A Concise Framework for Fairness: Navigating Disparate Impact in Healthcare AI

Pemla Jagtiani, Mert Karabacak, Konstantinos Margetis

Submitted to: JMIR AI
on: June 20, 2024

Disclaimer: This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

A Concise Framework for Fairness: Navigating Disparate Impact in Healthcare AI

Pemla Jagtiani[1*] BS; Mert Karabacak[2*] MD; Konstantinos Margetis[1] MD

[1]Department of Neurosurgery, Mount Sinai Health System New York City US
[*]these authors contributed equally

Corresponding Author:
Konstantinos Margetis MD

Abstract

As artificial intelligence (AI) increasingly permeates healthcare, it promises to enhance patient outcomes and operational efficiency. However, the integration of AI also introduces significant risks of perpetuating biases, necessitating careful consideration of fairness in these systems. This paper proposes a comprehensive framework aimed at mitigating biases and promoting fairness within healthcare AI. By outlining a structured approach encompassing all stages of the AI lifecycle—from data collection and preprocessing to model selection and continuous monitoring—we provide actionable guidance for developers, researchers, and healthcare professionals. Furthermore, we introduce specific fairness metrics such as False Positive/Group Size Parity and False Discovery Rate Parity, which are crucial for evaluating AI systems in various healthcare applications—from diagnostic tools to resource allocation.

(JMIR Preprints 20/06/2024:63465)
DOI: https://doi.org/10.2196/preprints.63465

Original Manuscript

Title: A Concise Framework for Fairness: Navigating Disparate Impact in Healthcare AI

Authors: Pemla Jagtiani, BS[a,*], Mert, Karabacak, MD[b,*]; Konstantinos Margetis, MD, PhD[b]

[*]*Equally contributing first authors*

Affiliations: [a]School of Medicine, SUNY Downstate Health Sciences University; [b]Department of Neurosurgery, Mount Sinai Health System, New York, NY, United States of America;

Corresponding author's name and current institution: Konstantinos, Margetis, MD, PhD; Department of Neurosurgery, Mount Sinai Health System, New York, NY, United States of America

Corresponding author's email: Konstantinos.Margetis@mountsinai.org

Abstract

As artificial intelligence (AI) increasingly permeates healthcare, it promises to enhance patient outcomes and operational efficiency. However, the integration of AI also introduces significant risks of perpetuating biases, necessitating careful consideration of fairness in these systems. This paper proposes a comprehensive framework aimed at mitigating biases and promoting fairness within healthcare AI. By outlining a structured approach encompassing all stages of the AI lifecycle—from data collection and preprocessing to model selection and continuous monitoring—we provide actionable guidance for developers, researchers, and healthcare professionals. Furthermore, we introduce specific fairness metrics such as False Positive/Group Size Parity and False Discovery Rate Parity, which are crucial for evaluating AI systems in various healthcare applications—from diagnostic tools to resource allocation.

Introduction

The integration of artificial intelligence (AI) in healthcare has the potential to transform patient care. However, concerns about fairness, transparency, and potential biases have been raised.[1] Bias, a systematic error in decision-making processes, can stem from various sources in an AI framework, including sample bias, outcome bias, and biases introduced during data handling and feature engineering stages.[2] If an AI system is trained on biased data, it may make biased decisions, leading to suboptimal care and disparate impact on certain protected groups.[3,4] As AI plays an increasingly significant role in healthcare decision-making, prioritizing fairness and mitigating biases is crucial.

A Concise Framework for Ensuring Fairness in Healthcare AI

A comprehensive and systematic approach to model development and deployment is crucial to effectively address potential biases and ensure fairness in healthcare AI systems. Here, we propose a concise approach encompassing all stages of the AI lifecycle, from data collection and preprocessing to model selection, evaluation, and ongoing monitoring. ***Figure 1*** presents a step-by-step framework that can guide researchers, developers, and healthcare professionals in their efforts to create fair and equitable AI models.[4,5]

1. ***Data Representation:*** Assess the representativeness of the training data. Underrepresentation of certain populations can lead to sample bias and unreliable predictions for those groups. Ensure the data accurately represents the diverse population the AI system will serve by collecting and incorporating data from underrepresented groups if possible.

2. ***Outcome Labels:*** Inaccuracies in the outcome labels used for training can introduce label bias, which can arise from misdefined outcomes or measurement errors.[5] Label bias can lead to systematic errors that disproportionately affect certain groups. Outcome bias occurs when labels reflect underlying systemic biases present in the data. Carefully review and validate the outcome labels to ensure they are unbiased, and redefine or correct them if needed to minimize bias.

3. ***Feature and Transformation Bias:*** Mitigate biases during feature engineering by carefully handling missing values and combining categories. Consider the methods used for collecting sensitive attributes, such as race and ethnicity indicators. The accuracy and potential errors in these attributes can vary depending on whether they are self-reported, recorded by a third party, or inferred from other data.[5] Addressing these issues is crucial for reducing biases introduced during data collection and preprocessing.

4. ***Model Selection:*** Evaluate potential models based on both accuracy and fairness metrics that are relevant to the model's intended use and the context of its application. Select models that balance performance and fairness, considering the potential impact on different patient populations and the specific goals of the AI system.[4]

5. ***Fairness Metrics:*** Define fairness for the specific context and select relevant metrics,

such as False Positive/Group Size (FP/GS) Parity, False Discovery Rate (FDR) Parity, False Positive Rate (FPR) Parity, Recall Parity, False Negative/Group Size (FN/GS) Parity, False Omission Rate (FOR) Parity, and False Negative Rate (FNR) Parity.[5] Incorporate fairness in model selection through various approaches:

a. *Performance versus Equitability:* Compare the equity of models with similar overall performance. Evaluate the trade-offs between performance and fairness, such as balancing FDR Parity and Recall Parity, to explicitly show these trade-offs.

b. *Subgroup Performance:* Consider models that perform best for specific subgroups (e.g., different races or sexes) and those that perform consistently across groups.

c. *Equity Penalty:* Develop a model selection parameter that penalizes performance based on deviation from equity, aggregating equity across multiple groups. This approach provides options for the final model, balancing overall performance and equity measures. Unlike traditional model selection, which is based solely on performance metrics, the final choice involves a judgment call reflecting the dual goals of accuracy and equity.[5]

6. ***Evaluate and Iterate:*** Assess the selected models using the chosen fairness metrics. Analyze the trade-offs between performance and fairness, and iteratively refine the models to improve both aspects simultaneously.

7. ***Deployment and Monitoring:*** After deploying the model, regularly monitor it for data drift, which refers to changes in data distribution over time that can degrade model performance. Data drift can occur due to various factors, such as changes in population demographics, genetic variation, imaging techniques, disease prevalence, and social determinants of health.[4] Continuously assess the model's predictions against new data to ensure accuracy and reliability. If data drift is detected, retrain the model with updated data to maintain performance and fairness.[4]

How to Define Fairness for Specific Contexts

Defining fairness in healthcare AI applications involves understanding the project's goals and the potential impacts of the model's decisions on different subgroups. The choice of fairness metrics often depends on the type of intervention and the intended use. When selecting fairness metrics, it is essential to consider the potential harms associated with different types of errors and how these harms may disproportionately affect certain subgroups. When choosing fairness metrics, the following guidance based on the context could be considered:

- For screening tools, it may be appropriate to prioritize metrics that focus on minimizing false negatives, such as FNR Parity and FN/GS Parity. This is because false negatives in screening can lead to delayed diagnosis and treatment, which may have serious consequences for patient outcomes.

- For diagnostic tools, it may be beneficial to balance metrics that consider both false positives and false negatives, such as FPR Parity, FNR Parity, and Recall Parity. This approach ensures that the tool is accurate in identifying the condition while minimizing the harms associated with both types of errors.
- For prognostic tools, it may be suitable to consider metrics that focus on the reliability of positive predictions, such as FDR Parity and Recall Parity. This is because false positives in prognostic tools can lead to unnecessary interventions or distress for patients, while false negatives may result in missed opportunities for early intervention.
- For resource allocation, it may be appropriate to consider metrics that account for group sizes, such as FP/GS Parity and FN/GS Parity. This ensures that the chances of being incorrectly included or excluded from receiving resources are similar across subgroups, promoting equitable access to care.
- For patient education, it may be beneficial to prioritize metrics that minimize false positives and false negatives, such as FPR Parity, FOR Parity, and Recall Parity. This approach ensures that patients receive accurate information about their conditions and the likelihood of different outcomes, empowering them to make informed decisions about their health.

Achieving parity on all fairness metrics simultaneously may not always be possible, especially when the prevalence of the condition differs across subgroups.[5] In such cases, stakeholders should prioritize the most relevant fairness metrics based on the specific context and the potential impact on different subgroups. By considering appropriate fairness metrics and iteratively developing, evaluating, and refining models, healthcare organizations can create AI systems that promote equity and fairness.

Conclusion

The integration of AI in healthcare has the potential to revolutionize patient care, but it must be approached with a commitment to fairness and equity. By adopting a comprehensive framework for bias mitigation and fairness promotion, healthcare organizations and AI developers can create AI systems that enhance care quality while avoiding the perpetuation of disparities. To fully realize AI's potential in healthcare, organizations must engage with patients and stakeholders, understand their needs and concerns, and prioritize fairness as a core value.

References

1. Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care — Addressing Ethical Challenges. N Engl J Med 2018;378(11):981–3.

2. Ferrara E. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. Sci 2023;6(1):3.

3. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. JAMA Intern Med 2018;178(11):1544.

4. Chen RJ, Wang JJ, Williamson DFK, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat Biomed Eng 2023;7(6):719–42.

5. Rodolfa K, Saleiro P, Ghani R. Chapter 11 Bias and Fairness. In: Big data and social science: data science methods and tools for research and practice. Boca Raton, FL: CRC Press; 2021.

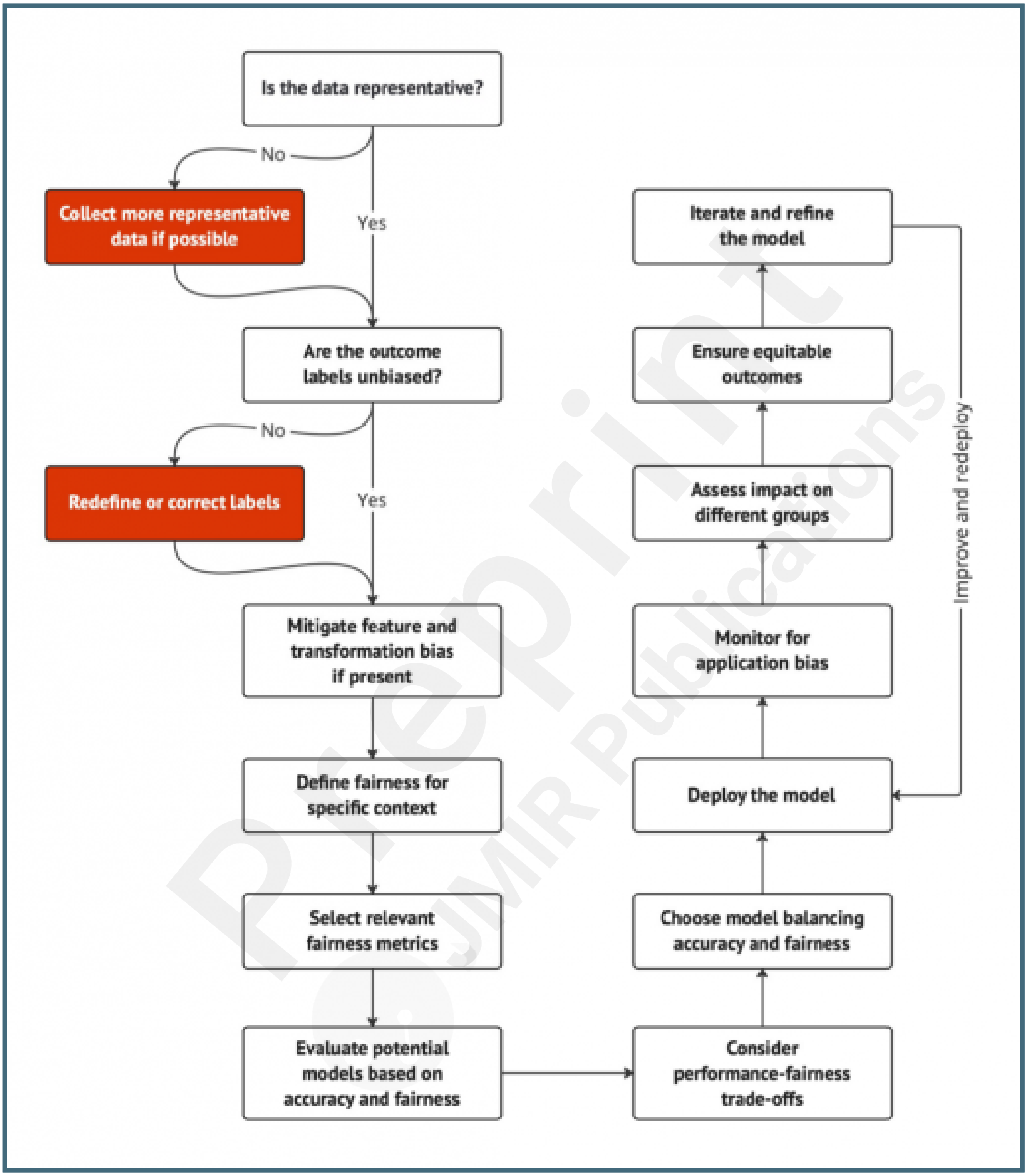
Figure Legend

Figure 1. Flowchart of the proposed framework for ensuring fairness in healthcare AI.

Supplementary Files

Figures

Flowchart of the proposed framework for ensuring fairness in healthcare AI.


Is the data representative?
No
Yes
Collect more representative data if possible
Are the outcome labels unbiased?
No
Yes
Redefine or correct labels
Mitigate feature and transformation bias if present
Define fairness for specific context
Select relevant fairness metrics
Evaluate potential models based on accuracy and fairness
Consider performance-fairness trade-offs
Choose model balancing accuracy and fairness
Deploy the model
Monitor for application bias
Assess impact on different groups
Ensure equitable outcomes
Iterate and refine the model
Improve and redeploy