Supplementary Information

Supplementary Table 1. Glossary

- **Anticlassification**: an antidiscrimination principle that holds that people should not be classified for differential treatment on the basis of a protected category (e.g., race), and seeks to eliminate the unfairness individuals experience due to bias arising from decision-makers' consideration of the protected category in question
- **Antisubordination**: an antidiscrimination principle that aims to ensure status-based equality across protected classes. The antisubordination principle sees differences in the status of minorities as the main concern and recognizes that completely 'race-blind' approaches to decisions may at times fail to alleviate and may even exacerbate disparities and unfairness in some circumstances.
- **Calibration**: refers to the agreement between observed outcomes and predictions
- **Clinical Prediction Model:** multiple predictors combined to estimate the presence of a specific condition (diagnostic) or risk of future event (prognostic) presented in the form of an equation, score, tool or related.
- **"Negatively" polar prediction:** prediction is used to target an intervention perceived as punitive or coercive (e.g., involuntary commitment, screening for child abuse, or quarantining patients at high infectious risk).
- **Non-polar prediction**: interests of subject and model user are aligned and goal is to maximize benefit and avoid harm for the individual patient, help a patient *balance* the benefits and harms of a decision to align decisions with their own values and preferences; the paramount interest to the patient is getting the most accurate prognosis.
- **Polar prediction**: subject has an interest in receiving a higher or lower/favorable or unfavorable prediction, rather than necessarily receiving the most accurate forecast
- **"Positively" polar prediction**: prediction for which patients may have an interest to be ranked high to receive a service that may be available only to some of those who can potentially benefit (e.g., allocation of scarce medical resources)
- **Race:** a construct used to divide people into groups based on physical appearance, ancestry, and sociocultural factors. Self-identification is the preferred means of obtaining this information. As a construct, it is not a direct cause of health outcomes. However, it may be an indirect cause of health outcomes through the health effects of racism. Additionally, it may be importantly predictive of health outcomes through correlated factors based on socioeconomic status, culture and genetic ancestry.
- **Race-aware:** is used to describe CPMs including race as a predictor variable
- **Race-unaware:** is used to describe CPMs that do not include race as a predictor variable.

Supplementary Table 2. Algorithmic Fairness Technical Expert Panel

Name	Affiliation	Expertise
John Cuddeback, MD, PhD	American Medical Group Association	Integration of predictive tools in health systems
O. Kenrik Duru, MD, MS	University of California Los Angeles	Health disparities, preventive medicine; shared decision making
Sharad Goel, PhD, MS	Harvard Kennedy School	Computer science and sociological impact, algorithmic fairness
William Harvey, MD, MSc	Tufts Medical Center	Health system leadership, clinical informatics integration, patient safety
David Kent, MD, MS*	Tufts Medical Center	Predictive modeling, comparative effectiveness research
Keren Ladin, PhD, MSc*	Tufts University	Medical ethics, health disparities research, health policy
Keith Norris, MD, PhD	University of California Los Angeles	Health policy, health disparities
Jessica K. Paulus, PhD*	OM1	Research methodology, comparative effectiveness research, real-world data
Joyce Sackey, MD	Stanford Medicine	Diversity and inclusion, medical innovations and impacts on marginalized groups
Richard Sharp, PhD, MA	Mayo Clinic	Biomedical ethics, integration of technologies in patient care, patient advocacy, individualized medicine
Kayte Spector-Bagdady, JD, MBE	University of Michigan	Law and bioethics, use of personal health data
Ewout Steyerberg, PhD	Leiden University Medical Centre	Predictive modeling, model evaluation, medical decision-making, statistical methods
Berk Ustun, PhD, MS	University of California San Diego	Machine learning, health informatics
Saul Weingart, MD, PhD, MPP	Tufts Medical Center	Clinical decision-making, healthcare management, predictive instruments for patient safety

*Technical expert panel co-chair

Supplementary Table 3. Stakeholder Participants

Characteristics	Session 1 participants (n = 10)	Session 2 participants (n = 7)
Age		
18-24	2	0
25-34	3	1
35-44	0	1
45-54	2	2
55-64	2	2
65+	1	1
Gender		
Female	8	5
Male	2	2
Race		
Asian or Asian American	2	1
Black or African American	3	3
Native Hawaiian/Pacific Islander	0	0
White or Caucasian	4	2
Other	1	1
Ethnicity		
Hispanic/Latino	2	1
Not Hispanic/Latino	8	6
History of Chronic Disease		
Yes	NC	4
No	NC	3
Education		
High school diploma/GED	NC	1
Associate or technical degree	NC	2
Some college	NC	2
Professional/advanced degree(s)	NC	2
Healthcare Work History		
Yes	NC	4
No	NC	3

NC indicates not collected.

Supplementary Table 4. TEP Voting Summary

Item	Agreement* (%)	Agree/Disagree		Statements
		Mean	SD	
1	92	4.25	0.60	**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 weak 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	92	4.58	1.11	**Distinction between predictive and causal inference** In understanding the use of race and other protected characteristics in clinical prediction models, it is important not to conflate the goal of predictive inference (which depends only on correlations) with 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	90	4.3	0.90	**Goals of clinical predictive inference** 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)
	80	4.2	0.98	B. Clinical prediction models can also be used to support efficient resource allocation to maximize population-wide benefits when resources are constrained. (P4)
	80	4.3	0.78	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)
4	77	4.38	0.84	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	83	3.92	0.76	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). (R1) Modelers should report clearly how race was obtained and defined in their sample. (R2)
6	75	3.83	1.21	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†	73	3.73	1.14	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)
	100	4.45	0.50	

Item	Agreement* (%)	Agree/Disagree		Statements
		Mean	SD	
8†	80	4.0	1.10	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 and 4) and for fairness (see Table 5) concerns. (R4)
	75	3.5	1.38	
9†	70	3.5	1.38	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)
	100	4.22	0.42	
10	100	4.73	0.45	Samples used for prediction model derivation should represent the underlying population consistent with intended use. (R6)
11	91	4.73	0.62	Prediction model development should adhere to best practice guidance[56,57,62], 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	91	4.36	0.64	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†	55	3.73	0.75	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)
	75	4.23	0.80	
14	75	4.0	0.95	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	75	4.09	0.79	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)
16	91	4.09	0.51	The hallmark of a non-polar prediction is that it is used only to optimize an individual's outcomes or align a decision with patient's own values and preferences. (P9)

Item	Agreement* (%)	Agree/Disagree		Statements
		Mean	SD	
17	82	4.45	0.78	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)
18†	100	4.36	0.48	While accurate prediction can guide optimally efficient resource allocation, accurate prediction does not ensure (or preclude) fair decisions. (P10)
	100	4.7	0.46	
19†	100	4.64	0.48	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)
	100	4.6	0.49	
20†	91	4.45	0.66	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)
	90	4.5	0.67	
21†	64	4.0	1.04	As a guiding principle, algorithms should neither exacerbate nor ignore existing disparities. (P12)
	80	4.4	0.8	
22†	73	3.91	0.67	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)
	75	4.0	0.77	
23†	82	4.18	0.72	Fairness assessment should be done with population samples reflecting the target population, since fairness results may not generalize across different settings. (R14)
	90	4.3	0.64	
24†	91	4.45	0.66	Fairness should be continuously audited, with corrective adjustments made to achieve predetermined (or evolving) fairness goals. (R15)
	75	4.1	0.83	
25	90	4.2	0.60	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†	50	3.8	1.08	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)
	92	4.54	0.63	

Item	Agreement* (%)	Agree/Disagree		Statements
		Mean	SD	
27	80	4.2	0.75	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)
28†	90	4.0	0.90	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)
	100	4.44	0.50	
29†	90	4.2	0.87	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)
	89	4.44	0.68	
30	75	4.0	1.26	In general, models used for resource allocation should employ logic that is open to human scrutiny. (R16)
31	80	4.3	0.78	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

*Voting scores 4 or 5

†Multiple votes are shown for items that required multiple rounds of revisions before consensus was reached. The statement shown in the table is the final consensus version, which is also presented in Tables 1-5 of the GUIDE.