The American Journal of Physical Education and Health Science (AJPEHS)

ISSN: 2992-9679 (ONLINE)

VOLUME 4 ISSUE 2 (2026)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Physical Education and Health Science (AJPEHS)

Volume 4 Issue 2, Year 2026
ISSN: 2992-9679 (Online)
DOI: https://doi.org/10.54536/ajpehs.v4i2.7562
https://journals.e-palli.com/home/index.php/ajpehs

Exploring AI Bias in Key Healthcare: The Role of Data Science

Yara Mohammed[1*], Manar Alsaid[2]

Article Information

Received: February 13, 2026

Accepted: April 23, 2026

Published: August 03, 2026

Keywords

Ai Bias, Algorithmic Fairness, Artificial Intelligence In Healthcare, Data Privacy, Systematic Review

ABSTRACT

Artificial Intelligence (AI) is increasingly integrated into healthcare to enhance decision-making; improve patient outcomes; and optimize healthcare operations. However, the adoption of AI in this sector presents challenges, particularly regarding ethical considerations, data privacy, and bias. The purpose of this study is to explore the types of AI bias reported in healthcare literature from 2018 to 2024 and to highlight the implications for AI implementation in medical practice. the paper discusses Data sciences strategies to mitigate these biases, highlighting the critical role for responsible and ethical AI development. It roles in role in mitigating and enhancing the fairness in ML and LLM. It then delves into the process of fine-tuning, outlining the steps taken to adapt the model to specific datasets and tasks. This study is a systematic review of papers published between 2018 and 2024. The literature was analyzed to identify various types of AI bias; including algorithmic bias; data bias; and ethical-human-related bias. Relevant articles were selected based on their contribution to understanding AI bias in healthcare and its effects on healthcare delivery. The analysis identified several types of AI bias prevalent in healthcare. Algorithmic bias; data bias; and ethical-human-related bias were the primary categories. These biases can distort the decision-making process; impact patient outcomes; and challenge the fundamental principles of medical ethics. The review also found that human and data biases are significant factors contributing to AI bias in healthcare systems. The study highlights the need of data sciences for addressing AI-related biases in healthcare to ensure the successful and ethical implementation of AI technologies. Special attention is given to avoiding overfitting and optimizing hyperparameters during the fine-tuning process. Future research should focus on developing strategies to mitigate these biases; ensuring that AI systems are fair; reliable; and trustworthy. Identifying and addressing these biases is essential for the ethical integration of AI in healthcare and for protecting the integrity of the medical profession.

INTRODUCTION

AI bias is an issue that might be overlooked when people discuss the adoption and implementation of AI in any field. In healthcare, however, these concerns are important because AI systems must operate within frameworks of transparency, accountability, and equity to avoid reinforcing harm in clinical practice (Bouderhem, 2026; Rahman, 2026). The healthcare field can be more critical than others as they involve people's lives, and it is, in some cases, a matter of life and death (Hoffman & Podgurski, 2019). AI healthcare is a critical issue that can significantly impact patient outcomes, equity, and trust in medical systems. One major source of bias is data bias, where AI models are trained on datasets that may not represent diverse populations, such as the underrepresentation of certain ethnicities, genders, or age groups (Bol, 2024). One of the most critical types of bias is Algorithmic bias where flawed design or training can cause algorithms to prioritize certain groups over others, leading to less accurate or misrepresented diagnoses for certain types of groups such as underrepresented populations (Bol, 2024, Norori *et al.*, 2021). Human bias also plays a role, as subjective decision-making during data labeling or interpretation by clinicians can influence the data used to train AI models. Examples of bias in AI healthcare include racial bias, where systems show lower accuracy in diagnosing conditions like skin cancer in people with darker skin tones due to training on predominantly light-skinned datasets, and gender bias, where models may underperform in diagnosing conditions that present differently in women, such as heart disease symptoms (Bol, 2024; Giovanola & Tiribelli, 2023). Examples of bias in AI healthcare include racial bias, where systems show lower accuracy in diagnosing conditions such as skin cancer in people with darker skin tones due to training on predominantly light-skinned datasets, and gender bias, where models may underperform in diagnosing conditions that present differently in women, such as heart disease symptoms (Bol, 2024; Giovanola & Tiribelli, 2023). Socioeconomic bias is another concern, as AI systems trained on data from wealthier populations may not generalize well to underserved communities. In addition, gender-related differences in healthcare engagement may contribute to downstream inequities in health data, since women and men can differ in clinic attendance, treatment engagement, and delays in seeking care (Iyanda *et al.*, 2026).

The consequences of bias in AI healthcare are far-reaching. It can exacerbate health disparities, leading to worse outcomes for marginalized groups, and eroding

[1] University of North Texas College of Information, United States
[2] East Texas A&M University, United States
[*] Corresponding author's e-mail: YaraMohammed@my.unt.edu

trust in AI systems if they are perceived as unfair or inaccurate (Giovanola & Tiribelli, 2023). Additionally, biased AI can create legal and ethical challenges, especially if it harms patients. To mitigate these biases, it is essential to use diverse and representative data for training AI models, ensuring they include diverse populations to improve generalizability. Transparency and explainability are also crucial, as developing interpretable AI models allows biases to be identified and addressed. Regular audits of AI systems can help monitor biased outcomes and update models as needed, while ethical guidelines should be established to emphasize fairness and equity in AI healthcare applications.

The goal of this study is to explore AI bias, as reported in the key healthcare literature. Understanding the issues reported in the literature will enhance our understanding of AI bias and its impact on the adoption and implementation of AI in critical areas such as healthcare. This research paper attempts to address the following research question that will form the basis for the identification and selection of healthcare key publication. The questions are: What types of AI bias are reported in AI literature? What is the impact of AI bias on the adoption and implementation of AI in Healthcare? What types of recommendations and considerations are reported in AI literature?

LITERATURE REVIEW

Several studies discussed the different types of bias and their impact on decision-making. They outlined elements of potential bias in the development and implementation of AI in healthcare. Some of the studies provided actionable recommendations for mitigating data bias to prevent exacerbation of health disparities. The study by Nazer *et al.* (2023) highlighted a unique perspective on Data Bias with practical implications for addressing these challenges in healthcare. The study identifies various types of bias, including data bias, algorithmic bias, and outcome bias. It provides a framework for identifying and mitigating these biases, such as fairness audits, data balancing techniques, and inclusive algorithm design. A study by Hague (2019) discussed the importance of recognizing and managing biases that can emerge when AI is implemented in healthcare settings. The study emphasizes the need for proper model risk management (MRM) protocols, similar to those used in the financial industry, to ensure that AI models in healthcare do not inadvertently enforce existing biases or create new ones.

A study by Adegunle *et al.* (2026) examined bias in AI-driven clinical decision support tools and found that bias may stem from unrepresentative datasets, the absence of equity auditing, and insufficient transparency in reporting. The study also emphasized the importance of stronger governance frameworks to ensure that AI tools are implemented fairly across diverse patient populations. Another major form of bias in healthcare AI is data bias. Data-centric approaches have been used to evaluate dataset bias, particularly in relation to racial bias in healthcare, with the aim of improving dataset fairness and model performance through targeted analysis. Celi *et al.* (2022) explored how biases in AI perpetuate healthcare disparities by focusing on data sources and their impact on marginalized communities. The study suggested actionable steps to mitigate these biases and promote more equitable AI use in healthcare. Similarly, recent research in population health AI has reinforced the argument that data bias and algorithmic bias are often rooted in unrepresentative training data, and that fairness audits and transparency are necessary to prevent the reproduction of health disparities (Chukwuemelie *et al.*, 2026). Ethical challenges posed by data bias in machine learning systems were also discussed by Zhao *et al.* (2022). The authors examined the societal implications of bias and emphasized the need for accountability. They explored how unintentional biases in data collection and labelling can exacerbate inequalities in decision-making systems. A unique aspect of the study is its focus on the ethical frameworks that organizations can adopt to address these challenges. It also highlights the importance of diverse representation in training datasets to minimize bias. Related forms of bias in healthcare also extend beyond datasets and algorithms. Olson *et al.* (2024) examined how implicit and explicit weight bias negatively affect healthcare experiences for individuals with larger bodies, leading to more complex provider interactions, lower care quality, and worse health outcomes. The study also highlighted how societal norms favoring thinness and Whiteness reinforce these biases. Additionally, it criticized reliance on the Body Mass Index (BMI) as a health measure, arguing that it is flawed, not universally applicable, and may result in inappropriate medical advice and treatment. Chen *et al.* (2024) proposed three different mitigation methods: preprocessing, in-processing, and post-processing. This approach was used in various domains including healthcare. The authors explored how biases in data collection and preprocessing can influence decision-making. The authors highlight the importance of transparency in bias detection and propose mitigation strategies, such as algorithmic adjustments and human oversight. It underscores the need for diverse datasets to reduce representational bias. The study identified different types of bias—data bias (caused by imbalanced datasets), algorithmic bias (arising from model design and assumptions), and outcome bias (leading to disparities across different populations). These biases can contribute to inconsistencies in diagnostic accuracy, treatment recommendations, and access to healthcare

Ethical issues play an important role in healthcare where the profession relies heavily on establishing professional code ethics. Ethical theories can serve as a foundation for analyzing AI ethics and its use in healthcare (Zhao *et al.*, 2022; Zhou *et al.*, 2024). Recent research has further emphasized that healthcare AI ethics must be approached through an integrated framework that includes privacy preservation, fairness, governance, transparency, and equitable benefit sharing, particularly in federated learning

environments (Mir *et al.*, 2026). Each ethical theory offers unique insights into how ethical values can be established and implemented. For example, utilitarianism aims to maximize overall well-being by prioritizing AI that benefits the majority, though it may sometimes compromise the rights of minority groups. Deontology, on the other hand, emphasizes adherence to moral duties, such as respecting privacy and autonomy, but can struggle to adapt to AI's rapidly evolving landscape. Virtue ethics focuses on developing AI with qualities like honesty and compassion but lacks clear guidelines for resolving ethical conflicts. Other frameworks such as care ethics, feminist ethics, and environmental ethics provide alternative perspectives on AI's moral considerations. While each theory offers valuable insights, none of them fully capture the complexity of AI ethics (Poli *et al.*, 2025).

Nsoesie and Galea (2022) in a study on how data science can be used to address racial bias and healthcare equity, emphasized the importance of robustly interrogating datasets to promote health equity. The article examined data bias implications in various contexts, such as healthcare, finance, and AI systems. It provided detailed examples of how bias in data collection or preprocessing affects decision-making. Specific methodologies, including algorithmic adjustments, human oversight, and diverse datasets can be used to reduce representational bias. Recommendations are provided for organizations to adopt proactive bias management strategies including the use of a framework for identifying and mitigating bias such as fairness audits, data balancing techniques, and inclusive algorithm design.

MATERIALS AND METHODS

The study used a systematic literature review to explore the degree of existing publications that exhibit AI bias in healthcare. The PRISMA method was used to include and exclude papers, as outlined in Figure 1 below.

For the period from 2018-2024, within these seven years a significant phase where AI began to prominently emerge in healthcare applications. The relevance of AI during this time is underscored by its diverse applications, as highlighted during the COVID-19 pandemic. AI technologies played a crucial role in enhancing diagnostic and treatment decisions, improving contact tracing, and deploying advanced AI-driven solutions in response to the pandemic (Van der Schaar *et al.*, 2021; Habermann, 2021; Vaishya *et al.*, 2020).

These papers were sourced from reputable databases including Google Scholar, PubMed, Scopus, Web of Science, and arXiv for the period from 2018 to 2024. Although the present review is limited to studies published between 2018 and 2024, more recent literature confirms that fairness, transparency, and governance remain central concerns in healthcare AI, reinforcing the continued relevance of these issues beyond the review period (Adegunle *et al.*, 2026; Mir *et al.*, 2026). These papers mark a significant phase in which AI began to emerge more prominently in healthcare applications. Various terms were used to search the databases. Terms such as data bias, AI bias, and data transparency in healthcare were derived from the research question and the problem statement. The dataset was grouped by 'Publication Year' and 'Types of Bias', treating the publication year as a temporal variable and the types of biases as categorical variables. counted the number of publications for each type of bias per year to measure the annual research output focused on each specific bias.

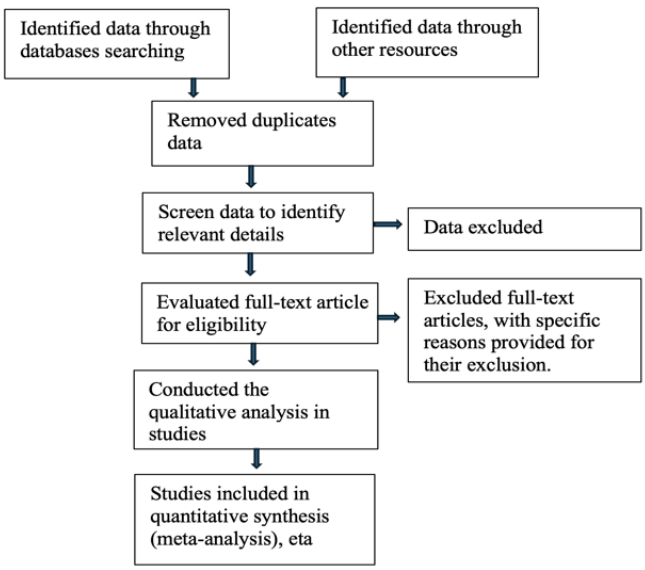

Figure 1: This figure explains the PRISMA process

The analysis was conducted using Python and involved identifying various types of biases recently found in healthcare by leveraging Large Language Models (LLMs). The process extracted relevant keywords and categorized different types of biases in healthcare AI, followed by a comprehensive trend analysis to observe how these

biases have been discussed and prioritized over the years. The goal was to track the evolution of research focus on different biases, offering insights into the shifting landscape and emerging priorities within the field of healthcare AI.

Furthermore, line graphs were generated to track the number of publications per bias type over time. This graphical representation facilitated the identification of trends, peaks, and shifts in research focus across different years. Additionally, a dendrogram was created to examine the thematic and relational patterns between bias types and keywords. This analysis helped uncover correlations and clusters within the dataset, highlighting significant thematic relationships.

A detailed correlation analysis was also performed to investigate the connections between various types of biases and their associated keywords. This approach identified patterns and co-occurrences, revealing interconnected themes and offering deeper insights into the dynamics of healthcare AI research.

RESULTS AND DISCUSSION

Figure 2 reflected the most significant research publications in AI healthcare over years. This surge can be attributed to several key factors, including heightened global awareness of healthcare inefficiencies exposed by the COVID-19 pandemic, increased funding and research grants targeting AI solutions in healthcare, and the maturation of AI technologies that have reached a level of sophistication necessary to address complex healthcare challenges. Furthermore, regulatory changes and heightened interest in ethical AI practices may have prompted more focused research into bias mitigation, reflecting a growing commitment within the academic and medical communities to ensure that AI technologies are both effective and equitable.

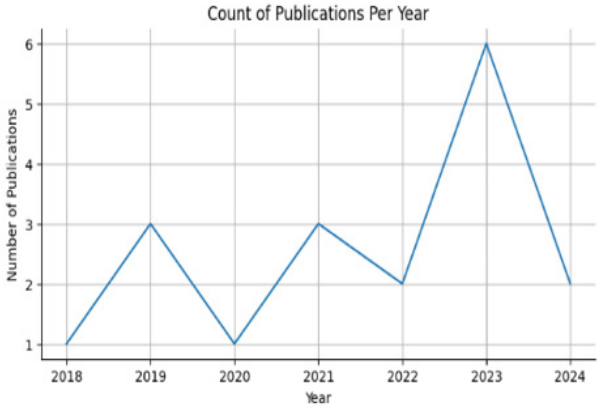

Figure 2: Recent Healthcare Publications

	Publication Year	Paper Title	Bias Keywords	Types of Bias	Link to Paper
0	2019	The Impact of Unconscious Bias in Healthcare: ...	unconcious bias	outcome bias	https://doi.org/10.1093/infdis/jiz214
1	2021	Implicit bias in healthcare: clinical practice...	implicit bias,cognitive bias	Algorithmic/outcome bias	https://doi.org/10.7861/fhj.2020-0233
2	2019	Benefits, Pitfalls, and Potential Bias in Heal...	not listed.	Algorithm Bias	Hague, D. C. (2019). Benefits, pitfalls, and p...
3	2021	Rising to the challenge of bias in health care AI	not listed.	outcome bias	Rising to the challenge of bias in health care...
4	2024	Mitigating Weight Bias in the Clinical Setting...	weight prejudice, weight bias, harm reduction,...	weight bias	https://onlinelibrary.wiley.com/doi/epdf/10.11...
5	2023	Unmasking Bias in AI: A Systematic Review of B...	Data bias, algorithmic bias	Data Bias	https://arxiv.org/abs/2310.19917
6	2023	An AI-Guided Data Centric Strategy to Detect a...	Data bias, AI, healthcare	Data Bias	https://arxiv.org/abs/2311.03425
7	2022	Towards Assessing Data Bias in Clinical Trials	Data bias, clinical trials	Data Bias	https://arxiv.org/abs/2212.09633
8	2021	Sources of bias in artificial intelligence tha...	Data bias, AI, healthcare	Data Bias	https://journals.plos.org/digitalhealth/articl...
9	2020	Addressing AI Algorithmic Bias in Health Care	Data bias, algorithmic bias	Data Bias	https://jamanetwork.com/journals/jama/fullarti...
10	2019	Bias in artificial intelligence algorithms and...	Data bias, AI algorithms	Data Bias	https://journals.plos.org/digitalhealth/articl...
11	2018	Bias in medical AI: Implications for clinical ...	Data bias, medical AI	Data Bias	https://journals.plos.org/digitalhealth/articl...
12	2022	Ethical Implications of Data Bias in Machine L...	Data bias, ethical AI, machine learning	Ethical Bias	https://doi.org/10.1016/j.patter.2022.100611
13	2024	Towards Better Data Science to Address Racial ...	Data bias, racial disparities, predictive anal...	Data Bias	https://academic.oup.com/pnasnexus/article/1/3...
14	2023	Unmasking Bias in Artificial Intelligence: A S...	Data bias, healthcare AI, mitigation	Data Bias	https://academic.oup.com/jamia/article/31/5/11...
15	2023	Gender Bias in AI: Impacts on Healthcare Outcomes	Gender bias, medical AI, fairness	Algorithmic Bias	https://doi.org/10.1016/j.jbi.2023.104021
16	2023	Data Bias Management	Data bias, fairness, machine learning	Data Bias	https://arxiv.org/abs/2305.09686
17	2023	Bias and Unfairness in Machine Learning Models...	Data bias, algorithmic bias, societal bias	Data Bias	https://www.mdpi.com/2504-2289/7/1/15

Figure 3: Evolution of Research on Bias in Healthcare AI

Starting in 2018, the research landscape began to intensively address the implications of biases in medical AI and their impact on clinical decision-making. This early focus laid the foundation for a deeper exploration into the nuances of how data biases could significantly skew healthcare outcomes. By 2019, pivotal papers such as "The Impact of Unconscious Bias in Healthcare: How to Recognize and Mitigate It" and "Benefits, Pitfalls, and Potential Bias in Health Care AI" were instrumental in setting the groundwork for understanding how both implicit and algorithmic biases can affect healthcare outcomes. These studies underscored the critical need for addressing unconscious biases that stem from stereotypes and the integration of AI in medical settings. As the discourse evolved, by 2021, attention shifted towards more specific applications of AI, as exemplified by studies like "Rising to the challenge of bias in health care AI," which focused on outcome biases related to AI-driven policy decisions during the COVID-19 pandemic. This trend towards addressing data biases in AI continued to expand through 2023 and 2024, with significant contributions such as "Unmasking Bias in AI: A Systematic Review of Bias Detection and Mitigation Strategies in Electronic Health Record-based Models" and "Towards Better Data Science to Address Racial Bias and Health Equity." These works explore comprehensive strategies to mitigate data and racial biases in AI applications, highlighting a growing acknowledgment of the complex interplay between AI technology and the perpetuation of biases within healthcare. This progression points towards an urgent need for robust bias management frameworks to ensure fair and equitable healthcare delivery, reflecting the evolving understanding and approaches to tackling these critical issues from 2018 onwards as shown in figure 3.

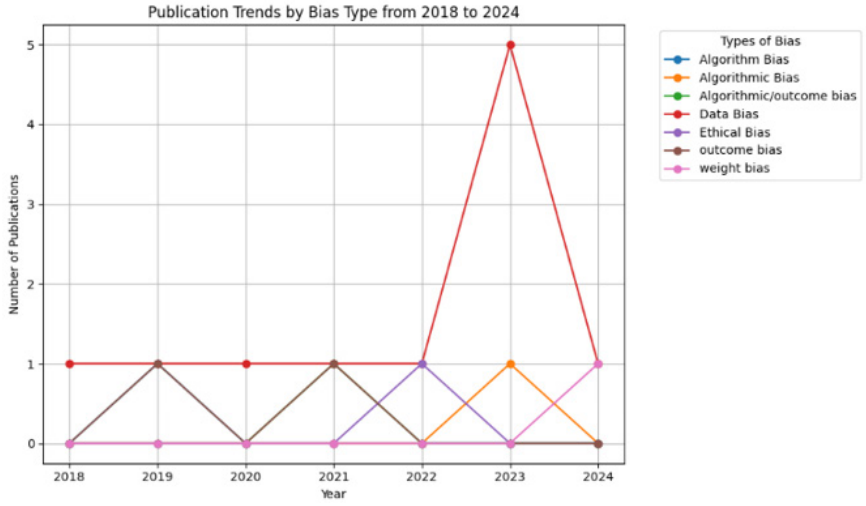

Figure 4: Publication Trends by Bias Type

The analysis reveals both consistency and fluctuations in research trends related to bias in AI and healthcare. Certain bias types, such as 'Algorithmic Bias' and 'Data Bias,' demonstrated relatively stable trends with minor variations, indicating a sustained interest in these areas over time. Figure 4 indicates a significant peak was observed in publications on 'Weight Bias' in 2023, suggesting a surge in attention likely influenced by external factors such as policy shifts, increased public awareness, or groundbreaking research findings. In contrast, 'Ethical Bias' and 'Outcome Bias' exhibited minimal activity throughout the observed period, indicating that these areas might be less prioritized or under-explored. These findings highlight key implications: the sharp rise in 'Weight Bias' research warrants further investigation into its underlying causes and broader societal impact. Meanwhile, the steady focus on 'Algorithmic Bias' and 'Data Bias' reinforces their critical role in AI ethics discussions. Lastly, the limited attention given to 'Ethical Bias' and 'Outcome Bias' suggests potential gaps in the research landscape, presenting opportunities for future studies.

Figure 4 provides a detailed exploration of publication trends categorized by specific bias types in healthcare AI from 2018 to 2024. While the overall surge in 2023 highlighted in Figure 2 is evident, this figure reveals the nuanced distribution of research focus across different bias categories. The data highlights that 'Data Bias' and 'Weight Bias' garnered significant attention in 2023, reflecting increased concerns about data quality and representation in healthcare AI. Conversely, other biases, such as 'Ethical Bias' and 'Outcome Bias,' experienced minimal activity, signaling potential research gaps in these important areas. This breakdown by bias type highlights the shifting priorities within the field, offering valuable insights into the areas of bias that have captured the academic community's interest and those requiring further exploration.

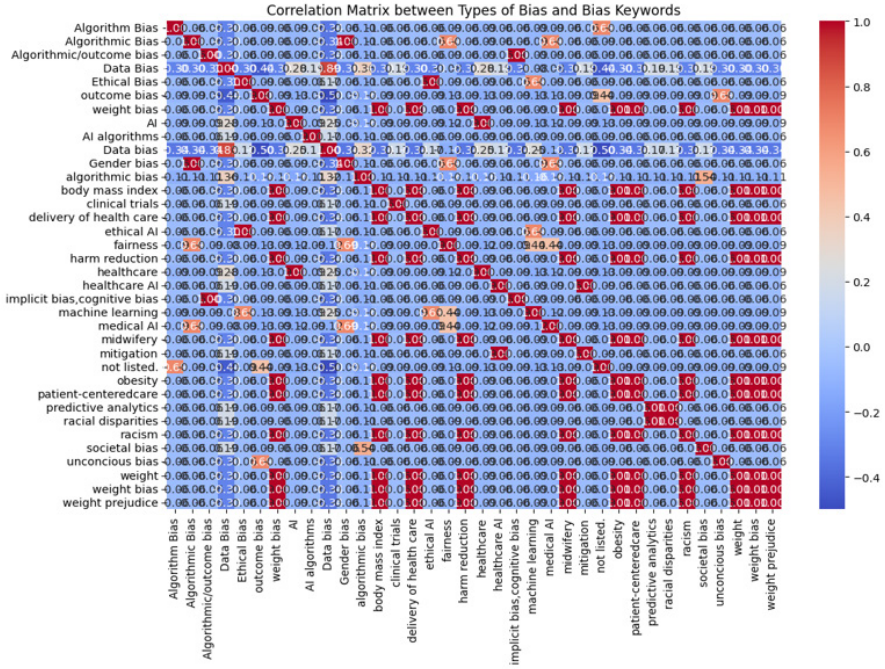

Figure 5: Heatmap Correlation Matrix in Healthcare AI

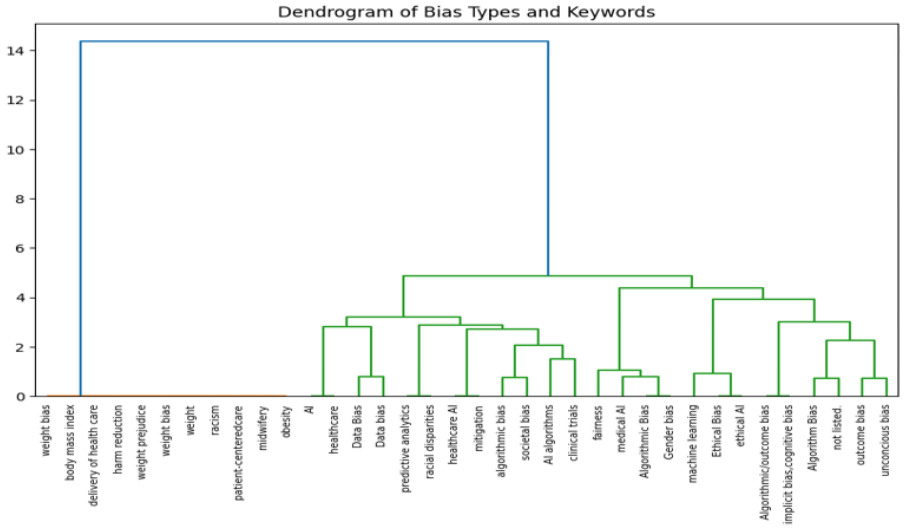

Figure 6: Exploring Connections Between Bias Types and Keywords in Healthcare AI

The heatmap correlation matrix in Healthcare AI (Figure 5) initially provided an overview of the relationships between various biases and keywords. However, due to the many keywords and overlapping correlations, the heatmap became visually complex and less interpretable. to understand the previous complexity the dendrogram was employed (Figure 6) to clearly specify and illustrate the hierarchical relationships and clustering between biases and keywords. The dendrogram simplifies the interpretation by grouping related biases and keywords into distinct clusters, offering a more structured and comprehensible view of their interconnections.

Figure 6 reveals several significant findings that shed light on the thematic and relational patterns within the dataset. "Algorithmic Bias" clusters strongly with "AI algorithms," "machine learning," and "data bias," reflecting the technical challenges associated with bias arising from the design of algorithms and the quality of training datasets. This underscores the role of unrepresentative datasets as a primary contributor to biased algorithmic outcomes in healthcare AI. Similarly, "Data Bias" forms a cluster with keywords such as "racial disparities," "predictive analytics," and "healthcare AI," highlighting the relationship between flawed data representation and inequities in healthcare delivery. These associations emphasize the need to address data-driven systemic inequities, particularly for marginalized groups.

"Outcome Bias" is observed to correlate with keywords like "implicit bias," "patient-centered care," and "healthcare," signifying a focus on the inequities in healthcare outcomes arising from clinician biases and organizational inefficiencies. Additionally, the clustering

of "weight bias" and "weight prejudice" with terms such as "body mass index" and "obesity" draws attention to biases tied to cultural norms and societal perceptions of weight, which directly affect treatment strategies and healthcare interactions. Moreover, "Ethical AI" and "fairness" are clustered with biases like "Algorithmic Bias" and "Data Bias," illustrating the growing emphasis on implementing ethical frameworks and fairness audits to mitigate these challenges. Interestingly, the findings also highlight interdisciplinary concerns, with keywords such as "midwifery," "racism," and "societal bias" forming connections with both technical (e.g., "AI algorithms") and clinical (e.g., "healthcare") themes, indicating a holistic approach to addressing biases.

Practical Implementation

The paper identifies different type of bias, with help of data science the paper proposed ideas and recommendation that will benefit the audients, or the readers limit the AI bias in health care. In the delves into the process of fine-tuning, outlining the steps taken to adapt the model to specific datasets and tasks. Special attention is given to avoiding overfitting and optimizing hyperparameters during the fine-tuning process. Include recommendation on enhancing large language models (LLMs) and Machine learning (ML) to improve natural language understanding, with a strong emphasis on addressing ethical challenges in AI development. Additionally, the proposed strategies to mitigate these biases, highlighting the critical need for responsible and ethical AI development.

Dataset preparation

The DS can help in dealing of different tape of Baies. One of the primary sources of AI bias is the dataset used to train models. If the datasets lack diversity and fail to represent all demographic groups, such as racial, ethnic, gender, or socioeconomic categories, the resulting models may be biased, particularly against underrepresented groups. To address this issue, it is essential to ensure that training datasets are both diverse and representative of the broader population. This includes ensuring adequate representation from all relevant groups, including different races, genders, ages, and medical conditions. A practical approach involves actively collecting and curating data from varied populations, ensuring that all subgroups are well-represented. This may require making a conscious effort to include data from minority and underserved populations to create a more balanced dataset, helping to prevent AI systems from disproportionately favoring one group over others. The data scenes technics such as Data-centric approaches can eliminate bias in machine learning models. González-Sendino *et al.* (2024) used causal Bayesian models to restructure datasets before training, keeping sensitive attributes for auditing. Their method is fair by removing unnecessary effects and dealing with multiple biased factors. It is essential to train LLM on various types of data that include text, image, and so on. The expanding availability of biomedical data from biobanks, electronic health records, medical imaging, and wearable devices has created new opportunities for multimodal AI in healthcare.

Kamatala *et al.* (2025) emphasize the importance of data preprocessing and algorithmic fairness constraints during training, allowing adversarial debiasing while limiting sensitive traits.

LLM and Machine Learning (ML) fairness

Even with diverse datasets, AI models can still develop biases during training due to underlying structural inequalities present in the data or within the model itself. These biases can lead to unfair predictions and discriminatory outcomes. Enhancing (LLMs) to improve natural language understanding, with a strong emphasis on addressing ethical challenges in AI development. It explores key ethical concerns related to LLMs, including the risks of bias and unintended outcomes (Radanliev, 2025).

To address this, fairness-enhancing algorithms can be implemented to explicitly target fairness in AI predictions. Techniques such as adversarial debiasing, fairness constraints, and bias correction methods are integrated during the training process to minimize discriminatory outcomes. By adjusting the model's learning mechanism, these methods help ensure that harmful biases are not learned or perpetuated, resulting in more equitable AI systems that provide fairer predictions across different demographic groups.

The definition of XAI extends past technical interpretability tools to incorporate institutional explanation methods, which include documentation systems and human oversight mechanisms. Through previses works demonstrates that fairness-aware ML systems and LLM require explanation across both their logical processes and their social and procedural dimensions. According to the combined analysis of these authors, the implementation of fairness and XAI requires dual-level strategies that combine explanations at both model and system levels (why specific outputs were generated and who controlled fairness decisions with their evolution). Chukwudi's (2024) work is very useful because it uses new methods but makes things harder to understand. The method works to reduce intersectional bias, but it also makes models that are more complicated than what normal XAI tools can explain meaningfully. The post hoc interpretability methods do not give enough information about how adversarial models work, which is why they try to hide sensitive information from the training process (Mohanarajesh *et al.*,2024).

study shows that model bias happens at many points in the AI process, from gathering data to making organizational decisions. The research focuses on two main sources of bias: automated hiring systems with feedback loops and credit scoring algorithms that use proxy discrimination. The implementation of Responsible AI governance frameworks requires auditing procedures and ethics boards to work alongside stakeholder inclusion at each stage of

the model lifecycle. The analysis highlights the importance of explainability through transparency by showing that model transparency must extend beyond algorithmic outputs to include details about model builders as well as data assumptions and decision implementation. The definition of XAI extends past technical interpretability tools to incorporate institutional explanation methods, which include documentation systems and human oversight mechanisms. Demonstrates that fairness-aware ML systems require explanation across both their logical processes and their social and procedural dimensions. According to the combined analysis of these authors, the implementation of fairness and XAI requires dual-level strategies that combine explanations at both model and system levels (why specific outputs were generated and who controlled fairness decisions with their evolution) (Varsha, 2023).

Multimodal AI in Healthcare

The paper recommend that generate multiped AI models for the various medical setting. Multimodal (MLLMs) significant research area, where LLMs handle tasks across different data types and exhibit advanced capabilities like mathematical reasoning and generating narratives from images. The multimodal model that combines a visual encoder with an LLM to understand and produce dialogue based on video content. Previous studies introduced HeLM, a framework that enables LLMs to assess disease risk by leveraging diverse clinical data (Belyaeva *et al.*, ??). the LLMs as a core component for handling multimodal tasks, exhibiting unexpected abilities like performing mathematical reasoning and generating stories from images. The use LLMs as a core component for handling multimodal tasks, exhibiting unexpected abilities like performing mathematical reasoning and generating stories from images. (Maaz *et al.*, 2023) in developed VideoChatGPT, combining a visual encoder with an LLM to comprehend and generate dialogues about videos. (Lyu *et al.*, 2023) proposed MACAW-LLM, which integrates text, audio, and visual data for seamless multimodal language modeling. LLaVA-Med introduces a cost-effective method to teach vision-language models to answer biomedical image-related queries. Meanwhile, systems like ELIXR aim to align language models with medical imaging, enabling broader tasks like visual question answering (Fu *et al.*, 2023; Li *et al.*, 2023). These advancements in multimodal biomedical AI hold great potential the multifacete nature of medical decision-making for enhancing and healthcare outcomes. However, to fully realize this potential, challenges related to data integration, privacy concerns, and technical limitations must be addressed.

Multimodal Model Evaluation

They are large number of way to available to assess the effectiveness of MLLM and evaluat it's standard (Li *et al.*, 2023). such as The SEED-Bench is another benchmark used to compare MLLMs with generative comprehension (Li *et al.*, 2023). Multi-modal pre-training and that the suggested benchmark dataset can efficiently assess the performance of vision-language (VL) models for report production and medical VQA medical modalities. BenchMD is a standard for integrated learning on sensors and medical pictures. MedPerf provides a secure and privacy-preserving platform for benchmarking medical AI models, allowing researchers to evaluate the performance of their models without compromising patient privacy (Fu *et al.*, 2023; Soenksen *et al.*, 2022; Wantlin *et al.*, 2023 ;Xu *et al.*2023). In additional, Governments and regulatory bodies should create guidelines that mandate fairness, transparency, and accountability in AI systems, especially in sensitive sectors like healthcare. This includes establishing ethical standards for AI development, ensuring compliance with data privacy laws.

Educating the Healthcare

It is very critical to educate the providers and heath clear providers about these AI Bais issues. The AI model might start performer Baise, if they are aware of bias types they will be able to idented and reported to the AI team. Which help of modify the AI, and the datasets. Also, the AI team need to be educated about the heather care police and Ethics before they fined tuned the AI mode.

Limitations

This study focuses on the period from 2018 to 2024, a foundational phase in which AI became more visibly integrated into healthcare and concerns regarding bias, fairness, and transparency became increasingly prominent. While newer studies published after 2024 continue to highlight these concerns, the present review provides a structured baseline for understanding how these issues emerged and developed during this important period. Additionally, while this study identifies key AI biases in healthcare, further work is needed to examine bias-related keywords within each identified category. A more detailed review of specific terms used in bias discussions would allow for a deeper understanding of how these biases emerge in different healthcare applications. This would also help develop more effective strategies to reduce bias and improve the fairness and reliability of AI-driven healthcare systems. Mandal *et al.* (2021) raise concerns about biases in visual datasets used for computer vision, showing that the location where data is collected affects the dataset's representation. Since search engines personalize results based on location, datasets tend to overrepresent the demographics of the region they are retrieved from. This leads to AI models that work well in certain areas but may be inaccurate or unfair when applied globally. Similar concerns apply to healthcare AI, where models trained on data from a single region may not generalize well to diverse populations, increasing the risk of biased medical decisions. Expanding the dataset to include geographical, institutional, and policy-based variations could further strengthen the study. Since AI biases can be influenced by regional regulations,

demographic factors, and evolving healthcare policies, a long-term and cross-regional study would provide a more nuanced understanding of bias trends and their effects on AI adoption in healthcare. Furthermore, Manvi *et al.* (2024) found that LLMs are shaped by biases in their training data based on how they interpret different regions. Their study showed that while LLMs can accurately recognize geographic patterns, they tend to favor high-income regions and show bias against lower-income areas, especially in assessments of intelligence and morality. This highlights the need to ensure that AI systems, including those in healthcare, do not reinforce biases that could negatively affect underrepresented communities. To address these geographic and economic biases in AI, future research should focus on expanding datasets to include more diverse sources and creating more equitable AI models that work effectively across different regions and populations.

CONCLUSION

This study provides valuable insights into the evolving landscape of healthcare AI research, focusing on biases identified in publications from 2018 to 2024. Through trend analysis, we observed shifts in the attention paid to different types of biases, with notable spikes in research on "Weight Bias" in 2023, alongside consistent interest in "Algorithmic Bias" and "Data Bias." These trends highlight the dynamic nature of research priorities, influenced by societal, technological, and policy-driven factors. The Analysis offers a snapshot of how the field addresses biases in AI, it also underscores certain gaps, such as the limited focus on "Ethical Bias" and "Outcome Bias." The paper practical implementation the role of data sciences to limit the AI bias. These gaps present opportunities for future research to explore underrepresented areas and ensure a more comprehensive approach to addressing biases in healthcare AI. Furthermore, the findings reveal the need for robust frameworks to manage and mitigate biases as AI systems become increasingly integral to healthcare delivery. While this study illuminates current trends in healthcare AI bias research, it serves as a foundation for future work to refine methodologies, address limitations, and ultimately foster the development of equitable and ethical AI systems in healthcare. Further study may use advance analysis such as machine learning technique to find the hidden bias pattern biases in AI applications, the healthcare sector can move toward more inclusive, fair, and trustworthy technologies that benefit all populations.

REFERENCES

Adegunle, F., Chhatwal, K., Arab, S., Alabdaljabar, M. S., Raslan, M. A., Sayed, O., & Goldsweig, A. M. (2026). Bias and oversight in clinical AI: A review of decision support tools and equity frameworks. *Journal of General Internal Medicine,* 1–12.

Belyaeva, A., Cosentino, J., Hormozdiari, F., McLean, C. Y., & Furlotte, N. A. (2023). *Multimodal LLMs for health grounded in individual-specific data* (arXiv:2307.09018). arXiv. https://arxiv.org/abs/2307.09018

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. (2023). Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125,.*

Bol, K. W. O. (2024). *Impact of bias in AI and human decision-making: A healthcare perspective* [Master's thesis, University of Twente].

Bouderhem, R. (2026). Artificial intelligence in healthcare: Ethical challenges and promoting clinical equity and transparency. In *Handbook of tissue reconstruction and regeneration* (pp. 1–29). Springer Nature Singapore.

Celi, L. A., Cellini, J., Charpignon, M. L., Dee, E. C., Dernoncourt, F., Eber, R., & Yao, S. (2022). Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. *PLOS Digital Health, 1*(3), Article e0000022.

Chen, F., Wang, L., Hong, J., Jiang, J., & Zhou, L. (2024). Unmasking bias in artificial intelligence: A systematic review of bias detection and mitigation strategies in electronic health record-based models. *Journal of the American Medical Informatics Association, 31*(5), 1172–1183.

Chukwudi, W. C. (2024). *Mitigating intersectional bias in machine learning: A novel approach to fairness in automated decision-making.* ResearchGate.

Chukwuemelie, N. D., Yussuf, M. D., & Abdul, S. O. (2026). Precision public health: Harnessing AI and big data for equitable health outcomes. *Center of Artificial Intelligence, 1*(1).

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., & Ji, R. (2023). *MME: A comprehensive evaluation benchmark for multimodal large language models* (arXiv:2306.13394). arXiv. https://doi.org/10.48550/arXiv.2306.13394

Giovanola, B., & Tiribelli, S. (2023). Beyond bias and discrimination: Redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. *AI & Society, 38*(2), 549–563.

Habermann, J. (2021). *Psychological impacts of COVID-19 and preventive strategies: A review.* [Publisher information needed]

Hague, D. C. (2019). Benefits, pitfalls, and potential bias in healthcare AI. *North Carolina Medical Journal, 80*(4), 219–223.

Hoffman, S., & Podgurski, A. (2019). Artificial intelligence and discrimination in healthcare. *Yale Journal of Health Policy, Law, & Ethics, 19*, 1.

Iyanda, E. O., Kakwagh, V. V., Gomment, T. I., Yunusa, E., & Owoyemi, J. O. (2026). Gender differences in health seeking behaviours among people living with HIV/AIDS in Benin City, Edo State-Nigeria. *American Journal of Medicine, 1*(1), 5–17.

Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., & Shan, Y. (2023). *SEED-Bench: Benchmarking multimodal LLMs with generative comprehension* (arXiv:2307.16125). arXiv.

https://arxiv.org/abs/2307.16125

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., & Gao, J. (2023). LLaVA-Med: *Training a large language-and-vision assistant for biomedicine in one day*. arXiv.

Lyu, C., Wu, M., Wang, L., Huang, X., Liu, B., Du, Z., Shi, S., & Tu, Z. (2023). *Macaw-LLM: Multi-modal language modeling with image, audio, video, and text integration* (arXiv:2306.09093). arXiv. https://arxiv.org/abs/2306.09093

Maaz, M., Rasheed, H., Khan, S., & Khan, F. S. (2023). *Video-ChatGPT: Towards detailed video understanding via large vision and language models* (arXiv:2306.05424). arXiv. https://arxiv.org/abs/2306.05424

Mandal, R., Kapoor, S., & Krishnaswamy, N. (2021). Geographical bias in visual datasets and its impact on facial analytics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Manvi, R., Khanna, S., Burke, M., Lobell, D., & Ermon, S. (2024). *Large language models are geographically biased* (arXiv:2401.01234). arXiv. https://arxiv.org/abs/2401.01234

Mir, B. A., Abbas, S. R., & Lee, S. W. (2026). Federated learning in healthcare ethics: A systematic review of privacy-preserving and equitable medical *AI*. *Healthcare, 14*(3), Article 306.

Mohanarajesh, K. (2024). *Develop new techniques for ensuring fairness in artificial intelligence and ML models to promote ethical and unbiased decision-making*. [Publisher information needed]

Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health, 2*(6), Article e0000278.

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for healthcare: A call for open science. *Patterns, 2(*10).

Nsoesie, E. O., & Galea, S. (2022). Towards better data science to address racial bias and health equity. *PNAS Nexus, 1*(3), Article pgac120.

Olson, S. M., Muñoz, E. G., Solis, E. C., & Bradford, H. M. (2024). Mitigating weight bias in the clinical setting: A new approach to care. *Journal of Midwifery & Women's Health, 69*(2), 180–190.

Poli, P. K. R., Pamidi, S., & Poli, S. K. R. (2025). Unraveling the ethical conundrum of artificial intelligence: A synthesis of literature and case studies. *Augmented Human Research, 10*(1), 2.

Radanliev, P. (2025). AI ethics: Integrating transparency, fairness, and privacy in AI development. *Applied Artificial Intelligence, 39(*1), Article 2463722. https://doi.org/10.1080/08839514.2025.2463722

Rahman, M. I. (2026). Algorithmic bias and transparency in artificial intelligence tools for nursing education: A scoping review. *Nurse Education in Practice*, Article 104756.

Ratwani, R. M., Sutton, K., & Galarraga, J. E. (2024). Addressing AI algorithmic bias in healthcare. *JAMA, 332*(13), 1051–1052. https://doi.org/10.1001/jama.2024.13486

Soenksen, L. R., Ma, Y., Zeng, C., Boussioux, L., Villalobos Carballo, K., Na, L., Wiberg, H. M., Li, M. L., Fuentes, I., & Bertsimas, D. (2022). Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ Digital Medicine, 5*(1), 149.

Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. (2020). Artificial intelligence (AI) applications for the COVID-19 pandemic. Diabetes & Metabolic Syndrome: *Clinical Research & Reviews, 14*(4), 337–339.

van der Schaar, M., Alaa, A. M., Floto, A., Gimson, A., Scholtes, S., Wood, A., McKinney, E., Jarrett, D., Lio, P., & Ercole, A. (2021). How artificial intelligence and machine learning can help healthcare systems respond to COVID-19. *Machine Learning, 110*(1), 1–14.

Varsha, P. S. (2023). How can we manage biases in artificial intelligence systems—A systematic literature review. *International Journal of Information Management Data Insights, 3*(1), Article 100165. https://doi.org/10.1016/j.jjimei.2023.100165

Wantlin, K., Wu, C., Huang, S.-C., Banerjee, O., Dadabhoy, F., Mehta, V. V., Han, R. W., Cao, F., Narayan, R. R., Colak, E., Adamson, A., Heacock, L., Tison, G. H., Tamkin, A., & Rajpurkar, P. (2023). BenchMD: *A benchmark for unified learning on medical images and sensors*. arXiv.

Xu, L., Liu, B., Khan, A. H., Fan, L., & Wu, X.-M. (2023). *Multi-modal pre-training for medical vision-language understanding and generation: An empirical study with a new benchmark*. arXiv.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2023). *A survey on multimodal large language models* (arXiv:2306.13549). arXiv. https://arxiv.org/abs/2306.13549

Zhao, T., Dai, E., Shu, K., & Wang, S. (2022). Towards fair classifiers without sensitive attributes: Exploring biases in related features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (pp. 1433–1442)