Advancing Equity in Healthcare AI: Mitigating Bias and Reducing Opacity

Yik Ting Agnes Chiu[1]

Received March 27, 2026
Accepted July 22, 2026
Electronic access September 15, 2026

The rise of Artificial Intelligence (AI) usage in the healthcare sector ranges from supporting clinical decisions to managing workflow and assisting on medical imaging. Yet, AI's ethical value should also be examined through whether it reproduces existing inequities on top of model performance. This review examines algorithmic bias and opacity within healthcare AI and reviews mitigation strategies that current literature supports. A qualitative literature review was conducted using peer-reviewed sources published between 2015 and 2025. These peer-reviewed sources were identified in PubMed, IEEE Xplore, and Google Scholar. A methodical strategy was used to screen the identified literature. Included studies were synthesized thematically to identify recurring concepts, representative case studies, and mitigation measures relevant to healthcare equity. Three dimensions of algorithmic bias were identified: cognitive and societal bias, data and measurement bias, and model and systemic bias. Algorithmic opacity interacts with these dimensions of bias instead of belonging to one of them. Mitigation measures identified in the literature address specific points of failure. However, the measures are dependent on the context and may not be evenly validated. This review's findings support a lifecycle-oriented perspective to examine ethical healthcare AI. A lifecycle-oriented view allows us to approach healthcare equity by identifying where bias enters the cycle, ensuring sufficient transparency in the system, and preserving accountability in governance. These frameworks and mitigation measures should be validated in more diverse clinical and geographical settings in future research.

Keywords: Artificial Intelligence (AI), Healthcare Equity, Algorithmic Bias, Algorithmic Opacity, AI Governance, Explainability, Ethical AI

Introduction

An increasing range of healthcare activities now embeds Artificial Intelligence, such as risk prediction, triage support, workflow management, patient communication, medical imaging and clinical decision support[1]. More than half of approved AI/ML-based medical devices between 2015 and 2020 were used for radiology, with 129 (58%) AI/ML devices in the USA and 126 (53%) in Europe targeting imaging applications[2]. There is a positive outlook on how AI can improve efficiency levels and enable earlier detection and a broader scale. This sentiment is especially strong in healthcare systems that suffer from resource constraints, workforce shortages, or unequal access to specialist expertise[1,3]. However, according to a 2025 survey across many countries, just 19% of institutions claimed a high degree of success for AI usage in clinical diagnosis despite much discussion and trial initiatives since around 2016[4].

It is important to recognize that high performance accuracy at a macro level may still replicate the racial, socioeconomic, linguistic, geographic, or institutional inequities in society. Maurud et al. found that although AI in clinical research informatics could reveal existing inequity in care, it could also pose a risk to health equity through biased design and use[5]. A similar conclusion is presented in Hussain et al.'s study where healthcare algorithms can exacerbate ethnic and racial disparities when faced with vulnerable or weak governance and stewardship[6]. For those reasons, the ethical question moves beyond whether healthcare should use AI and towards the conditions under which AI can be developed and deployed without worsening existing disparities.

Because algorithmic outputs assist in significant medical decisions involving diagnosis, eligibility, prioritization, and treatment, algorithms can impact the patients' access to care, relationships with clinicians, and they can distribute risks unevenly across already disadvantaged groups. This is where the significance of healthcare AI lies.

This review focuses on algorithmic bias and opacity's relationship and their implications for healthcare equity. Algorithmic bias is defined as the production of unfair or unequal outcomes for different groups by AI systems. Algorithmic opacity refers to the barriers to understanding, contesting, and governing how AI systems reach their output. Healthcare eq-

[1] *Canadian International School of Hong Kong, Hong Kong*

uity means the equal and fair conditions under which patients are not systematically disadvantaged in achieving health outcomes.

This review aims to avoid treating bias as a merely technical flaw and assessing opacity only on a surface explainability level. This aim is achieved by examining algorithmic harm in healthcare through an AI lifecycle-oriented view. This approach is then linked to questions of trust, governance, and human oversight.

Three questions are examined. First, how can algorithmic bias in healthcare AI be categorized that remains useful across the stages of AI lifecycle? Second, how do algorithmic opacity affect trust and accountability in healthcare equity? Third, what trade-offs are associated with each mitigation measure?

A qualitative literature review was conducted. Peer-reviewed studies from three databases, published between 2015 and 2025, were examined. No primary data was collected, and no specific algorithmic models were included. A thematic synthesis focusing on three dimensions of algorithmic bias and algorithmic opacity is presented. Afterwards, mitigation measures are evaluated. A discussion is presented at the end to explore implications, limitations, and unresolved questions for research.

Methodology

Search Strategy and Eligibility Criteria

The systematic search of this review began by examining how algorithmic bias and opacity arise in the healthcare AI lifecycle. Then, potential mitigation measures to lessen inequity were studied. The literature search was conducted on three scholarly literature search tools: PubMed was used mainly for biomedical and public health literature; IEEE Xplore was used for technical studies in AI and health data science; and Google Scholar was used to retrieve interdisciplinary studies across ethics, policy, and conceptual analyses.

Backward reference tracing was conducted on relevant studies. As each database's syntax is different, exact search strings were adapted (see Table 1).

Note that the earliest result returned from the search only started from 2022 for IEEE Xplore. Citations and patents were excluded from the Google Scholar search.

Studies were included if: (1) AI or other related computational systems in the context of healthcare or health-informatics are addressed; (2) Algorithmic bias, fairness, opacity, governance, or mitigation pertinent to healthcare equity is examined; (3) Explanation of inequity formation from algorithmic bias, healthcare empirical case, governance or implementation analysis, or a mitigation measure relevant to healthcare is contributed; (4) Sufficient detail in empirical studies was provided to understand the study design and relevance of the findings.

This review excluded studies outside healthcare. Within healthcare, studies that only examined model accuracy and architecture, or only ethics in passing were omitted. Duplicates or sources with insufficient methodological information to support appraisal were not used. Commentaries and non-peer-reviewed webpages were used only for general context.

Records retrieved from the three databases were screened in the following stages: title screening, abstract screening, and full-text evaluation. Duplicates were verified by matching author(s), year, and title before removal.

This review also draws on four additional peer-reviewed studies located through targeted searches outside the formal database query because they offer highly illustrative analyses of bias and equity in healthcare AI. They comprise work on race-adjusted eGFR and kidney transplantation wait-listing (Nissaisorakarn et al.), a systematic review of bias in neuroimaging-based AI models for psychiatric diagnosis (Chen et al.), an empirical evaluation of racial bias in a widely used healthcare cost–based risk stratification algorithm (Obermeyer et al.), and a randomized trial of AI-supported selection for supplemental MRI in breast cancer screening (Salim et al.).

Data Extraction and Thematic Synthesis

The three dimensions of bias were derived mainly from four anchor works: Hussain et al. describes the formation of an inequity feedback loop through biased clinical documentation, under-diagnosing algorithms, and inequitable tools[6]; Obra et al. demonstrates how historical research norms, underrepresentation, and predictors embed bias in the AI lifecycle[7]; Thomsen et al. demonstrates subgroup performance disparities resulting from uneven data composition and quality in the model training phase[8]; Maurud et al. provides a broader health-equity framing in the context of clinical research informatics[5]. Additional case studies and related governance or mitigation studies enhanced information from these anchors.

This process led to the classification of the three dimensions of bias in this review. They are not mutually exclusive. And the three dimensions also provide more clarity in where inequity enters and how it is reproduced. The first dimension of bias is cognitive and societal bias, referring to the existing values, presumptions, and social structures that shape the design of healthcare AI. The second dimension of bias is data and measurement bias, comprising bias in sampling and labelling or proxy selections that lead to inequitable. The third dimension is model and systemic bias, referring to the situations where inequities are amplified due to model objectives, optimization choices, and post-deployment feedback loops.

Table 1 Exact search phrases for each database.

Date	Database	Exact query strings
June 13, 2026	PubMed	("algorithmic bias" OR "machine learning bias" OR "algorithmic fairness") AND (healthcare OR "clinical decision support" OR "medical ethics") AND (equity OR "racial disparities")
June 13, 2026	IEEE Xplore	(("All Metadata":"algorithmic opacity" OR "All Metadata":"black box AI" OR "All Metadata":"explainable AI" OR "All Metadata":"XAI") AND ("All Metadata":healthcare OR "All Metadata":"medical imaging" OR "All Metadata":"clinical decision support") AND ("All Metadata":mitigation OR "All Metadata":"bias detection" OR "All Metadata":"fairness-aware"))
June 13, 2026	Google Scholar	intitle:"algorithmic bias" healthcare "ethical" (quantitative OR empirical) -McKinsey -WHO -Deloitte

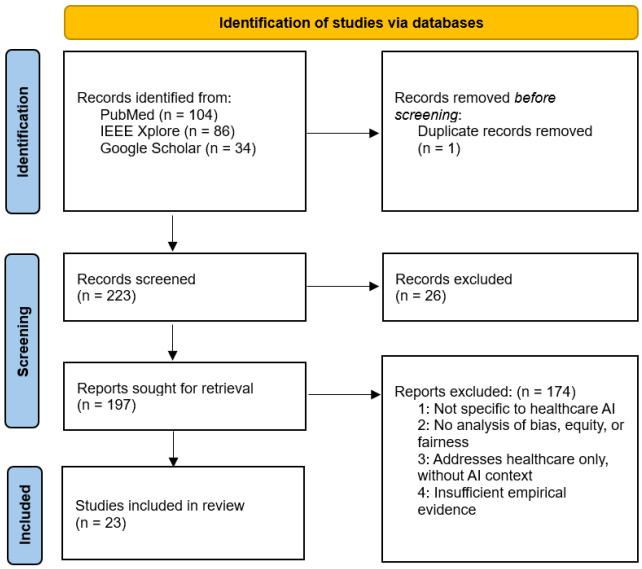

Figure. 1 PRISMA-style flow diagram of study selection process

Quality Appraisal

To assess the quality of each included literature study, research question clarity, study design transparency, and healthcare equity and AI bias relevance were assessed.

Quality appraisal was not used as a strict criterion of exclusion except when studies were extremely vague or did not have identifiable methodological substance. Reviewed studies span empirical studies, reviews, and policy or framework papers. Each type of study has heterogenous designs. Hence, a single numeric rating was not applied to evaluate their quality. In place of the numeric rating, a CASP-style criteria were used. The criteria include question clarity, method transparency, and healthcare equity relevance.

To reduce bias in the review process, the following approaches were taken: three complementary databases were covered on our search to ensure diversity; the search query combined terms such as conceptual, clinical, fairness, opacity, and governance to ensure the review would not lean towards engineering studies or ethics commentary only; and, reference tracing was used to capture influential studies missed by keyword searches. Nonetheless, the review has limitations as included literature remains concentrated on high-income settings, especially the United States, and many studies focus on a small number of highly cited case studies.

Results

Three Dimensions of Algorithmic Bias Undermine Equity

Hasanzadeh et al. classified the AI lifecycle into six phases[9]. Each lifecycle phase is mapped to a dominant bias dimension to clarify where equity enters the lifecycle. Each step in the AI lifecycle is vulnerable to algorithmic bias. These biases can be broadly categorized into three interrelated dimensions: cognitive and societal, data and measurement, model and systemic.

Cognitive and Societal Bias

Historical cognitive and societal biases involve certain perceptions, assumptions and preferences. These can become embedded in medical healthcare practices and translated directly into the dataset used to train healthcare AI. These biases span from an individual's subconscious processes like confirmation bias to society's enduring inequities such as racial stereotypes, gender norms, and discrimination.

One route through which social bias enters healthcare AI is medical perception. A study found that 2.5% of 48,651 hospital admission notes from an academic medical center

Table 2 Mapping three dimensions of bias onto a six-phase AI lifecycle.

Lifecycle Phase	Dominant Bias Dimension
Phase 1: Conception	Cognitive and societal bias
Phase 2: Data Collection	Data and measurement bias
Phase 3: Pre-processing	Data and measurement bias (transitioning into) Model and systemic bias
Phase 4: In-processing	Model and systemic bias
Phase 5: Post-processing	Model and systemic bias
Phase 6: Post-deployment Surveillance	Cognitive and societal bias Data and measurement bias Model and systemic bias

contained stigmatizing language[10]. Higher rates of stigmatizing language were found in more specific conditions, such as 6.9% for diabetes and 3.4% for substance use disorder. Across these subgroups, non-Hispanic Black patients also had a higher probability of stigmatizing language on their note. The effect on patients due to such language should not be overlooked. Stigmatizing language can influence or even modify patient treatments, alienate patients, and translate prejudice among medical professionals[6]. When these notes are used as training data, race-based and disease-specific stigma can be directly embedded into AI systems.

Historical research norms provide another way for social assumptions to enter algorithms. After the thalidomide tragedy, women of reproductive potential were excluded from clinical trials in the name of fetal protection[7]. Thus, clinical decision instruments (CDIs) tied persistent male dominant cohorts to Western regulatory procedures after the tragedy. Investigators continue to design and interpret studies relative to earlier male-only cohorts even after policies have been revised. This leads to the normalization of male bodies and experiences as the standard. When social values lead to protective regulatory response, it can become an ingrained research norm that shapes the representation on development datasets decades later.

Assumptions can encompass more than one aspect. A kidney-function case study demonstrates the consequences of assumptions that intersect race, biology and society. Through the use of race-adjusted eGFR (estimated glomerular filtration rate) equation, a widely used algorithm for prioritizing kidney transplant candidates deprioritized Black patients[11]. The eGFR equation has long been a common measurement tool for kidney health. The race-modified equation was created by researchers who found a difference in GFR rates between Black patients and non-Black patients. The researchers attributed this to fundamental biological differences instead of existing racial disparities in society. However, there is a dearth of evidence for biological differences that state Black patients have more muscle mass than white patients. As a result, after a race modifier is added to the equation to better fit the model, a higher GFR is assigned to Black patients. This means a Black patient's kidney health will be overestimated by the equation. This is an example of how race, treated as a biological category, encodes historical and social inequities into a seemingly objective formula.

Cognitive and societal biases do not only enter healthcare AI through open prejudice. These studies show that cognitive and societal biases can appear in many forms, including stigmatizing language, research norms, and race variables. Once these biases enter the algorithm, it becomes part of the training that AI systems treat as the truth, magnifying disparities in its outcome. The opacity created by this demands the need for explainability tools and governance structures to pinpoint hidden assumptions.

Data and Measurement Bias

Representation bias, sampling bias, selection bias, missing or filtered data, and differences in data quality across groups can all be categorized into data and measurement bias.

Many AI development datasets have representation bias. A review found that 97.5% of 555 neuroimaging-based AI models were exclusively trained on data from high-income populations rather than globally representative data[12]. When algorithms are trained to optimize majority-group data, they

can succeed on an aggregate basis but can underperform on marginalized patients as their characteristics may be absent or under-sampled during model training.

In addition, measurement can also be distorted through sex skew and geographic concentration. Obra et al.'s CDI review highlighted that 65% of CDIs had more male than female participants in mixed sex cohorts[7]. Similar cases are seen in widely used instruments such as the Hamilton Depression Scale (HAM-D). The HAM-D was developed using only male patients. However, depression is not a male-only condition. Depression affects all sexes and is manifested differently across sexes. By training exclusively on male patient symptoms, women's experience may be under-recognized. Moreover, Obra et al. find that CDI authors are mostly based in North America (52%) and Europe (31%), with 45% of all authors located in the United States. Yet, CDIs developed from these environments are often used in non-Western contexts. These contexts differ by disease patterns and population structures. The Kruis Score for irritable bowel syndrome (IBS) is an example of how health patterns differ in these contexts. Because there is a higher IBS prevalence in women in the West, during the development of the Kruis Score for IBS, women were over-represented in the development cohort. However, regional consensus statements show that IBS prevalence is approximately the same among both genders in Asian, South American, and African populations. When these scores are applied to regions where health patterns differ, in this case by gender, they may systemically under-detect IBS in men or over detect it in women. This shows how geographically narrow development samples can limit global generalizability.

Underdiagnosis and mismeasurement can also stem from data and measurement bias. As reported in Hussain et al., Seyyed-Kalantari et al. highlights the underdiagnosis of pulmonary disease in Hispanic and Black women[6]. This underdiagnosis was because chest-radiograph models were trained on majority population datasets. The disproportionate training data leads to delayed access to care for Hispanic and Black women. Goldstein et al. describe this as "differential observability," which is when certain groups receive fewer opportunities such as referrals, tests, or specialty evaluations, less information about them is recorded, any algorithm trained on observed utilization will learn these gaps as if they reflected lower need[13]. An example is that Black patients are less likely to receive nephrology care before dialysis compared to White patients. When a model trained to predict the need for nephrology services uses historical nephrology visits, it under predicts the need of Black patients. A similar case happens with electronic health record (EHR) data. Murthi et al. have argued that EHRs capture only people who access and remain in the healthcare system[14]. Hence, uninsured, under-insured, or medically underserved populations risks being systematically under represented. This links back to cognitive and societal bias and is amplified through data and measurement bias.

Even with a balanced data set, data and measurement bias can still exist. Thomsen et al.'s study shows that despite a similar number of Black and White participants (104 vs 101) being included to train glucose-prediction models for type 1 diabetes, the quality of data differed between Black and White participants[8]. Black participants had less continuous glucose monitoring (CGM) data on average (8.6 vs. 9.5 weeks) and more missing data (4.1 vs. 2.6 weeks). As White participants' data become more dominated in training, model performance improved for them but worsened for Black participants. Thomsen et al. conclude that diverse training data are essential for equitable AI-enabled glucose prediction even when overall sample sizes were balanced.

Finally, a self-reinforcing feedback loop is constructed by how data and measurement bias impact patient trust and utilization. Hussain et al. illustrates the feedback loop that can be formed between patients and algorithmic outputs[6]. Mistrust develops when patients experience biased or opaque algorithmic outputs. This mistrust reduces the patient's engagement with care, which in turn reduces the amount and quality of data collected from them. This further skews datasets toward majority populations. As training data become increasingly unrepresentative, AI models become less accurate for marginalized groups. This produces further bias and deeper mistrust, forming a feedback loop. This measurement layer of bias is often opaque to clinicians and patients who see only the final prediction.

Model and Systemic Bias

Model design decisions and deployment contexts can create new disparities or exacerbate preexisting ones, even in situations where data is diverse and of high quality. Model and systemic bias may result from decisions about proxy targets, thresholds, and operational objectives. These decisions determine how the AI models should optimize outcomes, which errors to tolerate, and how algorithmic recommendations translate into real world output.

Obermeyer et al. analyzed a commercial population-health algorithm that falsely deemed Black patients healthier than White patients at the same predicted risk[15]. In their study sample of 49,618 patients, the model used predicted healthcare costs as a proxy for health needs. Historically, Black patients faced structural barriers to access care. Thus, they incur lower costs even at similar or higher levels of illness. Because of this proxy, the algorithm systematically assigned them lower risk scores than those assigned to White patients with similar levels of illness. Black patients had 26% more chronic illnesses at the same risk score. Correcting this bias would increase the share of Black patients identified for extra care from 17.7% to 46.5% at the 97th percentile of the algo-

rithm's risk score. Here, racial bias was not explicitly embedded using race. Instead, a seemingly neutral design choice of the model, using cost as the prediction target, implicitly encoded decades of structural discrimination into the model.

Structural inequities can further be amplified by neutral variables. As Hussain et al. mentioned, in Shanklin et al.'s study, an appointment-scheduling system allocates appointment slots based on predicted no-show risk[6]. For Black patients, structural barriers such as transportation, inflexible work schedules, and mistrust raise no-show rates for them. The model learns that Black patients have higher risks. As a result, they are being systematically assigned to less convenient slots with longer wait times. This shows how a system optimizing for operational efficiency can optimize against equity in the process, creating further barriers to disadvantaged patients. Even when neutral variables are not optimized for efficiency in operations, they can also affect clinical decisions. Obra et al.'s review reveals that family history appears as a predictor in 1.4% of CDIs[7]. For patients such as immigrants or those without documented family history information, this missing information may reduce the number of risk factors input in calculations. Ultimately, this can influence the decision-making for screening or treatment. Even with improved data, model and systemic bias could enter through model design choices, underscoring the need for governance and fairness-aware model design as mitigation measures.

Opacity Traps Exacerbate Bias and Destroy Trust

Algorithmic opacity makes bias detection and interpretation challenging once AI systems move from development to practice. Opacity arises in healthcare when stakeholders including patients, clinicians, and institutions cannot inspect how a model was trained, the assumptions that determine its outputs, or how recommendations can be translated into clinical decisions. Decisions within healthcare should be transparent because it allocates time, resources and life-saving interventions.

A case in radiology illustrates the consequences of algorithmic opacity on clinical trust and usability. The clinical trial ScreenTrustMRI tested AISmartDensity to detect cancers missed by human radiologists[16]. While the AI effectively identified missed cancers, it only offered a risk score without explanation. This creates a barrier to clinician interpretation. Because clinicians could not fully understand or verify the rationale behind these AI decisions, it is difficult to use and/or explain the results to their patients. This can be inferred as a gradual erosion of clinicians' trust, even when the AI was correct, making them less willing to integrate and justify it into practice. This case shows how algorithmic opacity can limit AI's potential impact on healthcare.

Another problem associated with algorithmic opacity is false privacy protection. Current technical evidence no longer supports the idea that de-identified, pseudonymized or anonymized data are safe. Algorithmic opacity prevents patients from interrogating whether de-identification of their data is sufficiently robust or how their data is used. Hence, patient consent is not fully informed under this context.

Ahmed et al.'s study shows that the effectiveness of anonymization techniques is increasingly in doubt and that re-identification is becoming easier with data availability and advanced data mining methods[17]. Ahmed et al. also mentioned that Rocher, Hendrickx, and de Montjoye's 2019 *Nature Communications* study used a generative copula-based model to estimate re-identification risk in anonymized datasets. The results showed that with just 15 demographic attributes, 99.98% of Americans could be correctly re-identified in any available anonymized dataset.

From the perspective of healthcare equity, this technical reality is particularly important. The harms of re-identification are not distributed evenly across groups. Healthcare records contain uniquely identifying combinations of diagnoses, medications, procedures, dates, and more. Individuals or groups with rare conditions or uncommon treatment trajectories make up more unique combinations. This makes some individuals or groups especially "fingerprintable." The claim that data is anonymized can create a false sense of security. Yet, patients are unable to contest that claim.

Moreover, it is challenging to enforce accountability for re-identification risk. Throughout the AI lifecycle, stakeholders may only control parts of the opaque chain. Because not a single party has complete visibility in this chain, the risk of re-identification becomes unassignable and dispersed. For communities with a history of mistrusting healthcare organizations, such as those who have been subjected to discrimination, surveillance, or unfair study methods, the willingness to interact with health services is further undermined by this opacity. Such communities lack the means to assess if said protections are sufficient in the absence of algorithmic transparency. This may thereby worsen the very data inequities that underlie various forms of algorithmic bias.

Bias Mitigation Measures

Across the AI lifecycle, algorithmic bias and opacity cannot be mitigated with a single measure. Existing mitigation tends to target specific areas of susceptibility including opaque model logic, skewed training distributions, or weak institutional oversight. Yet, each mitigation measure has its own trade-offs. These trade-offs may be in validity, scalability, cost, and clinical usability. Mitigation measures are not universally effective tools. They should be viewed as context-dependent interventions, and their value depends on the clinical task and the healthcare environment in which they are deployed.

Mitigation Measure 1: Explainable AI (XAI) Techniques

XAI is a prominent mitigation measure focusing mainly on explainability and interpretability of machine learning. This primarily addresses algorithmic opacity and supports the detection of cognitive and societal bias and model and systemic bias. XAI techniques include Shapley Additive Explanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and Score-Weighted Class Activation Mapping (ScoreCAM). Although XAI techniques typically have low to moderate computational costs, they still require additional time from clinicians and informatics expertise to configure and interpret. As a result, their viability is greatest in settings with greater resources that can afford this overhead. According to Mandava et al., SHAP and LIME increased physician transparency and confidence in healthcare risk assessment without significantly reducing performance [18].

An ensemble model from Elias et al.'s report achieved 90.82% accuracy, 89.99% F1-score, 100% precision, and 81.63% recall for osteoporosis risk [19]. XAI helped identify age, hormonal changes, and family history as influential predictors.

A similar case is Kabir's breast-cancer classification study that reports 91.02% accuracy for the best CNN model with ScoreCAM and LIME visual explanations [20]. These studies demonstrated that XAI can make model behavior and outcomes less opaque. However, XAI does not inherently remove unfairness.

In fact, Tsolakis et al. demonstrated that current saliency maps and feature-importance plots for cardiac deep-learning models are often overly complicated and poorly aligned with cardiologists' workflows [21]. This limits their utility in real-time decision-making. They highlight that unstructured XAI output can be more perplexing than beneficial by introducing an ontology-based review process that systematically scores explanations for clarity and clinical relevance. While highly simplified explanations risk encouraging overconfidence, more comprehensive significance scores and saliency maps can quickly become too overwhelming for physicians, indicating a trade-off.

Similarly, in mental health screening, though XAI tools such as SHAP can show how the features that contributed to a prediction, they are still unable to translate results into actionable understanding to support concrete screening or triage decisions [22]. To address this, Kandala et al. proposed a Generative Operational Framework. This framework uses large language models and retrieval-augmented generation to translate XAI outputs and clinical guidelines into integrated and human-readable reports. Hence, without an additional translation layer, XAI can remain too complex to be used immediately in healthcare settings.

In addition, Alghamdi et al.'s FIXAIH framework emphasizes the "technical-proficiency gap" between the complexity of typical explanation methods and the practical needs of clinicians and regulators [23]. The framework is built from an analysis of 5,083 XAI and interpretable ML studies. It is critical that XAI is not simply treated as an add-on that makes a model trustworthy. Because of its resource-intensity, XAI tools need to be matched to specific clinical needs, legal requirements, and workflow realities. In high-volume and real-time healthcare settings, XAI viability depends on institution's technical infrastructure to generate explanations on a large scale and capacity to integrate explanations into routine decision support.

Mitigation Measure 2: Balancing Datasets

A second set of mitigation measures involves adding or reweighting examples in imbalanced datasets. This primarily addresses data and measurement bias. In Ugbomeh et al.'s review, a stroke study data set consists of more non-stroke cases than stroke cases [24]. This can cause models to miss high-risk patients. They apply SMOTE, a method that generates synthetic minority-class examples based on existing cases, and train several machine learning models.

After applying SMOTE, the best model had reached a high accuracy and F1-score. This case showed that synthetic balancing can indeed improve model performance on a dataset. Given its low computational cost and simple implementation, SMOTE can be sufficiently scalable across clinical datasets. Nonetheless, the study did not report model performance of synthetic samples on different demographic subgroups. It cannot claim that SMOTE made the system fair. It is only that it improved overall metrics.

A similar root cause is presented in Kabir's breast cancer work [20]. There is an unbalanced imaging dataset where some tumor classes are underrepresented. The authors augment the data to increase minority-class examples and then train several Convolutional Neural Networks (CNNs). The best model gained a 91% accuracy. Again, the impact of augmentation on performance for different patient groups was not examined.

Both studies showed how synthetic and resampling methods can bolster accuracy of models. At the same time, however, it is possible that they may also reproduce or exacerbate existing biases. These mitigation measures must be followed by subgroup-stratified validation and clinical review to achieve healthcare equity. Otherwise, a misleading impression of fairness can be given while concealing persistent inequities.

Another related method to this mitigation measure is fairness-aware optimization. Fairness-aware optimization includes subgroup-aware objectives, distributionally robust learning, or federated fairness constraints. These methods mainly target model and systemic bias. In particular, they address unequal error rates or performance gaps across groups.

Zhang et al.'s Federated Learning with Unified Fairness Ob-

jective (FedUFO) work identified that federated healthcare models trained across multiple institutions can perform unevenly across subpopulations[25]. A FedUFO adds a common fairness objective across four real digital-health tasks with the objective of maintaining stable performance across groups while preserving overall accuracy. As a result, the fairness-aware federated models had a similar accuracy level to standard baselines, but more consistent performance across subgroups. This result shows how fairness constraints can be built in without a large tradeoff of accuracy. However, this added layer of complexity may increase computational cost and lower viability in less equipped healthcare systems.

Li et al.'s FairFML tackled the specific equity issue of gender disparities in cardiac arrest outcome prediction[26]. A federated learning framework is designed so the training objective penalizes unequal performance between men and women. Their experiments show that fairness metrics improve up to 65% compared with conventional centralized models. Along with this improvement, FairFML still maintains predictive quality. However, the optimization itself can be computationally demanding and may require reliable labels for protected attributes. This makes it hard to deploy in hospitals with limited technical capacity. Therefore, for clinical deployment, these methods need to undergo ongoing auditing.

Mitigation Measure 3: AI Governance

AI Governance does not directly remove biases from models. It is used to address model and systemic bias and algorithmic opacity through structuring the design, evaluation, monitoring and revision of AI systems. Governance also determines how equity is treated. It can be addressed as a core requirement or an afterthought. Only a minority explicitly consider inequity, racism, or bias in AI governance in Nong et al.'s study, which encompasses 17 interviewees from 13 academic medical centers across four U.S. regions[27]. Equity is mostly framed as a property of the tool rather than a broader institutional problem. This result points to the existing inadequate governance even in leading centers.

Goldstein et al. illustrated a concrete model by the Algorithm-Based Clinical Decision Support (ABCDSS) oversight committee[13]. First, the reasons for including certain sensitive variables such as race, ethnicity, sex, and payer status are documented through retrospective evaluation. Stratified performance metrics are also required. Then, silent evaluation with tools running in the background is performed to see how they translate into workflow and whether data elements are differentially observed across demographic groups. Ongoing monitoring is used to track performance after deployment. This model is costly in terms of expert time and analytic effort. It may scale best in institutions that can support a standing committee.

From the lifecycle perspective, Kumar et al.'s Regulation-Aware Neural Network Lifecycle (RANNL) incorporated interpretability, fairness metrics, and regulatory readiness into a five-stage model development process[28]. External principles including the EU AI Act, UNESCO, and OECD recommendations are connected to each stage. This lengthy process likely incurs costs in design and documentation. However, compared to ad hoc governance, it scales better because templates and checkpoints that are already in place can be reused across models. Organizations with an existing quality-management or AI operations pipeline allow higher viability of RANNL. The nature of compliance is the main vulnerability to be wary of. When supervisory bodies lack equity literacy or execution power, compliance can risk turning into a box-ticking activity. External regulations also risk box-ticking if detailed operational criteria are not provided. Therefore, before putting models into clinical testing or deployment, even a regulation-aware AI operations pipeline still needs explicit thresholds and empirical confirmation of fairness and performance at each level.

Overall, AI governance operates across the entire AI lifecycle, making it powerful. However, the risks become high if committees only exist on paper or lack equity literacy, enforcement power, and resources for sustained review. Expert time, institutional coordination, and documentation efforts represent substantial implementation costs outside well-resourced systems.

Discussion

Three bias dimensions were mapped onto Hasanzadeh et al.'s six phases of the AI lifecycle in the results section[9]. Table 3 extends this mapping by adding where the three mitigation measures act in each phase of the AI lifecycle.

In the early conception phase, XAI reveals cognitive and societal assumptions. Balancing datasets become most effective where data and measurement bias enter during collection and pre-processing. Model and systemic bias dominates phases 4-5, in-processing and post-processing. This makes AI governance as a mitigation measure crucial for sustained supervision throughout the lifecycle. And all three bias dimensions re-enter in phase 6: post-deployment surveillance. Table 3 also implies that biases can reappear even after initial mitigation.

Similarly, opacity persists throughout the AI lifecycle and can be directly addressed by mitigation measures of XAI techniques and AI Governance. This review finds that if Artificial Intelligence is developed to address bias and opacity and is ethically managed, it can promote more equitable, transparent, and responsible healthcare. According to the reviewed literature, AI can improve diagnostic accuracy and support decision making, but these benefits remain uneven when the three di-

Table 3 Mapping AI Lifecycle Phases to the Mitigation Measures for the Bias Dimensions.

Lifecycle Phase	Dominant Bias Dimension	Mitigation Measures to those biases
Phase 1: Conception	Cognitive and societal bias	Explainable AI (XAI) Techniques
Phase 2: Data Collection	Data and measurement bias	Balancing Datasets
Phase 3: Pre-processing	Data and measurement bias (transitioning into) Model and systemic bias	Balancing Datasets AI Governance
Phase 4: In-processing	Model and systemic bias	AI Governance
Phase 5: Post-processing	Model and systemic bias	AI Governance
Phase 6: Post-deployment Surveillance	Cognitive and societal bias Data and measurement bias Model and systemic bias	Explainable AI (XAI) Techniques Balancing Datasets AI Governance

mensions of bias are unmitigated. The healthcare system is a high-stakes environment where lives may be on the line. Thus, the demand is not merely technical accuracy, but the risk mitigation of ethical consequences resulting from embedded biases in opaque systems.

The first implication is that technical design should not be the sole focus for healthcare AI. Institutional histories, proxy variables, and resource asymmetries that are embedded into healthcare systems also impact healthcare AI. Because of this, bias cannot be reduced to questioning model performance across demographic groups under a single definition of statistical fairness. As Goldstein et al. pointed out, qualitative assessments are often just as relevant as quantitative performance assessments in the context of clinical decision support[13]. The examination of variables' contributions, value, and interactions with underlying heterogeneity are all part of qualitative assessments.

In this review, though many mitigation measures appear to be promising, they have not been validated across large diverse healthcare systems. This review is a conceptual and qualitative review rather than a formal meta-analysis, it is better suited to identify mechanisms rather than ranking mitigation strategies with quantitative metrics. The reviewed literature is also skewed towards high-income institutions with comparatively strong data infrastructure and analytic capacity. Therefore, the inference about the feasibility of each approach in under-resourced healthcare settings is limited.

Although ethical obligations and liability in healthcare AI are beyond the scope of this review, they have important implications for how algorithms are designed, governed, and used. When algorithmic bias and opacity negatively impact patients, the moral and legal responsibility may be difficult to assign because accountability is dispersed among stakeholders such as data curators, model developers, vendors, institutional leaders, and clinicians throughout the AI lifecycle. Ethical duty can only be satisfied when roles of human oversight are clearly defined. Clinicians should have the power to question and override AI outputs, thereby providing clearer points of accountability. However, it can also function as a liability shield when clinicians are held responsible in the absence of system transparency. This may obscure the greater responsibility embedded in the design and deployment that ultimately shape an algorithm's behavior.

The results of this review suggest that we have to be cautious of mitigation measures as they are not definitive prescriptions. XAI techniques can increase interpretability of model behavior and help clinicians identify problematic features or proxies. Yet, the implementation process can be complex and interpreting the output from XAI tools is time-consuming. Synthetic balancing, augmentation, and fairness-aware optimization can reduce certain performance gaps. Even so, they may inadvertently amplify existing biases or shift error burdens onto other groups. Finally, through AI governance, equity checks can be incorporated into model development, evaluation, and monitoring.

On balance, this review encourages a cautiously optimistic attitude to healthcare AI systems where AI will not replace human judgement. Future research should explore hybrid human-AI models with operationalization of governance and accountability. It should also assess how equity-focused frameworks influence which tools are adopted, who is able to use them, and the circumstances under which they are deployed.

References

1 Y. A. Fahim, I. W. Hasani, S. Kabba, W. M. Ragab. Artificial intelligence in healthcare and medicine: clinical applications, therapeutic advances, and future perspectives. European Journal of Medical Research. Vol. 30,

pg. 848, 2025, https://doi.org/10.1186/s40001-025-03196-w.

2 J. Bajwa, U. Munir, A. Nori, B. Williams. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthcare Journal. Vol. 8, pg. e188–e194, 2021, https://doi.org/10.7861/fhj.2021-0095.

3 Z. Strika, K. Petkovic, R. Likic, R. Batenburg. Bridging healthcare gaps: a scoping review on the role of artificial intelligence, deep learning, and large language models in alleviating problems in medical deserts. Postgraduate Medical Journal. Vol. 101, pg. 4–16, 2024, https://doi.org/10.1093/postmj/qgae122.

4 E. G. Poon, C. H. Lemak, J. C. Rojas, J. Guptill, D. Classen. Adoption of artificial intelligence in healthcare: survey of health system priorities, successes, and challenges. Journal of the American Medical Informatics Association: JAMIA. Vol. 32, pg. 1093–1100, 2025, https://doi.org/10.1093/jamia/ocaf065.

5 S. Maurud, S. H. Henni, A. Moen. Health equity in clinical research informatics. Yearbook of Medical Informatics. Vol. 32, pg. 138–145, 2023, https://doi.org/10.1055/s-0043-1768720.

6 S. A. Hussain, M. Bresnahan, J. Zhuang. The bias algorithm: how ai in healthcare exacerbates ethnic and racial disparities - a scoping review. Ethnicity & Health. Vol. 30, pg. 197–214, 2025, https://doi.org/10.1080/13557858.2024.2422848.

7 J. K. Obra, C. Singh, K. Watkins, J. Feng, Z. Obermeyer, A. Kornblith. Potential for algorithmic bias in clinical decision instrument development. NPJ Digital Medicine. Vol. 8, pg. 762, 2025, https://doi.org/10.1038/s41746-025-02119-7.

8 H. B. Thomsen, L. Y. Li, A. A. Isaksen, B. Lebiecka-Johansen, C. Bour, G. Fagherazzi, W. P. T. M. van Doorn, T. V. Varga, A. Hulman. Racial disparities in continuous glucose monitoring-based 60-min glucose predictions among people with type 1 diabetes. PLOS Digital Health. Vol. 4, pg. e0000918, 2025, https://doi.org/10.1371/journal.pdig.0000918.

9 F. Hasanzadeh, C. B. Josephson, G. Waters, D. Adedinsewo, Z. Azizi, J. A. White. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. Npj Digital Medicine. Vol. 8, pg. 154, 2025, https://doi.org/10.1038/s41746-025-01503-7.

10 G. Himmelstein, D. Bates, L. Zhou. Examination of stigmatizing language in the electronic health record. JAMA Network Open. Vol. 5, pg. e2144967, 2022, https://doi.org/10.1001/jamanetworkopen.2021.44967.

11 P. Nissaisorakarn, H. Xiao, M. D. Doshi, N. Singh, K. L. Lentine, S. E. Rosas. Eliminating racial disparities in kidney transplantation: race-adjusted egfr in black candidates and its implications on preemptive listing. Clinical Transplantation. Vol. 35, pg. e14397, 2021, https://doi.org/10.1111/ctr.14397.

12 Z. Chen, X. Liu, Q. Yang, Y.-J. Wang, K. Miao, Z. Gong, Y. Yu, A. Leonov, C. Liu, Z. Feng, H. Chuan-Peng. Evaluation of risk of bias in neuroimaging-based artificial intelligence models for psychiatric diagnosis: a systematic review. JAMA Network Open. Vol. 6, pg. e231671, 2023, https://doi.org/10.1001/jamanetworkopen.2023.1671.

13 B. A. Goldstein, D. Mohottige, S. Bessias, M. P. Cary. Enhancing clinical decision support in nephrology: addressing algorithmic bias through artificial intelligence governance. American Journal of Kidney Diseases: The Official Journal of the National Kidney Foundation. Vol. 84, pg. 780–786, 2024, https://doi.org/10.1053/j.ajkd.2024.04.008.

14 S. Murthi, N. Martini, N. Falconer, S. Scahill. Evaluating ehr-integrated digital technologies for medication-related outcomes and health equity in hospitalised adults: a scoping review. Journal of Medical Systems. Vol. 48, pg. 79, 2024, https://doi.org/10.1007/s10916-024-02097-5.

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

16 M. Salim, Y. Liu, M. Sorkhei, D. Ntoula, T. Foukakis, I. Fredriksson, Y. Wang, M. Eklund, H. Azizpour, K. Smith, F. Strand. AI-based selection of individuals for supplemental mri in population-based breast cancer screening: the randomized screentrustmri trial. Nature Medicine. Vol. 30, pg. 2623–2630, 2024, https://doi.org/10.1038/s41591-024-03093-5.

17 M. M. Ahmed, O. J. Okesanya, M. Oweidat, Z. K. Othman, S. S. Musa, D. E. Lucero-Prisno Iii. The ethics of data mining in healthcare: challenges, frameworks, and future directions. BioData Mining. Vol. 18, pg. 47, 2025, https://doi.org/10.1186/s13040-025-00461-w.

18 R. Mandava, S. S. Vellela, N. Malathi, K. Haritha, S. Gorintla, L. Dalavai. Exploring the role of xai in enhancing predictive model transparency in healthcare risk assessment. in 2025 International Conference on Computational Robotics, Testing and Engineering Evaluation (ICCRTEE) pg. 1–5, 2025. https://doi.org/10.1109/ICCRTEE64519.2025.11052988.

19 F. Elias, M. S. Reza, M. Z. Mahmud, S. Islam, S. R. Alve. Interpretable machine learning for osteoporosis risk: a comparative evaluation of predictive models and xai analysis. in 2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN) pg. 1–6, 2025. https://doi.org/10.1109/QPAIN66474.2025.11172213.

20 Md. Z. I. Kabir. Breast cancer classification using pre-trained cnns with explainable ai for enhanced decision support. in 2025 International Conference on Electrical, Computer and Communication Engineering (ECCE) pg. 1–6, 2025. https://doi.org/10.1109/ECCE64574.2025.11012958.

21 N. Tsolakis, C. Maga-Nteve, S. Vrochidis, N. Bassiliades, G. Meditskos. Enhancing cardiac ai explainability through ontology-based evaluation. in 2024 15th International Conference on Information, Intelligence, Systems & Applications (IISA) pg. 1–4, 2024. https://doi.org/10.1109/IISA62523.2024.10786663.

22 R. Kandala, A. K. Moharir, D. A. Nayak. From explainability to action: a generative operational framework for integrating xai in clinical mental health screening. in 2025 IEEE International Conference on Data Mining Workshops (ICDMW) pg. 1316–1325, 2025. https://doi.org/10.1109/ICDMW69685.2025.00156.

23 S. Alghamdi, R. Mehmood, F. A. Alqurashi, A. Alzahrani. Paving the roadmap for xai and iml in healthcare: data-driven discoveries and the fixaih framework. IEEE Access. Vol. 13, pg. 174393–174427, 2025, https://doi.org/10.1109/ACCESS.2025.3616353.

24 O. Ugbomeh, V. Yiye, E. Ibeke, C. P. Ezenkwu, V. Sharma, A. Alkhayyat. Machine learning algorithms for stroke risk prediction leveraging on explainable artificial intelligence techniques (xai). in 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT) Vol. 1 pg. 1–6, 2024. https://doi.org/10.1109/ICEECT61758.2024.10739320.

25 F. Zhang, Z. Shuai, K. Kuang, F. Wu, Y. Zhuang, J. Xiao. Unified fair federated learning for digital healthcare. Patterns. Vol. 5, pg. 100907, 2024, https://doi.org/10.1016/j.patter.2023.100907.

26 S. Li, Q. Wu, X. Li, D. Miao, C. Hong, W. Gu, Y. Ning, Y. Shang, N. Liu. FairFML: a unified approach to algorithmic fair federated learning with applications to reducing gender disparities in cardiac arrest outcomes. Studies in Health Technology and Informatics. Vol. 329, pg. 1848–1849, 2025, https://doi.org/10.3233/SHTI251245.

27 P. Nong, R. Hamasha, J. Platt. Equity and ai governance at academic medical centers. The American Journal of Managed Care. Vol. 30, pg. SP468–SP472, 2024, https://doi.org/10.37765/ajmc.2024.89555.

28 G. Kumar, R. Kumari, A. Shukla, A. P. S. Yadav, R. Verma, J. Gaur. Regulatory frameworks and ethical governance of neural networks in computer science applications. in 2025 IEEE 3rd International Symposium on

Sustainable Energy, Signal Processing and Cybersecurity pg. 1–7, 2025.
https://doi.org/10.1109/iSSSC66652.2025.11388372.