Ensuring Equity in the Development of Health AI Algorithms

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Miranda Khoury
Spring 2024
Department of Biomedical Engineering

Technical Team Members:
None

On my honor as a UVA student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS
Dr. Joshua Earle, Department of Engineering and Society
Dr. Timothy Allen, Department of Biomedical Engineering
Dr. Kate Owen, AMPEL BioSolutions LLC

Introduction: What Is Health AI, and Why is Equitable Health AI Necessary?

The burgeoning field of bioinformatics promises to lead healthcare into an age of seemingly endless possibility, especially when incorporated with the powerful developing techniques of artificial intelligence and its subdiscipline machine learning. Biotechnology companies and hospitals have improved diagnosis and treatment of diseases via predictive bioinformatics tools that help clinicians understand a patient's disease and devise individualized treatment strategies (Hubbard et al., 2023; Panesar, 2023). At the same time, the immense power of artificial intelligence could bring about devastating impacts if it is not carefully deployed (Ghassemi & Nsoesie, 2022).

The technologies I discuss in this paper are artificial intelligence (AI) algorithms for predicting and qualifying disease, specifically algorithms that use machine learning (ML) techniques. AI is defined as the ability of computers to replicate human intelligence; when researchers try to create a machine with artificial intelligence, they want the machine to be able to make complex decisions and think the way a human can (Gil de Zúñiga et al., 2024). Machine learning, a subset of AI, is the ability for a computer to learn over time (Wang et al., 2009). ML algorithms can provide great utility to healthcare providers, but they also spark concerns about equity and algorithmic fairness (Khan et al., 2023). For instance, a study published in Nature determined that an ML system for evaluating breast cancer performed better than clinicians and reduced workload by 88%, demonstrating the positive possibilities of such tools (McKinney et al., 2020). On the other hand, a 2020 study revealed that an algorithm used for deciding whether to schedule over suspected no-show appointments was making access to care more difficult for people with mobility issues and lower socioeconomic status, demographics that already struggle with getting adequate care (Murray et al., 2020). This is only one of many case studies highlighting the negative potential of medical ML algorithms to make life worse for already marginalized populations if not implemented with care.

Zink and collaborators define health equity as "the state in which everyone has the opportunity to attain their best possible health" (Zink et al., 2023). From the above examples, it is clear that ML algorithms developed for healthcare — from this point onward described as "health AI" — have not regularly been achieving health equity. Researchers developing health AI algorithms, such as myself, must study how and why health AI algorithms become inequitable so we can prevent making the same mistakes in our own work. To seek these much-needed answers, my STS thesis investigates the following primary research question: how do we ensure that healthcare ML algorithms serve all kinds of people equitably, minimizing rather than exacerbating disparity in healthcare?

In the rest of this paper, I first describe my methodology in answering my research question. Then, I give an equity analysis of three case studies of health AI implementations and discuss what these health AI models did well and poorly. These case studies include the Impact Pro health AI for recommending additional care, the Sepsis Watch health AI for diagnosing sepsis, and a health AI created by researchers at the University of Florida for diagnosing bacterial vaginosis. Finally, I consolidate these findings and add my own insights to provide a set of guidelines for researchers to ensure equitable health AI moving forward.

Methods

Current reviews on equity in health AI offer limited utility in that many remain very broad, presenting general ethical principles to uphold as opposed to specific guidelines for researchers to follow (for instance: Khan et al., 2023; Rigby, 2019; Siala & Wang, 2022; Smith et al., 2020). To address this gap in the discussion, I analyzed case studies, reasoning that by investigating individual scenarios, I would be able to gain specific insights that the discussion is currently lacking. The specifics identified for a single case study may not be relevant to all cases of health AI; however, by comparing common themes across multiple case studies, I was able to generate a set of guidelines that are more detailed than the general moral precepts that make up the majority of the discussion around equitable health AI (fairness, accountability, etc.) but that are still broad enough to apply to multiple kinds of health AI implementations.

I chose case studies based on several criteria. First, information had to be available as to the positive or negative equitable effects of the given health AI implementation. Second, the model had to have been employed on actual patient data as opposed to being used purely in simulation. I also decided to include a health AI known as Impact Pro among my cases because the landmark paper dissecting its failure by Obermeyer and colleagues was instrumental in engendering more research into equity in health AI (Hickok et al., 2022), and I felt I would be remiss not to discuss such a historic case in a paper dedicated to the topic of health AI equity.

Upon identifying three case studies that met the above criteria, I employed the STS framework of social construction of technology (SCOT) to analyze each case. The SCOT framework holds that social groups impact the evolution and acceptance of a technology; various social groups have different goals for a technology, and the social groups whose views dominate ultimately shape the end result of the technology (Pinch & Bijker, 1984). Many groups stand to benefit from the adoption of health AI, and as such, there are several competing interests and goals for the technology. Due to these competing interests, SCOT is an ideal framework to use to investigate the trajectory of health AI.

To conduct SCOT analysis, I needed to identify stakeholder groups most relevant to my research question. To do this, I considered three main ways people may interact with health AI: who will use health AI? Who will health AI be used on behalf of? And finally, who designs health AI? Based on these questions, I have identified the social groups of clinicians, patients, and AI developers as populations for further study. It is important to note that I have excluded two stakeholder groups — academics and researchers who use health AI as well as governmental and regulatory bodies — due to practical limitations on my project scope. Using SCOT to outline the motives of these interest groups and their respective roles in specific cases allowed me to understand why certain implementations of health AI failed and some succeeded. In the following sections, I present the results of my SCOT analysis on the three case studies I selected.

Case 1: Impact Pro Model for Additional Care Prioritization

Impact Pro is a health AI model developed by healthcare information technology company Optum. Optum markets Impact Pro to hospitals as a tool for determining which patients should be recommended for intensive care (Optum, n.d.). Impact Pro faced controversy in 2019 when a landmark *Science* journal article reported that the researchers had uncovered algorithmic bias against Black patients (Obermeyer et al., 2019).

Impact Pro used healthcare spending as a proxy for how sick a person was, with the algorithm assuming that more healthcare spending meant a patient was more sick and should more likely be recommended for additional care. It is not unreasonable to use this as a proxy, but it overlooked a large issue: the Black population in the U.S. has historically had less access to healthcare because of discrimination from healthcare professionals and having more difficulty affording care than their white counterparts. As a result, Black patients do not spend as much on healthcare as equivalently sick white patients (Sargent, 2021). It is important to note that the developers of Impact Pro gave a good faith effort to develop an equitable algorithm — indeed, they excluded race as a feature of the algorithm — but since healthcare spending is correlated to race, and healthcare spending was the main feature used to determine whether patients needed additional care, the algorithm unintentionally took race into consideration despite the efforts of the developers.

The general consensus among those in the health AI field is that some of the algorithmic bias against Black patients could have been avoided had the proxy variable for patient sickness been selected more carefully (Hickok et al., 2022; Obermeyer et al., 2019; Sargent, 2021). For instance, avoidable healthcare costs or number of comorbidities might have picked up on legitimate differences in patients' overall health without stratifying by patient wealth. In hindsight, it seems obvious that spending is not as good an indicator of health as a biological marker like blood pressure or number of adverse health events like heart attacks, so why did the developers pick spending?

The goal of Impact Pro was, in part, to reduce costs for the healthcare systems using the algorithm (Sargent, 2021). Optum wants to entice healthcare systems to buy its software, and advertising that Impact Pro will reduce costs for hospitals will bring in customers for Optum. With the goal of reducing costs in mind, it becomes more clear why spending would be used as the key predictive variable; the developers may have thought it would act as a proxy that could both predict patient sickness (incorrectly) *and* help mitigate costs for their healthcare provider customers, and this may have led them to select spending over a different proxy (e.g. number of comorbidities) that would have been more equitable. Through a SCOT lens, the social group of AI developers (Optum) and healthcare providers (hospitals) had different priorities than the social group of patients in that they wanted to reduce costs as well as provide good care, whereas the priority of patients is simply to receive good care. Since Optum was the maker of Impact Pro, Optum's priorities triumphed over the priorities of the patient social group, resulting in the selection of healthcare spending as the primary predictive variable. This in turn incorporated bias into the algorithm.

Some approaches the developers could have taken that likely would have identified the algorithmic bias before Impact Pro was used for decision making processes are, first, to reference social science literature and discuss with the other stakeholder groups about the proxy variable selected before starting to build the health AI model. Hearing the perspectives of patients and healthcare providers and reviewing the literature could have illuminated unexpected relationships between the proxy variable and demographics variables, identifying poor choices of proxies before beginning work on the model. Second, running the algorithm on patient data during a test period in which it is not being used for decision making and analyzing the results could have revealed the systemic bias before it was deployed.

Case 2: Sepsis Watch Model for Sepsis Diagnosis

Sepsis Watch is a health AI model created at Duke University Hospital to detect and manage sepsis in patients. Health system leaders' goals were twofold: first, to detect sepsis, and second, to ensure that interventions for sepsis were completed successfully for patients diagnosed with it. After hearing about the Impact Pro case, the developers of Sepsis Watch used physiological markers like lab tests and vital signs as model features instead of financial data. They also tested for algorithmic fairness after every tweak to the model (Levi & Gorenstein, 2023). The developers hosted meetings with another social group, healthcare providers, to foster communication and train the healthcare providers about how to work with and interpret the algorithm (Sendak et al., 2020).

To the developers' view, Sepsis Watch was primed to perform equitably — but a colleague at the hospital pointed out a potential issue with using Sepsis Watch in Duke University Hospital's pediatrics unit. Hispanic children historically are not treated for sepsis as quickly as children of other ethnicities, perhaps due to the need for interpreters or doctors not taking Hispanic children's illness as seriously as other children's (Moorthy et al., 2023). As a result, the colleague suggested that it was possible that Sepsis Watch could learn to become biased against Hispanic children, not recommending sepsis procedures for them fast enough. In response, the developers spent 8 weeks testing Sepsis Watch's performance in the pediatric unit and ultimately concluded that it was not biased against Hispanic children.

From a SCOT perspective, the AI developer social group met with members of the clinician social group to decide on priorities together. While they did not reach out to members of the patient social group, they prioritized equity and accuracy in the algorithm, which aligns with patients' priorities (patients want to receive the best care and to be treated fairly). Because all three social groups were aligned on their priorities, the technology that developed to meet these priorities satisfied all three social groups and the technology stabilized into the successful Sepsis Watch algorithm still being used today.

What guidelines do we take away from this more positive implementation of health AI? First, focusing on physiological data like vital signs as opposed to financial or demographic information can prevent systemic biases from being baked into the algorithms. Second, ample communication with other social groups can align priorities before work begins on the model. Additionally, seeking out feedback from experts in diverse fields can lead to identifying previously unnoticed trends that may cause bias and therefore need to be accounted for. And finally, researchers have a responsibility to periodically test their algorithms for bias and respond to concerns brought up by other social groups.

Case 3: University of Florida Model for Bacterial Vaginosis Diagnosis

Researchers at the University of Florida developed a series of health AI models for diagnosing bacterial vaginosis in women, aiming to see if by tuning the way the models were created, they could create a diagnosis model without algorithmic bias. While testing the accuracy of their models, they found that all four of the model types they created performed with the worst accuracy and precision for Asian women, while it performed best for white women and middling for Hispanic and Black women. These initial one-size-fits-all models were trained on all the data points, but the researchers also experimented with other models that were only trained on data from one ethnicity to see if this would improve the equity of their models. Training models on subsets of ethnicity had varied effects, sometimes decreasing the model accuracy and/or precision compared to the one-size-fits-all-model and sometimes improving it. Interestingly, the accuracy of diagnosis for the Asian population improved when the Asian population data was run through the model designed for the Black population. The developers hypothesize that having a small number of positive cases in Asian women of bacterial vaginosis may have been the reason for the model's poor performance on this subset, as having an uneven split between positive and negative data points makes training algorithms more difficult (Celeste et al., 2023).

From a SCOT perspective, the developers initially defaulted to a one-size-fits-all model, prioritizing developer and clinician goals of ease of use over the goal of accuracy, which is the top priority of patients. Upon observing the inequity in their AI, the developers shifted priorities to accuracy. To accomplish this, they made multiple models, making ease of use worse but only improving accuracy in some cases. Thus, the technology failed to achieve either the goal of developers and clinicians or of patients and could not completely satisfy any social group. As a result, the technology did not stabilize as SCOT predicts it will when one social group wins out over the others; the developers acknowledged that the technology needs to be adjusted more before use on actual patients.

I consider this implementation of health AI moderately successful. The developers were successful in that they identified inequities before potentially incorporating the model into actual healthcare practice, but they were unsuccessful in that they were not able to completely remove inequity; therefore I would not consider it advisable to implement the model in hospitals. There are several guidelines we can take away from this case study. First, these developers' results show that one-size-fits-all models may lead to inequitable model performance, so developers should experiment with multiple model types and applications and test each one for algorithmic bias to find the best configuration. Second, this case study identified the susceptibility of health AI models to poor performance when the dataset is unbalanced; that is, when there are a disproportionate number of disease-negative or positive data points in the dataset. Researchers should strive to balance their datasets, either by randomly selecting data points from overrepresented groups or by collecting more data.

Looking Forward: Practical Guidelines for Ensuring Health AI Equity

The case studies discussed previously illuminate that despite good faith efforts on the part of AI developers and healthcare providers to take steps toward fairness in health AI, inequities can still pervade the technology. From analyzing these case studies and reviewing the literature for additional insights, I have developed the following list of guidelines that researchers should uphold when implementing health AI (Table 1).

One trend that could be easily rectified is that even in the successful implementations of health AI, the AI developers did not reach out to members of the patient social group, instead focusing primarily on hearing feedback from members of the clinician social group. I suggest that AI developers gain feedback from patients as well as clinicians to make sure all three stakeholder groups are aligned in their priorities.

Another point I would like to highlight is that of transparency. Transparency is key to ensuring equity in health AI. It is important to make one's AI development process public so social groups can make informed decisions about whether to employ the AI or not. Additionally, frank discussions about the successes and failures of one's health AI can greatly aid other developers. Indeed, I would not have been able to create my list of guidelines without the openness of the researchers whose work inspired this paper.

Table 1. Guidelines for Developing Equitable Health AI

Developer Guideline	**Reference**
Inform users of limits of AI algorithm's generalizability	(Murphy et al., 2021)
Require clinician input instead of decision making being entirely automated	(Murphy et al., 2021)
Prioritize the use of biological variables (labs, vitals, etc.) over claims or monetary variables as model features	(Zink et al., 2023)
If accuracy bias within certain demographics observed, experiment with adding more biological features	(Williams et al., 2021)
Test for demographic trends in AI performance before deploying	Mine
If inequity cannot be removed from AI, consider not deploying it if the AI is not sorely needed	Mine
Reference literature and field experts while selecting model features	Mine

to uncover potential relationships between features and demographics that could lead to bias	
Seek feedback from both clinicians and patients during AI development	(*Ethics & Governance of Artificial Intelligence for Health: WHO Guidance*, 2021)
Monitor AI after deployment and respond immediately to any inequity concerns brought up	Mine
When possible, use a balanced dataset for training the AI	(Gurevich et al., 2023)
Be transparent about how AI was developed, its performance, etc.	Mine

These guidelines will help make health AI more equitable, but they may not be able to ensure equity completely. This is due to the fact that the bias and inequity in health AI stem from systemic, institutional bias in our society. Long-term restriction of healthcare services to those of higher socioeconomic status and discrimination by doctors against minority populations have created a set of healthcare data and practices that are biased in favor of straight, white, male, and upper class people. Therefore, inequity cannot entirely be circumvented by even the most careful of design considerations until our society has fundamentally changed (Nadis, 2022). I now propose some call-to-action items to move healthcare to a more equitable place.

First, like all humans, healthcare providers carry explicit and implicit biases that contribute to bias in the healthcare system as a whole. We can reduce clinician bias by encouraging diversity in the healthcare space — such as by providing more scholarships and professional organizations for clinicians of color and queer clinicians — and increasing awareness in clinicians of their susceptibility to bias. Second, we must improve access to healthcare for poor and marginalized people. This will help counterbalance the historical lack of data on these populations which can cause algorithms to pick up bias. To do this, healthcare systems can invest in free shuttle systems to doctors' offices and hospitals and develop mobile clinics that operate in neighborhoods outside of normal working hours. Additionally, I argue for more diversity in clinical trials, as treatments being tested on a wider array of patients not only makes trial results more generalizable but improves patient trust in the medical system (Schwartz et al., 2023).

Equity remains a key concern in the development and use of health AI. We may never be able to ensure perfect equity in health AI, but we have a moral imperative to try. The discussion around health AI equity is ongoing and complex, and I have only scratched the surface of the proposed issues and solutions to this problem. Nonetheless, with the specific guidelines I have outlined in this paper and the wider healthcare changes I have suggested, it is my hope that readers will take away from this paper a framework for how to proceed ethically when developing their own health AI — and moreover, that they will be inspired and empowered to advocate for health equity in our society.

References

Celeste, C., Ming, D., Broce, J., Ojo, D. P., Drobina, E., Louis-Jacques, A. F., Gilbert, J. E., Fang, R., & Parker, I. K. (2023). Ethnic disparity in diagnosing asymptomatic bacterial vaginosis using machine learning. *Npj Digital Medicine*, *6*(1), Article 1. https://doi.org/10.1038/s41746-023-00953-1

Ethics & Governance of Artificial Intelligence for Health: WHO guidance. (2021). World Health Organization. https://iris.who.int/bitstream/handle/10665/341996/9789240029200-eng.pdf

Ghassemi, M., & Nsoesie, E. O. (2022). In medicine, how do we machine learn anything real? *Patterns*, *3*(1). https://doi.org/10.1016/j.patter.2021.100392

Gil de Zúñiga, H., Goyanes, M., & Durotoye, T. (2024). A Scholarly Definition of Artificial Intelligence (AI): Advancing AI as a Conceptual Framework in Communication Research. *Political Communication*, *41*(2), 317–334. https://doi.org/10.1080/10584609.2023.2290497

Gurevich, E., El Hassan, B., & El Morr, C. (2023). Equity within AI systems: What can health leaders expect? *Healthcare Management Forum*, *36*(2), 119–124. https://doi.org/10.1177/08404704221125368

Hickok, M., Dorsey, C., O'Brien, T., Baur, D., Ingram, K., Chauhan, C., & Gamundani, A. (2022). *Case Study: The Distilling of a Biased Algorithmic Decision System through a Business Lens* (SSRN Scholarly Paper 4019672). https://doi.org/10.2139/ssrn.4019672

Hubbard, E. L., Bachali, P., Kingsmore, K. M., He, Y., Catalina, M. D., Grammer, A. C., & Lipsky, P. E. (2023). Analysis of transcriptomic features reveals molecular endotypes of SLE with clinical implications. *Genome Medicine*, *15*(1), 84.

https://doi.org/10.1186/s13073-023-01237-9

Khan, B., Fatima, H., Qureshi, A., Kumar, S., Hanan, A., Hussain, J., & Abdullah, S. (2023). Drawbacks of Artificial Intelligence and Their Potential Solutions in the Healthcare Sector. *Biomedical Materials & Devices (New York, N.y.)*, 1–8. https://doi.org/10.1007/s44174-023-00063-2

Levi, R., & Gorenstein, D. (2023, June 6). AI in medicine needs to be carefully deployed to counter bias – and not entrench it. *NPR*. https://www.npr.org/sections/health-shots/2023/06/06/1180314219/artificial-intelligence-racial-bias-health-care

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., … Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*(7788), Article 7788. https://doi.org/10.1038/s41586-019-1799-6

Moorthy, G. S., Young, R. R., Smith, M. J., White, M. J., Hong, H., & Kelly, M. S. (2023). Racial Inequities in Sepsis Mortality Among Children in the United States. *Pediatric Infectious Disease Journal*, *42*(5), 361–367. https://doi.org/10.1097/INF.0000000000003842

Murphy, K., Di Ruggiero, E., Upshur, R., Willison, D. J., Malhotra, N., Cai, J. C., Malhotra, N., Lui, V., & Gibson, J. (2021). Artificial intelligence for good health: A scoping review of the ethics literature. *BMC Medical Ethics*, *22*(1), 14. https://doi.org/10.1186/s12910-021-00577-8

Murray, S. G., Wachter, R. M., & Cucina, R. J. (2020). Discrimination By Artificial Intelligence

In A Commercial Electronic Health Record—A Case Study. *Health Affairs Forefront*. https://doi.org/10.1377/forefront.20200128.626576

Nadis, S. (2022, February 1). *The downside of machine learning in health care*. MIT News | Massachusetts Institute of Technology. https://news.mit.edu/2022/marzyeh-ghassemi-explores-downside-machine-learning-health-care-0201

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

Optum. (n.d.). *Health Risk Analytics—Impact Pro | Optum*. Retrieved February 12, 2024, from https://www.optum.com/business/health-plans/data-analytics/predict-health-risk.html

Panesar, A. (2023). *Precision Health and Artificial Intelligence: With Privacy, Ethics, Bias, Health Equity, Best Practices, and Case Studies* (1st ed.). https://learning.oreilly.com/library/view/precision-health-and/9781484291627/

Rigby, M. J. (2019). Ethical Dimensions of Using Artificial Intelligence in Health Care. *AMA Journal of Ethics*, *21*(2), 121–124. https://doi.org/10.1001/amajethics.2019.121

Sargent, S. L. (2021). AI Bias in Healthcare: Using ImpactPro as a Case Study for Healthcare Practitioners' Duties to Engage in Anti-Bias Measures. *Canadian Journal of Bioethics*, *4*(1), 112–116. https://doi.org/10.7202/1077639ar

Schwartz, A. L., Alsan, M., Morris, A. A., & Halpern, S. D. (2023). Why Diverse Clinical Trial Participation Matters. *New England Journal of Medicine*, *388*(14), 1252–1254. https://doi.org/10.1056/NEJMp2215609

Sendak, M. P., Ratliff, W., Sarro, D., Alderton, E., Futoma, J., Gao, M., Nichols, M., Revoir, M., Yashar, F., Miller, C., Kester, K., Sandhu, S., Corey, K., Brajer, N., Tan, C., Lin, A., Brown, T., Engelbosch, S., Anstrom, K., … O'Brien, C. (2020). Real-World Integration of a Sepsis Deep Learning Technology Into Routine Clinical Care: Implementation Study. *JMIR Medical Informatics*, *8*(7), e15182. https://doi.org/10.2196/15182

Siala, H., & Wang, Y. (2022). SHIFTing artificial intelligence to be responsible in healthcare: A systematic review. *Social Science & Medicine*, *296*, 114782. https://doi.org/10.1016/j.socscimed.2022.114782

Smith, M. J., Axler, R., Bean, S., Rudzicz, F., & Shaw, J. (2020). Four equity considerations for the use of artificial intelligence in public health. *Bulletin of the World Health Organization*, *98*(4), 290–292. https://doi.org/10.2471/BLT.19.237503

Wang, H., Ma, C., & Zhou, L. (2009). A Brief Review of Machine Learning and Its Application. *2009 International Conference on Information Engineering and Computer Science*, 1–4. https://doi.org/10.1109/ICIECS.2009.5362936

Williams, W. W., Hogan, J. W., & Ingelfinger, J. R. (2021). Time to Eliminate Health Care Disparities in the Estimation of Kidney Function. *New England Journal of Medicine*, *385*(19), 1804–1806. https://doi.org/10.1056/NEJMe2114918

Zink, A., Morriss, S., & Gangopadhyaya, A. (2023). *Building Equitable Artificial Intelligence in Health Care*.