Toward an "Equitable" Assimilation of Artificial Intelligence and Machine Learning Into Our Health Care System

Ritu Agarwal, Guodong (Gordon) Gao

Enthusiasm about the promise of artificial intelligence and machine learning in health care must be accompanied by oversight and remediation of any potential adverse effects on health equity goals that these technologies may create. We describe five equity imperatives for the use of AI/ML in health care that require attention from health care professionals, developers, and policymakers.

Background

Artificial intelligence (AI) and machine learning (ML) are being increasingly heralded as transformational technologies that may potentially be able to address long-standing and persistent challenges in health care practice and delivery related to patient safety, health care quality, costs, equity, and access [1]. Promising use cases demonstrating the potential value of AI/ML have been described across the clinical and administrative spectrum [2], including drug discovery, analysis and interpretation of radiological images, early detection of sepsis, hospital resource management, automated generation of clinical encounter notes, and insurance claims processing. However, while the opportunities enabled by AI/ML are doubtless encouraging, we must temper the optimism and excitement generated by these tools with intentional and constant vigilance to the adverse interaction they may have with health equity goals [3].

The prevalence of health inequities, starkly underscored by the COVID-19 pandemic, remains a pernicious threat to marginalized populations [4]. Despite widespread acknowledgement of the serious nature of this threat and ongoing policy efforts, unequal health outcomes continue to be present among communities and sub-populations differentiated by a range of sociodemographic social determinants of health (SDOH), and racial and ethnic factors [5]. Marginalized populations experience worse outcomes across multiple disease conditions and interactions with the health care system, including cancer survival rates, blood pressure control in hypertensive patients [6], unmanaged diabetes, infant and maternal mortality, and access to medical resources [9]. Systematically eliminating such differences must be a major objective in our policy and practice efforts as AI/ML continues to be diffused throughout health care.

The juxtaposition of a system that is already rife with health inequities with the digital transformation of health care and its accelerating reliance on data and algorithms illustrates the potential danger posed by AI/ML tools. AI/ML tools are constructed on a foundation of data and incorporated into models and algorithms by data scientists and tool developers. It is important to recognize that data and models are fundamentally artifacts of human creation and inevitably reflect societal, structural, and individual biases that are root causes of health inequities [10]. When such data and models are institutionalized and broadly implemented into systems of care delivery, the concern is that they will further exacerbate biases and adversely affect health equity.

Policymakers and practitioners seeking to leverage the power of AI/ML to positively affect health care must ensure that the principle of health equity remains front and center in tool development, and that appropriate guardrails are constructed for AI/ML applications being proposed for deployment in clinical and administrative tasks. In previous work we have presented an "AI Bias-Aware Framework" to guide the efforts of data scientists as they develop AI applications for health care [11]. In this commentary, we build upon that work with guidance for policymakers and practitioners. Accomplishing health equity goals implies paying attention to five overarching issues. Three considerations are technical in nature and reflect concerns that may arise from the way in which specific AI/ML tools are developed. The final two issues relate to broader societal questions for which policies and guidelines need to be constructed.

The Equity Imperatives for AI/ML in Health Care

Data Quality and Representativeness

As the oil that fuels AI/ML applications, data present a critical source of bias in ML. Building these tools requires large training datasets that are culled together from a variety

Electronically published July 10, 2024.

Address correspondence to Ritu Agarwal, Center for Digital Health and Artificial Intelligence, 100 International Dr, Baltimore, MD 21202 (ritu.agarwal@jhu.edu).

N C Med J. 2024;85(4):246-250. 0029-2559/2024/85408

of sources that may include patient data stored in electronic health record systems, claims data maintained by payers, and daily activity data captured through smart sensors and other wearable technologies. ML algorithms use these datasets to learn patterns and make predictions such as how likely a particular patient is to be readmitted within 30 days after being discharged [12], the probability that a patient will respond positively to a specific therapeutic intervention, or whether a radiographic image shows signs of a malignant tumor.

The critical reliance on data for training makes ML applications especially vulnerable to data biases. Thus, there is a pressing need to carefully examine underlying sources of data and the degree to which they are representative of the true diversity present across subpopulations and communities. For example, a range of biases have been identified in EHR data alone, including missing data, inadequate sample sizes for certain vulnerable groups that contribute to lower predictive power, and misclassification or errors in the data. Marginalized populations have historically had lower trust in the health care enterprise and exhibited reticence in engaging with the health care system in general, and data sharing in particular [14]. This further reduces the availability of data for these communities that can feed the ML applications.

The literature documents numerous examples of the inadequate performance of ML algorithms for diverse patients arising from data limitations. These include underdiagnosis of minority populations in the analysis of chest radiography images [15], the presence of gender imbalance in image datasets used for automated diagnosis of thoracic diseases [16], limitations of publicly available datasets used for the diagnosis of skin cancer where dark-skinned patients are underrepresented, and lower performance of diabetic retinopathy applications in low-income countries [17].

The limited availability of data for subpopulations and the associated negative impact on health equity are vividly captured in the notion of health data poverty: "the inability for individuals, groups, or populations to benefit from a discovery or innovation due to insufficient data that are adequately representative" [18]. Health data poverty amplifies the difference in benefits that accrue to vulnerable populations, further damaging health equity. A recent study found that genome-wide association studies (GWAS) that are being utilized for understanding the relationship between genetic diversity and disease to enable the design of more targeted therapies conduct analyses in data that are dominated by populations with European ancestry. As a result, the opportunity for other races and ethnicities to benefit from these discoveries is limited [19].

ML Model Specification Biases

Bias in ML models can arise from data non-representativeness but can also be caused by the specific performance objectives and outcomes for which the model is optimized. During development, ML models are subjected to a process of iterative optimization to determine which model performs best for a pre-specified set of objectives. To illustrate, a model may be optimized for objectives such as predictive accuracy or minimization of false positives. However, such objectives can result in a penalty for underrepresented populations in the data: improving accuracy for the dataset as a whole may yield a model that is optimized for the majority group but underperforms for others. This "fairness-accuracy" tradeoff is widely acknowledged in the ML literature [20] and is increasingly being recognized in health care. Model developers as well as clinical and administrative leaders must pay close attention to not only the overall performance of ML algorithms but also the outcomes they yield for underrepresented populations.

A compelling example of how the choice of outcome variable or prediction target for an algorithm can unintentionally cause harm is described by Obermeyer and coauthors (2019) [21]. An algorithm widely used to predict risk scores for determining whether patients should be enrolled in care management programs used health costs as a proxy for health needs. Black patients historically had lower health care spending and were erroneously assumed to be in lower need of further care [21]. To ensure conformance with equity goals, policymakers need clear visibility into the comparative performance of models across subpopulations, as well as the specific performance objectives that developers sought to achieve. The US Food and Drug Administration (FDA) should establish a flexible, risk-based regulatory framework that encourages innovation while prioritizing patient safety, and address concerns about exacerbating existing disparities in health care access and outcomes.

ML Model Drift

The data that fuel ML applications are not static: they change over time. This can sometimes make the historical data used for training ML models obsolete, depending on the degree of volatility in the data. In health care, examples of data obsolescence can be found across the clinical spectrum. For example, new drugs routinely replace older therapies, and clinical guidelines for cancer treatment are consistently revised to reflect the latest scientific developments [22]. ML models can become inaccurate when the data used to determine the model parameters changes over time, a phenomenon referred to as "drift." When such drift occurs, model performance can degrade substantially and the generalizability of the model across a different context and setting (e.g., trained with data from one hospital system to be used in another) is likely to be low.

What types of drift can potentially occur in medical AI applications? Sahiner and coauthors (2023) describe three categories of drift: *input*, *clinical context of use*, and *concept* [23]. Input drift refers to changing characteristics of ML input data, such as differences in the instrumentation used to capture clinical data when they are generated. Input drift can only occur when there are differences in the patient populations used for model training and model deployment. An ML

model developed with EHR data from a large urban health system may perform poorly in a safety-net hospital. The second form of drift, clinical context of use, can occur when the environments for model development and deployment are heterogeneous [23]. Sahiner and coauthors (2023) use the illustration of cancer prevalence to demonstrate this form of drift: when changes occur in disease prevalence between the time of model training and development to the time of actual deployment, the model calibration may underperform and produce results that are inaccurate or incorrect. The final form of drift, concept drift, is a result of the first two categories and reflects the changes that have occurred in the relationship between input variables (e.g., a patient's health data and history) and output (the predicted probability of a certain disease being present).

To mitigate the potential for model drift, practitioners need to "stress test" AI models across different contexts and settings prior to broader deployment. It is necessary to fully disclose the sources of data, as well as details about the data collection and generation process used for different data elements. Policymakers need to exercise oversight over the safety and efficacy of AI methods and models that are likely to be increasingly bundled into EHR systems to ensure that they are addressing equity considerations [24].

Building Diverse Teams for AI/ML Development

The goals of health equity can only be met when diverse voices, experiences, and perspectives are shared and heard. This implies that teams responsible for the conceptualization, development, and deployment of AI/ML applications in health care must include diverse members that reflect the heterogeneity in race, gender, ethnicity, and SDOH in the nation's population. There are two challenges in accomplishing this. First, the underlying scientific disciplines for AI/ML are the science, technology, engineering, and mathematics (STEM) fields. Diversity in STEM education and the STEM workforce has been a longstanding concern in the United States and despite some improvements over the past decade, minority populations in STEM occupations are not proportionally represented (e.g., women are only one-third of the STEM workforce; Black, Hispanic, Alaska Native, and American Indian individuals comprise 35% of the US population but only 24% of the STEM workforce) [25]. Second, the medical profession in the United States does not reflect the full diversity of the population and, as in the case of the STEM workforce, although enrollment of minorities in medical education has been higher in recent years gaps in representation still remain [26]. In 2023, less than 6% of all US physicians identified as Black [27]. The lack of diversity in the two groups that will be leading the development of AI/ML applications in health care will doubtless have an adverse effect on health equity goals.

The construction of diverse teams must be an intentional activity of practitioners charged with AI/ML development efforts. From a policy perspective, the NIH's AIM-AHEAD program that is specifically seeking to address health equity through amplifying the representation of minorities in AI/ML research and development is a model that can potentially be expanded and replicated across states [28].

Patient Literacy and Autonomy

A key aspect of health equity is that all patients have the cognitive resources and literacy to fully understand the nuances of the care being provided to them, including diagnoses and recommended treatments. Patients also have the right to expect that providers are able to provide explanations or rationale underlying proposed care plans, and that final decisions about care will be made with joint discussion and deliberation [29].

The introduction of AI/ML in direct patient care poses two concerns for health equity goals. One: with the introduction of AI/ML, patients may experience a diminished personal connection with the provider, further attenuating candor and potentially resulting in them choosing to hide information from the physician or avoid seeking care even when necessary. To the degree that reticence in information sharing is more prevalent among marginalized populations, this can further intensify the issue of data poverty for these groups. Recent public opinion surveys indicate that more than 50% of patients believe that the use of AI would negatively affect their relationship with their doctor [30].

A second concern relates to patient autonomy: the ability of patients to have a voice and self-determination in their own care [31]. Patients can exercise autonomy only when they have the competence to evaluate the nature and quality of care being offered to them. With the infusion of AI/ML into direct care delivery, several ethical dilemmas arise. What happens when the care plan is constructed by an algorithm? If the algorithm is a "black box," as many ML models are [32], on what basis will the provider explain the recommendation to the patient? What is the nature of informed consent in such a setting? Does the patient have a right to know that the AI is augmenting the doctor? The consequence of diminished patient autonomy is likely to be experienced by vulnerable populations who are already burdened by unequal health outcomes.

Conclusion

The promise of AI/ML to address effectiveness, equity, and efficiency outcomes in health care is robust. In recent work we have shown the myriad of ways in which AI can augment the critical care delivery work of clinicians, potentially even helping to correct human biases they may exhibit [33]. Yet, this promise comes with the potential threat of further eroding equity in a system that already exhibits an unacceptable level of inequity. Practitioners, including leaders in health care delivery organizations, clinicians responsible for the delivery of direct patient care, and technology industry executives, must pay tenacious attention to the considerations discussed here and continually inspect and

audit AI/ML applications for fairness. Policymakers must design incentives that promote equity considerations, such as benchmarks for model performance across subgroups prior to the approval by the FDA and continued audits of AI/ML models used in clinical settings to ensure relevance and adaptation to new research findings, clinical guidelines, and treatment protocols [34]. Policies that encourage educational programs for the broader dissemination of AI/ML knowledge and understanding among vulnerable populations can be critical in bringing these groups to the table to collectively shape the development of applications. The Executive Order issued recently by the White House on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence [35] represents an important step toward ensuring that AI/ML further health equity goals, embodying principles and actions reflected in the OECD Collective Action for Responsible AI in Health [36]. NCMJ

Ritu Agarwal, PhD William Polk Carey Distinguished Professor and Co-Director, Center for Digital Health and Artificial Intelligence, Johns Hopkins University, Baltimore, Maryland.
Guodong (Gordon) Gao, PhD Professor and Co-Director, Center for Digital Health and Artificial Intelligence, Johns Hopkins University, Baltimore, Maryland.

Acknowledgments

We thank Jennifer Bagdasarian and Rui Han for research assistance.

The authors have no conflicts of interests related to the writing and publication of this article.

References

1. Topol E. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again.* Basic Books; 2019.
2. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med.* 2022;28(1):31-38. doi:10.1038/s41591-021-01614-0
3. Agarwal R, Bjarnadottir M, Clark J, Rhue L, Gao G (Gordon). Data Science for Health Equity: Perils and Promise of AI/ML in Healthcare. *Data Science for Health Equity.* Published online April 24, 2023. Accessed April 22, 2024. https://papers.ssrn.com/abstract=4428084
4. Lopez L III, Hart LH III, Katz MH. Racial and ethnic health disparities related to COVID-19. *JAMA.* 2021;325(8):719-720. doi:10.1001/jama.2020.26443
5. Hill L, Ndugga N, Published SA. Key Data on Health and Health Care by Race and Ethnicity. KFF. Published March 15, 2023. Accessed April 22, 2024. https://www.kff.org/racial-equity-and-health-policy/report/key-data-on-health-and-health-care-by-race-and-ethnicity/
6. Abrahamowicz AA, Ebinger J, Whelton SP, Commodore-Mensah Y, Yang E. Racial and ethnic disparities in hypertension: barriers and opportunities to improve blood pressure control. *Curr Cardiol Rep.* 2023;25(1):17-27. doi:10.1007/s11886-022-01826-x
7. Heard-Garris N, Yu T, Brody G, Chen E, Ehrlich KB, Miller GE. Racial discrimination and metabolic syndrome in young Black adults. *JAMA Netw Open.* 2024;7(4):e245288. doi:10.1001/jamanetworkopen.2024.5288
8. Dayo E, Christy K, Habte R. Health in colour: black women, racism, and maternal health. *Lancet Reg Health – Am.* 2023;17. doi:10.1016/j.lana.2022.100408
9. Caraballo C, Ndumele CD, Roy B, et al. Trends in racial and ethnic disparities in barriers to timely medical care among adults in the US, 1999 to 2018. *JAMA Health Forum.* 2022;3(10):e223856. doi:10.1001/jamahealthforum.2022.3856
10. Nazer LH, Zatarah R, Waldrip S, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digit Health.* 2023;2(6):e0000278. doi:10.1371/journal.pdig.0000278
11. Agarwal R, Bjarnadottir M, Rhue L, et al. Addressing algorithmic bias and the perpetuation of health inequities: an AI bias aware framework. *Health Policy Technol.* 2023;12(1):100702. doi:10.1016/j.hlpt.2022.100702
12. Talwar A, Lopez-Olivo MA, Huang Y, Ying L, Aparasu RR. Performance of advanced machine learning algorithms overlogistic regression in predicting hospital readmissions: A meta-analysis. *Explor Res Clin Soc Pharm.* 2023;11:100317. doi:10.1016/j.rcsop.2023.100317
13. 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):1544-1547. doi:10.1001/jamainternmed.2018.3763
14. Kennedy BR, Mathis CC, Woods AK. African Americans and their distrust of the health care system: healthcare for diverse populations. *J Cult Divers.* 2007;14(2):56-60.
15. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med.* 2021;27(12):2176-2182. doi:10.1038/s41591-021-01595-0
16. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci.* 2020;117(23):12592-12594. doi:10.1073/pnas.1919012117
17. Cleland CR, Rwiza J, Evans JR, et al. Artificial intelligence for diabetic retinopathy in low-income and middle-income countries: a scoping review. *BMJ Open Diabetes Res Care.* 2023;11(4):e003424. doi:10.1136/bmjdrc-2023-003424
18. Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Health data poverty: an assailable barrier to equitable digital health care. *Lancet Digit Health.* 2021;3(4):e260-e265. doi:10.1016/S2589-7500(20)30317-4
19. Fitipaldi H, Franks PW. Ethnic, gender and other sociodemographic biases in genome-wide association studies for the most burdensome non-communicable diseases: 2005–2022. *Human Molecular Genetics.* 2023;32(3):520–532. https://doi.org/10.1093/hmg/ddac245
20. Chen RJ, Wang JJ, Williamson DFK, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat Biomed Eng.* 2023;7(6):719-742. doi:10.1038/s41551-023-01056-8
21. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366(6464):447-453. doi:10.1126/science.aax2342
22. Zhang W, Lee AM, Jena S, et al. Computational drug discovery for castration-resistant prostate cancers through in vitro drug response modeling. *Proc Natl Acad Sci.* 2023;120(17):e2218522120. doi:10.1073/pnas.2218522120
23. Sahiner B, Chen W, Samala RK, Petrick N. Data drift in medical machine learning: implications and potential remedies. *Br J Radiol.* 2023;96(1150):20220878. doi:10.1259/bjr.20220878
24. EHR giant Epic plans to launch AI validation software. *Fierce Healthcare.* Published April 3, 2024. Accessed April 22, 2024. https://www.fiercehealthcare.com/ai-and-machine-learning/epic-plans-launch-ai-validation-software-healthcare-organizations-test
25. Diversity and STEM: Women, Minorities, and Persons with Disabilities 2023. NSF - National Science Foundation. Accessed April 22, 2024. https://ncses.nsf.gov/pubs/nsf23315/report
26. Medical student diversity sees uptick—for now. American Medical Association. Published December 21, 2023. Accessed April 22, 2024. https://www.ama-assn.org/education/medical-school-diversity/medical-student-diversity-sees-uptick-now
27. Howard J. Only 5.7% of US doctors are Black, and experts warn the shortage harms public health | CNN. Published February 21, 2023. Accessed April 22, 2024. https://edition.cnn.com/2023/02/21/health/black-doctors-shortage-us/index.html
28. Vishwanatha JK, Christian A, Sambamoorthi U, Thompson EL, Stinson K, Syed TA. Community perspectives on AI/ML and health equity: AIM-AHEAD nationwide stakeholder listening sessions. *PLOS Digit Health.* 2023;2(6):e0000288. doi:10.1371/journal.pdig.0000288
29. Robertson C, Woods A, Bergstrand K, Findley J, Balser C, Slepian MJ. Diverse patients' attitudes towards Artificial Intelligence (AI) in diagnosis. *PLOS Digit Health.* 2023;2(5):e0000237. doi:10.1371/journal.pdig.0000237
30. Tyson A, Pasquini G, Spencer A, Funk C. 60% of Americans Would Be Uncomfortable With Provider Relying on AI in Their Own Health Care. Policy Commons. Published online February

22, 2023. Accessed April 22, 2024. https://policycommons.net/artifacts/3453213/60-of-americans-would-be-uncomfortable-with-provider-relying-on-ai-in-their-own-health-care/4253539/
31. Entwistle VA, Carter SM, Cribb A, McCaffery K. Supporting patient autonomy: the importance of clinician-patient relationships. *J Gen Intern Med.* 2010;25(7):741-745. doi:10.1007/s11606-010-1292-2
32. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell.* 2019;1(5):206-215. doi:10.1038/s42256-019-0048-x
33. Agarwal R, Dugas M, Gao G (Gordon). Augmenting physicians with artificial intelligence to transform healthcare: challenges and opportunities. *J Econ Manag Strategy.* 2024;33(2):360-374. doi:10.1111/jems.12555
34. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat Med.* 2021;27(4):582-584. doi:10.1038/s41591-021-01312-x
35. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Published October 30, 2023. Accessed June 7, 2024. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/
36. Anderson B, Sutherland E. *Collective Action for Responsible AI in Health.* OECD; 2024. doi:10.1787/f2050177-en