Bias Prevention in Global Adoption of Artificial Intelligence Based Health Management Systems

Aditya Sood[1] and Ayush Sangari[2]

1. Emory University School of Medicine, 2. Renaissance School of Medicine at Stony Brook University

INTRODUCTION

Artificial intelligence is now capable of being used as a screening tool for a variety of purposes: distinguishing between benign nevi and melanoma, assessing diabetic and pressure ulcers, improving psoriasis and other inflammatory skin disease classifications, and more. Dermatology's embrace of artificial intelligence is good for both practitioners, as well as for patients. Our clinics have a mean wait time of 56 days. Accurate screening tools will inevitably allow for us to be able to better manage our patients, as well as see more of them. This technology offers massive benefits globally, but narrow artificial intelligence models pose massive bias when used on patients dissimilar to those which the datasets were created.

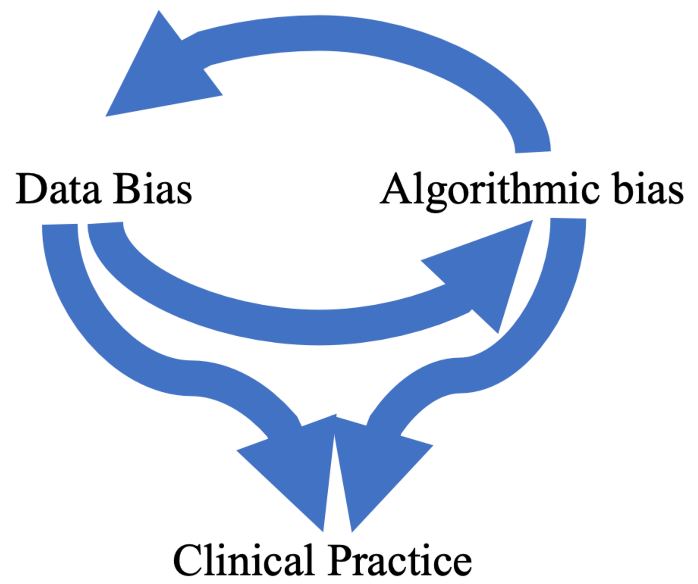

Figure 1. Bias Feedback Loop

OBJECTIVE

Objective: This study analyses current medical policy practices with regards to the bias prevention in artificial intelligence systems to mitigate unforeseen bias in global adoption of AI based screening technologies.

METHODS

1. This study began by surveying existing regulations and guidelines for the use of artificial intelligence in clinical care
2. These existing architecture came from prominent regulatory agencies, non-profits, and think tanks
3. Existing regulation and guidelines were collaged, duplicates removed, and all entries analyzed
4. Each guideline was then rated in terms of its ability to address bias prevention when applying artificial intelligence systems in diverse patient populations
5. Specific questions were then developed that addressed each important factor within machine learning in the space of global health and key metrics useful for evaluating these algorithms
6. Highly rated guidelines were then collated, condensed, and further researched.
7. Finally, 5 key questions and 5 guidelines based off the prior questions were developed

RESULTS

Key Question 1: How was the information presented to explain the model obtained?	This question focuses on the reliability of data inputted into the algorithm.
Key Question 2: How does each inputted attribute/feature contribute to the final prediction/output?	This question centers on the interpretability of algorithm outputs.
Key Question 3: What are some instances that the model considers to be similar to a specific instance?	This questions examines the internal validity of algorithm outputs.
Key Question 4: What are the defining characteristics of each identified output class?	This questions studies the external validity & consistency of algorithm outputs.
Key Question 5: How can a provided instance be modified so that the model output is a certain class?	This question tests the ecological validity of algorithm outputs.

Proposed Guideline 1: Contextual information must be taken into consideration whenever implementing recommendations made by an artificial intelligence system.	In the case of clinical practice, AI-generated predictions, diagnoses, and treatment plans must be analyzed with regards to the clinical correlates of each patient.
Proposed Guideline 2: Measures in model selection and fitting must be taken to ensure interpretability over explanatory power in all models in public and global health.	Techniques to ensure model overfitting does not occur, e.g. K-fold cross validation, shall be mandated.
Proposed Guideline 3: Model outputs for machine learning algorithms in global health should propose multiple, consistent courses of actions to ensure internal validity before adoption by public health agencies.	Mandating consistency in multiple courses of action in a given public health scenario limits sporadic behavior after implementation.
Proposed Guideline 4: An interdisciplinary, multinational panel of global health experts shall verify key characteristics of identified output classes insofar as applies within non-discriminatory supervised AI algorithms.	Verification with global health experts will ensure external validity and congruence with existing global health strategies.
Proposed Guideline 5: Algorithms implemented within the context space of global health must be trained on generalizable data while maintaining both external validity and ecological validity.	Ensuring ecological validity within the dynamic space of global health is required with the inherent fluidity of the global health emergencies.

DISCUSSION

Best practices to avoid bias in global adoption of artificial intelligence-based health systems involve a focus on two topics: dataset expansion and model explainability.

AI programs are only as good as the dataset they are trained on. For this reason, the training dataset size and variety matter. Those currently being used to train programs are often vastly over-represented by type I-II skin patients. One of the largest open-source archives of skin images - the International Skin Imaging Collaboration: Melanoma Project – is predominantly composed of fair skinned patients.

AI systems based on homogenous datasets do poorly on non-matching data, leading to poor outcomes and increased mortality rates.

This study also concluded that an integral aspect of bias prevention in adopting artificial intelligence systems globally is ensuring that the results of artificial intelligence systems are explainable and interpretable, allowing them to catch instances where bias may exist. Current enforcement of AI explainability depends on either satisfying a set of mathematical constraints or satisfying a general yet vague set of criteria. This study proposes that such systems be able to answer a baseline set of five fundamental questions in an accessible manner.

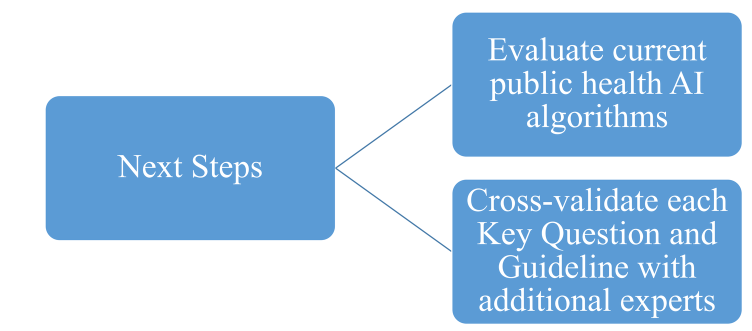

Figure 2. Additional steps to further implement bias prevention techniques in Global Health

REFERENCES

1. Panch T, Szolovitz P, Atun R. Artificial intelligence machine learning and health systems. J Glob Health. 2018;8:020303. 10.7189/jogh.08.020303
2. Johnson AE, Pollard TJ, Mark RG. 2017, November. Reproducibility in critical care: a mortality prediction case study. Machine Learning for Healthcare Conference 2017. JMLR W&C Track Volume 68. Available: http://proceedings.mlr.press/v68/johnson17a/johnson17a.pdf. Accessed: 20 March 2018.
3. Fraser H, Coiera E, Wong D. Safety of patient-facing digital symptom checkers. Lancet. 2018;392:2263-4. 10.1016/S0140-6736(18)32819-8
4. Badawi O, Brennan T, Celi LA, Feng M, Ghassemi M, Ippolito A, et al. Making big data useful for health care: a summary of the inaugural mit critical data conference. JMIR Med Inform. 2014;2:e22. doi: 10.2196/medinform.3447

ACKNOWLEDGEMENTS

We would like to thank the Emory University School of Medicine and Renaissance School of Medicine at Stony Brook University for their invaluable support.