Patterns

CellPress
OPEN ACCESS

Perspective

In medicine, how do we machine learn anything real?

Marzyeh Ghassemi[1,2,3,*] and Elaine Okanyene Nsoesie[4,5]
[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Institute for Medical Engineering & Science Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3]CIFAR AI Chair, Vector Institute, Toronto, ON M5G 1M1, Canada
[4]Department of Global Health, School of Public Health, Boston University, Boston MA, USA
[5]Center for Antiracist Research, Boston University, Boston MA, USA
*Correspondence: mghassem@mit.edu

https://doi.org/10.1016/j.patter.2021.100392

THE BIGGER PICTURE Evidence indicates that data acquired from and about human bodies in medicine and health do not always create equitable systems. Bias is pervasive in clinical devices, interventions, and interactions. These include devices that are designed without regard for sex, gender, and skin color; interventions that embed race; disease diagnoses that hinge on gender or ethnicity; and biased interactions between patients and health workers. Data from these systems when used in machine learning algorithms will promote or exacerbate these biases. Often there is a lack of education in computer science about the systemic impact of gender discrimination, racism, and socioeconomic inequalities on data used in developing machine learning algorithms for health. The solutions to addressing these engrained biases are not easy and require intentional efforts by those who develop algorithms and those who use algorithms including, computer scientists, engineers, clinicians, healthcare institutions, and others. However, these solutions cannot exist without education about the historical injustices against marginalized groups, a refusal to accept inequities as the norm, and shouldering the responsibility to create and apply algorithms that reduce rather than promote inequity.

1 2 3 4 **5** Mainstream: Data science output is well understood and (nearly) universally adopted

SUMMARY

Machine learning has traditionally operated in a space where data and labels are assumed to be anchored in objective truths. Unfortunately, much evidence suggests that the “embodied” data acquired from and about human bodies does not create systems that function as desired. The complexity of health care data can be linked to a long history of discrimination, and research in this space forbids naive applications. To improve health care, machine learning models must strive to recognize, reduce, or remove such biases from the start. We aim to enumerate many examples to demonstrate the depth and breadth of biases that exist and that have been present throughout the history of medicine. We hope that outrage over algorithms automating biases will lead to changes in the underlying practices that generated such data, leading to reduced health disparities.

INTRODUCTION

Machine learning is a group of algorithms, models, and techniques that broadly seeks to find patterns from observed data that could be usefully applied to predictions in unseen data.[1] Humans are imperfect, and society generally accepts that there must therefore be some bias in any human-driven processes, even those that may have life or death consequences. In a medical context, this means that there are potentially learnable biases that have been part of medical data as far back as important medical devices or interventions have existed.

As machine learning researchers in health, we have sometimes operated under the assumption that the labels in observed data are “real,” i.e., factual or unbiased.[2] For instance, we assume that the diagnoses extracted from electronic health care records (EHR) are accurate and then build a model that will predict those diagnoses as an output. We use the vital sign measurements collected from medical devices as inputs to models that predict future physiology. We assume that the interventions observed in a clinical setting are unbiased implementations of appropriate and expert medical care, and then build models to mimic the policies we see enacted. Unfortunately, much evidence suggests that the “embodied” data (i.e., data acquired from and about human bodies) in medicine and human health does not always create systems that are equitable, specifically these systems do not allow the same advantages for the diverse populations that are served.[3–5]

Bias, defined as systemic neglect, stereotypes, and beliefs and actions that create or promote health disparities,[5–8] are present in medical devices, clinical interactions, clinical diagnostics, clinical interventions, and policies. A clinical device biased toward certain demographics, clinical interventions, and diagnostics calibrated on a specific population might not work for others, while implicit bias at health care institutions and racist policies lead to a lower quality of care received by millions. Data collected from biased clinical devices and systems will lead to interpretation bias, and machine learning algorithms trained on these data will replicate or exacerbate these biases. But why do these biases exist and how can machine learning algorithms learn anything useful from these data?

Many articles have been written about bias in algorithms and how these biases promote or exacerbate inequities. This perspective aims to (1) summarize the breadth, depth, and incredible scope of pervasive biased practices, and (2) elaborate on how these issues are likely to be repeated or worsened in machine learning settings. We focus on algorithms as potential integrand parts of all portions of the clinical pipeline described, e.g., devices/interventions/diagnostics, but do not consider scenarios where algorithms independently constitute any of those components. We aim to enumerate many examples to demonstrate that researchers must be wary of the biases that exist, and that have been present throughout the history of medicine. We hope that the continuous outcry over algorithms automating such biases will lead to a larger examination and redefinition of practice that will reduce health disparities.

BIAS IN CLINICAL DEVICES, INTERVENTIONS, AND INTERACTIONS

To demonstrate the pervasiveness of bias in clinical devices, interventions, and interactions, we provide examples of devices that fail to work as intended across broad sections of humanity. At a very basic level, devices such as pulse oximeters do not correctly capture oxygenation levels for darker skin at low oxygen saturation,[9] exactly when it is crucial to capture oxygen levels. Feiner et al.[9] state that, ''in our 20 years of testing pulse oximeter accuracy, and probably in other testing laboratories, the majority of subjects have been light skinned. Most pulse oximeters have probably been calibrated using light-skinned individuals, with the assumption that skin pigment does not matter.'' The authors also note that this also interacts with gender, potentially impacting intersectional identities more. Another example is the failure of hand soap dispenser stations to recognize dark skin.[10] While this might be considered a minor inconvenience, it reflects broader issues around device failure.

Bias is also present in devices that are designed without considerations of sex or gender differences. For example, manufacturers make a majority of artificial hearts in a ''standard'' size too large for many women, despite a similar number of men and women suffering from heart disease.[11] Similarly, women face increased risk from metal-on-metal hip implants, due in part to anatomic differences between men and women that are not taken into account during implant design.[12] Another study reported that, among subjects treated with a ventricular assist device (VAD) to provide mechanical circulatory support for patients with heart failure, women had a higher rate of stroke. Device manufacturers finally designed a smaller VAD for women in 2010, almost 50 years after the device was first used.[13] Devices that are not the right size or shape for women can lead to avoidable health complications and death.

Interventions like anesthesia are also not immune to such variant impact, with red-haired subjects requiring more and dark-haired subjects less, to obtain the same sedation result.[14] Risk scores across the spectrum of clinical areas embed race into the basic data used to individualize risk assessments.[15] However, algorithms that include these data must recognize that adopting previous and current clinical decision support scores may propagate existing biases. We note that interventions are one example of resource allocation in clinical settings because clinical attention, medication, etc., are all finite resources. Such issues extend to clinical understanding of how to treat female patients—who are not a minority group—even at a basic level of medication dosages. Recent work has demonstrated that women frequently have higher concentrations of drugs in their blood, and take longer to eliminate drugs from their bodies, when given the same drug dose as men.[16] This is a potential contributor to the more than 90% of adverse drug event cases where women experienced worse side effects than men.

Furthermore, algorithms designed to predict patients' health risk and to allocate resources can also be biased toward minoritized populations. Obermeyer et al. showed that a commercial algorithm used for predicting patients' health risk was less likely to refer sicker black patients for additional care compared with white patients, thereby affecting the health care received by millions of patients.[3] By predicting need based on health care expenditure, the algorithm failed to account for disparities in access to health care between black and white patients, differences that are rooted in systemic racism. Such discriminatory algorithms have grave effects on individuals and entire populations since they determine who does, or does not, receive needed care.

Diagnoses also have well-established biases where the very definition of a condition or disease hinges on gender or ethnicity, and has been used in a discriminatory fashion.[15,17–20] Even for conditions that do not rely on race or gender, the gender- or race-specific presentation of a condition may be poorly understood, or ignored, in medical education and literature.[20,21] In dermatology, the low percentage of examples in darker skin (4%–18%) can lead to serious underdiagnosis,[22] as in the COVID-19 pandemic when skin-based manifestations of the condition were initially on light-skin only.[21] If disease presentation is never shown in dark skin tones in medical textbooks, how will clinicians recognize it? Similarly, prior work has shown that female patients disproportionately die from heart attacks, but only when male doctors treat them.[19] Note, the inverse is not true—male patients do not die more often when female doctors treat them post heart attack. As noted in the paper, ''mortality rates decrease when male physicians practice with more female colleagues or have treated more female patients in the past.'' This first effect—that more female *colleagues* help male doctors recognize heart attack in women—is a sobering argument for why representation in a care team helps all doctors improve. Machine learning algorithms that use medical images (such as X-rays) and other clinical data on disease diagnosis are likely to create or worsen these existing inequalities when trained with gender or race/ethnic imbalanced datasets.[23]

Finally, even the interactions between patients and health workers may be loaded with bias.[24,25] A study by Schulman et al. from 1999 found that doctors were much less likely to recommend cardiac catheterization to black patients with medical files that are statistically identical to those of white patients.[26] Recent work on linguistic features on clinical notes suggested that black patients may be subject to systematic bias in clinical perceptions of their credibility.[27] A study by Li et al. reported that, although women presented with higher rates of hypertension and heart failure than men, they were less likely to receive lipid-lowering medications and optimal care at hospital discharge compared with men.[28] The quality of care a patient receives may also vary based on their socioeconomic status, with patients from lower socioeconomic status more likely to receive less optimal care.[29] Other work has shown that end-of-life care disparity can be modeled algorithmically by examining the level of mistrust between patient and caregivers in clinical notes[30]—with higher levels of mistrust for black patients than white ones. Furthermore, elicitation of responses from frontline workers may strongly vary depending on their biases, and this biased elicitation will create data that represent a biased narrative and could be used to remove treatment options from patients.

We note that a key issue with all human-centric deployments of machine learning is the inherent noisiness of human judgments, even when we desire less variability in a judgment process.[31] Such noise is concerning in medicine because human cognitive biases shape judgments that will then dictate the clinical knowledge that is recorded, and the clinical care that is dispensed. In both cases, algorithms will mimic the knowledge and care that they are provided as learning fodder, and human experts will be hard-pressed to understand, audit, and address such biases that can easily be downloaded and deployed by anyone with an internet connection. In an account presented at a symposium on antiracism as public health policy, Patricia Williams, a legal scholar and proponent of critical race theory, described an incident where an algorithm used for making decisions about osteopenia and osteoporosis care could not generate recommendations because she had been labeled by the hospital as black (Antiracism as Health Policy: Race, COVID-19, and Policy Reform. Patricia Williams. University Distinguished Professor of Law and Humanities, Northeastern University School of Law. https://www.bu.edu/sph/conversations/uncategorized/part-2-antiracism-as-health-policy-race-covid-19-and-policy-reform/). Upon changing her race label to white, the algorithm proceeded to ask questions about her family history, whether she had previously broken any bones, and whether she showed signs of rheumatoid arthritis. This experience highlights two critical issues—the use of black box algorithms to disproportionately assign care and medical resources to whites, and the inherent human assumption that race is self-evident.

WHAT ROLE DOES MACHINE LEARNING HAVE?

These examples of the complexity present in all health care data can be linked to a long history of discrimination and racism, and unproven assumptions about race and biology. While clinical staff operate in environments where they can observe such issues (if they choose to), there is often a lack of understanding and education in computer science about how discrimination, racism, and inequalities have influenced the data used in training algorithms in health care.

Many papers and books have discussed how biased data and algorithms lead to biased interventions and policies, which disproportionately affect marginalized individuals and groups. Some algorithms will produce more accurate results for groups that are most represented in the data. For example, an algorithm trained to detect skin diseases using a dataset that mostly represents people with lighter skin will produce less accurate results for people with darker skin. However, increasing representation of diverse populations in such a dataset is not always sufficient since clinical and public health datasets can carry racial, ethnic, socioeconomic, and other social biases due to how they are collected.

Also, without consideration of social factors that affect and shape individual's health based on their membership in specific groups, such as, gender, race/ethnicity, income, or sexual orientation, algorithms used for allocating resources and automating medical decisions will make inaccurate predictions for the most vulnerable or at-risk populations. But adding or removing a patient's details will not automatically lead to unbiased outcomes. For example, inclusion of patient details, such as ethnicity, body mass index, or socioeconomic status, in health care decision algorithms can lead to explicit bias against certain groups. Algorithms that encode race are sometimes designed to reduce or increase risk of diagnosing disease in specific populations solely based on race.[15] However, exclusion of personal details does not automatically make an algorithm unbiased. One study demonstrated that an algorithm designed to predict patients who were likely to miss their hospital appointments based on EHR data did not become unbiased after removing personal details, because it still included prior observations, i.e., information on whether a patient had previously missed an appointment, as a variable in the algorithm.[4] Predictions made by the algorithm were likely to discriminate against particular individuals who may have missed an appointment in the past due to inability to afford transportation or childcare, thereby discriminating against people from lower income groups.

At the population level, algorithms used to develop policies that fail to address these socio-cultural forces that affect the health of marginalized groups are likely to have biased impacts that exacerbate existing health inequalities. Deference to algorithms or probabilities is dangerous—in a world where race, income, and gender concordance increase your probability of survival. Understanding individual patient needs and developing targeted interventions and policies aimed at addressing the needs of marginalized populations is one effective approach. In the previous example, targeted interventions (such as housing, transportation, or child care assistance) to reduce no-shows resulted in a 9% average reduction in no-show rates across 12 clinics, demonstrating that sometimes solutions to algorithmic bias lie beyond the algorithm.

Unfortunately, there are no simple solutions to such entrenched problems. We attempt to make recommendations for those who develop algorithms and those who use these algorithms for decision making, which includes computer scientists, engineers, clinicians, healthcare institutions, and others. Specifically, we suggest how we, as machine learning researchers, can develop an awareness that data labels are often misleading or inappropriate, and develop or adopt practices that lead to targeted solutions.

First, researchers must recognize that implicit biases against minoritized groups in health care forbids naive applications,[32] e.g., if we use clinical scores that—*as defined*—do not accurately estimate the severity of a condition in black patients as prediction targets, the ignoring of black pain is automated.[33]

Beyond the understanding that diverse groups should be in data, a focus on disaggregated data helps emphasize an understanding of who is included and who is excluded. Researchers should also question subgroup assumptions embedded in data or models, as these are often based on unproven assumptions about biology or environment, e.g., black people are more muscular.[34]

Second, the clinicians and health care institutions that work with machine learning systems should apply a ‘‘do-no-harm’’ approach to using algorithms. This requires health care providers to question how algorithms use patients’ information to assess risk and make care recommendations, and perform rigorous and regular audits to ensure that algorithms provide equal care to patients. Most algorithms are developed with the intention of solving a particular health challenge; however, algorithms do not always perform as expected. Regular and rigorous audits to assess how algorithms are affecting the populations they are meant to serve, including marginalized and poor populations, are important for identifying and addressing bias. It is common to quote the "do-no-harm" approach in medicine, but it can be difficult to apply, especially when practitioners lack a thorough understanding of how algorithms used to make decisions contribute to harm.

Third, those who develop and use algorithms for making health care decisions have a responsibility to reduce and not exacerbate health inequalities by studying and acknowledging historical injustices against marginalized groups and adopting systemic anti-racist policies and practices.[34] Researchers without an understanding of the history of discrimination and scientific racism are likely to blame negative health outcomes on biology and behavioral choices rather than the policies that restrict access to high-quality care, increase risk of certain diseases, and restrict access to opportunities and resources that contribute to good health. The impact of racism on individuals and populations is not solely due to individual biases but also institutions and policies that directly and indirectly affect health.[35–39] Policies that encourage residential segregation based on race and ethnicity to those that promote the belief that black people are naturally prone to having higher rates of disease have long existed in the US. These biased policies have been linked to negative health outcomes, including higher rates of cancer, tuberculosis, asthma, and mental health issues. Furthermore, without recognizing the impact of racism and other forms of discrimination, there will not be deliberate efforts to collect data that includes demographic details of groups that are affected; and, without these data, researchers cannot measure and describe the impact of biased policies on health. See Crear-Perry et al.[40,41] and Yousif et al.[40,41] for suggestions on how anti-racist practices can be incorporated into medical education and practice. Similarly, those who develop machine learning algorithms used in health care settings should study the history and impacts of racism, and receive training on how antiracism and decolonial methodologies can be incorporated in the development, evaluation, and deployment of algorithms.

Fourth, machine learning researchers can adopt approaches and ideas from fields where bias has been thoroughly studied. For instance, the causal inference literature has rigorous formal definitions of bias that can be addressed with various methods: selection bias, measurement bias, and confounding bias.[42] Many in the scientific community are working to harness the power of algorithms to improve understanding of, and personalized risk prediction in, disease heterogeneity.[43] For instance, current biobanks, such as the UK-Biobank[44] and All of Us, target ethnically diverse data to understand individual gender or genetic risk. Recently published standards for quality or performance, such as the CONSORT and CONSORT-AI statements,[45] can also guide researchers on how to avoid bias, and provide guidance on appropriate subgroup and secondary analyses.[46] By acknowledging label uncertainty and the veracity of data-driven learning in the health sciences, we can progress past the limitations of past research.[47]

We fully acknowledge that engaging with, and working through, a system with encoded bias is going to be a hard research process. However, the impact of publishing research that may harm a group due to misunderstanding the nuances of underlying data is enormous.[48] Researchers and institutions have the ability, and therefore a responsibility, to reduce health inequalities by acknowledging and redressing historical injustices against marginalized groups and adopting anti-racist practices.[49]

REFERENCES

1. Murphy, K.P. (2021). Machine Learning. In A Probabilistic Perspective, second edition (MIT Press), pp. 1–4.

2. Topol, E.J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. *25*, 44–56.

3. Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science *366*, 447–453.

4. Murray, S.G., Wachter, R.M., Cucina, and Russell, J. (2020). Discrimination by artificial intelligence in a commercial electronic health record—a case study. Health Aff. Blog *10*. https://doi.org/10.1377/hblog20200128.626576.

5. Risberg, G., Johansson, E.E., and Hamberg, K. (2006). Analysis of the risk of gender bias in medicine. Gend. Med. *3*, S32.

6. Dusenbery, M. (2018). Doing Harm: The Truth about How Bad Medicine and Lazy Science Leave Women Dismissed, Misdiagnosed, and Sick (HarperCollins).

7. Doherty, T.S., and Carroll, A.E. (2020). Believing in overcoming cognitive biases. AMA J. Ethics *22*, E773–E778.

8. Pritlove, C., Juando-Prats, C., Ala-Leppilampi, K., and Parsons, J.A. (2019). The good, the bad, and the ugly of implicit bias. Lancet *393*, 502–504.

9. Feiner, J.R., Severinghaus, J.W., and Bickler, P.E. (2007). Dark skin decreases the accuracy of pulse oximeters at low oxygen saturation: the effects of oximeter probe type and gender. Anesth. Analg. *105*, S18–S23, tables of contents.

10. Benjamin, R. (2020). Race after technology: abolitionist tools for the New Jim code. Social Forces, 98 1–3.

11. Eveleth, R. The Design Bias of Heart Failure. (2016).

12. Hartman, C.W., Gilbert, B.J., and Paprosky, W.G. (2009). Gender issues in total hip arthroplasty: length, offset, and osteoporosis. Semin. Arthroplasty, 20 62–65.

13. Eisen, H.J. (2019). Left ventricular assist devices (LVADS): history, clinical application and complications. Korean Circ. J. *49*, 568–585.

14. Liem, E.B., et al. (2004). Anesthetic requirement is increased in redheads. Anesthesiology *101*, 279–283.

15. Vyas, D.A., Eisenstein, L.G., and Jones, D.S. (2020). Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. N. Engl. J. Med. *383*, 874–882.

16. Zucker, I., and Prendergast, B.J. (2020). Sex differences in pharmacokinetics predict adverse drug reactions in women. Biol. Sex Differ. *11*, 32.

17. Diao, J.A., et al. (2021). Clinical implications of removing race from estimates of kidney function. JAMA *325*, 184–186.

18. Ahmed, S., et al. (2020). Examining the potential impact of race multiplier utilization in estimated glomerular filtration rate calculation on African-American care outcomes. J. Gen. Intern. Med. *36*, 464–471.

19. Greenwood, B.N., Carnahan, S., and Huang, L. (2018). Patient–physician gender concordance and increased mortality among female heart attack patients. Proc. Natl. Acad. Sci. U. S. A. *115*, 8569–8574.

20. (2020). Sex and gender: modifiers of health, disease, and medicine. Lancet *396*, 565–582.

21. Lester, J.C., Jia, J.L., Zhang, L., Okoye, G.A., and Linos, E. (2020). Absence of images of skin of colour in publications of COVID-19 skin manifestations. Br. J. Dermatol. *183*, 593–595.

22. Adelekun, A., Onyekaba, G., and Lipoff, J.B. (2021). Skin color in dermatology textbooks: an updated evaluation and analysis. J. Am. Acad. Dermatol. *84*, 194–196.

23. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., and Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc. Natl. Acad. Sci. U. S. A. *117*, 12592–12594.

24. Shen, M.J., et al. (2018). The effects of race and racial concordance on patient-physician communication: a systematic review of the literature. J. Racial Ethn. Health Disparities *5*, 117–140.

25. Williams, D.R. (2018). Stress and the mental health of populations of color: advancing our understanding of race-related stressors. J. Health Soc. Behav. *59*, 466–485.

26. Schulman, K.A., et al. (1999). The effect of race and sex on physicians' recommendations for cardiac catheterization. N. Engl. J. Med. *340*, 618–626.

27. Beach, M.C., et al. (2021). Testimonial injustice: linguistic bias in the medical records of black patients and women. J. Gen. Intern. Med. *36*, 1708–1714. https://doi.org/10.1007/s11606-021-06682-z.

28. Li, S., et al. (2016). Sex and race/ethnicity-related disparities in care and outcomes after hospitalization for coronary artery disease among older adults. Circ. Cardiovasc. Qual. Outcomes *9*, S36–S44.

29. Arpey, N.C., Gaglioti, A.H., and Rosenbaum, M.E. (2017). How socioeconomic status affects patient perceptions of health care: a qualitative study. J. Prim. Care Community Health *8*, 169–175.

30. Boag, W.G. Quantifying Racial Disparities in End-Of-Life Care. (2018).

31. Kahneman, D., Sibony, O., and Sunstein, C.R. (2021). Noise: A Flaw in Human Judgment (Little, Brown).

32. Hall, W.J., et al. (2015). Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. Am. J. Public Health *105*, e60–76.

33. Pierson, E., Cutler, D.M., Leskovec, J., Mullainathan, S., and Obermeyer, Z. (2021). An algorithmic approach to reducing unexplained pain disparities in underserved populations. Nat. Med. *27*, 136–140.

34. Schutte, J.E., et al. (1984). Density of lean body mass is greater in blacks than in whites. J. Appl. Physiol. *56*, 1647–1649.

35. Bailey, Z.D., Feldman, J.M., and Bassett, M.T. (2021). How structural racism works—racist policies as a root cause of U.S. racial health inequities. New Engl. J. Med. *384*, 768–773.

36. Brondolo, E., Gallo, L.C., and Myers, H.F. (2009). Race, racism and health: disparities, mechanisms, and interventions. J. Behav. Med. *32*, 1–8.

37. Bailey, Z.D., et al. (2017). Structural racism and health inequities in the USA: evidence and interventions. Lancet *389*, 1453–1463.

38. Gee, G.C., and Ford, C.L. (2011). Structural racism and health inequities: old issues, new directions. Du Bois Rev. *8*, 115–132.

39. Williams, D.R., Lawrence, J.A., and Davis, B.A. (2019). Racism and health: evidence and needed research. Annu. Rev. Public Health, 40 105–125.

40. Crear-Perry, J., Maybank, A., Keeys, M., Mitchell, N., and Godbolt, D. (2020). Moving towards anti-racist praxis in medicine. Lancet, 396 451–453.

41. Yousif, H., Ayogu, N., and Bell, T. (2020). The path forward—an antiracist approach to academic medicine. N. Engl. J. Med. *383*, e91.

42. Hernan, M.A., and Robins, J.M. (2019). Causal Inference (CRC Press).

43. Wang, M., et al. (2016). Statistical methods for studying disease subtype heterogeneity. Stat. Med. *35*, 782–800.

44. Bycroft, C., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.

45. Liu, X., et al. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. BMJ *370*, m3164.

46. Sun, X., Ioannidis, J.P.A., Agoritsas, T., Alba, A.C., and Guyatt, G. (2014). How to use a subgroup analysis. JAMA *311*, 405.

47. Ioannidis, J.P.A. (2019). Why most published research findings are false. Chance, 32 4–3213.

48. AlShebli, B., Makovi, K., and Rahwan, T. (2020). Retracted article. The association between early career informal mentorship in academic collaborations and junior author performance. Nat. Commun. *11*. https://doi.org/10.1038/s41467-020-19723-8.

49. Jones, C.P. (2018). Toward the science and practice of anti-racism: launching a national campaign against racism. Ethn. Dis. *28*, 231–234.

About the authors

Elaine O. Nsoesie is an assistant professor at Boston University School of Public Health and an Assistant Director of Research at Boston University Center for Antiracist Research. She is a founding member of the Faculty of Computing & Data Sciences, and a Data Science Faculty Fellow. She has a PhD in Computational Epidemiology, MS in Statistics, and BS in Mathematics. Her research is focused on the use of data and technology to address health inequity and improve health in communities globally. She was previously an assistant professor at the Institute for Health Metrics and Evaluation (IHME) in Seattle, Washington.

Dr. Marzyeh Ghassemi is an assistant professor at MIT in electrical engineering and computer science and Institute for Medical Engineering & Science, and a Vector Institute faculty member holding a Canadian CIFAR AI Chair. She holds a Herman L.F. von Helmholtz Career Development Professorship and was also named one of MIT Tech Review's 35 Innovators Under 35. Prior to her PhD in Computer Science at MIT, she received an MSc degree in biomedical engineering from Oxford University as a Marshall Scholar and BS degrees in computer science and electrical engineering as a Goldwater Scholar at New Mexico State University.