Mitigating Racial Bias in Machine Learning

Kristin M. Kostick-Quenet, I. Glenn Cohen, Sara Gerke, Bernard Lo, James Antaki, Faezah Movahedi, Hasna Njah, Lauren Schoen, Jerry E. Estep, and J.S. Blumenthal-Barby

Keywords: Algorithmic Bias, Racial Bias, Machine Learning, Artificial Intelligence, Ethics

Abstract: When applied in the health sector, AI-based applications raise not only ethical but legal and safety concerns, where algorithms trained on data from majority populations can generate less accurate or reliable results for minorities and other disadvantaged groups.

I. Introduction

Artificial intelligence (AI) is revolutionizing healthcare from improving diagnostics and imaging to enabling precision medicine, pharmaceutical discovery, and health management. Particularly, Machine Learning (ML) plays a crucial role for extracting complex patterns in data to provide descriptive or predictive information relevant to clinical decision making. For these benefits of AI/ML to be equal across socioeconomic, ethnic, racial, and gender lines, certain fundamental challenges must be addressed. Primary among these is the potential for algorithmic bias, or the presence of systematic and repeatable tendencies in an algorithm to generate unequal outcomes and disparate impacts across population subgroups.[1] In particular, *racial bias*

Kristin M. Kostick-Quenet, Ph.D., *is an Assistant Professor at Baylor College of Medicine in Houston, Texas, USA.* **I. Glenn Cohen, J.D.,** *is James A. Attwood and Leslie Williams Professor of Law, Deputy Dean, and Faculty Director of the Petrie-Flom Center for Health Law Policy, Biotechnology & Bioethics at Harvard Law School in Cambridge, Massachusetts, USA.* **Sara Gerke, Dipl.-Jur. Univ., M.A.,** *is an Assistant Professor of Law at Penn State Dickinson Law in Carlisle, Pennsylvania, USA.* **Bernard Lo, Ph.D.,** *is Professor of Medicine Emeritus and Director of the Program in Medical Ethics Emeritus at UCSF and President Emeritus of the Greenwall Foundation in New York, NY, USA.* **James Antaki, Ph.D.,** *is the Susan K. McAdam Professor of Heart Assist Technology in the Meinig School of Biomedical Engineering at Cornell University in Ithaca, New York, USA.* **Faezah Movahedi** *is a Ph.D. Candidate and Research Assistant at Cornell University in Ithaca, New York, USA.* **Hasna Njah, Ph.D.,** *is an Assistant Professor at in the Multimedia Information Systems and Advanced Computing Laboratory (MIRACL) at the University of Sfax in Tunisia. University of Gabes, Tunisia* **Lauren Schoen** *is a J.D. candidate at the University of Texas School of Law in Austin, Texas, USA.* **Jerry E. Estep, M.D.,** *is the Head of the Section of Heart Failure and Transplantation in the Tomsich Family Department of Cardiovascular Medicine in the Sydell and Arnold Miller Family Heart, Vascular & Thoracic Institute at the Cleveland Clinic in Ohio, USA.* **J.S. Blumenthal-Barby, Ph.D.,** *is the Cullen Professor of Medical Ethics and Associate Director of the Center for Medical Ethics and Health Policy at Baylor College of Medicine in Houston, Texas, USA.*

The Journal of Law, Medicine & Ethics, 50 (2022): 92-100. © 2022 The Author(s)
DOI: https://doi.org/10.1017/jme.2022.13

has been documented across a range of ML-based analyses[2] and is increasingly viewed as a civil rights issue.[3]

Recent calls for regulatory standards and review of large technology companies[4] point to the need to proactively mitigate algorithmic bias. However, very few of the potential negative impacts of ML are addressed by existing legal prohibitions. This is especially concerning for ML applications used in the health sector, where algorithmic bias raises not only ethical and legal issues but also may endanger patients. Algorithms trained on data from majority populations may generate less accurate or reliable results for minorities and other disadvantaged groups.[5]

New regulations and tools for identifying and mitigating algorithmic bias are emerging across the world,[6] yet they often do not include specific or concrete steps for developers to self-regulate and audit their systems. Meanwhile, they overemphasize developers' responsibility for mitigating bias, even though many sources of bias found in algorithms may be systemic, requiring context-dependent and system-level rather than product-focused solutions.[7]

This article describes a specific example that illustrates some of the challenges in applying existing guidelines for mitigating algorithmic bias in a ML tool for real-world clinical decision making by physicians and patients. We then discuss the existing legal regulation of AI/ML racial bias and future directions.

II. A Case Study Highlighting Regulatory Gaps and Difficult Normative Questions

Our team developed a decision support framework for patients with severe heart failure that includes a prognostic ML algorithm employing Bayesian probability models to calculate personalized estimates for patients about their likely outcomes after receiving a left ventricular-assist device (LVAD). An LVAD is a mechanical circulatory assist device that helps to propel blood from the heart's left ventricle to the body to prolong survival and improve functioning and quality of life for select patients. By providing personalized probabilities of survival and adverse events, our risk prediction tool is designed to help clinicians identify suitable candidates for LVAD and help inform patients about treatment decisions. Our efforts to identify potential for racial bias in the tool's algorithms identified practical challenges regarding algorithmic bias that other developers may also face. These challenges, in turn, highlight gaps in existing regulations to guide developers and mitigate algorithmic bias.

Addressing Technical Sources of Bias

DEMOGRAPHIC IMBALANCE IN THE DATASET

Given the importance of data quality for mitigating bias, our first step was to examine whether our algorithms' training dataset contains any potential for racial bias due to statistical disproportions by race in the data. The algorithm is trained on data from the Interagency Registry for Mechanically Assisted Circulatory Support (INTERMACS) representing outcomes of *all* patients in the U.S. who received an LVAD since 2006. The latest INTERMACS report (2020) shows that, despite recent increases in LVAD implantation among Black patients, Black patients still constitute only 27% of LVAD patients. Latinx populations make up about 8%, and Asian patients under 3%[8] of LVAD

Table 1[9]

Baseline Characteristics of Patients on Isolated Left Ventricular Assist Device Support*

Patient Characteristics	All Patients (N = 25,551)	2010-2014 Era (n = 10,944)	2015-2019 Era (n = 14,607)	*P* Value*
Demographics				
Age at implant, y	57.1 ± 13.0	57.3 ± 12.9	57.0 ± 13.0	.06
Female sex	5496 (21.5)	2338 (21.4)	3158 (21.6)	.003
Race				<.0001
White	16,753 (65.6)	7540 (68.9)	9213 (63.1)	
Black	6417 (25.1)	2505 (22.9)	3912 (26.8)	
Other	2381 (9.3)	899 (8.2)	1482 (10.1)	
Body mass index, kg/m^2	28.7 ± 7.2	28.7 ± 6.7	28.6 ± 7.5	.4
Medical history				
Severe diabetes	2215 (8.8)	828 (7.8)	1387 (9.5)	<.0001
Dialysis	686 (2.7)	268 (2.4)	418 (2.9)	.04
Current ICD	19,989 (78.7)	8884 (81.7)	11,105 (76.5)	<.0001
History of cardiac surgery	8281 (32.4)	3907 (35.7)	4374 (29.9)	<.0001
Current smoker	1455 (5.8)	569 (5.4)	886 (6.1)	.02

*Taken from the Society of Thoracic Surgeons INTERMACS 2020 Annual Report [30]

The Journal of Law, Medicine & Ethics, 50 (2022): 92-100. © 2022 The Author(s)

https://doi.org/10.1017/jme.2022.13 Published online by Cambridge University Press

patients, while White patients make up 63%. Despite mirroring imbalances in the general population, this disproportion creates potential for racial bias in the relevance and accuracy of predictions for minorities.

COMORBIDITY IMBALANCE

Other covariates such as comorbidity may complicate or bias estimates. Black LVAD patients experience higher prevalence of comorbidity with chronic renal failure, which strongly predicts in-hospital mortality after LVAD.[10] However, on the whole, Black patients have lower in-hospital mortality compared with White patients after adjusting for age and comorbidity.[11] Similarly, for patients receiving LVAD as a bridge to transplant, certain comorbidities were identified[12] as independent risk factors that mediate an increased incidence of graft failure among Black patients compared with White patients.[13] However, racial differences in survival disappear when controlling for these independent risk factors.[14] Another recent study[15] showed that among patients who received an LVAD as a bridge to transplantation, African and Hispanic patients had a higher risk of death while waiting for transplant or of being delisted due to deteriorating health status compared to Whites, even after adjusting for covariates. However, *after* transplantation, 5-year survival curves did not differ by race/ethnicity. Other comorbidities such as obesity[16] and body mass index that are greater on average among Black and Latinx patients,[17] respectively, may not directly impact survival but can be associated with adverse events. For example, a meta-analysis found that obese patients with LVAD had significantly greater risk of device-related infections, right heart failure, and pump thrombosis compared with non-obese patients.[18] However, after adjusting for clinical comorbidities, Black patients do not appear to experience significantly more adverse events and no differences in overall in-hospital mortality after implant[19] — findings that have been corroborated by other studies.[20]

Our own statistical analysis indicated that race was *not* a significant predictor of outcomes and found no identifiable causal mechanisms for racial disparities in outcomes among patients selected for LVAD. Our analysis shows that the average accuracy of predictions for survival and adverse events is not affected by the inclusion of race in the model (Table 2). Further, despite non-significant dependencies of some variables on race, such as Black patients more likely to be working for income (potentially also reflecting younger average age at implant), no other variables were significantly associated with race that might indicate the existence of proxy variables influencing our model's results. Out of a list of 82 variables ranked by information gain brought to the survival model, race ranks among the least informative at 61.

Together, the studies cited above do not offer a clear understanding of whether and how race influences LVAD outcomes, including bridge to transplant. Our attempt to dissect racial bias in our algorithm revealed no significant role of race as a director predictor; however, certain findings from the literature described above suggest an indirect but important relationship between race and LVAD outcomes.

While our findings corroborate the larger literature described above suggesting little direct evidence for racial and ethnic disparities in LVAD outcomes,[21] we nevertheless note that a majority of these did detect racial differences before adjusting for clinical comorbidities and health status. Further, that Black patients supported by LVAD on average need the device at a much younger age[22] and are significantly more likely to die or be delisted while waiting for transplant due to worsening health status[23] suggest that Black patients supported by LVAD suffer disproportionately from patterned health disparities unrelated to aging. New data are emerging that suggest Black patients receiving the newest Heartmate 3 device experience a higher morbidity burden and smaller gains in functional capacity and quality of life when compared with White patients receiving the same device.[24]

Together, the studies cited above do not offer a clear understanding of whether and how race influences LVAD outcomes, including bridge to transplant. Our attempt to dissect racial bias in our algorithm revealed no significant role of race as a direct predictor; however, certain findings from the literature described above suggest an indirect but important relationship between race and LVAD outcomes. As the next section

The Journal of Law, Medicine & Ethics, 50 (2022): 92-100. © 2022 The Author(s)

https://doi.org/10.1017/jme.2022.13 Published online by Cambridge University Press

shows, the seeming race-neutrality of the algorithm may hide the actual racial injustice in the underlying LVAD patient population.

Potential Societal and Systemic Sources of Bias

What complicates understanding of racial disparities in LVAD outcomes is the fact that the data sets that inform these understandings are comprised only of LVAD patients. That may sound like a tautology, because one might ask, "Who else would be in the dataset?"! But what matters is that individuals of all races do not have an equal probability of being included in these datasets because there is unequal access to LVAD therapy. As Ueyama et al.[25] write, the absence of substantial differences in survival outcomes in LVAD patients is likely due to "strict patient selection, preoperative work-up, and ... assessment of patient compliance and rigorous comorbid disease control before and after LVAD implantation." In other words, clinicians use strict selection criteria to determine who is most likely to benefit clinically from an LVAD, taking not only clinical but also non-clinical factors into account.[26] Black individuals may be subject to selection bias due to lack of access to local cardiologists whom they trust, lack of established referral patterns from their primary care physicians to experienced LVAD clinics, or lack of vigorous, effective programs for addressing challenges due to compliance, social support or RV failure.[27] Black patients may further experience selection bias on the basis of baseline health status and comorbidities and be more often deemed ineligible candidates for advanced cardiac therapies,[28] resulting in lower rates of referral to cardiology care despite demonstrating equal need. Further, candidacy for LVAD is influenced not only by clinical but also psychosocial and socioeconomic factors, particularly evaluations of social support,[29] which involve a high degree of interpretation and lack standardization and transparency in terms of how they influence clinicians' decision-making.[30] A recent study[31] found that, despite having similar scores of overall social support, Black patients were less likely to have a spouse as primary support, and that the absence of a spouse negatively impacted their eligibility for LVAD compared to white patients matched for the same criteria.

In the context of available literature on LVAD candidacy and outcomes, and a technical examination of race as a predictor variable in our own data set, our critical takeaway is that the apparent absence of racial differences may falsely convey equity in *outcomes* while masking socioeconomic inequities in *access or distribution*. Our calculator — and any calculator based on information from an LVAD-only dataset —will inevitably be less useful, accurate, and relevant for individuals who are excluded from access to LVAD and thus from an algorithms' training data set(s). The origin of potential bias appears to be at the point of access and perhaps attributed to more antecedent socioeconomic factors generating disparities in health status at the time of initial evaluation. Few guidelines exist to help developers mitigate racial bias whose origins are upstream of data collection and linked to systemic factors.

The normative analysis of such disparities is also more complicated than the current discourse on racial algorithmic bias might suggest. In the case of LVAD patients, does the bias "belong" to the algorithm that correctly reflects the "true north" of reduced access or distribution by Black patients, or does it "belong" to the health care system that produces the reduced access or distribution, or both (and to what extent)? To put it very practically: if the best dataset with which to build an algorithm is likely the LVAD one — accurately reflecting the data of LVAD patients but neglecting to account for systemic injustice in health care distribution and access influencing who becomes an LVAD patient — what should the developer do? Is it appropriate, desirable or necessary to try to "correct" the data set to try to reduce the systemic injustice in health care distribution? We put "correct" in quotes because such manipulations may make the algorithm *less accurate* in terms of reflecting the actual data, but perhaps *more just* in terms of the outcome it produces. Indeed, the situation becomes more complicated when one recognizes that the distributional consequences are not always so clear cut. For example, one could imagine situations where the accuracy loss hurts (leads to less accurate predictions for) some vulnerable racial or other groups (say, Latinx patients) while the distributional gain benefits a different group (say, Black patients). How much of this can developers determine ex ante, and how well equipped are they to make these determinations?

In the next Part we review some of the existing legal guidance and proposed initiatives to address racial algorithmic bias; but to forefront one of our conclusions: they may not adequately address the kind of problem we have flagged using the LVAD example.

III. Existing and Proposed Initiatives to Address Racial Algorithmic Bias through Regulation

A. Current Hard Law

Precious little current "hard law" in the U.S. speaks to racial bias in algorithms, let alone the more subtle versions of the problem we have identified in the LVAD case. Under Section 1557 of the U.S. Affordable Care

The Journal of Law, Medicine & Ethics, 50 (2022): 92-100. © 2022 The Author(s)

Act, health entities are prohibited from discriminating against patients on the basis of race, ethnicity, national origin, sex, age, or disability. However, some courts have interpreted the provision to prohibit disparate impact theories of liability, a significant problem for winning a lawsuit in this context.[32]

Table 2
Comparison of Prediction Accuracy In/Excluding Race

Outcomes:	Prediction Accuracy Including Race	Excluding Race
Death (average)	0.79	0.79
Death 30 day (1mo)	0.94	0.93
Death 90 day (3 mo)	0.89	0.89
Death 1yr	0.83	0.83
Death 2yr	0.77	0.77
Death 3yr	0.73	0.73
Death 4yr	0.71	0.71
Death 5yr (if avail)	0.70	0.70
EVENT: Infection	0.85	0.85
30 days	0.90	0.90
90 days	0.88	0.88
1 year	0.86	0.86
2 years	0.84	0.84
3 years	0.84	0.84
4 years	0.83	0.83
5 years	0.83	0.83
EVENT: Stroke	0.94	0.94
30 days	0.97	0.97
90 days	0.96	0.96
1 year	0.94	0.95
2 years	0.93	0.93
3 years	0.92	0.92
4 years	0.92	0.92
5 years	0.92	0.92
EVENT: Bleeding	0.65	0.65
30 days	0.76	0.77
90 days	0.70	0.70
1 year	0.63	0.63
2 years	0.61	0.61
3 years	0.60	0.60
4 years	0.63	0.63
5 years	0.60	0.60

Tort law may be similarly unavailing. Andrew Selbst[33] has argued that ordinary negligence law, as currently constituted, will have a hard time grounding liability for cases where "medical AI/ML can reproduce or potentially exacerbate human biases" for a number of reasons, including because negligence law is keyed to the notion of foreseeability; however, he points out that for medicine the "unforeseeable nature of AI errors risks being the exception that swallows the rule." How foreseeable these errors are may change as insights into AI performance increase over time.

U.S. Food and Drug Administration (FDA) regulation may solve the issue but is limited to AI/ML systems that are classified as medical devices under Section 201(h) of the Federal Food, Drug, and Cosmetic Act. However, many AI/ML systems never undergo FDA review because of how Congress has structured FDA's authority in this area and the way that FDA has constructed the category of Software as Medical Device.[34] As we discuss more fully below, even if FDA has jurisdiction, the current regulatory approach is insufficient to address racial bias.[35]

In April 2021, the Federal Trade Commission[36] signaled in a blog post an intent to consider the "sale or use of, for example, racially based algorithms" as potentially violating Section 5 of the FTCA Act prohibiting "unfair or deceptive practices," using healthcare ML examples as motivating examples. This may suggest a new avenue of legal liability, but it is too soon to tell whether the theories of liability the FTC will pursue under the Act will capture the kind of case we contemplate above.

State insurance regulators may also have a role to play in policing bias in some ML applications in medicine; New York announced it would investigate the users of an algorithm that helped payers and health systems target patients for high-risk care management that was found by a paper published in *Science* to treat Black patients worse than white patients with the same health status.[37] It is unclear if this was an unusual case or whether there will be robust interest in the issue by insurance regulators; but of course their jurisdiction is limited to bias related to insurance coverage/payment.

Overall, our assessment is that while there is some "hard law," it does not currently exert much of a push on AI/ML developers to detect and avoid racial bias in their algorithms. Moreover, what hard law there is may not clearly apply to more subtle forms of biased algorithms like the LVAD case.

B. Proposed New Directions

The US[38] and the European Union (EU)[39] have independently proposed initiatives for regulatory guidelines to ensure diversity, nondiscrimination, fairness and equity in ML from design to execution. Numerous other governments,[40] private companies and institutions[41] and non-governmental organizations[42] have similarly proposed high level standards to improve algorithmic fairness and accuracy, many revolving around improving data quality. In April 2021, the European Commission issued a Proposal for Regulation

The Journal of Law, Medicine & Ethics, 50 (2022): 92-100.

https://doi.org/10.1017/jme.2022.13 Published online by Cambridge University Press

Laying Down Harmonized Rules on AI (AI Act)[43] that would require that "high-risk AI systems which make use of techniques involving the training of models" use "high quality" data sets, defined in Recital 44 as data that are "sufficiently relevant, representative," and having "the appropriate statistical properties… as regards the persons or groups of persons on which the high-risk AI system is intended." As applied to our case, it would seem as though this standard may be met — the dataset includes all patients who receive an LVAD and thus appears representative; however, the problem is that who becomes an LVAD patient is itself not race-neutral.

The proposal further states that "… providers should be able to process also special categories of personal data … to ensure the bias monitoring, detection and correction in relation to high-risk AI systems." Further, they should be "complete in view of the intended purpose of the system." However, the AI Act provides little guidance about steps to take in cases where relevant data is missing that cannot be easily generated or synthesized, as with the LVAD case.

A number of scholars[44] have called for a focus on larger societal and systemic sources of bias. Gerke et al.[45] have pointed out that because the performance of an algorithm depends on how it is implemented in the broader healthcare system, ideally regulators' focus should take a more systemic rather than a product-focused approach. In some cases, this may require developers and regulators to "collect data on a myriad of information beyond its current regulatory gaze and perhaps even beyond its legal mandate, requiring additional statutory authority …" The authors argue that while this expanded gaze seems good in theory, it raises difficult questions about how far upstream developers and regulators need to go to ensure data quality, as well as how far downstream to look to account for context-dependent outcomes.[46]

Specific data quality criteria may be applicable for those specifically intended to support decisions about treatment, requiring developers and regulators to engage in even more due diligence. Evidence-based criteria outlined in the International Patient Decision Aid Standards (IPDAS)[47] state that decision aids should provide patients with a side-by-side view of information relevant to their treatment options. How might this latter approach work as applied to the LVAD case? It would lead us to provide personalized risk estimates for outcomes *with* versus *without* LVAD to transparently convey the (usually positive) role that LVAD therapy can play in ushering patients towards their desired treatment goals.

But notice that this does not solve the problem we have identified. A version of our algorithm trained only on INTERMACS would only account for patients *already selected* to receive an LVAD, not for those who *might* be suitable candidates (patients under evaluation for candidacy). This latter subset of the patient population (and/or their providers) is precisely the intended set of end-users for our algorithm.

An ideal regulatory requirement for decision support algorithms would thus be to expand the data set to ensure they are representative of the *full* range of patients being referred and evaluated (and perhaps never achieving candidacy) for a treatment, not just people who received the treatment. Doing so requires that developers collect further data, an oft-cited strategy for improving data quality.[48] However, in our field — and in many other fields of study in which developers may seek additional training data for their algorithms — these data are proprietary and require additional partnerships (e.g., to create standardized registries) and/or budgeting for data acquisition expenditures. While in some cases data acquisition needs are evident during the project design phase, in others they are realized only after a critical interrogation of the training dataset — precisely the kind of context-dependent and iterative validation testing that Gerke et al. advocate for. The associated expenditures necessary to realize this goal are a critical challenge to mitigating bias within development projects with time and budgetary constraints. Further, our larger point is that these data do not exist to be collected, for the reasons described above.

Price[49] has discussed the potential use of labeling as a solution to mitigating contextual bias in AI/ML in health care (defined as mismatches between a product's design and the characteristics, knowledge or cultural values of its users[50] or from structural constraints

Standards should be developed that clearly articulate what is required from AI/ML manufacturers to effectively mitigate racial bias. Labeling as such without broader regulatory reforms does not address the underlying justice-based concern that the algorithm simply does not apply very well to certain groups.

The Journal of Law, Medicine & Ethics, 50 (2022): 92-100. © 2022 The Author(s)

and biases that shape how it is applied in practice[51]), of which racial bias could be viewed as a subset. Could a possible solution to the problem we have identified be for FDA to use the label of AI/ML to better describe for whom it works, e.g., "not cleared for Black women under the age of 40"? This may sound desirable because it both provides clarification to medical users and patients and may provide incentives to developers to "do the work" to ensure that AI/ML works for a broader set of patients or to have its limitations explicitly stated. While labeling is essential to create transparency, it must be coupled with a robust regulatory approach to AI/ML. In other words, standards should be developed that clearly articulate what is required from AI/ML manufacturers to effectively mitigate racial bias. Labeling as such without broader regulatory reforms does not address the underlying justice-based concern that the algorithm simply does not apply very well to certain groups.

We do not purport to be exhaustive in our discussion of all the sources of U.S. law or major proposals available to try to regulate racial bias in AI/ML. But we are far from optimistic that the current regulatory options are likely to make significant headway on the general problem of racial bias in AI/ML. While some, like possible future FTC action, remain undefined, we think most such initiatives are particularly unlikely to exert pressure on cases like the LVAD example, where the source of the problem lies in the upstream patterns influencing which patients get LVAD and thus included in relevant training data sets.

IV. Bringing Together Our Case Study and the Regulatory Space

The authors are mostly bioethicists with grant funding to spend significant time and resources determining how our algorithm should handle issues of race, and worked with a team of algorithm developers committed to close scrutiny on this issue. However, neither will be true for many working in the clinical AI/ML space. As described above, there are no legal obligations at the moment that would have required our team or others like it to undertake the analysis we have, and no data on how many other ML systems in clinical use have undergone a similar pre-implementation analysis. If we found good reason that our algorithm should not be used for certain racial or other groups, there are currently no obligations to limit the label of the algorithm as such, and there does not seem to be a standardized way to explain or set out such limitations to practitioners or the general public. The use of "datasheets for datasets" that Gebru et al.[52] have proposed could potentially outline limitations of algorithmic performance for certain population subgroups but may be too technical for easy interpretation in clinical settings.

It would be nice to be able to close an article like this with some easy solutions. We are not so naïve. But we do think some directions are more promising than others. The most promising solutions seem to us to involve collecting more data, especially those related to larger systemic influences on health access and outcomes. However, the challenges in defining the scope and navigating availability of these data pose roadblocks for both developers and regulators seeking to mitigate algorithmic racial bias. Grant funding for AI/ML development in healthcare should integrate funding to examine the underlying data set for racial gaps and mitigation strategies. The notion of an algorithmic impact assessment is gaining currency as an important step in the building and deployment of algorithms more generally. Tools and guidelines for conducting these assessments of algorithmic bias in health already exist. For example, Chicago Booth's Algorithmic Bias Playbook developed by leaders in the field is an exemplary step-by-step approach to considering and rectifying such biases.[53] Hospital systems, payers, and others could voluntarily commit to only implementing AI/ML that has not only gone through such an assessment but also to publishing the results (and perhaps in some instances, an auditing of the process). Furthermore, journal editors could publish only AI algorithms that included a rich description of the derivation and validation data sets and how they differed from likely target populations, as well as a standardized assessment of algorithmic bias. Such guidelines for reporting on race were recently outlined in the *Journal of the American Medical Association*.[54]

These steps would be desirable, perhaps necessary, but they are not sufficient. Our efforts to examine bias potential in our own tool draws attention to hard-to-address upstream challenges in the fair allocation of LVAD and the societal and economic forces contributing to racial disparities in eligibility. Our findings provide cautionary evidence that data-improvement strategies are only partial and inadequate solutions. Correcting for underestimation and other imbalances in data sets cannot correct for what Cunningham et al.[55] have termed "negative legacy" in datasets, or scenarios where there are undesirable patterns in the historical factors shaping the datasets on which algorithms are trained. However, the process that developers engage in in good faith to evaluate racial bias in their algorithms can help to point out where the "real" targets for bias mitigation may lie (in the case of LVAD, at access points of care) and provide leverage in calling for greater awareness of and policy solutions to enhance inclusivity in these contexts. Without sys-

The Journal of Law, Medicine & Ethics, 50 (2022): 92-100. © 2022 The Author(s)

tematic investigation into the impacts of race as a predictor of outcomes, as well as upstream factors (e.g., social and structural determinants of health) that predict racial patterns in candidacy for LVAD, we can only speculate as to how greater inclusivity might promote algorithmic fairness. Our analysis highlights that even where algorithms themselves may not be biased, bias may nevertheless exist in the broader social and structural factors that determine inclusion in the training data of the algorithms.

We recommend that developers seeking to mitigate bias in ML use their algorithms as leverage to call upon stakeholders (physicians, societies, institutions, vendors of EHRs, research funders, and AI developers, etc.) who are responsible for generating relevant data sets to make a concerted effort to document race and associated variables (e.g., potential proxies and dependent variables) to enable systematic inquiries into sources of potential racial bias. Without such documentation, developers of ML algorithms are left to speculate on how best to balance data sets and where to target bias mitigation strategies. We believe an effective algorithm based on a critically examined dataset can serve as a broader call to researchers for a more thorough documentation of race-related variables and can also provide a tool to promote objectivity and fairness in clinical decision making.

Note

This project was supported by grant number R01HS027784 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality. I.G.C. and S.G. were supported by a grant from the Collaborative Research Program for Biomedical Innovation Law, a scientifically independent collaborative research program supported by a Novo Nordisk Foundation grant (NNF17SA0027784). I.G.C. reports, during the conduct of the study; personal fees from Otsuka, personal fees from Illumina, other from Dawnlight, personal fees from Bayer, and grants from Moore Foundation, outside the submitted work. B.L. reports personal fees from Takeda Pharmaceuticals, outside the submitted work. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

References

1. B. Friedman and H. Nissenbaum, "Bias in Computer Systems," 1996, *available at* <https://nissenbaum.tech.cornell.edu/papers/Bias%20in%20Computer%20Systems.pdf> (last visited December 6, 2021).
2. A. Caliskan, J.J. Bryson, and A. Narayanan, "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases," *Science* 356, no. 6334 (2017): 183-186; A. Hadhazy, "Biased Bots: Artificial-Intelligence Systems Echo Human Prejudices," *Princeton University* (April 2017): at 18; T. McSweeny, "Psychographics, Predictive Analytics, Artificial Intelligence, & Bots: Is The FTC Keeping Pace?" *Georgetown Law Technology Review* 2 (2018): 514, 516, 530; J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," paper presented at Conference on Fairness, Accountability and Transparency, 2018; I.A. Hamilton, "Why It's Totally Unsurprising That Amazon's Recruitment AI Was Biased against Women," *Business Insider*, 2018, *available at* <https://www.businessinsider.com/amazon-ai-biased-against-women-no-surprise-sandra-wachter-2018-10> (last visited December 6, 2021); T. Brennan, W. Dieterich, B. Ehret, "Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System," *Criminal Justice and Behavior* 36, no. 1 (2009): 21-40; J. Dressel and H. Farid. "The Accuracy, Fairness, and Limits of Predicting Recidivism," *Science Advances* 4, no. 1 (2018): eaao5580.
3. T. Ryan-Mosely, "The New Lawsuit That Shows Facial Recogntion Is Officially a Civil Rights Issue," *MIT Technology Review* (2021).
4. C. Ross, "As the FDA Clears a Flood of AI Tools, Missing Data Raise Troubling Questions on Safety and Fairness, " *STAT*, February 3, 2021, *available at* <https://www.statnews.com/2021/02/03/fda-clearances-artificial-intelligence-data/> (last visited December 6, 2021).
5. Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science* 366, no. 6464 (2019): 447-453.
6. European Commission, "Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts," 2021, *available at* <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206> (last visited December 6, 2021).
7. D.A. Ansell and E.K. McDonald, "Bias, Black Lives, and Academic Medicine," *New England Journal of Medicine* 372, no. 12 (2015): 1087-1089; I. Price, W. Nicholson, "Medical AI and Contextual Bias," 2019.
8. J. van Meeteren, S. Maltais, S.M. Dunlay, et al., "A Multi-Institutional Outcome Analysis of Patients Undergoing Left Ventricular Assist Device Implantation Stratified by Sex and Race," *Journal of Heart and Lung Transplantation* 36, no. 1 (2017): 64-70.
9. E.J. Molina, P. Shah, M.S. Kiernan, et al., "The Society of Thoracic Surgeons Intermacs 2020 Annual Report," *Annals of Thoracic Surgery* 111, no. 3 (2021): 778-792.
10. P. Goyal, T. Paul, Z.I. Almarzooq, et al., "Sex- and Race- Related Differences in Characteristics and Outcomes of Hospitalizations for Heart Failure with Preserved Ejection Fraction," *Journal of the American Heart Association* 6, no. 4 (2017): e003330.
11. *Id.*
12. H. Ueyama, A. Malik, T. Kuno, et al., "Racial Disparities in Hospital Outcomes after Left Ventricular Assist Device Implantation," *Journal of Cardiac Surgery* 35, no. 10 (2020): 2633-2639.
13. C. Lui, C.D. Fraser III, X. Zhou, et al., "Racial Disparities in Patients Bridged to Heart Transplantation with Left Ventricular Assist Devices," *Annals of Thoracic Surgery* 108, no. 4 (2019): 1122-1126.
14. Ueyma et al., *supra* note 12.
15. A.K. Okoh, M. Selevanny, S. Singh, et al., "Racial Disparities and Outcomes of Left Ventricular Assist Device Implantation as a Bridge to Heart Transplantation," *ESC Heart Failure* 7, no. 5 (2020): 2744-2751.
16. A. Jaiswal, L.K. Truby, A. Chichra, et al., "Impact of Obesity on Ventricular Assist Device Outcomes," *Journal of Cardiac Failure* 26, no. 4 (2020): 287-297.
17. Okoh et al., *supra* note 15.
18. M.S. Khan, M. Yuzefpolskaya, M.M. Memon, et al., "Outcomes Associated with Obesity in Patients Undergoing Left Ventricular Assist Device Implantation: A Systematic Review and Meta-analysis," *Asaio Journal* 66, no. 4 (2020): 401-408.
19. Ueyma et al., *supra* note 12.
20. van Meeteren et al., *supra* note 8; A. Itoh, "Impact of Age, Sex, Therapeutic Intent, Race and Severity of Advanced Heart Failure on Short-Term Principal Outcomes in the MOMENTUM

The Journal of Law, Medicine & Ethics, 50 (2022): 92-100. © 2022 The Author(s)

3 Trial," *Journal of Heart and Lung Transplantation* 37, no. 1 (2018): 7-14.
21. *Id.*; Ueyma et al., *supra* note 12.
22. E.J.A. Bowles, R. Wellman, H.S. Feigelson, et al., "Risk of Heart Failure in Breast Cancer Patients after Anthracycline and Trastuzumab Treatment: A Retrospective Cohort Study," *Journal of the National Cancer Institute* 104, no. 17 (2012): 1293-1305.
23. Okoh et al., *supra* note 15.
24. F.H. Sheikh, A.K. Ravichandran, D.J. Goldstein, et al., "Impact of Race on Clinical Outcomes after Implantation with a Fully Magnetically Levitated Left Ventricular Assist Device: An Analysis From the MOMENTUM 3 Trial," *Circulation: Heart Failure* 14, no. 10 (2021): e008360.
25. Ueyma et al., *supra* note 12.
26. A. Tsiouris, R.J. Brewer, J. Borgi, H. Nemeh, G. Paone, and J.A. Morgan, "Continuous-Flow Left Ventricular Assist Device Implantation as a Bridge to Transplantation or Destination Therapy: Racial Disparities in Outcomes," *Journal of Heart and Lung Transplantation* 32, no. 3 (2013): 299-304; X. Wang, A.A. Luke, J.M. Vader, T.M. Maddox, K.E. Joynt Maddox , "Disparities and Impact of Medicaid Expansion on Left Ventricular Assist Device Implantation and Outcomes," *Circulation: Cardiovascular Quality and Outcomes* 13, no. 6 (2020): e006284.
27. Okoh et al., *supra* note 15; Tsiouris et al., *supra* note 26.
28. Ueyma et al., *supra* note 12.
29. *Id.*; K. Breathett, E. Yee, N. Pool, et al., "Does Race Influence Decision Making for Advanced Heart Failure Therapies?" *Journal of the American Heart Association* 8, no. 22 (2019): e013592.
30. Q.M. Bui, L.A. Allen, L. LeMond, M. Brambatti, E. Adler, "Psychosocial Evaluation of Candidates for Heart Transplant and Ventricular Assist Devices: Beyond the Current Consensus," *Circulation: Heart Failure* 12, no. 7 (2019): e006058.
31. R.S. Steinberg, A. Nayak, T. Dong, A.A. Morris, "Primary Caregiver Relationships for Advanced Heart Failure Therapy Differ Based on Sex and Race and Predict Eligibility," *Journal of Cardiac Failure* 26, no. 10 (2020): S10.
32. S. Takshi, "Unexpected Inequality: Disparate-Impact from Artificial Intelligence in Healthcare Decisions," *Journal of Law and Health* 34, no. 2 (2021): 215.
33. A.D. Selbst, "Negligence and AI/MLs Human Users," *Boston University Law Review* 100, no. 1315 (2020): 1354-1360.
34. W.N. Price II, R. Sachs, and R.S. Eisenberg, "New Innovation Models in Medical AI," *Washington University Law Review* (forthcoming 2022); S. Gerke, B. Babic, T. Evgeniou, and I.G. Cohen, "The Need for a System View to Regulate Artificial Intelligence/Machine Learning-Based Software as Medical Device," *npj: Digital Medicine* 3, no. 1 (2020): 53.
35. Ross, *supra* note 4.
36. Federal Trade Commission. Aiming for truth, fairness, and equity in your company's use of AI. 2021.
37. M.E.A.W. Mathews, "New York Regulator Probes UnitedHealth Algorithm for Racial Bias," *Wall Street Journal*, Octover 26, 2019.
38. White House, *Executive Order on Maintaining American Leadership in Artificial Intelligence*, 2019.
39. I.G. Cohen, T. Evgeniou, S. Gerke, T. Minssen, "The European Artificial Intelligence Strategy: Implications and Challenges for Digital Health," *The Lancet Digital Health* 2, no. 7 (2020): e376-e379; AI HLEG, *Policy and Investment Recommendations for Trustworthy AI*, 2019; European Commission, *White Paper on Artificial Intelligence—A European Approach to Excellence and Trust*, 2020.
40. SPDP Commission, *Infocomm Media Development Agency Proposed Model AI Governance Framework*, 2020.
41. Microsoft, *Microsoft AI principles*, 2021.
42. Partnership on AI, *Tenets*, June 15, 2021.
43. European Commision, *supra* note 6.
44. See *supra* note 7; also Gerke et al., *supra* note 34.
45. Gerke et al., *supra* note 34.
46. A. Nayak, A.J. Hicks, and A.A. Morris , "Understanding the Complexity of Heart Failure Risk and Treatment in Black Patients," *Circulation: Heart Failure* 13, no. 8 (2020): e007264.
47. K.R. Sepucha, P. Abhyankar, A.S. Hoffman, et al., "Standards for Universal Reporting of Patient Decision Aid Evaluation Studies: The Development of SUNDAE Checklist," *BMJ Quality & Safety* 27, no. 5 (2018): 380-388.
48. N.T. Lee, P. Resnick, and G. Barton, "Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms," Brookings Institute, 2019.
49. Price et al., *supra* note 7.
50. Friedman and Nissenbaum, *supra* note 1.
51. Lui, *supra* note 7.
52. T. Gebru, J. Morgenstern, B. Vecchione, et al. "Datasheets for Datasets," 2018, Cornell University, *available at* <https://arxiv.org/abs/1803.09010> (last visited December 6, 2021).
53. E. Moss, E.A. Watkins, S. Singh, M.C. Elish, and J. Metcalf, "Assembling Accountability: Algorithmic Impact Assessment for the Public Interest," 2021, *available at* <https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/> (last visited December 6, 2021).
54. A. Flanagin, T. Frey, S.L. Christiansen, "Committee AMoS. Updated Guidance on the Reporting of Race and Ethnicity in Medical and Science Journals," *JAMA* 326, no. 7 (2021): 621-627.
55. P. Cunningham and S.J. Delany, "Algorithmic Bias and Regularisation in Machine Learning," 2020, Cornell University, *available at* <https://arxiv.org/abs/2005.09052arXiv preprint arXiv:2005.09052> (last visited December 6, 2021).

The Journal of Law, Medicine & Ethics, 50 (2022): 92-100. © 2022 The Author(s)