Building Equitable Artificial Intelligence in Health Care

Addressing Current Challenges and Exploring Future Opportunities

Anna Zink
UNIVERSITY OF CHICAGO BOOTH SCHOOL OF BUSINESS, CENTER FOR APPLIED AI

Sarah Morriss
URBAN INSTITUTE

Anuj Gangopadhyaya
LOYOLA UNIVERSITY CHICAGO

with Ziad Obermeyer
UNIVERSITY OF CALIFORNIA, BERKELEY

September 2023

AI applications in health care are prone to biases that could perpetuate health disparities. In this paper we study the ways in which AI may maintain, perpetuate, or worsen inequitable outcomes in health care. We review current approaches to evaluating and mitigating biased AI and potential applications of AI to address health equities. Finally, we discuss current incentives for equitable AI and potential changes in the regulation and policy space. As AI becomes increasingly embedded in the daily operations of health care systems, it is imperative that we understand its risks and evaluate its impact on health equity.

Introduction

Data has always been central to medicine, and practitioners have been using prediction models since before the 1990s (Gail et al. 1989; Kononenko 2001). However, the digitization of health records and claims data in the early 2000s greatly increased access to and use of health care data, spurring new investment in prediction tools for the health care industry. Artificial Intelligence (AI) algorithms, which

include the subfields of machine learning and deep learning, use large datasets to make predictions, and have been developed for insurers, hospitals, and physician groups to assist with decisions about patient care, resource staffing, diagnosing conditions, and more. While the authors were unable to find precise evidence on the full extent of AI diffusion in the health care sector today, we know that the FDA has approved over 500 AI-enabled devices as of 2023, and that AI models for predicting medication adherence, disease onset, and hospitalizations have been developed for insurers and population health management groups (Gervassi et al. 2022).

Unequal Treatment at 20

This work is part of a series of publications that commemorates the 20th anniversary of the 2003 Institute of Medicine report, Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. This report found that people of color received lower-quality health care than white patients, even when access-related factors were held constant. Two decades later, we still observe the same inequities, which has motivated thought leaders to imagine how to redesign the health care system so it works equitably.

The Institute of Medicine published *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care* in 2003 (Institute of Medicine 2003). The report contains no mention of AI or machine learning. It was not that AI did not exist at that time—IBM's AI algorithm Deep Blue beat the best chess player in the world that same year—but the authors and many others at the time were unaware of the role it would play in health care 20 years later. It was not until 2016, when ProPublica published an article showing that a common risk assessment tool used in bail decisions was more likely to wrongly flag Black defendants as high risk for reoffending, that the research community recognized the immediacy of the issue (O'Neil 2016).[1] Since then, there have been an overwhelming number of examples across health care settings and clinical areas warning that AI could perpetuate or widen disparities in health (AHRQ 2022). With the advent of ChatGPT and other generative AI models poised to become embedded in company and individual workflows, evaluating the impact of AI on health equity is as pressing as ever.[2]

In this paper, we explain the relationship between AI and health equity—in particular, the ways in which AI may maintain, perpetuate, or worsen inequitable outcomes in health. We discuss current approaches to understanding and addressing this issue. Our focus is on AI applications in health care; however, many of the concepts and solutions presented are applicable to a broader set of algorithms used in health care, including rule-based clinical decision support systems, which we also discuss. We further review examples of AI applications that could be built to address health equity. Finally, we discuss existing and possible incentives for equitable AI.

BOX 1

Definitions

Artificial Intelligence

The combination of computer science and robust datasets (e.g., structured data, images, text) to enable problem-solving by developing algorithms which seek to create expert systems that make predictions or classifications based on input data. In health care, AI is often used to analyze medical images or free text to automate clinical tasks such as diagnosis, triage, note-taking, and even communication.

Machine Learning

A branch of AI and computer science that focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. In health care, machine learning is often used to predict the likelihood of a health event or condition.

Sources: "What is Artificial Intelligence (AI)?" IBM. Accessed July 27, 2023. https://www.ibm.com/topics/artificial-intelligence; "What is Machine Learning?" IBM. Accessed July 27, 2023. https://www.ibm.com/topics/machine-learning?lnk=fle.

AI and Health Equity

AI has a wide range of applications in health care, including medical image analysis, virtual assistants, clinical decision support, predictive analytics, remote patient-monitoring, and health communication. Even AI applications that don't directly impact medical decisionmaking can affect health care use. For example, AI algorithms are used to optimize appointment schedules in doctors' offices. The potential impact of these applications on health outcomes is significant: a patient who never gets to the doctor's office loses an opportunity for diagnosis and treatment. In the office, AI-generated information could change a doctor's decision about a course of treatment, influence which patients an insurer approves for additional services or expensive treatments, or affect appointment wait times.

Given AI's potential influence on patient care, it is important to evaluate the quality of predictions generated by AI. We consider two common problem areas in AI applications that could perpetuate or worsen heath disparities: (1) failure to generalize and (2) incorrect or inadequate prediction targets. We highlight, using several examples, the consequences these errors have for health equity. We define health equity as the state in which everyone has the opportunity to attain their best possible health.[3] Equitable AI, therefore, is AI that moves us toward, rather than away from, that state, either by incorporating health equity principles when building and deploying AI applications, or by developing AI applications that target health equity directly.

When Predictions Don't Generalize

AI algorithms learn from existing data in order to make new predictions. In health care, many AI algorithms are trained on convenient samples of data collected for purposes outside of the prediction problem. When AI algorithms make predictions on populations that differ from the training data, the

predictive performance of the AI algorithm can decrease. Take, for example, AI for melanoma diagnosis: on its surface, this is an area well suited for AI, since pictures of melanoma can be analyzed for cancer risk. Many researchers rely on large skin-image repositories to train their AI algorithms; however, some of these databases primarily collect images from fair-skinned populations in the United States, Europe, and Australia. Research has shown that AI trained on these data performed worse on dark-skinned populations (Adamson and Smith 2018). Including more training data for darker-skinned populations improved the prediction quality for dark-skinned populations. However, generalizability is not just a question of assessing who is in the dataset, but of understanding how the AI application will be used, and whether predictions built using the training data will translate to these other settings. For example, skin-image data used to train AI predictions may be collected in dermatologists' offices, among patients who have access to this type of specialist care, but the AI application could be deployed widely (on, for example, a smartphone) to users who do not have similar access to specialty care and for whom the training dataset may not generalize.

The inability to generalize predictions often results in exacerbating health inequities, since the populations excluded or underrepresented in training data are those that have been historically excluded or disadvantaged. The lack of diversity in clinical trials is well documented, and many historical medical studies that inform clinical practice today predominately included white men (Dresser 1992). Furthermore, because of extensive barriers to accessing health care data, AI researchers often train AI models on a select few public databases (Johnson et al. 2016). These data are often collecting from single sites, and AI algorithms are at risk of reduced predictive performance when used at other sites (Röösli, Bozkurt, and Hernandez-Boussard 2022; Song et al. 2020). Even large national datasets like health claims data only collect data on beneficiaries *using* the health care system, meaning that populations without access, or with less access, are likely underrepresented in data compared to other groups.

While data being representative is important to improve the predictive performance of AI algorithms, it could also help increase clinician and patient trust in AI tools (Bibbins-Domingo Helman, and Dzau 2022; Schwartz et al. 2023). Research has found that diversity in clinical trials increased physician willingness to prescribe drugs, and patient trust in their efficacy, which could reduce disparities in prescribing rates and increase trust in the medical system (Alsan et al. 2022). A parallel argument could be made for AI: diversity in AI training data is not just a means to improve predictive performance but a necessary step to increase trust in AI.

Incorrect or Inadequate Prediction Targets

Prediction targets are the measures AI algorithms are trained to predict. Common prediction targets include the presence of a condition and the occurrence of an adverse event like rehospitalization. The use of mismeasured or proxy outcomes in AI can reinforce disparities in care if those outcomes are themselves influenced by factors driven by structural racism and inequities (Mullainathan and Obermeyer 2021).

Take, for example, the common practice of using health insurance claims data to measure the presence or absence of a condition like diabetes or heart disease. An individual is marked as having the condition if they have a claim with diagnosis codes for that condition within a specified time window. This means that observing heart disease or diabetes is conditional on the patient visiting the doctor and the doctor recording the diagnosis codes on the claims. Patients with less access to care or facing provider bias will be, on average, underdiagnosed.

There are many AI applications that use this type of data to predict the presence of a condition. For example, several AI-based chest X-ray prediction models use public radiology datasets with recorded diagnosis of a condition or event. One study evaluated these AI algorithms and found that female patients, Black patients, Hispanic patients, and patients with lower socioeconomic status (with Medicaid insurance) were more likely to be incorrectly predicted as not having a condition or event by the AI algorithm compared to other patients (Seyyed-Kalantari et al. 2021). Because underserved patients don't go to the doctor's office, they are less likely to have a chronic condition recorded on a health care claim and are therefore not identified in AI algorithms using this as a prediction target. The AI algorithm predicted the presence of the chronic condition on the claim, not the existence of the condition itself.

In the previous example, the correct prediction target had been selected—the occurrence of a condition—but it was mismeasured in the data used to train the AI algorithm. In other problems, the prediction target is not measurable. In such situations, one commonly selects a "proxy" target in place of the actual target of interest. Proxy targets are measured variables correlated with the target of interest. The use of proxy variables is very common in health care as well as in other fields, such as employment and housing, where targets like "ability" or "good tenant" are not measurable.

As an example, population health management models are frequently used by private health insurers to allocate care management services. The goal of these models is to predict patients who would benefit from more care management services, but because improvement in health is difficult to measure, health insurers use a proxy outcome. A common proxy outcome is annual health care costs. One study found that a population health management algorithm proxying health care need with costs allocated more care to white patients than to Black patients conditional on health needs (Obermeyer et al. 2019). Because the proxy target was correlated with access to and use of health care services, it identified frequent users of health care services, who were less likely to be Black patients given current inequities in health care access (Manuel 2018).

Similar issues occur for predictors (the variables used to predict the target), which might cause some variables to have differential power for different groups of patients. For example, family health history is a common risk factor for many cancers. As a result, it is used to determine the timing and frequency of preventive services and as an input in many risk models. However, family history has been found to be more likely mismeasured for Black patients relative to white patients (Chavez-Yenter et al. 2022; Andoh 2023). Therefore, predictions relying on family history will not work as well for Black patients, unless other risk predictors are able to make up for this loss of information.

Evaluation

For many, AI represents a black box that produces remarkable and sometimes frightening results. The challenges in understanding what AI is and how it works should not prevent physicians, clinicians, health systems, researchers, patients, and other stakeholders from evaluating it with the same rigor and precision as we do any other treatment, software, or tool in medicine.

How do we determine whether AI is equitable or not? Ideally, we would estimate the causal impact of the introduction of an AI algorithm on equity (Kasy and Abebe 2021). Identifying the causal impact would require either a randomized controlled trial or quasi-experimental methods and data on the intervention. We would need data on those treated (i.e., units exposed to AI), those not treated (i.e., units not exposed to AI), group identifiers, empirical measures of equity, and any confounders (Groos et al. 2018). These analyses require large investments to collect and measure data and long study time horizons to observe outcomes. There has been little empirical work in this vein, likely due to financial and operational constraints as well as challenges regarding data access.

Instead, research has focused on evaluations using more readily available data; typically, data collected predeployment or at the point of deployment. These analyses primarily focus on (1) the difference in predictive performance of an AI algorithm for different groups, and/or (2) the difference in AI-suggested allocation of care or services for different groups. Differences in performance or allocation across groups are quantitatively assessed using a set of measures, commonly referred to as fairness measures (Barocas, Hardt, and Narayanan 2019; Verma and Rubin 2018). Fairness measures often conflict with each other. For example, in the study on the population health management risk predictor (Obermeyer et al. 2019), the algorithm was very good at predicting health care costs, and fairness measures based on predictive performance would have found the algorithm fair: it was able to predict health care costs equally well for both Black and white patients. It wasn't until the researchers considered *how* services were allocated that they noticed that more white patients were being referred than Black patients given the same measured level of sickness. Because the prediction target was biased, fairness measures based on predictive performance were unable to detect any problems. Which fairness measures to prioritize will depend on the clinical settings and goals of the developer and/or policymaker and, in general, it is best to use a suite of measures.[4]

The focus on measurement in this space may feel familiar to the field of health care quality measurement.[5] Health care quality, like fairness, is hard to measure and define. To try to piece together a picture of health care quality more broadly, many different quality measures are used to capture not only what a provider does to maintain or improve health but also health outcomes themselves. Health care quality measures have become a central component of alternative payment models and other policy initiatives to date. They are viewed as a necessary but imperfect means to assess quality: measures are correlated with health care quality, but they can induce a "teach to the test" mentality, and it is hard to attribute quality differences to organizations rather than to differences in the health of the patients they serve. We can view fairness measures for AI with a similar lens. Measurement creates a clearer picture of equitable AI, but it would be misguided to over-rely on select metrics and lose sight

of the outcome we care about most: more equitable health. It is also difficult to estimate the effect of AI on health disparities through observation alone.

Defining Groups

A pressing challenge in the evaluation of AI is the lack of group data needed to evaluate whether an AI algorithm is equitable or not (Lu et al. 2022). These data are used to define and identify groups that experience systemic discrimination in health care. Race/ethnicity/language data (commonly referred to as REL data) and other group identifiers are not always collected or not collected with the same quality as other variables in health care data.[6] Without a means to measure groups, there is no way to evaluate the impact of AI on that group. What's more, there is the question of how to define groups. The federal government as well as many states have put forth new standards on the measurement and definition of REL data (SHADAC 2022).[7] Many are calling for more granular race and ethnicity categories, since coarse racial-ethnic group definitions commonly used in research fail to measure differences within groups (Movva et al. 2023). Measures of intersectionality or finer groups can provide more information but also raise statistical challenges (Ghasemi et al. 2021; Spielman, Folch, and Nagle 2014). In general, efforts to disaggregate race and ethnicity data have been met with support and a necessary amount of caution.[8]

Building More-Equitable AI

There are numerous resources available to help incorporate health equity principles into the AI workflow (Chen et al. 2021; Diakopoulos et al. n.d.; Nelson et al. 2020; Obermeyer et al. 2021; Rajkomar et al. 2018; Wang et al. 2022). One example is the bias evaluation checklist developed by Wang and colleagues that helps one classify the risk (low, medium, high) of specific sources of bias that might occur in the process of developing and deploying an algorithm (Wang et al. 2022). To best anticipate potential risks, it is important to have contextual knowledge about inequities that exist in the given prediction setting. There are also a number of technical solutions that have been proposed to reduce issues of bias (Huang et al. 2022). These solutions can be bucketed into categories based on where they are addressed in AI development (i.e., problem selection, data collection, defining the prediction target, algorithm development, and post-deployment). Most of the solutions that have been presented to date focus on outcome definition and algorithmic development, in part because these solutions can be addressed immediately at the point of AI development. Other solutions, such as collecting more—and better—data require time and financial investments. More evidence on how these strategies impact the design of AI and improve health equity is needed.

Governance

Operationalizing algorithmic bias solutions requires coordination across multiple stages of development and the work of many teams. What's more, evaluation and monitoring practices are ongoing: data shift, model degradation, and changes in user behavior can affect the fidelity of AI output over time (Subbaswamy and Saria 2020). Such efforts require organizational oversight and buy-in

(McCradden et al. 2020). Researchers have published internal auditing frameworks for organizations to use (Raji et al. 2020). For hospitals that already have governance structures in place, expanding processes to incorporate AI applications might be a natural extension: pioneers in this space have shared plans for AI oversight (Bedoya et al. 2022). The ability and interest of organizations to establish such processes remains to be seen.

Can AI Self-Correct?

Could health care AI be trained to recognize its own inequities and self-correct? AI alignment is a subfield in AI research that addresses whether and how to incorporate human preferences and ethical goals in the development of AI. Ethical preferences must be computed empirically and incorporated into AI algorithm objectives. However, we've discussed how fairness measures that encode these values can contradict each other. While a monitoring system could be set up to warn of changes in metrics or data shift, a human is still required to specify priorities. Priorities could be set by AI developers or regulated by policy, but in the end, human intervention will always be needed to imbue AI with equitable principles.

Race as a Predictor

In prediction problems, one is typically taught to include any variable that has predictive power. Race and ethnicity are commonly correlated with clinical outcomes and can be accurately predicted from most datasets. Therefore, they have historically been included as clinical risk predictors. However, a re-examination is underway regarding the use of patient racial and ethnic information in AI and other algorithms, out of concern that their inclusion in risk prediction models might increase health disparities (Vyas, Eisenstein, and Jones 2020).

The interpretation of race as a variable in a prediction model is often not clearly articulated. This has led to harmful misinterpretations of race differences as biological in origin in many medical models (Cerdeña, Plaisime, and Tsai 2020). As an alternative to these "race-based" medical models, race-conscious medicine explicitly defines race as a sociological and power construct (Cerdeña, Plaisime, and Tsai 2020). This focuses research on the effects of structural racism and reduces health inequities. Differences in predicted risk by race group reflect health disparities resulting from the deleterious effect of racism on health (Bailey et al. 2017; Borrell et al. 2021). Structural racism has been shown to be a key determinant of health disparities; for example, segregated housing has led to a higher risk of asthma (Bailey et al. 2017). Race, then, is a proxy for health disparities and the effects of structural racism.

If this is the case, then we could try to find better predictors than race to account for disparities in health due to racism and other social determinants of health. As an example, consider the recent update to calculating estimated glomerular filtration rate (eGFR) used to assess kidney functioning. Previously, the algorithm used a race-based adjustment to inflate eGFR estimates for Black patients; however, the removal of race from eGFR calculations underestimated eGFR in Black patients. Neither of these solutions is desired: one leads to potential underuse of treatment and the other to overuse. However,

researchers found that including cystatin C in the estimation of GFR removed differences in predictive accuracy by race (Williams et al. 2021). In other settings, the physiological risk factors or other clinical measures that explain risk differences between race groups might not always be identified or available for prediction, in which case researchers could try to measure the source of the health disparity directly: for example, if researchers believe that unobserved differences in health risks are a result of structural racism, then this could be incorporated into the risk predictor. How to measure structural racism in health care data is an open question that a number of researchers are actively working on (Groos et al. 2018; Hardeman et al. 2022).

When new predictor variables explain risk differences, race is no longer predictive and drops out of the AI algorithm on its own accord. But in the situation where new variables are not available or measured, race can still be an effective means to measure differences in risk for patients and may be the only way to account for health disparities in prediction models (Manski 2022). Thus, race as a predictor can help predict key outcomes of interest and point researchers in productive directions to understand and measure the source of disparities.

Research Applications

How can AI be harnessed to improve health equity? AI has already played a positive role in promoting equity by providing researchers with new tools to explore and address biases that exist in the health care system today (Chen, Joshi, and Ghassemi 2020). For example, AI-built techniques to analyze image data allowed researchers to discover that current methods for diagnosing knee pain in MRIs failed to identify knee pain experienced by many Black patients (Pierson et al. 2021). This study identified differences in patient-reported knee pain versus clinician diagnosis of knee pain, and attributed a portion of these differences to information in the knee X-ray that was not considered by doctors during diagnosis. This research could impact clinical practice (i.e., how doctors read knee X-rays) and the set of patients identified with osteoporosis, which would have implications on health outcomes if it affected patient treatment decisions. Using AI to discover new information in medical image data might change the way these data are interpreted by medical professionals.

This research suggests that there is more that could be learned from medical image data than what doctors are trained to look for. What's more, AI can be used to tap into these data, which might be less biased than traditional data sources that reflect inequities in our health care system. For example, what if, instead of using claims data to identify levels of sickness, we used image and lab data while correcting for racial differences in access to these services? These data have proven to be very predictive of health events and conditions and might provide a truer signal to what we want to measure: health. For instance, ECGs are commonly used to predict heart failure. What other conditions might they also be able to predict? Would it be possible to use the ECG as input to a predictor instead of documented conditions, which might be poorly or unevenly documented in claims data?

Image and lab data offer one opportunity for new, potentially less biased sources of data, but AI could also be used to collect data that is uncorrelated with access to health care settings. There are

many examples of applications like wearable or at-home monitoring devices that collect health data outside of the hospital or doctor's office. Patient access to these devices might be limited, but 97 percent of Americans own a cell phone, and 93 percent use the internet.[9] One study showed that search data could be used to understand health information needs (Abebe et al. 2019). Patient advocacy forums and other online spaces could provide data on conditions, including common symptoms, that are not documented in the claims data today. These data could be vulnerable to their own biases and should be thoroughly evaluated before use, in addition to requiring data privacy safeguards.[10]

AI can not only be used to analyze and collect new data but also to remove biases in existing data. Society has long recognized that humans make biased decisions. One solution has been to blind decisionmakers from the sources of information we don't want them to use. For example, we blind journal reviewers from authors' identities and orchestra judges from candidates' looks. AI can be used as a tool for similar purposes. For example, AI methods have been developed to remove gender and racial differences in the documentation of clinical notes before they are used for prediction problems (Zhang et al. 2020). Other researchers found that AI algorithms could detect which lab site had produced images for biopsy (Howard et al. 2021). While this information might be valuable in some settings, for certain prediction problems it should be removed. These methods do not address biases that arise from differences in access to care—and therefore the presence or absence of a clinical note or a lab test—but focus on removing the predictive link between the data and a sensitive attribute such as race or gender conditional on the data having been collected.

The use of AI for communication is another robust area of research (Butow and Hoque 2020). The ability of AI to interpret and analyze sound and text provides unique opportunities to recognize and identify cultural subtleties and nuances in communication with patients. Culture is a key factor in successful health communication for both medical decisionmaking and health promotion and has been identified as one strategy for reducing health disparities (Betsch et al. 2016; Kreuter and McClure 2004). Leveraging AI to understand and therefore improve health communication could help patients become more active participants in their health and health care decisionmaking.

Finally, AI can create data to explain the decisions of AI and humans alike (Singla et al. 2020). For example, by studying how AI-generated perturbations of data impact prediction results, we can learn more about what predictors are influencing the result (Chen, Joshi, and Ghassemi 2020). AI can also help form new hypotheses about human behavior (Ludwig and Mullainathan 2023). For example, researchers determined what specific features of a face affected bail decisions using AI-generated faces. These data hold facial features constant (e.g., skin color, gender representation, and aging) while varying one specific feature (e.g., thinness of face) to test for the effect of that feature on bail decisions.[11] Analogous studies could be done in health care (Miller et al. 2019). Understanding the source of potential biases can help identify biased decisionmaking.

Targeting Adoption

Research shows that 20 percent of doctors in the US treat 80 percent of Black patients and that these doctors are more resource-constrained (Bach et al. 2004). We also know that there is high variation in

diagnostic skill and that mistakes are more likely to occur among low-skilled, low-resourced doctors. Thus, moving low-skilled doctors to higher skill levels presents an opportunity to improve the average quality of care received among Black patients. AI offers a chance to close that gap (Chan, Gentzkow, and Yu 2019). As an example, research has found that AI can be used to help physicians diagnose heart attack (Mullainathan and Obermeyer 2022). AI can not only improve skill level within specialties, but also provide specialist tools for primary care doctors. For example, the AI-enabled medical software IDx-DR helps primary care doctors identify patients at risk of eye retinopathy, a condition typically diagnosed only by eye care professionals (Grzybowski and Brona 2021). Medicare and other private insurers have started paying for IDx-DR and other similar types of AI products. Reimbursement will likely improve adoption, but to ensure the biggest impact, adoption could be incentivized in areas where it would be most beneficial, through grant funding or increased reimbursement rates in geographic areas that lack specialists or have low-quality care.

Incentives for Equitable AI

Twenty years ago, the Institute of Medicine documented many examples of racial and ethnic disparities in *Unequal Treatment*. Knowledge of disparity and the moral imperative for equity has not, to date, spurred noteworthy changes in the trajectories of these outcomes: we continue to observe large racial and ethnic disparities in access to care, access to quality care, and health outcomes. Why then should we expect that the knowledge of inequitable AI alone would spur action?

What incentives exist to date for equitable AI? Equitable AI solutions take time and expertise: one study team found that an algorithmic audit on two hospital-based care models took 115 person-hours and 8–10 months to complete (Lu et al. 2022). Thus, from a business standpoint, the decision to invest in equitable AI solutions will depend on the private returns of equitable AI. This includes responding to changes in consumer sentiment and demand for equitable care plus reducing exposure to systematic and costly patient grievances cases. Entering 2023, hospitals faced higher expenses, depressed finances, and negative margins (Kaufman Hall 2022). Health insurers appear to be in slightly better financial positions. It is hard to predict how much investment will be made in this area without external pressure.

If there isn't a current business incentive for equitable AI, then government action might serve to reduce the costs and/or increase the benefits of equitable AI. Recent announcements of public and private funding will likely help incentivize more research and work in this area. The White House recently announced that it would initiate new investments to fund responsible AI research and development. This includes $140 million in funding for the National Science Foundation to launch several institutes focused on public assessments of existing generative AI systems.[12] Private organizations are also providing funding to evaluate AI used in diagnostic decisionmaking.[13]

Regulating Equitable AI

There are several regulatory options available to federal agencies to promote equity. As an example, let's consider approaches that the Centers for Medicare & Medicaid Services (CMS) could employ.

Under the Affordable Care Act, CMS has experimented with offering rewards and/or levying penalties to compel health systems to provide better-quality care. Health care quality is typically represented through measures that feed into public-facing scorecards or are used by CMS to set penalties and bonuses. These measures have not historically included social drivers of health, despite their huge importance for health outcomes (Hood et al. 2016). However, several new measures have been proposed to assess the social determinants of health in order to allocate resources to improve health equity.[14] CMS could also create a measure of equitable AI to incentivize the adoption of equitable AI practices directly. This measure could be incorporated into public reporting programs such as the Hospital Inpatient Quality Reporting Program so that patients might use information on whether organizations are adopting equitable AI practices to exert demand-side pressure. The measure could also be incorporated into bonus and penalty formulas to incentivize organizations to invest in better AI processes. While these approaches focus on solutions available to CMS and, therefore, only apply directly to the Medicare and Medicaid programs, they could still affect a wide range of organizations: the majority of Medicaid beneficiaries are on privately run managed care organizations, and soon the majority of Medicare beneficiaries are expected to be enrolled in Medicare Advantage plans (MedPAC 2023).[15] Furthermore, almost every hospital treats Medicare and Medicaid patients and would be affected by these policies.

Various agencies and branches of the federal government have indicated that regulating AI is a top concern for them. In 2022, President Joe Biden released a blueprint for an AI Bill of Rights outlining five principles to guide the design, use, and deployment of AI to protect people against its harms.[16] The Food and Drug Administration has published a beta version of how it plans to regulate AI used in health care (FDA 2021). Other organizations such as the Center for Medicare and Medicaid Innovation are considering how to incorporate these principles into pilot programs.[17] The Department of Health and Human Services has issued a notice of proposed rulemaking to revise Section 1557 of the Patient Protection and Affordable Care Act. This revision, if finalized, would explicitly prohibit discrimination in the use of clinical algorithms to support decisionmaking in covered entities.[18] State legislatures have also shown interest in this issue: in August 2022, the California Attorney General issued a letter to all hospitals requesting that they share how they are identifying and addressing racial and ethnic disparities in commercial decisionmaking tools.[19] It is still an open question how federal and local agencies will regulate algorithms moving forward.

Conclusion

AI is an increasingly important part of health care decisionmaking. It is therefore vital that AI applications are developed and monitored with health equity in mind. While there has been much work in this area to date, from research on identifying and improving algorithmic bias to resources for building more equitable AI, to the development of incentives for equitable AI—be that regulation or funding—there is still much to be done. We need more research measuring the impact of AI on health equity. This includes quasi-experimental methods and randomized controlled trials as well as qualitative studies engaging affected communities such as doctors and patients to understand the specific

mechanisms by which AI systems affect them. We should also be carefully monitoring adoption patterns of AI use among health care professionals and patients to understand whether AI's benefits are fairly distributed. Furthermore, we should not only be focused on the evaluation of AI applications that already exist, but also think critically about the set of problems we are trying to solve with AI: can we focus on applications that push us toward health equity rather than just ensuring new applications don't create further harm? Finally, while this article focuses on solutions to improving AI, it is imperative that we do not lose focus on addressing the underlying issues at the root of this discussion: the pervasive health inequities that exist in this country as a result of structural racism and discrimination.

Notes

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias," ProPublica, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[2] Benji Edwards, "GPT-4 Will Hunt for Trends in Medical Records Thanks to Microsoft and Epic," Ars Technica, April 18, 2023, https://arstechnica.com/information-technology/2023/04/gpt-4-will-hunt-for-trends-in-medical-records-thanks-to-microsoft-and-epic/.

[3] P. Braveman, E. Arkin, T. Orleans, D. Proctor, and A. Plough, "What Is Health Equity?," Robert Wood Johnson Foundation, May 1, 2017, https://www.rwjf.org/en/insights/our-research/2017/05/what-is-health-equity-.html.

[4] Deborah Hellman, "Measuring Algorithmic Fairness," *Virginia Law Review* 106 (4), June 1, 2020, https://virginialawreview.org/articles/measuring-algorithmic-fairness/.

[5] "Types of Health Care Quality Measures," Agency for Healthcare Research and Quality (AHRQ), last reviewed July 2015, https://www.ahrq.gov/talkingquality/measures/types.html.

[6] Neil P. Rowen, Brianna Van Stekelenburg, Raman Nohria, Robert S. Saunders, and Rebecca G. Whitaker, "How to Improve Race, Ethnicity, and Language Data and Disparities Interventions," *Health Affairs Forefront* (blog), Health Affairs, September 14, 2022, https://www.healthaffairs.org/content/forefront/improve-race-ethnicity-and-language-data-and-disparities-interventions.

[7] Office of Management and Budget (OMB), "Initial Proposals for Updating OMB's Race and Ethnicity Statistical Standards," *Federal Register* (blog), National Archives, January 27, 2023, https://www.federalregister.gov/documents/2023/01/27/2023-01635/initial-proposals-for-updating-ombs-race-and-ethnicity-statistical-standards.

[8] Farah Kader, Lan N. Đoàn, Matthew Lee, Matthew K. Chin, Simona C. Kwon, and Stella S. Yi, "Disaggregating Race/Ethnicity Data Categories: Criticisms, Dangers, and Opposing Viewpoints," *Health Affairs Forefront* (blog), Health Affairs, March 25, 2022, https://www.healthaffairs.org/content/forefront/disaggregating-race-ethnicity-data-categories-criticisms-dangers-and-opposing.

[9] "Mobile Fact Sheet," Pew Research Center, April 7, 2021, https://www.pewresearch.org/internet/fact-sheet/mobile/; "Internet/Broadband Fact Sheet," Pew Research Center, April 7, 2021, https://www.pewresearch.org/internet/fact-sheet/internet-broadband/.

[10] Emma Woollacott, "Apple Sued over 'Racial Bias' of Apple Watch," *Forbes*, December 29, 2022, https://www.forbes.com/sites/emmawoollacott/2022/12/29/apple-sued-over-racial-bias-of-apple-watch/.

[11] "The Face Effect: When a Face Is a Felony," Center for Applied Artificial Intelligence, accessed April 27, 2023, https://faceeffect.ai/.

[12] "FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation that Protects Americans' Rights and Safety," The White House, May 4, 2023, https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/.

[13] "Augmented Intelligence in Medicine and Healthcare Initiative (AIM-HI)," Kaiser Permanente—Division of Research, 2023, https://divisionofresearch.kaiserpermanente.org:443/projects/aim-hi?kp_shortcut_referrer=kp.org/aim-hi.

[14] Debbie Chang and Rachel Nuzum, "Now Is the Time for Measuring Social Drivers of Health in Medicare, Medicaid, and the Children's Health Insurance Program," *To the Point* (blog), The Commonwealth Fund, March 28, 2022, https://www.commonwealthfund.org/blog/2022/now-time-measuring-social-drivers-health-medicare-medicaid-and-childrens-health-insurance.

[15] "Share of Medicaid Population Covered under Different Delivery Systems," State Health Facts, KFF, as of July 1, 2022, https://www.kff.org/medicaid/state-indicator/share-of-medicaid-population-covered-under-different-delivery-systems/.

[16] "Blueprint for an AI Bill of Rights," Office of Science and Technology Policy, The White House, accessed January 16, 2023, https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

[17] Melissa Majerol and Dora Lynn Hughes, "CMS Innovation Center Tackles Implicit Bias," *Health Affairs Forefront* (blog), Health Affairs, July 5, 2022, https://www.healthaffairs.org/content/forefront/cms-innovation-center-tackles-implicit-bias.

[18] "Section 1557 of the Patient Protection and Affordable Care Act," US Department of Health and Human Services, last reviewed February 3, 2023, https://www.hhs.gov/civil-rights/for-individuals/section-1557/index.html.

[19] "Attorney General Bonta Launches Inquiry into Racial and Ethnic Bias in Healthcare Algorithms," Office of the Attorney General, State of California Department of Justice, August 31, 2022, https://oag.ca.gov/news/press-releases/attorney-general-bonta-launches-inquiry-racial-and-ethnic-bias-healthcare.

References

Abebe, Rediet, Shawndra Hill, Jennifer Wortman Vaughan, Peter M. Small, and H. Andrew Schwartz. 2018. "Using Search Queries to Understand Health Information Needs in Africa." arXiv doi:10.48550/arXiv.1806.05740.

Adamson, Adewole S., and Avery Smith. 2018. "Machine Learning and Health Care Disparities in Dermatology." *JAMA Dermatology* 154 (11): 1247–48. doi:10.1001/jamadermatol.2018.2348.

AHRQ (Agency for Healthcare Research and Quality). 2022. *Impact of Healthcare Algorithms on Racial and Ethnic Disparities in Health and Healthcare*. Rockville, MD: AHQR. https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/protocol.

Alsan, Marcella, Maya Durvasula, Harsh Gupta, Joshua Schwartzstein, and Heidi L. Williams. 2022. "Representation and Extrapolation: Evidence from Clinical Trials." Working Paper 30575. Cambridge, MA: National Bureau of Economic Research. doi:10.3386/w30575.

Andoh, Joana E. 2023. "The Stories We Don't Know." *JAMA* 329 (18): 1551. doi:10.1001/jama.2023.5891

Bach, Peter B., Hoangmai H. Pham, Deborah Schrag, Ramsey C. Tate, and J. Lee Hargraves. 2004. "Primary Care Physicians Who Treat Blacks and Whites." *New England Journal of Medicine* 351 (6): 575–84. doi:10.1056/NEJMsa040609.

Bailey, Zinzi D., Nancy Krieger, Madina Agénor, Jasmine Graves, Natalia Linos, and Mary T. Bassett. 2017. "Structural Racism and Health Inequities in the USA: Evidence and Interventions." *Lancet* 389 (10077): 1453–63. doi:10.1016/S0140-6736(17)30569-X.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. https://fairmlbook.org/.

Bedoya, Armando D., Nicoleta J. Economou-Zavlanos, Benjamin A. Goldstein, Allison Young, J. Eric Jelovsek, Cara O'Brien, Amanda B. Parrish, Scott Elengold, Kay Lytle, Suresh Balu, Erich Huang, Eric G. Poon, and Michael J. Pencina. 2022. "A Framework for the Oversight and Local Deployment of Safe and High-Quality Prediction Models." *Journal of the American Medical Informatics Association* 29 (9): 1631–36. doi:10.1093/jamia/ocac078.

Betsch, Cornelia, Robert Böhm, Collins O. Airhihenbuwa, Robb Butler, Gretchen B. Chapman, Niels Haase, Benedikt Herrmann, Tasuku Igarashi, Shinobu Kitayama, Lars Korn, Ülla-Karin Nurm, Bernd Rohrmann, Alexander J. Rothman, Sharon Shavitt, John A. Updegraff, and Ayse K. Uskul. 2016. "Improving Medical Decision Making and Health Promotion through Culture-Sensitive Health Communication: An Agenda for Science and Practice." *Medical Decision Making* 36 (7): 811–33. doi:10.1177/0272989X15600434.

Bibbins-Domingo, Kirsten, Alex Helman, and Victor J. Dzau. 2022. "The Imperative for Diversity and Inclusion in Clinical Trials and Health Research Participation." *JAMA* 327 (23): 2283–84. doi:10.1001/jama.2022.9083.

Borrell, Luisa N., Jennifer R. Elhawary, Elena Fuentes-Afflick, Jonathan Witonsky, Nirav Bhakta, Alan H. B. Wu, Kirsten Bibbins-Domingo, José R. Rodríguez-Santana, Michael A. Lenoir, James R. Gavin, III, Rick A. Kittles, Noah A. Zaitlen, David S. Wilkes, Neil R. Powe, Elad Ziv, and Esteban G. Burchard. 2021. "Race and Genetic Ancestry in Medicine—A Time for Reckoning with Racism." *New England Journal of Medicine* 384 (5): 474–80. doi:10.1056/NEJMms2029562.

Butow, Phyllis, and Ehsan Hoque. 2020. "Using Artificial Intelligence to Analyse and Teach Communication in Healthcare." *Breast* 50:49–55. doi:10.1016/j.breast.2020.01.008.

Cerdeña, Jessica P., Marie V. Plaisime, and Jennifer Tsai. 2020. "From Race-Based to Race-Conscious Medicine: How Anti-Racist Uprisings Call Us to Act." *Lancet* 396 (10257): 1125–28. doi:10.1016/S0140-6736(20)32076-6.

Chan Jr, David C., Matthew Gentzkow, and Chuan Yu. 2019. "Selection with Variation in Diagnostic Skill: Evidence from Radiologists." Working Paper 26467. Cambridge, MA: National Bureau of Economic Research. doi:10.3386/w26467.

Chavez-Yenter, Daniel, Melody S. Goodman, Yuyu Chen, Xiangying Chu, Richard L. Bradshaw, Rachelle Lorenz Chambers, Priscilla A. Chan, Brianne M. Daly, Michael Flynn, Amanda Gammon, Rachel Hess, Cecelia Kessler, Wendy K. Kohlmann, Devin M. Mann, Rachel Monahan, Sara Peel, Kensaku Kawamoto, Guilherme Del Fiol, Meenakshi Sigireddi, Saundra S. Buys, Ophira Ginsburg, and Kimberly A. Kaphingst. 2022. "Association of Disparities in Family History and Family Cancer History in the Electronic Health Record with Sex, Race, Hispanic or Latino Ethnicity, and Language Preference in 2 Large US Health Care Systems." *JAMA Network Open* 5 (10): e2234574. doi:10.1001/jamanetworkopen.2022.34574.

Chen, Irene Y., Shalmali Joshi, and Marzyeh Ghassemi. 2020. "Treating Health Disparities with Artificial Intelligence." *Nature Medicine* 26:16–17. doi:10.1038/s41591-019-0649-2.

Chen, Irene Y., Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2021. "Ethical Machine Learning in Healthcare." *Annual Review of Biomedical Data Science* 4:123–44. doi:10.1146/annurev-biodatasci-092820-114757.

Diakopoulos, Nicholas, Sorelle Friedler, Marcelo Arenas, Solon Barocas, Michael Hay, Bill Howe, H. V. Jagadish, Kris Unsworth, Arnaud Sahuguet, Suresh Venkatasubramanian, Christo Wilson, Cong Yu, and Bendert Zevenbergen. n.d. "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms." Fairness, Accountability, and Transparency in Machine Learning (FAT/ML). Accessed August 12, 2023. https://www.fatml.org/resources/principles-for-accountable-algorithms.

Dresser, Rebecca. 1992. "Wanted: Single, White Male for Medical Research." *Hastings Center Report* 22 (1): 24–29. doi:10.2307/3562720.

FDA (US Food & Drug Administration). 2021. *Artificial Intelligence/Machine Learning (AI/ML)–Based Software as a Medical Device (SaMD) Action Plan*. Silver Spring, MD: FDA.

Gail, Mitchell H., Louise A. Brinton, David P. Byar, Donald K. Corle, Sylvan B. Green, Catherine Schairer, and John J. Mulvihill. 1989. "Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually." *Journal of the National Cancer Institute* 81 (24): 1879–86. doi:10.1093/jnci/81.24.1879.

Gervasi, Stephanie S., Irene Y. Chen, Aaron Smith-McLallen, David Sontag, Ziad Obermeyer, Michael Vennera, and Ravi Chawla. 2022. "The Potential for Bias in Machine Learning and Opportunities for Health Insurers to Address It." *Health Affairs* 41 (2): 212–18. doi:10.1377/hlthaff.2021.01287.

Ghasemi, Elham, Reza Majdzadeh, Fatemeh Rajabi, AbouAli Vedadhir, Reza Negarandeh, Ensiyeh Jamshidi, Amirhossein Takian, and Zahra Faraji. 2021. "Applying Intersectionality in Designing and Implementing Health Interventions: A Scoping Review." *BMC Public Health* 21 (1407). doi:10.1186/s12889-021-11449-6.

Groos, Maya, Maeve Wallace, Rachel Hardeman, and Katherine P. Theall. 2018. "Measuring Inequity: A Systematic Review of Methods Used to Quantify Structural Racism." *Journal of Health Disparities Research and Practice* 11 (2). https://digitalscholarship.unlv.edu/jhdrp/vol11/iss2/13.

Grzybowski, Andrzej, and Piotr Brona. 2021. "Analysis and Comparison of Two Artificial Intelligence Diabetic Retinopathy Screening Algorithms in a Pilot Study: IDx-DR and Retinalyze." *Journal of Clinical Medicine* 10 (11): 2352. doi:10.3390/jcm10112352.

Hardeman, Rachel R., Patricia A. Homan, Tongtan Chantarat, Brigette A. Davis, and Tyson H. Brown. 2022. "Improving the Measurement of Structural Racism to Achieve Antiracist Health Policy." *Health Affairs* 41 (2): 179–86. doi:10.1377/hlthaff.2021.01489.

Hood, Carlyn M., Keith P. Gennuso, Geoffrey R. Swain, and Bridget B. Catlin. 2016. "County Health Rankings: Relationships between Determinant Factors and Health Outcomes." *American Journal of Preventive Medicine* 50 (2): 129–35. doi:10.1016/j.amepre.2015.08.024.

Howard, Frederick M., James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, Rita Nanda, Olufunmilayo I. Olopade, Jakob N. Kather, Nicole Cipriani, Robert L. Grossman, and Alexander T. Pearson. 2021. "The Impact of Site-Specific Digital Histology Signatures on Deep Learning Model Accuracy and Bias." *Nature Communications* 12 (4423). https://www.nature.com/articles/s41467-021-24698-1.

Huang, Jonathan, Galal Galal, Mozziyar Etemadi, and Mahesh Vaidyanathan. 2022. "Evaluation and Mitigation of Racial Bias in Clinical Machine Learning Models: Scoping Review." *JMIR Medical Informatics* 10 (5): e36388. doi:10.2196/36388.

Institute of Medicine Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care. 2003. "Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care". Washington, DC: National Academies Press. https://doi.org/10.17226/12875.

Johnson, Alistair E. W., Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. "MIMIC-III, a Freely Accessible Critical Care Database." *Scientific Data* 3 (160035). doi:10.1038/sdata.2016.35.

Kasy, Maximilian, and Rediet Abebe. 2021. "Fairness, Equality, and Power in Algorithmic Decision-Making." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, March 2021: 576–586. doi:10.1145/3442188.3445919.

Kaufman Hall. 2022. *The Current State of Hospital Finances: Fall 2022 Update*. Chicago: American Hospital Association. https://www.aha.org/guidesreports/2022-09-15-current-state-hospital-finances-fall-2022-update.

Kononenko, Igor. 2001. "Machine Learning for Medical Diagnosis: History, State of the Art and Perspective." *Artificial Intelligence in Medicine* 23 (1): 89–109. https://doi.org/10.1016/S0933-3657(01)00077-X.

Kreuter, Matthew W., and Stephanie M. McClure. 2004. "The Role of Culture in Health Communication." *Annual Review of Public Health* 25:439–55. doi:10.1146/annurev.publhealth.25.101802.123000.

Lu, Jonathan, Amelia Sattler, Samantha Wang, Ali Raza Khaki, Alison Callahan, Scott Fleming, Rebecca Fong, Benjamin Ehlert, Ron C. Li, Lisa Shieh, Kavitha Ramchandran, Michael F. Gensheimer, Sarah Chobot, Stephen Pfohl, Siyun Li, Kenny Shum, Nitin Parikh, Priya Desai, Briththa Seevaratnam, Melanie Hanson, Margaret Smith, Yizhe Xu, Arjun Gokhale, Steven Lin, Michael A. Pfeffer, Winifred Teuteberg, Nigam H. Shah. 2022. "Considerations in the Reliability and Fairness Audits of Predictive Models for Advance Care Planning." *Frontiers in Digital Health* 4. https://doi.org/10.3389/fdgth.2022.943768.

Ludwig, Jens, and Sendhil Mullainathan. 2023. "Machine Learning as a Tool for Hypothesis Generation." Working Paper No. 31017. Cambridge, MA: National Bureau of Economic Research. https://www.nber.org/papers/w31017.

Manski, Charles F. 2022. "Patient-Centered Appraisal of Race-Free Clinical Risk Assessment." *Health Economics*. https://pubmed.ncbi.nlm.nih.gov/35791466/.

Manuel, Jennifer I. 2018. "Racial/Ethnic and Gender Disparities in Health Care Use and Access." *Health Services Research* 53 (3): 1407–1429. https://pubmed.ncbi.nlm.nih.gov/28480588/.

McCradden, Melissa D., Shalmali Joshi, Mjaye Mazwi, and James A. Anderson. 2020. "Ethical Limitations of Algorithmic Fairness Solutions in Health Care Machine Learning." *Lancet Digital Health* 2 (5): e221–e223. https://doi.org/10.1016/S2589-7500(20)30065-0.

MedPAC (Medicare Payment Advisory Commission). 2023. *March 2023 Report to the Congress: Medicare Payment Policy*. Washington, DC: MedPAC.

Miller, Andrew, Ziad Obermeyer, John Cunningham, and Sendhil Mullainathan. 2019. "Discriminative Regularization for Latent Variable Models with Applications to Electrocardiography." *Proceedings of Machine Learning Research* 97:4585–94. https://proceedings.mlr.press/v97/miller19a.html.

Movva, Rajiv, Divya Shanmugam, Kaihua Hou, Priya Pathak, John Guttag, Nikhil Garg, and Emma Pierson. 2023. "Coarse Race Data Conceals Disparities in Clinical Risk Score Performance." arXiv. . https://doi.org/10.48550/arXiv.2304.09270.

Mullainathan, Sendhil, and Ziad Obermeyer. 2021. "On the Inequity of Predicting A While Hoping for B." *AEA Papers and Proceedings* 111:37–42. doi:10.1257/pandp.20211078.

Mullainathan, Sendhil, and Ziad Obermeyer. 2022. "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care." *The Quarterly Journal of Economics* 137 (2): 679–727. doi: 10.1093/qje/qjab046.

Nelson, Amy Hawn, Della Jenkins, Sharon Zanti, Matthew Katz, Emily Berkowitz, T. C. Burnett, and Dennis Culhane. 2020. *A Toolkit for Centering Racial Equity Throughout Data Integration*. Philadelphia, PA: Actionable Intelligence for Social Policy. https://aisp.upenn.edu/wp-content/uploads/2022/07/AISP-Toolkit_5.27.20.pdf.

Obermeyer, Ziad, R. Nissan, M. Stern, S. Eaneff, Emily Joy Bembeneck, and Sendhil Mullainathan. 2021. *Algorithmic Bias Playbook*. Chicago, IL: Chicago Booth, Center for Applied Artificial Intelligence.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464): 447–53. doi:10.1126/science.aax2342.

O'Neil, Cathy. 2016. *Weapons of Math Destruction*. Harlow, England: Penguin Books.

Pierson, Emma, David M. Cutler, Jure Leskovec, Sendhil Mullainathan, and Ziad Obermeyer. 2021. "An Algorithmic Approach to Reducing Unexplained Pain Disparities in Underserved Populations." *Nature Medicine* 27:136–40. doi:10.1038/s41591-020-01192-7.

Raji, Inioluwa Deborah, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* 33–44. doi:10.1145/3351095.3372873.

Rajkomar, Alvin, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. 2018. "Ensuring Fairness in Machine Learning to Advance Health Equity." *Annals of Internal Medicine* 169 (12): 866–72. doi:10.7326/M18-1990.

Röösli, Eliane, Selen Bozkurt, and Tina Hernandez-Boussard. 2022. "Peeking into a Black Box, the Fairness and Generalizability of a MIMIC-III Benchmarking Model." *Scientific Data* 9 (24). doi:10.1038/s41597-021-01110-7.

Schwartz, Aaron L., Marcella Alsan, Alanna A. Morris, and Scott D. Halpern. 2023. "Why Diverse Clinical Trial Participation Matters." *New England Journal of Medicine* 388:1252–54. doi:10.1056/NEJMp2215609.

Seyyed-Kalantari, Laleh, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. 2021. "Underdiagnosis Bias of Artificial Intelligence Algorithms Applied to Chest Radiographs in Under-Served Patient Populations." *Nature Medicine* 27:2176–82. doi:10.1038/s41591-021-01595-0.

SHADAC (State Health Access Data Assistance Center). *Collection of Race, Ethnicity, Language (REL) Data on Medicaid Applications: New and Updated Information on Medicaid Data Collection Practices in the States, Territories, and District of Columbia*. Issue Brief, November 30, 2022. State Health and Value Strategies.

Singla, Sumedha, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. 2020. "Explanation by Progressive Exaggeration." arXiv. Last revised February 10, 2020. doi:10.48550/arXiv.1911.00483.

Song, Xing, Alan S. L. Yu, John A. Kellum, Lemuel R. Waitman, Michael E. Matheny, Steven Q. Simpson, Yong Hu, and Mei Liu. 2020. "Cross-Site Transportability of an Explainable Artificial Intelligence Model for Acute Kidney Injury Prediction." *Nature Communications* 11 (5668). doi:10.1038/s41467-020-19551-w.

Spielman, Seth E., David Folch, and Nicholas Nagle. 2014. "Patterns and Causes of Uncertainty in the American Community Survey." *Applied Geography* 46:147–57. doi:10.1016/j.apgeog.2013.11.002.

Subbaswamy, Adarsh, and Suchi Saria. 2020. "From Development to Deployment: Dataset Shift, Causality, and Shift-Stable Models in Health AI." *Biostatistics* 21 (2): 345–52. doi:10.1093/biostatistics/kxz041.

Verma, Sahil, and Julia Rubin. 2018. "Fairness Definitions Explained." Proceedings of the International Workshop on Software Fairness. 1–7. https://doi.org/10.1145/3194770.3194776

Vyas, Darshali A., Leo G. Eisenstein, and David S. Jones. 2020. "Hidden in Plain Sight—Reconsidering the Use of Race Correction in Clinical Algorithms." *New England Journal of Medicine* 383:874–82. doi:10.1056/NEJMms2004740.

Wang, H. Echo, Matthew Landers, Roy Adams, Adarsh Subbaswamy, Hadi Kharrazi, Darrell J. Gaskin, and Suchi Saria. 2022. "A Bias Evaluation Checklist for Predictive Models and Its Pilot Application for 30-Day Hospital Readmission Models." *Journal of the American Medical Informatics Association* 29 (8): 1323–33. doi:10.1093/jamia/ocac065.

Williams, Winifred W., Joseph W. Hogan, and Julie R. Ingelfinger. 2021. "Time to Eliminate Health Care Disparities in the Estimation of Kidney Function." *New England Journal of Medicine* 385:1804–06. https://doi.org/10.1056/NEJMe2114918.

Zhang, Haoran, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. "Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings." arXiv. Published online March 11, 2020. https://doi.org/10.48550/arXiv.2003.11515.

About the Authors

Anna Zink is a principal researcher at the Center for Applied AI at the University of Chicago Booth School of Business where she works on their Algorithmic Bias Initiative along with other projects related to the development and adoption of AI in health care. She is interested in the possibilities and pitfalls of machine learning techniques to evaluate and improve decisionmaking in health care and insurance design. Her research on health plan payment studies how to balance fairness and efficiency goals in developing risk-adjustment formulas. Zink received her PhD in health policy from Harvard University.

Sarah Morriss is a research assistant in the Health Policy Center at the Urban Institute. She analyzes data and provides assistance with questionnaire development for Urban's Health Reform Monitoring Survey and Well-Being and Basic Needs Survey. She also contributes to policy briefs and papers on topics related to health equity, health care access, and families' experiences with federal safety net programs. Her research interests include disability and mental health policy issues. Moriss has a bachelor's degree in economics and public policy from the University of Chicago.

Anuj Gangopadhyaya is an assistant professor of economics at Loyola University Chicago and was previously a senior research associate in the Health Policy Center at the Urban Institute. His research focuses on the impact of safety net programs on health and well-being, family income, and education achievement outcomes for children in low-income families. He has focused on the impact of Medicaid eligibility expansion on children's education achievement, maternal and child health effects of the earned income tax credit program, and the impact of the Affordable Care Act Medicaid expansion on adult labor supply and fertility rates of women of reproductive age. Gangopadhyaya received his PhD in economics from the University of Illinois at Chicago.

Ziad Obermeyer is an associate professor and Blue Cross of California distinguished professor at UC Berkeley, where he works at the intersection of machine learning and health. His research uses machine learning as a tool to help doctors make better decisions, and help researchers make new discoveries—by "seeing" the world the way algorithms do. He has also shown how widely used algorithms affecting millions of patients automate and scale up racial bias. That work has affected how many organizations build and use algorithms, and how lawmakers and regulators hold AI accountable. He is a Chan Zuckerberg Biohub investigator and a faculty research fellow at the National Bureau of Economic Research, and he was named an Emerging Leader by the National Academy of Medicine. Previously, he was assistant professor at Harvard Medical School, and continues to practice emergency medicine in underserved communities.

Acknowledgments

This paper was prepared for the Urban Institute's "Unequal Treatment at 20" initiative with generous support from the Robert Wood Johnson Foundation, the Commonwealth Fund, the Episcopal Health Foundation, and the California Health Care Foundation. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission. The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at urban.org/funding principles. The authors gratefully acknowledge helpful comments, suggestions, and guidance from the Chicago Booth Center for Applied AI, Advisory Committee members Nicole Stern and Jose Guillem, Urban's Community Advisory Board, and Urban Institute reviewers Judah Axelrod, Brian Smedley, Kima Joy Taylor, and Faith Mitchel.

The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at urban.org/fundingprinciples.

500 L'Enfant Plaza SW
Washington, DC 20024

www.urban.org

ABOUT THE URBAN INSTITUTE

The Urban Institute is a nonprofit research organization that provides data and evidence to help advance upward mobility and equity. We are a trusted source for changemakers who seek to strengthen decisionmaking, create inclusive economic growth, and improve the well-being of families and communities. For more than 50 years, Urban has delivered facts that inspire solutions—and this remains our charge today.

Copyright © September 2023. Urban Institute. Permission is granted for reproduction of this file, with attribution to the Urban Institute.