npj | digital medicine

Review article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01503-7

Bias recognition and mitigation strategies in artificial intelligence healthcare applications

Fereshteh Hasanzadeh[1], Colin B. Josephson[2], Gabriella Waters[3], Demilade Adedinsewo[4], Zahra Azizi[5] & James A. White[1,2]

Artificial intelligence (AI) is delivering value across all aspects of clinical practice. However, bias may exacerbate healthcare disparities. This review examines the origins of bias in healthcare AI, strategies for mitigation, and responsibilities of relevant stakeholders towards achieving fair and equitable use. We highlight the importance of systematically identifying bias and engaging relevant mitigation activities throughout the AI model lifecycle, from model conception through to deployment and longitudinal surveillance.

As of May 13, 2024, the Food and Drug Administration (FDA) update indicated an unprecedented surge in the approval of AI-enabled Medical Devices, listing 191 new entries while reaching a total of 882, predominantly in the field of radiology (76%), followed by cardiology (10%) and neurology (4%)[1]. These approvals reflect AI's growing role in healthcare, including applications such as analyzing medical images, monitoring health metrics through wearable devices, and predicting outcomes from Electronic Medical Records. This illustrates the rapid growth of AI technologies to improve and personalize patient care, not only in the field of medical imaging and diagnostics but across all aspects of healthcare delivery. This growth has been driven by the unique adeptness of AI models to navigate large healthcare datasets and learn complex relationships with reduced processing speed, enabling superior task performance compared to traditional statistical methods or rule-based techniques. However, these models can and have gained complexity, presenting challenges distinct from those encountered by simpler or traditional statistical tools. Specifically, deep learning (DL) models are commonly opaque (i.e., black-box) in nature, lacking explainability or a clear identification of features that influence model performance, thus limiting opportunities for human oversight or an evaluation of biological plausibility[2].

Central to these challenges is the issue of bias, which can manifest itself in numerous forms to exacerbate existing healthcare disparities. Regulatory bodies, including the European Commission, FDA, Health Canada, the World Health Organization (WHO), have intensified their efforts to establish stricter frameworks for the development and deployment of AI in healthcare, recognizing a critical need to uphold the core principles of fairness, equity, and explainability[3–7]. Such frameworks aim to systematically identify and mitigate bias to ensure that AI models adhere to ethical principles and do not perpetuate or amplify historical biases or discrimination against vulnerable patient populations.

This comprehensive narrative review delves into the unique types of bias commonly encountered in AI healthcare modeling, explores their origins, and offers appropriate mitigation strategies. We begin by highlighting the central role that bias mitigation plays in achieving fairness, equity, and equality for healthcare delivery. This is followed by a detailed review of where and how bias can introduce itself throughout the AI model lifecycle and conclude with strategies to quantify and mitigate bias in healthcare AI.

Methods

We followed a critical review methodology to objectively explore and consolidate literature related to AI bias in healthcare[8]. Our review focused on English articles published from 1993 to 2024, sourced from databases such as PubMed, Google Scholar, and publisher platforms like Elsevier. To refine our search, Boolean operators were used, combining keywords inclusive of "Medical AI," "Healthcare AI," "AI bias," "AI ethics," "Responsible AI," "Healthcare disparities," "Fairness in medical AI," "Equity, equality, and fairness in healthcare," "Bias mitigation strategies in AI," "Algorithmic bias," "AI model life cycle," and "Data diversity." Each specific type of bias in AI (e.g., selection bias, representation bias, confirmation bias, etc.) was also searched. Additionally, reference lists from selected articles were examined to identify further relevant studies on bias in healthcare AI applications. Studies offering differing perspectives or contradictory evidence were carefully selected to ensure a balanced and comprehensive view of bias in healthcare AI.

[1]Libin Cardiovascular Institute, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada. [2]Department of Medicine, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada. [3]Morgan State University, Center for Equitable AI & Machine Learning Systems, Baltimore, MD, USA. [4]Department of Cardiovascular Medicine, Mayo Clinic, Jacksonville, FL, USA. [5]Department of Cardiovascular Medicine, Stanford University, Stanford, CA, USA.
e-mail: jawhit@ucalgary.ca

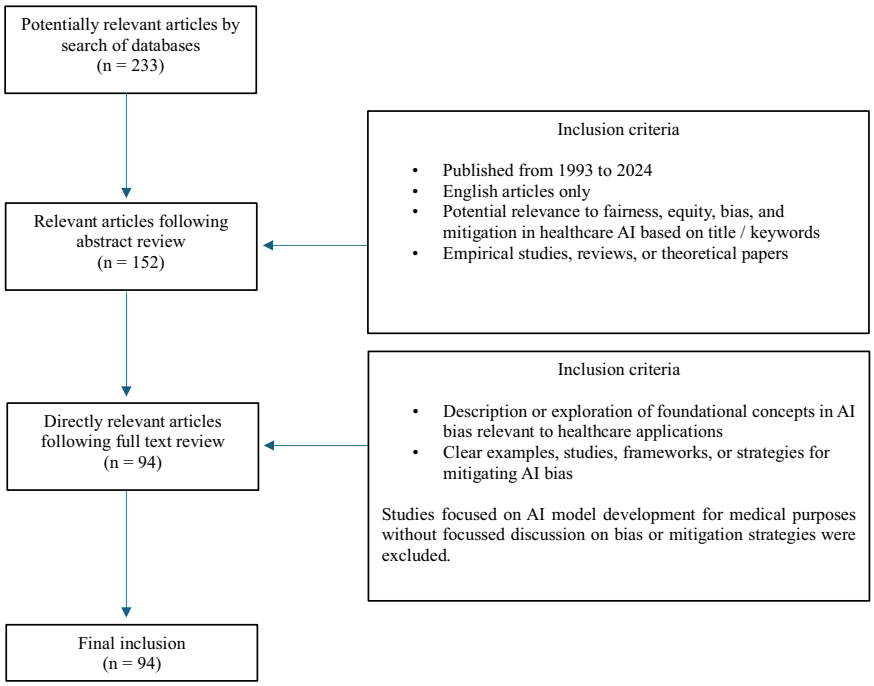

Fig. 1 | Study selection procedure for identifying relevant articles on AI bias in healthcare. This figure illustrates the process used to filter and review literature from 233 initial articles, narrowing down to 94 articles that directly explored AI bias and mitigation strategies in healthcare.

We included studies published within a broad timeframe to capture both historical and contemporary perspectives on fairness, equity, and bias in healthcare AI. Incorporating foundational references, such as a 1993 publication on equity and equality in health[9] was crucial, as these concepts form the basis for understanding and evaluating fairness in contemporary AI applications.

Our search yielded 233 potentially relevant articles. After initial screening of titles and abstracts, this was narrowed down to 152 articles with direct relevance to AI bias in healthcare. Following thorough full-text review, 94 articles were selected for our final review, as shown in Fig. 1. We employed a thematic-based approach to identify and categorize recurring themes, patterns, and insights related to bias types and mitigation strategies.

Bias

In the context of healthcare AI, bias can be defined as any systematic and/or unfair difference in how predictions are generated for different patient populations that could lead to disparate care delivery[10]. Through this, disparities related to benefit or harm are introduced or exacerbated for specific individuals or groups, eroding the capacity for healthcare to be delivered in a fair and equitable manner[11]. The concept of "bias in, bias out", a derivative of the classic adage "garbage in, garbage out," is often implicated when AI model failures are observed in real world settings[12], highlighting how biases within training data often manifest as sub-optimal AI model performance out in the wild[13]. However, bias may be introduced into all stages of an algorithm's life cycle, including their conceptual formation, data collection and preparation, algorithm development and validation, clinical implementation, or surveillance. This complexity is compounded by the inadequacy of methods for routinely detecting or mitigating biases across various stages of an algorithm's life cycle, emphasizing a need for comprehensive and holistic bias detection frameworks[14,15].

In 2023 Kumar, et al., conducted a study systematically evaluating the burden of bias in contemporary healthcare AI models[14]. Using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) selection strategy[16], a standardized methodology to estimate risk of bias (ROB), they sampled 48 studies distributed across tabular, imaging, and hybrid data models. They reported that 50% of these studies demonstrated a high ROB, often related to absent sociodemographic data, imbalanced or incomplete datasets, or weak algorithm design. Only 20%, or 1 in 5 studies were considered to have a low ROB. A similar study, performed by Chen, et al., using the PROBAST (Prediction model Risk Of Bias ASsessment Tool) framework[17] (For further details, refer to Supplementary Table 1), examined 555 published neuroimaging-based AI models for psychiatric diagnosis, identifying only 86 studies (15.5%) included external validation while 97.5% included only subjects from high-income regions. Overall, 83% of studies were rated at high ROB[18]. These studies emphasize a critical need for improved awareness of bias in healthcare AI, and the routine adoption of mitigation strategies capable of bridging model conception through to fair and equitable clinical adoption.

Fairness, Equality and Equity

Fairness, equality, and equity are core principles of healthcare delivery that are directly influenced by bias. Fairness in healthcare encompasses both distributive justice and socio-relational dimensions, requiring a holistic consideration of each individual's unique social, cultural, and environmental factors, these going beyond the concept of equality - which aims to ensure equal access and outcomes[19,20]. Equity recognizes that certain groups may require tailored resources or support to attain comparable health benefits[9]. Navigating these nuances and potential trade-offs between core principles is essential, as blanket approaches to fairness may inadvertently reinforce existing disparities[21].

Defining and measuring fairness metrics, such as demographic parity, equalized odds, equal opportunity, and counterfactual fairness, is a complex challenge that requires a deep understanding of the healthcare context as well as the lived experiences of diverse patient populations. Failure to apply these metrics appropriately can lead to unintended consequences that may undermine the ethical foundations of equitable care, such as perpetuating healthcare disparities, misallocating resources, or reinforcing systemic biases that disproportionately impact vulnerable populations[22].

Differentiating equality from equity is essential to understanding the influences that AI bias can impose on healthcare disparities. These often-competing principles, as illustrated in Fig. 2, must be iteratively considered to achieve the best possible balance[9,23].

Types of Bias

In the following sections, we explore the various forms of bias commonly encountered in healthcare AI. Our discussion, supplemented by relevant

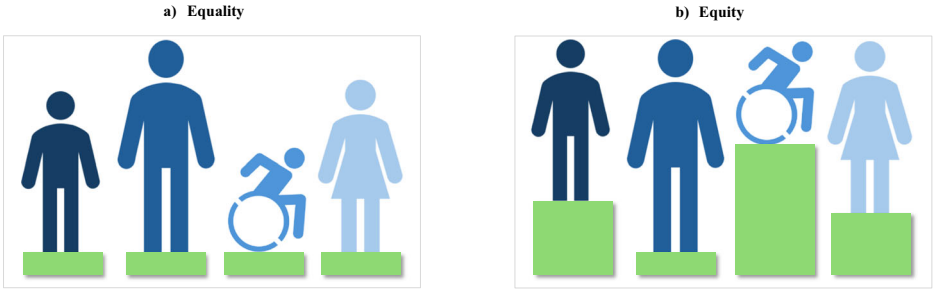

Fig. 2 | Schematic description of equality and equity in healthcare. This figure illustrates the key differences between Equality and Equity in healthcare support. It presents two scenarios: **a** "Equality" is depicted where each individual, regardless of their needs, receives the same level of support, symbolized by equal height green platforms for all figures. **b** "Equity" is shown where supports are varied according to individual needs, represented by green platforms of different heights, ensuring that each person reaches the support they need. This visualization underscores the importance of tailoring healthcare resources to address specific needs to achieve true equity.

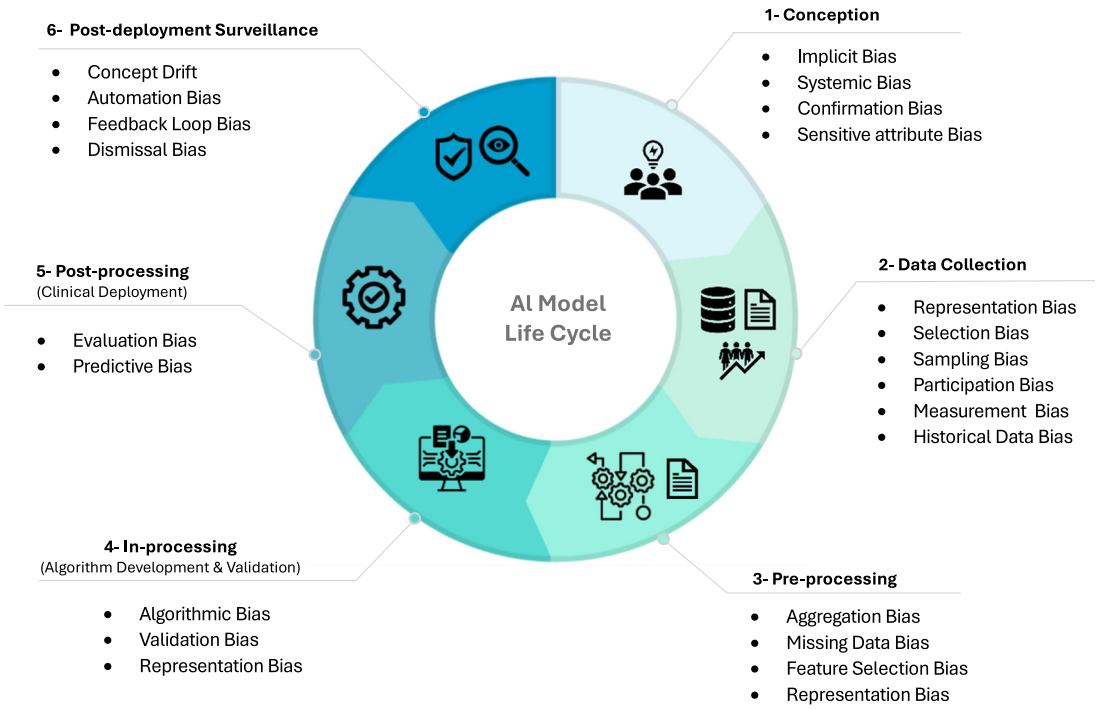

Fig. 3 | The AI model life cycle and common biases across each phase. This figure maps the stages of the AI model life cycle in healthcare, highlighting the common phase at which biases can be introduced. The AI life cycle is divided into six phases: conception, data collection, pre-processing, in-processing (algorithm development and validation), post-processing (clinical deployment), and post-deployment surveillance. Each phase is prone to specific biases that can affect the fairness, equity, and equality of healthcare delivery.

examples, has been structured to consider biases introduced from; (i) human origin; (ii) algorithm development; and (iii) algorithm deployment that may exist within each stage of the AI model lifecycle. A graphical illustration of these stages is provided in Fig. 3.

Human Biases

The dominant origin of biases observed in healthcare AI are human. While rarely introduced deliberately, these reflect historic or prevalent human perceptions, assumptions, or preferences that can manifest across various future stages of AI model development, potentially with profound impact[24]. For example, data collection activities influenced by human bias can lead to the training of algorithms that replicate historical healthcare inequalities, leading to cycled reinforcement where past injustices are perpetuated into future practice[25]. The different types of human biases that can be introduced are discussed below and summarized in Table 1.

Implicit bias. Implicit bias occurs when subconscious attitudes or stereotypes about a person's or group's characteristics, such as birth sex, gender identity, race, ethnicity, age, and socioeconomic status, become embedded in how individuals behave or make decisions. This can present surreptitious but powerful influences on medical AI systems trained from these decisions, particularly when features contributing to this bias are not routinely captured by training data[26]. This is commonly the case for patient self-reported characteristics, such as gender identity and ethnicity (referring to shared cultural practices, perspectives, and beliefs) that are commonly absent from or inconsistently coded by Electronic Health Records (EHR)[27].

Systemic bias. Systemic bias represents an important dimension of human bias. While related to implicit bias, systemic bias extends to encompass broader institutional norms, practices, or policies that can lead to societal harm or inequities. The origins of systemic bias are more structural in nature and act at higher societal levels than implicit bias, often requiring modification in legislation or institutional policies to address[28,29]. For example, systemic bias may manifest as inadequate medical resource funding for un-insured individuals, underserved

Table 1 | Relevant forms of human-introduced biases

Name	Definition	Example
Implicit bias	Subconscious attitudes or stereotypes that become embedded in how individuals behave or make decisions.	Women with cirrhosis awaiting liver transplantation are less likely to receive a transplant and more likely to die compared to men[85]. Women experiencing myocardial infarction are more likely to be misdiagnosed and less likely to receive timely treatment compared to men, due to atypical symptom presentation[86].
Systemic bias	Societal or institutional norms, practices, or policies that can lead to societal harm or inequities.	Racial disparities in Intensive Care Unit (ICU) outcomes are linked to structural inequalities, with mortality differences between racial groups explained by broader societal factors[87].
Confirmation bias	Conscious or subconscious selection, interpretation or weighting of data / results that confirms pre-formed or underlying beliefs.	In dermatology AI, annotation bias results in inconsistent disease labeling, especially related to differences in skin tone or race in 36% of cutaneous malignant neoplasms studies. Inter-annotator bias, due to varying expertise levels, further exacerbates this inconsistency by reinforcing pre-existing beliefs about normality or abnormalities[88].
Concept shift	The data or concepts change over time.	Changing from ICD-9 to ICD-10 (International Classification of Diseases) led to a notable increase in identified opioid-related hospital stays[89].

communities, or racial and ethnic minority groups, while implicit bias may be observed in a clinician's own subconscious assumptions about a patient's capacity to comply with medical care based on a stereotype.

Confirmation bias. It is important to recognize that human biases can provide influence beyond the contamination of training data. How each model is conceived and designed, and how it is ultimately used or monitored in clinical practice may similarly be influenced by human implicit or systemic biases. During model development, developers may consciously or subconsciously select, interpret, or give more weight to data that confirms their beliefs, overemphasizing certain patterns while ignoring other patterns that don't fit expectations. This form of human-mediated bias is called confirmation bias[30,31].

How human biases change over time is an important consideration given strong dependencies on the use of large-scale historic training data that are accrued over extended periods of time. While implicit and systemic biases shift with societal influence, models trained from historical data may inadvertently re-introduce unwanted biases into contemporary care. For example, older datasets influenced by an ethnicity bias may lead to AI models that generate skewed predictions across minority groups, despite more inclusive contemporary ideologies. This results in training-serving skew, where shifts in bias alter data distributions between the time of training and model serving[32]. A similar temporally sensitive bias, known as "Concept shift," can occur due to changes in the perceived meanings of data. How clinicians perceive and code diseases or outcomes can change dynamically over time, introducing a unique, human-level bias surrounding what data means and is being used for. Addressing this bias can be challenging, requiring involvement from clinicians with experience spanning these shifting practice periods. Effort must be undertaken to ensure awareness of historical data that contains outdated practices, underrepresented groups, or past healthcare disparities not representing current real-world scenarios[33].

Data bias

Beyond human influences, numerous additional biases can be introduced to data used in AI model training, altering its representation of the target environment or population and leading to skewed or unfair outcomes[34]. A summary of biases introduced during data collection are provided in Table 2.

Representation bias. Representation bias describes a lack of sufficient diversity in training data and is a dominant form of bias limiting the generalizability of healthcare AI models into unique environments or populations[25]. This bias can arise from an underlying implicit or systemic bias in the healthcare system or a history of underrepresentation for a minority group, either directly or indirectly from a reluctance to share information or participate in clinical trials[35–37]. These factors can lead to historical gaps in healthcare data, which can now be projected forward into AI healthcare models. For example, convolutional neural networks (CNNs) trained from large chest X-ray datasets sourced from academic healthcare facilities have been shown to underdetect disease in specific patient populations, inclusive of females, Black, Hispanic, and patients of low socioeconomic status[38].

Selection and sampling bias. Representation bias can also be established through selection or sampling bias. Selection bias occurs when the process of how data is chosen or collected inadvertently favors certain groups or characteristics, leading to a non-representative sample. An example of this is the "healthy volunteer" selection bias observed in the UK Biobank, where participants are generally healthy and therefore do not represent patients typically encountered by healthcare systems[39]. Sampling bias is a form of selection bias resulting from non-random sampling of subgroups, establishing data patterns that are non-generalizable to new populations. Participation bias or self-selection bias are also strong contributors to representation bias in research generated datasets. For example, patients who are sicker, have higher comorbidity, or are unable to participate in longitudinal research protocols for geographic or economic reasons may not be represented[10].

Measurement bias. Large scale data resources required for AI model development and validation are often sourced from multiple hospital sites, each with unique methods of data acquisition. Such variations are often not related to biological factors, rather are related to differences in data acquisition or processing. These measurement biases result in systematic differences that alter the representation of variables, being most commonly encountered in medical diagnostics[33]. In radiology, for example, imaging hardware manufacturer, model, software versions and acquisition parameters all meaningfully alter the characteristics of diagnostic images[40]. Similarly, in pathology, variations in tissue preparation, staining protocols, and scanner-specific parameters can impact the data used in cancer-diagnostic tasks[41]. AI models can inadvertently learn patterns associated with these non-biological variations, causing them to deviate from their primary objective of diagnosing medical conditions. In contrast, models trained from only one data source, adhering to a consistent set of acquisition parameters, may underperform when applied to data from another source[34].

Algorithmic bias

Algorithmic biases can be considered as those inherent to the pre-processing of a training dataset or during the conceptual design, training, or validation phases of an algorithm. They can stem from the inappropriate selection of non-diverse datasets, features, or selection of algorithm processes set by model developers[34]. These biases, along with examples, are further detailed in Table 3.

Table 2 | Relevant forms of data bias

Bias	Definition	Example
Representation bias	The training data for an AI system does not adequately represent all parts of the target population, exhibiting underrepresentation of certain birth sexes, genders, ages, races, ethnicities, etc.	1-Racial bias in DL-based cardiac MRI (Magnetic Resonance Imaging) segmentation was observed in the UK Biobank database as the training data was not race-balanced, leading to lower segmentation accuracy for racial and minority ethnic groups compared to White participants[90]. 2-Skin cancer detection AI models trained with datasets with insufficient diversity of skin types, particularly darker skin tones, have shown poorer performance among skin of color[91].
Sampling Bias	A non-random sampling of patients leads to an increased or decreased likelihood of selecting one group more frequently than another.	The sampling of patients based on having sufficient laboratory results and medication orders in EHR-based research inherently favors sicker patients, leading to a non-random sample that may not accurately represent the broader patient population[92].
Selection Bias	The process of selecting a sample causes it to be not representative of the population, leading to skewed results that do not accurately reflect the whole.	The LIFE-Adult study exhibited selection bias, as the sample comprised participants with higher social status and healthier lifestyles than the general Leipzig population. This led to non-representative results, potentially underestimating health conditions in broader communities[93].
Participation Bias or Self-selection Bias	The decision to participate in a study is left entirely up to individuals.	Differences between volunteers and non-participants for psychological research, especially those attracted to studies on negative life events, show higher rates of personality and affective disorders. This demonstrates a self-selection bias, where individuals' psychological characteristics influence their decision to participate, leading to samples not representative of the general population[94].
Measurement Bias	Variations in data acquisition and processing which inaccurately represent the true variable of interest, leading to skewed results.	In The Cancer Genome Atlas (TCGA), measurement bias was observed due to each contributing medical institution employing different protocols for tissue processing, staining, and imaging. These discrepancies introduced systematic variations in the dataset, which were learned by deep learning models, affecting their accuracy, and demonstrating how measurement methods can introduce bias in AI research[41].

Table 3 | Relevant forms of algorithmic biases

Bias	Definition	Example
Aggregation bias	An inappropriate combination of distinct groups or populations during data pre-processing for model development leads to aggregation bias, where the model's performance is only optimized for the majority. In other words, one single model is unlikely to suit all groups.	Machine learning clinical prediction studies may not adequately handle missing data, leading to aggregation bias. This uniquely affects model performance among subgroups less likely to have complete data resources[43].
Feature selection bias	When the set of features chosen to train a model do not adequately represent the underlying problem or are not equally relevant across all subpopulations within the dataset.	AI models developed for COVID-19 patient risk prediction, triage, and contact tracing have shown feature selection bias through inadequate representation or consideration of social determinants of health (SDOH), such as race, socio-economic status, and access to technology. The underrepresentation of these factors has led to biased outcomes, particularly those most vulnerable or marginalized[95].

Aggregation bias. A type of algorithmic bias strongly impacting model generalizability is aggregation bias, which occurs during the data pre-processing phase[10]. Data aggregation is the act of transforming patient data into a format more suitable for algorithm development, including imputing missing values, selecting key variables, combining data from various sources, or engineering new data features. When population data is merged to form a common, model-ready input, biases can emerge through the selection of input features that are maximally available across all subjects, establishing a "one-size-fits-all" approach[42]. One example of this bias is managing missing or outlier values, such as patient weight. This variable may not be available for certain patients with disabilities, particularly those using wheelchairs, or may be under-representative for those with limb amputations or terminal illnesses[43]. Models trained using aggregate data may mistakenly assume uniformity across diverse patient groups, may impute missing variables, and inherently overlook unique characteristics or needs within specific subpopulations.

Feature selection bias. Feature selection bias is the selective introduction or removal of variables during model development based on pre-conceived ideas or beliefs surrounding the planned task. Extending beyond human confirmation bias, the forced inclusion of variables hypothesized to influence performance or deemed of high priority, commonly referred to as Sensitive Variables, is a common practice foundationally motivated to avoid bias. Variables, such as age, sex, and race are often deemed "sensitive" because they represent personal characteristics that predictive algorithms should ideally not discriminate against. Accordingly, their inclusion in models may be aimed at representing priority demographic groups who might behave uniquely due to a variety of differences not described by other input features. These sensitive variables therefore act as "proxy variables" (i.e. surrogates) for more complex features not adequately described by training data to enhance the accuracy and equity of healthcare AI models[30,44]. As an example, including birth sex in cancer prediction models may be suggested to ensure they do not predict a male sex-specific cancer (e.g., prostate cancer) in a female patient[45]. However, the use of proxy variables over more appropriate source data can paradoxically propagate bias. For example, a study by Obermeyer, et al., exposed racial bias in an algorithm used for predicting healthcare resource needs using a proxy variable of healthcare cost consumption. This model systematically predicted lower healthcare resource need for Black versus White patients despite similar risk levels due to lower historic healthcare spending in Black patient populations[46]. Expanded discussion of this is provided in Case Study 1 (Box 1). This highlights the profound impact that proxy variable choice can have on algorithmic bias, emphasizing a need for careful selection in the context of both study questions and intended model use.

Box 1 | Real-World Case Study 1

Mitigating Feature Selection Bias in AI Risk Prediction.

A widely used AI risk prediction algorithm in the U.S. healthcare system, analyzed by Obermeyer et al. in 2019, included data from 43,539 White patients and 6,079 Black patients (2013–2015). The algorithm, designed to identify high-risk patients based on predicted healthcare costs, exhibited racial bias, underestimating the needs of Black patients. The study found that Black patients had 26.3% more chronic illnesses than White patients at the same risk score level (4.8 vs. 3.8 conditions). This bias stemmed from using healthcare costs as a proxy for illness severity; systemic barriers like reduced healthcare access, financial constraints, and lower trust levels led to lower costs for Black patients, causing the algorithm to misjudge their risk. To address this, researchers recalibrated the algorithm to use direct health indicators, such as chronic condition counts, instead of costs. This change nearly tripled the enrollment of high-risk Black patients in care management programs, from 17.7% to 46.5%, promoting more equitable healthcare. However, ongoing surveillance is necessary, as reliance on historical data and evolving healthcare dynamics could allow biases to re-emerge[46].

Table 4 | Relevant forms of model deployment biases

Bias	Definition	Example
Automation bias	Healthcare professionals rely too heavily on an AI system's guidance and no longer search for confirming evidence.	Dratsch et al.[96] studied automation bias in mammography screening, finding that radiologist reporting accuracy declined with exposure to incorrect AI BI-RADS (Breast Imaging Reporting and Data System) suggestions. Accuracy dropped from 79.7% to 19.8% for inexperienced radiologists, 81.3% to 24.8% for moderately experienced radiologists, and 82.3% to 45.5% for very experienced radiologists when AI suggested an incorrect BI-RADS category. This demonstrated the pronounced susceptibility to automation bias, particularly for less experienced clinicians[96].
Feedback loop bias	Clinicians unconditionally trust and follow AI recommendations, even when they are incorrect. This can lead to a cycle where the algorithm continues to learn from mistakes, perpetuating and reinforcing them in future model cycles.	A model for clinical decision support regarding COVID-19 patient admission, CORONET, was intentionally programmed to provide incorrect recommendations. Healthcare professionals were expected to override these. However, only 35% of those provided with the risk score alone, and 52% of those provided with the risk score and explanation critically evaluated and considered overriding the model's incorrect recommendations[97].
Dismissal bias (Alarm Fatigue)	A tendency to ignore or undervalue critical warnings from automated systems, often due to past false alarms.	One study conducted a comprehensive observational analysis over 31 days across five ICUs, involving 461 adults and monitoring over 2.5 million alarm instances. Of 1,154,201 unique alarms, 88.8% were for arrhythmias with a majority being false positive[98]. The majority of these alarms are routinely ignored due to staff fatigue.

Model Deployment Biases

With AI system adoption incentives in the form of workflow efficiency, an over-reliance on AI systems and a progressive de-skilling of the workforce is of genuine concern[47]. Several biases can be introduced following model deployment directly related to these factors. These biases are listed with examples in Table 4.

Automation bias. In the context of healthcare AI applications, automation bias reflects the inappropriate user adoption of inaccurate AI predictions due to the nature of their automated delivery. This bias can manifest in two forms: omission errors and commission errors. Omission errors occur when clinicians fail to notice or ignore errors made by AI tools, especially when the AI makes decisions based on complex details difficult for humans to detect or easily interpret. This is often exacerbated in fast-paced environments, such as radiology. Commission errors, on the other hand, arise when clinicians inappropriately place greater trust in, or follow an AI model's judgment despite conflicting clinical evidence[48].

Feedback Loop bias. Feedback Loop bias occurs when clinicians consistently adopt and accept AI recommendations, even when inaccurate, and these labels are then captured and reinforced by future training cycles. For example, the presentation of an AI labeled dataset to physicians for adjustment, rather than de-novo labeling of the raw data, can reinforce and propagate prior model biases[49].

Dismissal bias. In contrast to an over-reliance on AI systems, dismissal bias, commonly referred to as "alert fatigue", can occur when end-users begin to overlook or de-value AI-generated alerts or suggestions. This end-user introduced bias is often exacerbated by high rates of false positives that lead to a progressive distrust of a system's capabilities. This deployment bias has been shown to lead to the dismissal of appropriate critical warnings, with potential risk of patient harm[47,49].

Mitigating Bias in Healthcare AI: A Model Life-cycle Approach

Establishing standardized and repeatable approaches for mitigating bias is an expanding societal responsibility for AI-healthcare developers and providers[23]. This task must be recognized as both longitudinal and dynamic in nature, shifting throughout time based on evolving clinical practice, local population needs, and broader societal influence. A valuable concept for establishing frameworks for bias mitigation in AI healthcare applications is considering the 'AI Model life cycle', defining key stages where biases can enter, propagate, or persist[10].

An AI model's life cycle, as illustrated in Fig. 3, includes a conception, data collection and pre-processing, in-processing (algorithm development and validation), post-processing (clinical deployment), and post-deployment surveillance phase. Systematically considering bias across these sequential phases requires a multifaceted approach tailored to identify, quantify, and mitigate its impact on the core principles of fairness, equity, and equality, and maintain the ethical integrity of healthcare AI. Establishing definitions for what a meaningful bias is (e.g., one that is sufficient to mandate mitigation or usage warnings) is not a uniform task and must be independently assessed on a case-by-case basis. However, considerations should be given to setting nominal accuracy performance difference thresholds (e.g., demographic group differences) to enable the routine,

automated, and iterative surveillance of AI healthcare models across relevant demographics[15].

In the following sections, we provide a concise overview of each phase of the AI model lifecycle, respective recommendations for bias mitigation, and discuss the potential challenges and limitations of such strategies. Figure 4 provides an overview of detection and mitigation strategies relevant to each of these phases. Recommendations have been consolidated from various research teams and organizations, offering complementary insights into the effective implementation, management, and potential obstacles of adopting AI systems. We have also compiled a comprehensive overview of additional bias mitigation strategies, guidelines, and frameworks in AI healthcare models, which can be found in Supplementary Table 1.

Conception phase. As illustrated in Fig. 3, bias surveillance should begin at time of model conception, demanding a clear, clinically oriented research question from which areas of bias can then be envisioned. This, and all future phases of the model life cycle, should be systematically reviewed by a diverse and representative AI healthcare team including clinical experts, data scientists, institutional stakeholders, and members of underrepresented patient populations. During conception, teams should align with and document adherence to Diversity, Equity, and Inclusion (DEI) principles of the local institution while identifying plans for mitigating imbalances in team membership[50] (For further details, refer to Supplementary Table 1). Meetings focussed on refining the research question, populations affected, datasets available, and intended outcomes should concurrently consider any unintended consequences for specific groups defined by race, ethnicity, gender, sex at birth, age, or other socio-demographic characteristics. In these discussions, it is critical to actively eliminate implicit, systemic, and confirmation biases ensuring research questions and AI models are designed to maximally satisfy the principles of fairness and equity across diverse socio-demographic and socio-economic groups while maintaining ethical practice[51].

Bias mitigation strategies for the conception phase experience unique challenges given their need for upstream introduction. Embedding bias awareness during conceptualization requires prior education and training for all contributing members of the AI development team, which can be challenging to implement and sustain. Beyond improved awareness, critical thinking activities should be routinely engaged to overcome confirmation bias that can exist within teams, maintaining mindfulness of sensitive attribute biases such as age, gender, or ethnicity. This requires constant vigilance and willingness to question established norms. These challenges highlight a need for standardized approaches to stewarding early concepts prior to data processing, demanding a pause for consideration of systemic biases ingrained into historical policies or practice, and ensuring that target populations are fairly and equitably considered[50].

Data Collection Phase. Data collection efforts should aim to generate datasets that best reflect the diversity of the population each model is intended to serve. It is essential to design data collection processes to capture an appropriate breadth of demographics, thereby allowing for the reporting and meaningful consideration of each dataset's relevance to the target population. Special consideration should be paid to nuances in patient subgroups, such as ethnic diversity or unique characteristics, such as disabilities[52]. To address common biases of this phase, such as representation, sampling, selection, and measurement biases, practitioners are advised to refer to the specific mitigation strategies outlined in Fig. 4. For example, when training is dominated by historical data known to have inherent limitations or biases, a prospectively captured external validation dataset should be considered to ensure generalizability across relevant sub-cohorts[53]. During data collection, it is recommended to; use a variety of data sources to enhance diversity; engage accessible data initiatives such as Open Science Practices[34] and STANDING together[54] (For further details, refer to Supplementary Table 1); make informed decisions on the use of retrospective versus prospective data (acknowledging their respective biases and challenges); assess the accuracy and reliability of data to identify potential biases; and carefully consider inclusion and exclusion criteria as well as recruitment procedures of any clinical trial data being used. The latter commonly leads to an exclusion of specific groups of patients, limiting generalizability to real world practice[23].

Achieving unbiased, broadly representative healthcare datasets remains a formidable challenge. Addressing data sparsity for under-represented populations may not be feasible, while laudable efforts to prospectively improve data quality is time consuming. Both sampling and participation biases are notably difficult to eliminate, given that certain groups may be less inclined to participate in data collection or sharing activities, despite outreach efforts. Additionally, standardizing data collection methods to reduce measurement bias is resource intensive. While initiatives like Open Science Practices and STANDING[34,54] aspire to improve the diversity of data resources available for healthcare AI, vulnerable groups may remain under-represented given a relative abundance of resources from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations. Finally, certain socio-demographic features, such as gender orientation, remain poorly captured by health systems and are not reliably inferred. These ongoing limitations emphasize a rapidly evolving need for targeted data collection and democratization strategies to facilitate fair and equitable healthcare AI[23,52].

Pre-processing (Model Planning and Preparation) Phase. The pre-processing phase encompasses a range of tasks that are undertaken to clean and prepare raw data for model development. The mitigation of bias during this phase requires careful attention to the management of missing data, selection of relevant variables, and feature engineering to ensure appropriate data diversity, representation, and the generation of balanced sub-samples for model validation[10,55]. Failure to execute these techniques appropriately can introduce variance or sensitivity to data shifts, underscoring the importance of meticulous attention to prevent bias propagation[33]. Specifically, it is recommended to review data collection methods and demographic distributions to maximize demographic representation, assess accuracy and stability of input variables across minority groups, and consider appropriate data augmentation techniques to address sub-group imbalance. Biases such as aggregation, missing data, feature selection, and representation biases are particularly prevalent during this phase, demanding focused attention to ensure these do not compromise model integrity[15,56]. Details surrounding specific pre-processing bias mitigation strategies are presented in Fig. 4.

Applying these strategies must be done with appropriate understanding of their limitations and should be tailored to specific contexts as inappropriate choices or execution can inadvertently amplify rather than mitigate bias. For example, while aiming to reduce aggregation bias, shifts toward disaggregated data can lead to overly granular datasets prone to noise, leading to reduced model generalizability[42]. When handling missing data bias, multiple imputation may introduce inaccuracies, particularly for non-random missingness or substantial data gaps[43]. Data augmentation to address sparsity can generate synthetic data that fails to appropriately reflect true diversity, thus reinforcing biases[42]. Accordingly, nuanced understandings of bias mitigation techniques and their appropriate use is required during this phase.

In-processing (Algorithm Development & Validation) Phase. In-processing represents all activities surrounding the training and validation phase of an AI algorithm. Potential biases introduced during this phase, including algorithmic, validation, and representation bias, must be intentionally sought and addressed[10]. We provide an example in Case Study 2 (Box 2). This demands iterative participation from the healthcare AI team to anticipate scenarios that could seed biased model behavior[50] Beyond stratified sub-group analyses, counterfactual examples should be considered to test these hypotheses during validation, purposely altering a candidate feature (e.g., ethnicity) to assess its systematic (biased) influence on model performance[57] (For further details, refer to

Fig. 4 | Bias mitigation strategies across AI model life cycle. This figure outlines the key strategies for addressing biases at different phases of the AI model life cycle. It highlights critical interventions from the initial conception phase through to post-deployment surveillance, ensuring that each phase in the AI development process incorporates practices aimed at promoting fairness, equity, and effectiveness. The strategies are categorized by the model's lifecycle phases, providing a roadmap for systematic bias mitigation in AI applications.

Box 2 | Real-World Case Study 2

Addressing Representation Bias (Racial) in Deep Learning-Based Cardiac MRI Segmentation.

A deep learning model (nnU-Net) for cardiac MRI segmentation, trained on UK Biobank data of 5,903 subjects (80% White, 20% Black, Asian, Chinese, Mixed, and Other), initially showed racial bias, with the Dice Similarity Coefficient (DSC) at 93.5% for White subjects but as low as 84.5% for Black and Mixed-race subjects. Researchers used three distinct strategies to address this: (1) stratified batch sampling (balancing racial groups in each batch during pre-processing phase) improved DSC for Black subjects from 85.88% to 93.07% and for Mixed-race from 84.52% to 93.84%, though overall accuracy decreased slightly; (2) fair meta-learning (using a secondary classifier to predict race during in-processing phase) raised DSC to 89.60% for Black and 88.03% for Mixed-race but increased complexity; and (3) protected group models (training separate models for each group during in-processing phase) achieved the best results, with DSC reaching 92.15% for Black and 93.17% for Mixed-race subjects, reducing bias to a standard deviation of 0.89. However, this approach required racial information during inference, limiting its practicality[99].

Supplementary Table 1). However, generating meaningful counterfactual examples requires a deep understanding of the dataset and relationships between features. Moreover, if not carefully implemented, counterfactual examples can lead to overfitting or produce unrealistic scenarios, thereby reducing the model's real-world applicability. Additional limitations include difficulties in finding sufficient representative data for counterfactual examples, which could skew the results[57]. An alternate, albeit resource and time intensive method to identifying algorithmic bias is "Red Teaming"[58]. This is a process where an independent team attempts to identify biases or other vulnerabilities in an AI model, determining whether certain conditions, such as unique demographic distributions, alters performance. This may not be practical for all organizations, especially for small sized teams with limited budgets.

Under-representation of minority classes is a ubiquitously encountered challenge for healthcare datasets that should be acknowledged during model training and validation[59]. To address class imbalance, strategies such as resampling to balance class distributions[60], synthetic data generation (such as Synthetic Minority Over-sampling Technique (SMOTE))[61], and the application of cost-sensitive learning to emphasize minority class errors should be considered[62]. The latter is a technique where misclassifying examples from the minority class is penalized more heavily than examples from the majority class, assisting in balancing model performance. However, these strategies have limitations. SMOTE, for example, generates synthetic samples by interpolating between minority class case examples, which can result in unrealistic data points that do not accurately reflect true variability[61]. Random Under-Sampling, on the other hand, risks discarding potentially valuable data from the majority class, leading to information loss[60]. Cost-sensitive learning can lead to overfitting, especially when costs assigned to the minority classes are high[62]. Regardless of the techniques employed, appropriate evaluation metrics for imbalanced datasets should be used, such as F1 score and precision-recall curves. These can also serve as cost functions during training to improve model generalizability[63].

Stratified cross-validation can assist in establishing representative class proportions within each fold to improve model generalizability[64]. This technique is reliant on the availability of large data resources and may not fully account for minority subgroups when not sufficiently represented. Fairness metrics like demographic parity, equal opportunity, equalized odds, and causal fairness (Table 5) can be leveraged to quantify and monitor for algorithmic bias. However, the application of these approaches can result in a "fairness-accuracy trade-off", as striving for equitable treatment across different groups may result in a reduction in overall model accuracy[59,60].

Federated learning techniques have gained popularity to improve model access to diverse datasets across unique healthcare environments, enabling a more collaborative and decentralized approach that ensures data privacy while reducing resource needs for model training[65] (For further details, refer to Supplementary Table 1). However, while enhancing model generalizability, federated learning inherently limits team access for data pre-processing or quality assurance tasks, limiting its appropriateness for specific applications.

Adversarial training, which involves training a model to be less influenced by sensitive attributes (e.g., race, gender) by introducing adversarial examples, can be effective in reducing representation bias (For further details, refer to Supplementary Table 1). However, this technique can be computationally intensive and may lead to reduced model accuracy if adversarial examples do not accurately reflect real-world variation, again introducing a "fairness-accuracy trade-off"[44].

Transfer learning is an approach that can be used to efficiently train models to perform tasks leveraging knowledge gained from historic models trained to perform similar or unique tasks. This can be used to fine-tune externally trained models using smaller quantities of local data, reducing the potential for external algorithmic bias. However, it must be recognized this may transfer unrecognized biases inherent to the original environment's dataset, potentially perpetuating these biases into new healthcare systems[53].

Model architecture choices can directly influence the transparency or interpretability of generated predictions. For example, while decision tree models provide clear feature usage insights, deep learning models offer limited insights[66]. However, techniques like LIME (Local Interpretable Model-agnostic Explanations)[67] and SHAP (Shapley Additive exPlanations) can help decipher feature importance in complex models[68] (For further details, refer to Supplementary Table 1). While valuable, these techniques can sometimes be inconsistent or fail to capture the true decision-making process of highly complex models, potentially leading to a false sense of interpretability[66]. The use of ensemble methods, where multiple models are combined to improve prediction accuracy, is one example where computational complexity is increased and interpretability reduced[59]. In general, model complexity should be minimized, and architectures chosen that maximize explainability, permitting the greater detectability and awareness of bias[66].

Following model optimization, external validation should be used whenever feasible to assess performance across diverse environments, patient demographics, and clinical characteristics. Validation across multiple independent settings is recommended for models intended for use beyond the local institution, such as for commercial distribution. Obtaining access to diverse external datasets for validation is often challenging due to privacy concerns, data-sharing restrictions, and variability of data quality across different institutions, however, is of paramount importance when models are intended for use beyond local environments[25]. The size and number of unique cohorts required for external validation varies based on the model's application and target population's diversity, but remains central to confirming performance consistency[53].

Finally, it is essential to meticulously document each algorithm's development methodologies, ensuring clear descriptions of the target population, its accuracy, and limitations[69].

Post-processing (Clinical Deployment) Phase. This phase encompasses a model's implementation in live clinical environments[70]. Human-

Table 5 | Available Types of Fairness Metrics

Fairness Metric	
Demographic Parity	Equation: $\|P(\hat{Y} = 1\|A = a) - P(\hat{Y} = 1\|A = a')\| \leq \varepsilon$ Description: • $\hat{Y}$ is the predicted outcome • A is the sensitive attribute (e.g., race, gender) • a and a' are different values of the sensitive attribute • ε is a small constant that defines the maximum acceptable difference in the probability of the positive outcome between the different groups. Focus: Demographic parity measures the statistical independence between the predicted outcome $\hat{Y}$ and the sensitive attribute A. It aims to ensure that the probability of the positive outcome is equal across different groups defined by the sensitive attribute.
Equal Opportunity	Equation: $\|P(\hat{Y} = 1\|A = a, Y = 1) - P(\hat{Y} = 1\|A = a', Y = 1)\| \leq \varepsilon$ Description: • $\hat{Y}$ is the predicted outcome • A is the sensitive attribute. • Y is the true outcome • a and a' are different values of the sensitive attribute • ε is a small constant that defines the maximum acceptable difference in the true positive rate between the different groups. Focus: Equal opportunity focuses on ensuring that the true positive rate (the probability of a positive prediction given a positive true outcome) is equal across different groups defined by the sensitive attribute. This metric aims to provide equal chances of being correctly identified as positive for all groups.
Equalized Odds	Equation: $P(\hat{Y} = 1\|A = a, Y = y) = P(\hat{Y} = 1\|A = a', Y = y)$ Description: • $\hat{Y}$ is the predicted outcome. • A is the sensitive attribute. • Y is the true outcome. • a and a′ are different values of the sensitive attribute, representing different groups. Focus: Equalized Odds ensures that a machine learning model performs consistently across different groups by equalizing the rates of false positives and true positives. This metric is vital as it balances prediction accuracy and errors, promoting fairness and preventing any group from suffering disproportionately from prediction mistakes.
Causal Fairness	Equation: $P(\hat{Y} = 1\|do(A = a)) = P(\hat{Y} = 1\|do(A = a'))$ Description: • $\hat{Y}$ is the predicted outcome • A is the sensitive attribute. • do(A = a) and do(A = a') represent the interventional distributions, where the sensitive attribute A is set to a and a' respectively, while all other variables are held constant. Focus: Causal fairness is based on the principles of causal inference and aims to ensure that the predicted outcome $\hat{Y}$ is independent of the sensitive attribute A, even when intervening on A. This metric focuses on the causal relationship between the sensitive attribute and the predicted outcome rather than just the statistical independence.

machine interface design and choice of platform deployment will influence accessibility, while planned reimbursement models can threaten equitable delivery across socio-economic groups. Adherence to Human-in-the-loop (HITL) strategies, where human experts review all model predictions, is recommended for clinical decision making[71,72].

Transparent disclosure of each model's training population demographic distributions is recommended to declare potential biases and to avoid using models in under-represented populations. The reporting of model performance for relevant sub-populations, as permitted by data resources, is also strongly recommended. Model threshold adjustments can deliver improved responsiveness to user-entered data, such as a patient's or clinician's preferences, recognizing individual opinions or beliefs surrounding a given prediction task[73].

As stated earlier, tools to enhance model explainability for end-users are available and may improve both trust and adoption. Simple but informative ways to incorporate such insights should be considered to improve transparency and interpretability. For image-based predictions, saliency maps can be employed to highlight regions where model predictions are most strongly influenced. For traditional machine learning (ML) models, SHAP values can be used to demonstrate the relative importance and influence of data features[74]. It must be recognized that these tools may fall short in explaining complex model behaviors and may provide oversimplified or misleading insights that could falsely increase end-user trust.

Structured pre-deployment testing across different clinical environments and populations is recommended to identify unforeseen biases in human-machine interactions. This includes shadow deployment in live clinical environments where model results do not influence clinical behavior but are assessed for their calibration, end-user adoption, and user experience to identify barriers to fair and equitable use[75,76]. This process can take time and delay the model's full implementation.

An example of implementing these strategies is seen in the DECIDE-AI guidelines (Developmental and Exploratory Clinical Investigation of DEcision support systems driven by Artificial Intelligence)[77], which offer a structured approach for the early-stage clinical deployment of AI decision support systems (For further details, refer to Supplementary Table 1). By focusing on human-AI interaction, transparent reporting, and iterative validation, these guidelines aim to ensure AI models are both effective and safe in clinical practice. Such frameworks are critical to ensure the transparency and safety of AI models in clinical environments, however, are resource-intensive and time-consuming. Therefore, centralization of institutional resources and processes to support these pathways is essential.

Post-deployment Surveillance Phase. This final phase encompasses post-deployment activities for AI models in active healthcare environments, inclusive of performance surveillance, model maintenance, and re-calibration (data augmentation and re-weighting)[78]. Mechanisms to monitor user engagement, decision impact, and model accuracy versus standard care pathways should be adopted and purposely bridged to patient demographics to identify biased sub-group behavior or downstream inequities in clinical benefit. This is a life-long process, recognizing the potential for concept drift, feedback loop-bias, degradation in

fairness metrics, or new biases emerging over time (Fig. 4). As a novel and emerging challenge for healthcare institutions, attention must be raised to administrators and practitioners that data destined for live AI models must be considered as a regulated data product, demanding ongoing quality assurance and maintenance. In this context, adhering to established guidelines and frameworks is essential to ensure sustained accuracy and equity of AI algorithms[79]. The FDA's Proposed Regulatory Framework for AI/ML-based Software as a Medical Device (SaMD) emphasizes a need for real-world performance monitoring, inclusive of tracking model performance to identify established or emerging bias[80]. While this places meaningful responsibility on AI model providers, healthcare institutions must hold responsibility to maintain an awareness of their local populations and clinical practice shifts that are commonly opaque to commercial platforms. Accordingly, programmatic approaches to AI model surveillance in clinical environments is an expanding priority for healthcare providers[81].

Future Directions

Given the rapid expansion of AI healthcare innovation, there is immediate need for the meaningful incorporation of DEI principles across all phases of the AI model life cycle, inclusive of structured bias surveillance and mitigation frameworks[82]. This must be inclusive of actively educating and training a more diverse and representative AI developer community, implementing, and expanding institutional programs focussed on AI model assessment and surveillance, and development of AI healthcare-specific clinical practice guidelines. Adequately addressing these needs will remain a challenge due to the rapid pace of AI advancement relative to legislative, regulatory, and practice guideline development. Regardless, policy makers, clinicians, researchers, and patient advocacy groups must coordinate to enhance diversity in AI healthcare models.

Incorporating DEI principles must be perceived as an essential priority, particularly given a lack of representation in current regulatory guidelines for AI applications[50,83]. Moreover, there is a critical need to integrate AI and machine learning content into medical training curricula. This will prepare healthcare professionals for a future where data-driven decision-making is increasingly considered the standard of care. Understanding AI, its potential biases, and ethical implications will therefore be crucial for these individuals to appropriately contribute to its refinement and appropriate clinical use[84].

Conclusion

In the evolving landscape of healthcare delivery, one increasingly influenced by AI technology, recognizing, and mitigating bias is a priority. While essential for achieving accuracy and reliability from AI innovations, addressing bias is core to upholding the ethical standards of healthcare, ensuring a future where care is delivered with fairness and equity. This ensures that AI will serve as a tool for bridging gaps in healthcare, not widening them.

Data availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study. This review article is based on previously published studies and does not contain original research data. All data supporting the findings of this study are available within the article and its references.

Received: 17 June 2024; Accepted: 6 February 2025;
Published online: 11 March 2025

References

1. Health, C. for D. and R. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. *FDA* (2024).
2. Al Kuwaiti, A. et al. A review of the role of artificial intelligence in healthcare. *J. Pers. Med* **13**, 951 (2023).
3. WHO calls for safe and ethical AI for health. https://www.who.int/news/item/16-05-2023-who-calls-for-safe-and-ethical-ai-for-health.
4. WHO outlines considerations for regulation of artificial intelligence for health. https://www.who.int/news/item/19-10-2023-who-outlines-considerations-for-regulation-of-artificial-intelligence-for-health.
5. Da Silva, M., Flood, C. M., Goldenberg, A. & Singh, D. Regulating the Safety of Health-Related Artificial Intelligence. *Health. Policy* **17**, 63–77 (2022).
6. AI pitfalls and what not to do: mitigating bias in AI | British Journal of Radiology | Oxford Academic. https://academic.oup.com/bjr/article/96/1150/20230023/7498925.
7. Directorate-General for Parliamentary Research Services (European Parliament), Lekadir, K., Quaglio, G., Tselioudis Garmendia, A. & Gallin, C. *Artificial Intelligence in Healthcare: Applications, Risks, and Ethical and Societal Impacts*. (Publications Office of the European Union, 2022).
8. Grant, M. J. & Booth, A. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Inf. Libraries J.* **26**, 91–108 (2009).
9. Culyer, A. J. & Wagstaff, A. Equity and equality in health and health care. *J. Health Econ.* **12**, 431–457 (1993).
10. Nazer, L. H. et al. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health* **2**, e0000278 (2023).
11. DeCamp, M. & Lindvall, C. Latent bias and the implementation of artificial intelligence in medicine. *J. Am. Med. Inform. Assoc.* **27**, 2020–2023 (2020).
12. Hanson, B. et al. Garbage in, garbage out: mitigating risks and maximizing benefits of AI in research. *Nature* **623**, 28–31 (2023).
13. Burlina, P., Joshi, N., Paul, W., Pacheco, K. D. & Bressler, N. M. Addressing Artificial Intelligence Bias in Retinal Diagnostics. *Transl. Vis. Sci. Technol.* **10**, 13 (2021).
14. Kumar, A. et al. Artificial intelligence bias in medical system designs: a systematic review. *Multimed. Tools Appl* **83**, 18005–18057 (2024).
15. Chin, M. H. et al. Guiding Principles to Address the Impact of Algorithm Bias on Racial and Ethnic Disparities in Health and Health Care. *JAMA Netw. Open* **6**, e2345050 (2023).
16. Page, M. J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **372**, n71 (2021).
17. Wolff, R. F. et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann. Intern Med* **170**, 51–58 (2019).
18. Chen, Z. et al. Evaluation of Risk of Bias in Neuroimaging-Based Artificial Intelligence Models for Psychiatric Diagnosis: A Systematic Review. *JAMA Netw. Open* **6**, e231671 (2023).
19. Giovanola, B. & Tiribelli, S. Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. *AI Soc.* **38**, 549–563 (2023).
20. Pu, L. Fairness of the Distribution of Public Medical and Health Resources. *Front Public Health* **9**, 768728 (2021).
21. Fletcher, R. R. Nakeshimana, A. & Olubeko, O. Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health. *Front. Artif. Intell.* **3**, 561802 (2021).
22. Xu, J. et al. Algorithmic fairness in computational medicine. *eBioMedicine* **84**, 104250 (2022).
23. Abràmoff, M. D. et al. Considerations for addressing bias in artificial intelligence for health equity. *npj Digit. Med.* **6**, 1–7 (2023).
24. Jackson, M. C. Artificial Intelligence & Algorithmic Bias: The Issues with Technology Reflecting History & Humans Notes & Comments. *J. Bus. Tech. L.* **16**, 299–316 (2021).
25. Celi, L. A. et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. *PLOS Digital Health* **1**, e0000022 (2022).
26. FitzGerald, C. & Hurst, S. Implicit bias in healthcare professionals: a systematic review. *BMC Med Ethics* **18**, 19 (2017).
27. Implicit bias of encoded variables: frameworks for addressing structured bias in EHR–GWAS data | Human Molecular Genetics |

Oxford Academic. https://academic.oup.com/hmg/article/29/R1/R33/5899023?login=true.
28. Feagin, J. & Bennefield, Z. Systemic racism and U.S. health care. *Soc. Sci. Med.* **103**, 7–14 (2014).
29. Payne, B. K. & Hannay, J. W. Implicit bias reflects systemic racism. *Trends Cogn. Sci.* **25**, 927–936 (2021).
30. Elston, D. M. Confirmation bias in medical decision-making. *J. Am. Acad. Dermatol.* **82**, 572 (2020).
31. Althubaiti, A. Information bias in health research: definition, pitfalls, and adjustment methods. *J. Multidiscip. Healthc.* **9**, 211–217 (2016).
32. Feng, Q., Du, M., Zou, N. & Hu, X. Fair Machine Learning in Healthcare: A Review. Preprint at https://doi.org/10.48550/arXiv.2206.14397 (2024).
33. Chen, R. J. et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat. Biomed. Eng.* **7**, 719–742 (2023).
34. Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D. & Tzovara, A. Addressing bias in big data and AI for health care: A call for open science. *Patterns (N. Y)* **2**, 100347 (2021).
35. Ekpo, E. et al. Underrepresentation of Women in Reduced Ejection Heart Failure Clinical Trials With Improved Mortality or Hospitalization. *JACC: Adv.* **3**, 100743 (2024).
36. Gomez, S. E., Sarraju, A. & Rodriguez, F. Racial and Ethnic Group Underrepresentation in Studies of Adverse Pregnancy Outcomes and Cardiovascular Risk. *J. Am. Heart Assoc.* **11**, e024776 (2022).
37. Scharff, D. P. et al. More than Tuskegee: Understanding Mistrust about Research Participation. *J. Health Care Poor Underserved* **21**, 879–897 (2010).
38. Gaube, S. et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digit. Med.* **4**, 1–8 (2021).
39. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv. (CSUR)* **54**, 1–35 (2021).
40. Castro, D. C., Walker, I. & Glocker, B. Causality matters in medical imaging. *Nat. Commun.* **11**, 3673 (2020).
41. Dehkharghanian, T. et al. Biased data, biased AI: deep networks predict the acquisition site of TCGA images. *Diagnostic Pathol.* **18**, 67 (2023).
42. Cirillo, D. et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *npj Digit. Med.* **3**, 1–11 (2020).
43. Nijman, S. et al. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *J. Clin. Epidemiol.* **142**, 218–229 (2022).
44. Siddique, S. et al. Survey on Machine Learning Biases and Mitigation Techniques. *Digital* **4**, 1–68 (2024).
45. Lu, M. Y. et al. AI-based pathology predicts origins for cancers of unknown primary. *Nature* **594**, 106–110 (2021).
46. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
47. Yoo, J., Hur, S., Hwang, W. & Cha, W. C. Healthcare Professionals' Expectations of Medical Artificial Intelligence and Strategies for its Clinical Implementation: A Qualitative Study. *Health. Inf. Res* **29**, 64–74 (2023).
48. Neri, E., Coppola, F., Miele, V., Bibbolino, C. & Grassi, R. Artificial intelligence: Who is responsible for the diagnosis? *Radio. Med* **125**, 517–521 (2020).
49. Ueda, D. et al. Fairness of artificial intelligence in healthcare: review and recommendations. *Jpn J. Radio.* **42**, 3–15 (2024).
50. Cachat-Rosset, G. & Klarsfeld, A. Diversity, Equity, and Inclusion in Artificial Intelligence: An Evaluation of Guidelines. *Appl. Artif. Intell.* **37**, 2176618 (2023).
51. Wiens, J. et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med* **25**, 1337–1340 (2019).
52. Ramirez, A. H. et al. The All of Us Research Program: Data quality, utility, and diversity. *Patterns (N. Y)* **3**, 100570 (2022).
53. Yang, J., Soltan, A. A. S. & Clifton, D. A. Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. *npj Digit. Med.* **5**, 1–8 (2022).
54. Ganapathi, S. et al. Tackling bias in AI health datasets through the STANDING Together initiative. *Nat. Med* **28**, 2232–2233 (2022).
55. Kamiran, F. & Calders, T. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* **33**, 1–33 (2012).
56. Albahra, S. et al. Artificial intelligence and machine learning overview in pathology & laboratory medicine: A general review of data preprocessing and basic supervised concepts. *Semin. Diagnostic Pathol.* **40**, 71–87 (2023).
57. Wachter, S., Mittelstadt, B. & Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* **31**, 841 (2017).
58. Feffer, M., Sinha, A., Deng, W. H., Lipton, Z. C. & Heidari, H. Red-Teaming for generative AI: Silver bullet or security theater? In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Vol. 7, 421–437 https://doi.org/10.1609/aies.v7i1.31647 (2024).
59. Wang, Y.-C. & Cheng, C.-H. A multiple combined method for rebalancing medical data with class imbalances. *Computers Biol. Med.* **134**, 104527 (2021).
60. Kim, A. & Jung, I. Optimal selection of resampling methods for imbalanced data with high complexity. *PLoS One* **18**, e0288540 (2023).
61. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
62. Ling, C. X. & Sheng, V. S. Cost-Sensitive Learning and the Class Imbalance Problem.
63. Scholz, D. et al. (2024). Imbalance-aware loss functions improve medical image classification. in.
64. Wilimitis, D. & Walsh, C. G. Practical Considerations and Applied Examples of Cross-Validation for Model Development and Evaluation in Health Care: Tutorial. *JMIR AI* **2**, e49023 (2023).
65. Nguyen, D. C. et al. Federated Learning for Smart Healthcare: A Survey. *ACM Comput. Surv.* **55**, 1–37 (2023).
66. Linardatos, P. Papastefanopoulos, V. & Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **23**, 18 (2021).
67. Zhang, Y., Song, K., Sun, Y., Tan, S. & Udell, M. 'Why Should You Trust My Explanation?' Understanding Uncertainty in LIME Explanations. Preprint at https://doi.org/10.48550/arXiv.1904.12991 (2019).
68. Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent Individualized Feature Attribution for Tree Ensembles. Preprint at https://doi.org/10.48550/arXiv.1802.03888 (2019).
69. Amann, J. et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inf. Decis. Mak.* **20**, 310 (2020).
70. Challen, R. et al. Artificial intelligence, bias and clinical safety. *BMJ Qual. Saf.* **28**, 231–237 (2019).
71. Mittermaier, M., Raza, M. & Kvedar, J. C. Collaborative strategies for deploying AI-based physician decision support systems: challenges and deployment approaches. *NPJ Digital Med.* **6**, 137 (2023).
72. Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J. & Fernández-Leal, Á. Human-in-the-loop machine learning: a state of the art. *Artif. Intell. Rev.* **56**, 3005–3054 (2023).
73. Pfohl, S. et al. Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare. in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* 1039–1052 https://doi.org/10.1145/3531146.3533166 (Association for Computing Machinery, New York, NY, USA, 2022).
74. Dykstra, S. et al. Machine learning prediction of atrial fibrillation in cardiovascular patients using cardiac magnetic resonance and electronic health information. *Front. Cardiovasc. Med.* **9**, 998558 (2022).

75. Bizzo, B. C. et al. Addressing the Challenges of Implementing Artificial Intelligence Tools in Clinical Practice: Principles From Experience. *J. Am. Coll. Radiol.* **20**, 352–360 (2023).
76. Daye, D. et al. Implementation of Clinical Artificial Intelligence in Radiology: Who Decides and How? *Radiology* **305**, 555–563 (2022).
77. Vasey, B. et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat. Med* **28**, 924–933 (2022).
78. Thomas, L. et al. Real-world post-deployment performance of a novel machine learning-based digital health technology for skin lesion assessment and suggestions for post-market surveillance. *Front Med (Lausanne)* **10**, 1264846 (2023).
79. Widner, K. et al. Lessons learned from translating AI from development to deployment in healthcare. *Nat. Med* **29**, 1304–1306 (2023).
80. Health, C. for D. and R. Artificial Intelligence and Machine Learning in Software as a Medical Device. *FDA* (2023).
81. Abràmoff, M. D. et al. Foundational Considerations for Artificial Intelligence Using Ophthalmic Images. *Ophthalmology* **129**, e14–e32 (2022).
82. Siala, H. & Wang, Y. SHIFTing artificial intelligence to be responsible in healthcare: A systematic review. *Soc. Sci. Med.* **296**, 114782 (2022).
83. Nyariro, M., Emami, E. & Abbasgholizadeh Rahimi, S. Integrating Equity, Diversity, and Inclusion throughout the lifecycle of Artificial Intelligence in health. in *13th Augmented Human International Conference* 1–4 https://doi.org/10.1145/3532530.3539565 (Association for Computing Machinery, New York, NY, USA, 2022).
84. Grunhut, J., Marques, O. & Wyatt, A. T. M. Needs, Challenges, and Applications of Artificial Intelligence in Medical Education Curriculum. *JMIR Med. Educ.* **8**, e35587 (2022).
85. Lai, J. C., Pomfret, E. A. & Verna, E. C. Implicit bias and the gender inequity in liver transplantation. *Am. J. Transpl.* **22**, 1515–1518 (2022).
86. Zbierajewski-Eischeid, S. J. & Loeb, S. J. Myocardial infarction in women: promoting symptom recognition, early diagnosis, and risk assessment. *Dimens Crit. Care Nurs.* **28**, 1–6 (2009). ; quiz 7–8.
87. McGowan, S. K., Sarigiannis, K. A., Fox, S. C., Gottlieb, M. A. & Chen, E. Racial Disparities in ICU Outcomes: A Systematic Review. *Crit. Care Med.* **50**, 1 (2022).
88. Daneshjou, R. et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Sci Adv* **8**, eabq6147.
89. Heslin, K. C. et al. Trends in Opioid-related Inpatient Stays Shifted After the US Transitioned to ICD-10-CM Diagnosis Coding in 2015. *Med. Care* **55**, 918 (2017).
90. Puyol-Antón, E. et al. Fairness in Cardiac Magnetic Resonance Imaging: Assessing Sex and Racial Bias in Deep Learning-Based Segmentation. *Front. Cardiovasc. Med.* **9**, 859310 (2022).
91. Guo, L. N., Lee, M. S., Kassamali, B., Mita, C. & Nambudiri, V. E. Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection-A scoping review. *J. Am. Acad. Dermatol* **87**, 157–159 (2022).
92. Rusanov, A., Weiskopf, N. G., Wang, S. & Weng, C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. *BMC Med Inf. Decis. Mak.* **14**, 51 (2014).
93. Enzenbach, C., Wicklein, B., Wirkner, K. & Loeffler, M. Evaluating selection bias in a population-based cohort study with low baseline participation: the LIFE-Adult-Study. *BMC Med. Res. Methodol.* **19**, 135 (2019).
94. Kaźmierczak, I., Zajenkowska, A., Rogoza, R., Jonason, P. K. & Ścigała, D. Self-selection biases in psychological studies: Personality and affective disorders are prevalent among participants. *PLOS ONE* **18**, e0281046 (2023).
95. Delgado, J. et al. Bias in algorithms of AI systems developed for COVID-19: A scoping review. *J. Bioeth. Inq.* **19**, 407–419 (2022).
96. Dratsch, T. et al. Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance. *Radiology* **307**, e222176 (2023).
97. Wysocki, O. et al. Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artif. Intell.* **316**, 103839 (2023).
98. Drew, B. J. et al. Insights into the Problem of Alarm Fatigue with Physiologic Monitor Devices: A Comprehensive Observational Study of Consecutive Intensive Care Unit Patients. *PLoS One* **9**, e110274 (2014).
99. Puyol-Anton, E. et al. Fairness in cardiac MR image analysis: an investigation of bias due to data imbalance in deep learning-based segmentation. *Med. Image. ComputComput. Assist. Interv. – MICCAI* **12903**, 413–423 (2021).

Author contributions

F.H. conducted all literature reviews, synthesized co-author contributions, and drafted all components of the manuscript. J.A.W. led conceptual framework development, contribution coordination, and performed structural revisions throughout the writing process. C.B.J., G.W., D.A., and Z.A. each provided domain-relevant expertise and guidance toward manuscript structure and content, performed content revision, and made edits to the final collated manuscript. All authors have read and approved the final version of the manuscript.

Competing interests

Dr. Demilade Adedinsewo is supported by the Mayo Building Interdisciplinary Research Careers in Women's Health (BIRCWH) Program funded by the National Institutes of Health [grant number K12 AR084222]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. All other authors declare no competing financial or non-financial interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01503-7.

Correspondence and requests for materials should be addressed to James A. White.