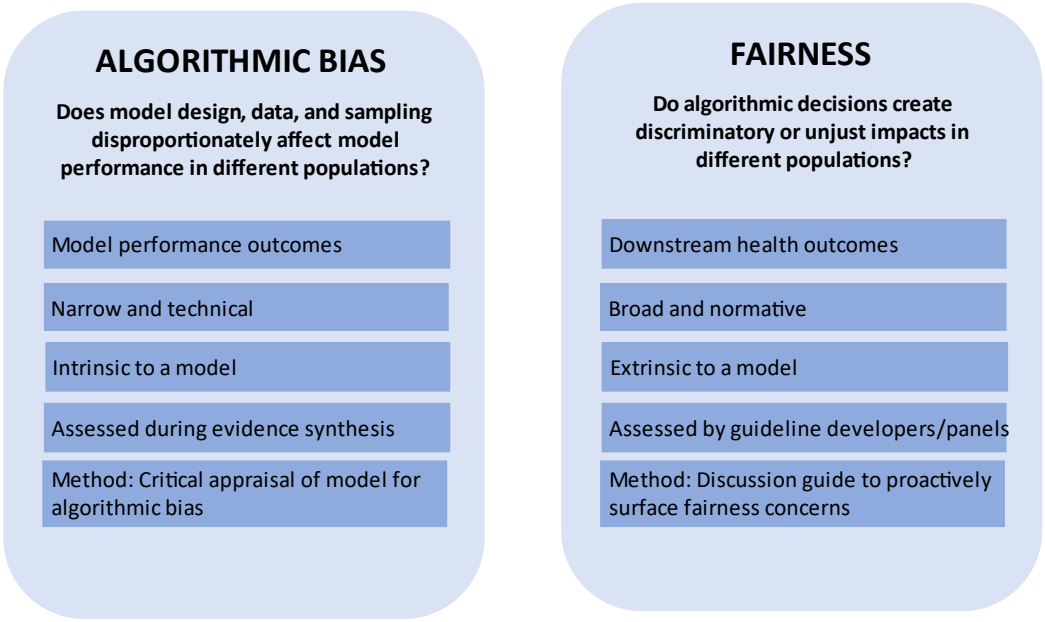

Figure 2. Mitigating algorithmic bias and unfairness in clinical decision making; reproduced from Paulus and Kent 2020[12]

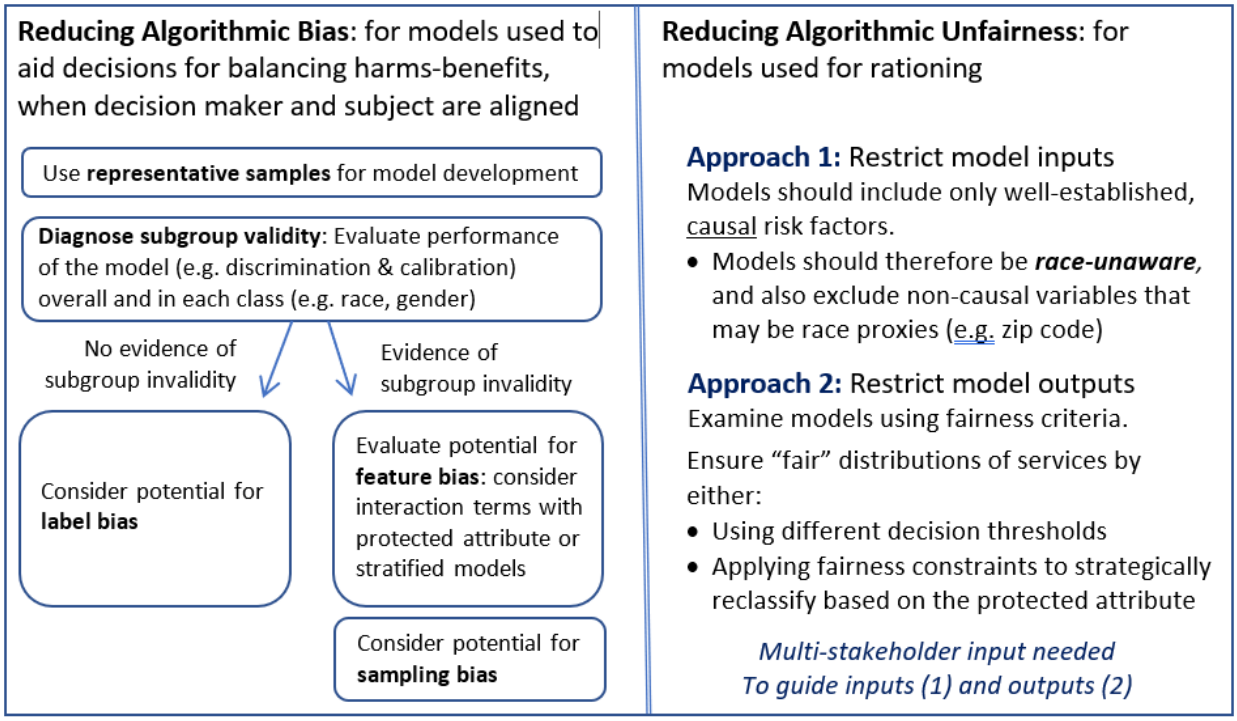

Table 1. Signaling questions to identify risks of bias in race-aware clinical prediction models

PROBAST Domain	Added Signaling Questions	Rationale and Considerations
1. Participants	1.3a Were data on racial and ethnic groups gathered using consistent definitions or categories with adequate response options?	The classification of race and ethnicity is complex and there are not best practices for collecting these data.[41] For example, self-reported race, race as categorized by others, and family racial history each reflect distinct meanings of racism. In the absence of best practices, at minimum, the collection of racial and ethnic data should use consistent categories and definitions. Considerations include the availability of a category for multiracial/multiethnic individuals and whether heterogeneity within groups is addressed (e.g., the identification of Black or Indigenous heritage within the broader Hispanic category, country of origin, or immigration status).[41] Ideally, the collection of race and ethnicity should adhere to evidence-based REAL (Race, Ethnicity, and Language) standards with the ability to select multiple categories.[42]
	1.3b Was the racial and ethnic distribution of the population in the development data similar to the distribution in the target population?	Representation bias occurs when estimates from one population are inappropriately extrapolated to other populations.[32] Underrepresentation of racial and ethnic groups may contribute to differential predictive performance. The population distribution from development data could be compared to recent Census estimates at the national, state, or local level to assess representativeness in the target population.
	1.4 Were racial and ethnic groups classified/categorized in a similar way in the development data and population to whom model is applied? (Validation studies only)	Similar to 1.3a, but in this case, were the categories consistent between development and validation datasets, or was there further opportunity for misclassification?
2. Predictors	2.4a Was a transparent rationale provided for including race and ethnicity as a predictor?	In the absence of consensus or clear criteria for the inclusion of race in clinical prediction models,[12] at minimum, is the rationale for inclusion of race transparent? For example, do authors state that race and ethnicity was included to improve calibration because of known differences in incidence? While prediction modelers are not asked to justify the inclusion of other variables, it is appropriate in the case of race and ethnicity because of concerns regarding potential misuse. Race should not be included in models with a causal aim because the notion of racial and ethnic groups as genetically distinct has been invalidated.[43] For causal inference, the social construct of race could be decomposed into causal elements such as more direct measures of racism,[42] health care access, socioeconomic status, or biologic differences due to chronic stressors.
	2.5 Was differential missingness of predictor data in racial and ethnic groups reported?	Missing data may be associated with contact with the healthcare system, which in turn reflects care patterns and structural racism. Exclusion of participants with missing data provides a selective sample of data for model development.[21] Bias may be further

Table 1. Signaling questions to identify risks of bias in race-aware clinical prediction models

PROBAST Domain	Added Signaling Questions	Rationale and Considerations
		exacerbated in the context of differential missingness among groups. The more missing data there are in a population, the more selected and less representative the development data become when used in a complete case analysis where participants with missing data are excluded. Multiple imputation is the preferred method for handling missing data so that data from all participants can be included; this is further addressed in PROBAST item 4.4.[21]
3. Outcome	3.4a Was differential followup or ascertainment of the outcome in racial and ethnic groups reported?	Differential ascertainment of outcomes among groups may lead to systematic over or underprediction. Similarly, differential loss to followup among groups may result in differences in censoring among groups. Censoring can bias predicted risks because of overrepresentation of those experiencing the outcome.[21]
	3.7 Were proxy outcomes avoided as the predicted outcome, where the proxy may be subject to encoded bias (label choice bias)?	Label choice bias is a mismatch between the ideal target the algorithm should be predicting and a biased proxy variable the algorithm is actually predicting.[44] Proxy outcomes may reflect encoded bias where some racial and ethnic groups have experienced less access to a service and the use of this measure as a predicted outcome could reinforce inequities. For example, healthcare cost may reflect access rather than true health needs.[38] Similarly, revascularization as an event of interest may reflect practice patterns favoring intervention in specific groups.[45]
4. Analysis	4.1a Were there sufficient outcomes occurring in specific racial and ethnic groups to assess model performance separately in these groups? (Model validation studies)	The effective sample size in prediction models is the number of outcome events. This question asks whether the events per variable are adequate when assessed separately in racial and ethnic groups. Investigators have suggested 10 to 20 events per variable in development studies, but the actual calculation can be more complex.[21] In validation studies, at least 100 events is recommended.[46]
	4.6a Were competing risk methods used in the prediction model?	Overestimation and bias can occur in prediction models not accounting for prominent competing risks, such as all-cause mortality in elderly populations.[21] Standard predictions using Cox regression models can overestimate absolute risk because individuals with a competing event (e.g., all-cause death) are censored and treated as if the predicted outcome could occur in the future.[47] The potential for bias is further exacerbated in the context of well documented differential life expectancy among racial and ethnic groups. A similar question was recently added in the QUAPAS tool.[29]
	4.7a Were relevant model performance measures evaluated appropriately in racial and ethnic groups? How does model performance (calibration,	Both calibration and discrimination should be assessed and reported separately by racial and ethnic group, allowing for comparison among groups. Reported measures of calibration should be meaningful, such as calibration plots or expected to observed ratios, which allow for a quantification of the direction and magnitude of any miscalibration.

Table 1. Signaling questions to identify risks of bias in race-aware clinical prediction models

PROBAST Domain	Added Signaling Questions	Rationale and Considerations
	discrimination) compare amongst racial and ethnic groups?	Additional guidance available in Moons et al 2019.[21] Because model performance is quantified using a number of different measures, assessments of algorithmic bias may differ for each measure with respect to population characteristics (i.e., case mix) and varying condition prevalence or incidence. For example, discrimination can be higher in samples with greater variation in the explanatory variable.[27] While differences in discrimination measures are expected in different groups,[12] accurate models will always have good calibration across groups.

Table 2. Discussion guide for considering fairness and health equity when making clinical practice recommendations involving clinical prediction models

Domain	Critical Question

Background	• Are there known inequities in access to and quality of care and health outcomes (such as morbidity/mortality) from the condition?
	  ○ Which specific population(s) are most affected?
	  ○ Are there differences in the uptake of a service/intervention among different population(s)?

Clinical Decision Context	• To what extent is the prediction model used for resource allocation (prioritization of goods and services that are not severely limited) or rationing (prioritization of goods and services that are severely limited)?[48] • Is there effect modification by underlying risk (i.e., does the relative benefit or harm depend on underlying risk)? • Is there evidence to suggest that a decision using a clinical prediction model results in a more or less efficient or equitable allocation of services than a decision guided by other criteria (i.e., clinical judgement* or an alternative model)? • Is there evidence about the fairness of usual care without the use of an algorithm? • Are there alternative interventions that could be considered to reduce potential harm?
Model Performance and Limitations	• Does the risk model have differential model performance (calibration and discrimination) when assessed across racial and ethnic groups? • Are there data to suggest differential performance in models that include race and ethnicity as a predictor or stratifying factor compared to models that do not? • Are there concerns about the equitable availability of data needed to perform risk scoring, such as the availability of testing for novel risk markers? • Are there limitations to the model that do not appear as model performance issues; for example, equations are not available for all population groups, or the predicted outcome is subject to label bias?
Implementation of Risk Models and Use of Thresholds	• If there is a stated decision threshold, to what extent did guideline developers justify the decision threshold specific to clinical context? • To what extent is heterogeneity within population groups addressed in the communication of risk and considered in assigning threshold? • Is uncertainty of the decision threshold addressed? • If there are concerns about the ability to adequately risk score due to data availability or model limitations, what alternatives are available for scoring or communicating risk uncertainty? • Are there alternatives to risk thresholds[49] (e.g., presentation of continuous risk scores, context-specific thresholds, or establishing a range of acceptable thresholds) that could be considered to improve risk communication? • If miscalibration exists, what is the clinical significance and potential damaging effect to different groups? ○ What is the direction of bias (e.g., would miscalibration worsen or mitigate inequities among specific populations)? • If there are no ways of mitigating model performance differences (algorithmic bias) through modifications of the prediction model itself, are there other ways in which inequitable outcomes can be mitigated, such as application of supplementary decision criteria independent of the model? • As the prediction model is implemented: what arrangements will be put in place to monitor disparate impact, and possibly adjust use?

Table 2. Discussion guide for considering fairness and health equity when making clinical practice recommendations involving clinical prediction models

Research and Policy Gaps	• Are there important evidence gaps that should be addressed to improve the assessment of risk in specific population(s) to equitably guide care?

*Assumption is that clinician judgement for assessment of prognosis may have implicit bias, lack transparency, may not be reproducible, or could be incomplete.