JAMIA Open, 2026, **9**(3), ooag081
https://doi.org/10.1093/jamiaopen/ooag081
Review

Strategies for mitigating artificial intelligence bias in healthcare: a systematic review

Kais Gadhoumi, PhD[1], Rashaud Senior, MD, MMCi[2,3], Sijia Wei, PhD, RN[4], Leila Ledbetter, MLIS[5], Michael D. Green[3], Bethany D. Bonner, PharmD, MMCi[3,6], Yvonne Mosley, MBA, CSSBB, CPHQ, CPPS[7], Katie Seidler, DPT, MSCI, MMCi[3,8], Kristle Green, PharmD, MMCi[8], Donghwan Lee, BSN, RN[1], Chuan Hong, PhD[3], Vincent Guilamo Ramos, PhD, MSN, MPH, MS, MSW, BS, RN[9], Michael P. Cary, Jr., PhD, RN[1,3,*]

[1]Duke University School of Nursing, Durham, NC, 27710, United States
[2]Avance Care, Raleigh, NC, United States
[3]Duke University School of Medicine, Durham, NC, 27710, United States
[4]Feinberg School of Medicine and Kellogg School of Management, Northwestern University, Chicago, IL, United States
[5]Evidence Synthesis Information Scientist at the Virginia Tech Blacksburg, VA, 24061, United States
[6]Triad HealthCare Network, Greensboro, NC, 27401, United States
[7]Belk College of Business, University of North Carolina at Charlotte, Charlotte, NC, 28223, United States
[8]Food and Drug Administration, Silver Spring, MD, 20993, United States
[9]Johns Hopkins School of Nursing, Baltimore, MD, 20001, United States
*Corresponding author: Michael P. Cary, Jr., PhD, RN, Duke University, 307 Trent Drive, Durham, NC 27707, United States (michael.cary@duke.edu)

Abstract

Objectives: **Artificial intelligence is used in healthcare to identify and manage health conditions across diverse patient populations and clinical settings, but biases in algorithms can perpetuate and exacerbate inequalities in health and healthcare delivery. While some research has focused on addressing this critical issue, there is a lack of comprehensive information on the types of strategies employed for bias mitigation in the use of artificial intelligence in healthcare and the effectiveness of these strategies. The objective of this review was to address this lack of information by identifying, categorizing, and describing the effectiveness of bias reduction strategies and fairness metrics in healthcare algorithms.**

Materials and Methods: **Following the Preferred Reporting Items for Systematic reviews and Meta-Analyses guidelines, this study categorizes and evaluates the effectiveness of bias reduction strategies and fairness metrics identified in a previous scoping review.**

Results: **The review included 35 studies. The findings are related to the stages of the algorithm lifecycle and are aligned with the authors' institutional governance structure for algorithm development, silent evaluation, effectiveness evaluation, and deployment. The majority (60%) of the identified strategies were implemented during the algorithm development and silent evaluation stages, and all studies utilized group fairness metrics for performance measurement. Most studies (85%) reported effective bias reduction, while only a few reported ineffectiveness (8%) or no effect (5%).**

Conclusion: **There is a significant opportunity for model developers and end users to identify and reduce bias, particularly during model design. When evaluating strategy effectiveness, efforts should be measured using evidenced-based fairness metrics—such as group-based metrics—to ensure effectiveness and interpretability.**

Key words: artificial intelligence, algorithms, delivery of healthcare, socioeconomic factors

Lay Summary

Artificial intelligence (AI) is increasingly used in healthcare to detect diseases, improve diagnosis, and support clinical decisions. However, algorithms can unintentionally reproduce existing inequalities if they contain bias. This review examined how researchers have tried to identify and reduce bias in healthcare algorithms and how effective those efforts have

Received: August 19, 2025. **Revised:** March 24, 2026. **Accepted:** May 7, 2026

been. Following accepted systematic review standards, the study analyzed 35 published investigations on bias reduction strategies and fairness measures. Most strategies (about 60%) were applied during the early stages of algorithm design and testing rather than after the models were introduced into real-world clinical settings. Every study evaluated bias using group fairness metrics, which compare performance by demographic variables such as age, sex, or race. While a large majority of studies reported success in reducing bias (85%), few described limited (8%) or no improvement (5%). The findings highlight the importance of identifying and mitigating bias early in the development process and of using clear, evidence-based fairness metrics to ensure more equitable and trustworthy algorithmic tools in healthcare.

Background and significance

Algorithms play a pivotal role in identifying and managing health conditions across diverse patient populations and clinical settings. However, research has highlighted the presence of biases in algorithms, leading to downstream issues such as differential disease detection or care provision among patient populations, often unbeknownst to users.[1–11] These users—encompassing researchers, data scientists, and clinicians—must be able to identify and reduce these biases because bias detection and mitigation are essential to ensuring equitable and trustworthy artificial intelligence (AI) systems in healthcare.[12–14] It is imperative to identify and address biases that may be inadvertently incorporated into algorithms because such biases can perpetuate and exacerbate existing inequalities in health and healthcare delivery. When algorithms inherit biases from their training data or design, they risk producing discriminatory outcomes, such as unequal access to care, misdiagnoses, or suboptimal treatment recommendations for marginalized groups. Moreover, these biases can undermine trust in AI technologies, create ethical and legal challenges, and ultimately hinder the equitable advancement of healthcare. Addressing these biases is not only a technical requirement but also a moral obligation to ensure AI systems promote fairness, improve patient outcomes across diverse populations, and contribute to reducing longstanding disparities in healthcare.

Given the profound potential of bias in AI to influence health outcomes, there have been efforts to develop strategies to identify and address such biases. For example, Section 1557 of the Patient Protection and Affordable Care Act issued in August 2022 prohibits discrimination on the basis of race, color, national origin, sex, age, or disability in healthcare.[15] Since its issuance, the United States Department of Health and Human Services has emphasized the need to address discrimination and potential bias in clinical algorithms.

Despite the contributions on bias mitigation in healthcare AI in the literature, there remain 2 notable gaps. First, there is a lack of comprehensive information on the types of strategies employed for bias mitigation. To address this gap, we previously conducted a scoping review that identified 45 scientifically tested tools and applications aimed at mitigating bias,[16] specifically focusing on racial and ethnic bias. Second, there is a dearth of knowledge regarding the effectiveness of these bias mitigation strategies. As outlined in its research protocol,[17] the Agency for Healthcare Research and Quality published a systematic review that describes several strategies used in the literature[18] while noting the difficulty of quantifying and extrapolating the results for broader use. Two other recent reviews conducted similar research—one review focused only on models using electronic health record data,[19] and the other focused only on changes in racial/ethnic disparities.[20]

In the present review, we build on these efforts by identifying, categorizing, and describing the effectiveness of the bias reduction strategies and fairness metrics in healthcare algorithms, as articulated in our previous work. Specifically, we dissect studies from our scoping review that applied and evaluated bias mitigation algorithms and strategies to healthcare analytics in different care settings to examine qualitative and quantitative aspects of bias reduction approaches and metrics, the type of bias they address, and the stage at which these approaches are applied in the lifecycle of the healthcare algorithm. Our aim is to identify and ultimately recommend bias reduction approaches that are effective against algorithmic discrimination.

Our review includes a broader array of models trained on heterogeneous data than those examined in previous reviews[21–24]; as such, our review incorporates more bias types and places greater emphasis on advancing equity, an ethical principle that reflects our shared values as a society and provides the foundation on which other principles (ie, transparency, safety, regulatory compliance) can be effectively implemented. We also propose OPTIMIZE-AI[2] (Optimizing Predictive Tools and Intelligence Models to Improve Zero Errors in AI [Accountability and Impact]), a framework to guide the development, evaluation, and deployment of equitable AI models in healthcare. To support their practical implementation, we relate these strategies to Duke Health's Algorithm-Based Clinical Decision Support (ABCDS) Oversight,[25] a framework that provides a structured approach for governing and evaluating clinical algorithms through 4 stages designed to guide development teams in the algorithmic lifecycle.

Materials and methods

Information sources

The databases searched include Medline (PubMed), Embase (Elsevier), and Web of Science Core Collection (Clarivate), ProQuest Computer Science Database, and ProQuest Dissertations & Theses Global.

Definitions

We adopt definitions consistent with current guidance for AI in clinical prediction research. Artificial intelligence is defined following the TRIPOD+AI statement as a field of computer science that focuses on developing models and algorithms capable of performing tasks that typically require human intelligence.[26] Algorithmic bias denotes systematic differences in the accuracy

or calibration of predictive models across groups defined by protected characteristics such as race, sex, or age, and algorithmic fairness refers to clinical decision-making processes that do not consistently favor or disadvantage members of one protected class over another.[27] To minimize ambiguity, we use "bias" to describe such systematic performance inequities rather than statistical estimation bias.

Search strategy

This review categorizes and evaluates the effectiveness of bias reduction strategies and fairness metrics; the authors followed the same search strategy as that utilized in a previously conducted scoping review.[16] The literature search was developed and conducted by an experienced medical librarian (LL) with input from the other authors and included a mix of keywords and subject headings: "algorithm," "bias," "mitigation/assessment," "healthcare," and "race/ethnicity."

The original search, which was conducted on August 24, 2022, was extended but limited to studies published up to November 30, 2022, to capture the state of AI bias in healthcare prior to the widespread adoption of foundational models and transformer-based architectures, sparked by the public release of ChatGPT in 2022 (OpenAI, San Francisco, CA, United States), which introduced a new paradigm in AI development and application, known as generative AI.[28,29] Empirically, analysis of Dimensions bibliometric data (Dimensions.ai, Digital Science, London, United Kingdom) shows that publications explicitly referencing "foundational model(s)," "large language model(s)," and related AI terminology increased more than 10-fold beginning in December 2022 (Figure S1).

The pre-foundational-model era reflects the state of bias mitigation prior to the proliferation of transformer-based architectures (eg, generative pretrained transformer and related models). Restricting our search to the pre-foundational-model era of healthcare AI enables a coherent synthesis of pre-generative AI mitigation strategies and establishes a foundation for subsequent reviews incorporating developments after 2022.

The full, reproducible search strategies for all included databases are provided in Supplementary Material S1.

Eligibility criteria and study selection process

We followed the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines and the same study selection process and inclusion criteria as those employed in the previous scoping review.[16,30,31] To evaluate the effectiveness of the mitigation strategies reported in empirical studies, we added new exclusion criteria to the present systematic review: (1) studies not focusing on clinical applications and (2) studies with insufficient details on the model(s) used. Titles/abstracts and full texts were independently screened by at least 2 reviewers to exclude studies clearly outside the scope of this review, including those not addressing racial or ethnic bias mitigation, not involving algorithmic applications in healthcare, or representing non-empirical publications (eg, editorials, commentaries, or conference abstracts). Artificial intelligence tools were considered to be related to healthcare if they were algorithmic systems designed to inform or guide patient care or diagnostic, prognostic, or population health management decisions within clinical or public health settings. Any discrepancies between reviewers were resolved through discussion with a third reviewer. Detailed exclusion categories and counts for full-text screening are shown in Figure 1.

Data collection and synthesis

Data extraction matrices were developed by M.P.C. and S.W. They included key components that helped define and contextualize the purpose, fairness metrics, and overall performance of an algorithm. All coauthors pilot-tested the extraction matrices on 2 studies. After team discussions, the matrices were revised, and an extraction manual was developed to ensure consistent and accurate coding for the full extraction phase. Two reviewers independently extracted data from each paper. Conflicts that were unresolvable between 2 reviewers were resolved by K.G Kais Gadhoumi.

Quality/risk of bias assessment

We used the Standard Quality Assessment Criteria for Evaluating Primary Research Papers from a Variety of Fields (QualSys) tool to compare quality across the included studies (Table S1).[32]

ABCDS framework

Duke Health's ABCDS Oversight is a framework designed for the governance and evaluation of clinical algorithms.[25] This framework ensures that these algorithms are innovative, safe, equitable, and high quality by introducing checkpoints throughout their development and deployment lifecycle. It involves 4 phases as follows. The Model Development phase necessarily includes clinical performance metrics for retrospective evaluation as well as the more traditional definition of model validation, for example, internal/external validation. The Silent Evaluation phase involves a prospective deployment of a model within a real-world clinical setting without any alerts generated for the clinical teams, which is used to evaluate model performance using real-life clinical data. The Effectiveness Evaluation phase involves small-scale deployment of the model for a subset of users, comparing its performance and user adoption to an existing standard. In the General Deployment phase, the model is deployed at a larger scale in typical clinical workflows with regular monitoring for performance drift or deviations.

Results

Summary of the included studies

Of the 18 028 studies identified from our scoping review,[16] 11 579 unique records were screened, of which 346 studies were selected for full-text retrieval and eligibility assessment. Nine additional studies were identified through targeted web searches and assessed against our eligibility criteria. After excluding 319 studies that failed to meet eligibility criteria, 35

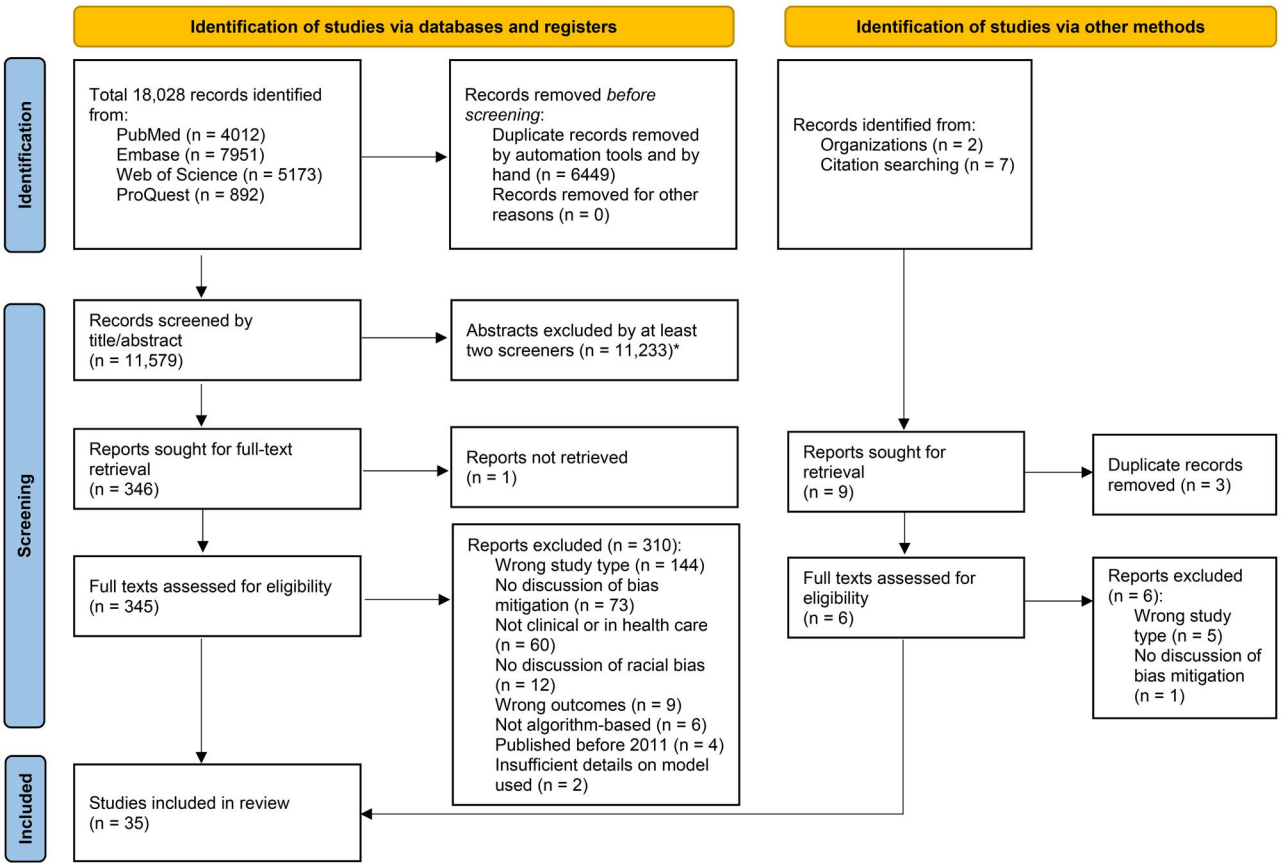

Figure 1. PRISMA diagram of the literature search and selection process. *Common reasons for exclusion at the title/abstract screening stage included (1) studies not related to bias mitigation in algorithmic models, (2) studies without clinical or population-level applications, (3) studies not in healthcare settings, and (4) studies lacking empirical data (eg, editorials, commentaries, conference abstracts).

studies were retained for inclusion. Figure 1 depicts the selection process and exclusion criteria flowchart following the PRISMA guidelines.

Table S2 provides a summary of the studies included. Of the 35 studies included in this review,[9,33–66] all had an observational retrospective cohort design, and all but 6 (83%) were based on data from US sites; non-American sites included Israel, the United Kingdom, South Korea, and the Netherlands.[37,43,49,52,56,59] Moreover, the included studies spanned various clinical settings across the care continuum, including inpatient, outpatient, labor and delivery, as well as academic and community acute care facilities. A few studies included in-person surveys[45,46,50,54]; they also encompassed many different clinical contexts and outcomes, such as inpatient mortality, pediatric postoperative mortality, cancer survival, pre-eclampsia, electrocardiogram analysis, and 30-day hospital readmission.

Clinical categories

The studies evaluated bias and bias mitigation across multiple clinical subject areas, mostly in cardiology (23%), oncology (11%), healthcare utilization (11%), and nephrology (9%). Fewer studies involved other specialties and clinical settings, such as obstetrics (6%), critical care (6%), orthopedics (3%), and infectious disease (3%). Figure 2A depicts the number of studies by clinical subject area.

Type of bias

The studies reported different types of bias. Racial and ethnic bias was the most frequently investigated form of bias in all studies ($n = 28$, 80%), followed by gender bias ($n = 8$, 23%) and age bias ($n = 6$, 17%). Most included studies used the terms "sex" and "gender" interchangeably in reference to bias without clearly defining or distinguishing between them. Other types of bias reported include colorism/shadeism, bias based on socioeconomic status, and bias toward underrepresented subpopulations. Figure 2B depicts the complete list of types of bias reported in the reviewed studies.

Categorization of bias mitigation strategies

A variety of approaches for identifying and mitigating bias in clinical algorithms were reported (Table 1, Table S3). A useful categorization of approaches proposed in the literature is based on the stage of algorithm development at which they are applied. Pre-processing, in-processing, and post-processing approaches tackle debiasing at the algorithm data input, development, and output stages, respectively. In our previous scoping review,[16] we proposed an additional category of "algorithmic design" specifically to implement the concept of *health equity by design*,[67] embedding a fairness lens in the

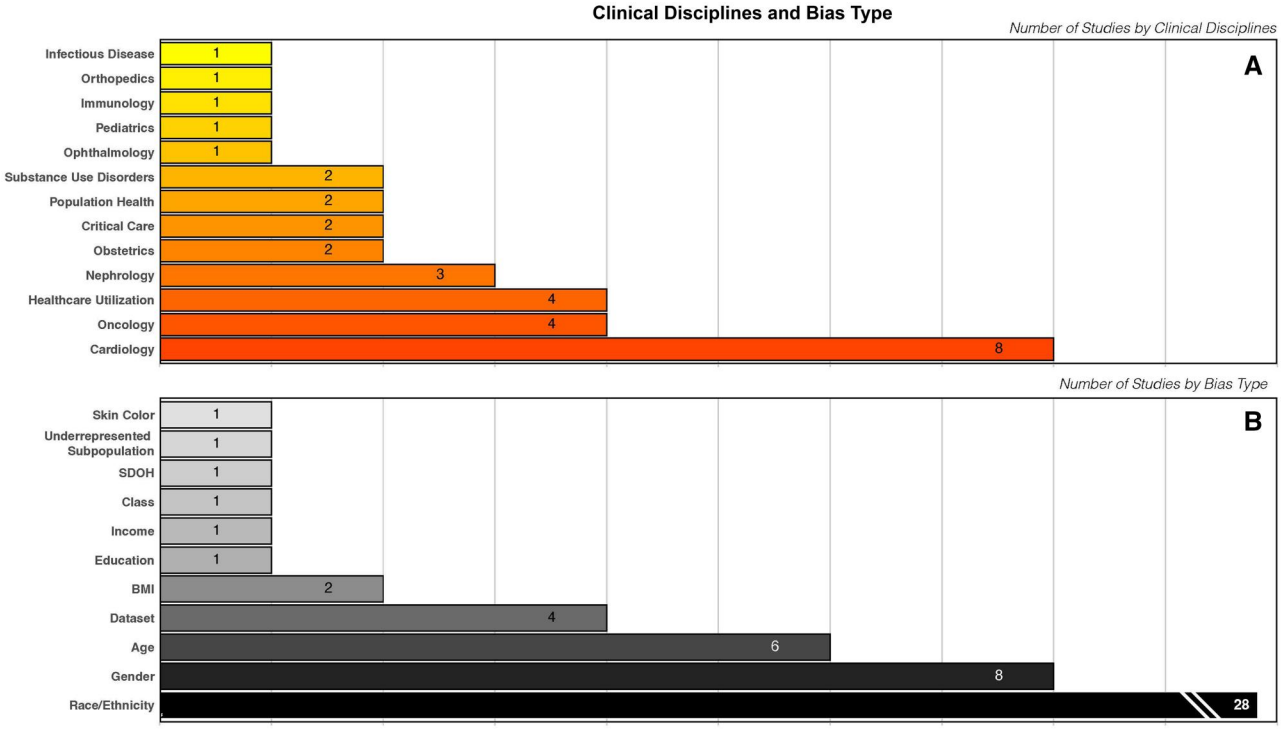

Figure 2. Distribution of studies by clinical subject area.

algorithm's conceptualization phase rather than in its development phase.

Because none of the included studies were evaluated in real-world clinical settings, all bias mitigation strategies and methods identified in this review focus on model development and fit within the Model Development phase of the ABCDS Oversight framework lifecycle. These strategies and methods are primarily implemented to enhance fairness and reduce bias during model creation and internal validation. Mitigation strategies applicable to subsequent phases of the ABCDS Oversight Framework, such as effectiveness evaluation, general deployment, and ongoing monitoring, were not observed in this review.

Algorithm design strategies

Twenty-one (60%) studies applied a bias mitigation strategy embedded in the design of the clinical algorithm. Several included the protected variables in the model,[46,50,57,62,66] a concept known as *fairness through awareness*,[68] while others excluded the protected variables,[39,43,48,55] also known as *fairness through unawareness*.[69] A few studies used an adjusted learning technique that modified the outcome variable[9,51,58] or stratified the protected variables.[34,35,38,41,42,47,53,59,62,63]

Pre-processing strategies

Ten (29%) studies applied strategies prior to model training. Most studies applied a data resampling or reweighing technique to correct sampling bias by rebalancing the proportions of subgroups.[34,36,40,52,55,57,58,60,61] One study used a disparate impact remover,[54] a data perturbation technique that modifies attribute values to increase group fairness while preserving the rank order within groups.[70]

In-processing strategies

These strategies operate on the optimization procedure of the algorithm to achieve fairness and accuracy. Typically, this involves the modification of the objective function of the model through constraints and regularization, adversarial learning, or adjusted learning.[71–73] Nine studies applied such strategies. Two studies[52,55] applied a regularization approach based on prejudice removal,[74] while one study added a bias-penalizing term to the loss function to achieve regularization and an ultimately fair outcome.[56] Two other studies applied an adversarial learning approach, one by introducing a new adversarial loss function to minimize bias from protected variables[33] and the other by using an adversarial neural network.[61] Finally, 4 studies employed an adjusted learning approach to minimize bias: through meta-learning,[57] through the modification of the objective function,[41] or through the use of transfer learning.[44,65]

Post-processing strategies

Post-processing approaches act on and adjust the model outcome on a per-group basis, primarily by either correcting the outcome or by correcting the model itself. Seven studies (20%) corrected the outcome by calibration of the model output,[37,38,42,45,49,57,64] and one study[61] corrected the model by using a calibrated equalized odds technique.[75]

Strategy evaluation: Fairness metrics

Numerous metrics were proposed to uncover bias, measure its effect, and quantify fairness improvement that debiasing strategies may accomplish.[76] Fairness metrics serve the dual purpose of identifying and assessing bias in models, on the one hand, and evaluating the effectiveness of bias mitigation strategies, on

Table 1. Bias mitigation approaches and techniques across the algorithmic lifecycle stages.

Type/stage	Approach	Technique	Studies
Algorithm design	Adjusted learning	Stratification of protected variables	Afrose et al.,[34] Akbilgic et al.,[35] Borgese et al.,[38] Do et al.,[41] Foryciarz et al.,[42] Howard et al.,[47] Noseworthy et al.,[53] Puyol-Antón et al.,[59] Segar et al.,[63] and Segar et al.[62]
		Modification of outcome variables	Lin et al.,[51] Obermeyer et al.,[9] and Pierson et al.[58]
	Fairness through unawareness	Exclusion of protected variables	Buckley et al.,[39] Gama et al.,[43] Huang et al.,[48] and Park et al.[55]
	Fairness through awareness	Inclusion of protected variables	Hammond et al.,[46] Landy et al.,[50] Pfohl et al.,[57] Segar et al.,[62] and Weissman et al.[66]
Preprocessing	Sampling/reweighing		Afrose et al.,[34] Allen et al.,[36] Burlina et al.,[40] Mosteiro et al.,[52] Park et al.,[55] Pfohl et al.,[57] Pierson et al.,[58] Radovanović et al.,[61] and Reeves et al.[60]
	Relabeling/ perturbation		Park et al.[54]
In-processing	Regularization/ constraints	Prejudice removal	Mosteiro et al.[52] and Park et al.[55]
		Regularization through added bias-penalizing term	Perez Alday et al.[56]
	Adversarial learning		Adeli et al.[33] and Radovanović et al.[61]
	Adjusted learning	Meta-learning	Puyol-Antón et al.[59]
		Modification of objective function	Do et al.[41]
		Transfer learning	Gao and Cui[44] and Toseef et al.[65]
Post-processing	Output correction	Calibration	Barda et al.,[37] Borgese et al.,[38] Foryciarz et al.,[42] Gianattasio et al.,[45] Joo et al.,[49] Pfohl et al.,[57] and Thompson et al.[64]
	Model correction	Calibrated equalized odds	Radovanović et al.[61]

the other. Fairness metrics extracted included demographic parity, equal opportunity, predictive parity, and calibration within groups. These metrics characterize equitable performance across demographic subgroups rather than overall model accuracy, emphasizing balanced utility for diverse patient populations. Although they differ on multiple levels, fairness metrics may be generally grouped into 2 main categories: group-based fairness metrics vs individual- and counterfactual-based fairness metrics[68,77] (Supplementary Material S1). Notably, all 35 studies used group fairness metrics to evaluate the performance of the mitigation strategies they employed, and none of the studies employed an individual fairness metric.

Conceptually, group fairness metrics (Table 2) can be divided into 4 main categories that apply different but complementary statistical criteria to assess fairness in a decision support model[78]—parity-based, confusion matrix-based, calibration-based, and score-based metrics.[79] Nine studies (26%) applied parity-based metrics[9,51,52,54,55,57,61,65,66]; 16 studies (46%) applied a version of an accuracy equality metric, which is a subtype of confusion matrix-based metric[38–41,43–45,47,48,50,53,54,56,58,62,64]; 8 studies (23%) used calibration as a fairness measure[9,37,38,42,48,49,57,63]; and only one study employed a score-based metric.[59] A list of metrics reported by each study is provided in Table 2.

Parity-based metrics examine the statistical independence of protected variables and compare predicted positive rates across protected/sensitive groups. Examples of parity-based metrics include statistical parity, demographic parity, and disparate impact. Eleven of 35 (31%) studies applied parity-based metrics.

Confusion-based metrics measure statistical separation between protected variables and compare protected/sensitive groups by considering potential underlying differences between groups. Examples of such metrics include equalized odds, equalized opportunity, and accuracy equality. Many studies (16 of 35; 46%) applied a version of an accuracy equality metric (eg, by measuring the false discovery or omission rate instead of accuracy) to detect and mitigate bias.

Calibration-based metrics consider the statistical sufficiency of protected variables and compare outcome probability scores between groups to evaluate fairness. This definition of fairness may be met if participants with protected characteristics and those with unprotected characteristics are both equally likely to belong to the positive classification. Calibration, test fairness, and well calibration are examples of calibration-based metrics.[80,81] Eight of the 35 studies (23%) used calibration to measure fairness.

Finally, score-based metrics compare protected groups based on expected scores. Examples of these metrics include statistical similarity, balance for positive and negative class, and Bayesian fairness. Only 1 of the 35 studies used a score-based metric.[59]

Strategy evaluation: Effectiveness

Figure 3 summarizes the reported effectiveness of the bias mitigation strategies identified across the included studies.

Table 2. Fairness metrics reported.

	Parity-based metrics	Confusion matrix-based metrics	Calibration-based metrics	Score-based metrics
Definition	Compare predicted positive rates across groups	Compare groups by considering potential underlying differences between groups	Compare groups based on predicted probability	Compare groups based on expected scores
Statistical property tested	Independence	Separation	Sufficiency	—
Examples (studies)	Statistical parity: Lin et al.,[51] Mosteiro et al.,[52] Obermeyer et al.,[9] Park et al.,[54] Pfohl et al.,[57] Radovanović et al.[61] Demographic parity: Toseef et al.,[65] Weissman et al.,[66] Disparate impact: Mosteiro et al.,[52] Park et al.,[54] Park et al.[55]	Accuracy equality: Borgese et al.,[38] Buckley et al.,[39] Burlina et al.,[40] Do et al.,[41] Gama et al.,[43] Gao and Cui,[44] Gianattasio et al.,[45] Howard et al.,[47] Huang et al.,[48] Landy et al.,[50] Noseworthy et al.,[53] Park et al.,[54] Perez Alday et al.,[56] Pierson et al.,[58] Segar et al.,[62] Thompson et al.[64] Equalized odds: Mosteiro et al.,[52] Radovanović et al.,[61] Reeves et al.[60] Equal opportunity: Adeli et al.,[33] Allen et al.,[36] Hammond et al.,[46] Lin et al.,[51] Park et al.,[55] Reeves et al.[60]	Test fairness, well calibration: Barda et al.,[37] Borgese et al.,[38] Foryciarz et al.,[42] Huang et al.,[48] Joo et al.,[49] Obermeyer et al.,[9] Pfohl et al.,[57] Segar et al.[63]	Balance for positive and negative class, Bayesian fairness, statistical similarity: Puyol-Antón et al.[59]

No studies in the current review used individual- or counterfactual-based fairness metrics.

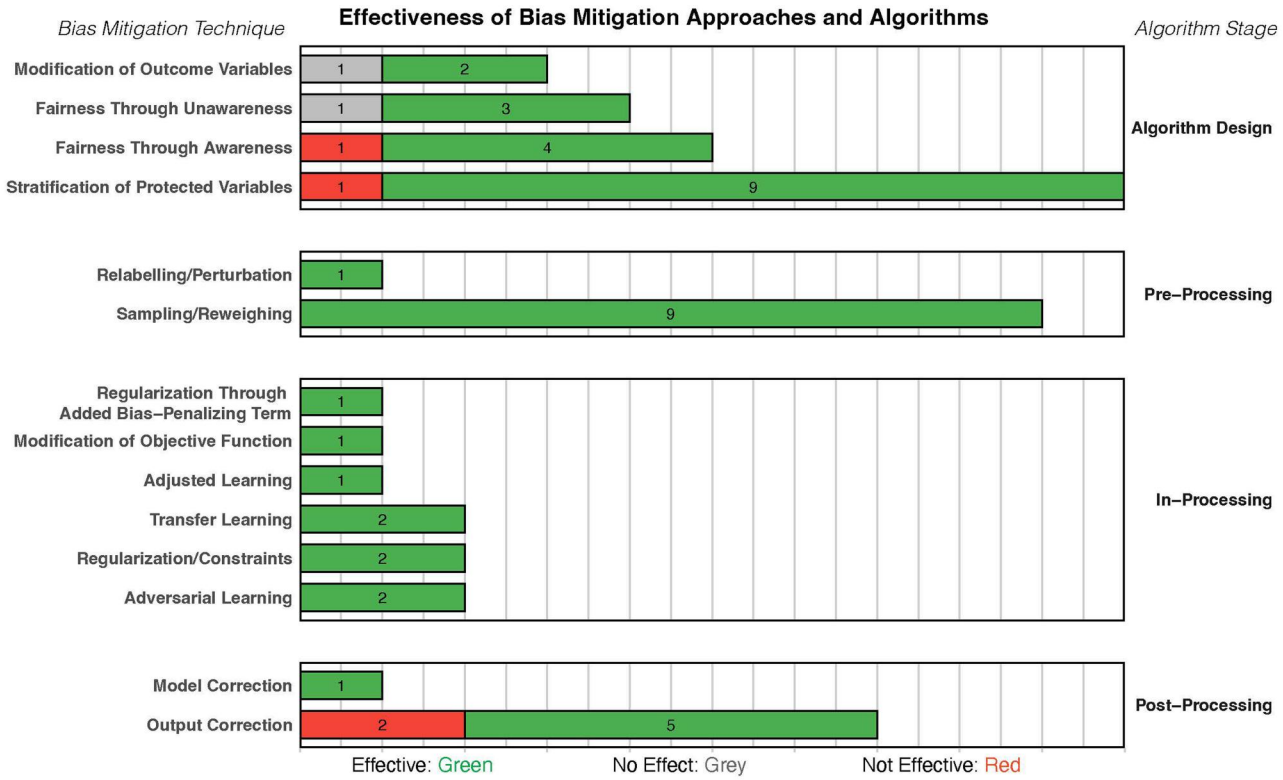

Figure 3. Bias mitigation strategies and their author-reported effectiveness.

Importantly, in this context, "effectiveness" refers to author-reported effectiveness as described in the original publications. None of the included studies employed standardized or comparable definitions, criteria, or thresholds to evaluate bias reduction, and therefore, direct comparisons between methods or strategies are not possible. Likewise, no study reported stakeholder engagement or consultation to establish what would constitute a clinically meaningful reduction in algorithmic bias.

Most studies (86%) reported strategy effectiveness in terms of reducing bias or fairness improvement (eg, reduced subgroup disparity, improved model performance across subgroups). Only 3 studies (9%) reported ineffectiveness of strategies used,[38,64,66] while 2 studies (6%) reported that no effect could be observed.[9,39] Interestingly, the ineffective strategies used either algorithmic design approaches (eg, stratification of protected variables and inclusion of protected variables) or post-processing approaches (eg, calibration). All pre- and in-processing strategies were reported to be effective.

The specific outcomes and metrics used to report bias reduction in each study are detailed in Table S3.

Discussion

In this study, we present a synthesis of 35 studies that explore various bias mitigation strategies and the application of fairness metrics designed to address bias in algorithms used for decision-making in healthcare. Our review includes a wide range of models trained on heterogeneous data to examine the contexts in which different strategies are utilized and their effectiveness, offering initial insight into potential relationships between the effectiveness of bias mitigation strategies and their clinical subject area. To support practical implementation, we aligned these strategies with the stages of Model Development, the first phase of the ABCDS Oversight framework lifecycle. Further, our review highlights the numerous decisions required to mitigate bias in clinical algorithms, such as selecting appropriate data pre-processing techniques, model optimization strategies, and post-processing adjustments tailored to specific contexts. The literature does not clearly define which clinical settings benefit most from each strategy, thereby highlighting the need for ongoing evaluation to better understand which approaches are most effective in different clinical scenarios.

The intersection of bias reduction and fairness metrics

A fairness metric aims to evaluate the equity of model performance across different population subgroups rather than the overall clinical accuracy or quality of predictions. A model that demonstrates comparable predictive performance (eg, sensitivity, specificity, or predictive value) across groups defined by protected characteristics such as race, sex, or age may not necessarily be considered "fair." All studies in this review assessed group fairness, typically by comparing model performance by demographic variables such as race, sex, or age. This predominant focus likely reflects the practical and regulatory alignment of group-level metrics, which are more established, interpretable, and feasible than individual or counterfactual fairness measures that require modeling unobservable patient-level scenarios. However, this emphasis may overlook inequities within groups and limit understanding of fairness at the individual level. Future work should extend fairness evaluation beyond group comparisons to incorporate individual and counterfactual perspectives for a more comprehensive and ethically grounded assessment of algorithmic equity in healthcare.

There is significant variability in the application of group-based fairness metrics, which highlights the complexity involved in selecting appropriate metrics for each context and the need for a deep understanding of both the metrics and the clinical environments they are applied in. Moreover, not all fairness metrics can be used with all bias mitigation strategies, as their compatibility depends on the type of model output, the stage at which mitigation is applied, and the specific fairness goals. Bias mitigation strategies, whether applied before, during, or after model training, are often designed to optimize specific fairness metrics, and improving one metric may compromise another. As a result, selecting appropriate combinations requires careful consideration of the clinical context and tradeoffs between fairness objectives.

No single bias mitigation approach has proven universally superior across all contexts.[82] The effectiveness of a given strategy often depends on factors such as the dataset characteristics, model type, clinical application, and the specific definitions of fairness being prioritized. For example, methods that improve statistical parity may inadvertently reduce predictive accuracy or introduce new disparities across other subgroups.[75,83,84] Similarly, fairness metrics can sometimes conflict with one another, making it challenging to optimize for all simultaneously.[57,85] These tradeoffs underscore the importance of context-aware decision-making and highlight the need for ongoing evaluation, stakeholder engagement, and transparency when selecting and implementing bias mitigation approaches in healthcare AI.

Bias reduction strategies lifecycle and effectiveness

Our categorization aligns model development processes (design, preprocessing, in-processing, and postprocessing) with the ABCDS lifecycle stages of Model Development. None of the included studies evaluated and mitigated bias in post-development phases (ie, Silent Evaluation, Effectiveness Evaluation, and General Deployment), as none had undergone prospective real-world testing. However, bias detection and mitigation should not end at the Model Development phase. Fairness assessment must be treated as an iterative and continuous process that extends throughout the entire lifecycle of algorithm-based clinical decision support systems. Current oversight and governance frameworks[25,86,87] highlight the necessity of revisiting and updating bias assessments as models transition to prospective evaluation and postdeployment phases, where new or unanticipated biases may emerge due to shifts in data, clinical practice, or patient populations. Consistent with the ABCDS Oversight Framework, bias should be assessed at multiple entry points within each stage, and mitigation strategies should be adapted according to the type, source, and impact of the bias identified. This underscores the importance of viewing fairness as a dynamic, lifecycle-wide responsibility rather than a discrete activity confined to the Model

Development phase. More broadly, the field must approach algorithmic tools with appropriate skepticism and demand rigorous, evidence-based prospective evaluation before integrating them into clinical practice. Postdeployment monitoring should only occur after robust predeployment validation to avoid repeating cycles of unrealistic expectations and unintended harm.[88]

Notably, while 85% of included studies reported positive or partially successful bias reduction, a small proportion (8% ineffective, 5% no effect) indicated limited or null results. This suggests potential publication bias favoring positive findings and underscores the importance of reporting neutral or negative outcomes to provide a full picture of mitigation efficacy. The few studies that achieved only partial bias reduction offer valuable insight into persistent challenges such as small sample sizes, limited subgroup representation, and lack of external validation. Recognizing these less successful efforts helps highlight methodological gaps in the current evidence base and underscores the need for more balanced reporting of both effective and ineffective bias mitigation approaches to strengthen transparency and reproducibility in future research. Moreover, assessing the magnitude of success in mitigating bias is notably challenging due to the absence of standardized methods and metrics. This diversity in methodologies and the lack of consensus on fairness metrics complicate the ability to compare the effectiveness of different strategies, thereby hindering a comprehensive understanding of their impact across various clinical contexts and algorithmic applications.

The interpretation of the "effectiveness" of mitigation strategies should be approached cautiously, as this term was based solely on author-reported results rather than standardized or externally validated criteria. The lack of uniform definitions and stakeholder involvement in evaluating what constitutes meaningful bias reduction limits the comparability and generalizability of these findings. Moving forward, developing consensus measures and reporting standards for assessing the effectiveness of bias-mitigation strategies will be critical to enable objective evaluation and translation of these approaches into clinical practice.

Recommendations

Our findings emphasize that the strategic selection of strategies and fairness metrics is crucial to the success of bias mitigation efforts, thus directly influencing the assessment and enhancement of equity throughout different stages of an algorithm's lifecycle. Practitioners should carefully consider the specific requirements and constraints of their application domain, as well as the potential tradeoffs between fairness and predictive performance. Importantly, bias-related concepts and terminologies should be employed with greater precision and transparency. Studies should adopt precise conceptual definitions and transparent methodological approaches to clearly differentiate biological sex from gender-related factors when assessing bias in healthcare algorithms.

To guide future development, deployment, and governance of AI models in healthcare, we propose the OPTIMIZE-AI^2 framework. This framework provides a holistic approach to developing, deploying, and governing AI models in healthcare. It emphasizes outcome-focused evaluation, consistent performance across diverse groups, transparency, inclusivity, bias mitigation, seamless workflow integration, a zero-harm goal, continuous monitoring, and strong accountability structures. Together, these principles aim to ensure that AI systems are safe, equitable, effective, and aligned with ethical and regulatory standards. A detailed description of the OPTIMIZE-AI^2 framework can be found in Supplementary Material S1.

Limitations

This systematic review has several limitations that may affect the applicability and generalizability of its findings. First, there may be inherent patterns associated with certain features or variables that were not extracted, thus possibly causing the review to overlook nuanced data elements crucial for comprehensive bias analysis. Additionally, while our review mainly focused on the primary reported bias mitigation strategies, a few studies may have included secondary interventions that were not discussed in detail, thereby potentially underrepresenting the scope of efforts undertaken. A significant limitation arises from the measurement of fairness in many studies, which did not utilize validated or widely accepted fairness metrics. Instead, general metrics such as the area under the receiver operating characteristics curve, sensitivity, and specificity were often employed. While these metrics are useful for certain assessments, they do not provide a robust measure of fairness and may not accurately reflect biases in clinical algorithms.

Last, the studies included in this review were restricted to those published up to November 2022. This cutoff was intentionally selected to capture the landscape of bias in AI technologies in healthcare prior to a pivotal shift in the field. Empirical analysis of bibliometric records from Dimensions.ai (Digital Science, London, United Kingdom) indicates a pronounced and sustained surge in scholarly output referencing "foundational model(s)," "large language model(s)," and related generative AI terminology (Figure S1). Specifically, beginning in December 2022, the volume of publications containing these terms increased by more than an order of magnitude, reflecting a marked surge in research activity and dissemination within this domain. Around this time, the emergence and rapid adoption of foundational models, particularly large transformer-based architectures, marked a significant evolution in both AI capabilities and deployment strategies. These models introduced new methodological complexities, data dependencies, and ethical considerations that distinguish them from earlier AI systems. Including studies beyond this point would have introduced a qualitatively different set of technologies and challenges, necessitating a distinct analytic framework. Rather than mixing fundamentally distinct generations of AI technologies, this review focuses on the pre-foundational model era to provide a coherent and analytically consistent synthesis. Studying the pre-foundational-model period isolates traditional bias sources (ie, representation imbalance and model transparency) and provides a baseline for future comparative analyses. Although newer studies are not captured, this temporal boundary enhances the interpretability of our findings and lays the groundwork for future reviews that will address the implications of this new generation of AI technologies in healthcare.

Conclusion

This systematic review highlighted the effectiveness of bias reduction strategies implemented during different stages of the algorithm lifecycle; the use of fairness metrics highlights their importance in promoting health equity and achieving the best possible health for all populations. For healthcare systems, these findings emphasize the need to critically evaluate the fairness of algorithms prior to their implementation and to continuously monitor their performance across diverse patient populations. Policymakers should consider developing guidelines for technology developers that mandate fairness assessments and bias mitigation strategies in the use of artificial intelligence in healthcare. Researchers are encouraged to include patients, families, and communities in codesign processes and employ multidisciplinary teams to develop fairness metrics and bias reduction techniques while continuously examining their impact on care delivery and outcomes. Developers, implementers, and regulators should interpret fairness evaluation as a continuous process integrated throughout the model lifecycle (from data collection to postdeployment surveillance) to maintain equitable and trustworthy AI in healthcare.

Author contributions

Kais Gadhoumi (Conceptualization, Formal analysis, Writing—original draft, Writing—review & editing), Rashaud Senior (Conceptualization, Formal analysis, Writing—original draft, Writing—review & editing), Sijia Wei (Conceptualization, Formal analysis, Writing—original draft, Writing—review & editing), Leila Ledbetter (Conceptualization, Formal analysis, Writing—original draft, Writing—review & editing), Michael D. Green (Formal analysis, Writing—review & editing), Bethany D. Bonner (Formal analysis, Writing—review & editing), Yvonne Mosley (Formal analysis, Writing—review & editing), Katie Seidler (Formal analysis, Writing—review & editing), Kristle Green (Formal analysis, Writing—review & editing), Donghwan Lee (Formal analysis, Writing—review & editing), Chuan Hong (Formal analysis, Writing—original draft, Writing—review & editing), Vincent Guilamo Ramos (Formal analysis, Writing—review & editing), and Michael P. Cary Jr. (Conceptualization, Formal analysis, Supervision, Writing—original draft, Writing—review & editing)

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Conflicts of interest

The authors have no competing interests to declare.

Funding

This study was partially funded by National Institutes of Health (NIH) awards, UL1TR002553 and R25NR021365. The funder played no role in the study design, data collection, analysis and interpretation of data, or writing of this manuscript. M.D.G. was supported by the National Institute On Aging of the NIH under award number F99AG088695. The contents of this article represent the views of the author(s) and do not necessarily represent the official views of the U.S. Food and Drug Administration, Department of Health and Human Services or the U.S. Government.

Data availability

The data underlying this article are available in the article and in Supplementary Material S1.

References

1. Blair IV, Steiner JF, Havranek EP. Unconscious (implicit) bias and health disparities: where do we go from here? *Perm J*. 2011;15:71-78. https://doi.org/10.7812/TPP/11.979
2. FitzGerald C, Hurst S. Implicit bias in healthcare professionals: a systematic review. *BMC Med Ethics*. 2017;18:19. https://doi.org/10.1186/s12910-017-0179-8
3. Gopal DP, Chetty U, O'Donnell P, et al. Implicit bias in healthcare: clinical practice, research, and decision making. *Future Healthc J*. 2021;8:40-48. https://doi.org/10.7861/fhj.2020-0233
4. Shah HS, Bohlen J. *Implicit Bias*. StatPearls Publishing; 2024. Accessed February 16, 2024. http://www.ncbi.nlm.nih.gov/books/NBK589697/
5. Greenwald AG, McGhee DE, Schwartz JLK. Measuring individual differences in implicit cognition: the implicit association test. *J Pers Soc Psychol*. 1998;74:1464-1480. https://doi.org/10.1037/0022-3514.74.6.1464
6. Charlesworth TES, Navon M, Rabinovich Y, et al. The project implicit international dataset: measuring implicit and explicit social group attitudes and stereotypes across 34 countries (2009-2019). *Behav Res Methods*. 2023;55:1413-1440. https://doi.org/10.3758/s13428-022-01851-2
7. Jimenez N, Seidel K, Martin LD, et al. Perioperative analgesic treatment in Latino and non-Latino pediatric patients. *J Health Care Poor Underserved*. 2010;21:229-236. https://doi.org/10.1353/hpu.0.0236
8. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. *N Engl J Med*. 2020;383:874-882. https://doi.org/10.1056/NEJMms2004740
9. Obermeyer Z, Powers B, Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science (1979).* 2019;366:447-453. https://doi.org/10.1126/science.aax2342
10. Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. *J Glob Health*. 2019;9:010318. https://doi.org/10.7189/jogh.09.020318
11. Seyyed-Kalantari L, Zhang H, McDermott MBA, et al. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med*. 2021;27:2176-2182. https://doi.org/10.1038/s41591-021-01595-0

12. National Institute for Health Care Management. Racial Bias in Health Care Artificial Intelligence 2021. Accessed January 11, 2024. https://nihcm.org/publications/artificial-intelligences-racial-bias-in-health-care
13. Grant C. Algorithms Are Making Decisions About Health Care, Which May Only Worsen Medical Racism. American Civil Liberties Union, News & Commentary. Updated October 3, 2022. Accessed January 11, 2024. https://www.aclu.org/news/privacy-technology/algorithms-in-health-care-may-worsen-medical-racism
14. Colón-Rodríguez CJ. Shedding Light on Healthcare Algorithmic and Artificial Intelligence Bias. U.S. Department of Health and Human Services, Office of Minority Health. Updated July 12, 2023. Accessed January 11, 2024. https://www.minorityhealth.hhs.gov/news/shedding-light-healthcare-algorithmic-and-artificial-intelligence-bias
15. Department of Health and Human Services. Nondiscrimination in health programs and activities. 2022; 87:47824-47920. Accessed January 11, 2024. https://www.federalregister.gov/documents/2022/08/04/2022-16217/nondiscrimination-in-health-programs-and-activities
16. Cary MP, Zink A, Wei S, et al. Mitigating racial and ethnic bias and advancing health equity in clinical algorithms: a scoping review. *Health Aff*. 2023;42:1359-1368. https://doi.org/10.1377/hlthaff.2023.00553
17. Agency for Healthcare Research and Quality (AHRQ). Research Protocol: Impact of Healthcare Algorithms on Racial and Ethnic Disparities in Health and Healthcare. Agency for Healthcare Research and Quality (AHRQ); Updated 2022. Accessed October 27, 2023. https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/protocol
18. Tipton K, Leas BF, Flores E, et al. Impact of healthcare algorithms on racial and ethnic disparities in health and healthcare. Agency for Healthcare Research and Quality (AHRQ). 2023. https://doi.org/10.23970/AHRQEPCCER268
19. Chen F, Wang L, Hong J, et al. Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. *J Am Med Inform Assoc*. 2024;31:1172-1183. https://doi.org/10.1093/jamia/ocae060
20. Siddique SM, Tipton K, Leas B, et al. The impact of health care algorithms on racial and ethnic disparities. *Ann Intern Med*. 2024;177:484-496. https://doi.org/10.7326/M23-2960
21. Huang J, Galal G, Etemadi M, et al. Evaluation and mitigation of racial bias in clinical machine learning models: scoping review. *JMIR Med Inform.* 2022;10:e36388. https://doi.org/10.2196/36388
22. Daneshjou R, Smith MP, Sun MD, et al. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. *JAMA Dermatol.* 2021;157:1362-1369. https://doi.org/10.1001/jamadermatol.2021.3129
23. Bear Don't Walk OJ, Reyes Nieva H, Lee SS, et al. A scoping review of ethics considerations in clinical natural language processing. *JAMIA Open.* 2022;5:ooac039. https://doi.org/10.1093/jamiaopen/ooac039
24. Kaur D, Uslu S, Rittichier KJ, et al. Trustworthy artificial intelligence: a review. *ACM Comput Surv*. 2023;55:1-38. https://doi.org/10.1145/3491209
25. Bedoya AD, Economou-Zavlanos NJ, Goldstein BA, et al. A framework for the oversight and local deployment of safe and high-quality prediction models. *J Am Med Inform Assoc.* 2022;29:1631-1636. https://doi.org/10.1093/jamia/ocac078
26. Collins G, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024; 385:e078378. https://doi.org/10.1136/bmj-2023-078378
27. Ladin K, Cuddeback J, Duru OK, et al. Guidance for unbiased predictive information for healthcare decision-making and equity (GUIDE): considerations when race may be a prognostic factor. *NPJ Digit Med.* 2024;7:290. https://doi.org/10.1038/s41746-024-01245-y
28. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med (Lond)*. 2023;3:141. https://doi.org/10.1038/s43856-023-00370-1
29. Busch PA, Hausvik G, Nielsen J. The early wave of ChatGPT research: a review and future agenda. *Comput Hum Behav*. 2025;6:100213. https://doi.org/10.1016/j.chbah.2025.100213
30. Page MJ, Moher D, Bossuyt PM, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ.* 2021;372:n160. https://doi.org/10.1136/bmj.n160
31. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med*. 2018;169:467-473. https://doi.org/10.7326/M18-0850
32. Kmet LM, Lee RC, Cook LS. *Standard Quality Assessment Criteria for Evaluating Primary Research Papers from a Variety of Fields*. Alberta Heritage Foundation for Medical Research; 2004.
33. Adeli E, Zhao Q, Pfefferbaum A, et al. Representation learning with statistical independence to mitigate bias. *IEEE Winter Conf Appl Comput Vis*. 2021;2021:2512-2522. https://doi.org/10.1109/wacv48630.2021.00256
34. Afrose S, Song W, Nemeroff CB, et al. Subpopulation-specific machine learning prognosis for underrepresented patients with double prioritized bias correction. *Commun Med (Lond)*. 2022;2:111. https://doi.org/10.1038/s43856-022-00165-w
35. Akbilgic O, Langham MR, Davis RL. Race, preoperative risk factors, and death after surgery. *Pediatrics.* 2018;141:e20172221. https://doi.org/10.1542/peds.2017-2221
36. Allen A, Mataraso S, Siefkas A, et al. A racially unbiased, machine learning approach to prediction of mortality: algorithm development study. *JMIR Public Health Surveill*. 2020;6:e22400. https://doi.org/10.2196/22400
37. Barda N, Yona G, Rothblum GN, et al. Addressing bias in prediction models by improving subpopulation calibration. *J Am Med Inform Assoc*. 2021;28:549-558. https://doi.org/10.1093/jamia/ocaa283
38. Borgese M, Joyce C, Anderson EE, et al. Bias assessment and correction in machine learning algorithms: a use-case in a natural language processing algorithm to identify hospitalized patients with unhealthy alcohol use. *AMIA Annu Symp Proc*. 2022;2021:247-254.
39. Buckley A, Sestito S, Ogundipe T, et al. Racial and ethnic disparities among women undergoing a trial of labor after cesarean delivery: performance of the VBAC calculator with and without patients' race/ethnicity. *Reprod Sci*. 2022;29:2030-2038. https://doi.org/10.1007/s43032-022-00959-2

40. Burlina P, Joshi N, Paul W, et al. Addressing artificial intelligence bias in retinal diagnostics. *Transl Vis Sci Technol*. 2021; 10:13. https://doi.org/10.1167/tvst.10.2.13
41. Do H, Nandi S, Putzel P, et al. A joint fairness model with applications to risk predictions for underrepresented populations. *Biometrics.* 2023;79:826-840. https://doi.org/10.1111/biom.13632
42. Foryciarz A, Pfohl SR, Patel B, et al. Evaluating algorithmic fairness in the presence of clinical guidelines: the case of atherosclerotic cardiovascular disease risk estimation. *BMJ Health Care Inform*. 2022;29:e100460. https://doi.org/10.1136/bmjhci-2021-100460
43. Gama RM, Clery A, Griffiths K, et al. Estimated glomerular filtration rate equations in people of self-reported black ethnicity in the United Kingdom: inappropriate adjustment for ethnicity may lead to reduced access to care. *PLoS One.* 2021; 16:e0255869. https://doi.org/10.1371/journal.pone.0255869
44. Gao Y, Cui Y. Deep transfer learning for reducing health care disparities arising from biomedical data inequality. *Nat Commun*. 2020;11:5131. https://doi.org/10.1038/s41467-020-18918-3
45. Gianattasio KZ, Ciarleglio A, Power MC. Development of algorithmic dementia ascertainment for racial/ethnic disparities research in the US Health and Retirement Study. *Epidemiology (Fairfax).* 2020;31:126-133. https://doi.org/10.1097/EDE.0000000000001101
46. Hammond G, Johnston K, Huang K, et al. Social determinants of health improve predictive accuracy of clinical risk models for cardiovascular hospitalization, annual cost, and death. *Circ Cardiovasc Qual Outcomes*. 2020;13:e006752. https://doi.org/10.1161/CIRCOUTCOMES.120.006752
47. Howard FM, Dolezal J, Kochanny S, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat Commun*. 2021;12:4423. https://doi.org/10.1038/s41467-021-24698-1
48. Huang C, Murugiah K, Li X, et al. Effect of the new glomerular filtration rate estimation equation on risk predicting models for acute kidney injury after percutaneous coronary intervention. *Circ Cardiovasc Interv*. 2023;16:e012831. https://doi.org/10.1161/CIRCINTERVENTIONS.122.012831
49. Joo YS, Kim HW, Baek CH, et al. External validation of the international prediction tool in Korean patients with immunoglobulin a nephropathy. *Kidney Res Clin Pract*. 2022;41: 556-566. https://doi.org/10.23876/j.krcp.22.006
50. Landy R, Young CD, Skarzynski M, et al. Using prediction models to reduce persistent racial and ethnic disparities in the draft 2020 USPSTF lung cancer screening guidelines. *J Natl Cancer Inst*. 2021;113:1590-1594. https://doi.org/10.1093/jnci/djaa211
51. Lin YC, Mallia D, Clark-Sevilla AO, et al. Preeclampsia predictor with machine learning: a comprehensive and bias-free machine learning pipeline. medRxiv 22276107. https://doi.org/10.1101/2022.06.08.22276107, June 9, 2022, preprint: not peer reviewed.
52. Mosteiro P, Kuiper J, Masthoff J, et al. Bias discovery in machine learning models for mental health. *Information*. 2022; 13:237. https://doi.org/10.3390/info13050237
53. Noseworthy PA, Attia ZI, Brewer LC, et al. Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ECG analysis. *Circ Arrhythm Electrophysiol*. 2020;13:e007988. https://doi.org/10.1161/CIRCEP.119.007988
54. Park J, Arunachalam R, Silenzio V, et al. Fairness in mobile phone–based mental health assessment algorithms: exploratory study. *JMIR Form Res*. 2022;6:e34366. https://doi.org/10.2196/34366
55. Park Y, Hu J, Singh M, et al. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. *JAMA Netw Open.* 2021;4:e213909. https://doi.org/10.1001/jamanetworkopen.2021.3909
56. Perez Alday EA, Rad AB, Reyna MA, et al. Age, sex, and race bias in automated arrhythmia detectors. *J Electrocardiol*. 2022;74:5-9. https://doi.org/10.1016/j.jelectrocard.2022.07.007
57. Pfohl SR, Foryciarz A, Shah NH. An empirical characterization of fair machine learning for clinical risk prediction. *J Biomed Inform*. 2021;113:103621. https://doi.org/10.1016/j.jbi.2020.103621
58. Pierson E, Cutler DM, Leskovec J, et al. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat Med*. 2021;27:136-140. https://doi.org/10.1038/s41591-020-01192-7
59. Puyol-Antón E, Ruijsink B, Piechnik S, et al. Fairness in cardiac MR image analysis: an investigation of bias due to data imbalance in deep learning based segmentation. In: M de Bruijne, et al., eds. *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021:413-423*. Springer. 2021. https://doi.org/10.1007/978-3-030-87199-4_39
60. Reeves M, Bhat HS, Goldman-Mellor S. Resampling to address inequities in predictive modeling of suicide deaths. *BMJ Health Care Inform*. 2022;29:e100456. https://doi.org/10.1136/bmjhci-2021-100456
61. Radovanović S, Petrović A, Delibašić B, et al. Making hospital readmission classifier fair—what is the cost? In: Central European Conference on Information and Intelligent Systems (CECIIS 2019). 2019:325-331.
62. Segar MW, Hall JL, Jhund PS, et al. Machine learning-based models incorporating social determinants of health vs traditional models for predicting in-hospital mortality in patients with heart failure. *JAMA Cardiol.* 2022;7:844-854. https://doi.org/10.1001/jamacardio.2022.1900
63. Segar MW, Jaeger BC, Patel KV, et al. Development and validation of machine learning-based race-specific models to predict 10-year risk of heart failure: a multicohort analysis. *Circulation.* 2021;143:2370-2383. https://doi.org/10.1161/CIRCULATIONAHA.120.053134
64. Thompson HM, Sharma B, Bhalla S, et al. Bias and fairness assessment of a natural language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across racial subgroups. *J Amer Inform Assoc*. 2021;28:2393-2403. https://doi.org/10.1093/jamia/ocab148
65. Toseef M, Li X, Wong KC. Reducing healthcare disparities using multiple multiethnic data distributions with fine-tuning of transfer learning. *Brief Bioinform*. 2022;23:bbac078. https://doi.org/10.1093/bib/bbac078
66. Weissman GE, Teeple S, Eneanya ND, et al. Effects of neighborhood-level data on performance and algorithmic equity of a model that predicts 30-day heart failure readmissions

at an urban academic medical center. *J Card Fail*. 2021;27:965-973. https://doi.org/10.1016/j.cardfail.2021.04.021
67. Argentieri R, Mason T, Hefcart J, et al. Embracing health equity by design. Health IT Buzz. 2022. Updated February 22. Accessed March 22, 2024. https://www.healthit.gov/buzz-blog/health-it/embracing-health-equity-by-design
68. Dwork C, Hardt M, Pitassi T, et al. Fairness through awareness. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS 12)*. 2012:214-226. https://doi.org/10.48550/ARXIV.1104.3913
69. Chen J, Kallus N, Mao X, et al. Fairness under unawareness: assessing disparity when protected class is unobserved. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019:339-348. https://doi.org/10.1145/3287560.3287594
70. Feldman M, Friedler S, Moeller J, et al. Certifying and removing disparate impact. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2014*. Published online July 15, 2015. Accessed April 26, 2024. http://arxiv.org/abs/1412.3756
71. Celis LE, Huang L, Keswani V, et al. Classification with fairness constraints: a meta-algorithm with provable guarantees. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency Association for Computing Machinery*. Updated April 15, 2020. Accessed April 26, 2024. http://arxiv.org/abs/1806.06055
72. Zhang BH, Lemoine B, Mitchell M. Mitigating unwanted biases with adversarial learning. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society Association for Computing Machinery*. Updated January 22, 2018. Accessed April 26, 2024. http://arxiv.org/abs/1801.07593
73. Kilbertus N, Gascón A, Kusner MJ, et al. Blind justice: fairness with encrypted sensitive attributes. In: *Proceedings of the 35th International Conference on Machine Learning*. Updated June 8, 2018. https://doi.org/10.48550/arXiv.1806.03281
74. Kamishima T, Akaho S, Asoh H, et al. Fairness-aware classifier with prejudice remover regularizer. In: PA Flach, T De Bie, N Cristianini, eds. *Machine Learning and Knowledge Discovery in Databases. ECML PKDD*. Springer; 2012:35-50. https://doi.org/10.1007/978-3-642-33486-3_3
75. Pleiss G, Raghavan M, Wu F, et al. On fairness and calibration. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems Curran Associates Inc*. Updated November 3, 2017. Accessed April 26, 2024. http://arxiv.org/abs/1709.02012.
76. Hort M, Chen Z, Zhang JM, et al. Bias mitigation for machine learning classifiers: a comprehensive survey. *ACM J Responsib Comput*. 2024;1:1-52. https://doi.org/10.48550/ARXIV.2207.07068
77. Kusner MJ, Loftus JR, Russell C, et al. Counterfactual fairness. In: *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems.* 2017:4069-4079. https://doi.org/10.48550/ARXIV.1703.06856
78. Barocas S, Hardt M, Narayanan A. Fairness and machine learning: limitations and opportunities MIT Press; 2023. https://books.google.com/books? id=HuGwEAAAQBAJ
79. Caton S, Haas C. Fairness in machine learning: a survey. *CM Comput Surv*. 2024;56:1-38. https://doi.org/10.48550/ARXIV.2010.04053
80. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2016.
81. Kleinberg JM, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores. Information Technology Convergence and Services. arXiv, arXiv:1609.05807, 2016, preprint: not peer reviewed.
82. Long R. Fairness in machine learning: against false positive rate equality as a measure of fairness. *J Moral Philos.* 2021;19:49-78.
83. Chen Z, Zhang JM, Sarro F, et al. A comprehensive empirical study of bias mitigation methods for machine learning classifiers. *ACM Trans Softw Eng Methodol.* 2023;32:1-30. https://doi.org/10.1145/3583561
84. Biswas A, Barman S, Deshpande A, et al. Quantifying inframarginality and its trade-off with group fairness. CoRR, 2019; abs/1909.00982.
85. Wang X, Yang CC. Balancing fairness and performance in healthcare AI: a gradient reconciliation approach. arXiv, arXiv:2504.14388, 2025https://doi.org/10.48550/arXiv.2504.14388, preprint: not peer reviewed.
86. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med.* 2019;25:1337-1340. https://doi.org/10.1038/s41591-019-0548-6
87. Matheny ME, Thadaney Israni S, Whicher D. *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril*. National Academy of Medicine; 2020.
88. Rose S. AI hype cycles and reality in health care. *JAMA Health Forum.* 2025;6:e251904. https://doi.org/10.1001/jamahealthforum.2025.1904

JAMIA Open, 2026, 9, 1–13
https://doi.org/10.1093/jamiaopen/ooag081
Review