Preventing Bias and Inequities in AI-Enabled Health Tools

Duke | MARGOLIS CENTER for Health Policy

July 6, 2022

Authors

Trevan Locke, Assistant Research Director, Duke-Margolis Center for Health Policy

Valerie J. Parker, Senior Policy Analyst, Duke-Margolis Center for Health Policy

Andrea Thoumi, Health Equity Policy Fellow, Duke-Margolis Center for Health Policy

Benjamin A. Goldstein, Associate Professor of Biostatistics and Bioinformatics, Duke University

Christina Silcox, Research Director, Digital Health, Duke-Margolis Center for Health Policy

Acknowledgments

We would like to thank the wide range of stakeholders who spoke with us in individual interviews and over the course of this project as well as the public meeting hosted in December 2021 for providing their expertise, experiences, and insights on the topics discussed herein. We would also like to acknowledge Pat Baird, Buz Waitzkin, and Kadija Ferryman for their thoughtful comments and edits while reviewing this paper..

Additionally, we would like to thank our colleagues at Duke-Margolis for their time, insights, and support including Luke Durocher, Patricia Green, Marianne Hamilton Lopez, Gillian Sanders Schmidler, Mark McClellan, Katie Huber, Brian Canter, and Rebecca Ray.

Support for this project was provided by The Pew Charitable Trusts.

About Duke-Margolis

The Robert J. Margolis, MD, Center for Health Policy at Duke University is directed by Mark McClellan, MD, PhD, and brings together expertise from the Washington, DC, policy community, Duke University, and Duke Health to address the most pressing issues in health policy. The mission of Duke-Margolis is to improve health and the value of health care through practical, innovative, and evidence-based policy solutions. Duke-Margolis catalyzes Duke University's leading capabilities, including interdisciplinary academic research and capacity for education and engagement, to inform policy making and implementation for better health and health care. For more information, visit healthpolicy.duke.edu.

Recommended Citation Format

Locke T, Parker V, Thoumi A, Goldstein B, Silcox C (2022). Preventing Bias and Inequities in AI-enabled Health Tools. Washington, DC: Duke-Margolis Center for Health Policy.

Executive Summary

Artificial Intelligence (AI) has shown great potential across a variety of areas in our society, including within the health care system. AI is used to analyze images in radiology, assess patients to provide decision support to providers, and flag patients at a high risk of deterioration. While AI can be a useful tool, it is built by humans and with data collected by humans. As such, it is susceptible to reproducing and potentially scaling the effects of the biases and inequities which pervade our society. As such, it is unsurprising that there are many documented cases in which health AI has shown disparate performance amongst different patient subgroups and has otherwise been shown to worsen health inequities. But if carefully built and tested, AI has the potential to reduce biased care and improve health equity, for example through increased access, nudging health care professionals past subconscious bias, and more personalized care. This paper explores how bias enters into an AI-enabled health tool throughout various stages of the development and implementation process, identifies mitigation and testing practices that can reduce the likelihood of building a tool that is biased or inequitable, and describes gaps where more research is needed.

Building from insights from numerous stakeholder interviews conducted in Fall 2021 through Spring 2022, a public convening held during December 2021, and a literature review, we identified four areas in which bias can be introduced:

- Inequitable framing of the health care challenge or the user's next steps
- The use of unrepresentative training data
- The use of biased training data
- Insufficient care with choices in data selection, curation, preparation, and model development

From there, the paper provides recommendations for identifying and mitigating biased AI in health care focused on different stakeholder groups: developers, purchasers of the AI tools such as providers or payers, data originators such as health systems, and regulators, with a particular focus on the U.S. Food and Drug Administration (FDA).

Developers should be aware of how bias or inequitable outcomes can be caused by the development process. They need to follow consensus standards (where they exist) and help develop good machine learning practices (GMLP) to prevent tools from biased performance or contributing to inequitable outcomes. They need to work with teams with diverse expertise, including a deep understanding of the problem being solved, the data being used, the differences that can occur across subgroups within the population of interest, and how the AI tool output is likely to be used.

Purchasers and users need to test tools within their own subpopulations, both during implementation, but also over time to monitor any drift towards bias or inequity. This includes not just the accuracy of the tool itself, but also patient outcomes resulting from use of the tool.

Data originators have a responsibility to ensure that their data is recorded in just and equitable ways. Data originators include multiple groups: real-world data generators like health systems, payers, and tech companies making wearables and remote monitoring home devices, as well as private and public consortiums building health databases for health research and AI development purposes. All data originators should prioritize standardization, reductions of bias in subjective descriptions, and annotation of where their data may differ across populations. These differences can be due to access challenges, differing performance of data collection tools such as sensors, or other reasons. This is a responsibility that is not simply about building better AI but also ensuring the highest quality care decisions and improving the overall learning health care system.

The FDA, working in tandem with other federal agencies, should ensure that AI-enabled medical devices perform well across subgroups. They should also require clear and accessible labeling of the products regarding subgroup testing and populations intended for use and work to build systems to monitor for biased performance of AI-enabled devices.

As a developing field, some of the best practices, data, and tests needed to facilitate implementation of these recommendations are not yet built or are still in development. The health care ecosystem as a whole needs to work together to ensure that AI tools are purposefully built to create a more just and equitable health care system, rather than replicating or worsening the current state.

Terminology

Throughout this paper we utilize the terms justice, equity, and equality to refer to aspects of health care systems and delivery. While these terms are similar, it is important to note their differences:

Equality involves giving everyone the same resources and opportunities. Equality, however, does not ensure that everyone reaches the same outcomes because it does not account for systemic barriers that disproportionately impact some groups over others.[1]

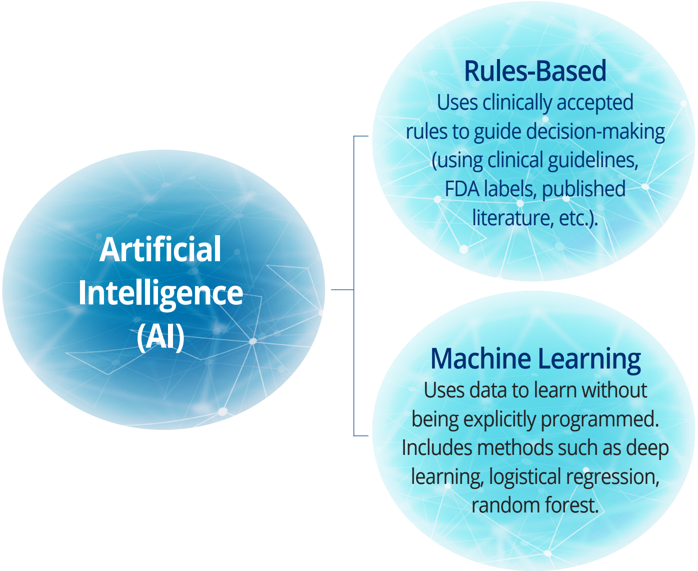
Equity addresses unjust social systems themselves, recognizing that individuals have different circumstances and need different resources and opportunities to reach equal outcomes.[1] Health equity has been defined by the Robert Wood Johnson Foundation as "the state in which everyone has a fair and just opportunity to be as healthy as possible. This requires removing obstacles to health such as poverty, discrimination, and their consequences, including powerlessness and lack of access to good jobs with fair pay, quality, education and housing, safe environments, and health care."[2,3]

(Health) justice means giving every community an equitable chance at being healthy by addressing persistent political, economic, and social inequities and injustices that affect all sectors that serve our communities and have a disproportionate impact on the health of marginalized communities.[4]

Introduction

Artificial Intelligence (AI) has shown considerable promise for improving health and medical treatments, but it can reflect and scale biases and inequities already prevalent throughout our society. This is particularly a concern for machine-learning (**Figure 1**) trained on real-world health data that reflects the biased and inequitable care given to certain subgroups within patient populations due to factors attributable to structural racism and institutionalized inequities such as lack of access to affordable health insurance and care, as well as prejudice and implicit biases.[5–9] But rules-based algorithms have also been shown to have biased performance and outcomes.[10] As the health care field works to leverage innovative technology into new approaches to delivering care, all stakeholders must commit to ensuring that current and past inequities and biases do not become more ingrained. These stakeholders include developers of AI health tools, but also purchasers, regulators, and other contributors to the development of AI health tools including health systems, payers, companies that collect health and wellness data such as wearables and health monitors, and regulators (**Table 1**). These stakeholders often have multiple roles regarding data origination, development, evaluation, and use of these AI tools. Patients that are impacted by these tools are also critical stakeholders that the rest of the community must work with to help ensure equitable and just health care.

FIGURE 1



Uses clinically accepted rules to guide decision-making (using clinical guidelines, FDA labels, published literature, etc.).

Uses data to learn without being explicitly programmed. Includes methods such as deep learning, logistical regression, random forest.

AI is a broad category that includes both rules-based AI and machine learning-based tools. While rules-based tools are built using clinically accepted guidelines and well-defined clinical expertise, machine learning (ML) tools are built by computer-derived relationships in data used to train the underlying algorithms.

TABLE 1 Health AI Stakeholders and Their Touchpoints in the Development, Evaluation, and Implementation of AI Health Tools

	Health Systems	Payers	Wearables/ Monitoring Tech Companies	AI Companies	Regulators	Patients
Develop AI Health Tools	✓ (homegrown)	✓ (homegrown)	✓ (within products)	✓		
Use AI Health Tools	✓	✓	✓			✓
Purchase AI Health Tools	✓	✓				✓
Evaluate AI Health Tools	✓	✓		✓	✓	
Originate Data Used to Build AI Health Tools	✓	✓	✓	✓		✓
Regulate AI Health Tools					✓	

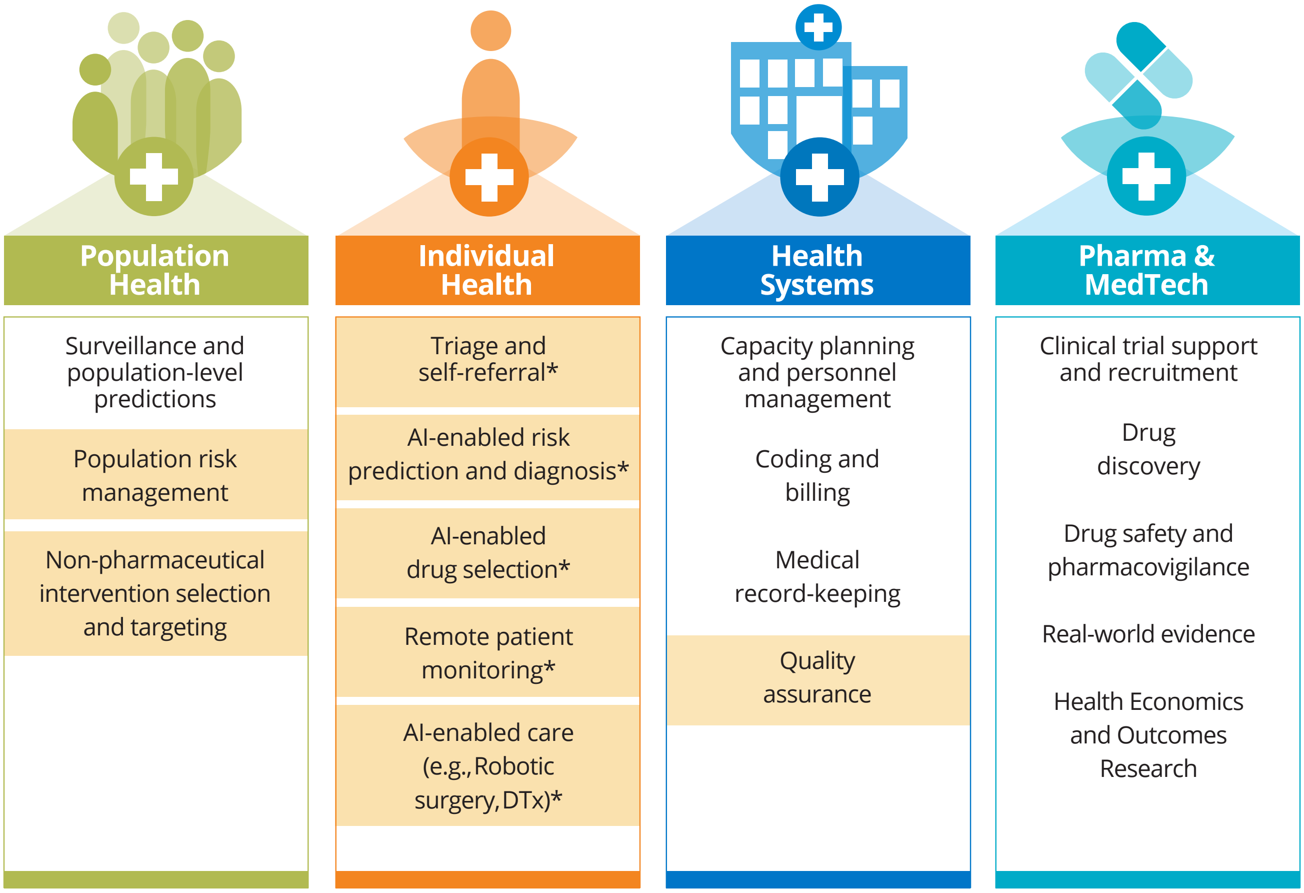
This responsibility of ensuring that AI tools lead to more equitable health care is critical as more health tools are being built and implemented. According to the consulting firm Advisory Board, use of AI by provider organizations has increased in recent years with 18 percent for precision medicine, 16 percent of organizations using AI for protocol compliance, and 14 percent for risk and care gap identification.[11] Provider systems, especially larger or academic-based health systems, may be using a combination of commercial products and "homegrown" tools developed by their own staff and researchers. AI is also being used in public health, in payer systems, and in the development and surveillance of other medical products; for example, to identify drug targets and for clinical trial data collection and management. **Figure 2**, adapted from a United States Agency for International Development report, shows the diversity of AI uses within the health space.[12] AI applications in the health care setting include clinical decision support (CDS), wearable remote monitoring and analyses, digital therapeutics, administrative software (such as scheduling software that can predict "no-shows", software used to determine home nursing aid hours, and supply chain management), and population health management. Some of these tools and products are classified as medical devices, under U.S. Food and Drug Administration (FDA) authority, while others are used for administrative, wellness, or business purposes and are outside of FDA review. These AI-enabled tools have the potential to significantly improve outcomes as well as increase access and efficiency in health care by improving decision-making, increasing access, reducing costs, or hastening diagnosis and treatment. But, if developed without care to the prevention of bias, inequity, and injustice, they also can scale inequities and further entrench health care disparities.

In a recent blog post, the Office of the National Coordinator for Health Information Technology (ONC) defined an unbiased and equitable AI tool as one that "does not exhibit prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics. The impact of using the [tool] is similar across same or different populations or groups."[13] We believe this definition includes protected classes around age, race, and gender, but also includes other subgroups where performance may differ due to geography, insurance status, or comorbidities. At the same time, developers, evaluators, purchasers, and, ultimately, users should be particularly focused on subgroups which have been subject to discrimination, as bias in AI tools could then systematically perpetuate these inequities and lead to unjust outcomes.

There is nuance in this topic. An AI tool can be biased in ways that are not necessarily inequitable as long as performance remains clinically acceptable for the full population of interest. For example, a mortality prediction tool may have much higher accuracy in individuals with conditions that have clearly defined disease courses, but still work acceptably well in the rest of the indicated population. On the other hand, AI tools can also be accurate across subpopulations, but the typical use of tool leads to inequitable outcomes because the problem has not been well defined or the actions taken with respect to the prediction cause injustice. It can also be difficult to measure if a tool has biased performance or causes inequities if the true patient outcome of interest is not standardly defined, not typically collected in health care systems, or takes an extended time to occur.

It is critically important for processes and tests to be put in place throughout the development cycle to allow developers and manufacturers to mitigate potential issues before they become a problem or, alternatively, “fail fast” and go back to drawing board. This includes involving other stakeholders and experts on the development team, including people knowledgeable about the data being used to train AI/machine learning (ML) tools, potential users, and patients. Additionally, it is important to test for bias after the tool is developed, both before deployment but also in regular intervals after deployment. Biased performance could result later in time from updates to the AI tool, but also changes in the data used by the tool to compute the results.[14]

FIGURE 2 | AI Use Cases in the Health Care Ecosystem



Population Health

- Surveillance and population-level predictions
- 🟨 Population risk management
- 🟨 Non-pharmaceutical intervention selection and targeting

Individual Health

- 🟨 Triage and self-referral*
- 🟨 AI-enabled risk prediction and diagnosis*
- 🟨 AI-enabled drug selection*
- 🟨 Remote patient monitoring*
- 🟨 AI-enabled care (e.g.,Robotic surgery,DTx)*

Health Systems

- Capacity planning and personnel management
- Coding and billing
- Medical record-keeping
- 🟨 Quality assurance

Pharma & MedTech

- Clinical trial support and recruitment
- Drug discovery
- Drug safety and pharmacovigilance
- Real-world evidence
- Health Economics and Outcomes Research

🟨 Products within the scope of this paper

* Products that may be under FDA authority

Categories and examples have been modified from USAID’s 2019 paper “AI in Global Health: Defining a Collective Path Forward”[12]

Bias in AI tools is not an issue specific to health care and has been seen in applications from facial recognition to finance to criminal sentencing.[15] However, there are unique challenges to addressing bias in health care settings. Health data is often complicated by a lack of common definitions and standardization, resulting in challenges with making a cohesive, interoperable data system. Because of historical and ongoing systemic racism and discrimination in health care and the subjective nature of much of health data, AI can be biased in many ways. While other sectors can fully leverage the "move fast and break things" mantra of Silicon Valley, doing so in the health care setting could have significant consequences for life and well-being. With the rise of utilization of AI tools in health care settings, it is crucial that there are processes in place to remove or minimize bias in existing and newly built systems, rigorous tests to detect bias when it does arise, and a shared commitment from all stakeholders, including developers, purchasers, data originators, and regulators to ensuring justice and equity in health care.

Bias in health care also long predates the advent of AI. For example, systemic racism and discrimination have resulted in racial and gender biases in how providers rate and interpret patients' pain in real-world settings.[16] Only in recent years have strides been made to ensure clinical trials are representative of real-world populations, but more work remains. There are also wide geographic gaps in access to tests, procedures, specialists, and treatments both within and across countries.[17] With these historical and contemporary experiences in mind, it is clear that we need to examine how bias within AI can perpetuate human biases that already exist in health care, what steps need to be taken to ensure health care uses of AI do not ingrain the same biases in a newer system, and how AI might be leveraged to mitigate existing biases.

With these challenges in mind, this white paper explains how bias and inequities can be introduced and incorporated into AI-enabled health tools during the origination and development process, along with strategies for testing and detecting bias to inform regulators and other evaluators of AI, policymakers, and those responsible for creating, purchasing, implementing, or maintaining AI tools.

How AI Tools Become Biased or Inequitable

Bias and inequities in health care AI can arise in many ways. Here, we categorize how bias and inequity can be introduced and incorporated into AI tools throughout the origination and development process, and how Good Machine Learning Practice (GMLP) and other tools can help identify and mitigate potential issues. Others have also explored this topic in some depth.[18,19] This section will walk through four major areas where bias or inequities can emerge, highlight examples, and propose potential solutions while identifying where research on best practices is still needed.

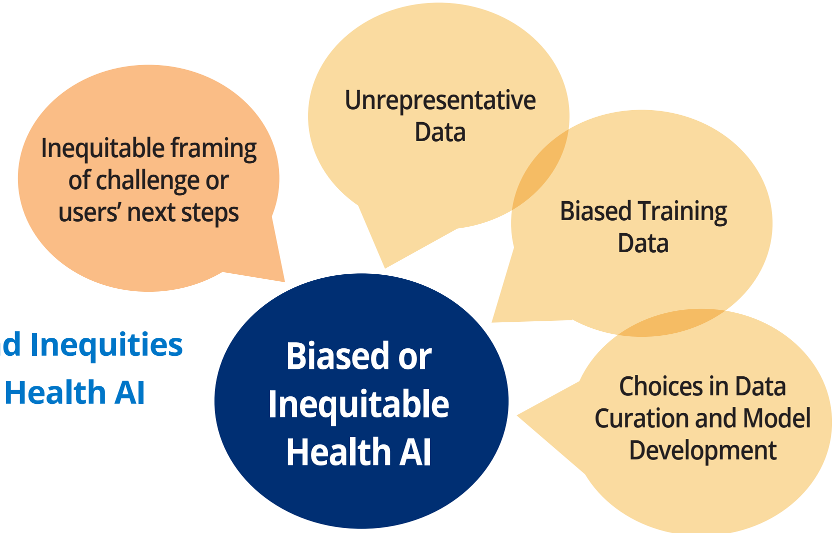

Inequitable framing of challenge or users' next steps
Unrepresentative Data
Biased Training Data
Choices in Data Curation and Model Development
Biased or Inequitable Health AI

FIGURE 3 How Bias and Inequities Can Arise in Health AI

Inequitable Framing of the Health Care Challenge or the Users' Next Steps

The first area is one of the most difficult to address because it includes the risk of building an AI tool that does not necessarily have biased performance across subgroups but does cause overall worsening of health disparities, contributing to an unjust health system. It is the developer's responsibility to ensure that the tool is equitably solving problems, both in the identification of the problem being solved and the subsequent actions taken by the users of the tool. As such, the first step of the development process is to clarify the problem that needs to be solved, including understanding what factors may be causing the challenge and identifying subgroups more likely to be affected by the problem. The second step, which is equally important to do at this initial development stage, is to identify what actions will be taken in response to the AI-generated output/prediction. This could be an automated action, done without a qualified health care professional's involvement (e.g., a cardiac defibrillator detecting an abnormal rhythm and delivering an electric shock or making a diagnosis that would normally require a specialist exam). In these cases, the initial actions will be fully defined. In contrast, many if not most AI tools are meant to assist or augment health care workers.[20] These tools may only produce a list of possible actions or even just display a risk score, leaving the decision-making to the user. Even if the tool is not going to be making specific recommendations for action, developers have a responsibility to work with potential users to understand how the results of the AI tool will be used in the real world. Developers need to interrogate how their tools will impact care decisions throughout patient journeys and their ultimate outcomes. This analysis is critical to ensuring that the development team is developing the most useful product possible, and that it won't exacerbate health disparities.

In a December 2021 public meeting held by Duke-Margolis on this topic, participants highlighted a good example of this definitional challenge – the "no-show" prediction tool. Health care systems often lose money when patients do not show up for appointments, since they cannot bill for services that were not performed. Both health systems and commercial developers have built AI tools to predict which patients may not show up to appointments. Schedulers then often double-book appointments for patients at high risk of not appearing. However, individuals who are more likely to be a "no-show" are often Black, Latino, and American Indian/Alaskan Native patients, who disproportionately experience systemic barriers to accessing care such as lack of reliable transportation, limited access to paid sick leave or affordable health insurance that may prevent them from being able to get to an appointment.[21,22] Additionally, many patients from these same racial and ethnic communities already experience disproportionally worse health outcomes in comparison to white patients.[22] Some of these tools were accurate in predicting the risk of a no-show, but these predictions were just probabilities. So, when both double-booked patients did come to their appointments, they were either not seen promptly or were rushed through their appointments, reducing the quality of their care. This negative experience could also impact the likelihood of these patients attending their next appointment, continuing a cycle of worse outcomes.

Identifying the people most likely to not show up could be an equitable tool if, rather than double-booking, the actions taken in response are supportive efforts such as reminders or ride-shares to appointments. But a team at University of California, San Francisco explored this issue within their own health system and determined it was more useful to reframe the challenge to "which supportive measure is most likely to help this patient attend their appointment?", a framing that directly supports taking more equitable actions.[23] Framing the question as "who won't show up" was provider-focused and concerned about maximizing hospital revenues; the reframed challenge was patient-focused and concerned maximizing patient care. As stated by Sara Murray during the aforementioned Duke-Margolis public meeting: "for example, let's not predict no shows hoping that the folks working at the clinic desk know what to do with that. Let's predict who would benefit with from conversion of their in-person appointment to a video visit. Let's predict who needs [transportation] or their parking to be compensated. Let's predict who would actually just benefit from a reminder call. So that's what we're working on now, really designing the questions in a way that mitigates bias and discrimination in the application."[24]

Examples like this show why developers need to work with providers and patients to understand the actual problem at hand and why certain populations may be more or less likely to have that problem. Developers should frame problems in ways that create the most just and equitable solutions possible to the identified challenge, so they don't build a tool that worsens health inequities even if the direct outputs are not biased. Developers need to conduct rigorous user testing of products to fully understand how users will respond to AI predictions, whether or not specific actions are recommended. Likewise, to evaluate the potential impact of those responses, a diverse set of patients, caregivers, clinical staff, and others that will be impacted by those actions should be approached for input as to the likely results. Even if the AI tool itself won't be making specific suggestions for actions to take, this analysis should be done to identify and evaluate the likely actions that will be taken to understand the potential for inequitable outcomes.

Another way in which bias can be introduced during this phase of development is through assumptions made by the development team. For instance, are the developers assuming that the AI tool will only be used in a particular patient population subgroup? How might this assumption directly impact the functionality of the tool? How will they make the limitations clear to users? Alternatively, has the problem been defined with respect to a system's ability to respond? Consider an AI tool intended to determine who is at high risk of hospitalization so health systems can provide additional support to those identified to potentially prevent that hospitalization. Because many health care systems have limited resources available for providing increased support to populations experiencing this higher risk, they may not be able to provide support to all the identified patients. Therefore, only patients at the very highest risk are offered the additional support. Developers could consider designing a tool to predict which patients will benefit the most from additional support, ensuring that limited resources are used more effectively to improve overall population health. Should such an approach be taken, much care would be needed to ensure the extra resources are provided in a way that encourages more equitable outcomes, as external factors will also affect ultimate outcomes.

Development teams need to have a process for involving representatives from all relevant stakeholder groups, including patients, in identifying the current problem to be solved, understanding how the results from the AI tool will be used, and considering potential adverse impacts, with particular attention to impacts that may worsen racial, ethnic, or socio-economic health disparities. This process should clearly map out both the user and patient journeys that would include the use of this tool, and identify downstream effects where use of the tool may lead to continued or worsened inequities. Engaging bioethicists, health equity researchers, or health anthropologists can help the development team consider additional viewpoints, broaden discussions, and identify underlying assumptions.

Unrepresentative Data

Once the problem that is being solved has been carefully and properly framed, data need to be collected to train the algorithm(s) used in the AI tool. It is important that every effort is made to find broadly representative data, consistent with the locations and populations with which the tool will be used to ensure that the data used to train the AI tool has the same heterogeneity as the data that will be used with it after implementation.[25] For example, there are geographic biases to much of the data used to train AI. One study showed that the majority of papers describing AI tools in health relied on data from just 3 states (California, New York, and Massachusetts) for training those tools.[26] Much of the data comes from large hospital systems and academic health centers, rather than community hospitals, ambulatory care, or public health departments, where data may be less easily accessible to developers. For example, the Medical Information Mart for Intensive Care (MIMIC) is a database of de-identified ICU data that have been used to train many AI algorithms, and consists entirely of data from a single large academic medical center in Boston, MA.[27] Unrepresentative training sets can also lead to bias if there are differences in presentation or risk between different subpopulations, but insufficient data to train the tool to understand those different presentations or risks. One infamous example of this source of bias is the lack of darker skin tones in training sets for AI tools intended to detect melanoma or other skin lesions, where signs of disease may present in different ways.[5,28]

When training sets are unrepresentative of real-world diversity in race, geography, socioeconomic status, medical status, gender, or other factors, there is a higher risk that predictions generated by the model will be less accurate for those unrepresented subpopulations. Every effort should be made to ensure a representative dataset.. If a developer is not able to use representative data, they have an even greater responsibility to carefully test the trained tool for bias. Tools built on unrepresentative data have been shown to have generalizable performance. However, if the tests show biased performance, developers should work to improve the representation within the training data and retrain the algorithm. Then retest to see if the performance is now equitable among subpopulations

Checking training datasets to determine if they are representative of the relevant patient population should be straightforward if those classifiers are present within the data. Developers should document how representative their training data is and evaluators should routinely examine this information. But there are challenges to doing this that the overall health ecosystem needs to work together to solve. Developers often use de-identified real-world datasets for training, as it is an efficient way to gather large datasets. Classifiers on race, socioeconomic status, or other social determinants of health are generally not present in those datasets due to patient privacy laws or because the information is not collected during the general course of care. When those classifying data elements are not present, it becomes challenging to determine if an algorithm under development is being influenced by underlying bias in the data. Privacy laws are being re-examined as technological advances have changed the risks, benefits, and utility of current practices. A concerted effort and discussion among all stakeholders are needed to determine how best to collect and allow use of health information in ways that preserve privacy but provide a better understanding of bias and equity within health systems broadly, which will also benefit AI development.

A separate challenge is to define what it means to be representative. First, simple representativeness of small subpopulations may not be sufficient to adequately train a model (for example, training a model to predict breast cancer risk for both women and men, given the low rate of breast cancer in men). In those cases, oversampling of the smaller populations may be required. Second, more research is also needed to understand how much intersectional representativeness is needed. For example, rather than just individually looking at race, then gender, then each other factor individually, do we need to look at the proportion of Black women and white men? Or potentially even more granular, combining race, age, geography, co-morbidities, or socioeconomic status? When only a single factor is examined, important intersectional differences can be masked.29 But perfect representation along all dimensions is impossible and it is unclear how developers should identify key subpopulations for any given project. Health equity researchers are a useful resource to understand which intersectional factors should be considered, and disease specialists may have a better understanding of how co-morbidities or disabilities may affect representativeness.

Going forward, government and private sector efforts to collect high quality data for broad training of health AI should take care to include data on relevant subpopulation characteristics and work to ensure that these datasets are representative of the patient population. Endeavors such as the Data Nutrition Project are assisting in this effort by advocating for nutritional-like labeling for datasets so that developers are readily aware of what the dataset contains.[30] Similarly, Sendak et al. advocate for the creation of a “Model Facts” label to provide transparency for clinicians and clinical end users so they know exactly when and how to incorporate model output into clinical decisions in order to prevent harm to patients.[31] Beyond efforts to test for and identify biases, government efforts should incentivize the creation of more representative and informative datasets for training new algorithms. Initiatives like AIM-AHEAD and the National Artificial Intelligence Research Resource (NAIRR) may help contribute to these efforts.[32,33] Projects like NIH’s All Of Us have worked to create methods for broad inclusion and participation, which should act as models.[34]

Biased Training Data

Even once a representative training dataset is collected, AI tools may still incorporate and reflect bias recorded within the data itself. Because of a history of inequity in our health care system that persists today, data from real-world and other sources available to train new AI tools reflect human biases and inadequate supports for historically marginalized communities. For this reason, researchers advocate for an understanding of structural racism in health care and how it impacts AI tools.[7] Without closely working with individuals that input data into these systems, it is difficult to have a complete understanding of the biases that may be present in individual data elements that are being used to train their systems. Even then, subconscious bias and other factors may not be apparent to those entering data. For example, several tools that predicted Covid-19 deterioration used oxygen saturation sensor data from fingertip devices, which use laser-based sensors that have less accurate performance in individuals with darker skin or other attributes.[35–37] Sensor accuracy can be affected by "poor circulation, skin pigmentation, skin thickness, skin temperature, current tobacco use, and use of fingernail polish".[38] This is an example of why each data element needs to be critically examined for relevant subpopulations where that data element may be less accurate or informative.

Bias and access issues within health care can also affect recorded data, for example in what diagnoses are considered and what labs or imaging tests are ordered or completed. Certain groups may have less well-understood signs and symptoms for common diagnoses, again causing misdiagnoses or delayed diagnoses. For example, women often present with or describe symptoms of heart attack differently than men, and dermatological issues present differently on darker skin. Other subgroups may be more distrustful of medical advice due to the historical and ongoing issues we have described and therefore they access fewer services despite having similar disease severity and after controlling for income. All of these issues, depending on how the training data is labeled or annotated, can build existing biases into AI systems. These issues with biased data are also challenging because training data does need to reflect the challenges of the real-world data that it will use after implementation to make predictions. However, careful design of systems which are trained to expect those differences could also act to reduce overall bias and increase more equitable care. For example, the Visual Dx program gives users visual examples of how different dermatological conditions appear on skin with different levels of pigmentation and considers how conditions appear in different individuals when suggesting possible diagnosis based on descriptive text.[39]

One method to prevent this bias is to carefully interrogate the data elements planned for use in training and avoid using data elements known to have bias. Unfortunately, while systematic bias within health care is well documented, bias within specific data elements is less understood and documented. While some disparities are well documented, such as in pain management, other disparities that likely exist given the historic and systemic inequities in the health care system have not yet been uncovered. Development teams should carefully consider what data elements to include when training algorithms. Automated tools like those produced by IBM and Carnegie Mellon researchers (formerly from University of Chicago) can help identify some biases that emerge in training data, generally for a single subgroup type and for qualitative data.[40–42] For objective data measured in standardized ways by sensors or other devices, developers should research any populations for whom the sensor or measurement systems may be less accurate. Developers should also consider if the AI tools may use this data in ways that the sensors have not been tested. For example, a device may be used clinically to give real-time warnings of abrupt changes, and individual measurement values may not be considered clinically relevant. AI tools, however, may find patterns in the individual measurements that get weighed more heavily in predictions. In those cases, any differential performance issues in individual accuracy numbers of those devices may be less well known because that is not the intended purpose.

For other types of data, especially subjective data, teams should interview individuals that collect and use that type of data, ideally from the system(s) from which it came. Patients, particularly those with complex or chronic conditions, also often have deep knowledge of challenges with health data relevant to their health and their expertise should also be consulted.[43,44] Teams could also include experts that have examined how human biases impact the types of health data being used for training. For instance, Park et al. have shown that some clinicians express negative attitudes towards patients in their provider notes through the use of stereotypes based on the patients' race or social class.45 Peer-reviewed publications on particular subsets of data such as disparities in documented pain between Black and white patients or heart attacks in women, can help provide guidance on how biases impact a specific target or disease area.[6,8,9,16,46,47] When training data does have subgroup information, developers can also look to see if there are subgroup differences within data elements. However, it may still be necessary to use data that is known to have a bias. In these cases, careful evaluation of affected subgroup performance is particularly critical. Developers can also analyze performance across different health systems that have features (different geographies, patient demographics, dominant insurance types, etc.) that may make those biases more or less likely to better understand the generalizability of the tool.

Those evaluating AI tools for use should also carefully consider what data elements were used for training and how the developers examined that data for potential bias. Regulators, providers, patients, or other users considering the use of AI tools should ask detailed questions about the data required to operate the tools, where the training data came from, and how those systems collected and defined the data elements included. At the same time, health systems need to also rigorously examine and understand the challenges within their own data and work to mitigate data quality disparities as well as improve quality overall. More research and transparency are still needed to understand how these biases impact health care data.

Choices in Data Selection, Curation, Preparation, and Model Development

Data preparation and model development choices can also contribute to bias in the resulting algorithm. Choice of proxy variables[a] for labeling/annotation purposes, combining data inappropriately, removing subjects with missing data, and optimization choices can all cause bias in the resulting AI tool, as this section explains.

One prominent example of bias caused during model development was highlighted by Ziad Obermeyer and team's analysis of a tool that hospital systems used to identify patients at high risk of serious illness to allocate supportive services.[49] That tool was built using historical health care expenditures, which was used as a proxy of the risk of developing serious illness. Because Black patients historically have lower rates of accessing the health care system for a variety of reasons, even when significantly ill, costs are a racially biased proxy for severity of illness. As such, the use of the tool to predict illness severity systemically underestimated the likelihood of serious health conditions for Black patients, who were then not offered preventative and supportive services.

Using proxy variables for labeling/annotating is often critical to the efficient development of AI tools. Proxy variables and computable phenotypes[b] can also be used to address missing data or signs, symptoms, and diagnoses that are not coded in common ways across health systems. In each case of proxy variables or data curation, development teams need to very carefully consider the method being used for potential introduction of bias. The Algorithmic Bias Playbook, developed by Obermeyer and Chicago Booth's Center for Applied Artificial Intelligence, explains this issue as a "label choice bias" where developers use a proxy variable that is associated with the outcome they actually want to predict.[51] This becomes an issue when the proxy variable is not consistently associated with the actual outcome across different subgroups, as in the example above with care costs and seriousness of illness. They suggest that developers (and users) clearly articulate the actual outcome that they wish to predict (the "ideal target"), then identify what the algorithm was specifically trained to predict (the proxy variable), and then analyze and interrogate potential biases in the prediction of the ideal target when using that proxy (see the Playbook for more details).[51]

One common challenge is whether to include race or other subgroup classifiers as inputs into the tool at all. This is particularly being discussed regarding rules-based AI tools where adjustments for race or other variables have been added to the calculations within the tool in an attempt to correct the final score to better statistically fit historical data. There has been much discussion about removing race variables and the associated adjustments from assessments of kidney function, ICU risk scores, and other common performance measures.[10,52,53] As a social construct, it is not clear why race would have an effect on these scores, and it is generally believed that race is acting as an imperfect proxy for other environmental or medical risk factors, by which certain races are disproportionally affected. It is clear that, when possible, the actual risk factor should be used to better personalize care. It should be noted that simply removing subgroup classifiers

[a] Proxy variables are data elements used in place of something that may be more pertinent, but also more difficult to measure. This may be because the actual element of interest is difficult to quantify, people may be sensitive to disclosure, or it is not routinely recorded in the database(s) being used. Examples include using blood pressure as a proxy for cardiovascular health or educational attainment as a proxy for income.[48]

[b] "computable phenotype" is a definition of a condition, disease, or characteristic or clinical event that is based solely on data that can be processed by a computer.[50]

does not prevent ML-based systems from being able to impute that data from other information and continue to weight it. For example, ML tools were shown to be able to determine race from imaging files with no other identifying information.[54] Therefore, testing the final product remains a key step, even with these mitigation measures.

Another data curation challenge relates to whether to include subjects with significant missing data. As discussed above, systemic inequities and bias can affect the "completeness" of a medical record. For example, providers may choose not to prescribe certain medications or diagnostic labs if they know the patient can't afford the co-pay or aren't able to travel to a specialist. Lacking access may also result in missed diagnoses or no treatments, despite symptoms being present. Removing these individuals from the training dataset may reduce performance of the tool for these populations.

Once the training dataset has been collected and curated, development teams must choose how to train the model and optimize the results. Without going too deeply into technical specifics, these choices have significant impacts on how well the tool works for subjects whose data is significantly different from the majority of the data in the training set. For example, development teams may choose to use model compression to allow AI tools to run on mobile devices or improve patient privacy by adding noise to an algorithm's analysis to limit the possibility of reidentification. But both of these choices can degrade the performance of algorithms in subpopulations that "look different" to the AI because they have data elements with values that are less common. While these modifications can have important benefits, it is important to make sure the accuracy of the algorithm on attributes that do not appear frequently in the data set is not compromised. Hooker has described how a given algorithm handles these low frequency attributes can affect whether the algorithm is sensitive to small changes in subpopulations with low representation in a given dataset.[55]

This work exploring exactly how these model development decisions impact bias is in early stages, so this will be an important area for AI researchers to consider moving forward. Developers should work closely with health equity researchers when choosing proxy variables, and carefully test performance using the actual outcomes of interest. More work is also needed to understand how missing data affects accuracy, how competing technical priorities can affect performance in biased ways, and how to develop optimization methods that prioritize equity.

Addressing Biased or Inequitable Performance in AI Tools After Development

Given these sources of biases and inequities, what can be done if a tool does show biased performance after development? The best option is to determine the cause of the bias and to go back to development and retrain the model. A second option is to make clear (within the product use instructions and during marketing and training) that the AI tool is only intended for use in certain populations. In the context of FDA regulated devices, the device label would need to make this clear. However, device labels may not be enough if users are unlikely to see or retain the information in the label. FDA would also need to consider if other steps were necessary to ensure the user knows when they may be using a device off-label. For example, a user may have to enter certain patient data that matches the population of use before the device will report results.

But there are also important ethical and legal considerations to taking this label-based approach, more so if the lower performance is seen in subgroups that are considered protected classes. For instance, if an AI tool only accurately predicts the risk of heart attacks in non-Hispanic white men, is it equitable to allow marketing of that product if its use is restricted to only that population? Does that risk systematically perpetuate inequities for marginalized populations for which this tool is inaccessible? On the other hand, what if the tool is purposefully optimized to better diagnose conditions in historically marginalized or medically underserved populations, and therefore does not perform as well on populations that historically and currently have had better outcomes? This type of bias, as long as the user was properly trained on who to use it for, may be acceptable from a health equity standpoint, although there could be legal challenges if the subgroups involved protected classes. The Federal Trade Commission (FTC) has made clear that the "FTC Act prohibits unfair or deceptive practices. That would include the sale or use of – for example – racially biased algorithms."[56] However, it is unclear if that prohibition would also be true for non-protected subpopulations that are based on geography or socioeconomic status. So, while proper documentation and labeling may help prevent biased outcomes from AI tools trained on limited data, additional health equity and health justice issues will arise if only some parts of the population have access to AI tools. Developers considering this approach would also need to be aware of the potential for reputational damage, even if the approach is legal. The complicated reception of BiDil, approved in 2005 for the treatment of heart failure in "self-identified" African Americans, may offer some insights into potential physician, patient, and media reactions.[57,58]

Regulators and policymakers will need to examine the potential for disparate outcomes when AI tools are not trained to work with the entire patient population. There is an argument that we should allow tools, even if they don't work on the entire population, that would have a significant positive impact as long as the use doesn't change the risks for others. One could even argue that certain triage or screening tools used on a portion of the population could allow physicians to spend more time on complicated cases where the AI tool may not work as well, improving overall outcomes. But there are other considerations, such as whether provider skills are likely to degrade as they start to depend on the AI tool over time, leading to poorer patient outcomes when the tool can't be used. For example, if an AI tool assists in proper imaging positioning from technicians for most of their patients, will they know how to correctly position patients for whom they don't have AI assistance? Or will those patients suffer from having lower-quality images throughout the rest of their care journey? Long-term outcome-based research would be needed to fully understand the implications.

Federal Efforts to Address Bias and Inequities in AI

As AI use in health care settings expands and bias issues are increasingly recognized, regulators are grappling with how best to review and approve these products. In the U.S., the FDA has released the Artificial Intelligence/Machine Learning-Based Software as a Medical Device (SaMD) Action Plan.[59] This action plan highlights the goals of the FDA related to furthering the regulation of AI products. Key goals include developing an updated regulatory framework, harmonizing development of GMLP, working with patients to support transparency in AI devices, and advancing pilots to provide additional clarity on evidence generation for AI devices.[60] After experts highlighted that it is critical that the FDA include addressing bias as part of their regulatory framework, the FDA also included in their action plan support for regulatory science efforts to develop evaluation methodologies for AI devices, including those that will assist with the identification and elimination of bias.[61] In 2021, the FDA, in partnership with Health Canada and the United Kingdom's Medicines and Healthcare products Regulatory Agency, released guiding principles for the development of GMLP that, among other principles, stresses the need for representative datasets and the importance of monitoring deployed AI tools. Elsewhere, businesses in the AI sector have shared that the FDA has asked them to conduct extensive subgroup analyses to include in submission for regulatory review and approval.[24]

However, as noted above in **Figure 2**, the FDA's ability to regulate AI in the health care setting is limited to certain kinds of AI tools based on the definition of what constitutes a medical device. Population health management tools that don't give person-specific recommendations regarding the prevention, diagnosis, or treatment of a specified disease or condition (e.g., tools that make predictions around cost, hospitalization, or death) are generally not under FDA authority. The 21st Century Cures Act, passed in 2016, also limits the types of CDS software that is under FDA authority.[62,63] CDS software that processes data from medical images, laboratory tests, or sensors is generally considered a medical device when used as part of predictions for a specified disease or condition.

However, CDS that works with electronic health records and claims data that is only meant to support or provide recommendations to a provider about the prevention, diagnosis, or treatment of a disease or condition while also enabling that provider to review and understand the basis for that recommendation was removed from FDA authority under this Act.

Other federal agencies have regulatory power over AI tools outside FDA authority, and others are working to build guidance and standards to facilitate building more just AI tools. The FTC, for example, has provided guidelines to avoid discriminatory outcomes and conducted enforcement actions on companies falling short.[56] The National Institute for Standards and Technology (NIST), while not an enforcement body, has been actively working across the entire AI space to develop a risk management framework.[64,65] This framework aims to foster the development of innovative approaches to address characteristics of AI trustworthiness including accuracy, interpretability, reliability, privacy, robustness, safety, security, and mitigation of harmful bias. NIST also published a separate report on identifying and managing bias in AI that describes three broad challenges for mitigating bias and introduces preliminary guidance for addressing them.[66] The Office of Science and Technology Policy has championed the need for an AI Bill of Rights and hosted a range of events, including one focused on health care, to seek public input on the impacts of AI technology.[32,67]

Within other agencies and offices overseen by the Department of Health and Human Services, additional work on how to use AI is underway. ONC has championed "health equity by design" where equity is a core design feature of the office's collective health IT endeavors.[68] In the Center for Medicare and Medicaid Services' Artificial Intelligence Health Outcomes Challenge, finalists had to consider bias while competing to accelerate development of AI solutions for predicting patient health outcomes for Medicare beneficiaries.[69,70] At the National Institutes of Health, the AIM-AHEAD program seeks to establish new partnerships to increase participation and representation of researchers and communities that are currently underrepresented in AI development.[33]

Meanwhile in Congress, there are legislative efforts underway to require that companies assess the impacts of AI and other automated systems that they deploy for a range of factors of relevance to consumers, including bias. The proposed Algorithmic Accountability Act of 2022 would empower the FTC to create and enforce regulations for this assessment.[71]

Regulation, Evaluation, and Implementation

In the previous sections, we emphasized the responsibilities of developers to build equitable AI tools and continually test for and mitigate bias. However, there are other stakeholders that have responsibilities as well, including regulators, purchasers, users, and data originators. Each of these stakeholders has a different role and perspective in the landscape. Developers have a responsibility to build high-quality AI tools, and show through the use and documentation of GMLP and rigorous testing of the tool that it is generalizable across relevant subgroups. Evaluators should provide a secondary assessment of whether the tool was developed with best practices and whether the information provided shows that the tool functions as claimed across relevant subgroups. Health systems procuring the AI tool, along with regulators such as the FDA and third-party evaluators are within this group. Below we make recommendations for both the FDA and purchasers to consider for evaluating AI tools to ensure they are unbiased and that critical performance measures are available to users.

For Health Systems and Other Purchasers:

Health systems and payers are in the unique position of being involved in multiple points of the development and use of AI tools. Many health systems and payers, particularly academic health systems and large payers, build and deploy "homegrown" AI tools. More are evaluating and implementing commercial products, sometimes as part of a co-development process or a purchasing decision. Finally, health systems and payers are also data originators, generating the real-world health data that is or could be used for training and testing new products. They are in a position to improve equitable and interoperable health data collection and to collect data to identify relevant subpopulations, which will ultimately lead to better AI-enabled tools and more efficient subgroup performance testing. ONC is encouraging this through their work on "health equity by design" mentioned earlier.

Purchasers, particularly health systems that will use AI tools on large numbers of patients, have a responsibility to evaluate these tools both before and after implementation and examine both the accuracy of the outputs and the ultimate clinical outcomes for biased performance or inequitable outcomes among subgroups among their patient populations. Evaluating ultimate clinical outcomes is critically important, as seen in the Optum and UCSF no-show use cases described earlier. There have not been many systematic attempts to ensure all tools are evaluated pre- and post-implementation, particularly for longer term outcomes, although pilot stage testing is not uncommon. Some academic systems are starting processes to catalog the AI tools being used within their health systems and ensure they are regularly evaluated.[72] The Mayo Clinic has started an accelerator platform to partner with companies to do testing.[73] In contrast, smaller systems may have trouble finding the resources for this type of testing or even finding enough patients to perform subgroup evaluations for tools that are used for specific diagnostic or treatment decision-making. There may be a role for third-party reviewers that can aggregate data from several similar smaller systems to test for overall performance and bias in these cases.

Health systems and other large purchasers also have a unique ability to use their purchasing power to demand evidence from manufacturers that these tools work equitably across patient populations. After purchase, they also should be testing performance themselves within their own unique workflows and data recording practices, particularly for higher-risk tools or tools that allocate resources. This will require a balance – while purchasers such as health systems and insurers can and should be responsible for this when deploying tools widely across their patient populations, it would not be reasonable to place this type of responsibility on tools purchased and used by individual laypeople, such as non-prescription digital therapeutics or wearables, or tools that may be used by too few patients within a health system to allow rigorous evaluation of performance between subgroups. In these cases, there may be utility in the creation of third-party review systems to test for biased performance, particularly for tools that are not under FDA authority. For example, pharmacy benefit managers have begun to test digital therapeutics in order to add them to their formularies.[74]

For FDA:

The FDA should continue to work with standards groups and collaborative communities[c] to develop GMLP and best practices for testing AI-enabled software performance, with a focus on mitigating the potential of and testing for bias.[75]

FDA has emphasized the importance of transparency in AI tools, including holding a public meeting on the topic in October 2021 that highlighted how transparency can combat bias.[60] This included presentation on "Data Nutrition Labels" and the importance of standardized, interoperable data that includes marginalized or underrepresented community subsets. FDA should continue to embrace transparency, including by publishing guidance on what subpopulation data should be labeled and how it should be presented, and clearly specifying the populations used in training and in testing. Once marketed, FDA should require clear information on how data was annotated during training and the comparison used during testing. There should be clear warnings about any significant differences in tool performance across relevant subpopulations. This information should also be included and expanded upon in the label for these devices, but FDA also needs to consider how to ensure potential users, patients, and researchers can easily access this information since labeling for devices is not required to be made public.

While the FDA can't currently require subgroup analyses, they should and do strongly suggest performance testing be done in diverse settings and with diverse participants.[76] If they are reviewing products where manufacturers have chosen not to do so, there should be a risk-based determination of whether the label should reflect only

[c] Collaborative Communities are continuing forums for stakeholders to work together on medical device challenges to achieve common objectives and outcomes that include but are not led by FDA's Center for Devices and Radiological Health.

the population on which it was tested and if use can be restricted to those populations through automated or other means.

The FDA does not have the ability to control off-label use by health care professionals, so it is hard for them to know if AI tools are being consistently used off-label in populations or uses for which the device was not intended. It is also difficult for FDA to know if off-label use is being done purposefully or unknowingly. Off-label use that leads to adverse outcomes should be reported through FDA's MAUDE system, which allows FDA some limited insight into how devices are being used. When FDA becomes aware of safety issues due to improper or off-label use, they may issue communications to physicians as warnings.[77] They can also ask manufacturers to add safeguards to their software. For example, if a low-resolution image is loaded into an AI tool that is meant only to analyze high-resolution images, it will commonly display an error message. If the population of use is restricted, input data could be required that ensure the device will only report results for the intended population. For devices of higher concern, FDA could ask for post-market surveillance data to identify the populations in which the device is being used and if more safety features are needed to ensure that users know when they are using a device on someone for which it is not indicated.

During pre-market review and post-market surveillance, if the FDA finds evidence that AI-enabled SaMD may not work as accurately for a particular subgroup, they should require a root cause analysis. If possible, the issue should be mitigated. If not, depending on the risk of continued use, the label should be changed to acknowledge the performance issue and user notified or the device should be removed from the market.

Continued Need to Build Consensus Standards and Frameworks

Throughout this work, experts repeated that there are not simple tests or checklists that can be used to ensure tools will not be biased or lead to inequitable outcomes. Because AI tools are made for a broad range of use cases, each tailored to a specific patient population or specialty, it can be difficult to create clear directives or overarching checklists. Similarly, experts note their concerns that a checklist-based approach may discourage developers from thinking deeply and critically about how their specific algorithm has been conceptualized, built, and implemented and instead encourages shallow thinking that "checks off" a vague general list. However, frameworks like the Algorithmic Bias Playbook and others can be considered for approaches to evaluating AI tools to ensure that patients that receive the same tool "score" have the same need or outcome, regardless of sensitive attributes.[51,78]

Stakeholders, led by the government agencies described above, are working collaboratively on GMLP and risk management frameworks that can help prevent bias from entering AI tools during the development process. These frameworks can help by standardizing the development process and creating a set of questions and processes to help development teams think through where there may be potential challenges. Teams that incorporate individuals with diverse expertise and lived experiences regarding health data and health care inequities can identify additional concerns and potential adverse outcomes throughout the development process, particularly in the ideation and data selection stages.[79,80] These tools are being created and implemented within a complex medical ecosystem and much remains unknown about systemic inequities in health care and how that affects data and workflows. So ultimately, AI requires teams of humans to prioritize thinking through these nuanced and complex situations throughout the development cycle.

Looking Forward

While there is much work to be done to ensure AI tools in the health care setting don't ingrain or exacerbate existing biases, we should also acknowledge that AI has the potential to help solve longstanding issues. For example, where there is detectable bias in patient records, AI might be able to flag that bias and ensure institutions are living up to their stated values.[45] There are also ample examples of the potential of AI to address other issues such as biased pain measurement and gaps in image datasets used to teach clinicians on how to identify dermatological conditions. It is important to recognize that AI is a tool developed by humans and can be fallible like any other. However, when used effectively for the intended purpose and with proper oversight, AI tools can provide important value.

This paper focused mostly on how AI is used in health care settings under FDA authority. However, it is also worth considering the role of AI in other similar settings and purposes, such as promoting population health or identifying social determinants of health and how to best meet the needs of patients that have historically had and continue to have adverse interactions with the health care system.

We hope that the information and recommendations provided here will be useful for stakeholders across the health care AI space as they look for strategies and opportunities to mitigate or eliminate bias in health AI to ensure equitable access to quality care for all patients.

Appendix: Methods

We conducted semi-structured interviews with 35 stakeholders representing regulators, academic institutions, health systems and industry. We identified stakeholders by soliciting expert recommendations, through literature reviews, and utilizing snowball sampling to identify additional interviewees. Additionally, our team hosted a public meeting in December 2021, convening 19 experts for 4 panel discussions on relevant topics which also informed the content of this paper.[24]

Interviews were confidential, lasted 30 to 60 minutes and were conducted by web-conferencing software with the authors (two or three authors present per interview), with detailed notes taken. Interviews consisted of semi-structured questions designed by the authors on the basis of literature reviews and prior experience. Open-ended questions explored ways in which bias is introduced into AI, the types of biases which exist in artificial intelligence, how to mitigate bias—especially to resolve inequities, and whether or not artificial intelligence can be useful in reducing human biases which occur in health care. Authors analyzed the noted and identified persistent and important themes spanning across interviews and the public meeting. Those findings are conveyed in this paper.

References

[1] Equity vs. Equality: What's the Difference? GW-UMT. Published November 5, 2020. Accessed May 3, 2022. https://onlinepublichealth.gwu.edu/resources/equity-vs-equality/

[2] Braveman PA, Kumanyika S, Fielding J, et al. Health Disparities and Health Equity: The Issue Is Justice. *Am J Public Health.* 2011;101(Suppl 1):S149-S155. doi:10.2105/AJPH.2010.300062

[3] Braveman P, Arkin E, Orleans T, Proctor D, Plough A. What is Health Equity? RWJF. Published May 1, 2017. Accessed May 27, 2022. https://www.rwjf.org/en/library/research/2017/05/what-is-health-equity-.html

[4] AAMC Center for Health Justice. Who We Are and What We Do. AAMC Center For Health Justice. Accessed June 27, 2022. https://www.aamchealthjustice.org/who-we-are-and-what-we-do

[5] Adamson AS, Smith A. Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatology.* 2018;154(11):1247-1248. doi:10.1001/jamadermatol.2018.2348

[6] Schulman KA, Berlin JA, Harless W, et al. The Effect of Race and Sex on Physicians' Recommendations for Cardiac Catheterization. *New England Journal of Medicine.* 1999;340(8):618-626. doi:10.1056/NEJM199902253400806

[7] Robinson WR, Renson A, Naimi AI. Teaching yourself about structural racism will improve your machine learning. Biostatistics. 2020;21(2):339-344. doi:10.1093/biostatistics/kxz040

[8] Badreldin N, Grobman WA, Yee LM. Racial Disparities in Postpartum Pain Management. *Obstet Gynecol.* 2019;134(6):1147-1153. doi:10.1097/AOG.0000000000003561

[9] Green CR, Anderson KO, Baker TA, et al. The unequal burden of pain: confronting racial and ethnic disparities in pain. Pain Med. 2003;4(3):277-294. doi:10.1046/j.1526-4637.2003.03034.x

[10] Vyas DA, Eisenstein LG, Jones DS. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. *New England Journal of Medicine.* 2020;383(9):874-882. doi:10.1056/NEJMms2004740

[11] League J. Confronting the Reality of AI/ML in Care Delivery. *NEJM Catalyst Innovations in Care Delivery.* Published online March 16, 2022. Accessed May 3, 2022. https://catalyst.nejm.org/doi/full/10.1056/CAT.22.0072

[12] Artificial Intelligence in Global Health: Defining a Collective Path Forward. Published April 1, 2019. Accessed June 10, 2022. https://www.usaid.gov/cii/ai-in-global-health

[13] Getting the Best out of Algorithms in Health Care. Health IT Buzz. Published June 15, 2022. Accessed June 23, 2022. https://www.healthit.gov/buzz-blog/electronic-health-and-medical-records/getting-the-best-out-of-algorithms-in-health-care

[14] AI gone astray: How subtle shifts in patient data send popular algorithms reeling, undermining patient safety. STAT. Published February 28, 2022. Accessed June 14, 2022. https://www.statnews.com/2022/02/28/sepsis-hospital-algorithms-data-shift/

[15] AI is sending people to jail—and getting it wrong. MIT Technology Review. Accessed May 25, 2022. https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/

[16] Hoffman KM, Trawalter S, Axt JR, Oliver MN. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences.* 2016;113(16):4296-4301. doi:10.1073/pnas.1516047113

[17] Celi LA, Cellini J, Charpignon ML, et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. *PLOS Digital Health.* 2022;1(3):e0000022. doi:10.1371/journal.pdig.0000022

[18] Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical Machine Learning in Health Care. *Annu Rev Biomed Data Sci.* 2021;4(1):123-144. doi:10.1146/annurev-biodatasci-092820-114757

[19] Wang HE, Landers M, Adams R, et al. A bias evaluation checklist for predictive models and its pilot application for 30-day hospital readmission models. *Journal of the American Medical Informatics Association.* Published online May 17, 2022:ocac065. doi:10.1093/jamia/ocac065

[20] American Medical Association. CPT® Appendix S: Artificial Intelligence Taxonomy for Medical Services and Procedures. Published online 2021. https://www.ama-assn.org/system/files/cpt-appendix-s.pdf

[21] Agency for Healthcare Research and Quality, Boonyasai R, Azam I, et al. 2021 National Healthcare Quality and Disparities Report. *Agency for Healthcare Research and Quality.* Published online 2021:316.

[22] Shimotsu S, Roehrl A, McCarty M, et al. Increased Likelihood of Missed Appointments ("No Shows") for Racial/Ethnic Minorities in a Safety Net Health System. *J Prim Care Community Health.* 2016;7(1):38-40. doi:10.1177/2150131915599980

[23] Discrimination By Artificial Intelligence In A Commercial Electronic Health Record—A Case Study | Health Affairs Forefront. Accessed May 3, 2022. https://www.healthaffairs.org/do/10.1377/forefront.20200128.626576/full/

[24] Understanding Bias and Fairness in AI-enabled Healthcare Software. Margolis Center for Health Policy. Accessed May 27, 2022. https://healthpolicy.duke.edu/events/understanding-bias-and-fairness-ai-enabled-healthcare-software

[25] Sabharwal P, Hurst JH, Tejwani R, Hobbs KT, Routh JC, Goldstein BA. Combining adult with pediatric patient data to develop a clinical decision support tool intended for children: leveraging machine learning to model heterogeneity. *BMC Med Inform Decis Mak.* 2022;22(1):84. doi:10.1186/s12911-022-01827-4

[26] Kaushal A, Altman R, Langlotz C. Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. *JAMA.* 2020;324(12):1212-1213. doi:10.1001/jama.2020.12067

[27] Rogers P, Wang D, Lu Z. Medical Information Mart for Intensive Care: A Foundation for the Fusion of Artificial Intelligence and Real-World Data. *Frontiers in Artificial Intelligence.* 2021;4. Accessed May 27, 2022. https://www.frontiersin.org/article/10.3389/frai.2021.691626

[28] Kinyanjui NM, Odonga T, Cintas C, et al. Fairness of Classifiers Across Skin Tones in Dermatology. In: Martel AL, Abolmaesumi P, Stoyanov D, et al., eds. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020.* Lecture Notes in Computer Science. Springer International Publishing; 2020:320-329. doi:10.1007/978-3-030-59725-2_31

[29] Homan P, Brown TH, King B. Structural Intersectionality as a New Direction for Health Disparities Research. *J Health Soc Behav.* 2021;62(3):350-370. doi:10.1177/00221465211032947

[30] Holland S, Hosny A, Newman S, Joseph J, Chmielinski K. *The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards.* arXiv; 2018. doi:10.48550/arXiv.1805.03677

[31] Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. *NPJ Digit Med.* 2020;3:41. doi:10.1038/s41746-020-0253-3

[32] National Artificial Intelligence Initiative Office. *Envisioning a National Artificial Intelligence Research Resource (NAIRR): Preliminary Findings and Recommendations - An Interim Report by the NAIRR Task Force.;* 2022. https://www.ai.gov/wp-content/uploads/2022/05/NAIRR-TF-Interim-Report-2022.pdf

[33] AIM-AHEAD | Data Science at NIH. Accessed June 10, 2022. https://datascience.nih.gov/artificial-intelligence/aim-ahead

[34] National Institutes of Health. About: All of Us Research Program. All of Us Research Program | NIH. Published July 21, 2020. Accessed June 24, 2022. https://allofus.nih.gov/about

[35] Wong AKI, Charpignon M, Kim H, et al. Analysis of Discrepancies Between Pulse Oximetry and Arterial Oxygen Saturation Measurements by Race and Ethnicity and Association With Organ Dysfunction and Mortality. *JAMA Network Open.* 2021;4(11):e2131674. doi:10.1001/jamanetworkopen.2021.31674

[36] Fawzy A, Wu TD, Wang K, et al. Racial and Ethnic Discrepancy in Pulse Oximetry and Delayed Identification of Treatment Eligibility Among Patients With COVID-19. *JAMA Internal Medicine.* Published online May 31, 2022. doi:10.1001/jamainternmed.2022.1906

[37] Faulty oxygen readings delayed Covid treatments for darker-skinned patients, study finds. STAT. Published May 31, 2022. Accessed June 14, 2022. https://www.statnews.com/2022/05/31/faulty-oxygen-readings-delayed-covid-treatments-darker-skin-patients/

[38] Health C for D and R. Pulse Oximeter Accuracy and Limitations: FDA Safety Communication. *FDA.* Published online June 21, 2022. Accessed June 24, 2022. https://www.fda.gov/medical-devices/safety-communications/pulse-oximeter-accuracy-and-limitations-fda-safety-communication

[39] How VisualDx Works. VisualDx. Accessed June 24, 2022. https://www.visualdx.com/resources/how-visualdx-works/

[40] Introducing AI Fairness 360, A Step Towards Trusted AI - IBM Research. IBM Research Blog. Published September 19, 2018. Accessed June 24, 2022. https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/

[41] Google AI. Responsible AI practices. Google AI. Accessed June 24, 2022. https://ai.google/responsibilities/responsible-ai-practices/

[42] University of Chicago. Aequitas. Data Science and Public Policy. Accessed June 24, 2022. http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/

[43] Bell SK, Delbanco T, Elmore JG, et al. Frequency and Types of Patient-Reported Errors in Electronic Health Record Ambulatory Care Notes. *JAMA Network Open.* 2020;3(6):e205867. doi:10.1001/jamanetworkopen.2020.5867

[44] Organizing Committee for Assessing Meaningful Community Engagement in Health & Health Care Programs & Policies. Assessing Meaningful Community Engagement: A Conceptual Model to Advance Health Equity through Transformed Systems for Health. *NAM Perspectives.* Published online February 14, 2022. doi:10.31478/202202c

[45] Park J, Saha S, Chee B, Taylor J, Beach MC. Physician Use of Stigmatizing Language in Patient Medical Records. *JAMA Network Open.* 2021;4(7):e2117052. doi:10.1001/jamanetworkopen.2021.17052

[46] Tamayo-Sarver JH, Hinze SW, Cydulka RK, Baker DW. Racial and ethnic disparities in emergency department analgesic prescription. *Am J Public Health.* 2003;93(12):2067-2073. doi:10.2105/ajph.93.12.2067

[47] Akinlade O. Taking Black Pain Seriously. *New England Journal of Medicine.* 2020;383(10):e68. doi:10.1056/NEJMpv2024759

[48] Proxy Variables. Data Science Ethics. Published January 15, 2019. Accessed May 27, 2022. https://datascienceethics.com/podcast/proxy-variables/

[49] 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

[50] Richesson RL, Smerek MM, Blake Cameron C. A Framework to Support the Sharing and Reuse of Computable Phenotype Definitions Across Health Care Delivery and Clinical Research Applications. *EGEMS (Wash DC).* 2016;4(3):1232. doi:10.13063/2327-9214.1232

[51] Obermeyer Z, Nissan R, Stern M, Eaneff S, Bembeneck EJ, Mullainathan S. Algorithmic Bias Playbook. Published online June 2021.

[52] Ashana DC, Anesi GL, Liu VX, et al. Equitably Allocating Resources during Crises: Racial Differences in Mortality Prediction Models. *Am J Respir Crit Care Med.* 2021;204(2):178-186. doi:10.1164/rccm.202012-4383OC

[53] Ioannidis JPA, Powe NR, Yancy C. Recalibrating the Use of Race in Medical Research. *JAMA.* 2021;325(7):623-624. doi:10.1001/jama.2021.0003

[54] Gichoya JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health.* 2022;4(6):e406-e414. doi:10.1016/S2589-7500(22)00063-2

[55] Hooker S. Moving beyond “algorithmic bias is a data problem.” *Patterns.* 2021;2(4):100241. doi:10.1016/j.patter.2021.100241

[56] Aiming for truth, fairness, and equity in your company’s use of AI. Federal Trade Commission. Published April 19, 2021. Accessed May 3, 2022. http://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai

[57] Kahn J. Race in a Bottle. Scientific American. doi:10.1038/scientificamerican0807-40

[58] Maglo KN, Rubinstein J, Huang B, Ittenbach RF. BiDil in the Clinic: An Interdisciplinary Investigation of Physicians’ Prescription Patterns of a Race-Based Therapy. *AJOB Empir Bioeth.* 2014;5(4):37-52. doi:10.1080/23294515.2014.907371

[59] U.S. Food and Drug Administration. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. Published online January 2021.

[60] Virtual Public Workshop - Transparency of Artificial Intelligence/Machine Learning-enabled Medical Devices - 10/14/2021 - 10/14/2021. FDA. Published November 26, 2021. Accessed May 27, 2022. https://www.fda.gov/medical-devices/workshops-conferences-medical-devices/virtual-public-workshop-transparency-artificial-intelligencemachine-learning-enabled-medical-devices

[61] Ferryman K. Addressing health disparities in the Food and Drug Administration’s artificial intelligence and machine learning regulatory framework. *Journal of the American Medical Informatics Association.* 2020;27(12):2016-2019. doi:10.1093/jamia/ocaa133

[62] 114th Congress. *Public Law 144-255.* Accessed June 24, 2022. https://www.congress.gov/114/plaws/publ255/PLAW-114publ255.pdf

[63] Commissioner O of the. 21st Century Cures Act. FDA. Published January 31, 2020. Accessed June 24, 2022. https://www.fda.gov/regulatory-information/selected-amendments-fdc-act/21st-century-cures-act

[64] National Institute of Standards and Technology. AI Risk Management Framework: Initial Draft. Published online March 17, 2022. https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf

[65] National Institute of Standards and Technology. AI Risk Management Framework. NIST. Published July 12, 2021. Accessed June 24, 2022. https://www.nist.gov/itl/ai-risk-management-framework

[66] Schwartz R, Vassilev A, Greene K, Perine L, Burt A, Hall P. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.* National Institute of Standards and Technology; 2022. doi:10.6028/NIST.SP.1270

[67] Join the Effort to Create A Bill of Rights for an Automated Society. The White House. Accessed May 3, 2022. https://www.whitehouse.gov/ostp/news-updates/2021/11/10/join-the-effort-to-create-a-bill-of-rights-for-an-automated-society/

[68] Embracing Health Equity by Design. Health IT Buzz. Published February 22, 2022. Accessed June 24, 2022. https://www.healthit.gov/buzz-blog/health-it/embracing-health-equity-by-design

[69] Lessons Learned from the CMS Artificial Intelligence Health Outcomes Challenge | CMS. Accessed June 10, 2022. https://www.cms.gov/blog/lessons-learned-cms-artificial-intelligence-health-outcomes-challenge

[70] Centers from Medicare and Medicaid Services. Artificial Intelligence (AI) Health Outcomes Challenge. cms.gov. Accessed June 10, 2022. https://innovation.cms.gov/innovation-models/artificial-intelligence-health-outcomes-challenge

[71] Wyden, Booker and Clarke Introduce Algorithmic Accountability Act of 2022 To Require New Transparency And Accountability For Automated Decision Systems | U.S. Senator Ron Wyden of Oregon. Accessed May 25, 2022. https://www.wyden.senate.gov/news/press-releases/wyden-booker-and-clarke-introduce-algorithmic-accountability-act-of-2022-to-require-new-transparency-and-accountability-for-automated-decision-systems

[72] Algorithm-Based Clinical Decision Support (ABCDS) Oversight – Duke AI Health. Accessed June 24, 2022. https://aihealth.duke.edu/algorithm-based-clinical-decision-support-abcds/

[73] Mayo Clinic Platform_Accelerate program begins with four AI startups. Mayo Clinic News Network. Published March 23, 2022. Accessed June 24, 2022. https://newsnetwork.mayoclinic.org/discussion/3-23-mayo-clinic-platform_accelerate-program-begins-with-four-ai-startups/

[74] CVS Health program for PBM clients adds five new digital health programs. MobiHealthNews. Published March 11, 2020. Accessed May 27, 2022. https://www.mobihealthnews.com/news/cvs-health-program-pbm-clients-adds-five-new-digital-health-programs

[75] Center for Devices and Radiological Health. Collaborative Communities: Addressing Health Care Challenges Together. *FDA.* Published online December 10, 2021. Accessed May 27, 2022. https://www.fda.gov/about-fda/cdrh-strategic-priorities-and-updates/collaborative-communities-addressing-health-care-challenges-together

[76] U.S. Food and Drug Administration. Diversity Plans to Improve Enrollment of Participants from Underrepresented Racial and Ethnic Populations in Clinical Trials Guidance for Industry. *Clinical Trials.* Published online April 2022:12.

[77] Center for Devices and Radiological Health. Intended Use of Imaging Software for Intracranial Large Vessel Occlusion - Letter to Health Care Providers. *FDA.* Published online April 11, 2022. Accessed May 16, 2022. https://www.fda.gov/medical-devices/letters-health-care-providers/intended-use-imaging-software-intracranial-large-vessel-occlusion-letter-health-care-providers

[78] Estiri H, Strasser ZH, Rashidian S, et al. An objective framework for evaluating unrecognized bias in medical AI models predicting COVID-19 outcomes. *Journal of the American Medical Informatics Association.* Published online May 12, 2022:ocac070. doi:10.1093/jamia/ocac070

[79] Neff G, Tanweer A, Fiore-Gartland B, Osburn L. Critique and Contribute: A Practice-Based Framework for Improving Critical Data Studies and Data Science. *Big Data.* 2017;5(2):85-97. doi:10.1089/big.2016.0050

[80] Cury M, Whitworth E, Barfort S, et al. Hybrid Methodology: Combining Ethnography, Cognitive Science, and Machine Learning to Inform the Development of Context-Aware Personal Computing and Assistive Technology. *Ethnographic Praxis in Industry Conference Proceedings.* 2019;2019(1):254-281. doi:10.1111/1559-8918.2019.01284