Biostatistics (2020) **21**, 2, *pp.* 339–344
doi:10.1093/biostatistics/kxz040
Advance Access publication on November 19, 2019

Teaching yourself about structural racism will improve your machine learning

WHITNEY R. ROBINSON*

Department of Epidemiology, UNC Gillings School of Global Public Health, University of North Carolina at Chapel Hill; Carolina Population Center, University of North Carolina at Chapel Hill, CB #7345 McGavran-Greenberg, Chapel Hill, NC 27599-7435, USA

whitney_robinson@unc.edu

AUDREY RENSON

Department of Epidemiology, University of North Carolina at Chapel Hill, 135 Dauer Dr., Chapel Hill, NC 27599, USA

ASHLEY I. NAIMI

Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh, 130 DeSoto Street, 5131 Public Health Building, Pittsburgh, PA 15261-3100, USA

Summary

In this commentary, we put forth the following argument: Anyone conducting machine learning in a health-related domain should educate themselves about structural racism. We argue that structural racism is a critical body of knowledge needed for generalizability in almost all domains of health research.

Keywords: Causal inference; Directed acyclic graphs; Machine learning; Structural racism.

1. Motivation

In this commentary, we put forth the following argument: anyone conducting machine learning in a health-related domain should educate themselves about structural racism. As Domingos and others have argued, "Every learner must embody some knowledge or assumptions beyond the data it is given in order to generalize beyond it" (Domingos, 2012). We argue here that structural racism is a critical body of knowledge needed for generalizability in almost all domains of health research. We believe that this is especially true when inference relies on algorithms ("learners") to choose statistical models.

We make the recommendation to incorporate structural racism based on our experiences as epidemiologists. Epidemiologists are fundamentally interested in causing changes that result in improved individual and population health and that reduce health disparities (Glass *and others*, 2013). Quantifying statistical associations are central to these objectives, yet they are not sufficient. As we observe the challenges facing applied machine learning, we see echoes of debates that epidemiology has encountered. For instance, epidemiology's disease screening literature has imparted the intuition that even highly sensitive and specific

*To whom correspondence should be addressed.

screening algorithms can produce more false positives (analogous to the high "false discovery rate" in machine learning terminology) than true positives when an outcome is rare.

In particular, we are inspired by our field's 1980s- and 1990s-era debates about "black box epidemiology" (Weed, 1998). In many respects, this debate mirrors debates in machine learning about the trade-offs between improved prediction versus greater model interpretability (Seligman *and others*, 2018). The earlier epidemiology debate contrasted the use of multivariable-adjusted regression models to identify behavioral risk factors for cancer incidence to target in prevention efforts ("black box") versus research elucidating biological pathways of cancer development, particularly at the level of molecular biology ("mechanistic") (Weed, 1998). However, the "mechanistic" side's focus on molecular mechanisms ignored all the parts of the causal structure that were "above" the molecular level. A key insight of this debate was that, even a mechanistically oriented research orientation has its blind spots. Specifically, the integration of sociopolitical forces with a consideration of biology and behavior was missing in early debates in the field (Weed, 1998). The need to integrate factors from across the full breadth of the causal structure is likely even more crucial when making causal inference.

2. WHY STRUCTURAL RACISM IS CRITICAL KNOWLEDGE

Structural racism refers to "the totality of ways in which societies foster [racial] discrimination, via mutually reinforcing [inequitable] systems...(e.g., in housing, education, employment, earnings, benefits, credit, media, health care, criminal justice, etc.) that in turn reinforce discriminatory beliefs, values, and distribution of resources," reflected in history, culture, and interconnected institutions (Bailey *and others*, 2017). Below we present two reasons why expanding one's knowledge base of structural racism will improve the quality of work produced by machine learning applications to health.

First, even when racial categorization is not a topic of interest in an analysis, structural racism will shape associations of health-related processes. For instance, in a U.S. context, the sociodemographic variable White/Black race is a variable that is frequently available in health-related datasets and often highly predictive of health outcomes. For instance, Seligman *and others* (2018) attempted to predict body mass index (BMI), blood pressure, and waist circumference in a population of 15,784 longitudinally assessed participants using four machine learning methods (linear regression, penalized regressions, random forests, and neural networks). With access to 458 variables, "Black/African-American Race" was one of the top five variables selected by all four machine learning models Seligman *and others* (2018). In non-U.S. contexts, other markers of social stratification, such as caste, ethnicity, religion, social class, country of birth, home village, gender, or sexual identity, will be similarly predictive in health processes that involves human agency and social organization. An analyst striving to produce the most predictive and illuminating models ignores structural racism (and other axes of inequality) at his or her own peril. Ignoring structural racism is a decision to ignore structures that give rise to many and varied associations with health.

A second motivation for educating oneself about structural racism is the common occurrence of "algorithmic bias." When "race-neutral" approaches are employed in model development, prediction will tend to be poorer for racial minority populations. Greater error rates, and even failure of algorithms to perform at all, for racial minorities have been widely reported. Examples include facial recognition software that misidentifies gender and even species when presented with dark-skinned women of African descent (Buolamwini and Gebru, 2016) and proprietary formulas used in criminal sentencing that misclassify defendants as high risk for recidivism at a greater rate for Black versus White defendants (Rudin *and others*, 2018). Two explanations for differentially poorer model performance can be addressed by collecting more data: too few observations of members of racial minority groups and unrepresentative sampling that can differentially limit generalizability (Kreatsoulas and Subramanian, 2018).

However, an additional cause of algorithmic bias is not well appreciated and cannot be overcome simply by adding more of the same kind of data to a learner. The problem is the data generation process itself. The data generation process "[is] an inherently subjective enterprise in which a discipline's norms and conventions help to reinforce existing racial (and other) hierarchies" (Ford and Airhihenbuwa, 2010). As explicated by the Public Health Critical Race Praxis, without an explicit focus on social equity, the concerns of the most privileged members of society are overrepresented in data and research (Ford and Airhihenbuwa, 2010). One empirical demonstration of this phenomenon is Tehranifar *and others*' (2009) investigation of racial disparities in survival ranked by the degree of knowledge that had been generated about a cancer. Ranked from "nonamenble" (little knowledge, uniformly high mortality rates) to "mostly amenable" (well-studied, high survival), "the hazard ratios (95% confidence intervals) for African Americans versus Whites from nonamenable, partly amenable, and mostly amenable cancers were 1.05 (1.03–1.07), 1.38 (1.34–1.41), and 1.41 (1.37–1.46), respectively" (Tehranifar *and others*, 2009). Absent much knowledge about prevention or treatment, Blacks and Whites had similar mortality outcomes. The kind of knowledge that was produced disproportionately benefited the health of the White populations. One factor is the ability of more privileged groups to access knowledge and leverage financial and medical resources (Phelan *and others*, 2010). But a more fundamental factor is the research enterprise itself tends to collect more data and advance more quickly on problems that disproportionately affect those in society with more power and resources. For instance, over the past decades, medical treatment for the "triple-negative" subtype of breast cancer that disproportionately affects Black women has advanced much more slowly than for the hormone-receptor positive subtypes that are disproportionately diagnosed among White women in the Unites States (Foulkes *and others*, 2010). Similarly, in the case of multiple myeloma wherein Black patients make up 20% of all people diagnosed, in a review of 21 clinical trials in patients with multiple myeloma, the median proportion of Black patients enrolled was only 4.5% (Bhatnagar *and others*, 2017).

One solution to algorithmic bias is to follow Doug Weed's suggestion from the "black box" epidemiology debates: devote special attention to causal structures that act at different levels. A tool that we find valuable for this effort are directed acyclic graphs, or DAGs. In a DAG, arrows between nodes represent our beliefs about the presence of causal relationships among factors under study. D-separation, a set of rules for drawing and analyzing the relationships in the DAG (http://bayes.cs.ucla.edu/BOOK-2K/d-sep.html), results in concrete guidance for model building and interpretation. For instance, DAG analysis can identify biases, such as uncontrolled confounding ("omitted variable bias"), unaddressed in an analysis or even biases introduced when problematic variables are included in the adjustment set (e.g., instruments or mediators) (Keil *and others*, 2018; Hernán *and others*, 2001), which is not the same as, but related to, the concept of bias from "overfitting" (Seligman *and others*, 2018).

3. Empirical example: algorithms and lung function

Here we present an example from the pulmonary health literature. Lung function is typically assessed using a tool called a spirometer (Braun, 2015). Internationally, most commercially available spirometers require the operator to "select the race of an individual, as well as indicate their age, sex/gender and height. These data are fed into an algorithm that "corrects" for each factor, based on the assumption that normal levels of lung function differ by age, sex, height, and race. The first formulas were produced in the United States in the 1920s, "during a period when eugenic policies rooted in hereditarianism were popular" (Braun, 2015). The formulas have been updated over time, most recently based on data from the 1988–1994 NHANES exam (Braun, 2015). Today, for example, in the United States, compared to a "Caucasian" population, correction factors for individuals labeled "black" range from 10% to 15%, on the assumption that Black people have constitutionally poorer lung function; for people labeled "Asian," correction factors are between 4% and 6% (Braun, 2015).

Figure 1 shows a DAG depicting how the spirometer manufacturers and most lung function researchers tacitly conceptualize the relationship between race and normal lung function. In addition, White workers

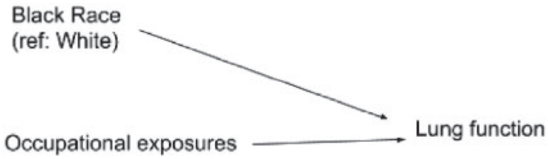

Fig. 1. Directed acyclic graph (DAG) depicting naive conceptualization of causal relationships among race, occupational exposures, and lung function.

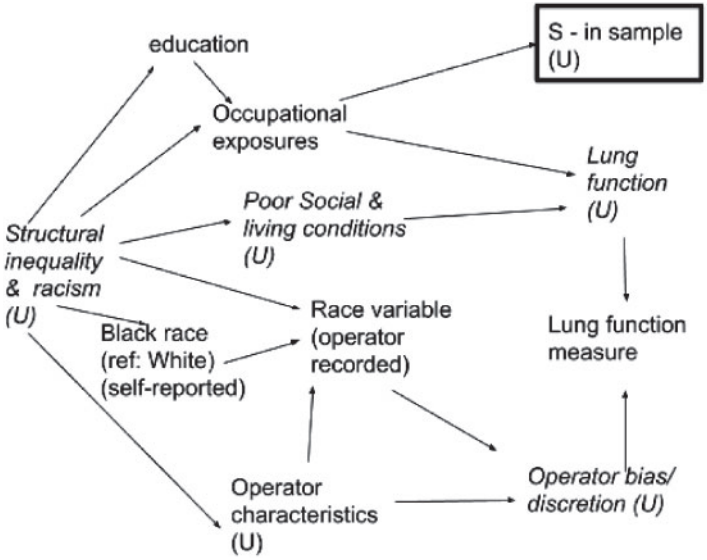

Fig. 2. Directed acyclic graph (DAG) of relationships among race, occupational exposures, and lung function, incorporating structural racism and theories from the Public Health Critical Race Praxis.

are centered in this conceptualization (Ford and Airhihenbuwa, 2010). Their lung function is considered to be the norm from which non-White people deviate. Race is being conceptualized here as a characteristic of a person that independently and innately depresses lung function. In addition, this DAG encodes the prediction that the effect of harmful occupational exposure on lung function would differ for Black and White people. Therefore, it would makes sense to correct for race when predicting lung function based on occupational exposures. The result is that a given level of poor lung function would be considered abnormally low for a White person but normal for a Black person. The implications are profound: Black and Asian people must exhibit lower levels of actual lung function than White people to cross clinical thresholds for qualifying for therapeutic services or disability benefits. Further, the model predictions could easily justify allowing greater levels of exposures to occupational hazards for Black workers who do not exhibit the depressed levels of lung function associated with harm.

Figure 2 presents a similar DAG informed by knowledge of structural racism. This DAG explicitly incorporates possible operator bias, interrogates the source of the "race" variable and its meaning (e.g., self-reported forced choice among categories, operator perception). Further, the DAG makes the data generation process and variations in data quality more explicit. Finally, by acknowledging that observed associations between race and lung function could be entirely explained by racial disparities in living and working conditions, the DAG animates the analyst to collect or include additional, more proximally causal, and likely more statistically predictive, variables, such as living conditions, for which race is operating as a surrogate. Crucially, incorporating more proximal and predictive variables into models, rather than relying on race variables to act as proxies, will improve transportability of algorithms across contexts.

4. Conclusion

"The fundamental goal of machine learning is to generalize beyond the algorithm training set" (Kreatsoulas and Subramanian, 2018). In particular, when applied to health and health care, even models intended only for prediction will have causal impacts. Model results will be used for decision-making about the allocation of health care, access to social welfare and disability systems, and acceptable limits of medical exposures for vulnerable populations. We have argued that grounding one's work in an understanding of structural racism will improve model accuracy and help avoid the pitfalls of limited application to racial minority populations, algorithmic bias, limited transportability and reinforcing racial inequities.

Acknowledgments

Conflict of Interest: None declared.

Funding

Whitney R. Robinson is supported by the National Institute of Minority Health and Health Disparities (NIMHD) of the National Institutes of Health (NIH) under award number NIH-R01MD011680.

Audrey Renson is supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development (NICHD) of the National Institutes of Health (NIH) under award number T32HD091058-01.

Ashley I. Naimi is supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development (NICHD) of the National Institutes of Health (NIH) under award number NIH-R01HD093602.

References

Bailey, Z. D., Krieger, N., Agénor, M., Graves, J., Linos, N. and Bassett, M. T. (2017). Structural racism and health inequities in the USA: evidence and interventions. *The Lancet* **389**, 1453–1463.

Bhatnagar, V., Gormley, N., Kazandjian, D., Goldberg, K., McKee, A. E., Blumenthal, G., Farrell, A. T. and Pazdur, R. (2017). FDA analysis of racial demographics in multiple myeloma trials. *Blood* **130**(Suppl 1), 4352.

Braun L. (2015). Race, ethnicity and lung function: a brief history. *Canadian Journal of Respiratory Therapy: CJRT* **51**, 99.

Buolamwini, J. and Gebru, T. (2018). Gender shades: intersectional accuracy disparities in commercial gender classification. In: Friedler, S. A. and Wilson, C. (editors), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, in Proceedings of Machine Learning Research, New York University, NYC, pp. 1–15.

Domingos, P. M. (2012). A few useful things to know about machine learning. *Communications of the ACM* **55**, 7887.

Ford, C. L. and Airhihenbuwa, C. O. (2010). The public health critical race methodology: praxis for antiracism research. *Social Science & Medicine* **71**, 1390–1398.

Foulkes, W. D., Smith, I. E. and Reis-Filho, J. S. (2010). Triple-negative breast cancer. *New England Journal of Medicine* **363**, 1938–1948.

Glass, T. A., Goodman, S. N., Hernán, M. A. and Samet, J. M. (2013). Causal inference in public health. *Annual Review of Public Health* **34**, 1–75.

Hernán, M. A., Hernández-Díaz, S., Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology* **15**, 615–625.

Keil, A. P., Mooney, S. J., Jonsson Funk, M., Cole, S. R., Edwards, J. K., Westreich, D. (2018). Resolving an apparent paradox in doubly robust estimators. *American Journal of Epidemiology* **187**, 891–892.

Kreatsoulas, C. and Subramanian, S. V. (2018). Machine learning in social epidemiology: learning from experience. *SSM-Population Health* **4**,347.

Phelan, J. C., Link, B. G. and Tehranifar, P. (2010). Social conditions as fundamental causes of health inequalities: theory, evidence, and policy implications. *Journal of Health and Social Behavior* **51**(Suppl 1), S28–S40.

Rudin, C., Wang, C. and Coker B. (2018). The age of secrecy and unfairness in recidivism prediction. arXiv preprint arXiv:1811.00731.

Seligman, B., Tuljapurkar, S. and Rehkopf, D. (2018). Machine learning approaches to the social determinants of health in the health and retirement study. *SSM-Population Health* **4**, 95–99.

Tehranifar, P., Neugut, A. I., Phelan, J. C., Link, B. G., Liao, Y., Desai, M. and Terry, M. B. (2009). Medical advances and racial/ethnic disparities in cancer survival. *Cancer Epidemiology and Prevention Biomarkers* **18**, 2701–2708.

Weed, D. L. (1998). Beyond black box epidemiology. *American Journal of Public Health* **88**, 12–14.

[*Received September 25, 2019; revised September 25, 2019; accepted for publication September 25, 2019*]