EBioMedicine 67 (2021) 103358

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.elsevier.com/locate/ebiom

Review

Ensuring that biomedical AI benefits diverse populations

James Zou[a], Londa Schiebinger[b,*]

[a] Department of Biomedical Data Science, Stanford University, United States
[b] History of Science, Stanford University, United States

ARTICLE INFO

Article History:
Received 19 January 2021
Revised 12 April 2021
Accepted 12 April 2021
Available online 4 May 2021

Keywords:
Health disparities
Artificial intelligence
Machine learning
Health policy
Race/ethnicity
Genetic ancestry
Sex
Gender

ABSTRACT

Artificial Intelligence (AI) can potentially impact many aspects of human health, from basic research discovery to individual health assessment. It is critical that these advances in technology broadly benefit diverse populations from around the world. This can be challenging because AI algorithms are often developed on non-representative samples and evaluated based on narrow metrics. Here we outline key challenges to biomedical AI in outcome design, data collection and technology evaluation, and use examples from precision health to illustrate how bias and health disparity may arise in each stage. We then suggest both short term approaches—more diverse data collection and AI monitoring—and longer term structural changes in funding, publications, and education to address these challenges.

© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

1. Introduction

The use of computer-aided diagnoses dates back to the 1950s with significant gains in the 1970s and exponential growth since 2008 [1]. Fueled by burgeoning health data from devices, genomics, and electronic health records as well as increasing computational capabilities, digital technologies have the potential to improve diagnostic and therapeutic efficacy.

The expanding use of AI in health and medicine, limited here to data-driven algorithms for making predictions, is promising. Deep learning, for example, allows digital technologies to diagnose melanoma, breast cancer lymph node metastasis and diabetic eye disease better than specialists when it is working well [2]. Ambient intelligence—where healthcare spaces are fitted with passive, contactless sensors—can assist clinicians and surgeons to improve the quality of healthcare delivery [3]. AI-driven techniques can achieve cardiologist-level diagnosis of heart conditions from electrocardiogram waveforms [4].

The concern, however, is that these new technologies may exacerbate existing structural health disparities [5]. Sex, gender, and race bias has long plagued health and biomedicine, stretching back to the advent of modern medicine in the eighteenth century and beyond [6–10]. Health inequalities related to socioeconomic status, ethnicity, gender, geographical area, and other social factors are stark both within countries such as the U.S. and the U.K, and between developed and developing countries [11–14]. Most papers on AI in health and medicine were conducted in the US, Europe, and China, [1] raising the possibility that, if exported, these technologies may not work as well in other parts of the world.

Researchers have begun to investigate steps where bias may enter digital technologies: at problem conception where critical decisions about cohorts and outcomes are made; at intermediate stages where data is collected, models developed, and where attributes like race may affect model predictions [15,16]; and finally at deployment where clinicians may be unaware that particular models are poor predictors for particular groups and where less privileged patients have less access to technologies and may distrust mainstream medicine [17–19]. In this article, we use concrete biomedical examples to illustrate how algorithmic bias and disparity can arise due to inadequate outcome choice, data collection, and model evaluation. We highlight personalised medicine as an especially salient challenge because of its growing importance and because it builds on individually specific data. We then discuss both short-term technical approaches and longer-term structural changes to improve the reliability and broaden the benefit of biomedical AI. This article focuses on statistical AI algorithms; embodied AI (e.g. robotics) may have additional challenges of bias due to physical appearance that may reinforce social or cultural stereotypes [20].

* Corresponding author.
E-mail address: schieb@stanford.edu (L. Schiebinger).

https://doi.org/10.1016/j.ebiom.2021.103358
2352-3964/© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

1.1. Outcome choice

Obermeyer et al. found that commercial algorithms designed to predict the health needs of patients with multiple comorbidities encoded bias when cost was used as a proxy for illness. They identified outcome choice as one of the most important decisions made in developing predictive algorithms. The team tested algorithms designed to identify patients in need of coordinated care in efforts to reduce catastrophic events, such as visits to emergency rooms. In the U.S., Black patients have more health needs, but when cost is used as a proxy for need, these needs are not adequately identified [21]. At equal levels of health (measured by number of chronic illnesses), Black patients generate lower costs than whites potentially because they receive fewer inpatient surgeries and outpatient specialist consults. After experiments testing how label choice affected both predictive performance and racial bias, the team recommended using an index variable that combined health prediction with cost prediction [22].

1.2. Instrumentation and data collection

Medical devices, such as the pulse oximeter, collect a wealth of health-related information, including oxygen levels, sleep heart rate, arrhythmia, etc. The problem with devices that use infrared and red light signaling is that these signals interact with skin pigmentation, and accuracy may vary with skin tone. This physical bias in devices means basic data collection can be flawed [23].

Let's take a common medical device, the pulse oximeter, first patented in Japan in 1972 [24]. Oxygen saturation has become a vital sign along with temperature, blood pleasure, pulse rate, and respiratory rate. Pulse oximeters—able to measure oxygen levels without drawing arterial blood—are among the first defenders in emergency rooms, for example, against COVID-19. Yet, pulse oximeters may overestimate arterial oxyhemoglobin saturation at low SaO_2 in patients with darker pigmented skin, [25] meaning that patients may not get the supplemental oxygen needed to avoid damage to vital organs, such as heart, brain, lungs, and kidneys [26]. A recent study compared oxygen saturation measures taken with pulse oximeters with those taken from arterial blood gas. Analysis of over 47,000 paired readings found that oximeters misread blood gases 12 percent of the time in Black patients compared to 4 percent of the time in white patients [27].

Medical researchers have known since 1989 that both deoxyhemoglobin and melanin in skin are primary light absorbers [28]. An early oximeter patent to adjust for skin tone was filed in 1999 [29]. Further patents were filed in 2017 [30] and 2019, [31] both of which take tissue colour (e.g. skin tone or melanin content) into account. Device developers, however, have been slow to take action [5].

Pulse oximeter data are fed into algorithms that increasingly guide hospital decisions. Algorithmic tools can only be as good as the devices feeding data into them. The pulse oximeter case exemplifies a broader phenomenon whereby darker-skin data is often underrepresented in evaluations of medical devices, findings, and algorithms. A recent survey of 36 papers describing cutaneous manifestation associated with COVID-19 showed that none of these papers included photos of darker skins [32].

1.3. Post-development evaluation and monitoring of medical AI

The growing emphasis on reliable and fair AI has exposed limitations in how medical AI algorithms are evaluated and monitored by the community. A recent study reviewing 130 medical AI devices approved for use by the U.S. Food and Drug Administration (FDA) found that 126 out of the 130 devices were evaluated on retrospective data collected before the devices were developed [33]. Retrospective evaluation is more vulnerable to overfitting by ML algorithm compared to randomized prospective studies that are the norm for evaluating new medicines. Moreover only 28% of the approved devices publicly reported evaluation results from more than one clinical site. Less than 13% of the approved AI devices did not report their performance stratified by sex, gender, or race/ethnicity in the public summary. While the FDA may have additional internal performance metrics, the lack of public information makes it challenging for physicians, hospitals, and patients to assess the reliability of the algorithm. This is especially concerning as AI algorithms are known to have heterogeneous performance in different subpopulations due to imbalances in the training data. This study of FDA approved AI further demonstrated how an algorithm predicting pneumothorax that passes the bar when evaluated at one hospital may be much less accurate when tested at a different hospital—the algorithm's prediction accuracy can drop by more than 10% across sites [34]. One contributing factor to the drop in performance is the different composition of race/ethnicity, sex, and gender in the test patients across different sites, and the fact that the thoracic detector had different performances across demographic groups.

1.4. Algorithms for personalised health and medicine

Personalised health is a major application area of machine learning. The idea of personalised health is to discover attributes—e.g. omic biomarkers or wearable sensor readouts—that characterizes an individual's health status and to recommend actionable interventions to reduce future disease risks. Because personalised health efforts typically require large datasets and predictive modeling of future outcomes, it is a key domain for machine learning algorithms and exemplifies the outcome design, data collection, and monitoring challenges we discuss here.

Precision diet based on microbiome is one intriguing example of personalised health and medicine. Models using features of an individual's microbiota—the collection of 100 trillion microbes that live on or within a person's body—have been shown to accurately predict that individual's response to food, e.g. glycemic response to different types of bread [34]. Several startups now offer personalised diet to manage diabetes and other diseases powered by microbiome-based algorithmic predictions. Similarly, recent studies have explored deep monitoring of individuals using wearables and longitudinal measurements of multi-omics including transcriptome, microbiome, and proteome [35,36]. Computational analyses integrating these data have identified personalised trajectories of diabetes progression, though these studies have been validated primarily on data from small number of participants.

The high-dimensional data used in personalising health, such as multi-omics, can greatly vary across individuals and are profoundly affected by environment and behaviours. For example, jetlag can alter the microbiome composition by increasing the relative representation of Firmicutes, a type of bacteria associated with obesity and metabolic diseases [37]. Studies have also demonstrated systematic variation in the microbiome across different ethnic groups [38]. Individuals in each part of the world have unique taxa of microbes that are more abundant in their body, likely due to both genetic and environmental differences. A major machine learning challenge is to ensure that the personalised predictive modeling for diet and disease management work reliably across diverse race/ethnicity, sex, gender, and geography. The danger for algorithmic bias can be especially high for personalised health since the underlying data and individual outcomes are highly heterogeneous. Many of the recent personalised health projects have been conducted in the U.S., Europe, and Israel, and have involved relatively small number of participants often recruited near universities where the researchers are affiliated. How well the technology works in other ethnic groups needs to be carefully evaluated.

2. Short-term technical solutions

2.1. Increasing the diversity of medical data resources

Appropriate data collection is a first crucial step toward developing and evaluating medical AI algorithms. On the one hand, many of the commonly available public datasets, especially for medical imaging, fail to properly represent minorities [39]. On the other hand, private datasets that may be more diverse are typically restricted to a specific hospital or academic center, and do not capture variation across sites. Remedying this gap is an important step towards improving the reliability of medical AI across different populations. Several encouraging efforts are under way. For example, as discussed in case studies above, the paucity of annotated photos of darker-skin individuals is a significant barrier for dermatology and telehealth algorithms. The Stanford Skin of Color Project is an ongoing crowd science effort to collect and curate the largest publicly available dataset of dermatologically relevant images from darker skin tones. This data can help to train and assess machine learning models. In genetics, there are similar efforts to prioritize the collection and analysis of non-European genetics data, which is necessary for genetic understandings of diseases such as polygenetic risk scores to benefit diverse populations [40]. For example, the recent PAGE study demonstrates a robust framework to identify new genetic correlates of phenotypes using over 49,000 non-European individuals [41]. Especially when collecting data from underrepresented groups, it is important to have transparent consent so that participants gain trust on how the data will be used. As we discussed in this article, collecting diverse data is one important step toward bias mitigation; carefully assessing algorithm design, evaluation metrics, and the accessibility of the technology in deployment is also important.

2.2. Monitoring medical AI algorithms post deployment

A key challenge for AI algorithms is that its performance can quickly change over time. This may be due to the algorithm itself being updated as more data is used to train it. This may also be due to changes in user characteristics over time (e.g. as the algorithm becomes more widely-adopted in low resource settings). The usage of the algorithm itself can also lead to changes in the user behavior, thus creating a feedback loop [42]. These phenomena together are called data drift. Careful and continuous monitoring of medical AI systems is therefore critical for ensuring their safe and unbiased application. Post deployment monitoring of AI algorithms is an emerging area of research. Researchers have developed statistical tests to detect if the data that an algorithm is applied to is substantially different from the data that the algorithm was trained on. These tests trigger real-time warnings that the deployed AI algorithm may have biases due to its training data. This can be a practical approach for monitoring [43]. We recommend that hospitals and regulators such as the FDA consider the monitoring framework and the AI algorithm as a holistic package to be evaluated and deployed together.

3. Long-term structural solutions

3.1. Policy and regulatory agencies

Given the growing importance of AI in biomedicine and health care, some regulatory agencies have taken action. The American Medical Association, for example, recommends that AI for health care be "thoughtfully designed, high-quality, [and] clinically validated" [44]. This has been criticised, however, as not going far enough to advance health equity. The American Heart Association, by contrast, has issued a "call to action" to overcome structural racism [45] complete with a structural racism and health equity language guide [46].

The FDA has set out five criteria for excellence in its Digital Health Innovation Action Plan. The specific guidelines to certify medical computer-aided systems, however, do not mention sex, gender, or other axes of health disparities in data collection [47]. Ferryman recommends expanding the current FDA guidelines for software as a medical device (SaMD) to include a four-part pre- and post-market review of ML health tools: an analysis of health disparities in the clinical domain of interest; a review of training data for bias; transparency surrounding decisions made regarding model performance, especially in relation to health disparities; and post-market review of health equity outcomes [48]. It is worth noting that most of the medical AIs approved by the FDA focus on automating relatively low-level aspects of the clinical workflow and that human experts are still responsible for making final decisions. It is therefore important to assess the algorithm not just in isolation but in the context of human usage.

Where regulatory agencies such as the FDA, International Medical Device Regulators Forum, or European Medicines Agency have no jurisdiction, Institutional Review Boards (IRB) may provide oversight to ensure that sex, gender, race, and ethnicity analyses are appropriately integrated into research with enhanced and rigorous reviews. An important part of US IRBs' remit is to ensure that study designs are sound [49]. If study participants are limited to the institution's geographic location, however, the limits of the study should be clearly stated.

3.2. Curriculum and team development

Universities have joined the effort to develop socially responsible AI by reforming curricula. In 2019, Harvard University initiated an Embedded EthiCS curriculum—in response to student demand—that integrates ethical issues into the core computer science (CS) curriculum. Barbara Grosz et al. argue that embedding ethical reasoning throughout the entire CS curriculum has the potential to "habituate students to thinking ethically as they develop algorithms and build systems, both in their studies and as they pursue technical work in their careers" [50]. These courses are taught by interdisciplinary teams of humanists and computer scientists. Stanford University and the Technical University of Munich have implemented similar approaches [51,52].

Across the European Union, a 2020 survey of 61 universities showed that two thirds teach ethics in CS courses (this includes professional codes of ethics in addition to the social issues that are the focus of our interests here). Ethics is most often taught in stand-alone modules (39%) and not linked to core curricula (28%–the other 33% combine both approaches) [53]. The AI for biomedicine community can also benefit from curriculum that incorporates ethics into computational training by drawing from the relevant material in both CS and medical ethics.

Additional problems may arise from the lack of gender and ethnic diversity in AI teams, a situation that can contribute to perpetuating unconscious biases in research design and outcomes. Teams should be diverse in terms of participants and also in terms of skill sets and methods [54]. AI researchers themselves should understand the basics of sex, gender, diversity, and intersectional analysis as these relate to their technical work [55]. In addition, McLennan et al. recommend embedding ethicists directly into AI development teams to enhance robust analysis of social issues—from the beginning of development (and not after the fact) [56].

Next steps will be to review what is included in "ethics". Much the discussion of ethics revolves around narrow, formalist principles or, as we saw in university curricula, codes of professional ethics. Rather than ethics per se, what is needed are robust analyses of social, cultural, and legal issues brought by humanists, social scientists, and legal experts that can anticipate how technologies might reinforce social inequities and suggest structural solutions for overcoming

them [56]. This type of interdisciplinary collaboration may require changing how universities are structured and how researchers are trained. Researchers need to learn to collaborate across disciplinary divides in seamless and productive ways.

3.3. Funding agencies

Funding agencies can serve as gatekeepers to excellence in science, including CS [20]. The European Commission has embraced "ethics by design" for the development of AI systems since 2018. This requires that ethical and legal principles based on the General Data Protection Regulation be implemented in the design process [57]. In 2019, the Stanford University Human-Centered AI Institute implemented an Ethics Review Board, modeled loosely on Institutional Review Boards, to assess grant proposals for social implications of the research before funding [58]. In 2020, the German Research Foundation (DFG)—which includes biomedical research—implemented guidelines for sex, gender, and diversity analysis in proposals, where relevant, in efforts to enhance fairness in research outcomes [59]. Implementation of any of these policies depends on training researchers, evaluators, and staff in issues surrounding social equities and regular reviews of the efficacy of such policies.

3.4. Conferences and peer-reviewed journals

A number of ML and AI conferences have policies to encourage diversity, equity, and inclusion in participation. NeurIPS (Neural Information Processing Systems) conference requires in its call for papers that authors detail "the potential broader impact of their work, including its ethical aspects and future societal consequences." [60,61]. As they state, "regardless of scientific quality or contribution, a submission may be rejected for ethical considerations, including methods, applications, or data that create or reinforce unfair bias or that have a primary purpose of harm or injury." [62]. Although some worry that this may simply lead to "ethics washing" [52] and not be taken seriously, we recommend that other conferences adopt similar policies and that all conferences plan to evaluate this policy every five years.

Editorial boards of peer-reviewed journals and peer-review conferences can also require sophisticated ethical analysis when selecting papers. The journal *Nature Machine Intelligence* is considering requiring authors to include a statement on the broader societal impacts and ethics of their work—but this has yet to be implemented [63]. *The Lancet* has pledged to advance racial equality through editorial oversight in publications [64] and also to conduct editorial checks for appropriate use of sex and gender analysis before accepting manuscripts for publication [65].

3.5. Outstanding questions

The numerous articles we reviewed reveal important examples of how health disparities surrounding race/ethnicity, sex and gender, geographic location, socioeconomic status, etc. are often amplified by AI. These studies, however, tend to treat each factor separately—either race/ethnicity, or sex, or gender, or socioeconomic status, or abilities, etc. What is needed now are intersectional analyses in health and medical research. Intersectionality analyses how overlapping or intersecting forms of discrimination related to a patient's social and cultural life course function in health outcomes [66].

An iconic example of intersectional analysis from facial recognition found that systems that analysed sex and race separately failed to capture the full severity of the bias for Black women. The sex analysis found that systems performed better on men's faces than on women's faces. The race analysis found that systems performed better on lighter-skin than darker-skin. Intersectional analysis found that these single axes missed that systems performed significantly worse for Black women. Error rates were 35% for darker-skinned women, 12% for darker-skinned men, 7% for lighter-skinned women and less than 1% for lighter-skinned men [67].

Returning to pulse oximetry, given that outcomes are worse for Black patients (undifferentiated by sex), would they be significantly worse for darker-skinned women or gender-diverse individuals? Findings are inconclusive. Feiner et al. suggested in 2007 that women, especially those with smaller fingers, exhibit greater variability in oximeter performance, especially at low SaO_2. One study that took an intersectional approach found that SpO_2 values less than 97% were 6 times more frequent in white males than in Black females [68]. A very small 2020 study pointed out, however, that low hemoglobin levels were prevalent in females; thus it was not possible to statistically separate the contributions of sex and low hemoglobin to oximeter bias [69]. A recent study of 47,000 plus patients found that oximeters misread hypoxemia more often in female than in males—but the difference was slight compared to differences related to skin tone [70]. Researchers should take steps to learn what these findings might mean for an intersectional analysis of sex and skin tone. Intersectionality is a technique that needs to be incorporated into analyses to overcome health inequalities. Device makers should consider relevant intersectional analyses when calibrating medical instruments.

While the impact of sex on pulse oximetry remains unclear, that of gender may be significant. A randomized, blind study found that nail polish (worn more often by people who identify as women than as men) interfered with oximetry, with black, blue, and green colours lowering the accuracy of reading more than purple or red [71]. The recommendation is that medical professionals remove patient nail polish before employing an oximeter.

More research is needed to understand intersecting human characteristics of sex, gender, race/ethnicity, socioeconomic status, age, etc. to enhance health outcomes across the whole of society. It is also important to perform life course analysis to assess the impact of algorithms over time, where possible.

4. Conclusion

AI holds tremendous potential to improve many aspects of human health, from early stage biomarker discovery to more effective personalised diagnosis and treatment. As with other biomedical technologies such as genome sequencing and editing, it is critical that innovation in biomedical AI is complemented by efforts to reduce human risk and to ensure that its benefits are broadly shared by diverse countries and populations. Here we have outlined key challenges that can lead to bias and disparity in biomedical AI. Many of these challenges are fundamentally linked to how we design and collect the data used to train and evaluate the algorithms. We also propose both short term approaches—e.g. better post-deployment monitoring of medical AI can be readily adopted—as well as long term structural changes—e.g. ensuring that social, cultural, and ethical analyses are integrated into medical AI curriculum—as steps toward this goal. Clearly, technology alone is not the fix; large social problems that undergird structural inequality need to be addressed [72,73]. Nonetheless, researchers and educators can do their part to develop education and technologies that strive toward social justices.

4.1. Search strategy and selection criteria

Data for this Review were identified by searches of PubMed and Google Scholar, and references from relevant articles using the search terms "artificial intelligence", "machine learning", "sex", "gender", "race", "ethnicity," "health disparities," "precision health". Importantly, to find articles about sex in biomedicine, one often needed to search "gender"; medical researchers tend to use these distinct terms interchangeably and often incorrectly [74]. Only articles published in English between 2000 and 2021 were included.

4.2. Contributors

LS conceptualized the paper. LS and JZ contributed to the contents. Both authors read and approved the final version of the manuscript.

Declaration of Competing Interest

The authors have nothing to disclose.

Acknowledgments

J.Z. is supported by NSF CCF 1763191, NSF CAREER 1942926, NIH P30AG059307, NIH U01MH098953. Funders had no role in data collection, analysis, interpretation, writing or the decision to submit this manuscript for publication.

References

[1] Tran BX, Vu GT, Ha GH, et al. Global evolution of research in artificial intelligence in health and medicine: a bibliometric study. J Clin Med 2019;8(3):360.
[2] Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology 2018;125(8):1264–72.
[3] Haque A, Milstein A, Fei-Fei L. Illuminating the dark spaces of healthcare with ambient intelligence. Nature 2020;585(7824):193–202.
[4] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, Ng AY. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019;25(1):65.
[5] Colvonen PJ, DeYoung PN, Bosompra NO, Owens RL. Limiting racial disparities and bias for wearable devices in health science research. Sleep 2020.
[6] Jones JH. Bad blood: the Tuskegee syphilis experiment. Simon and Schuster; 1981.
[7] Schiebinger L. The mind has no sex? Women in the origins of modern science. Harvard University Press; 1989.
[8] Wailoo K. How cancer crossed the color line. Oxford University Press; 2010.
[9] Smith Taylor J. Women's health research: progress, pitfalls, and promise: institute of medicine of the national academies. Washington, DC: National Academies Press; 2011.
[10] Hogarth RA. Medicalizing Blackness: making racial difference in the Atlantic world, 1780-1840. UNC Press Books; 2017.
[11] National Center for Health Statistics. Health, United States, 2018. Hyattsville, MD.
[12] Public Health England, Health profile for England, 2019. https://www.gov.uk/government/publications/health-profile-for-england-20. Accessed 16 Mar 2021.
[13] OECD/European Union. Health at a glance: Europe 2020: state of health in the EU cycle. Paris: OECD Publishing; 2020 Accessed 16 Mar 2021. doi: 10.1787/82129230-en.
[14] OECD. Health for everyone?: Social inequalities in health and health systems, OECD health policy studies. Paris: OECD Publishing; 2019 Accessed 16 Mar 2021. doi: 10.1787/3c8385d0-en.
[15] Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. N Engl J Med 2020;383:874–82.
[16] DeCamp M, Lindvall C. Latent bias and the implementation of artificial intelligence in medicine. J Am Med Inform Assoc 2020;27(12):2020–3.
[17] Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med 2018;169(12):866–72.
[18] Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in health. arXiv preprint arXiv:2009.10576 2020 forthcoming in Annual Reviews.
[19] Röösli E, Rice B, Hernandez-Boussard T. Bias at warp speed: how AI may contribute to the disparities gap in the time of COVID-19. J Am Med Inform Assoc 2021;28(1):190–2.
[20] Tannenbaum C, Ellis RP, Eyssel F, Zou J, Schiebinger L. Sex and gender analysis improves science and engineering. Nature 2019;575(7781):137–46.
[21] Benjamin R. Assessing risk, automating racism. Science 2019;366(6464):421–2.
[22] 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):447–53.
[23] Kadambi A. Achieving fairness in medical devices. Science 2021;372(6537):30–1.
[24] Severinghaus JW, Honda Y. History of blood gas analysis. VII. Pulse oximetry. J Clin Monit 1987;3(2):135–8.
[25] Feiner JR, Severinghaus JW, Bickler PE. Dark skin decreases the accuracy of pulse oximeters at low oxygen saturation: the effects of oximeter probe type and gender. Anesthesia Analgesia 2007;105(6):S18–23.
[26] Moran-Thomas Amy. How a popular medical device encodes racial bias. Boston Review 2020:157–72. http://bostonreview.net/science-nature-race/amy-moran-thomas-how-popular-medical-device-encodes-racial-bias. Accessed 5 August 2020.
[27] Sjoding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS. Racial bias in pulse oximetry measurement. N Engl J Med 2020;383(25):2477–8.
[28] Ries AL, Prewitt LM, Johnson JJ. Skin color and ear oximetry. Chest 1989;96(2):287–90.
[29] Chin RP, inventor; Nellcor Puritan Bennett LLC, assignee. Oximeter sensor with user-modifiable color surface. United States patent US 5,924,982. 1999.
[30] Bechtel KL, Shultz KM, Margiott AM, Kechter GE, inventors; ViOptix Inc, assignee. Determining Tissue Oxygen Saturation with Melanin Correction. United States patent application US 15/494,444. 2017.
[31] Barker A, Chapman D, Dickin E, Cervi M, inventors; Kent Imaging, assignee. Automatic compensation for the light attenuation due to epidermal melanin in skin images. United States patent US 10,395,352. 2019.
[32] Lester JC, Jia JL, Zhang L, Okoye GA, Linos E. Absence of skin of colour images in publications of COVID-19 skin manifestations. Br J Dermatol 2020;183(3):593–5.
[33] Wu E, Wu K, Daneshjou R, Ouyang D, Ho D, Zou J. How medical AI devices are evaluated: limitations and recommendations from analysis of FDA approvals. Nat Med 2021;5:1–3.
[34] Leshem A, Segal E, Elinav E. The gut microbiome and individual specific responses to diet. mSystems 2020;5:e00665. -20.
[35] Rose SM, Contrepois K, Moneghetti KJ, et al. A longitudinal big data approach for precision health. Nat Med 2019;25(5):792–804.
[36] Ahadi S, Zhou W, Rose SM, et al. Personal aging markers and ageotypes revealed by deep longitudinal profiling. Nat Med 2020;26(1):83–90.
[37] Li Y, Hao Y, Fan F, Zhang B. The role of microbiome in insomnia, circadian disturbance and depression. Front Psychiatry 2018;9:669.
[38] Gupta VK, Paul S, Dutta C. Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. Front Microbiol 2017;8:1162.
[39] Lester JC, Taylor SC, Chren MM. Under-representation of skin of colour in dermatology images: not just an educational issue. Br J Dermatol 2019;180(6):1521–2.
[40] Fernández-Rhodes L, Young KL, Lilly AG, et al. Importance of genetic studies of cardiometabolic disease in diverse populations. Circ Res 2020;126(12):1816–40.
[41] Wojcik GL, Graff M, Nishimura KK, et al. Genetic analyses of diverse populations improves discovery for complex traits. Nature 2019;570(7762):514–8.
[42] Izzo Z, Ying L, Zou J. How to learn when data reacts to your model: performative gradient descent. arXiv preprint arXiv:2102.07698 2021.
[43] Kim MP, Ghorbani A, Zou J. Multiaccuracy: Black-box post-processing for fairness in classification. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society; 2019. p. 247–54.
[44] American Medical Assocation, Aubmented Intelligence in Health Care H-480.940. https://policysearch.ama-assn.org/policyfinder/detail/augmented%20intelligence?uri=%2FAMADoc%2FHOD.xml-H-480.940.xml. Accessed 16 March 2021.
[45] Churchwell K, Elkind MS, Benjamin RM, et al. Call to action: structural racism as a fundamental driver of health disparities: a presidential advisory from the American Heart Association. Circulation 2020 :CIR-0000000000000936.
[46] AHA, Structural racism and health equity language guide. 2020. https://professional.heart.org/-/media/phd-files-2/science-news/s/structural_racism_and_-health_equity_language_guide.pdf. Accessed 16 Mar 2021.
[47] 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):12592–4.
[48] Ferryman K. Addressing health disparities in the food and drug administration's artificial intelligence and machine learning regulatory framework. J Am Med Inform Assoc 2020;27(12):2016–9.
[49] Duffy KA, Ziolek TA, Epperson CN. Filling the regulatory gap: potential role of institutional review boards in promoting consideration of sex as a biological variable. Women's Health 2020;29(6):868–75.
[50] Grosz BJ, Grant DG, Vredenburgh K, Behrends J, Hu L, Simmons A, Waldo J. Embedded EthiCS: integrating ethics across CS education. Commun ACM 2019;62(8):54–61.
[51] Miller, K. Building an Ethical Computational Mindset. Stanford Report. 2020.
[52] Interview with Alena Buyx, An embedded ethics approach for AI development, *NewsRx Health* & Science. Sept 20, 2020.
[53] Ethics4EU, Existing Competencies in the Teaching of Ethics in Computer Science Faculties – Research Report. 2020. http://ethics4eu.eu/outcomes/existing-competencies-in-the-teaching-of-ethics-in-computer-science-faculties-research-report /. Accessed 16 Mar 2021.
[54] Nielsen MW, Bloch CW, Schiebinger L. Making gender diversity work for scientific discovery and innovation. Nature Human Behav 2018;2(10):726–34.
[55] Schiebinger L, Klinge I. Gendered innovations 2: how inclusive analysis contributes to research and innovation Eds. Luxembourg: Publications Office of the European Union; 2020.
[56] McLennan S, Fiske A, Celi LA, Müller R, Harder J, Ritt K, Haddadin S, Buyx A. An embedded ethics approach for AI development. Nature Mach Intell 2020;2(9):488–90.
[57] Annex to the communication from the commission to the European Parliament, the European Council, ... Coordinated Plan on Artificial Intelligence Brussels, 7.12.2018. https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-795-F1-EN-MAIN-PART-1.PDF. Accessed 16 Mar 2021.
[58] Both authors participated in this process. A review is in progress.
[59] German Research Foundation, proposal guidelines. https://www.dfg.de/en/research_funding/principles_dfg_funding/diversity_dimensions/index.html. Accessed 16 Mar 2021.
[60] Gibney E. The battle to embed ethics in AI research. Nature 2020;577(7792):609.
[61] Hecht B, Wilcox L, Bigham JP, Schöning J, Hoque E, Ernst J, Bisk Y, De Russis L, Yarosh L, Anjum B, Contractor D, Wu C. It's time to do something: mitigating the negative impacts of computing through a change to the peer review process. ACM Future of Computing Blog 2018. https://acm-fca.org/2018/03/29/negativeimpacts/. Accessed 7 April 2021.

[62] NeurIPS, Thirty-fourth conference on neural information processing systems, Call for papers, 2020 https://nips.cc/Conferences/2020/CallForPapers. Accessed 23 October 2020.
[63] Van Noorden R. The ethical questions that haunt facial-recognition research. Nature 2020;587:354–8.
[64] *The Lancet*. Advancing racial equality. 2020. https://www.thelancet.com/racial-equality. Accessed 16 Mar 2021.
[65] Schiebinger L, Leopold SS, Miller VM. Editorial policies for sex and gender analysis. Lancet (London, England) 2016;388(10062):2841–2 Adopted by the.
[66] Schiebinger L, Klinge I, Sánchez de Madariaga I, Paik HY, Schraudner M, Stefanick M. Gendered innovations in science, health & medicine, engineering and environment Eds.. Method, Intersectional Approaches; 2020. http://genderedinnovations.stanford.edu/terms/intersectionality.html. Accessed 16 Mar 2021.
[67] Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. Conference on fairness, accountability and transparency, proceedings of machine learning research; 2018. p. 1–15.
[68] Witting MD, Scharf SM. Diagnostic room-air pulse oximetry: effects of smoking, race, and sex. Am J Emerg Med 2008;26(2):131–6.
[69] Choi BM, Kang BJ, Yun HY, Jeon B, Bang JY, Noh GJ. Performance of the MP570T pulse oximeter in volunteers participating in the controlled desaturation study: a comparison of seven probes. Anesthesia Pain Med 2020;15(3):371–7.
[70] Private communications with Michael Sjoding, 17 December 2020.
[71] Coté CJ, Goldstein EA, Fuchsman WH, Hoaglin DC. The effect of nail polish on pulse oximetry. Anesthesia Anal 1988;67(7):683–6.
[72] Editorial. Technology can't fix this. Nature Mach Intell 2020:363.
[73] McLennan S, Lee MM, Fiske A, Celi LA. AI ethics is not a panacea. Am J Bioeth 2020;20(11):20–2.
[74] Madsen TE, Bourjeily G, Hasnain M, et al. Sex-and gender-based medicine: the need for precise terminology. Gender Genome 2017;1(3):122–8.