Unmasking racial bias in medical AI: a narrative review of evidence and implications

by *Francesco Mastrocola*[*], *Elisabetta Ferrara*, *Oscar Genovesi*[**]

Artificial Intelligence (AI) is transforming healthcare, promising improvements in diagnosis, treatment, and patient outcomes. However, racial bias persists and feeds inequities. This review inspects how bias manifests in medical AI domains, identifying unrepresentative training data and proxy variables. It explores mitigation strategies and knowledge gaps, spotting the interdisciplinary approach to fortify equitable and accountable medical AI.

Keywords: AI; bias; clinical; data; healthcare; racial bias.

Smascherare i pregiudizi razziali nell'intelligenza artificiale medica: una revisione narrativa delle prove e delle implicazioni

L'intelligenza artificiale (IA) sta trasformando l'assistenza sanitaria, promettendo miglioramenti nella diagnosi, nel trattamento e negli esiti clinici dei pazienti. Tuttavia, i pregiudizi razziali persistono alimentando disuguaglianze. Questa revisione esamina come i pregiudizi si manifestino nell'IA medica, identificando dati di formazione non rappresentativi e variabili proxy. Esplora strategie di mitigazione e lacune di conoscenza, individuando l'approccio interdisciplinare per rafforzare un'IA medica equa e responsabile.

Parole chiave: IA; pregiudizio; clinico; dati; assistenza sanitaria; pregiudizio razziale.

Introduction

The integration of artificial intelligence (AI) in healthcare represents a revolution holding both promise and peril for health equity. While AI technologies offer unprecedented diagnostic capabilities, mounting evidence suggests these systems may strengthen racial disparities in healthcare delivery.

DOI: 10.5281/zenodo.17523874

[*] Università Telematica "Leonardo da Vinci". francesco.mastrocola@unidav.it, elisabetta.ferrara@unidav.it.

[**] Fondazione Università "G. d'Annunzio". genovesioscar@gmail.com.

Paragraphs "Introduction", 1, 2, and 3 are attributed to F. Mastrocola; paragraph 3.1, 3.2, 3.3 are attributed to E. Ferrara; paragraphs 4 and "Conclusions" are attributed to E. Ferrara and O. Genovesi. All authors reviewed the work, approving the final version.

Healthcare data reflects historical inequities in access, delivery, and documentation. Machine learning algorithms may learn and perpetuate discrimination patterns embedded within medical records (Parasuraman, Manzey, 2010). Automation bias exacerbates these challenges, as healthcare providers may exhibit excessive reliance on AI-generated recommendations, overlooking clinical information outside algorithmic parameters (Parikh, Teeple, Navathe, 2019; Gigerenzer, Hoffrage, Kleinbölting, 1991).

Clinical data sources, such as EHRs and diagnostic testing patterns, reflect historical utilization differences across racial groups, representing disparities in care access rather than biological variations (Montavon, Samek, Müller, 2018). AI systems trained on such data may encode systemic inequities, creating self-reinforcing cycles of discriminatory healthcare delivery. This narrative review examined mechanisms through which AI can foster racial bias in medicine, analyzing current approaches to bias detection and mitigation.

1. Intersectional Framework for Understanding AI Bias in Healthcare

Algorithmic discrimination operates through interconnected systems of oppression rather than isolated identity categories. Crenshaw's foundational work demonstrates that experiences at intersections of multiple identities cannot be understood as simply the sum of separate discriminations (Crenshaw, 1989). His influence helps to understand AI bias, as highlighted by recent discussions on 'Real Talk: Intersectionality and AI' (Howard, 2023). Current AI fairness approaches have critical limitations – even when language biases based on race, ethnicity, and gender are mitigated in word embedding models, biases persist against intersectional groups such as "Mexican American females" (Guo, Caliskan, 2021).

Building on Collins's matrix of domination theory, AI bias operates within existing power structures that simultaneously privilege and marginalize different groups (Collins, 2019). Contemporary examples include image recognition applications that identify gender particularly poorly for dark-skinned women (Buolamwini, Gebru, 2018).

Critical technology studies position technology as inherently political rather than neutral. Benjamin's "New Jim Code" argues that automation can hide, speed, and deepen discrimination while appearing benevolent (Benjamin, 2019). Noble's "Algorithms of Oppression" demonstrates how search engines reinforce racism (Noble, 2018), while Crawford's "Atlas of AI" reveals AI as "a technology of extraction" (Crawford, 2021). These systems embed structural inequities from healthcare data and institutions into algorithmic processes.

The sociology of risk provides insights into how AI implementation creates new uncertainty and trust relationships. Brown and van Voorst's analysis reveals how AI technologies generate "cultures of hope" among professionals seeking technological solutions amid resource constraints, masking systematic disadvantages for intersectional populations (Brown, van Voorst, 2024). The "opacity" challenges of AI systems further complicate clinical understanding and decision-making (Burrell, 2016; Grote, Berens, 2020; Hawley, 2015).

2. Methodology

Literature was identified through PubMed, Scopus, and Web of Science databases using terms related to artificial intelligence, healthcare, and racial bias, covering publications through January 2025. We included peer-reviewed articles, theoretical frameworks, and policy analyses addressing algorithmic bias in healthcare contexts.

A narrative review approach was selected to synthesize diverse literature types and provide a comprehensive analysis across technical, ethical, and clinical perspectives.

3. Mechanisms of Algorithmic Bias in Healthcare

Racial bias in AI within healthcare operates through interrelated mechanisms. AI systems derive learning from historical medical records that embed documented patterns of healthcare inequities, creating differential misclassification bias (Gianfrancesco *et al.,* 2018). Algorithms trained on historically biased data systematically magnify existing healthcare disparities.

EHRs present three areas of bias amplification: missing data bias affecting marginalized populations with fragmented care, sample size bias when insufficient minority data leads to algorithmic defaulting to majority trends, and measurement error bias from suboptimal care patterns (Burrell, 2016). These biases manifest through "automation complacency", where healthcare providers demonstrate excessive reliance on algorithmic recommendations while overlooking clinical information outside algorithmic parameters (Topol, 2019).

Mitigation strategies include preprocessing approaches such as weighting methods and data augmentation for underrepresented groups (Cary *et al.*, 2023). However, debate persists regarding race and ethnicity variables in clinical algorithms (Norgeot, Glicksberg, Butte, 2019; Cross, Choma, Onofrey, 2024), reflecting the challenge of balancing demographically aware algorithms against perpetuating societal biases.

3.1. Evolution and Current Challenges of AI Prediction Models in Healthcare

Healthcare prediction models have evolved from scoring systems with small datasets to advanced AI algorithms analyzing complex, multimodal data. While these systems show promise in diagnostic accuracy and personalized medicine, mounting evidence reveals racial bias concerns (Gianfrancesco *et al.*, 2018). Obermeyer *et al.* (2019) demonstrated how a widely used healthcare algorithm exhibited racial bias, reducing identification of Black patients needing additional care by more than half compared to White patients.

Such biases occur through unrepresentative training data, measurement classification bias, and encoding of historical healthcare disparities into algorithmic systems. Data quality issues compound these problems, including missing data disproportionately affecting marginalized populations (Nijman *et al.*, 2022) and measurement errors from medical devices performing differently across racial groups (Charpignon *et al.*, 2023). The WHO (2021) emphasized that AI's promise for improving healthcare worldwide can only be achieved by placing ethics and human rights at the center of design, deployment, and use.

3.2. Clinical Implementation and Healthcare Provider Perspectives

AI tools operate within environments where clinical judgment, situated cognition, and systemic biases converge. Chowdhury and Lake (2018) distinguish between explainability (mathematical aspects) and understandability (interpreting AI recommendations), crucial for recognizing bias perpetuation. The impact is evident in risk-stratification algorithms – Vyas, Eisenstein, and Jones (2020) found that race-adjusted algorithms in cardiac surgery could steer Black patients away from life-saving procedures based on poorly understood racial adjustments. The addition of AI risks can repropose a paternalistic model of medical decision-making, where the 'computer knows best,' potentially undermining shared decision-making inside the inter-relations between clinicians and patients (McDougall, 2019).

Van de Sande *et al.* (2022) identify four key intervention domains: healthcare provider education in bias recognition, standardized data quality protocols, evolving regulatory frameworks with mandatory equity audits, and meaningful stakeholder engagement throughout AI development (Dankwa-Mullan *et al*., 2021).

3.2.1. Global Perspectives: AI Bias Challenges in Low and Middle-Income Countries

While much literature on AI bias emerges from high-income countries, unique challenges appear in low – and middle – income countries (LMICs), where AI systems may exacerbate rather than address health disparities. Critical "contextual bias" emerges when AI models trained on high-income country data are deployed in LMIC settings without adequate validation (Alami *et al.*, 2020). More than half of clinical AI datasets originate from the US or China, with almost all top databases affiliated with high-income countries (Larrazabal *et al*., 2021).

Studies across Latin America, Sub-Saharan Africa, and South Asia reveal that AI models trained on high-income country data may introduce substantial bias, leading to poor performance, particularly harmful in resource-limited settings (Schwalbe, Wahl, 2020). Dermatology AI systems trained predominantly on lighter skin tones showed reduced accuracy for darker skin conditions prevalent in African populations (Kamulegeya *et al*., 2019).

Limited local training data creates "data poverty" feedback loops where populations in data-rich regions benefit substantially more from AI healthcare applications, entrenching global health disparities. Cultural and linguistic factors further complicate implementation, as health concepts, symptom descriptions, and care-seeking behaviors vary across contexts, yet most AI systems are developed with limited consideration of these variations.

3.3. Implementation Strategies and Quality Assurance Frameworks

The regulatory frame for AI in healthcare is evolving, with animated debates concerning its classification as a medical device and the associated ethical and legal implications (Pesapane *et al.*, 2018; Gerke, Minssen, Cohen, 2020).

Addressing clinical implementation and global context challenges requires structured frameworks guiding AI system development, validation, and deployment while maintaining an equity focus. Van de Sande *et al.* (2022) propose step-by-step AI development with bias detection at each stage, aligning with the "roadmap for responsible machine learning in healthcare, "emphasizing data preprocessing, model development, and validation stages to prevent harm to vulnerable populations (Wiens *et al.*, 2019).

Quality assurance frameworks serve as core implementation tools. The PROBAST tool (Wolf *et al.*, 2019) offers structured bias risk assessment in prediction model studies, while the CONSORT-AI extension (Liu *et al.*, 2020) provides clinical trial reporting guidelines for AI interventions. These frameworks emphasize transparent reporting of model development processes and thorough bias assessment across population subgroups.

Bellamy et al. (2019) describe the AI Fairness 360 toolkit, offering an extensible framework for detecting and mitigating algorithmic bias through preprocessing techniques for data debiasing and post-processing methods adjusting model outputs for demographic group fairness. The WHO's AI ethics and governance guidance (2021) recommends standardized bias detection approaches, including regular equity audits, diverse development team representation, continuous demographic performance monitoring, and transparent limitation reporting.

Validation processes must address disparities in model performance across racial and ethnic groups (Navarro *et al.*, 2021). Future investigations should combine predominantly European and Anglo-Saxon studies with LMIC research to avoid Western-centric perspectives limiting generalizability and embrace approaches that prevent geographical, socio-economic, and cultural biases.

4. Limitations

Study limitations include the narrative review methodology, which may introduce selection bias in literature inclusion. The predominantly Western-centric literature limits generalizability to global contexts, and the rapid evolution of AI technology may outpace current findings. Success requires ongoing collaboration between technology developers, healthcare providers, policy makers, and affected communities.

Conclusions

This review examined racial bias in medical AI, revealing that technical sophistication does not inherently protect against bias. Effective mitigation requires intervention at multiple levels, from data collection to clinical implementation. Key findings indicate AI systems risk amplifying existing disparities through algorithmic processes. Future research should focus on developing standardized bias detection methods and validation frameworks for diverse populations.

References

Alami H., Lehoux P., Auclair Y., de Guise M., Gagnon M.P., Fortin J.P., et al. (2020). Artificial intelligence in health care: laying the foundation for responsible, sustainable, and inclusive innovation in low- and middle-income countries. *Globalization and Health*, 19: 52. https://doi.org/10.1186/s12992-020-00584-1

Bellamy R.K.E., Dey K., Hind M., Hoffman S.C., Houde S., Kannan K., Lohia P., Martino J., Mehta S., Mojsilović A., Nagar S., Natesan Ramamurthy K., Richards J., Saha D., Sattigeri P., Singh M., Varshney K.R., Zhang Y. (2019). AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research & Development*, 63(4/5): 1-15. https://doi.org/10.1147/JRD.2019.2942287

Benjamin R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Cambridge: Polity Press.

Brown P., van Voorst R. (2024). The influence of artificial intelligence within health-related risk work: a critical framework and lines of empirical inquiry. *Health, Risk & Society*, 26(7-8): 301-316. https://doi.org/10.1080/13698575.2024.2412374

Buolamwini J., Gebru T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81: 77-91.

Burrell J. (2016). How the machine 'thinks': understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1): 1-12. https://doi.org/10.1177/2053951715622512

Cary M.P. Jr., Zink A., Wei S., Olson A., Yan M., Senior R., Bessias S., Gadhoumi K., Jean-Pierre G., Wang D., Ledbetter L.S., Economou-Zavlanos N.J., Obermeyer Z., Pencina M.J. (2023). Mitigating racial and ethnic bias and advancing health equity in clinical algorithms: a scoping review. *Health Affairs (Millwood)*, 42(10): 1359-1368. https://doi.org/10.1377/hlthaff.2023.00553

Charpignon M.L., Byers J., Cabral S., Celi L.A., Fernandes F., Gallifant J., Lough M.E., Mlombwa D., Moukheiber L., Ong B.A., Panitchote A., William W., Wong A.I., Nazer L. (2023). Critical bias in critical care devices. *Critical Care Clinics*, 39(4): 795-813. https://doi.org/10.1016/j.ccc.2023.02.005

Chowdhury R., Lake M. (2018). Is explainability enough? Why we need understandable AI. *Forbes*. https://www.forbes.com/sites/rummanchowdhury/2018/06/04/is-explainability-enough-why-we-need-understandable-ai/?sh=33ed372d62f4

Collins P.H. (2019). *Intersectionality as Critical Social Theory*. Durham: Duke University Press.

Crawford K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

Crenshaw K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, 1989(1): 139-167.

Cross J.L., Choma M.A., Onofrey J.A. (2024). Bias in medical AI: implications for clinical decision-making. *PLOS Digital Health*, 3(11): e0000651. https://doi.org/10.1371/journal.pdig.0000651

Dankwa-Mullan I., Scheufele E.L., Matheny M., Quintana Y., Chapman W., Jackson G., South B.R. (2021). A proposed framework on integrating health equity and racial justice into the artificial intelligence development lifecycle. *Journal of Health Care for the Poor and Underserved*, 32(2): 300-317. https://doi.org/10.1353/hpu.2021.0065

Gerke S., Minssen T., Cohen G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial Intelligence in Healthcare*, 295-336. https://doi.org/10.1016/B978-0-12-818438-7.00012-5

Gianfrancesco M.A., Tamang S., Yazdany J., Schmajuk G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*, 178(11): 1544-1547. https://doi.org/10.1001/jamainternmed.2018.3763

Gigerenzer G., Hoffrage U., Kleinbölting H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological Review*, 98(4): 506-528. https://doi.org/10.1037/0033-295X.98.4.506

Grote T., Berens P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics*, 46(3): 205-211. https://doi.org/10.1136/medethics-2019-105586

Guo W., Caliskan A. (2021). Detecting emergent intersectional biases: contextualized word embeddings contain a distribution of human-like biases. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*: 122-133.

Hawley K. (2015). Trust and distrust between patient and doctor. *Journal of Evaluation in Clinical Practice*, 21(5): 798-801. https://doi.org/10.1111/jep.12374

Howard A. (2023). Real talk: intersectionality and AI. *MIT Sloan Management Review*. https://sloanreview.mit.edu/article/real-talk-intersectionality-and-ai/

Kamulegeya L.H., Okello M., Bwanika J.M., Musinguzi D., Nakibuuka J., Bassajja A., et al. (2019). Using artificial intelligence on dermatology conditions in Uganda: a case for diversity in training data sets for machine learning. *bioRxiv*. https://doi.org/10.1101/826057

Larrazabal A.J., Nieto N., Peterson V., Milone D.H., Ferrante E. (2021). Sources of bias in artificial intelligence that perpetuate healthcare disparities – a global review. *PLOS Digital Health*, 1(1): e0000022. https://doi.org/10.1371/journal.pdig.0000022

Liu X., Cruz Rivera S., Moher D., Calvert M.J., Denniston A.K., SPIRIT-AI and CONSORT-AI Working Group (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine*, 26: 1364-1374. https://doi.org/10.1038/s41591-020-1034-x

McDougall R.J. (2019). Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics*, 45(3): 156-160. https://doi.org/10.1136/medethics-2018-105118

Montavon G., Samek W., Müller K.R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73: 1-15. https://doi.org/10.1016/j.dsp.2017.10.011

Navarro C.L.A., Damen J.A.A., Takada T., Nijman S.W.J., Dhiman P., Ma J., Collins G.S., Bajpai R., Riley R.D., Moons K.G.M., Hooft L. (2021). Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ*, 375(2281): 1-9. https://doi.org/10.1136/bmj.n2281

Nijman S., Leeuwenberg A.M., Beekers I., Verkouter I., Jacobs J., Bots M.L., Asselbergs F.W., Moons K., Debray T. (2022). Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *Journal of Clinical Epidemiology*, 142: 218-229. https://doi.org/10.1016/j.jclinepi.2021.11.023

Noble S.U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Norgeot B., Glicksberg B.S., Butte A.J. (2019). A call for deep-learning healthcare. *Nature Medicine*, 25(1): 14-15. https://doi.org/10.1038/s41591-018-0320-3

Obermeyer Z., Powers B., Vogeli C., Mullainathan S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447-453. https://doi.org/10.1126/science.aax2342

Parasuraman R., Manzey D.H. (2010). Complacency and bias in human use of automation: an attentional integration. *Human Factors*, 52(3): 381-410. https://doi.org/10.1177/0018720810376055

Parikh R.B., Teeple S., Navathe A.S. (2019). Addressing bias in artificial intelligence in health care. *JAMA*, 322(24): 2377-2378. https://doi.org/10.1001/jama.2019.18058

Pesapane F., Volonté C., Codari M., Sardanelli F. (2018). Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. *Insights into Imaging*, 9(5): 745-753. https://doi.org/10.1007/s13244-018-0645-y

Schwalbe N., Wahl B. (2020). Artificial intelligence and the future of global health. *The Lancet*, 395(10236): 1579-1586. https://doi.org/10.1016/S0140-6736(20)30226-9

Topol E.J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1): 44-56. https://doi.org/10.1038/s41591-018-0300-7

Van de Sande D., Van Genderen M.E., Smit J.M., Huiskens J., Visser J.J., Veen R.E.R., van Unen E., Ba O.H., Gommers D., Bommel J.V. (2022). Developing, implementing, and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter. *BMJ Health & Care Informatics*, 29(1): e100495. https://doi.org/10.1136/bmjhci-2021-100495

Vyas D.A., Eisenstein L.G., Jones D.S. (2020). Hidden in plain sight — reconsidering the use of race correction in clinical algorithms. *New England Journal of Medicine*, 383(9): 874-882. https://doi.org/10.1056/NEJMms2004740

Wiens J., Saria S., Sendak M., Ghassemi M., Liu V.X., Doshi-Velez F., Jung K., Heller K., Kale D., Saeed M., Ossorio P.N., Thadaney-Israni S., Goldenberg A. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9): 1337-1340. https://doi.org/10.1038/s41591-019-0548-6

Wolf R.F., Moons K.G.M., Riley R.D., Whiting P.F., Westwood M., Collins G.S., Reitsma J.B., Kleijnen J., Mallett S., PROBAST Group (2019). PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, 170(1): 51-58. https://doi.org/10.7326/M18-1376

World Health Organization (2021). *Ethics and governance of artificial intelligence for health: WHO guidance*. Geneva: World Health Organization. https://apps.who.int/iris/bitstream/handle/10665/341996/9789240029200-eng.pdf