Algorithmic Bias

Addressing Equity and Fairness in Healthcare

Remle Newton-Dame

Marywood University

SLAS 6013: Qualifying Seminar

PhD in Strategic Leadership and Administrative Studies

Fall 2023

Algorithmic Bias

Addressing Equity and Fairness in Healthcare

Abstract

Bias in medicine is a significant threat to health equity in the United States. Algorithms promise objectivity and impartiality. However, they have not always delivered on that promise within healthcare. Since 2019, there has been an explosion of literature on the identification and mitigation of algorithmic bias, including 18 reviews. This umbrella review retains seven of those reviews, highlighting commonly cited methods, emerging areas of consensus, and gaps in the literature that must be addressed.

The emerging literature reveals active methodological development in bias identification and mitigation. Key metrics of bias identification from reviews included average odds, balanced accuracy, calibration, disparate impact, equal opportunity, equalized odds, error rate, and statistical parity. Notably, no gold standard method of bias identification was established. Mitigation measure recommendations included dataset augmentation to improve representativeness, resampling, reweighting, and label adjustment (pre-processing); adversarial debiasing, regularization, and inclusion of fairness constraints (in-processing); and label flipping, group-specific threshold setting, implementation of human-in-the-loop processes, robust continuous monitoring, inclusion of diverse stakeholders and incorporation of clinical feedback (post-processing). This umbrella review analyzes these findings from an equity, governance, and innovation adoption lens. It evaluates the threat of algorithmic bias against the ethical principles for research laid out in the Belmont Report. Finally, it provides modest recommendations that focus on practical, near-term improvements in data representativeness and bias mitigation at a local level.

Keywords: algorithmic fairness, algorithmic bias, health equity, bias mitigation

Table of Contents

Introduction

Background

In this age of data proliferation, in which an estimated 120 zettabytes of new data will be generated in a single year, human beings can no longer absorb and parse all information available in the workplace (Greifeneder, 2023). Enter the mighty algorithm. Algorithms use data to make an educated guess about an outcome or underlying state. Algorithms now drive business decision-making, advertisement targeting in web browsers, social media targeting, and traffic pattern optimization. Per a recent Brookings report, “In the pre-algorithm world, humans and organizations made decisions in hiring, advertising, criminal sentencing, and lending. These decisions were often governed by federal, state, and local laws that regulated the decision-making processes in terms of fairness, transparency, and equity. Today, some of these decisions are entirely made or influenced by machines…to influence decisions affecting people in a range of tasks, from making movie recommendations to helping banks determine the creditworthiness of individuals” (Turner Lee et al., 2019). Algorithms in production are often invisible to the end consumer, even as they subtly guide behavior. Many algorithms are now created via machine learning, using ever-larger datasets and more sophisticated computational processing. Artificial intelligence (AI) is also increasingly driving algorithmic development. As algorithms are further integrated into daily life, it is essential that we examine whether they are operating fairly.

Algorithmic bias is defined as “the instances when the application of an algorithm compounds existing inequities in socioeconomic status, race, ethnic background, religion, gender, disability or sexual orientation to amplify them and adversely impact inequities” (Panch et al., 2019). Biased algorithms have led to well-documented, grave consequences for communities of color (M. Wang et al., 2022). As clinical medicine increasingly adopts algorithms via machine learning and AI to drive staffing, bed allocation, appointment scheduling, and clinical decision-making, it is essential to identify areas of threat for algorithmic bias and source solutions that can be implemented across the development chain.

Problem Statement

Biased algorithms currently being deployed in the healthcare setting threaten to further entrench systemic health and healthcare disparities for patients of color, other legally protected groups (referred to here as "protected class status"), and unprotected groups who experience systemic discrimination, including low income and uninsured patients. Bias can be introduced during data collection, by the software developers or analysts, or during implementation within the healthcare system. Algorithms that were unbiased during deployment can also become biased as data entry patterns, patients, or the algorithms themselves change. This paper will examine how algorithmic bias may worsen existing healthcare disparities, identify techniques for assessing bias during algorithm deployment and evaluation, and delineate both statistical procedures and policy levers that can blunt or eliminate the impact of that bias. This paper hypothesizes that appropriate governance and guidelines, applied rigorously, can help address this crucial issue.

Literature Review

Context: Health Disparities

In the United States, race/ethnicity is strongly correlated with both quality of care and health outcomes. Black Americans have higher mortality rates for most kinds of cancer than other race/ethnicity groups, according to estimates by the American Cancer Society (Mitchell et al., 2022). From 2019-2020, Black people were 30% more likely to have an asthma diagnosis but almost 300% more likely to die from asthma or its sequelae (*Asthma and African Americans |*

Office of Minority Health, n.d.). A national sample of COVID-19 hospitalization data from 2020 found that Black men fared worse in acute kidney injury and in-hospital mortality as compared to white men (Pal et al., 2022), and college-educated Black women are more likely than high school-educated white women to die in childbirth (Hill et al., 2022). Although disparities in care among Black Americans are particularly stark, these kinds of findings have been replicated across other race/ethnicity groups, among patients living with lower income, and among LGBTQ+ patients.

These disparities reflect systemic inequities in the distribution of housing and clean air, exposure to pollutants, access to safe and appealing green space, and access to healthy food, as well as the impact of years of divestment in communities of color. They also reflect disparities in healthcare itself. A patient of color entering a hospital may receive different treatment than his or her white neighbor. A recent framework highlighted that Black patients received sub-optimal care on 40% of 2018 National Healthcare Quality and Disparities Report measures. Lower-income Black patients needing cardiac care were less likely than similar white patients to get a crucial bypass treatment in a national hospital survey (Javed et al., 2022). Black patients are less likely than white patients to receive adequate pain treatment (Tamayo-Sarver et al., 2003) or appropriate mental health care (Timmons et al., 2023). A 2017 meta-analysis found evidence of implicit bias in 35 of 42 studies examined, and in all 35 studies, that bias was associated with lower quality of care (FitzGerald & Hurst, 2017). These healthcare disparities are not limited to Black patients, but the disparities observed in the care received by Black patients are particularly stark.

In the wake of a national reckoning around race, racism, and bias, medicine has increasingly looked to identify and ameliorate bias in healthcare. Although the Institute of

Medicine published its landmark report "Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care" in 2003, progress has been halting (Institute of Medicine (US) Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care, 2003). The increasing adoption of algorithms can be an opportunity to reduce biased human treatment, both conscious and unconscious. However, if the algorithms adopted encode structural biases into their predictions, they will further widen healthcare disparities and decrease trust in the medical system.

Bias in Algorithms: Definition of the Field

Fundamentally, biased algorithms have more accurate predictions for some groups than others. Per a Brookings report on algorithmic bias and consumer harm, biased algorithms generate "outcomes which are systematically less favorable to individuals within a particular group…where there is no relevant difference between groups….[and if] left unchecked, biased algorithms can lead to decisions which can have a collective, disparate impact on certain groups of people" (Turner Lee et al., 2019). There are many commonly described forms of bias within machine learning and algorithmic development. Frequent culprits include historical bias, measurement bias, representation bias, temporal bias, population bias, and aggregation bias (Huang et al., 2022). However, the relatively new term "algorithmic bias" references bias generated by the algorithm itself, which results in unequal results for sub-populations. Per Huang et al., "[b]ias in a model may lead to unfairness if not appropriately evaluated and accounted for. Fairness is achieved when algorithmic decision-making does not favor an individual or group based on protected attributes." An expanding field is now considering how to minimize bias and maximize algorithmic fairness as algorithms proliferate in healthcare via machine learning and AI.

Perhaps the first instance in which algorithmic bias caught public attention was the 2016 ProPublica expose of bias in a commonly used sentencing algorithm. The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system is an algorithmic tool that uses survey and police data to make predictions about risk of recidivism to guide sentencing and parole decisions. In 2016, ProPublica found that COMPAS systematically overestimated risk among Black inmates and underestimated risk among white inmates (Kehl et al., 2017). In one sample of 10,000 inmates in Broward County, FL, COMPAS was twice as likely to mistakenly classify a Black inmate as “high risk” for violent recidivism as compared to a white inmate (Jeff Larson et al., n.d.). In this instance, an algorithm deployed in government operations contributed to systematically increasing incarceration for one group of Americans over another (Kehl et al., 2017). However, this phenomenon has been replicated across diverse settings, from mortgage rate determination (Breen et al., 2019) to smartphone facial recognition (Fisher et al., 2023).

Within healthcare, the advent of precision medicine, the rise of AI and machine learning, and the explosion of healthcare data have increasingly opened opportunities for algorithmic deployment. According to Obermeyer et al., algorithms may address crucial healthcare challenges but also must be monitored for fairness: “In health care, we are often faced with a limited supply of resources: tests, treatments, or other forms of care or extra help. Algorithms are used to help decision-makers identify who needs these resources. More generally, in many important social sectors, algorithms guide decisions about who gets what” (Ziad Obermeyer et al., 2021). As a leading voice sounding the alarm on algorithmic bias in healthcare, Obermeyer and his team have identified bias across multiple algorithms. In 2019, they found that a commonly used care management algorithm that predicted total cost of care as a proxy for healthcare need systematically misclassified the risk of Black patients, who have less access to expensive healthcare on average than comparably sick white patients (Obermeyer et al., 2019). This study, published in Science, served as a wake-up call to health services researchers and data scientists. The past three years have seen an exponential growth of this emerging discipline in this literature (Y. Wang et al., 2023).

Literature Review Approach

To evaluate the current state of algorithmic bias knowledge, this umbrella literature review sought to identify review articles focused on algorithmic bias in the medical literature. An initial search in Pubmed for the term "algorithmic bias" in title, abstract, or keyword yielded 121 results. This composite term was chosen to eliminate articles discussing other forms of bias, such as descriptions of statistical performance. The earliest article returned in the search was from 2019, suggesting that the search aligned well with the development of this new field.

The literature search identified 18 reviews, of which two were systematic reviews. Of those, 13 were retained after title review, and seven were retained after abstract review. Exclusion criteria included a primary focus on genomics, drug development, device performance, or non-healthcare applications. Reviews focusing on a single disease state or medical specialty were also excluded. All retained reviews examined the issue of algorithmic bias and fairness within a healthcare context.

Bias in Algorithms: Umbrella Review Findings

The seven retained reviews included scoping reviews, methods-oriented reviews, policy-oriented reviews, and framework reviews. Although many statistical concepts are referenced in this paper, they are not all defined in detail for brevity. The appendix includes abstracted term tables from relevant reviews to augment and define the terms referenced in this paper.

Panch et al.'s 2019 review is the earliest review included in this umbrella review. The aim of this high-level review was to define algorithmic bias in healthcare, explore the impact of bias, and identify emerging mitigation approaches. The authors stressed that significant challenges to equitable AI and algorithms within healthcare include "lack of a clear standard of fairness," "lack of contextual specificity," and "the black box nature of deep learning," which produces algorithms that cannot be explained or easily dissected by humans (Panch et al., 2019). The authors recommended that algorithms be developed with specific contexts and populations in mind, using rich data from that population. Rather than recommending specific technical solutions, they suggested bounding development with " 'human in the loop' systems, where algorithmic outputs are passed to a human decision-maker with necessary caveats"; providing bias mitigation training to data science teams; hiring staff that mirror the demographics of target populations; implicit bias education for all team members; and inclusion of clinicians in the data science development team. The authors highlighted that while transparency and explainability are key, the complexity of many currently deployed models makes this difficult in practice. They suggested focusing instead on "counterfactuals," whereby an input (race or SES, for example) is modified to understand the impact on prediction. The authors closed by calling for public sector participation in algorithms and AI via the development of fairness standards, regulation of deployment, and fomenting of public-private partnerships.

In their 2022 scoping review, Huang et al. sought to codify current methods to quantify fairness and mitigate bias from the published literature (Huang et al., 2022). Of 635 results included in their search, 12 were retained for the review. Eighty percent of the studies that looked for bias found it. Although the most common methods of bias identification were disparate impact, accuracy, and equal opportunity difference, the range of methods used across the studies was wide. Other measures of bias included balanced accuracy, statistical parity, average odds, error rate, sensitivity/specificity, AUROC, calibration, mean fairness measure, and regression of race on risk score. This heterogeneity is not surprising, given the infancy of this discipline within the healthcare field. Bias mitigation strategies employed by the review studies were deployed during data collection and curation, aka pre-processing (via resampling, augmenting existing data with other real or synthetic data, label adjustment or removal of variables); during development of the model, aka in-processing (via prejudice remover regularization, adversarial debiasing, introduction of fairness parameters or testing of multiple modeling modalities); and during deployment, aka post-processing (cutoff modification by sub-group, output adjustment). Pre-processing strategies were the most common. Post-processing strategies appeared to be the easiest to deploy to tailor models to context appropriately. Both pre- and post-processing strategies were successful. Huang et al. closed their review with a call for medicine to publish more datasets and code to increase cross-collaboration and enrich model development across the field.

In 2023, Wang et al. also published a scoping review, but their purview was broader than that of Huang et al., extending from fairness in medical AI data to cover society, law, and ethics (Y. Wang et al., 2023). This review incorporated Chinese and English literature and retained 95 articles from 824 papers reviewed. The authors noted that unlike the Chinese literature, which considered broader systems and regulation, English language articles primarily focused on technical aspects of AI solutions. Wang's review emphasized that uniform standards for evaluating, enforcing, and regulating fairness do not yet exist in the Chinese or English literature, to the detriment of AI development efforts. Furthermore, Wang et al. assert that there is a "lack of basic consensus on the concept of fairness." The Wang review did not cover measures of fairness and bias mitigation in as much detail as Huang or Xu. However, within "data fairness," they identify disparate impact and demographic parity as frequently used fairness metrics in medical AI. In the pre-processing phase, the authors call out the importance of training on unbiased or augmented data, sharing datasets across studies, actively de-biasing the data when necessary, defining variables in an inclusive way, including participants in the development lifecycle, and selecting a validation population that is diverse and representative. They also stress that routine issues such as sparse data or interactions within the model must be accounted for during development. In the in-processing phase, the authors recommend prioritizing fairness within the model development process. Possibly beneficial approaches include regularization terms, tuning parameters to balance fairness and performance, and forcing the model to prioritize equal odds or disparate impact. For post-processing, the authors cite decision-flipping to improve equal odds and differential thresholds by demographic sub-group, corresponding to fair risk prediction allocation. As an extension of post-processing, the authors suggest developing tailored monitoring and governance around algorithm deployment. They cite routine audits, regular updates to improve calibration and account for temporal trends and changes, and incorporating human review to ensure the results remain clinically relevant.

Xu et al.'s comprehensive review focused on how to concretely identify and mitigate bias within computational medicine and healthcare AI (Xu et al., 2022). Xu et al. categorized bias broadly as data bias, including sampling bias, and allocation bias; attrition bias; measurement bias; and algorithmic bias. Xu et al. argued that choice of fairness metric should be contextual. They recommended prioritizing equalized odds, which "allows the prediction Y^O to depend on protected attribute A, but only through the target variable Y," as well as "equal opportunity," which "checks whether the positive label is equally and accurately predicted by the classifier for all values of the protected attribute." For pre-processing bias mitigation approaches, Xu et al. recommended dynamic reweighting, resampling, or under-sampling the majority group in large datasets. The authors cautioned that using synthetic healthcare data to adjust the sample may lead to overfitting and impact generalizability. They found the removal of bias-indicating variables such as race to be ineffective and the imposition of demographic parity, or "requir[ing] that the overall proportion of individuals in a protected group predicted to be positive or negative [at comparable rates to] the overall population," to significantly lower prediction performance. For in-processing measures, the authors discussed emerging methodology in statistical prejudice remover approaches, independent learning (aka separating models by sub-group), transfer learning from the genomics field, and adversarial learning that prioritizes equalized odds. They also recommended interpretable learning, aka models that can be explained and understood, over black-box models, which may obscure sources of bias. For post-processing, the authors cited innovative studies employing equalized odds post-processing and calibration equalized odds (from Hardt et al.), risk score adjustments using a parameterized monotonically increasing function (from Kallus et al.), rank order adjustment via dynamic programming (from Cui et al.), and causal analysis (from Pan et al.). Although this review highlighted many mitigation methods, they did not identify a gold standard approach.

Norori et al.'s 2021 review highlighted that bias can enter the model at all stages of development and deployment. The authors point out that much algorithmic development happens using research datasets in which affluent, Caucasian, middle-aged men are overrepresented (Norori et al., 2021). They emphasize the importance of uniform data standards to foster data exchange and measure representation (*Welcome to the Open Standards for Data Handbook*, n.d.). In addition to metrics identified by other reviews, they suggest the F1 score to better target class imbalance (true positives / true positives + (false negatives + false positives)/2). They suggest that using synthetic data may show promise in improving performance in protected classes. They highlight the option of constraining the loss function to stay below a threshold for each protected class "so that no single group is systematically misclassified." They argue that open science, or "openly sharing multiple facets of the research process, including data, methods, and results under terms that allow reuse, redistribution, and reproduction of all findings," is critical to improving trust and performance in AI and machine learning. However, the authors note that healthcare data require different privacy safeguards. Federated learning models (where each unit controls its own data) or allowing localities to retrain algorithms on their own data may be more feasible than universal data exchange. The authors finish by recommending that AI developers work to produce explainable, transparent models (aka "white box") in which humans can review model decision-making while including patients in the development process.

Thomasian et al.'s 2021 viewpoint review took a more pragmatic frame (Thomasian et al., 2021). This review was developed primarily for regulators but also included guidance for developers. As with other reviews, the authors identified data representativeness as a crucial issue and proposed large open databases, the use of synthetic data, and federated learning or cyclical weight transfer as options to address the challenge. They also encouraged developers to think critically about the role of bias in underlying missing data and data quality. The authors recommended that developers weigh the value of the multiple available metrics mentioned in other reviews against the specific context and needs of their project. AI developed to operate in high-jeopardy situations like ICU resource deployment should be required to complete clinical trials, similar to therapeutics. The authors highlighted interpretability as an essential bias mitigation tool that's crucial to promote trust and encourage uptake in the implementation phase.

In addition to evaluating uptake and collecting user feedback, the authors recommended the implementation of "continual bias auditing and surveillance" as part of the overall tuning process. The authors felt that regulators should also develop mandatory regular reporting of the results of bias and quality surveillance, which could be developed by task forces at the federal, state, and local levels. Perhaps because the review was developed for a policy audience, recommendations remained broad and high-level.

Paulus and Kent's 2020 review used ethical and legal rationales to "introduce a practical framework for evaluating algorithmic bias and fairness in clinical decision-making and prediction in healthcare" (Paulus & Kent, 2020). Their high-level framework was designed for practitioners, administrators, and clinicians alike. Paulus and Kent illustrated that all fairness metrics cannot be satisfied at once. Extremely biased algorithms such as the COMPAS recidivism tool can appear unbiased if an inappropriate fairness metric is reported. The authors flag label bias (or differences in rate of observation or meaning of outcome by group) and feature bias (differences in rate of observation or meaning of predictor variable by group) as sources of algorithmic bias, in addition to differential missingness and sampling bias. Paulus and Kent argue that when determining which identification and mitigation strategies to use, developers should consider whether the algorithm use case is polar (e.g., patient with high risk score gains access to a scarce resource) or non-polar (e.g., patient and clinician get risk information to aid in collaborative decision-making). For non-polar indications where an outcome does not suffer from label bias, the authors argue that the inclusion of a protected class variable or interaction terms in the model or the use of stratified algorithms can sufficiently address bias. To address label bias, they recommend jettisoning proxy variables and identifying a non-biased outcome to predict, a strategy that may be less feasible in practice than in theory. In addition to highlighting the importance of the above steps, the Paulus and Kent framework proposes that developers looking to mitigate unfairness should include only well-validated causal variables that are not proxies for the protected class or statistically enforce fairness in the in-processing and post-processing stages. Throughout, they stress the need to tailor the approach to context by incorporating significant stakeholder input.

Analysis

Major Themes

The literature review highlighted that the study of algorithmic bias in healthcare is a rapidly expanding area of study, even as it builds on more established machine learning literature from other disciplines. There were many areas of consensus among the authors, including the need to build algorithms using high-quality and representative data, the need to quantify and address fairness within machine learning and AI, and the need to actively mitigate bias in the pre-, in-, and post-processing phases to produce equitable results. Additionally, most authors agreed on key techniques in the pre-processing (data augmentation, resampling or reweighting, and selection of outcomes that do not suffer from label bias), in-processing (prioritization of equalized odds, adversarial debiasing, regularization, and creation of interpretable models) and post-processing (implementation of human review, stakeholder engagement, and continuous monitoring) phases. However, consensus around monitoring and mitigating bias was elusive, and multiple frameworks and data quality standards were referenced without indication of a gold standard. Below, several overarching themes are explored in more detail through an equity, governance, and innovation adoption lens.

Equity Perspective: Removing Race Is Insufficient to Solve Bias

Much of the conversation currently underway in the clinical algorithm space has been focused recently on removing race. The NYC Department of Health and Mental Hygiene launched the Coalition to Confront Racism in Medical Algorithms in 2021, for example, with the goal of tackling "race norming," or the setting of different therapeutic guidelines by race/ethnicity (*Health Department Launches Coalition to Confront Racism in Medical Algorithms - NYC Health*, n.d.). Visweswaran et al. recently published a database to track race-based clinical algorithms, including risk calculators, medication recommendations, and laboratory range guidelines, to encourage the removal of race from medical decision-making tools. Within machine learning and AI, the strategy of simply excluding variables for protected classes such as race is termed "fairness through unawareness." By virtue of its simplicity to implement, this strategy is an attractive one.

Most authors in these reviews argued that excluding race from predictive algorithms was ineffective. As Kehl et al. explain, "excluding race itself does not necessarily mean that factors that correlate heavily to an individual's race—serving essentially as proxies for race—are excluded from these algorithms" (Kehl et al., 2017). Panch and Xu both found variable suppression to be unsuccessful. Huang et al. highlighted that "models may be able to infer protected group membership from other data features [and]...omission of protected attributes may cause bias if a legitimate relationship exists between the attribute and outcome of interest" (Huang et al., 2022). Norori et al. recommended including all available protected class variables, such as race or SES, to be able to test whether the predictions are independent of these variables (Norori et al., 2021). Many reviews cited Obermeyer et al., who identified outcomes as a more significant source of bias than predictors, due to issues in differential labeling or data capture within an observed outcome that can systematically disadvantage one group over another (Ziad

Obermeyer et al., 2021). Obermeyer et al. argued for the prioritization of direct outcome measures (e.g., death) over proxy outcome measures (e.g., cost) to address this issue. They recommended including protected class predictors in modeling. Paulus and Kent allowed that it was appropriate to suppress protected class variables when algorithms were leveraged to determine allocation of scarce resources or receipt of negative consequences (aka polar decisions) but argued that algorithms designed to optimize patient decision-making should incorporate these data (Paulus & Kent, 2020). Holistically, the literature review suggested that simply suppressing protected class variables is inadequate to remove bias.

Enhancing Equity Through Representation

Since the term "digital divide" was coined in 1999, concerns have been raised that the increasing digitization of medicine would lead to differential impacts by race/ethnicity, SES, and other protected classes. As van Dijk's Theory of Digital Technology Access and Societal Impacts outlines, "categorical inequalities in society produce an unequal distribution of resources…[and] an unequal distribution of resources causes unequal access to digital technologies" (Van Dijk, 2017). Much of the digital divide conversation has focused on differential access to medical practices with an electronic medical record, adoption of patient portals, or use of telehealth visits. However, the impact of the digital divide can extend from differential access to the differential impact of technology. The widespread use of medical record data by machine learning and AI to direct care will inherently produce biased results if some groups have less access to the underlying technology. Said another way, an algorithm cannot be fair if some groups of patients are not afforded the chance to contribute to the datasets upon which the algorithm is trained.

Addressing this issue was a focus across the included reviews. All review authors highlighted the need to build algorithms and AI on datasets that represent the diversity of a target population. As stated by Norori et al., "[v]ulnerable groups have a long history of being absent or misrepresented in existing datasets. When AI algorithms are trained with datasets in which vulnerable groups are not well represented, their predictive value may be limited" (Norori et al., 2021). Some authors focused on US representation and inclusion of historically excluded groups. Other reviews highlighted the need to capture international populations to ensure that algorithms and AI built for the developed world can benefit the developing world. Multiple studies cited the challenge of generalizing findings from genomics studies that include predominantly European or Asian populations.   In addition to synthetic data solutions, studies stressed the importance of data sharing using a common set of standards to improve the representativeness of machine learning datasets. However, they stopped short of identifying mechanisms by which these data could be identified and included.

Creating more diverse development datasets is essential. Safety net healthcare systems have an important role to play when it comes to augmenting traditionally used academic medical center clinical datasets to include diverse populations. Safety net systems are anchors for patients in urban and rural areas who are uninsured or under-insured, as well as those who have structural barriers to healthcare access at other systems. Incorporation of data from safety net systems into research collaboratives would be a step toward data equity. However, the financial barriers to such participation are significant. Safety net clinics and hospitals are the primary sources of care for uninsured patients. However, they are reimbursed for only 65% of the uncompensated care that they deliver (Khullar et al., 2018). While academic medical centers may have a venture capital arm that can profit from gains realized from AI innovation using their data, safety net systems do not. If the field of AI development hopes to address bias and fairness, it will be crucial to subsidize safety net systems' contribution of data to research repositories. To garner trust and encourage participation, it will also be necessary to guarantee access to that data among diverse partners within the safety net and the community.

Lack of Uniform Standards Hinders Bias Identification

Multiple authors highlighted the need for additional work to define a uniform set of evaluation standards for fairness and bias. Huang et al. found that metrics used across studies were heterogeneous and called for a more standardized approach for the developing field (Huang et al., 2022). Panch et al. found that "there is no broadly recognized quantitative summary metric for fairness and hence evaluation is ultimately qualitative" (Panch et al., 2019). Huang et al.'s solution was to incorporate multiple measures of bias within study design, as different biases might be highlighted by different measures. Other authors like Kleinberg et al. argued that there is no way to satisfy all possible metrics of algorithmic fairness, and each study must prioritize its own goals (Kleinberg et al., 2016). Xu et al. noted that "different strategies have different assumptions, [and] therefore it is challenging to have a gold standard" and recommended that developers test multiple methods and optimize for their context (Xu et al., 2022). The lack of uniform standards or metrics makes evaluating bias across studies or AI applications difficult. Without uniform measures adopted across the field and reinforced by journals and funders, it will be impossible to quantify the scale of the problem.

Governance Perspective: Include Diverse Stakeholders and Address Local Context

Often, corporate governance evolves in response to regulatory demands. This is true in the routine healthcare data and IT sphere, where governance that enshrines privacy and reporting requirements is well established. However, regulation in the algorithmic bias space has been slower to emerge. In their Framework for Responsible Innovation, Owen et al. theorized that a regulatory Responsibility Gap emerges when the usual mechanisms of regulation cannot keep up with the demands of "novel science and technology which are highly uncertain in terms of their current and future impacts, or which, by virtue of their novelty, have no historical precedent" (Owen et al., 2013). This theory highlighted that "[r]egulation, put simply, struggles with innovations that it has not encountered before." The advent of generative AI and the speed of machine learning diffusion into healthcare have created a regulatory vacuum that policy is slowly starting to fill. Strong governance at the healthcare system level may be an interim answer.

The theme of governance, while not addressed directly in all reviews, was omnipresent in the authors' calls to engage stakeholders, develop human-in-the-loop processes, craft continuous monitoring protocols, and engage clinicians for implementation feedback. Multiple articles recommended connecting with stakeholders to determine how to prioritize which metrics to use to identify bias, which protected classes to prioritize for bias mitigation, and how to design continuous monitoring systems to catch bias during the post-implementation period. Some reviews cited Obermeyer's Algorithmic Bias Playbook, which provides a qualitative guide to stakeholder engagement and evaluation. However, concrete governance guidance was largely lacking across reviews.

Identifying and engaging the right stakeholders with the right mandate will be essential to establish the kind of collaborative decision-making that the literature review suggests will mitigate bias in practice. In the absence of guidance from regulators or funders, models of algorithmic governance currently being established by early adopter systems like Duke Health may offer a replicable model. Duke informaticists identified “a lack of consensus in establishing governance to deploy, pilot, and monitor algorithms within operational healthcare delivery workflows” (Bedoya et al., 2022)**.** Their algorithm-based clinical decision support (ABCDS) process provides a governance structure for "model development…silent evaluation…effectiveness evaluation…and general deployment." Models must be registered and triaged to an appropriate level of review based on the strength of evidence, standard of practice, and whether the model could be designated "Software as a Medical Device" (SaMD) under FDA criteria. In order to progress between phases, models must pass checkpoints with explicit criteria, monitored by committees with expertise appropriate to those checkpoints. Equity and fairness assessments across sex, race/ethnicity, age, and insurance status are explicitly required to move from model development to silent evaluation and from effectiveness evaluation to general deployment. The ABCDS process provides a concrete guide for systems looking to establish governance that balances algorithmic equity, quality, and performance. Governance processes like ABCDS are crucial as more systems attempt to systematize their adoption of machine learning and AI.

Innovation Perspective: Moving the Field Forward from Methods Innovation to Adoption

For this burgeoning discipline to have maximum impact, it will be essential that the bias identification and mitigation techniques identified in these reviews be adopted by healthcare systems and AI companies. According to Rogers' Diffusion of Innovation Theory, new ideas are not adopted uniformly but by waves of uptake, in which people or systems can be roughly classified as innovators, early adopters, early majority, late majority, or laggards (Diffusion of Innovation Theory, n.d.). To diffuse beyond innovators and early adopters, a technology must be testable and straightforward to implement. Because machine learning and AI are perhaps more mature in fields outside of medicine, the reviews cited the use of discipline-agnostic tools and code repositories as facilitators of adoption. Xu et al. detailed code repositories from other disciplines that can be leveraged within healthcare to monitor and address bias. IBM's Artificial

Intelligence 360, available in both Python and R, showed particular promise in terms of "lack of complexity" and "trialability," two facilitators of adoption, according to Rogers.

However, the availability of tools is unlikely to be enough to push bias mitigation into universal practice. Per institutional theory, organizations make change through coercive pressures (requirements or public pressure), mimetic pressures (pressure to adopt best practices of peer organizations), or normative pressures (pressure from professionals, aka physicians or researchers, to adopt a practice) (Birken et al., 2022). The revelation of extensive health and healthcare disparities ranging from maternal mortality to COVID-19 outcomes over the last decade has led to significant coercive, mimetic, and normative pressures on healthcare institutions to address health equity, from the public, from gold standard institutions, and from physician's organizations like the American Medical Association (Advancing Health Equity: A Guide to Language, Narrative and Concepts, n.d.). By recognizing the power of algorithms to either democratize access to healthcare resources and address human bias or perpetuate those biases, healthcare organizations can further their own health equity goals while ensuring that they implement AI and machine learning correctly from the outset. While formal legislation develops, informal regulators like funders, institutional review boards, and academic journals have a role to play in establishing new norms for research and practice.

Ethical Implications

Ignoring the potential for bias in algorithms and machine learning produces serious ethical concerns. Authors such as Paulus and Kent have attempted to tease apart the challenges that biased algorithms pose at a legal level. As they highlight, this is a fraught endeavor: “There are two competing principles or goals in antidiscrimination law:…anticlassification, [which aims] to eliminate the unfairness individuals experience due to bias in decision makers' choices,

[and] antisubordination [which] seeks to eliminate status-based inequality across protected classes. [However, enforcing] balance in outcomes or results can only indirectly address anticlassification concerns…[and] ensuring fair process…requires adjudicating the degree of difference between groups that a fair society should tolerate" (Paulus & Kent, 2020). Because the body of law surrounding algorithms and AI deployment is only starting to emerge, the application of ethical principles may be fruitful. The impact of biased algorithms can fruitfully be examined using the Belmont Report's three ethical principles: respect for persons, beneficence, and justice.

The Belmont Report's principle of justice requires that the benefits and costs of research accrue equally to all participants in research. By extension, justice requires that the development of a new algorithm or AI tool provide equal opportunities for benefit or harm to each patient, regardless of their membership in a protected class. The healthcare spending algorithm flagged by Obermeyer et al. systematically underestimated how sick Black patients were and did not flag them for care management support at rates comparable to white peers (Obermeyer et al., 2019). This violates the principle of justice.

The Belmont Report's principle of respect for persons requires that research be careful not to "show lack of respect for an autonomous agent...[by] repudiat[ing] that person's considered judgments, [by] den[ing] an individual the freedom to act on those considered judgments, or [by] withhold[ing] information necessary to make a considered judgment" (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979). Many algorithms are created in order to identify the risk of disease for an individual patient in order to help the healthcare team decide the best course of treatment. The Framingham Risk Score is an example of a healthcare algorithm used in this way. However, it systematically underestimates the risk of heart disease among non-white patients (Paulus & Kent, 2020). This is functionally withholding available information, which could result in a patient not receiving a drug, lifestyle, or surgical intervention that could save their lives. This clearly violates the principle of respect for persons.

The Belmont Report's principle of beneficence requires that "[p]ersons are treated in an ethical manner not only by respecting their decisions and protecting them from harm, but also by making efforts to secure their well-being…[the] two general rules [are] 1) do not harm and 2) maximize possible benefits and minimize possible harms." The two examples cited above clearly violate the principle of beneficence. If the Framingham risk score underestimates cardiovascular risk among non-white patients, it is impossible for the healthcare system to maximize possible benefits with appropriate treatment. The Obermeyer example shows that deploying a racially biased algorithm to make decisions about scarce resource deployment results in measurable harm, with white patients more likely than commensurately sick Black patients to get access to services tailored to their needs.

Addressing algorithmic bias in the pre-processing, in-processing, and post-processing stages upholds the principles of respect for persons, beneficence, and justice. Responsibility for safeguarding these principles extends beyond software companies and algorithm developers to deploying healthcare systems, insurance companies, and researchers. Because algorithmic bias can enter at any stage in the algorithm life cycle, it must be identified and mitigated by multiple players in the healthcare information ecosystem. The Fairness, Accountability, and Transparency in Machine Learning project produced "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms," which provides health system players with a guide for how to ask ethical questions of their algorithm development and deployment. Forthcoming legislation at the national and state levels may further codify ethical requirements for algorithms and machine learning as the field advances. Strong governance that prioritizes ethics and equity will be necessary to ensure that the Belmont Report principles are upheld, inside and outside of research applications.

Policy Recommendations

To address biased algorithms, policy will be necessary at multiple levels of the regulatory and governance ecosystem. Some policy development is already underway. In 2021, the FDA published its "Artificial Intelligence/Machine Learning-Based Action Plan," a precursor to more concrete guidelines, under the aegis of their existing "Software as Medical Devices" classification (Thomasian et al., 2021). The Algorithmic Accountability Act of 2022 (H.R 6580) and Health Equity and Accountability Act of 2022 (H.R 7585) were introduced in Congress in 2022, and in 2023 Senate Majority Leader Schumer announced that the Senate would work to create an AI regulation framework. The White House also created a Blueprint for an AI Bill of Rights, which lists algorithmic discrimination protections as one of its five main principles (*Blueprint for an AI Bill of Rights | OSTP*, n.d.). These federal efforts could provide a crucial floor, but they are insufficient to address the many sources of algorithmic bias during the development lifecycle.

Every review included in this literature summary identified non-representative datasets as a fundamental threat to equitable algorithms. Requiring algorithm developers to augment their base datasets with data that mirror the receiving population is a crucial step. This will likely require an infusion of funding at the federal level to augment current large homogeneous research datasets with data that has sufficient sample size across different protected classes, including race/ethnicity, gender identity, and sexual orientation. Safety net healthcare systems often care for a disproportionate share of low-income patients, uninsured patients, and communities of color. Funding safety net systems to contribute data to research and development collaboratives is one avenue to enrich datasets while strengthening the public systems required to deploy them equitably. Funding the evaluation of algorithms deployed within safety net healthcare systems would surface algorithmic bias that might not otherwise be evident while supporting those systems to adopt innovations to benefit safety net patients equitably.

Although federal policy may ultimately incentivize more representative data repositories, more is needed to ensure that algorithms are deployed equitably in healthcare. As Thomasian et al. flag, the FDA only regulates algorithms used to "drive clinician decision making or to analyze patient health data or medical images" (Thomasian et al., 2021). This ignores the many operational algorithms currently being deployed in healthcare that impact patients. For example, appointment no-show predictive algorithms may systematically over-predict no-show risk among Black patients, leading to systematic overbooking and longer in-clinic wait times. It is essential that health system-level governance be developed to address bias at each stage of the algorithmic life cycle.

Healthy algorithmic governance should include identification and remediation of bias not only during development but also during deployment and via continuous quality monitoring. Governance should require the collection of performance feedback from developers, clinical partners, and end users. Patient advocates must also have a seat at the table, and health systems must make an effort to enhance transparency of these algorithms via "white box" approaches so that patients can participate in decision-making. Health systems must identify *which* protected classes to prioritize for bias remediation (race/ethnicity vs. gender vs. sexual orientation, for example). This needs to be both context-specific and organization-specific. One health system might prioritize groups that have experienced significant discrimination in the specific outcome, while another might prioritize patients facing structural barriers to health within their city. Community groups and other trusted brokers can be instrumental in this prioritization. That participation can be incentivized via Institutional Review Boards, professional societies, funders, and other research-oriented standard bearers. As the reviews emphasize, no one strategy is sufficient to overcome biased algorithms. However, with safeguards and thoughtful processes at each stage of the life cycle, healthcare may be able to improve upon its checkered history and share the benefits of innovation with all patients equitably.

Summary

Bias in medicine is a significant issue that impacts patients across disease states, specialties, and economic strata. In a healthcare environment in which college-educated Black women are more likely to die in childbirth than high school-educated white women, it is clear that health and healthcare disparities are pernicious and unacceptable (Hill et al., 2022). Algorithms promise objectivity and impartiality. They may ultimately play a role in reducing healthcare disparities. However, they have not necessarily delivered on that promise. In 2019, Obermeyer et al. demonstrated that a well-vetted algorithm predicting total cost of care to estimate underlying sickness systematically disadvantaged Black patients in favor of white patients (Obermeyer et al., 2019). Their findings have been replicated, indicating that this was not an isolated problem but rather a feature of algorithmic development and deployment.

Since 2019, there has been an explosion of literature on the identification and mitigation of algorithmic bias. Algorithmic bias is defined as "the instances when the application of an algorithm compounds existing inequities in socioeconomic status, race, ethnic background, religion, gender, disability or sexual orientation to amplify them and adversely impact inequities"

(Panch et al., 2019). By October 2023, 121 articles indexed in Pubmed featured the exact term "algorithmic bias," including 18 reviews. This umbrella literature review retained seven reviews and summarized each author team's approach.

Reviews identified key metrics of bias identification, including average odds, balanced accuracy, calibration, disparate impact, equal opportunity, equalized odds, error rate, and statistical parity (Huang et al., 2022). However, there was consensus across reviews that no gold standard method of bias identification in algorithmic development has been established, and a call for development in the field was issued.

Reviews also detailed pre-processing, in-processing, and post-processing techniques of bias mitigation. Key pre-processing techniques included dataset augmentation to improve representativeness, resampling, reweighting, and label adjustment. Frequently mentioned in-processing techniques included adversarial debiasing, regularization, and inclusion of fairness constraints. More heterogeneity was observed for post-processing techniques, which ranged from statistical (label flipping, group-specific threshold setting) to process-oriented (implementation of human-in-the-loop, robust continuous monitoring strategies, inclusion of stakeholders, feedback solicitation from clinicians). Only one review detailed the success or failure of algorithmic bias mitigation techniques in practice.

It is essential that algorithms and AI be developed on datasets that include all segments of a target population to improve health equity. Strategies include data sharing using universal data standards, sampling, and the inclusion of synthetic data, as well as distributed research collaboratives, which allow more local data control to protect healthcare data. Enhanced focus on the inclusion of safety net systems in research and development efforts, with commensurate funding, may yield significant dividends in improving data parity.

Simply ignoring the presence of bias (or addressing it through only the easiest-to-implement strategies, such as the removal of protected class predictor variables) is ethically untenable. Biased algorithms put into clinical care violate respect for persons, beneficence, and justice, which are ethical principles not only in research but in the practice of healthcare more broadly. Moreover, while machine learning is more advanced in other fields, the field of algorithmic bias and mitigation in healthcare is still nascent. Frameworks like the Principles for Accountable Algorithms are a helpful guide, as are publicly available code repositories such as the IBM AI Fairness 360 (Bellamy et al., 2019; *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms :: FAT ML*, n.d.). However, large-scale adoption of these innovations in middle and late-adopter health systems will require policy, governance, and support at multiple levels.

Policy to regulate algorithmic development, including identification of bias, is beginning to emerge. The FDA published an action plan in 2021 under the broader Software as a Medical Device umbrella, which was followed by a White House AI bill of rights and the introduction of two congressional AI-focused bills (*Blueprint for an AI Bill of Rights | OSTP*, n.d.; Health, 2023; Thomasian et al., 2021). However, to mitigate the many sources of bias throughout the lifecycle, bias identification and mitigation must also happen at the local level (Norori et al., 2021). Strong local governance, prioritizing the inclusion of multiple stakeholders as well as patients, will be crucial for bias mitigation to be successful in practice. In addition to the incorporation of high-quality bias identification metrics and the application of vetted pre-, in-, and post-processing mitigation methods, this governance must require robust stakeholder participation, evaluation of local context and priorities, and continuous monitoring post-implementation. Funders, IRBs, academic journals, and accreditation bodies can all support bias mitigation through guidance, requirement creation, and funding. Improving algorithmic fairness will require action across the healthcare ecosystem, but the reviews included here have illustrated multiple paths to improving data-driven health equity.

References

Advancing Health Equity: A Guide to Language, Narrative and Concepts. (n.d.). *Office of Minority Health*. (n.d.). Retrieved October 8, 2023, from https://minorityhealth.hhs.gov/asthma-and-african-americans

Bedoya, A. D., Economou-Zavlanos, N. J., Goldstein, B. A., Young, A., Jelovsek, J. E., O'Brien, C., Parrish, A. B., Elengold, S., Lytle, K., Balu, S., Huang, E., Poon, E. G., & Pencina, M. J. (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), 1631–1636. https://doi.org/10.1093/jamia/ocac078

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5), 4:1-4:15. https://doi.org/10.1147/JRD.2019.2942287

Birken, S. A., Ko, L. K., Wangen, M., Wagi, C. R., Bender, M., Nilsen, P., Choy-Brown, M., Peluso, A., & Leeman, J. (2022). Increasing Access to Organization Theories for Implementation Science. *Frontiers in Health Services,* 2, 891507. https://doi.org/10.3389/frhs.2022.891507

Blueprint for an AI Bill of Rights | OSTP. (n.d.). The White House. Retrieved October 14, 2023, from https://www.whitehouse.gov/ostp/ai-bill-of-rights/

Breen, N., Berrigan, D., Jackson, J. S., Wong, D. W. S., Wood, F. B., Denny, J. C., Zhang, X., & Bourne, P. E. (2019). Translational Health Disparities Research in a Data-Rich World. *Health Equity,* 3(1), 588–600. https://doi.org/10.1089/heq.2019.0042

Diffusion of Innovation Theory. (n.d.). Retrieved October 8, 2023, from https://sphweb.bumc.bu.edu/otlt/MPH-Modules/SB/BehavioralChangeTheories/BehavioralChangeTheories4.html

Fisher, E., Flynn, M. A., Pratap, P., & Vietas, J. A. (2023). Occupational Safety and Health Equity Impacts of Artificial Intelligence: A Scoping Review. *International Journal of Environmental Research and Public Health*, 20(13), 6221. https://doi.org/10.3390/ijerph20136221

FitzGerald, C., & Hurst, S. (2017). Implicit bias in healthcare professionals: A systematic review. *BMC Medical Ethics*, pp. 18, 19. https://doi.org/10.1186/s12910-017-0179-8

Greifeneder, B. (2023, April 24). Getting Tool Sprawl Under Control To Enable Data-Driven Business And Cloud-Scale Growth. *Forbes*. https://www.forbes.com/sites/forbestechcouncil/2023/04/24/getting-tool-sprawl-under-control-to-enable-data-driven-business-and-cloud-scale-growth/?sh=27c7b8eb68dc

Health, C. for D. and R. (2023). Artificial Intelligence and Machine Learning in Software as a Medical Device. *FDA*. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

Health Department Launches Coalition to Confront Racism in Medical Algorithms—NYC Health. (n.d.). Retrieved October 14, 2023, from https://www.nyc.gov/site/doh/about/press/pr2021/health-department-launches-cerca.page

Hill, L., Ranji, U., & Artiga, S. (2022). Racial Disparities in Maternal and Infant Health: Current Status and Efforts to Address Them | KFF. *Kaiser Family Foundation*. https://www.kff.org/racial-equity-and-health-policy/issue-brief/racial-disparities-in-maternal-and-infant-health-current-status-and-efforts-to-address-them/

Huang, J., Galal, G., Etemadi, M., & Vaidyanathan, M. (2022). Evaluation and Mitigation of

Racial Bias in Clinical Machine Learning Models: Scoping Review. *JMIR Medical Informatics*, 10(5), e36388. https://doi.org/10.2196/36388

Institute of Medicine (US) Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care. (2003). *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care* (B. D. Smedley, A. Y. Stith, & A. R. Nelson, Eds.). National Academies Press (US). http://www.ncbi.nlm.nih.gov/books/NBK220358/

Javed, Z., Haisum Maqsood, M., Yahya, T., Amin, Z., Acquah, I., Valero-Elizondo, J., Andrieni, J., Dubey, P., Jackson, R. K., Daffin, M. A., Cainzos-Achirica, M., Hyder, A. A., & Nasir, K. (2022). Race, Racism, and Cardiovascular Health: Applying a Social Determinants of Health Framework to Racial/Ethnic Disparities in Cardiovascular Disease. *Circulation: Cardiovascular Quality and Outcomes*, 15(1), e007917. https://doi.org/10.1161/CIRCOUTCOMES.121.007917

Jeff Larson, Mattu, S., Julia Angwin, & Lauren Kirchner. (n.d.). How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*. Retrieved September 14, 2023, from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Kehl, D., Guo, P., & Kessler, S. (2017). Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing. *Berkman Klein Center for Internet & Society*, Harvard Law School, Responsive Communities Initiative.

Khullar, D., Song, Z., & Chokshi, D. A. (2018). Safety-Net Health Systems At Risk: Who Bears The Burden Of Uncompensated Care? *Health Affairs Forefront*. https://doi.org/10.1377/forefront.20180503.138516

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores (arXiv:1609.05807). arXiv.

http://arxiv.org/abs/1609.05807

Mitchell, E., Alese, O. B., Yates, C., Rivers, B. M., Blackstock, W., Newman, L., Davis, M., Byrd, G., & Harris, A. E. (2022). Cancer healthcare disparities among African Americans in the United States. *Journal of the National Medical Association*, 114(3), 236–250. https://doi.org/10.1016/j.jnma.2022.01.004

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). The Belmont report: Ethical principles and guidelines for the protection of human subjects of research. *U.S. Department of Health and Human Services*. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html

Nicole Turner Lee, Paul Resnick, & Genie Barton. (2019). Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms | Brookings (Artifical Intelligence and Emerging Technology Initiative). *Brookings Institute*. https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns* (New York, N.Y.), 2(10), 100347. https://doi.org/10.1016/j.patter.2021.100347

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

Owen, R., Stilgoe, J., Macnaghten, P., Gorman, M., Fisher, E., & Guston, D. (2013). A Framework for Responsible Innovation. In R. Owen, J. Bessant, & M. Heintz (Eds.), *Responsible Innovation* (1st ed., pp. 27–50). Wiley. https://doi.org/10.1002/9781118551424.ch2

Pal, S., Gangu, K., Garg, I., Shuja, H., Bobba, A., Chourasia, P., Shekhar, R., & Sheikh, A. B. (2022). Gender and Race-Based Health Disparities in COVID-19 Outcomes among Hospitalized Patients in the United States: A Retrospective Analysis of a National Sample. *Vaccines*, 10(12), Article 12. https://doi.org/10.3390/vaccines10122036

Panch, T., Mattie, H., & Atun, R. (2019). Artificial intelligence and algorithmic bias: Implications for health systems. *Journal of Global Health*, 9(2), 010318. https://doi.org/10.7189/jogh.09.020318

Paulus, J. K., & Kent, D. M. (2020). Predictably unequal: Understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ Digital Medicine*, 3, 99. https://doi.org/10.1038/s41746-020-0304-9

Principles for Accountable Algorithms and a Social Impact Statement for Algorithms: FAT ML. (n.d.). Retrieved October 8, 2023, from https://www.fatml.org/resources/principles-for-accountable-algorithms

Tamayo-Sarver, J. H., Hinze, S. W., Cydulka, R. K., & Baker, D. W. (2003). Racial and ethnic disparities in emergency department analgesic prescription. *American Journal of Public Health*, 93(12), 2067–2073. https://doi.org/10.2105/ajph.93.12.2067

Thomasian, N. M., Eickhoff, C., & Adashi, E. Y. (2021). Advancing health equity with artificial intelligence. *Journal of Public Health Policy*, 42(4), 602–611. https://doi.org/10.1057/s41271-021-00319-5

Timmons, A. C., Duong, J. B., Simo Fiallo, N., Lee, T., Vo, H. P. Q., Ahle, M. W., Comer, J. S., Brewer, L. C., Frazier, S. L., & Chaspari, T. (2023). A Call to Action on Assessing and

Mitigating Bias in Artificial Intelligence Applications for Mental Health. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 18(5), 1062–1096. https://doi.org/10.1177/17456916221134490

Van Dijk, J. A. G. M. (2017). Digital Divide: Impact of Access. In P. Rössler, C. A. Hoffner, & L. Zoonen (Eds.), *The International Encyclopedia of Media Effects* (1st ed., pp. 1–11). Wiley. https://doi.org/10.1002/9781118783764.wbieme0043

Wang, M., Zhang, Y., & Deng, W. (2022). Meta Balanced Network for Fair Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11), 8433–8448. https://doi.org/10.1109/TPAMI.2021.3103191

Wang, Y., Song, Y., Ma, Z., & Han, X. (2023). Multidisciplinary considerations of fairness in medical AI: A scoping review. *International Journal of Medical Informatics*, 178, 105175. https://doi.org/10.1016/j.ijmedinf.2023.105175

Welcome to the Open Standards for Data Handbook. (n.d.). *Open Standards for Data Guidebook*. Retrieved October 8, 2023, from https://standards.theodi.org/

Xu, J., Xiao, Y., Wang, W. H., Ning, Y., Shenkman, E. A., Bian, J., & Wang, F. (2022). Algorithmic fairness in computational medicine. *EBioMedicine*, 84, 104250. https://doi.org/10.1016/j.ebiom.2022.104250

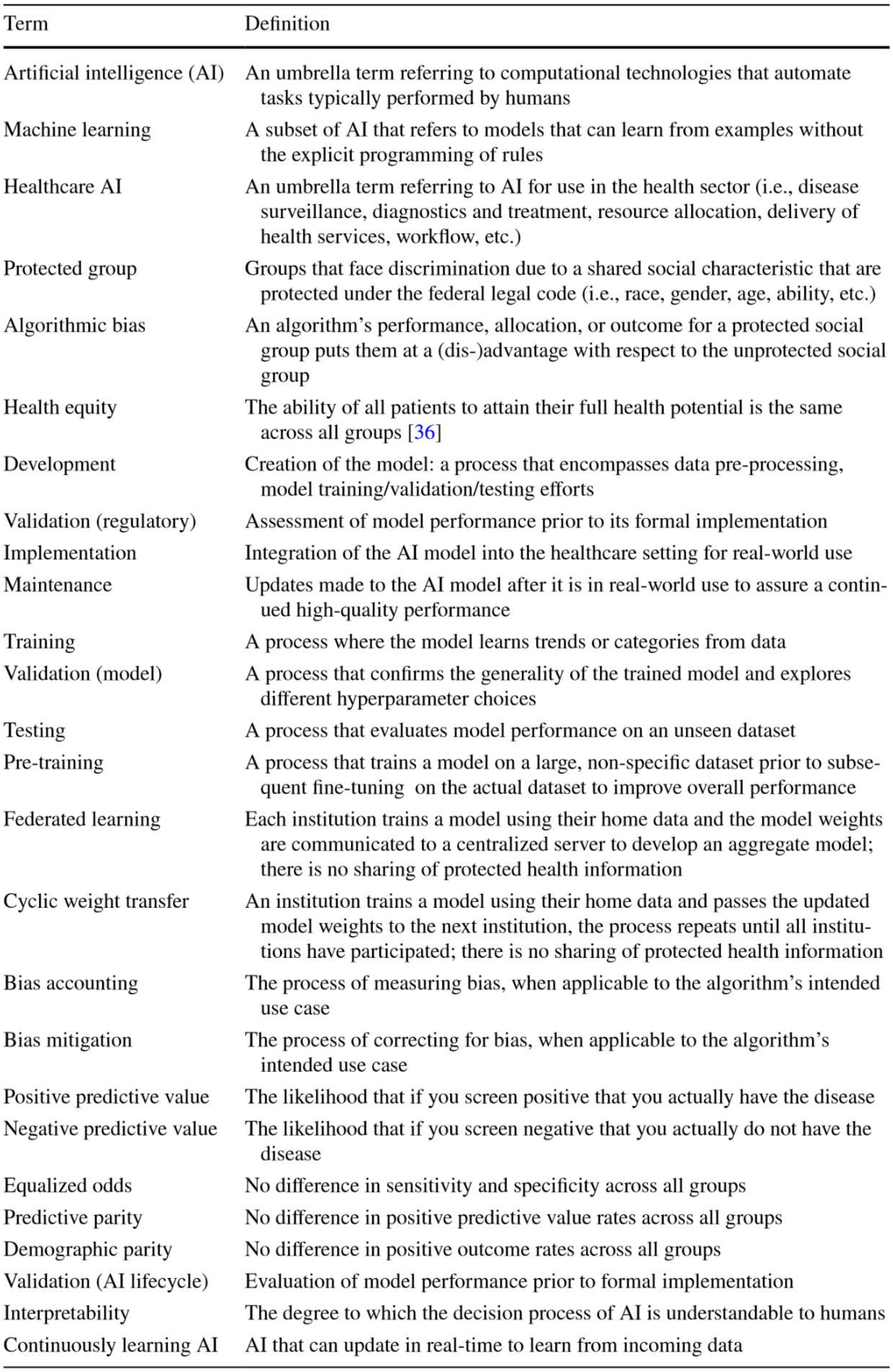
Ziad Obermeyer, Rebecca Nissan, Michael Stern, Stephanie Eaneff, Emily Joy Bembeneck, & Sendhil Mullainathan. (2021). *Algorithmic Bias Playbook* (pp. 1–21). Chicago Booth, Center for Applied Artificial Intelligence. https://www.chicagobooth.edu/-/media/project/chicago-booth/centers/caai/docs/algorithmic-bias-playbook-june-2021

Appendix: Definitional Tables from the Literature

“Group fairness metrics encountered in this review” (Huang et al., 2022)

Term	Description
AUROC[a]	Assesses overall classifier performance by measuring the TPR [True Positive Rate] and FPR [False Positive Rate] of a classifier at different thresholds.
Average odds	Compares the average of the TPR and FPR for the classification outcome between protected and unprotected groups.
Balanced accuracy	A measure of accuracy corrected for data imbalance, calculated as the average of sensitivity and specificity for a group.
Calibration	Assesses how well the risk score or probability predictions reflect actual outcomes.
Disparate impact	Measures deviation from statistical parity, calculated as the ratio of the rate of the positive outcome between protected and unprotected groups. Ideally, the disparate impact is 1.
Equal opportunity	For classification tasks in which one outcome is preferred over the other, equal opportunity is satisfied when the preferred outcome is predicted with equal accuracy between protected and unprotected groups. Ideally, the TPR or FNR[d] disparity between groups is 0.
Equalized odds	The TPR and FPR are equal between protected and unprotected groups.
Error rate	Compares the error rate of predictions, calculated as the number of incorrect predictions divided by the total number of predictions, between protected and unprotected groups. Ideally, the error rate disparity between groups is 0.
Statistical parity	Statistical parity (also known as demographic parity) is satisfied when the rate of positive outcomes is equal between protected and unprotected groups.

“Glossary of Key Terms” (Thomasian et al., 2021)



Term	Definition
Artificial intelligence (AI)	An umbrella term referring to computational technologies that automate tasks typically performed by humans
Machine learning	A subset of AI that refers to models that can learn from examples without the explicit programming of rules
Healthcare AI	An umbrella term referring to AI for use in the health sector (i.e., disease surveillance, diagnostics and treatment, resource allocation, delivery of health services, workflow, etc.)
Protected group	Groups that face discrimination due to a shared social characteristic that are protected under the federal legal code (i.e., race, gender, age, ability, etc.)
Algorithmic bias	An algorithm’s performance, allocation, or outcome for a protected social group puts them at a (dis-)advantage with respect to the unprotected social group
Health equity	The ability of all patients to attain their full health potential is the same across all groups [36]
Development	Creation of the model: a process that encompasses data pre-processing, model training/validation/testing efforts
Validation (regulatory)	Assessment of model performance prior to its formal implementation
Implementation	Integration of the AI model into the healthcare setting for real-world use
Maintenance	Updates made to the AI model after it is in real-world use to assure a continued high-quality performance
Training	A process where the model learns trends or categories from data
Validation (model)	A process that confirms the generality of the trained model and explores different hyperparameter choices
Testing	A process that evaluates model performance on an unseen dataset
Pre-training	A process that trains a model on a large, non-specific dataset prior to subsequent fine-tuning on the actual dataset to improve overall performance
Federated learning	Each institution trains a model using their home data and the model weights are communicated to a centralized server to develop an aggregate model; there is no sharing of protected health information
Cyclic weight transfer	An institution trains a model using their home data and passes the updated model weights to the next institution, the process repeats until all institutions have participated; there is no sharing of protected health information
Bias accounting	The process of measuring bias, when applicable to the algorithm’s intended use case
Bias mitigation	The process of correcting for bias, when applicable to the algorithm’s intended use case
Positive predictive value	The likelihood that if you screen positive that you actually have the disease
Negative predictive value	The likelihood that if you screen negative that you actually do not have the disease
Equalized odds	No difference in sensitivity and specificity across all groups
Predictive parity	No difference in positive predictive value rates across all groups
Demographic parity	No difference in positive outcome rates across all groups
Validation (AI lifecycle)	Evaluation of model performance prior to formal implementation
Interpretability	The degree to which the decision process of AI is understandable to humans
Continuously learning AI	AI that can update in real-time to learn from incoming data

“Candidate Criteria to Assess Algorithmic Fairness” (Paulus & Kent, 2020)

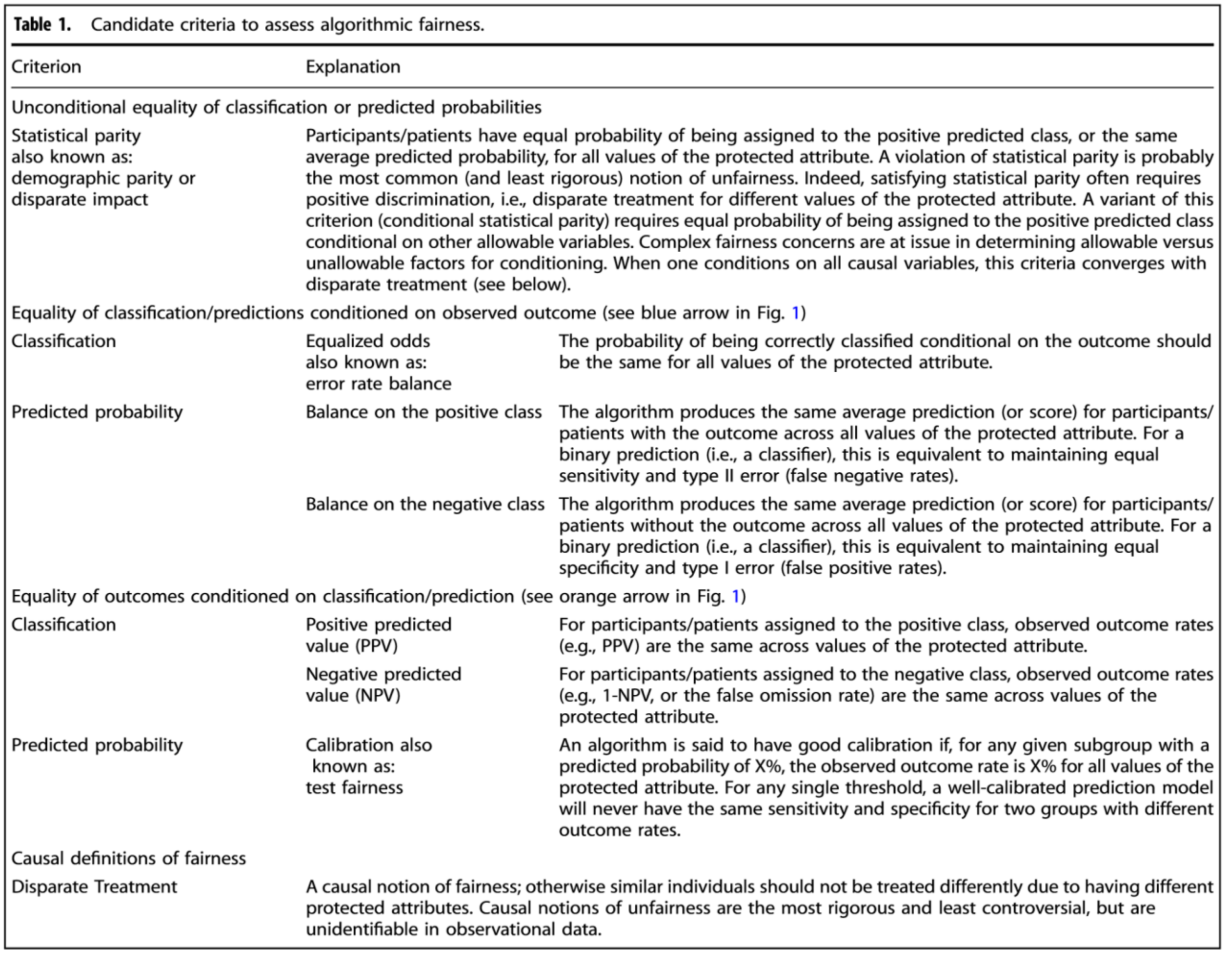


Table 1. Candidate criteria to assess algorithmic fairness.

Criterion	Explanation	
Unconditional equality of classification or predicted probabilities		
Statistical parity also known as: demographic parity or disparate impact	Participants/patients have equal probability of being assigned to the positive predicted class, or the same average predicted probability, for all values of the protected attribute. A violation of statistical parity is probably the most common (and least rigorous) notion of unfairness. Indeed, satisfying statistical parity often requires positive discrimination, i.e., disparate treatment for different values of the protected attribute. A variant of this criterion (conditional statistical parity) requires equal probability of being assigned to the positive predicted class conditional on other allowable variables. Complex fairness concerns are at issue in determining allowable versus unallowable factors for conditioning. When one conditions on all causal variables, this criteria converges with disparate treatment (see below).	
Equality of classification/predictions conditioned on observed outcome (see blue arrow in Fig. 1)		
Classification	Equalized odds also known as: error rate balance	The probability of being correctly classified conditional on the outcome should be the same for all values of the protected attribute.
Predicted probability	Balance on the positive class	The algorithm produces the same average prediction (or score) for participants/patients with the outcome across all values of the protected attribute. For a binary prediction (i.e., a classifier), this is equivalent to maintaining equal sensitivity and type II error (false negative rates).
	Balance on the negative class	The algorithm produces the same average prediction (or score) for participants/patients without the outcome across all values of the protected attribute. For a binary prediction (i.e., a classifier), this is equivalent to maintaining equal specificity and type I error (false positive rates).
Equality of outcomes conditioned on classification/prediction (see orange arrow in Fig. 1)		
Classification	Positive predicted value (PPV)	For participants/patients assigned to the positive class, observed outcome rates (e.g., PPV) are the same across values of the protected attribute.
	Negative predicted value (NPV)	For participants/patients assigned to the negative class, observed outcome rates (e.g., 1-NPV, or the false omission rate) are the same across values of the protected attribute.
Predicted probability	Calibration also known as: test fairness	An algorithm is said to have good calibration if, for any given subgroup with a predicted probability of X%, the observed outcome rate is X% for all values of the protected attribute. For any single threshold, a well-calibrated prediction model will never have the same sensitivity and specificity for two groups with different outcome rates.
Causal definitions of fairness		
Disparate Treatment	A causal notion of fairness; otherwise similar individuals should not be treated differently due to having different protected attributes. Causal notions of unfairness are the most rigorous and least controversial, but are unidentifiable in observational data.	