Special Communication | Health Informatics

Guiding Principles to Address the Impact of Algorithm Bias on Racial and Ethnic Disparities in Health and Health Care

Marshall H. Chin, MD, MPH; Nasim Afsar-Manesh, MD, MBA, MHM; Arlene S. Bierman, MD, MS; Christine Chang, MD, MPH; Caleb J. Colón-Rodríguez, DrPH, MHSA; Prashila Dullabh, MD; Deborah Guadalupe Duran, PhD; Malika Fair, MD, MPH; Tina Hernandez-Boussard, PhD, MPH, MS; Maia Hightower, MD, MPH, MBA; Anjali Jain, MD; William B. Jordan, MD, MPH; Stephen Konya; Roslyn Holliday Moore, MS; Tamra Tyree Moore, JD; Richard Rodriguez, MPH; Gauher Shaheen, PhD; Lynne Page Snyder, PhD, MPH; Mithuna Srinivasan, PhD; Craig A. Umscheid, MD, MS; Lucila Ohno-Machado, MD, PhD, MBA

Abstract

IMPORTANCE Health care algorithms are used for diagnosis, treatment, prognosis, risk stratification, and allocation of resources. Bias in the development and use of algorithms can lead to worse outcomes for racial and ethnic minoritized groups and other historically marginalized populations such as individuals with lower income.

OBJECTIVE To provide a conceptual framework and guiding principles for mitigating and preventing bias in health care algorithms to promote health and health care equity.

EVIDENCE REVIEW The Agency for Healthcare Research and Quality and the National Institute for Minority Health and Health Disparities convened a diverse panel of experts to review evidence, hear from stakeholders, and receive community feedback.

FINDINGS The panel developed a conceptual framework to apply guiding principles across an algorithm's life cycle, centering health and health care equity for patients and communities as the goal, within the wider context of structural racism and discrimination. Multiple stakeholders can mitigate and prevent bias at each phase of the algorithm life cycle, including problem formulation (phase 1); data selection, assessment, and management (phase 2); algorithm development, training, and validation (phase 3); deployment and integration of algorithms in intended settings (phase 4); and algorithm monitoring, maintenance, updating, or deimplementation (phase 5). Five principles should guide these efforts: (1) promote health and health care equity during all phases of the health care algorithm life cycle; (2) ensure health care algorithms and their use are transparent and explainable; (3) authentically engage patients and communities during all phases of the health care algorithm life cycle and earn trustworthiness; (4) explicitly identify health care algorithmic fairness issues and trade-offs; and (5) establish accountability for equity and fairness in outcomes from health care algorithms.

CONCLUSIONS AND RELEVANCE Multiple stakeholders must partner to create systems, processes, regulations, incentives, standards, and policies to mitigate and prevent algorithmic bias. Reforms should implement guiding principles that support promotion of health and health care equity in all phases of the algorithm life cycle as well as transparency and explainability, authentic community engagement and ethical partnerships, explicit identification of fairness issues and trade-offs, and accountability for equity and fairness.

JAMA Network Open. 2023;6(12):e2345050. doi:10.1001/jamanetworkopen.2023.45050

➕ **Invited Commentary**

➕ **Multimedia**

Author affiliations and article information are listed at the end of this article.

Introduction

Health care algorithms, defined as mathematical models used to inform decision-making, are ubiquitous and may be used to improve health outcomes. However, algorithmic bias has harmed minoritized communities in housing, banking, and education, and health care is no different.[1] Thus, addressing algorithmic bias is an urgent issue, as exemplified by a Biden Administration Executive Order stating that "agencies shall consider opportunities to prevent and remedy discrimination, including by protecting the public from algorithmic discrimination."[2]

An unbiased algorithm is one that ensures patients who receive the same algorithm score or classification have the same basic needs.[3] Health care algorithms are used for diagnosis, treatment, prognosis, risk stratification, triage, and resource allocation. A biased algorithm that used race to estimate kidney function resulted in higher estimates for Black patients compared with White patients, leading to delays in organ transplant referral for Black patients.[4] A commercial algorithm that risk-stratified patients to determine eligibility for chronic disease management programs effectively required Black individuals to be sicker than White individuals to qualify for such services.[5] Potentially biased algorithms have been developed for heart failure, cardiac surgery, kidney transplantation, vaginal birth after cesarean delivery, rectal cancer, and breast cancer, often affecting access to or eligibility for interventions or services, and resource allocation.[4]

The Agency for Healthcare Research and Quality (AHRQ) and the National Institute on Minority Health and Health Disparities (NIMHD) convened a panel to recommend core guiding principles for the development and use of clinical algorithms in health care, including data-driven, probability-based algorithms such as those using artificial intelligence and machine learning approaches. The panel's core guiding principles also apply to rules-based approaches derived from data (eg, if acute myocardial infarction, give aspirin), since these rules may reflect the specific data sets and patient populations from which they were generated and the potential biases within.

The Council on Artificial Intelligence of the Organization for Economic Cooperation and Development defines an artificial intelligence system as "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. Artificial intelligence systems are designed to operate with varying levels of autonomy."[6] Machine learning is a subset of artificial intelligence that analyzes data using mathematical modeling to learn patterns that can make predictions or guide tasks.[7] Traditional statistical regression techniques, often used in earlier risk prediction models, estimate relationships between predictors and outcomes. In contrast, machine learning models can "learn" by using mathematical techniques that infer relationships within large data sets to inform predictions.[8]

This article describes guiding principles for health care algorithms and key operational considerations. This work is not exhaustive because synergistic efforts, such as those of the Office of the National Coordinator for Health Information Technology (ONC), are ongoing.[9,10] Algorithmic bias is neither inevitable nor merely a mechanical or technical issue. Conscious decisions by algorithm developers, algorithm users, health care industry leaders, and regulators can mitigate and prevent bias and proactively advance health equity.

Methods

The AHRQ received a congressional letter in fall 2020 inquiring about the contribution of clinical algorithms to racial and ethnic bias in health care. In response, the AHRQ published a request for information to elicit perspectives from public stakeholders on this topic and commissioned an evidence review to examine the impact of health care algorithms on health disparities and to identify potential solutions to mitigate biases.[11] The subsequent evidence review underscored the limits of current knowledge and research about health care algorithms in the literature.

The AHRQ, the NIMHD, the US Department of Health and Human Services (HHS) Office of Minority Health, and the ONC collaboratively recruited 9 stakeholders with diverse backgrounds and

expertise to serve on a panel to develop guiding principles to address racial and ethnic bias in health and health care resulting from algorithms. The panel heard from a group of national and international thought leaders involved in algorithm design, development, implementation, and oversight during a 2-day hybrid public meeting and received feedback on draft principles from patient and community representatives and the public during a subsequent virtual meeting.[12] These perspectives were particularly important for the panel's recommendations, given the limitations of the published literature. The panel's work, including this article, was developed iteratively.

Results

Conceptual Framework for Mitigating and Preventing Bias in Health Care Algorithms

The conceptual framework to mitigate and prevent bias in health care algorithms (**Figure**) built on a National Academy of Medicine[13] algorithm life cycle framework adapted by Roski et al.[14] Within the context of structural racism and discrimination,[15] the goal is to promote health and health care equity for patients and communities. An algorithm's life cycle comprises 5 phases that typically occur sequentially.[16] Problem formulation (phase 1) defines the problem that the algorithm is designed to address, relevant actors, and priority outcomes. Problem formulation is followed by selection and management of the data used by the algorithm (phase 2) and subsequent development, training, and validation of the algorithm (phase 3). The algorithm is deployed and integrated in its intended setting (phase 4). Mechanisms should monitor performance and outcomes and maintain, update, or deimplement the algorithm accordingly (phase 5).

Guiding principles apply at each phase to mitigate and prevent bias in an algorithm. Operationalization of principles takes place at 3 levels: individual (developers and users), institutional (organizational policies and procedures), and societal (legislation, regulation, and private policy).

Figure. Conceptual Framework for Applying Guiding Principles to Mitigate and Prevent Bias Across an Algorithm's Life Cycle
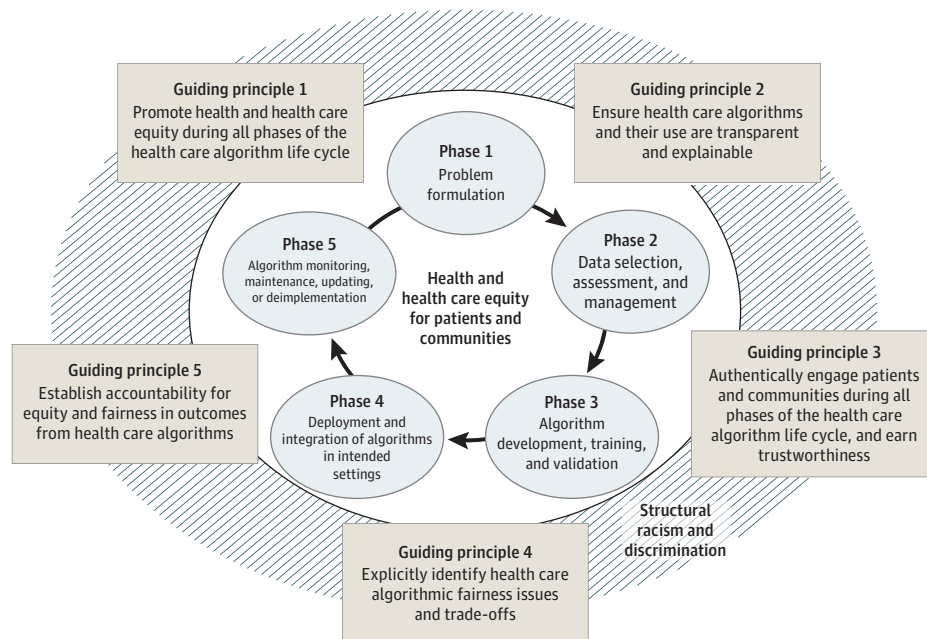
Guiding principles apply at each phase to mitigate and prevent bias in an algorithm.
Operationalization of the principles takes place at 3 levels: individual, institutional, and societal.

This conceptual framework builds on a National Academy of Medicine[13] algorithm life cycle framework adapted by Roski et al.[14]

Guiding Principles for Mitigating and Preventing Racial and Ethnic Bias in Health Care Algorithms

Tables 1 and 2 list the guiding principles and their operational considerations. Each principle is described hereinafter.

Guiding Principle 1: Promote Health and Health Care Equity During All Phases of the Health Care Algorithm Life Cycle

Advancing health equity should be a fundamental objective of any algorithm used in health care.[7] The World Health Organization defines equity as the "absence of unfair, avoidable, or remediable differences among groups of people, whether those groups are defined socially, economically, demographically, or geographically or by other dimensions of inequality (e.g., sex, gender, ethnicity, disability, or sexual orientation)."[17] Algorithms should be designed with goals of advancing health equity, promoting fairness, and reducing health disparities.

Formulating the problem appropriately is critical (phase 1), and improving health and health care equity for patients and communities should be central.[3] During the data selection, assessment, and management phase of the algorithm life cycle (phase 2), data used for algorithm development should be assessed for biases, accuracy, fitness for the intended purpose, and representativeness of the intended population. Engagement of key diverse stakeholders—which includes communities—during problem formulation (phase 1) and data selection (phase 2) is critical to avoid knowledge gaps. Any issues identified should be documented, and corrective actions should be taken before moving to algorithm development, training, and validation (phase 3).

It is critical to use rigorous methods, wise human judgment, and checks and balances in algorithm development to mitigate and prevent bias and ensure that conclusions are accurate, robust, and reproducible.[24] Compared to traditional statistical techniques in which statisticians have more manual control over the analyses, artificial intelligence models can be more opaque and more difficult to interpret. They risk being overfitted to the data at hand, threatening generalizability. Artificial intelligence models sometimes lack common sense and are more difficult to audit. Thus, rigorous methods and processes are essential for algorithm development.[25-29]

Algorithms should be validated across populations to ensure fairness in performance. After an algorithm is deployed, continuous monitoring for performance and data drift is necessary. Monitoring should assess the fairness and equity of the algorithm output as well as the impact of the algorithm on patients, populations, and society, including data privacy and resource allocation. Measurement and comparison of outcomes between advantaged and historically marginalized populations such as racial and ethnic minoritized groups or individuals with lower income should be assessed routinely by health care systems, algorithm vendors, and the research community and supported by research sponsors (eg, funders, scientific journals). Algorithm end users should supplement model outputs with human judgment. Furthermore, access to information technology for all should be ensured.

Guiding Principle 2: Ensure Health Care Algorithms and Their Use Are Transparent and Explainable

Algorithm developers, health care institutions, algorithm users, and regulators are responsible for ensuring that algorithms are transparent, easy to explain, and readily interpretable at all steps in the algorithm life cycle for diverse audiences.[30,31] The HHS states that "all relevant individuals should understand how their data is being used and how AI systems make decisions; algorithms, attributes and correlations should be open to inspection."[20] Development of transparent and explainable algorithms requires algorithm developers and stewards to present evidence for impact on processes and outcomes and to provide understandable and accurate explanations to clinicians and patients to enable informed decision-making.[32] In addition, an algorithm should only operate under the conditions for which it was designed, and outputs should only be used when there is confidence in the results.[31]

Table 1. Guiding Principles and Subprinciples for the Use of Algorithms in Health Care

Guiding principle	Framing for guiding principle	Subprinciples
1. Promote health and health care equity during all phases of the health care algorithm life cycle.	"Equity is the absence of unfair, avoidable or remediable differences among groups of people, whether those groups are defined socially, economically, demographically, or geographically or by other dimensions of inequality (e.g., sex, gender, ethnicity, disability, or sexual orientation)."[17]	Algorithms should be fair. One goal of health care algorithms should be to ensure and promote equitable outcomes for health and health care for all persons and populations. Bias in algorithms should be detected, mitigated, and prevented. Algorithm performance for all populations should be continuously monitored.
	"Health equity means that everyone has a fair and just opportunity to be as healthy as possible."[18]	Health decisions regarding patient care should not be made automatically by an algorithm and instead should involve human input.
	Health equity is defined "as the attainment of the highest level of health for all people. Achieving health equity requires valuing everyone equally with focused and ongoing societal efforts to address avoidable inequalities, historical and contemporary injustices—which includes systemic racism—and the elimination of health and healthcare disparities."[19]	Benefits from algorithms should not be impeded by a "digital divide" (eg, across providers located in areas with differential access to information technology, or across patients with differential access to and knowledge of technology).
	"AI applications should include checks from internal and external stakeholders to help ensure equitable application across all participants."[20] Key considerations include data, problem formulation, stakeholder needs, AI outputs.[20]	
2. Ensure health care algorithms and their use are transparent and explainable.	"All relevant individuals should understand how their data is being used and how AI systems make decisions; algorithms, attributes, and correlations should be open to inspection."[20]	Algorithm developers, health care institutions, and regulators are responsible for ensuring transparency, explainability, and interpretability of algorithms at all phases and to diverse audiences. There should be close collaboration between algorithm developers, health care system decision-makers, and researchers across the health care system to ensure that health care algorithms and their use are transparent and explainable.
		Explainability includes the following: (1) accompanying evidence or reasons for outcomes and approaches or processes[21]; (2) explanations that are understandable to individual users; (3) explanations that correctly reflect the system's process for generating the output; and (4) information that ensures that a system only operates under conditions for which it was designed, and outputs are only used when the system achieves sufficient confidence in its results.
3. Authentically engage patients and communities during all phases of the health care algorithm life cycle, and earn trustworthiness.	"The key to closing equity gaps … is direct participation by impacted communities in the development and implementation of solutions and policy decisions that directly impact them."[22]	Patients and communities, including historically marginalized populations such as racial and ethnic minoritized groups and individuals with lower income, should be engaged authentically and ethically when identifying and assessing a problem requiring the use of an algorithm as part of its solution, and during algorithm data selection, development, deployment, and monitoring.
	"Trustworthy AI refers to the design, development, acquisition, and use of AI in a manner that fosters public trust and confidence while protecting privacy, civil rights, civil liberties, and American values, consistent with applicable laws."[20]	Patients and the public should be aware of algorithm-driven decisions and decision tools, the context in which they are used, and how these tools impact their care and experience.
		Algorithms should be bound by concepts of data sovereignty.
		Trustworthiness should be earned through authenticity, ethical practices, security of data, and timely disclosures of algorithm use.
4. Explicitly identify health care algorithmic fairness issues and trade-offs.	"Ensure that all persons are treated fairly, which includes the requirement to ensure that no person or group is subject to discrimination, neglect, manipulation, domination or abuse (sometimes called 'justice' or 'fairness')."[7]	Algorithmic fairness and bias issues arise from both ethical choices and technical decisions and should be considered at each phase of the algorithm's life cycle.
	"There can be tradeoffs between model performance and both fairness and explainability. A model may have a high percentage of accurate predictions, but the model may be replicating historical biases present in the data. Similarly, a deep learning or other similarly complex model may have strong performance metrics, but it may be more difficult to understand and explain the model's outputs."[20]	Distributive justice metrics should be considered, including clinical outcomes, resource allocation, and algorithms' performance metrics.
		When unfairness is identified, bias should be mitigated using social (eg, diverse teams and stakeholder co-development, policies) and technical (eg, algorithmic fairness toolkits, fairness metrics, data set collection, deimplementation) mitigation methods.
		Algorithms and accompanying policies and regulations should also be viewed through the frames of equity of harms and risks and through explicit identification of trade-offs among different competing values and options.
5. Establish accountability for equity and fairness in outcomes from health care algorithms.	"Accountability: assign moral, legal, and organizational responsibilities to the individuals who control entities, goals, process, and outcomes affecting people."[23]	Individuals and organizations must accept responsibility to achieve equity and fairness in outcomes from health care algorithms and be accountable.
	"Appropriate mechanisms should be adopted to ensure questioning by and redress for individuals and groups adversely affected by algorithmically informed decisions. This should include access to prompt, effective remedies and redress from governments and companies that deploy AI technologies for healthcare."[7]	Organizations should establish processes at each phase of the algorithm life cycle to facilitate equity and fairness in outcomes. The processes should include model developers, end users, clinicians, health care administrators, patient advocates, and community representatives.
		Governmental, institutional, and business regulations and incentives should support equity and fairness.
		Persons and communities who have been harmed by unfair algorithms should be redressed.
		Transparency is critical throughout the life cycle of the algorithm to identify bias and inequity.

Abbreviation: AI, artificial intelligence.

Transparency includes multiple domains, such as availability of technical information, algorithm oversight, and communication of impact to stakeholders.[20,31,32] Algorithm developers should create profiles of the data used to train the algorithm, describing distribution of key aspects of the population in the data set (eg, race and ethnicity, gender, socioeconomic status, and age); they should also make data exploration analysis readily available for independent review. Algorithm developers should disclose types, sizes, and overall distributions in data sets used in their formulation, testing, and validation. Regulation should require algorithm information labels or model cards sufficient to assess design, validity, and the presence of bias.[10,21] Implementers should disclose the purpose of algorithms and their impact. If biases have been identified in an algorithm, the developers, implementers, and users should disclose such biases. Any bias mitigation attempts should also be disclosed to all with a stake in the algorithm, including patients, caregivers, and

Table 2. Considerations for Operationalizing Guiding Principles for Algorithm Use in Health Care

Guiding principle	Considerations for operationalizing guiding principle
1. Promote health and health care equity during all phases of the health care algorithm life cycle.	Researchers and research sponsors (eg, funders, scientific journals) should routinely assess the impact of health care algorithms on health equity.
	Validate algorithms for the specific purpose for which they are being deployed and across populations. Evaluate training data sets for representativeness of specific populations.
	Document any lack of representativeness. If appropriate, take mitigation measures before training the algorithm.
	Continually monitor algorithm performance for equitable impact across populations.
2. Ensure health care algorithms and their use are transparent and explainable.	Develop profiles of algorithm training data with the distributions of key features of the population (eg, race and ethnicity, gender, socioeconomic status, and age), and make the distributions available for independent review.
	Design regulations to ensure transparency, explainability, and interpretability. For example, require algorithm information labels to clearly communicate design features and the intent of the algorithm to stakeholders. Enough information should be provided to assess validity and bias.
	Develop reporting guidelines for publications examining algorithms that are explicit about bias, similar to proposals to address equity in observational studies and randomized clinical trials. Explain algorithm biases and mitigation measures to the stakeholder community.
	Make information available for patients and communities when an algorithm is used in their care, including what aspects of their personal data were used in the algorithm, what impact the algorithm had on their care (eg, diagnosis, prognosis, or treatment), and how the algorithm performs for their sociodemographic group.
3. Authentically engage patients and communities during all phases of the health care algorithm life cycle, and earn trustworthiness.	Engage patients and communities in decisions about those problems best addressed by algorithm solutions.
	Algorithm development teams should include a diverse group of people who are involved in decision-making.
	Put safeguards in place to protect patient autonomy and privacy in health care algorithm development, deployment, and monitoring.
	Speak to those most impacted by algorithmic bias to acknowledge potential or demonstrated harms and agree on methods of redress and closure.
4. Explicitly identify health care algorithmic fairness issues and trade-offs.	Model performance across patient cohorts should be measured using multiple objective measures (eg, accuracy measures such as sensitivity, specificity, area under the receiver operating curve, predictive values, calibration, and residuals) that are appropriate for the intended use of the algorithm.
	Fairness of the model output across patient cohorts should be measured using metrics such as demographic parity (same proportion of groups assigned to positive or negative class) and equalized odds (groups have same false positive rate and same false negative rate).
	Model fairness should be optimized for equity in clinical outcomes or resource allocation using bias mitigation methods (eg, disparate impact remover, label choice experiments, reweighing) and human judgment.
5. Establish accountability for equity and fairness in outcomes from health care algorithms.	Governmental agencies, accreditation organizations, and professional associations should implement regulatory processes, policies, and standards to mandate transparency and regular monitoring and validation of health care algorithms for equity and fairness. Incentives for fairness in health care algorithms should be created. Equity and fairness checks should be built into each phase of the algorithm life cycle for both technical bias and human bias that lead to inequities in model performance, clinical outcomes, and resource allocation. Unfair algorithms should be deactivated, removed, or discontinued. A structured reporting process could identify signals of emerging problems both locally and nationally and facilitate addressing the problems systematically.
	A legal and administrative framework and culture should be created to redress harm caused by algorithms. The framework should encourage quality improvement, collaboration, and transparency, as is recommended in the patient safety field.
	Algorithm developers, implementers, and users (including but not limited to health care delivery organizations) should adopt policies, procedures, and processes to monitor for equity and fairness at each phase of the algorithm life cycle. They should implement effective and transparent data collection mechanisms to support monitoring. They should identify clear algorithm stewardship and bias mitigation roles for each involved stakeholder group.
	Health care delivery organizations and algorithm vendors should invest in infrastructure, governance, and teams with diverse skills and experiences to support equity and fairness in algorithm development and use.
	Algorithms should not be deployed before validation on the impacted population. Underresourced institutions with limited technical capability should be supported in validation.
	Journals, funders, and research professional associations should identify standards for ensuring accountability for equity and fairness in outcomes from health care algorithms, for the algorithms to be published, funded, and rated as high quality.

communities. A structured reporting process could identify signals of emerging problems both locally and nationally and facilitate addressing such problems systematically.

Several reporting guidelines promote transparency of research examining algorithms.[33] However, these guidelines do not include concrete ways to report on fairness, and they rarely make explicit mention of equity.[34] Reporting guidelines for algorithms should therefore be updated with specific equity approaches as has been done for observational studies and randomized clinical trials.[35,36]

Guiding Principle 3: Authentically Engage Patients and Communities During All Phases of the Health Care Algorithm Life Cycle, and Earn Trustworthiness

Authentically engaging and partnering with patients and communities is essential to understand both a problem affecting them and its solutions.[22] Moreover, it is an ethical imperative to engage with patients and communities around health care algorithms and earn their trust, as these tools can provide great benefit or harm. Patients and communities, including populations who have been historically marginalized, should be engaged authentically and ethically when identifying and assessing a problem that requires use of an algorithm as part of its solution and during algorithm data selection, development, deployment, and monitoring.

Early and intentional engagement can help identify priorities of patients and communities and any concerns they have regarding algorithm use.[37] All patients and communities should be informed when an algorithm is used in their care, should be advised about impact of the algorithm on their treatment, and should be provided alternatives if appropriate.[38] They should know how the algorithm performs for their demographic group compared with other groups and be made aware of any opportunities to opt out of algorithms or to pursue alternatives to algorithm-driven decisions.

Algorithms should be bound by concepts of data sovereignty, the idea that data are subject to legal regulations of countries, nations, or states. Sovereignty is of particular importance to Indigenous nations.[39] Health care organizations, vendors, and other model developers earn trustworthiness through authenticity, ethical and transparent practices, security and privacy of data, and timely disclosures of algorithm use.

Guiding Principle 4: Explicitly Identify Health Care Algorithmic Fairness Issues and Trade-offs

The panel recommends that advancing health and health care equity for patients and communities should be the goal of health care algorithms. Advancing health equity requires expertise in algorithmic fairness—the field of identifying, understanding, and mitigating bias.[40-42] Health care algorithmic fairness issues arise from both ethical choices and technical decisions at different phases of the algorithm life cycle.[16,43] For example, fundamental ethical choices can arise during problem formulation (phase 1; eg, Is the goal of the algorithm to improve and advance equitable outcomes or is the primary goal to maximize profit?). Additionally, if a particular algorithm use involves choosing a cutoff point for action during model development and implementation, should that cutoff be chosen to maximize sensitivity of the tool to identify someone who might benefit from an intervention, or should it be chosen to maximize specificity of the tool so inappropriate patients are not exposed to unnecessary risk from the intervention? Trade-offs among competing fairness metrics and values are common. Different technical definitions of algorithmic fairness, such as sufficiency, separation, and independence, are mathematically mutually incompatible, trading off maximizing accuracy of an algorithm and minimizing differences among groups across definitions.[44] It is critical to make health care algorithm fairness issues and trade-offs explicit, transparent, and explainable. Thus, solutions to advance health equity with health care algorithms require ethical, technical, and social approaches—there is no simple cookie-cutter technical solution.[43,45]

Technical methods for improving fairness in algorithms can be divided into stages of modeling: preprocessing (eg, repair biased data set), in-processing (eg, use fairness metrics in the model optimization process to maximize accuracy and fairness), and postprocessing (eg, transform model output to improve prediction fairness).[46,47] Key issues for fairness metrics include prioritization of

fairness for group or individual, binary classification (eg, qualifies for service or not) vs continuous classification (eg, regression output), and use of regularization methods (fairness metrics to balance accuracy and fairness), reweighting methods (weight samples from underrepresented groups more highly), or both.[48] Of note, technical definitions and metrics of fairness often do not translate clearly or intuitively to ethical, legal, social, and economic conceptions of fairness.[46,47] Thus, close collaboration and discussion are essential among stakeholders, including algorithm developers, algorithm users, and the communities to whom the algorithm will be applied.

We recommend considering fairness of algorithms through the lens of distributive justice, the socially just distribution of outcomes and allocation of resources across different populations.[49] Distributive justice metrics include clinical outcomes, resource allocation, and performance measures of algorithms (eg, sensitivity, specificity, and positive predictive value).[8,43,50] When unfairness is identified, bias should be mitigated using both social (eg, diverse teams and stakeholder co-development) and technical (eg, algorithmic fairness toolkits, fairness metrics, data set collection, and deimplementation) mitigation methods.[51] Algorithms and accompanying policies and regulations should also be viewed through frames of equity of harms and risks and explicit identification of trade-offs among different competing values and options.[41-43] Algorithms with a higher risk of substantial harm and injustice should have stricter internal oversight by organizations and more stringent external regulation.[20]

Guiding Principle 5: Establish Accountability for Equity and Fairness in Outcomes From Health Care Algorithms

Model developers and users, including vendors, health care organizations, researchers, and professional societies, should accept responsibility to achieve equity and fairness in outcomes from health care algorithms and be accountable for the performance of algorithms in different populations. Institutions such as vendors and health care provider organizations should establish processes at each phase of the algorithm life cycle to promote equity and fairness. Transparency in the types of training data, processes, and evaluations used is paramount. For example, an academic medical center recently published its framework for oversight and deployment of prediction models, which includes checkpoint gates and an oversight governance structure.[52] Current evidence suggests that such governance infrastructure is rare.[53]

Organizations should have an inventory of their algorithms and have local, periodic evaluations and processes that screen for and mitigate bias. It is crucial for organizations to engage stakeholders throughout the entire algorithm life cycle to ensure fairness and promote trust. This means incorporating model developers, end users, health care administrators, clinicians, patient advocates, and community representatives. Different organizations and experts have recommended various accountability metrics and oversight structures.[3]

Regulations and incentives should support equity and fairness while also promoting innovation.[54] There should be redress for persons and communities who have been harmed by biased algorithms. An ethical, legal, social, and administrative framework and culture should be created that redresses harm while encouraging quality improvement, collaboration, and transparency, similar to what is recommended for patient safety.[55]

Conclusions

ChatGPT and other artificial intelligence language models have spurred widespread public interest in the potential value and dangers of algorithms. Multiple stakeholders must partner to create systems, processes, regulations, incentives, standards, and policies to mitigate and prevent algorithm bias in health care.[47] Dedicated resources and the support of leaders and the public are critical for successful reform. It is our obligation to avoid repeating errors that tainted use of algorithms in other fields.

ARTICLE INFORMATION

Corresponding Authors: Marshall H. Chin, MD, MPH, University of Chicago, 5841 S. Maryland Ave, MC2007, Chicago, IL 60637 (mchin@bsd.uchicago.edu); Lucila Ohno-Machado, MD, PhD, MBA, Yale School of Medicine, 333 Cedar St, New Haven, CT 06510 (lucila.ohno-machado@yale.edu).

Author Affiliations: University of Chicago, Chicago, Illinois (Chin); Oracle Health, Austin, Texas (Afsar-Manesh); Agency for Healthcare Research and Quality, Rockville, Maryland (Bierman, Chang, Jain, Umscheid); US Department of Health and Human Services Office of Minority Health, Rockville, Maryland (Colón-Rodríguez, R. H. Moore); NORC at the University of Chicago, Bethesda, Maryland (Dullabh, Snyder, Srinivasan); National Institute on Minority Health and Health Disparities, Bethesda, Maryland (Duran); Association of American Medical Colleges, Washington, DC (Fair); Stanford University, Stanford, California (Hernandez-Boussard); Equality AI, Park City, Utah (Hightower); American Medical Association, Chicago, Illinois (Jordan); Office of the National Coordinator for Health Information Technology, Washington, DC (Konya); Prudential Financial, Arlington, Virginia (T. T. Moore); NORC at the University of Chicago, Chicago, Illinois (Rodriguez); Elevance Health, Indianapolis, Indiana (Shaheen); Yale School of Medicine, New Haven, Connecticut (Ohno-Machado).

REFERENCES

1. O'Neil C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown; 2016.

2. The White House. Executive Order on further advancing racial equity and support for underserved communities through the federal government. Updated February 16, 2023. Accessed August 31, 2023. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/02/16/executive-order-on-further-advancing-racial-equity-and-support-for-underserved-communities-through-the-federal-government/

3. Obermeyer Z, Nissan R, Stern M, Eaneff S, Bembeneck EJ, Mullainathan S. Algorithmic Bias Playbook. Chicago Booth Center for Applied Artificial Intelligence; 2021. Accessed November 10, 2023. https://www.chicagobooth.edu/research/center-for-applied-artificial-intelligence/research/algorithmic-bias/playbook

4. 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

5. 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

6. Organization for Economic Cooperation and Development. Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449. 2019. Accessed November 10, 2023. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449#mainText

7. World Health Organization. Ethics and Governance of Artificial Intelligence for Health. World Health Organization; 2021. Accessed November 10, 2023. https://www.who.int/publications/i/item/9789240029200

8. Makhni S, Chin MH, Fahrenbach J, Rojas JC. Equity challenges for artificial intelligence algorithms in health care. *Chest*. 2022;161(5):1343-1346. doi:10.1016/j.chest.2022.01.009

9. Marchesini K, Smith J, Everson J. Increasing the transparency and trustworthiness of AI in health care. *HealthITbuzz* blog. April 13, 2023. Accessed August 31, 2023. https://www.healthit.gov/buzz-blog/health-innovation/transparent-and-trustworthy-ai-in-health-care

10. Office of the National Coordinator for Health Information Technology. Health data, technology, and interoperability: certification program updates, algorithm transparency, and information sharing (HTI-1) proposed rule. HealthIT.gov. June 22, 2023. Accessed August 31, 2023. https://www.healthit.gov/topic/laws-regulation-and-policy/health-data-technology-and-interoperability-certification-program

11.   Jain A, Brooks JR, Alford CC, et al. Awareness of racial and ethnic bias and potential solutions to address bias with use of health care algorithms. *JAMA Health Forum*. 2023;4(6):e231197. doi:10.1001/jamahealthforum.2023.1197

12.   Agency for Healthcare Research and Quality. Meetings examine impact of healthcare algorithms on racial and ethnic disparities in health and healthcare. Accessed August 31, 2023. https://effectivehealthcare.ahrq.gov/news/meetings

13.   Matheny M, Thadaney IS, Ahmed M, Whicher D, eds. Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. National Academy of Medicine; 2022. Accessed November 10, 2023. https://nam.edu/artificial-intelligence-special-publication/

14.   Roski J, Maier EJ, Vigilante K, Kane EA, Matheny ME. Enhancing trust in AI through industry self-governance. *J Am Med Inform Assoc*. 2021;28(7):1582-1590. doi:10.1093/jamia/ocab065

15.   Bailey ZD, Krieger N, Agénor M, Graves J, Linos N, Bassett MT. Structural racism and health inequities in the USA: evidence and interventions. *Lancet*. 2017;389(10077):1453-1463. doi:10.1016/S0140-6736(17)30569-X

16.   Ng MY, Kapur S, Blizinsky KD, Hernandez-Boussard T. The AI life cycle: a holistic approach to creating ethical AI for health decisions. *Nat Med*. 2022;28(11):2247-2249. doi:10.1038/s41591-022-01993-y

17.   World Health Organization. Health equity. Accessed August 31, 2023. https://www.who.int/health-topics/health-equity#tab=tab_1

18.   Braveman P, Arkin E, Orleans T, Proctor D, Plough A. What is health equity? Robert Wood Johnson Foundation. May 1, 2017. Accessed November 10, 2023. https://www.rwjf.org/en/insights/our-research/2017/05/what-is-health-equity-.html

19.   Health Care Payment Learning & Action Network. Advancing health equity through APMs: guidance for equity-centered design and implementation. 2021. Accessed November 10, 2023. http://hcp-lan.org/workproducts/APM-Guidance/Advancing-Health-Equity-Through-APMs.pdf

20.   US Department of Health and Human Services. Trustworthy AI (TAI) Playbook. US Department of Health and Human Services; 2021. Accessed November 10, 2023. https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook.pdf

21.   Zuckerman BL, Karabin JM, Parker RA, Doane WEJ, Williams SR. Options and Opportunities to Address and Mitigate the Existing and Potential Risks, as well as Promote Benefits, Associated With AI and Other Advanced Analytic Methods. OPRE Report 2022-253. US Department of Health and Human Services Office of Planning, Research, and Evaluation, Administration for Children and Families; 2022. Accessed November 10, 2023. https://www.acf.hhs.gov/opre/report/options-opportunities-address-mitigate-existing-potential-risks-promote-benefits

22.   Gonzalez R. The spectrum of community engagement to ownership. Movement Strategy Center. 2019. Accessed November 10, 2023. https://movementstrategy.org/resources/the-spectrum-of-community-engagement-to-ownership/

23.   Loi M, Heitz C, Christen M. A comparative assessment and synthesis of twenty ethics codes on AI and big data. In: *Proceedings of the 2020 7th Swiss Conference on Data Science (SDS)*. IEEE; 2020:41-46. doi:10.1109/SDS49233.2020.00015

24.   Hunter DJ, Holmes C. Where medical statistics meets artificial intelligence. *N Engl J Med*. 2023;389(13):1211-1219. doi:10.1056/NEJMra2212850

25.   Parikh RB, Obermeyer Z, Navathe AS. Regulation of predictive analytics in medicine. *Science*. 2019;363(6429):810-812. doi:10.1126/science.aaw0029

26.   Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ*. 2020;368:l6927. doi:10.1136/bmj.l6927

27.   Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020;26(9):1320-1324. doi:10.1038/s41591-020-1041-y

28.   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

29.   Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. *Annu Rev Biomed Data Sci*. 2021;4:123-144. doi:10.1146/annurev-biodatasci-092820-114757

30.   Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. *Entropy (Basel)*. 2020;23(1):18. doi:10.3390/e23010018

31.   Phillips PJ, Hahn CA, Fontana PC, et al. *Four Principles of Explainable Artificial Intelligence. NIST Interagency/Internal Report (NISTIR)*. National Institute of Standards and Technology; 2021.

32. Vasse'i RM, McCrosky J. AI Transparency in Practice. Mozilla; 2023. Accessed August 31, 2023. https://foundation.mozilla.org/en/research/library/ai-transparency-in-practice/ai-transparency-in-practice/

33. CONSORT-AI and SPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat Med*. 2019;25(10):1467-1468. doi:10.1038/s41591-019-0603-3

34. Wawira Gichoya J, McCoy LG, Celi LA, Ghassemi M. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ Health Care Inform*. 2021;28(1):e100289. doi:10.1136/bmjhci-2020-100289

35. Antequera A, Lawson DO, Noorduyn SG, et al. Improving social justice in COVID-19 health research: interim guidelines for reporting health equity in observational studies. *Int J Environ Res Public Health*. 2021;18(17):9357. doi:10.3390/ijerph18179357

36. Welch VA, Norheim OF, Jull J, Cookson R, Sommerfelt H, Tugwell P; CONSORT-Equity and Boston Equity Symposium. CONSORT-Equity 2017 extension and elaboration for better reporting of health equity in randomised trials. *BMJ*. 2017;359:j5085. doi:10.1136/bmj.j5085

37. Joosten YA, Israel TL, Williams NA, et al. Community engagement studios: a structured approach to obtaining meaningful input from stakeholders to inform research. *Acad Med*. 2015;90(12):1646-1650. doi:10.1097/ACM.0000000000000794

38. The White House. Fact sheet: Biden-Harris administration announces key actions to advance tech accountability and protect the rights of the American public. October 4, 2022. Accessed November 10, 2023. https://www.whitehouse.gov/ostp/news-updates/2022/10/04/fact-sheet-biden-harris-administration-announces-key-actions-to-advance-tech-accountability-and-protect-the-rights-of-the-american-public/

39. National Congress of American Indians. Tribal nations & the United States: an introduction. Accessed November 27, 2023. https://archive.ncai.org/about-tribes

40. Drukker K, Chen W, Gichoya J, et al. Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. *J Med Imaging (Bellingham)*. 2023;10(6):061104. doi:10.1117/1.JMI.10.6.061104

41. Xu J, Xiao Y, Wang WH, et al. Algorithmic fairness in computational medicine. *EBioMedicine*. 2022;84:104250. doi:10.1016/j.ebiom.2022.104250

42. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput Surv*. 2021;54(6):1-35. doi:10.1145/3457607

43. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med*. 2018;169(12):866-872. doi:10.7326/M18-1990

44. Weinkauf D. When worlds collide—the possibilities and limits of algorithmic fairness (part 1). *Privacy Tech-Know* blog. April 5, 2023. Accessed August 31, 2023. https://www.priv.gc.ca/en/blog/20230405_01/

45. Pfeiffer J, Gutschow J, Haas C, et al. Algorithmic fairness in AI. *Bus Inf Syst Eng*. 2023;65:209-222. doi:10.1007/s12599-023-00787-x

46. Caton S, Haas C. Fairness in machine learning: a survey. *arXiv*. Preprint posted online October 4, 2020. doi:10.48550/arXiv.2010.04053

47. Cary MP Jr, Zink A, Wei S, et al. Mitigating racial and ethnic bias and advancing health equity in clinical algorithms: a scoping review. *Health Aff (Millwood)*. 2023;42(10):1359-1368. doi:10.1377/hlthaff.2023.00553

48. Jung S, Park T, Chun S, Moon T. Re-weighting based group fairness regularization via classwise robust optimization. *arXiv*. Preprint posted online March 1, 2023. doi:10.48550/arXiv.2303.00442

49. Daniels N. Justice, health, and healthcare. *Am J Bioeth*. 2001;1(2):2-16. doi:10.1162/152651601300168834

50. Rojas JC, Fahrenbach J, Makhni S, et al. Framework for integrating equity into machine learning models: a case study. *Chest*. 2022;161(6):1621-1627. doi:10.1016/j.chest.2022.02.001

51. Weinkauf D. When worlds collide—the possibilities and limits of algorithmic fairness (part 2). *Privacy Tech-Know* blog. Office of the Privacy Commissioner of Canada. April 5, 2023. Accessed August 31, 2023. https://www.priv.gc.ca/en/blog/20230405_02/

52. Bedoya AD, Economou-Zavlanos NJ, Goldstein BA, et al. A framework for the oversight and local deployment of safe and high-quality prediction models. *J Am Med Inform Assoc*. 2022;29(9):1631-1636. doi:10.1093/jamia/ocac078

53. Rojas JC, Rohweder G, Guptill J, Arora VM, Umscheid CA. Predictive analytics programs at large healthcare systems in the USA: a national survey. *J Gen Intern Med*. 2022;37(15):4015-4017. doi:10.1007/s11606-022-07517-1

54. Eggers W, Walsh S, Joergensen C, Kishnani P. Regulation that enables innovation. *Deloitte Insights*. March 23, 2023. Accessed November 10, 2023. https://www2.deloitte.com/us/en/insights/industry/public-sector/government-trends/2023/regulatory-agencies-and-innovation.html

55. McCradden MD, Joshi S, Anderson JA, Mazwi M, Goldenberg A, Zlotnik Shaul R. Patient safety and quality improvement: ethical principles for a regulatory approach to bias in healthcare machine learning. *J Am Med Inform Assoc*. 2020;27(12):2024-2027. doi:10.1093/jamia/ocaa085