PLOS DIGITAL HEALTH

REVIEW

AI-driven healthcare: A review on ensuring fairness and mitigating bias

Sribala Vidyadhari Chinta[1], **Zichong Wang**[1], **Avash Palikhe**[1], **Xingyu Zhang**[2], **Ayesha Kashif**[3], **Monique Antoinette Smith**[4], **Jun Liu**[5*], **Wenbin Zhang**[1*]

1 Florida International University, Miami, Florida, United States of America, **2** University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America, **3** Jose Marti MAST 6-12 Academy, Hialeah, Florida, United States of America, **4** Emory University, Atlanta, Georgia, United States of America, **5** Carnegie Mellon University, Pittsburgh, Pennsylvania, United States of America

* liujun@cmu.edu (JL); wenbin.zhang@fiu.edu (WZ)

Check for updates

OPEN ACCESS

Citation: Chinta SV, Wang Z, Palikhe A, Zhang X, Kashif A, Smith MA, et al. (2025) AI-driven healthcare: A review on ensuring fairness and mitigating bias. PLOS Digit Health 4(5): e0000864. https://doi.org/10.1371/journal.pdig.0000864

Editor: Po-Chih Kuo, National Tsing-Hua University: National Tsing Hua University, TAIWAN

Published: May 26, 2025

Copyright: © 2025 Chinta et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported in part by the National Science Foundation (NSF) (Grant No. 2404039 to WZ) and the National Institutes of Health (NIH) (Grant No. R01MD019814 to WZ). The funders provided support for Wenbin Zhang (WZ) who contributed to the study through funding acquisition, conceptualization, formal analysis, investigation, methodology, project administration, supervision, writing of the original draft, and review and editing of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Artificial intelligence (AI) is rapidly advancing in healthcare, enhancing the efficiency and effectiveness of services across various specialties, including cardiology, ophthalmology, dermatology, emergency medicine, etc. AI applications have significantly improved diagnostic accuracy, treatment personalization, and patient outcome predictions by leveraging technologies such as machine learning, neural networks, and natural language processing. However, these advancements also introduce substantial ethical and fairness challenges, particularly related to biases in data and algorithms. These biases can lead to disparities in healthcare delivery, affecting diagnostic accuracy and treatment outcomes across different demographic groups. This review paper examines the integration of AI in healthcare, highlighting critical challenges related to bias and exploring strategies for mitigation. We emphasize the necessity of diverse datasets, fairness-aware algorithms, and regulatory frameworks to ensure equitable healthcare delivery. The paper concludes with recommendations for future research, advocating for interdisciplinary approaches, transparency in AI decision-making, and the development of innovative and inclusive AI applications.

Author summary

In this paper, we investigate the rapid advancement of artificial intelligence (AI) in healthcare, focusing on its role in improving efficiency and effectiveness across specialties such as cardiology, ophthalmology, and dermatology. We note that AI technologies enhance diagnostic accuracy, treatment personalization, and patient outcome predictions. However, these developments also pose significant ethical challenges, particularly concerning biases in data and algorithms that can create disparities in healthcare delivery. We examine the integration of AI in healthcare, highlighting the critical challenges related to bias and exploring strategies for mitigation. We emphasize the need for diverse

datasets, fairness-aware algorithms, and regulatory frameworks to ensure equitable healthcare delivery. Our paper concludes with recommendations for future research, advocating for interdisciplinary approaches, transparency in AI decision-making, and the development of innovative and inclusive AI applications.

Introduction

Artificial intelligence (AI) is revolutionizing modern healthcare, dramatically transforming the ways we diagnose, treat, and manage diseases. The integration of AI into healthcare began in the late 20th century with systems like MYCIN [1] in the 1970s, which helped diagnose infections and recommend antibiotics, and CADUCEUS [2] in the 1980s, which emulated human diagnostic reasoning. These early systems laid the groundwork for today's advanced machine learning and deep learning techniques, which now significantly enhance diagnostic accuracy, treatment personalization, and patient outcome predictions.

As AI technologies have advanced, their impact on healthcare has grown exponentially. Modern AI applications, particularly deep learning, have enhanced image recognition, significantly improving diagnostic accuracy in fields such as radiology and pathology [3]. Predictive analytics, powered by AI, are essential in patient monitoring and management, using real-time data to forecast potential patient deteriorations [4]. Additionally, natural language processing (NLP) tools have revolutionized the handling of unstructured data, improving the functionality of electronic health record systems and facilitating more comprehensive patient care [5].

Several key algorithms and technologies underpin these advancements. Neural networks, particularly convolutional neural networks (CNNs), are extensively used for medical image analysis, aiding in the detection and characterization of various pathological findings [6]. Decision support systems incorporate diverse data, including genetic profiles and prior health records, to optimize treatment strategies [7]. Among the notable AI tools, IBM Watson stands out for its application in cancer treatment, although its widespread adoption faces challenges [8].

The integration of AI into healthcare has the potential to enhance diagnostic and operational efficiencies while assisting in reducing human error [9]. This is due to AI's ability to manage and analyze large datasets, enabling more effective resource allocation and efficient patient scheduling, both of which can lead to improved patient outcomes and satisfaction [10]. In addition, AI-driven predictive models have proven useful in public health for tracking disease patterns and aiding in the management of epidemics, as seen during the COVID-19 pandemic [11]. These capabilities demonstrate AI's potential to enhance healthcare delivery.

However, alongside these advancements, substantial ethical and fairness challenges have emerged. Biases embedded within training data can lead to skewed AI models, resulting in disparities in healthcare outcomes across different demographic groups [12]. For instance, an algorithm used in US hospitals was biased against black patients in resource allocation [13], and dermatological AI showed lower diagnostic accuracy for conditions like melanoma in darker-skinned individuals due to training primarily on fair-skinned images [14]. Similarly, AI tools for diagnosing depression have faced challenges when applied across different linguistic and cultural backgrounds because they were primarily trained on English-speaking, Western populations, leading to potential misdiagnoses in non-Western patients [15]. AI models trained predominantly on data from specific populations have exhibited lower diagnostic accuracy for underrepresented groups, exacerbating existing health disparities. These

examples highlight the urgent need for diverse datasets and transparent AI systems to ensure fairness and equity in healthcare AI applications.

Efforts to address these challenges have increasingly gained attention from governments and regulatory bodies. Recognizing the potential of AI to both improve and exacerbate healthcare disparities, policymakers have taken initial steps toward creating safeguards. The European Union's General Data Protection Regulation (GDPR), for example, provides a framework for ethical considerations in AI applications by addressing issues like data privacy and transparency [16]. In the United States, the Food and Drug Administration (FDA) has begun implementing guidelines to evaluate the safety and effectiveness of medical AI systems [17]. While these initiatives mark critical progress, they primarily focus on high-level governance and oversight, leaving significant gaps in addressing the technical challenges of bias detection and mitigation. Without comprehensive strategies that integrate policy, technical methodologies, and ethical considerations, these efforts risk being insufficient to tackle the systemic biases present in AI systems.

To bridge these gaps, this paper provides a significant contribution to the field of AI-driven healthcare by offering a comprehensive framework for understanding and addressing fairness and bias. It categorizes biases across the machine learning pipeline and aligns them with targeted detection and mitigation strategies. By integrating technical, ethical, and policy perspectives, the paper addresses key challenges in creating equitable AI systems and highlights actionable solutions. Through real-world examples from diverse healthcare domains, it bridges theoretical concepts with practical implementation. This work not only deepens the understanding of bias in healthcare AI but also provides researchers, practitioners, and policymakers with valuable tools and strategies for fostering fairness and inclusivity in AI applications.

The paper is structured to help readers follow its key contributions. Sect 1, discusses the applications of AI in various healthcare dimensions. The biases exhibited by those AI applications, their root causes, potential consequences, and fairness metrics and trade-offs have been discussed in Sect 2. Sect 3 explores various approaches to address and mitigate bias. Sect 4 identifies research gaps and future directions for enhancing AI in healthcare. This paper is concluded in Sect 5.

1. Applications of AI in healthcare

AI has emerged as a transformative force in healthcare, leveraging advanced algorithms, data analytics, and machine learning techniques. By processing vast amounts of health data with remarkable speed and precision, AI systems are enhancing diagnostic accuracy, enabling personalized treatment plans, and ultimately elevating patient outcomes. This section explores the diverse applications of AI across several key areas of healthcare, highlighting its impact and potential for future innovations.

1.1. Cardiology

AI has the potential to transform cardiology by enhancing diagnostic accuracy, supporting personalized treatment, and contributing to improved patient care [18,19]. Specifically, AI algorithms have proven effective in analyzing cardiovascular data to aid in the early detection and diagnosis of conditions such as arrhythmias, heart failure, and coronary artery disease [20,21]. For instance, machine learning models have shown high precision in interpreting echocardiograms, at times surpassing human experts in diagnostic speed and accuracy [22]. Additionally, AI holds potential for managing cardiovascular risk by integrating

diverse patient data to predict individual risk factors [23]. It also contributes to the development of personalized treatment plans by analyzing patient data and predicting responses to various treatment modalities, optimizing therapeutic decisions [24]. In oncology, AI tools like MesoNet have been used to predict patient survival by analyzing whole-slide digitized images. These tools have demonstrated greater accuracy in predicting patient survival compared to current pathology practices, offering new insights and potentially guiding better treatment decisions [25]. However, the integration of AI in clinical practice requires rigorous validation and careful ethical considerations to prioritize patient safety and uphold data security [26]. The continuous evolution of AI technologies holds promise for advancing diagnostic tools and therapeutic strategies, potentially improving precision and predictive capabilities in whole-person care. This could include recommending lifestyle changes tailored to individual contexts alongside medications adjusted to their genomic profiles [27]. Moreover, AI could play a role in supporting health equity by facilitating access to advanced cardiac care in underserved areas through telemedicine platforms and remote monitoring, which could help reduce disparities in cardiovascular health outcomes [28].

1.2. Ophthalmology

AI has demonstrated significant potential in ophthalmology by improving diagnostic accuracy for common eye diseases and streamlining clinical workflows. Deep learning models have been particularly effective in analyzing complex visual data from imaging techniques such as fundus photography and optical coherence tomography (OCT) [29]. These models are very good at diagnosing common eye diseases like diabetic retinopathy, glaucoma, and age-related macular degeneration, sometimes achieving expert-level sensitivity and specificity [30]. Automated analysis using AI not only speeds up the diagnostic process but also reduces human error, facilitating earlier and more precise interventions [31]. Furthermore, AI's predictive capabilities are being harnessed to forecast disease progression, which is crucial for conditions like glaucoma, where early detection can prevent severe vision loss [6]. Using AI tools in clinical practice also includes using AI-driven decision support systems that help plan treatments by guessing how different types of treatments will work [32]. Despite these advancements, challenges remain, including data privacy concerns, the need for large annotated datasets for training algorithms, and the integration of AI into existing clinical workflows [33]. However, ongoing research and collaboration between AI technologists and ophthalmic experts are likely to overcome these obstacles, solidifying AI's role in modern ophthalmology [34]. This further opens the frontier for the expansion of access to ophthalmological care in existing deserts through technology-enabled care models and non-specialist operators at the point of service, thereby addressing disparities in eye care availability and quality.

1.3. Dermatology

AI is advancing dermatology by improving diagnostic accuracy, personalizing treatments, and streamlining patient management. Deep learning models, such as CNNs, are effective in diagnosing skin cancer, demonstrating performance comparable to dermatologists in identifying melanomas and other skin conditions [35,36]. Moreover, AI applications in dermatology extend beyond cancer detection to managing chronic conditions like psoriasis and atopic dermatitis. These algorithms assist in monitoring disease progression and treatment response, enhancing decision-making in telemedicine and clinical workflows [37]. AI systems are also

utilized in cosmetic dermatology, optimizing treatment recommendations based on individual facial analysis [38]. Ensuring the efficacy and safety of AI tools requires rigorous validation processes with diverse datasets to mitigate bias and enhance generalizability [14]. Integration into clinical workflows and user training is crucial to maximizing the benefits of AI in dermatological practice [39]. As AI continues to evolve, continuous collaboration between technologists and clinicians is essential to address ethical considerations and improve patient outcomes in dermatology. Importantly, AI can democratize access to high-quality dermatological care by enabling remote consultations and diagnostics, thus reducing barriers for individuals in remote or underserved regions.

1.4. Neurology

AI has contributed to advancements in neurology by improving diagnostic accuracy and supporting research efforts. Techniques like machine learning and deep learning assist in analyzing imaging data for disorders such as Alzheimer's and epilepsy [5,40]. These technologies enhance the interpretation of MRI and CT scans, often achieving higher accuracy than traditional methods [6]. In treatment, AI algorithms personalize therapies based on patient data, optimizing outcomes for diseases like Parkinson's [41]. Additionally, AI has advanced neuroprosthetics by enabling adaptive brain-machine interfaces (BMIs), which, when combined with physiotherapy, show potential to improve motor function in patients with severe motor disabilities [42]. Predictive analytics in healthcare highlight the importance of transparent algorithms for reliable validation and monitoring, which support applications in neurology, such as tracking disease progression in conditions like multiple sclerosis [43]. AI's applications extend to research, facilitating the understanding of complex neurological phenomena and the development of innovative treatments [44]. For instance, StrokeSight, an EEG-based system, demonstrates how AI can improve neurology by providing fast, affordable stroke diagnoses and personalized treatments using deep learning [45]. Overall, AI's integration into neurology not only enhances clinical practices but also opens new avenues to address rare diseases and develop more finely tuned medications tailored to individual patients [46].

1.5. Radiology and cancer treatments

AI is advancing radiology and cancer treatment by enhancing imaging accuracy, enabling personalized therapies, and improving diagnostic workflows. Deep learning algorithms have shown significant progress in interpreting medical images, supporting the early detection and diagnosis of cancers such as breast and lung cancer [47]. Machine learning models assist radiologists by identifying complex patterns in imaging data, aiding in faster and more efficient diagnoses [48]. In cancer treatment, AI assists in formulating personalized treatment plans based on patient data and predictive analytics, which can predict treatment outcomes and suggest optimal therapies [49]. AI applications in radiology extend to prognostic evaluations, predicting disease progression and survival rates, refining treatment protocols and developing follow-up strategies [50]. The development of AI systems capable of outperforming human experts in tasks like breast cancer prediction highlights their potential in clinical settings. These advancements are supported by large datasets and strict validation processes, ensuring reliability and clinical applicability [51].

1.6. Emergency medicine and critical care

AI stands to significantly transform emergency medicine and critical care by enhancing patient triage, treatment efficacy, and time to diagnosis. As AI-powered tools like diagnostic

algorithms have been incredibly good at the early identification of conditions like sepsis and its progression, there is the potential to improve patient outcomes [52,53]. The prediction of patient outcomes has also positioned AI to facilitate the optimization of critical resource allocation. During the COVID-19 pandemic, AI-based self-triage tools had a critical role to play in predicting cases and hospitalizations at the population level [54]. AI-powered chatbots and virtual assistants are streamlining patient intake and symptom assessment [55,56]. Furthermore, AI's role in triaging patients based on urgency has revolutionized care prioritization, improving ED operations and patient flow [57,58]. These advancements underscore AI's potential to not only enhance emergency medical services but also pave the way for a more efficient, accurate, patient-centered healthcare system that extends care beyond traditional brick-and-mortar facilities.

2. Fairness concerns in healthcare

This section explores the fairness concerns associated with the use of AI in healthcare through three key areas. It begins by examining the sources of bias that can emerge at various stages of the machine learning pipeline. Next, it discusses the potential consequences of these biases, including inequitable healthcare outcomes, loss of trust, and ethical and legal challenges. Finally, it introduces fairness metrics and trade-offs, which provide a framework for addressing bias while balancing the inherent tensions between fairness and performance in healthcare applications. By delving into these topics, this section aims to underscore the critical need to address biases to ensure AI technologies contribute to fair and equitable healthcare for all.

2.1. Sources of bias

Bias in AI healthcare systems can arise at multiple stages of the machine learning (ML) pipeline, necessitating phase-specific mitigation strategies. To address this comprehensively, researchers have developed three primary approaches, each targeting a distinct phase of the pipeline: (i) pre-processing, which focuses on reducing bias in the input data before model training [59]; (ii) in-processing, which applies fairness constraints during the training phase [60]; and (iii) post-processing, which modifies model outputs to enhance fairness [61]. We adopt this three-stage framework to capture the end-to-end process of data collection, model development, and decision-making in AI healthcare systems. This structure, widely used in the fairness literature, supports systematic bias analysis and enables targeted detection and mitigation strategies at each stage [59–64]. The division also facilitates clearer reasoning about when and how fairness interventions can be most effectively applied throughout the ML pipeline.

A critical first step in designing effective interventions is to identify the underlying sources of bias at each stage of the ML pipeline. This section outlines those sources, with a summary provided in Table 1 and detailed discussions presented in the following sections.

2.1.1. Pre-processing stage. In the pre-processing stage, bias can originate from data collection, preparation, or representation before modeling begins. Because both the performance and fairness of an AI system depend on the quality of its data, bias at this stage can have widespread consequences. One common form of bias is *selection bias*, which arises from how the data is chosen or sampled for training, leading to unrepresentative datasets [65]. For instance, selection bias occurs when an AI model is trained predominantly on data from one demographic (e.g., white patients), while underrepresented groups, such as Black patients, have limited representation or inaccessible health records. This can lead to false negatives,

Table 1. Summary of biases across each stage of the machine learning (ML) pipeline.

Stage	Bias Type	References	Description
Pre-processing	Selection Bias	[65,66]	Arises from how the data is chosen or sampled for training, leading to unrepresentative datasets.
	Measurement Bias	[62,65,67]	Results from the ways features are chosen, utilized, or measured, leading to the systematic over- or under-representation of certain groups or variables in the data.
	Representation Bias	[14,65]	Occurs when datasets fail to adequately capture the diversity of the populations they aim to serve.
	Explicit Bias	[62]	Involves intentional or overt decisions that result in discriminatory patterns in data preparation or selection.
	Implicit Bias	[68]	Stems from unintentional or unconscious decisions that lead to skewed datasets, frequently influenced by systemic inequities.
In-processing	Algorithmic Bias	[69,70]	Arises when algorithms magnify existing biases in the training data or when they are inherently biased due to their design.
	Explicit Bias	[62]	Involves deliberate manipulations within the model's algorithms or training process that introduce or amplify disparities.
	Implicit Bias	[71]	Arises from model design decisions that unintentionally embed existing societal inequities, often due to insufficient awareness or oversight.
Post-processing	Evaluation Bias	[72]	Arises when performance assessments fail to account for disparities across demographic groups, leading to inequities in clinical outcomes.
	Explicit Bias	[62]	Involves intentional manipulation of outputs or decision thresholds that disproportionately affect specific groups.
	Implicit Bias	[68]	Occurs when decision-making criteria unintentionally create disparities, often because important fairness factors are overlooked.

https://doi.org/10.1371/journal.pdig.0000864.t001

fewer follow-up scans, and undiagnosed conditions, ultimately worsening health inequities for disadvantaged populations [66].

Another form of bias is *measurement bias*, which results from the ways features are chosen, utilized, or measured, leading to the systematic over- or under-representation of certain groups or variables in the data [62,65]. An example is the use of pulse oximeters, devices that estimate blood oxygen levels by emitting light through the skin. Research indicates that pulse oximeters often provide less accurate readings for individuals with darker skin tones, frequently overestimating their oxygen saturation levels [67]. This discrepancy can result in the underdiagnosis or delayed treatment of hypoxemia in these populations, rendering the models unreliable or biased in their predictions. Similarly, *representation bias* occurs when datasets fail to adequately capture the diversity of the populations they aim to serve [65]. For example, dermatological AI systems trained primarily on images of lighter skin tones often underperform when diagnosing conditions on darker skin [14]. This lack of representation can lead to inequitable healthcare outcomes and limit the generalizability of the AI system.

Within pre-processing, biases can also appear explicitly or implicitly. *Explicit bias* involves intentional or overt decisions that result in discriminatory patterns in data preparation or selection [62]. For instance, consider a dataset for training an AI model to predict heart disease. If the dataset is curated to include only data from male patients and excludes female patients entirely, this introduces explicit bias. This exclusion is deliberate and creates a model that cannot generalize to female patients, potentially leading to disparities in healthcare outcomes. *Implicit bias*, on the other hand, often stems from unintentional or unconscious decisions that lead to skewed datasets, frequently influenced by systemic inequities [68]. For example, using historical healthcare data that disproportionately underrepresented women in clinical trials introduces an implicit bias into the AI model. The unconscious reliance on historically biased datasets perpetuates existing inequalities, such as the underdiagnosis of heart attacks in women.

2.1.2. In-processing stage. In the in-processing stage, biases can emerge during model training or can be introduced through algorithmic choices, significantly impacting an AI system's outputs. Although the data itself can carry biases, the algorithms that learn from this data can also amplify or introduce a form of bias often referred to as *algorithmic bias*. It arises when algorithms magnify existing biases in the training data or when they are inherently biased due to their design [69]. For example, the Vaginal Birth After Cesarean (VBAC) calculator included race-based correction factors that systematically assigned lower success probabilities to African American and Hispanic women, discouraging VBAC attempts for these groups without robust scientific justification [70]. Such a design flaw not only reflected existing disparities in maternal healthcare but also exacerbated them, highlighting how biased algorithms can influence decision-making in critical areas.

Explicit and implicit biases can also arise during in-processing. *Explicit bias* involves deliberate manipulations within the model's algorithms or training process that introduce or amplify disparities [62]. For instance, intentionally adjusting the algorithm to prioritize predictions for one demographic group over another (e.g., optimizing cancer detection accuracy only for men) constitutes explicit bias. Such direct alterations can skew the model's performance metrics, favoring specific groups while disadvantaging others, resulting in ethical and clinical concerns. *Implicit bias*, however, arises from model design decisions that unintentionally embed existing societal inequities, often due to insufficient awareness or oversight [71]. For example, training a diabetes prediction model using an optimization metric that inadvertently prioritizes accuracy over fairness for underserved populations can perpetuate health disparities. These unconscious design choices may cause the model to fail to serve vulnerable groups adequately.

2.1.3. Post-processing stage. Finally, the post-processing stage involves how model outputs are evaluated, interpreted, and deployed in real clinical settings. *Evaluation bias* arises when performance assessments fail to account for disparities across demographic groups, leading to inequities in clinical outcomes [72]. For example, if an AI model for detecting diabetic retinopathy is evaluated primarily on data from urban hospitals, its performance metrics may overstate its accuracy and fail to reflect reduced performance in rural or underserved populations. This oversight in the evaluation process can lead to biased deployment decisions, disproportionately affecting access to reliable diagnostics in marginalized communities.

As in earlier stages, biases at this point can be explicit or implicit. *Explicit bias* involves intentional manipulation of outputs or decision thresholds that disproportionately affect specific groups [62]. For instance, adjusting the risk score thresholds for approving diagnostic tests based on race or socioeconomic status. Such actions, when deliberate, can reinforce discriminatory practices in clinical decision-making and resource allocation. *Implicit bias*, on the other hand, occurs when decision-making criteria unintentionally create disparities, often because important fairness factors are overlooked [68]. Consider an AI system that predicts which patients are at risk of readmission after hospital discharge. During the post-processing stage, the hospital decides to prioritize follow-up care for patients with private insurance, based on an assumption that they are more likely to comply with medical recommendations. However, this approach unintentionally introduces implicit bias, as patients with public insurance or no insurance who may face greater barriers to accessing care are deprioritized. The bias does not stem from a deliberate decision to disadvantage these groups, but rather from an unconscious oversight of fairness considerations in the post-processing criteria.

By examining biases according to their emergence in the ML pipeline, this structured framework facilitates clearer understanding of their origins and supports targeted mitigation

strategies. Such an approach underscores the importance of addressing biases early and continuously from data collection and algorithm design to final model deployment and evaluation to ensure equitable and responsible AI-driven healthcare solutions.

2.2. Potential consequences of biases

Biases in AI systems, particularly in healthcare, can have profound and far-reaching consequences. These biases usually stem from the data on which the AI systems are trained, the design of the algorithms themselves, and the contexts in which they are deployed (as discussed in Sect 2.1). Here are several potential consequences of biases in AI systems in healthcare:

Misdiagnosis and inequitable healthcare outcomes: AI integration in healthcare brings significant advancements but also risks of misdiagnosis and inequitable outcomes due to biased algorithms. AI systems often rely on non-representative data, leading to biased decision-making. For instance, Obermeyer highlighted that an algorithm used in healthcare disproportionately favored white patients over black patients because it used healthcare costs as a proxy for healthcare needs, indirectly embedding racial biases in its predictions [12]. Moreover, these biases in AI can exacerbate existing healthcare disparities. Rajkomar and others discussed how AI applications, if not carefully designed and monitored, could inherit and amplify socioeconomic and racial disparities. This is particularly problematic in diagnostics, where AI systems are trained predominantly on data from specific demographic groups, potentially leading to poorer diagnostic accuracy for underrepresented groups [4]. This was evident in a study by Adamson and Smith, which found that dermatology AI systems demonstrated lower accuracy rates in skin lesion diagnosis for dark-skinned individuals compared to those with lighter skin [14]. Such biased AI not only risks misdiagnosis but also contributes to inequitable healthcare outcomes by potentially steering healthcare resources away from those who may need them most. Vayena and his team emphasized the ethical imperative to ensure that AI tools in healthcare are developed with consideration for fairness and equity, advocating for diverse and inclusive data sets and algorithmic transparency to mitigate these biases [73].

Loss of trust in healthcare systems: Recent studies have echoed the concern that the deployment of biased AI in healthcare could significantly undermine public trust in medical systems. Trust is foundational to the patient-provider relationship and is crucial for the effective delivery of healthcare services [74]. When AI tools exhibit bias, whether in diagnosis, treatment recommendations, or patient management, they can lead to misdirected care, fostering distrust among patients, particularly in marginalized communities [4]. Kherbache and his team pointed out that when patients perceive AI-driven processes as opaque or unfair, their trust in the overall healthcare system may decline [75]. Incidents of AI failures that receive public attention can exacerbate this erosion of trust, leading patients to question the reliability and ethics of using AI in medical decision-making. Researchers like Kerasidou argue that trust is not only about the accuracy of AI but also its alignment with ethical principles that govern healthcare, such as beneficence and non-maleficence [76]. Furthermore, the lack of transparency in how AI models make decisions can be a major barrier to trust. Blease's study [77] suggests that without clear communication about how AI tools contribute to healthcare decisions, patients may become skeptical of diagnoses and treatments, fearing that their personal healthcare data could be misused or misunderstood. This potential trust deficit could have severe implications, not just for individual health outcomes but also for public health at large, as mistrust in healthcare systems can lead to lower rates of healthcare utilization, vaccine hesitancy, and poor adherence to medical advice [78]. Veinot highlights

the need for healthcare systems to maintain high standards of accountability and transparency as they integrate AI technologies to mitigate such risks [79].

Legal and ethical implications: The use of biased AI in healthcare not only poses clinical risks but also entails significant legal and ethical implications. Deploying biased AI systems could lead to legal breaches of anti-discrimination laws. In the United States, for example, the Civil Rights Act and the Americans with Disabilities Act set legal standards that could be violated by biased AI algorithms, which fail to provide equitable care across different patient demographics [80]. Ethically, biased AI conflicts with the fundamental medical ethics principles of justice and non-maleficence, demanding fairness and avoidance of harm, respectively [81]. Moreover, biased algorithms in healthcare could potentially expose medical practices and institutions to litigation related to malpractice or negligence, especially if these algorithms contribute to substandard care outcomes [82]. For instance, if an AI system were to consistently provide inferior diagnostic support for certain racial groups, this could be viewed as a form of systemic negligence or malpractice [73]. Additionally, the ethical implications extend to the breach of patient trust and the compromise of patient autonomy. The ethical medical practice relies heavily on the principles of informed consent and respect for patient's autonomy—principles that are challenged by opaque AI systems that do not make their decision-making processes or inherent biases clear to patients or practitioners [77]. These potential legal and ethical failures underscore the necessity for rigorous oversight and transparent development processes for AI in healthcare, aiming to ensure that these technologies adhere to both existing legal frameworks and ethical standards of practice.

Resource misallocation: The deployment of biased AI in healthcare settings can lead to resource misallocation, a critical issue that impacts both the efficiency and fairness of medical services. Biased algorithms may misdirect resources by prioritizing certain groups over others based on flawed data inputs or biased training procedures. For example, a study by Obermeyer demonstrated how an algorithm used for managing healthcare resources inadvertently favored healthier white patients over sicker black patients due to biased data inputs that did not accurately reflect patient needs [12]. This misallocation can exacerbate existing healthcare disparities by diverting necessary medical attention and resources away from those who are most in need. As Chen pointed out, such disparities are not just a matter of clinical outcomes but are deeply tied to social and economic inequalities that AI tools can inadvertently perpetuate [13]. Furthermore, biased AI can influence the allocation of resources within healthcare facilities, potentially resulting in inefficiencies that strain healthcare systems. This includes misallocating medical staff, diagnostic tools, and hospital beds, which can degrade the quality of care delivered while increasing wait times and healthcare costs [4]. Resource misallocation also raises ethical questions about fairness and equity in healthcare provisioning. It challenges the ethical principle of justice, which demands that healthcare resources be distributed based on need rather than biased algorithms [81]. This ethical breach can lead to further mistrust and reluctance among underserved populations to engage with healthcare systems, perpetuating a cycle of disadvantage.

Stifling innovation: The presence of bias in AI systems used in healthcare not only affects the accuracy and fairness of medical services but can also stifle innovation. When AI models are made using biased data, they might not accurately reflect the different needs of the population. This could make these innovations less useful and applicable to a wider range of demographic groups. Liang talks about how biased training datasets in AI development make it harder for these systems to work with different types of data. This could stymie innovation by preventing it from solving larger or more complex health problems that affect many people. Moreover, reliance on biased AI can deter investment in developing new technologies that are inclusive and equitable. Potential investors may be cautious about funding projects that

might not meet regulatory standards for fairness or could lead to public backlash — a concern highlighted by Corbett-Davies and Goel, who note the legal and social implications of deploying biased AI [83]. In addition, the perpetuation of bias in AI could solidify existing disparities in healthcare innovation. As Benjamin notes, if innovation is directed predominantly at solving problems specific to well-represented groups, less common but equally pressing issues in underrepresented groups are neglected, thereby limiting the scope and impact of technological advancements in healthcare [84]. The studies of Rajkomar [4], Hardt [85], and Howell [86] made things even more complicated by pointing out that biased algorithms can make it harder to find new treatments that work for everyone. This feeds a cycle of innovation that helps groups that are already overrepresented in data sets more than others. This restricted focus on innovation not only affects the equity of healthcare delivery but also limits the development of truly innovative, comprehensive healthcare solutions. These factors combined suggest that biased AI not only hinders the progression of medical technologies but also potentially locks the healthcare sector into a cycle of uneven innovation where only the needs of the majority are systematically addressed.

2.3. Fairness metrics and trade-offs

Biases in AI healthcare systems, as discussed in the previous sections, can significantly impact both the quality and equity of healthcare delivery. To address these challenges, fairness metrics provide a structured framework for evaluating and mitigating disparities in AI model predictions. Metrics such as Statistical Parity [87], Equal Opportunity [88], and Equalized Odds [89] serve as foundational tools for detecting bias. However, these metrics often reflect distinct fairness priorities, leading to inherent trade-offs when applied simultaneously [90]. For example, optimizing Statistical Parity to ensure equal outcomes across groups may conflict with Equal Opportunity, which focuses on ensuring equal access to positive outcomes for individuals who qualify. These conflicts underscore the complexity of achieving fairness in healthcare AI.

Furthermore, tensions arise not only between fairness metrics but also between different dimensions of fairness, such as group fairness and individual fairness. While group fairness seeks equitable treatment across demographic groups, individual fairness emphasizes similar outcomes for individuals with comparable features. These trade-offs become even more challenging in healthcare scenarios where fairness must often be balanced with performance, such as ensuring accurate predictions while maintaining equity. By navigating these conflicts, we aim to design AI systems that uphold fairness as a core principle in healthcare. To provide a clear understanding, we introduce these three key fairness metrics along with their mathematical foundations.

Statistical parity (SP): Statistical Parity [87] ensures that the probability of a positive predicted outcome is equal across different groups defined by a sensitive attribute. Mathematically, it is defined as:

$$SP = P(\hat{Y} = 1 \mid S = s_1) - P(\hat{Y} = 1 \mid S = s_2) \tag{1}$$

where $\hat{Y}$ is the predicted outcome and S is the sensitive attribute. Statistical Parity is satisfied when $SP = 0$, indicating no disparity between the groups.

Equal opportunity (EO): Equal Opportunity [88] ensures that individuals who are likely to achieve the advantaged outcome (*e.g.*, repaying a loan, being admitted to college) have an equal chance of receiving the positive prediction, regardless of their sensitive attribute (*e.g.*,

race, gender). Mathematically, it is defined as:

$$EO = P(\hat{Y} = 1 \mid S = s_1, Y = 1) - P(\hat{Y} = 1 \mid S = s_2, Y = 1) \tag{2}$$

where $\hat{Y}$ is the binary prediction, S is the sensitive attribute, and $Y = 1$ represents the advantaged outcome. Equal Opportunity is satisfied when $EO = 0$ indicates no disparity in predictions for the advantaged outcome across sensitive groups.

Equalized odds (Eq.Odds): Equalized Odds [89] ensures that for each ground truth label (Y = y), the predicted outcome $\hat{Y}$ is equally likely across groups defined by the sensitive attribute (S). Mathematically:

$$\text{Eq.Odds} = \sum_{y \in \{0,1\}} \left| P(\hat{Y} = 1 \mid S = s_1, Y = y) - P(\hat{Y} = 1 \mid S = s_2, Y = y) \right| \tag{3}$$

where $\hat{Y}$ is the predicted outcome, S is the sensitive attribute, and Y is the ground truth label. Equalized Odds is satisfied when $Eq.Odds = 0$, indicating no disparity across groups for all ground truth labels.

These group fairness metrics Statistical Parity, Equal Opportunity, and Equalized Odds reflect fundamentally different fairness priorities, leading to inherent trade-offs when attempting to satisfy them simultaneously. Statistical Parity requires predictions to be independent of sensitive attributes, ensuring equal positive outcome rates across groups, regardless of differences in their underlying characteristics, which can sometimes lead to disparities in how individuals within a group are treated. Consider a healthcare scenario where males have a higher prevalence of heart disease compared to females, enforcing Statistical Parity would mean predicting the same proportion of males and females as being at risk. However, this might conflict with Equal Opportunity, which requires equal true positive rates across groups. Achieving Statistical Parity in this scenario may lower the true positive rate for females, as increasing their positive predictions could include many individuals without heart disease. This highlights how prioritizing Statistical Parity can undermine Equal Opportunity by creating disparities in prediction accuracy within groups.

To achieve Equalized Odds, a model must ensure that true positive and false positive rates are the same across groups, which often requires adjusting prediction thresholds differently for each group. For example, if females are less likely to have heart disease compared to males, the model might lower the prediction threshold for females to classify more of them as "at risk" increasing the true positive rate for females. At the same time, it might raise the threshold for males to reduce false positives. While these adjustments equalize predictive performance across groups, they disrupt Statistical Parity by creating unequal proportions of positive predictions for males and females, reflecting the differences in the underlying likelihood of heart disease between the groups. This trade-off shows how achieving Equalized Odds can violate Statistical Parity by altering the balance of positive predictions across groups.

The trade-offs between these fairness metrics are formalized in the Impossibility Theorem [90], which demonstrates that it is often impossible to satisfy all fairness criteria simultaneously. For instance, achieving Statistical Parity might require selecting more from a group with a lower base rate, sacrificing Equalized Odds or Equal Opportunity, while prioritizing Equal Opportunity or Equalized Odds might allow selection rates to differ across groups, violating Statistical Parity. These inherent conflicts require careful consideration when deciding which fairness metric to apply. Ultimately, the choice of fairness metric should depend on the specific context and the ethical or practical priorities of the application. For instance, in critical healthcare scenarios like diagnosing heart disease, Equal Opportunity or Equalized

Odds might be prioritized to ensure equal access to accurate treatment and minimize disparities in predictive performance. On the other hand, Statistical Parity could be more relevant in preventive care programs, where equitable outreach and representation are paramount goals [91].

The conflict between fairness metrics can also be seen in the difference between individual and group fairness, which focuses on fairness at different levels. Group fairness [92] aims to ensure equitable treatment across different subgroups defined by sensitive attributes, *i.e.*, group fairness aims to ensure that the outputs of ML algorithms are independent of the sensitive attribute. In contrast, individual fairness [93] enforces fairness at a finer granularity by ensuring that similar individuals receive similar predictions based on their input features. Unfortunately, the simultaneous achievement of both maxima is practically impossible due to the trade-offs involved. Fig 1 illustrates this trade-off in the context of organ allocation, where gender is the sensitive attribute, and medical urgency determines priority. In the performance-driven result (left side of the Fig 1), organs are allocated solely based on urgency and likelihood of survival, prioritizing patients P_1, P_2, P_3, and P_4 due to their higher need and better survival prospects, while rejecting P_5, P_6, P_7, and P_8 because of their lower urgency or expected benefit. This approach maximizes utility by allocating resources to those expected to achieve the greatest overall benefit, such as improved survival rates and long-term health outcomes. However, it may inadvertently perpetuate systemic disparities between subgroups. The group fairness-driven result (right side of the Fig 1) adjusts the decision boundary to ensure equal approval rates for males and females, resulting in females P_7 and P_8 being approved despite their lower urgency, while males P_3 and P_4, with higher urgency, are denied. While this adjustment addresses equity at the group level, it compromises individual fairness, as highlighted by the purple rectangle, where individuals near the decision boundary disproportionately bear the group fair loss of the entire subgroup.

Building on the discussion of trade-offs between individual and group fairness, the tension between fairness and performance further illustrates these challenges. In contexts such as organ allocation, performance-driven approaches prioritize outcomes for individuals with the highest urgency or expected benefit. However, enforcing group fairness constraints such as ensuring equal approval rates across subgroups defined by sensitive attributes like gender can lead to a reduction in overall performance. For example, in the organ allocation scenario, ensuring that approval rates are equal for males and females might result in allocating organs to individuals like P_7 and P_8 (females, lower urgency) while denying organs to P_3 and

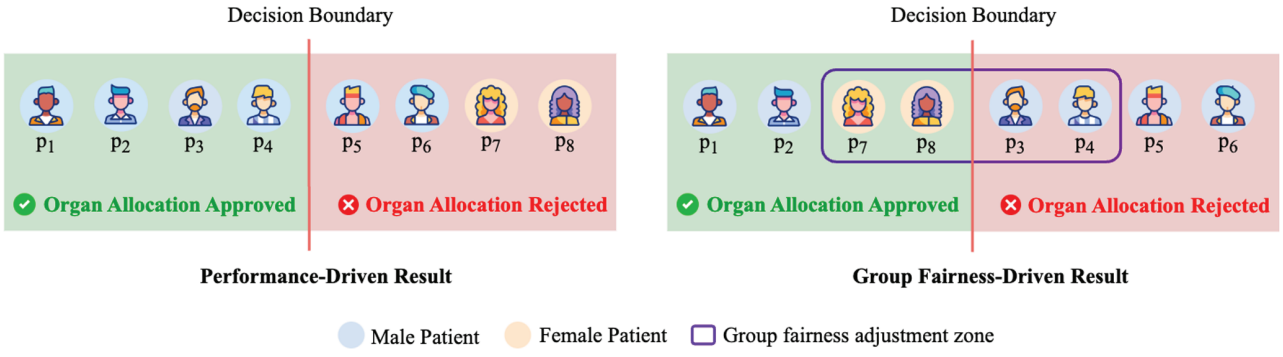

Fig 1. Illustration of the trade-off between performance-driven result and group fairness-driven result in the context of organ allocation, where gender is the sensitive attribute.

https://doi.org/10.1371/journal.pdig.0000864.g001

P_4 (males, higher urgency). This adjustment ensures equity at the group level but comes at the cost of reduced overall utility, as fewer lives may be saved or long-term health outcomes may be compromised. This trade-off highlights the inherent difficulty of simultaneously achieving fairness and optimal performance, as prioritizing one objective often necessitates sacrifices in the other.

3. Addressing and mitigating unfairness in AI

As AI technologies increasingly influence healthcare delivery, it becomes essential to scrutinize and refine these systems to prevent disparities in care and outcomes. This section explores various bias-detecting and mitigating strategies in AI systems in healthcare.

3.1. Bias detection methods

Bias detection is a crucial step in identifying and addressing inequities in AI systems, especially in healthcare given its high-stakes nature. This section discusses various tools and techniques for detecting bias, categorized by their roles in the pre-processing, in-processing, and post-processing stages of the machine learning (ML) pipeline. The aim is to provide a clear and actionable understanding of bias detection.

3.1.1. Pre-processing stage. The pre-processing stage focuses on detecting biases in datasets before model training begins. Identifying these biases at this stage is essential to prevent their propagation and amplification through the ML pipeline if left unchecked. Specifically, bias can be detected using *Fairness Metrics Techniques*, such as Equalized Odds and Equal Opportunity Difference, which measure disparities in clinical outcomes or feature distributions across demographic groups. These metrics help identify biases in healthcare datasets, such as selection bias and measurement bias by revealing whether certain patient populaitons are disadvantaged or overrepresented. For instance, if one group consistently experiences worse outcomes in the data, these tools can pinpoint the issue, enabling targeted corrections [88].

Another detection method is *Dataset Visualization Techniques*, such as t-SNE (t-Distributed Stochastic Neighbor Embedding) and PCA (Principal Component Analysis), which allow researchers to examine bias in high-dimensional healthcare data in more intuitive ways. These tools uncover patterns, such as clusters or representation gaps, that may signal clinical representation bias. For example, a t-SNE plot could reveal that certain skin tones are underrepresented in a dermatological dataset, prompting actions to balance the dataset [94]. Another approach is the use of *Causal Graphs*, which focus on identifying and understanding the relationships between variables in healthcare datasets. Causal Graphs are graphical representations of causal relationships among variables, allowing researchers to distinguish true causation from correlations and uncover confounding variables that may introduce bias. For instance, [95] proposed a framework leveraging Causal Variational Autoencoders (CVAEs) to indirectly reconstruct sensitive information, even when such attributes are unavailable due to privacy constraints. By identifying clinical biases at the data level, this method proves particularly valuable in addressing challenges within complex, real-world medical datasets.

3.1.2. In-processing stage. The in-processing stage focuses on detecting biases introduced or amplified during model training. This stage prioritizes fairness by incorporating bias identification strategies directly into the training process. One effective method is *Adversarial Learning*, which employs adversarial techniques to identify bias in models during training. A primary model predicts clinical outcomes (e.g., disease risk), while an adversary attempts to infer sensitive attributes such as patient demographics (e.g., race, gender) based on the model's predictions. The adversary's success in predicting these attributes reveals the extent

of algorithmic bias embedded in the model. By analyzing this feedback, healthcare developers can identify where and how protected attributes influence clinical predictions, enabling precise bias detection and promoting equitable care [96].

Another approach involves *Explanation Methods*, such as LIME (Local Interpretable Model-Agnostic Explanations), which provide insights into how clinical features influence predictions in healthcare models. These methods can uncover both explicit and implicit biases in the model, such as disproportionately relying on sensitive attributes like race, gender, or socioeconomic status or unintentionally reinforcing biased patterns that disadvantage certain patient groups. By iteratively applying these explanations during training, healthcare practitioners can identify and correct biases, improving fairness and transparency in patient care [97]. A further technique for bias detection is *Causal Modeling*, which identifies algorithmic bias by analyzing how sensitive attributes, such as gender, ethnicity, or socioeconomic status, influence predictive outcomes in healthcare applications. Causal Modeling involves constructing mathematical or computational models that map out causal relationships between variables, enabling researchers to trace pathways where biases emerge during the training of models used in medical decision-making. For instance, researchers utilized causal modeling with the Adult dataset from the UC Irvine Machine Learning Repository to detect gender bias in binary classification outcomes [98]. By treating the prediction model as a black box and constructing a corresponding causal model, they identified how gender impacted the model's decisions. This process uncovers both explicit and implicit biases in the model's decision-making framework, providing actionable insights for detecting inequities. By identifying biases at the model level, causal modeling promotes the development of fair and robust AI systems in health care.

3.1.3. Post-processing stage. The post-processing stage evaluates and mitigates any biases in the model's outputs after training is complete. This stage ensures equitable treatment in the model's final decisions. One effective approach for bias detection is *Counterfactual Analysis*, which assesses whether a model's decisions remain consistent even if sensitive attributes, such as race or gender, are changed. By using causal inference, this method determines whether changing a sensitive attribute would alter the model's outcome. If the outcome remains the same, the decision is considered fair, making this method effective for identifying and correcting both explicit and implicit biases in the model's outputs [99]. For instance, in a healthcare model that predicts eligibility for advanced treatments, if changing a sensitive attribute like race results in different treatment outcomes, it highlights bias in the model's decision-making process.

Another tool is *Model Cards*, which are standardized documents that summarize a model's performance across various clinical conditions and demographic groups. They include details about the model's intended use in healthcare settings, limitations, and evaluation results, including the detection and analysis of evaluation bias. By promoting transparency, Model Cards help healthcare stakeholders make informed decisions about deployment and encourage responsible use [100]. Another detection approach, *Reject Option* involves identifying cases near the decision boundary where classification decisions are uncertain and potentially influenced by bias. This approach helps detect patterns where sensitive demographic groups might be disproportionately affected by misclassifications. In healthcare settings, where errors in diagnosis or treatment can have significant consequences, the reject option highlights instances that require further inspection to ensure equitable outcomes. By focusing on borderline cases, this method provides valuable insights into potential biases in model predictions, aiding in the identification of inequities within healthcare systems [101].

This framework provides a clear and actionable approach to tackling biases in AI systems. Detecting and identifying biases at each stage of the ML pipeline is essential for creating fair, reliable, and equitable AI-driven healthcare solutions.

3.2. Mitigation strategies

Mitigating biases in AI systems within healthcare is crucial for ensuring equitable treatment across all patient demographics. This section outlines key strategies to address and reduce biases, including the use of diverse and representative data and fairness-aware approaches, thereby enhancing the fairness and accuracy of AI models.

3.2.1. Diverse and representative data. Mitigating biases in AI systems within healthcare is crucial for ensuring equitable treatment across all patient demographics. A fundamental strategy involves using diverse and representative datasets in the training phase of these systems. Research indicates that AI models can only be as unbiased as the data they are trained on [102]. Thus, the inclusion of comprehensive data from a wide array of patient groups, particularly those historically underrepresented in medical research, is essential [13].

The significance of representative data extends beyond just including diverse groups; it also involves the depth and quality of the data collected from these groups. Data must capture a breadth of variables that influence health outcomes, such as socioeconomic factors, environmental conditions, and genetic differences [4]. This approach helps in creating models that are better tuned to the nuances of various patient needs and conditions [14].

Additionally, involving stakeholders from diverse backgrounds in the data collection and AI development process can enhance the relevancy and sensitivity of the datasets [70]. This inclusion helps in identifying and addressing potential blind spots in AI training datasets, which, if overlooked, could perpetuate biases and inequalities in healthcare delivery [103]. Employing these strategies not only improves the fairness and effectiveness of AI tools but also builds trust in these technologies among all user groups, thereby fostering a more inclusive healthcare ecosystem [81].

3.2.2. Fairness-aware approaches. Incorporating various fairness-aware approaches is crucial for developing AI systems that are equitable and unbiased. These approaches can be categorized into three main strategies: pre-processing, in-processing, and post-processing. Each strategy targets different stages of the machine learning pipeline to address potential biases and ensure fair treatment across diverse demographic groups.

Pre-processing. In healthcare, pre-processing involves techniques to adjust the input data for AI models to ensure it accurately represents diverse patient demographics before model training begins. This process aims to eliminate any inherent biases that may skew AI predictions and outcomes, thus fostering equity in healthcare treatments and diagnostics [104]. By refining the dataset upfront, pre-processing helps in building AI systems that perform fairly across all patient groups [13].

- *Re-Sampling:* Re-sampling is a common pre-processing method that manipulates data samples to create a more balanced dataset. This can be done through either over-sampling minority classes or under-sampling majority classes [105]. An example of this in healthcare is the use of the Synthetic Minority Oversampling Technique (SMOTE) [106] in medical datasets where certain conditions or outcomes are rare. Chawla applied SMOTE to a dataset of imbalanced classes to improve the prediction of minority-class outcomes in medical diagnostics, resulting in more reliable predictive models for underrepresented conditions. This technique helps to ensure that the AI does not become biased towards the more frequent classes.

- *Re-Weighting:* To reduce bias, instances in the training data are reweighted. We assign a weight to each instance in a dataset based on its representation, thereby increasing the influence of underrepresented groups in the model training process. This technique was demonstrated by Calmon in [107], who developed a data pre-processing method that modifies features and outcomes to improve fairness. In healthcare, this could mean adjusting weights so that data from minority ethnic groups has a greater influence on the model, helping to prevent the AI from developing biases that might affect diagnosis or treatment recommendations [104].

In-processing. In-processing methods involve integrating fairness directly into the learning algorithm itself. This technique adjusts the algorithm during the training phase to minimize bias. Zemel in his study introduced a method where they developed a representation for data that is invariant to protected attributes like race or gender while maintaining the informative characteristics necessary for prediction [108]. In healthcare, this can be crucial for ensuring that diagnostic or treatment recommendations are not adversely skewed by underlying biases in the training data [109].

- *Adversarial Debiasing:* Adversarial debiasing involves the simultaneous training of a predictor and an adversary. The predictor learns to make accurate predictions, and the adversary learns to determine whether the predictions are biased toward certain groups. Deldjoo et al. [109] explored this in the context of healthcare, aiming to create clinical decision support systems that are unbiased towards patients' demographic attributes. This technique helps ensure that the AI's treatment recommendations do not reflect discriminatory patterns that might exist in the training data [110].
- *Constraint-Based Optimization:* Constraint-based optimization involves modifying the learning algorithm to satisfy fairness constraints while minimizing prediction error. You can implement this method by adding fairness constraints to the objective function that the algorithm is optimizing. Elzayn et al. [111] applied this method to healthcare datasets, ensuring that treatment recommendations do not disproportionately favor one group over another by enforcing demographic parity or equality of opportunity as constraints during model training [112].

Post-processing. After training a model, post-processing techniques are applied to modify the outputs of AI systems, ensuring fair treatment across different demographic groups. These adjustments address any residual biases to promote equitable outcomes. For example, Hardt et al. [88] applied different decision thresholds for various groups to equalize treatment effects. This method effectively adjusted risk scores in a healthcare dataset, ensuring that treatment recommendations were equitable [113].

- *Threshold Adjustment:* Threshold adjustment in AI healthcare involves setting an optimal cutoff point for predictive models to classify outcomes effectively, such as predicting disease presence or patient risk levels. A significant example of this is in cardiovascular disease prediction. In a study by Siontis [114], the researchers examined the application of different thresholds for predicting cardiovascular events using a logistic regression model. They aimed to optimize the sensitivity and specificity of predictions by adjusting the threshold to better serve clinical decision-making. This adjustment proved crucial in identifying higher-risk patients who might benefit from preventive treatments. Another notable case is in breast cancer screening, where Fenton [115] applied threshold adjustment to mammography interpretation models. By altering the decision threshold, they were able to reduce false

positives without substantially missing cases of cancer, thereby balancing the need for early detection with the risk of overdiagnosis and unnecessary anxiety for patients.

- *Output Recalibration:* Output recalibration in AI healthcare is critical for adapting predictive models to local contexts and ensuring that the predicted probabilities match actual clinical outcomes in a new patient population. A notable case involves the recalibration of a sepsis prediction model at Johns Hopkins Hospital, as detailed by Henry et al. [116]. The model was originally developed with a dataset from one patient demographic, but recalibration using local demographic data was required to maintain accuracy across diverse populations within the hospital. This recalibration helped align the model's predictions with the actual rates of sepsis observed, improving clinical decision-making and patient care. Another example is from the University of California, San Francisco (UCSF), where Rajkomar [4] recalibrated Google's deep learning model used for predicting medical outcomes. The model, initially trained on a national dataset, was recalibrated with UCSF-specific patient data to ensure its predictions accurately reflected the local patient population's risk factors and outcomes. This was particularly important for diseases with variable manifestations across different demographics.

4. Research gaps and future directions

This section delves into the research gaps in implementing fair AI in healthcare and future directions to enhance its efficacy. Ensuring fairness in AI systems within healthcare is a complex and multifaceted challenge that requires addressing various deficiencies and promoting interdisciplinary collaboration.

4.1. Ethical considerations

The integration of AI in healthcare raises significant ethical challenges, including privacy concerns, algorithmic bias and accountability. One critical gap lies in ensuring patient data privacy while enabling AI models to be trained on diverse datasets. Ensuring the confidentiality and security of this data against breaches is paramount [117]. Current approaches, such as anonymization, may not adequately prevent re-identification risks, particularly when combined with external data sources [118]. Future research should explore privacy-preserving techniques, such as federated learning and differential privacy, to enable secure training of AI systems without compromising data confidentiality [119]. Algorithmic bias is another major concern. AI models trained on non-representative data can perpetuate systemic inequalities, leading to unfair treatment outcomes for underrepresented groups [4]. Solutions like adversarial debiasing and fairness-aware algorithms should be further developed and rigorously tested in healthcare settings [120]. Furthermore, the deployment of AI in healthcare must take into account the potential displacement of healthcare professionals, raising concerns about job security and the loss of valuable human expertise and empathy in patient care [121]. Lastly, accountability for AI-driven decisions in healthcare is a critical ethical issue. It is critical to define who is responsible—the developers, the users, or the AI itself—when an AI system's decision results in patient harm [122].

4.2. Legal considerations

Legal considerations in AI healthcare systems encompass intellectual property rights, liability for errors, regulatory compliance, and policy gaps that hinder bias mitigation. A notable gap exists in the legal frameworks governing intellectual property rights as there is an ongoing debate over whether AI-generated medical inventions can be patented and who owns

the rights to AI-generated data and algorithms [82]. Future research should focus on developing frameworks that equitably address the rights and interests of developers, institutions, and patients. Furthermore, liability issues arise when AI systems make erroneous decisions that could harm patients. Determining whether the healthcare provider, the software developer, or another party is liable requires intricate legal analysis and possibly new legal frameworks [123]. Solutions such as liability insurance for AI developers and a staggered approach to liability, from strict liability for high-risk applications to simple fault-based liability for consumers, could address this gap [124]. Policy gaps further hinder bias mitigation. Future directions also include promoting policy changes that enforce transparency and fairness in data collection and algorithmic processing in healthcare applications [81]. Governments and institutions should implement policies mandating fairness audits and bias detection tools to ensure equitable outcomes. AI applications in healthcare must adhere to existing healthcare regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) in the U.S., which governs the privacy and security of patient data, and the General Data Protection Regulation (GDPR) in the EU, which sets stringent guidelines for data protection and privacy [125]. As AI technologies evolve, there may also be a need for specific regulations that address AI's unique aspects in healthcare to ensure that these innovations are safely and effectively integrated into medical practice [126]. Lastly, issues of informed consent are magnified with AI, as patients must understand the role of AI in their care, including how their data will be used and the implications of AI-driven decisions [122].

4.3. Transparency and interpretability

The lack of transparency and interpretability in AI models remains a major obstacle to fairness in healthcare. Many advanced models, such as deep neural networks, function as "black boxes," making their decision-making processes hard to understand and verify. This lack of clarity can lead to mistrust among healthcare providers and patients, as well as unidentified biases in outcomes [127]. For instance, an AI diagnostic tool might predict a higher risk of disease for one patient over another with nearly identical medical histories, without providing any rationale for the differing assessments. To address this, future research should focus on developing "glass-box" approaches that are inherently interpretable, such as decision trees or linear models, fostering greater trust among healthcare providers and patients [127]. Future research should focus on adopting model-agnostic explanation frameworks like LIME [97] or SHAP [128], which provide insights into the decision-making processes of any AI model, regardless of its underlying architecture. Developing these tools will enable clinicians to better understand and oversee AI-driven decisions, potentially leading to more equitable healthcare outcomes. Additionally, implementing model documentation standards like Model Cards can provide essential transparency by detailing the capabilities and limitations of AI models to all stakeholders [100].

4.4. Diversity and representation

The lack of diversity in training datasets is a significant research gap that impacts AI fairness in healthcare. Models trained on data that underrepresents certain demographic groups risk perpetuating biases, thereby worsening healthcare inequalities [4]. For example, a skin cancer detection AI trained predominantly on lighter skin tones may fail to accurately identify conditions on darker skin, highlighting the impact of underrepresentation in training data on healthcare outcomes. To mitigate this issue, techniques for synthetically augmenting underrepresented classes in datasets or incentivizing the collection of more comprehensive

and inclusive data must be developed [129]. Additionally, implementing algorithmic fairness approaches during model training, such as fairness constraints or adversarial debiasing, can help mitigate biases introduced by skewed datasets [62]. Studies suggest that geospatial AI models, when applied to geographic information systems (GIS), can uncover regions with inadequate healthcare services, thus guiding interventions to those most in need [79]. However, predictive analytics must be carefully designed to ensure they do not perpetuate biases present in underlying data [12]. Integrating feedback mechanisms from healthcare providers into AI systems offers a pathway to refine these models, ensuring they adapt to serve diverse populations effectively. The future direction in bridging this gap involves developing AI technologies that are not only predictive but also prescriptive, providing actionable insights to improve universal healthcare accessibility [4].

4.5. Interdisciplinary approaches

The integration of AI in healthcare requires interdisciplinary approaches that combine insights from medical ethics, data science, social sciences, and clinical practice to comprehensively address fairness [130]. However, current AI applications often lack this multidimensional perspective, resulting in solutions that are technically adequate but socially or ethically inappropriate [81]. For example, an AI diagnostic tool developed with only technical expertise might fail to account for socio-economic factors, leading to misdiagnoses in underserved communities that lack access to quality healthcare infrastructure. Future research should focus on frameworks that incorporate not only technical accuracy but also ethical, legal, and social implications from the outset, ensuring that AI systems are developed with a holistic view of fairness [79]. These collaborative frameworks can help prevent the exacerbation of existing disparities and promote a more equitable distribution of healthcare resources [13].

4.6. Longitudinal effects

The longitudinal effects of AI systems on healthcare fairness remain underexplored, with existing studies often focusing on short-term outcomes and immediate biases in algorithmic decision-making [131]. Understanding how these technologies affect health disparities over time is crucial, as initial biases can amplify if not properly addressed [4]. For example, an AI system predicting patient readmissions might keep giving more resources to well-funded hospitals because they had better outcomes in the past, making the gap worse for less-supported hospitals over time. Future research should involve continuous monitoring of AI systems post-deployment to assess their impact on various demographic groups across different time intervals [132]. This approach will help identify evolving biases and enable timely modifications to algorithms, thereby ensuring more equitable healthcare outcomes [12]. Additionally, incorporating feedback loops that allow healthcare providers to report disparities can further refine AI applications in real-world settings.

5. Conclusion

In conclusion, AI in healthcare holds transformative potential for improving diagnosis, treatment, and patient management. However, its integration also brings significant ethical, legal, and operational challenges, especially regarding fairness and equity. To balance the benefits of AI with these challenges, it is essential to address biases, ensure ethical practices, and foster interdisciplinary collaboration. Addressing biases involves implementing more inclusive data collection practices to ensure diverse representation and developing bias-aware AI

models to mitigate disparities in healthcare delivery. Ensuring ethical practices requires establishing robust regulatory frameworks to oversee AI implementation, guaranteeing compliance with ethical standards and legal requirements. Additionally, promoting transparency in AI decision-making processes is essential to building trust among healthcare providers and patients. To achieve these goals, fostering interdisciplinary collaboration is vital. Encouraging ongoing cooperation between technologists, clinicians, legal experts, and ethicists will help address the multifaceted challenges of AI in healthcare. Interdisciplinary research should focus on developing innovative and inclusive AI applications that effectively serve diverse populations without perpetuating existing disparities. Future research should aim to refine AI tools to ensure their effectiveness across diverse populations, continuously monitoring and adjusting to evolving biases. Investigating the long-term impacts of AI systems on healthcare equity is crucial for making necessary adjustments over time. By addressing these critical areas, we can harness the capabilities of AI to create a more equitable and effective healthcare system. As we advance, it is imperative to remain vigilant about the ethical implications of AI, ensuring that these technologies benefit all segments of society equally. This requires a committed effort to integrate fairness, transparency, and inclusivity into every stage of AI development and deployment in healthcare.

Author contributions

Conceptualization: Sribala Vidyadhari Chinta, Zichong Wang, Wenbin Zhang.

Data curation: Sribala Vidyadhari Chinta, Zichong Wang, Avash Palikhe, Xingyu Zhang, Ayesha Kashif.

Formal analysis: Sribala Vidyadhari Chinta, Zichong Wang, Avash Palikhe, Xingyu Zhang, Ayesha Kashif, Monique Antoinette Smith, Jun Liu, Wenbin Zhang.

Funding acquisition: Wenbin Zhang.

Investigation: Sribala Vidyadhari Chinta, Zichong Wang, Avash Palikhe, Xingyu Zhang, Ayesha Kashif, Monique Antoinette Smith, Jun Liu, Wenbin Zhang.

Methodology: Sribala Vidyadhari Chinta, Zichong Wang, Avash Palikhe, Jun Liu, Wenbin Zhang.

Project administration: Jun Liu, Wenbin Zhang.

Supervision: Wenbin Zhang.

Visualization: Avash Palikhe, Xingyu Zhang.

Writing – original draft: Sribala Vidyadhari Chinta, Zichong Wang, Avash Palikhe, Xingyu Zhang, Ayesha Kashif, Jun Liu, Wenbin Zhang.

Writing – review & editing: Sribala Vidyadhari Chinta, Zichong Wang, Avash Palikhe, Xingyu Zhang, Ayesha Kashif, Monique Antoinette Smith, Jun Liu, Wenbin Zhang.

References

1. Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. Comput Biomed Res. 1975;8(4):303–20. https://doi.org/10.1016/0010-4809(75)90009-9 PMID: 1157471
2. Miller RA. INTERNIST-1/CADUCEUS: problems facing expert consultant programs. Methods Inf Med. 1984;23(1):9–14. https://doi.org/10.1055/s-0038-1635320 PMID: 6374372

3. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24–9. https://doi.org/10.1038/s41591-018-0316-z PMID: 30617335
4. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347–58. https://doi.org/10.1056/NEJMra1814259 PMID: 30943338
5. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2(4):230–43. https://doi.org/10.1136/svn-2017-000101 PMID: 29507784
6. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005 PMID: 28778026
7. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2014;13:8–17. https://doi.org/10.1016/j.csbj.2014.11.005 PMID: 25750696
8. Strickland E. IBM Watson, heal thyself: how IBM overpromised and underdelivered on AI health care. IEEE Spectr. 2019;56(4):24–31. https://doi.org/10.1109/mspec.2019.8678513
9. Jha S, Topol EJ. Adapting to artificial intelligence: radiologists and pathologists as information specialists. JAMA. 2016;316(22):2353–4. https://doi.org/10.1001/jama.2016.17438 PMID: 27898975
10. Feldman B, Martin E, Skotnes T. Big data in healthcare: hype and hope. Dr. Bonnie 360. 2012.
11. Bachtiger P, Peters NS, Walsh SL. Machine learning for COVID-19-asking the right questions. Lancet Digit Health. 2020;2(8):e391–2. https://doi.org/10.1016/S2589-7500(20)30162-X PMID: 32835197
12. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–53. https://doi.org/10.1126/science.aax2342 PMID: 31649194
13. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care?. AMA J Ethics. 2019;21(2):E167-179. https://doi.org/10.1001/amajethics.2019.167 PMID: 30794127
14. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. JAMA Dermatol. 2018;154(11):1247–8. https://doi.org/10.1001/jamadermatol.2018.2348 PMID: 30073260
15. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. JAMA. 2019;322(24):2377–8. https://doi.org/10.1001/jama.2019.18058 PMID: 31755905
16. Voigt P, von dem Bussche A. The EU General Data Protection Regulation (GDPR). Springer. 2017. https://doi.org/10.1007/978-3-319-57959-7
17. US Food and Drug Administration. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). 2019. https://www.fda.gov/media/122535/download
18. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial intelligence in cardiology. J Am Coll Cardiol. 2018;71(23):2668–79. https://doi.org/10.1016/j.jacc.2018.03.521 PMID: 29880128
19. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. J Am Coll Cardiol. 2017;69(21):2657–64.
20. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nat Med. 2019;25(1):70–4. https://doi.org/10.1038/s41591-018-0240-2 PMID: 30617318
21. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med. 2019;25(1):65–9. https://doi.org/10.1038/s41591-018-0268-3 PMID: 30617320
22. Madani A, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of echocardiograms using deep learning. NPJ Digit Med. 2018;1:6. https://doi.org/10.1038/s41746-017-0013-1 PMID: 30828647
23. Ambale-Venkatesh B, Lima JAC. Cardiac MRI: a central prognostic tool in myocardial fibrosis. Nat Rev Cardiol. 2015;12(1):18–29. https://doi.org/10.1038/nrcardio.2014.159 PMID: 25348690
24. Krittanawong C, Johnson KW, Rosenson RS, Wang Z, Aydar M, Baber U, et al. Deep learning for cardiovascular medicine: a practical primer. Eur Heart J. 2019;40(25):2058–73. https://doi.org/10.1093/eurheartj/ehz056 PMID: 30815669

25. Courtiol P, Maussion C, Moarii M, Pronier E, Pilcer S, Sefta M, et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nat Med. 2019;25(10):1519–25. https://doi.org/10.1038/s41591-019-0583-3 PMID: 31591589

26. Farhud DD, Zokaei S. Ethical issues of artificial intelligence in medicine and healthcare. Iran J Public Health. 2021;50(11):i–v. https://doi.org/10.18502/ijph.v50i11.7600 PMID: 35223619

27. Fuster V, Kelly BB, Vedanthan R. Global cardiovascular health: urgent need for an intersectoral approach. J Am Coll Cardiol. 2011;58(12):1208–10. https://doi.org/10.1016/j.jacc.2011.05.038 PMID: 21903051

28. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7 PMID: 30617339

29. Ting DSW, Cheung CY-L, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA. 2017;318(22):2211–23. https://doi.org/10.1001/jama.2017.18152 PMID: 29234807

30. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402–10. https://doi.org/10.1001/jama.2016.17216 PMID: 27898976

31. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med. 2018;24(9):1342–50. https://doi.org/10.1038/s41591-018-0107-6 PMID: 30104768

32. Balyen L, Peto T. Promising artificial intelligence-machine learning-deep learning algorithms in ophthalmology. Asia Pac J Ophthalmol (Phila). 2019;8(3):264–72. https://doi.org/10.22608/APO.2018479 PMID: 31149787

33. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172(5):1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010 PMID: 29474911

34. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med. 2018;1:39. https://doi.org/10.1038/s41746-018-0040-6 PMID: 31304320

35. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–8. https://doi.org/10.1038/nature21056 PMID: 28117445

36. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J Invest Dermatol. 2018;138(7):1529–38. https://doi.org/10.1016/j.jid.2018.01.028 PMID: 29428356

37. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human-computer collaboration for skin cancer recognition. Nat Med. 2020;26(8):1229–34. https://doi.org/10.1038/s41591-020-0942-0 PMID: 32572267

38. Gomolin A, Netchiporouk E, Gniadecki R, Litvinov IV. Artificial intelligence applications in dermatology: where do we stand?. Front Med (Lausanne). 2020;7:100. https://doi.org/10.3389/fmed.2020.00100 PMID: 32296706

39. Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, et al. A deep learning system for differential diagnosis of skin diseases. Nat Med. 2020;26(6):900–8. https://doi.org/10.1038/s41591-020-0842-3 PMID: 32424212

40. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44. https://doi.org/10.1038/nature14539 PMID: 26017442

41. Dorsey ER, Bloem BR. The Parkinson pandemic-A call to action. JAMA Neurol. 2018;75(1):9–10. https://doi.org/10.1001/jamaneurol.2017.3299 PMID: 29131880

42. Ramos-Murguialday A, Broetz D, Rea M, Läer L, Yilmaz O, Brasil FL, et al. Brain-machine interface in chronic stroke rehabilitation: a controlled study. Ann Neurol. 2013;74(1):100–8. https://doi.org/10.1002/ana.23879 PMID: 23494615

43. Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works?. J Am Med Inform Assoc. 2019;26(12):1651–4. https://doi.org/10.1093/jamia/ocz130 PMID: 31373357

44. Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. Biol Psychiat Cogn Neurosci Neuroimag. 2018;3(3):223–30. https://doi.org/10.1016/j.bpsc.2017.11.007 PMID: 29486863

45. Kalahasty R, Motati L. Strokesight: a novel eeg-based diagnostic system for strokes using spectral analysis and deep learning. arXiv preprint 2022. https://arxiv.org/abs/2203.14296

46. Fornito A, Zalesky A, Bullmore E. Fundamentals of brain network analysis. London: Academic Press; 2016. https://doi.org/10.1016/C2012-0-06036-X

47. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer. 2018;18(8):500–10. https://doi.org/10.1038/s41568-018-0016-5 PMID: 29777175

48. Wang S, Summers RM. Machine learning and radiology. Med Image Anal. 2012;16(5):933–51. https://doi.org/10.1016/j.media.2012.02.005 PMID: 22465077

49. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018;2(10):719–31. https://doi.org/10.1038/s41551-018-0305-z PMID: 31015651

50. Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. CA Cancer J Clin. 2019;69(2):127–57. https://doi.org/10.3322/caac.21552 PMID: 30720861

51. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89–94. https://doi.org/10.1038/s41586-019-1799-6 PMID: 31894144

52. Wardi G, Carlile M, Holder A, Shashikumar S, Hayden S, Nemati S. Predicting progression to septic shock in the emergency department using an externally generalizable machine-learning algorithm. Ann Emerg Med. 2021;77(4):395–406.

53. Brann F, Sterling NW, Frisch SO, Schrager JD. Sepsis prediction at emergency department triage using natural language processing: retrospective cohort study. JMIR AI. 2024;3(1):e49784.

54. Lee Y-TH, Di M, Schrager JD, Buckareff Z, Patzer RE, Yaffee AQ. The use of a self-triage tool to predict COVID-19 cases and hospitalizations in the state of Georgia. West J Emerg Med. 2022;23(4):532–5. https://doi.org/10.5811/westjem.2022.4.55001 PMID: 35980417

55. Miner AS, Laranjo L, Kocaballi AB. Chatbots in the fight against the COVID-19 pandemic. NPJ Digit Med. 2020;3:65. https://doi.org/10.1038/s41746-020-0280-0 PMID: 32377576

56. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ. 2015;351:h3480. https://doi.org/10.1136/bmj.h3480 PMID: 26157077

57. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020;20(5):533–4. https://doi.org/10.1016/S1473-3099(20)30120-1 PMID: 32087114

58. Sterling NW, Brann F, Patzer RE, Di M, Koebbe M, Burke M, et al. Prediction of emergency department resource requirements during triage: an application of current natural language processing techniques. J Am Coll Emerg Physicians Open. 2020;1(6):1676–83. https://doi.org/10.1002/emp2.12253 PMID: 33392576

59. Caton S, Haas C. Fairness in machine learning: a survey. ACM Comput Surv. 2024;56(7):1–38. https://doi.org/10.1145/3616865

60. Hort M, Chen Z, Zhang JM, Harman M, Sarro F. Bias mitigation for machine learning classifiers: a comprehensive survey. ACM J Responsib Comput. 2024;1(2):1–52. https://doi.org/10.1145/3631326

61. Feldman T, Peake A. End-to-end bias mitigation: removing gender bias in deep learning. arXiv preprint 2021. https://arxiv.org/abs/2104.02532

62. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv. 2021;54(6):1–35. https://doi.org/10.1145/3457607

63. Pessach D, Shmueli E. A review on fairness in machine learning. ACM Comput Surv. 2022;55(3):1–44. https://doi.org/10.1145/3494672

64. Pagano TP, Loureiro RB, Lisboa FVN, Peixoto RM, Guimarães GAS, Cruz GOR, et al. Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. BDCC. 2023;7(1):15. https://doi.org/10.3390/bdcc7010015

65. Ferrara E. Fairness and bias in artificial intelligence: a brief survey of sources, impacts, and mitigation strategies. Sci. 2023;6(1):3. https://doi.org/10.3390/sci6010003

66. Mittermaier M, Raza MM, Kvedar JC. Bias in AI-based models for medical applications: challenges and mitigation strategies. NPJ Digit Med. 2023;6(1):113. https://doi.org/10.1038/s41746-023-00858-z PMID: 37311802

67. Gottlieb ER, Ziegler J, Morley K, Rush B, Celi LA. Assessment of racial and ethnic differences in oxygen supplementation among patients in the intensive care unit. JAMA Intern Med. 2022;182(8):849–58. https://doi.org/10.1001/jamainternmed.2022.2587 PMID: 35816344

68. Friedman B, Nissenbaum H. Bias in computer systems. ACM Trans Inf Syst. 1996;14(3):330–47. https://doi.org/10.1145/230538.230561

69. Ueda D, Kakinuma T, Fujita S, Kamagata K, Fushimi Y, Ito R, et al. Fairness of artificial intelligence in healthcare: review and recommendations. Jpn J Radiol. 2024;42(1):3–15. https://doi.org/10.1007/s11604-023-01474-3 PMID: 37540463

70. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight - reconsidering the use of race correction in clinical algorithms. N Engl J Med. 2020;383(9):874–82. https://doi.org/10.1056/NEJMms2004740 PMID: 32853499

71. Suresh H, Guttag J. A framework for understanding unintended consequences of machine learning. arXiv preprint 2019. https://arxiv.org/abs/1901.10002

72. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: implications for clinical decision-making. PLOS Digit Health. 2024;3(11):e0000651. https://doi.org/10.1371/journal.pdig.0000651 PMID: 39509461

73. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. PLoS Med. 2018;15(11):e1002689.

74. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: mapping the debate. Big Data Soc. 2016;3(2):2053951716679679. https://doi.org/10.1177/2053951716679679

75. Kherbache A, Mertens E, Denier Y. Moral distress in medicine: an ethical analysis. J Health Psychol. 2022;27(8):1971–90. https://doi.org/10.1177/13591053211014586 PMID: 33938314

76. Kerasidou A. Trust me, I'm a researcher!: the role of trust in biomedical research. Med Health Care Philos. 2017;20(1):43–50. https://doi.org/10.1007/s11019-016-9721-6 PMID: 27638832

77. Blease C, Kaptchuk TJ, Bernstein MH, Mandl KD, Halamka JD, DesRoches CM. Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. J Med Internet Res. 2019;21(3):e12802. https://doi.org/10.2196/12802 PMID: 30892270

78. Larson HJ, Jarrett C, Eckersberger E, Smith DM, Paterson P. Understanding vaccine hesitancy around vaccines and vaccination from a global perspective: a systematic review of published literature, 2007–2012. Vaccine. 2014;32(19):2150–9.

79. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc. 2018;25(8):1080–8. https://doi.org/10.1093/jamia/ocy052 PMID: 29788380

80. Carter SM, Rogers W, Win KT, Frazer H, Richards B, Houssami N. The ethical, legal and social implications of using artificial intelligence systems in breast cancer care. Breast. 2020;49:25–32. https://doi.org/10.1016/j.breast.2019.10.001 PMID: 31677530

81. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med. 2018;378(11):981–3. https://doi.org/10.1056/NEJMp1714229 PMID: 29539284

82. Price WN, Cohen IG. Privacy in the age of medical big data. Nat Med. 2019;25(1):37–43. https://doi.org/10.1038/s41591-018-0272-7 PMID: 30617331

83. Corbett-Davies S, Gaebler J, Nilforoshan H, Shroff R, Goel S. The measure and mismeasure of fairness. J Mach Learn Res. 2023;24(1):14730–846.

84. Benjamin R. Race after technology: abolitionist tools for the new Jim code. Medford, MA: Polity Press. 2019.

85. Hardt M, Cila N, Desmet P. Love. The forgotten dimension for just and democratic AI Futures. In: Proceedings of DRS. Design Research Society. 2024. https://doi.org/10.21606/drs.2024.909

86. Howell MD, Corrado GS, DeSalvo KB. Three epochs of artificial intelligence in health care. JAMA. 2024;331(3):242–4. https://doi.org/10.1001/jama.2023.25057 PMID: 38227029

87. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. ACM. 2012. p. 214–26. https://doi.org/10.1145/2090236.2090255

88. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. Adv Neural Inf Process Syst. 2016;29.

89. Le Quy T, Roy A, Iosifidis V, Zhang W, Ntoutsi E. A survey on datasets for fairness-aware machine learning. WIREs Data Min Knowl. 2022;12(3):e1452. https://doi.org/10.1002/widm.1452

90. Saravanakumar KK. The impossibility theorem of machine fairness—a causal perspective. arXiv preprint. 2020. https://arxiv.org/abs/2007.06024

91. Kleinberg J, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores. arXiv preprint 2016. https://arxiv.org/abs/1609.05807

92. Zhang W, Weiss J. Fair decision-making under uncertainty. In: Proceedings of the IEEE International Conference on Data Mining (ICDM). 2023. p. 886–95. https://doi.org/10.48550/arXiv.2301.12364

93. Wang Z, Ulloa D, Yu T, Rangaswami R, Yap R, Zhang W. Individual fairness with group constraints in graph neural networks. Frontiers in artificial intelligence and applications. IOS Press. 2024. https://doi.org/10.3233/faia240679

94. vanderMaaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(86):2579–605.

95. Grari V, Lamprier S, Detyniecki M. Fairness without the sensitive attribute via causal variational autoencoder. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization. 2022. p. 696–702. https://doi.org/10.24963/ijcai.2022/98

96. Zhang BH, Lemoine B, Mitchell M. Mitigating unwanted biases with adversarial learning. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2018. p. 335–40.

97. Ribeiro M, Singh S, Guestrin C. Why should I trust you?: explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM. 2016. 1135–44.

98. Hui W, Lau WK. Detecting and mitigating algorithmic bias in binary classification using causal modeling. In: 2024 4th International Conference on Computer Communication and Information Systems (CCCIS). IEEE. 2024. p. 47–51. https://doi.org/10.1109/cccis63483.2024.00016

99. Kusner M, Loftus J, Russell C, Silva R. Counterfactual fairness. Adv Neural Inf Process Syst. 2017;30:4066–76.

100. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM. 2019. p. 220–9. https://doi.org/10.1145/3287560.3287596

101. Ruano-Ordás D, Fdez-Riverola F, Méndez JR. Concept drift in e-mail datasets: An empirical study with practical implications. Inf Sci. 2018;428:120–35. https://doi.org/10.1016/j.ins.2017.10.049

102. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216–9. https://doi.org/10.1056/NEJMp1606181 PMID: 27682033

103. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, III HD, et al. Datasheets for datasets. Commun ACM. 2021;64(12):86–92. https://doi.org/10.1145/3458723

104. Kamiran F, Calders T. Data preprocessing techniques for classification without discrimination. Knowl Inf Syst. 2012;33(1):1–33.

105. Lum K, Johndrow J. A statistical framework for fair predictive algorithms. arXiv preprint 2016. https://arxiv.org/abs/1610.08077

106. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.

107. Calmon F, Wei D, Vinzamuri B, Natesan Ramamurthy K, Varshney KR. Optimized pre-processing for discrimination prevention. Adv Neural Inf Process Syst. 2017:30.

108. Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C. Learning fair representations. In: Proceedings of the 30th International Conference on Machine Learning. PMLR. 2013. p. 325–33.

109. Deldjoo Y, Noia TD, Merra FA. A survey on adversarial recommender systems. ACM Comput Surv. 2021;54(2):1–38. https://doi.org/10.1145/3439729

110. Zafar M, Valera I, Rodriguez M, Gummadi K. Fairness constraints: mechanisms for fair classification. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. PMLR. 2017. p. 962–70.

111. Elzayn H, Jabbari S, Jung C, Kearns M, Roth A. Fair algorithms for learning in allocation problems. In: Proceedings of the Conference on Fairness Accountability and Transparency. ACM. 2019. p.170–9.

112. Donini M, Oneto L, Ben-David S, Shawe-Taylor JS, Pontil M. Empirical risk minimization under fairness constraints. Adv Neural Inf Process Syst. 2018:31.

113. Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger K. On fairness and calibration. Adv Neural Inf Process Syst. 2017:30.

114. Siontis KC, Zhang X, Eckard A, Bhave N, Schaubel DE, He K, et al. Outcomes associated with apixaban use in patients with end-stage kidney disease and atrial fibrillation in the United States. Circulation. 2018;138(15):1519–29. https://doi.org/10.1161/CIRCULATIONAHA.118.035418 PMID: 29954737

115. Fenton JJ, Xing G, Elmore JG, Bang H, Chen SL, Lindfors KK, et al. Short-term outcomes of screening mammography using computer-aided detection: a population-based study of medicare enrollees. Ann Intern Med. 2013;158(8):580–7. https://doi.org/10.7326/0003-4819-158-8-201304160-00002 PMID: 23588746

116. Henry C. Hospital closures: the sociospatial restructuring of labor and health care. Ann Assoc Am Geogr. 2015;105(5):1094–110.

117. Mittelstadt B. Principles alone cannot guarantee ethical AI. Nat Mach Intell. 2019;1(11):501–7.

118. Lubarsky B. Re-identification of "anonymized" data. GeoLTechRev. 2017;1:202.

119. Ling X, Fu J, Wang K, Li H, Cheng T, Chen Z. Fedfdp: fairness-aware federated learning with differential privacy. arXiv preprint 2024. https://arxiv.org/abs/2402.16028

120. Poulain R, Tarek M, Beheshti R. Improving fairness in AI models on electronic health records: the case for federated learning methods. arXIv preprint 2023. https://arxiv.org/abs/2305.11386

121. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019;6(2):94–8. https://doi.org/10.7861/futurehosp.6-2-94 PMID: 31363513

122. Luxton D. Artificial intelligence in psychological practice: current and future applications and implications. Prof Psychol Res Pr. 2014;45(5):332–9.

123. Hall MA, Orentlicher D, Bobinski MA, Bagley N, Cohen IG. Medical liability and treatment relationships. Aspen Publishing. 2018.

124. Hacker P. The European AI liability directives – critique of a half-hearted approach and lessons for the future. arXiv preprint 2023. https://doi.org/10.48550/arXiv.2211.13960

125. Terry NP. Regulatory disruption and arbitrage in health-care data protection. Yale J Health Policy Law Ethics. 2017;17(1):143–208. PMID: 29756756

126. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. Artif Intell Healthc. Elsevier. 2020. p. 295–336. https://doi.org/10.1016/b978-0-12-818438-7.00012-5

127. Holzinger A, Biemann C, Pattichis C, Kell D. What do we need to build explainable AI systems for the medical domain? arXiv preprint 2017. https://arxiv.org/abs/1712.09923

128. Lundberg S, Lee S. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30.

129. Chen I, Johansson F, Sontag D. Why is my classifier discriminatory? Adv Neural Inf Process Syst. 2018:31.

130. Goodman B, Flaxman S. European union regulations on algorithmic decision making and a "right to explanation". AI Magazine. 2017;38(3):50–7. https://doi.org/10.1609/aimag.v38i3.2741

131. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. Practical guidance on artificial intelligence for health-care data. Lancet Digit Health. 2019;1(4):e157–9. https://doi.org/10.1016/S2589-7500(19)30084-6 PMID: 33323184

132. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319(13):1317–8. https://doi.org/10.1001/jama.2017.18391 PMID: 29532063