Fairness and Bias in Biomedical AI/ML: Defining Goals and Putting Them Into Practice

Nicole Martinez-Martin
Center for Biomedical Ethics, Stanford University School of Medicine, Stanford, California 94025, USA
Email: nicolemz@stanford.edu

Abdoul Jalil Djiberou Mahamadou
Center for Biomedical Ethics, Stanford University School of Medicine, Stanford, California 94025
Email: abdjiber@stanford.edu

Magdalena Ng
Center for Biomedical Ethics, Stanford University School of Medicine, Stanford, California 94025
Email: madelena@stanford.edu

Mildred K. Cho
Center for Biomedical Ethics, Stanford University School of Medicine, Stanford, California 94025
Email: micho@uchicago.edu

Concerns regarding generalizability and ensuring that artificial intelligence and machine learning (AI/ML) work effectively and accurately across different populations continue to present challenges for the ethical development and deployment of biomedical. Even though bias and fairness have been prioritized as issues for biomedical AI/ML, underlying differences in how researchers conceptualize and operationalize bias and fairness can contribute to difficulties in achieving goals for addressing fairness and mitigating bias. This session of the 2026 Pacific Symposium on Biocomputing offers the opportunity for interdisciplinary discussion and perspectives on addressing fairness in biomedical AI/ML.

Keywords: Ethics; Bias; Fairness; Machine Learning.

1. Introduction

The role of artificial intelligence (AI) in biomedical research and healthcare delivery has grown rapidly over the several years. Even as public and private investment in AI for healthcare has surged, however, issues such as bias and fairness in AI continue to present challenges for realizing the promise of AI to improve healthcare services. Studies over recent years have demonstrated ways that AI tools may systematically exclude some population groups from benefits and exacerbate

existing inequalities and harms in healthcare delivery. Biomedical researchers, clinicians and data scientists have recognized fairness as a central concern for artificial intelligence and machine learning (AI/ML) technologies in medicine.

Broadly speaking, fairness for AI in healthcare aims to ensure that AI models and tools work in ways that are generalizable across individuals or groups – that they work for the populations in which they will be used. Fairness in AI has generally referred to the idea that AI should treat similarly-situated individuals or groups similarly. However, there is disagreement in practice regarding what constitutes fairness – such as whether the goal of fairness should be equality in outcomes or simply proportionate representation of different groups in datasets. AI for healthcare and biomedical research is an interdisciplinary field and different disciplines bring different orientations to interpreting bias and fairness. For example, quantitative fields tend to see fairness in mathematical terms, such as finding technical solutions to adjust algorithms to account for bias. On the other hand, social scientists tend to view fairness through a lens of hierarchy and relational power dynamics, and bioethicists see fairness as based on rights [1]. Thus, fairness may be conceptualized differently by the different actors involved in developing AI/ML for biomedical research and healthcare. These disconnects between different ways of conceptualizing, justifying, and operationalizing fairness, even within the development of the same project, contribute to the difficulties in putting fairness into practice in AI for healthcare and biomedical research.

An interdisciplinary approach that includes computer science, clinical medicine, ethics and social sciences will be valuable for examining different approaches to operationalizing fairness in projects for AI in healthcare. Furthermore, there are political shifts that are impacting funding and practices relevant to fairness in AI/ML projects for healthcare. This session provides an opportunity for researchers to examine practices in AI/ML that support fairness and mitigate bias from an interdisciplinary perspective and discuss best practices in developing AI/ML for healthcare that provides benefit across diverse populations.

2. Session Contributions

2.1. *Bias Detection*

Ansari et al. [2] examine the use of large language models (LLMs) for bias detection. They note that LLMs are being applied to healthcare tasks such as decision support, text summarization, and question-answering, but have shown to demonstrate bias in relation to demographic categories such as race, gender identity, and sexual orientation. LLMs have been applied to audit models for bias, because of their ability to evaluate the large amounts of data. Ansari et al. investigated how model size and prompting techniques affect bias detection with GPT-3.5-turbo, GPT-4o, llama3.3,and o1-mini. They found that the best model for bias detection depends on the metric chosen by the auditor and that smaller models can offer a cost-effective alternative to larger models.

Mottez et al. [3] present a comprehensive framework for bias detection and mitigation that addresses disparities in relation to sex, age, and race in relation to diagnostic tasks with chest X-rays. Even

though deep learning models show promise in improving diagnostic accuracy from chest X-rays, they also may exacerbate healthcare disparities when their performance varies across demographic groups. Mottez et al. found that replacing the final layer of CNN with an eXtreme Gradient Boosting classifier improves the fairness of the subgroup while maintaining or improving the overall predictive performance. They describe a practical and effective path toward equitable deep learning deployment in clinical radiology.

2.2. *Social and Environmental Factors Impacting Bias*

Coggan et al. [4] examine the issue of how subconscious biases of healthcare providers may contribute to persistent demographic disparities identified in the treatment of patients in the emergency department (ED), with the goal of better understanding how subconscious biases influence clinical interpretations and decisions throughout a patient's stay. They conduct a retrospective cross-sectional analysis of 301,116 ED visits to a US pediatric medical center between 2019–2024. After adjusting analyses for confounders, including chief complaint, patient comorbidities, insurance type, socio-economic deprivation, and patient visit history, they trained gradient boosting models to predict admission and inspected feature importances across demographic groups for evidence of learned care disparities. They find significant demographic disparities in hospital admission and conclude that many visit characteristics, clinical and otherwise, may influence the operation of subconscious biases and affect ML-driven decision support tools.

Sun et al. [5] investigate the contextual factors that impact the uptake and adoption of ecological momentary assessment (EMA) and wearable sensors by adults in Hawai'i. The data streams from these types of devices are increasingly used to train AI/ML models for digital phenotyping and predictive intervention, and thus lower adoption rates by marginalized populations raises questions about fairness, bias, and inclusivity in model development. Sun et al. conducted a four-week observational study with adults in Hawai'i, combining continuous Fitbit monitoring and daily EMA surveys in order to identify primary barriers to study participation and adherence. Sun et al. then propose a set of design guidelines aimed at advancing the inclusivity, engagement, and fairness of wearable-based EMA research.

2.3. *Framework for Supporting Fairness in AI Development*

Foti et al. [6] set out a framework for fairness in AI that is adapted from work they originally developed for precision medicine research. The Trustworthy AI Decision Map can anchor and structure dialogue among stakeholders about the ethical implications of specific AI tools. The map identifies key decision points across the AI life cycle that impact fairness and trustworthiness in order to facilitate dialogue among stakeholders. The map that Foti et al. developed is meant to enable teams to anticipate downstream consequences, integrate multiple perspectives, and support institutional accountability.

Acknowledgments

This work is supported by NIH/NHGRI Grant 5R01HG014227.

References

1. Nicole Martinez-Martin and Mildred K. Cho. Bridging the AI Chasm: Can EBM Address Representation and Fairness in Clinical Machine Learning? *The American Journal of Bioethics: AJOB*, *22*(5), 30–32 (2022).
2. Zara N. Ansari, Aaron Fanous, Jesutofunmi A. Omiye, Ank Agarwal and Roxana Daneshjou.. Using Large Language Models to Audit Model Healthcare Biases. *Pacific Symposium on Biocomputing* (2026).
3. Helena Coggan, Anne Bischops, Pradip Chaudhari, Yuval Barak-Corren, Andrew M. Fine, Ben Y. Reis, Jaya Aysola and William G. La Cava. Deciphering the influence of demographic factors on the treatment of pediatric patients in the emergency department. *Pacific Symposium on Biocomputing* (2026).
4. Nicole Foti, Janet Shim, Caitlin McMahon and Sandra Soo-Jin Lee. Building Fair and Trustworthy Biomedical AI: A Tool for Identifying Key Decision Points. *Pacific Symposium on Biocomputing* (2026).
5. Clemence Mottez, Louisa Fay, Maya Varma, Sophie Ostmeier and Curtis Langlotz. From Detection to Mitigation: Addressing Bias in Deep Learning Models for Chest X-Ray Diagnosis. *Pacific Symposium on Biocomputing* (2026).
6. Aditi Jaiswal, Ali Kargarandehkordi, Christopher Slade, Roberto M. Benzo, Kristina T. Phillips and Peter Washington. Barriers to Designing Inclusive Ecological Momentary Assessment and Wearable Data Collection Protocols for AI-Driven Substance Use Monitoring in Hawai‘i. *Pacific Symposium on Biocomputing* (2026).