Exploring machine learning in health and its impacts on racialized populations

By: Irene Duah-Kessie

Introduction

Machine learning (ML) is increasingly becoming adopted across the health care sector. It has been used to identify new patterns and relationships in data sets, produce new hypotheses and inform researchers and health care providers of opportunities that can advance their work (Noorbakhsh-Sabet et al., 2019). Machine learning systems can improve their performance without any explicit programming, which assists in the development of automated clinical decision-making processes (Marr, 2016).

The clinical application of ML in health care is mainly focused on cancer, nervous system, cardiovascular, infectious, and chronic diseases as they tend to be leading causes of mortality (Noorbakhsh-Sabet et al. 2019). These technologies are often employed to achieve early diagnosis, assist with the accuracy of outcome predictions, to detect potential complications, which informs resource utilization and improves quality of patient care (Razavian & Tsirigos, 2018; Schnyer et al., 2017; Jiang, Jiang, & Zhi, 2017; Cruz & Wishart, 2007). Although the ultimate goal of ML is to develop automated processes that enable an efficient health care system, there are various practical and methodological challenges.

The widespread application and use of ML in clinical settings produces a wide range of ethical and technical challenges that require novel approaches to address concerns about data sizes, heterogeneity, privacy, and biases. Fairness definitions are becoming increasingly critical as biased algorithms deployed at scale can replicate or exacerbate racial disparities (Angwin et al., 2016). The links between health technology and social justice should therefore be a focal point for developing ML algorithms that function to produce equitable predictions and reducing barriers to good health.

The aim of this paper is to introduce concepts of bias and fairness of machine learning in health, current approaches to measuring fairness within these models, and limitations with current indicators of fair algorithms.

Bias in Machine Learning Algorithms

Machine learning algorithms are the brains behind any model (Costa, 2020). An algorithm is an explicit set of instructions designed to perform a specific task (Marr, 2016). The algorithm is trained using sets of relevant data and repeatedly tested until data points are ranked and aggregated to generate predictions (Kumar, 2018). The process of consistently exposing an algorithm to billions of data points and new experiences improves efficiency and accuracy of the machine being used (Costa, 2020). The output of the algorithm is referred to as a prediction – in many cases this can be either a medical diagnosis or treatment options. Figure 1 demonstrates the process of how algorithms are developed and enhanced through a machine learning system.

An algorithm is considered biased when systematic error occurs in the computing and translation processes, leading to unfair predictions that tend to give privileges to one group over another. (Mehrabi et al., 2019). The figure below highlights some of the common biases that show up at different stages of the cycle. Systemic issues of bias and discrimination arises when such predictions are used to further train the ML system, amplify inaccurate results, and normalizes inequitable decision-making processes. This phenomenon is known as a feedback loop, where a model's predicted outputs are reused to train subsequent versions of the model and can occur between the data and algorithm and algorithms and user interactions (Mehrabi et al., 2019). Although some biases may seem harmless, or somewhat inconvenient, if not addressed they become a part of the underlying rationale for algorithms that inform critical decisions.

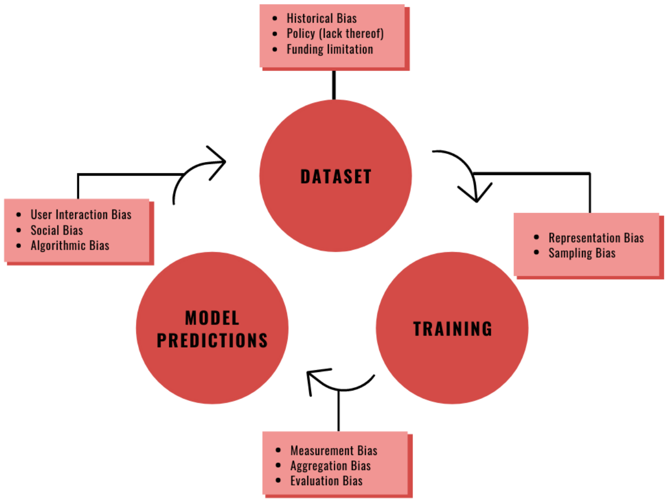

Figure 1: Machine learning algorithm process and the potential biases that arises at each stage of the process

Biases in Data Sets and Training Process

A wide range of biases are discussed in the literature that emerge within datasets, data collection, annotated and implementation processes. *Historical bias* is a form of bias that refers to the existing biases and socio-technical issues that seep into the data generation process and implicate data sets (Suresh et al, 2019; Crawford, 2016; Vaughn et al., 2020). For example, Goyal et al., 2015 found that Black, Hispanic, and Asian patients are given less pain medication than White patients for similar injuries and reported pain level, however there were no differences in overall analgesia administration. Their findings suggest that pain may be equally assessed, however, physicians tend to treat Black patients' level of pain differently due to historical, cultural and stereotypical notions that Black people have "thicker skin". This study demonstrates how implicit biases does not only influence health care delivery and health outcomes, but overtime discriminatory patterns in treatment options for racialized patients will make up data that influences predicted outputs of ML systems.

These historical biases are often interrelated to *policy limitations* or the lack thereof in failing to implement more rigorous approval processes for new and emerging ML algorithms and devices. For example, in 2018 the Food and Drug Administration (FDA) approved a "Software as a Medical Device" to detect diabetic retinopathy, however the data set used to train this device lacked Native Americans and Asian Americans, despite high prevalence of diabetes in these subpopulations (Vaughn et al., 2020). Although the FDA proposes guidelines for data quality assurance, there is very little emphasis on reducing bias and these regulatory frameworks do not rely on the input of physicians or health care providers in general. In addition, *funding bias* arises when biased outcomes are reported to support the interests of industry sponsors for the research study or technology. This source of bias is an underlying issue that can affect various stages of the process from data collection to implementation. There are very few studies that discuss this form of bias influencing ML in health care, however Fabbri et al. (2018) demonstrates that industry funding in health care generally is associated with commercial applications.

Findings from projects funded by industry leaders can be used in legislative settings to influence development of policies that align with the interests of institutions, rather than those patients and communities who are impacted (Fabrri et al., 2018). The structures of policy development and funding distribution and often interrelated and guided by ideas and principles deeply rooted in history. These factors are a great contribution to systemic biases that influence data collection and training processes.

Representation bias is a common type of bias, where the training data is not representative of the data used in practice (Muhrabi et al., 2019; Vaughn et. al, 2020). For example, many health care researchers frequently use the Medical Information Mart for Intensive Care IV (MIMIC-IV) data set in their work. However, it is not reflective of many populations and regions in various ways; it is based on critical care data from one-hospital in the city of Boston, (Vaughn et al., 2020). *Sampling bias* is another form of bias that contributes to issues of under representative data sets due to the non-random sampling of subgroups (Mehrabi et al., 2019). Consequently, trends gathered for one population may not be applicate to data collected from another population. Therefore, researchers must ensure their selection processes for data sets and subgroups accurately characterize those impact. If they do not, further iteration of the given model or data collection process should be considered.

Biases in Model Predictions and Interpretation

Measurement bias is the way particular features are selected, used, and measured (Suresh et al, 2019; Davies & Goel, 2018). Biases within the labels and indicators pose significant threats to generating equitable predictions. Electronic health records (EHR) are used routinely to document a patient's condition and also used as data points for ML models. However, studies show that EHR systems tend to vary and can affect the precision of recorded data. Hripcsak et al. (2011) found variations in recorded notes and lab results for severely ill patients compared to healthier patients undergoing similar treatments due to high demands on physicians' time and resources. As a result, nuanced experiences and conditions of various patients may not exist in sufficient number for a predictive algorithm to make better informed decisions for (Gianfrancesco et al., 2019). The underestimation or overestimation can lead to uninformative predictions that can have influence decisions around clinical support.

Aggregation bias is another form of bias within the process of analysis and deployment of algorithms. It occurs when incorrect conclusions arise for a subgroup based on observations of another subgroup (Suresh et al, 2019). For instance, it is known that certain biomarkers for diabetes appears differently across ethnicities (Herman & Cohen, 2012); however, many physicians administer HbA1 c levels in monitoring the complexities of diabetes among diverse patients. This treatment selection process is a form of the one-size fits all model, where one large label is made up of many heterogeneous populations despite indicators of distinct differences across subgroups (Vaughn et.al, 2020).

Evaluation bias refers to the use of inappropriate and disproportionate benchmarks (Mehrabi et al., 2019). This form of bias has been well-documented within facial recognition systems that are often biased towards skins of colour, genders, and various features (Buolamwini & Gebru, 2018). Adamson & Smith (2018) conducted multiple studies and found that Black Americans had lower incidence rates of skin cancer yet were less likely to survive (66%) than their White counter parts (94%). This finding had been attributed to poor screening processes that failed to catch signs of skin cancer cells early due to the lack of Black and diverse skin types used in training algorithms and diagnostic tools. These three forms of biases simultaneously promote biased predictions as they can shift, generalize, and/or modify the true nature of the data point that account for under- or misrepresentation subpopulations.

User Interaction Bias is a type of bias that can be triggered through the user interface and the user itself by imposing one's self-directed biased behavior and interaction (Mehrabi et al., 2019). This form of bias can be influenced by how predictions are presented or the feedback loop effect, where results most relevant to the users will seem most attractive and true. For instance, pre-existing beliefs and experiences among physicians, health care providers and engineers can influence how predictions are interpreted and the actions that follow. Sometimes, that action can be influenced by other people's actions or content, and this is referred to as *social bias.* Social and user interaction bias combined can exacerbate and confirm the historical issues that many racialized people are subjected to. However, when all the aforementioned biases are accounted for, *algorithmic bias* can still be apparent. Algorithmic bias is when bias is not present in the data set, rather it is added solely by the algorithm itself (Mehrabi et al., 2019). This form of bias demonstrates the ways in which hidden stereotypes and discrimination is deeply embedded into the fabric of our society and institutions. Because knowledge production and data collection processes are not always equitable, computer systems can replicate such errors and systematically create unfair outcomes.

As such several forms of biases tend to work simultaneously and contribute to one another at various stages of the process and can amplify the effects of bias not just on the algorithm's predictions but those who are those who are interacting with the algorithm (Chaney, Stewart, Engelhardt, 2018).

Measurements of Fairness in Health Care Data Sets

Within the literature, fairness is broadly defined as "the absence of any bias based on an individual's inherent or acquired characteristics such as race or gender that are irrelevant in the context of decision making" (Chouldechova, 2017). McDonald & Pan (2020) points out that the absence of intentional bias does not equate fairness, rather unfairness is a feature of biased data inputs or systems and the inequitable outcomes produced when models are deployed.

There have been three distinct definitions of fairness used within research. The first is an *anti-classification*, where social identities such as race, gender are not explicitly referenced. Several studies suggest that broad notions of anti-classification expressed through algorithms can lead to the exclusion of important information and use of harmful proxy variables leading to discriminatory decisions (Johnson et al., 2016; Davies & Goel, 2018). Fairness through blindness could therefore develop a predictive value, where unknown features are correlated with race and other socially oppressed identities. Zemel et al. (2013) describes representation learning, where one learns subtle representation to minimize the groups information, however, issues arise when there is misinformation or a lack of information occurs. Davies & Goel (2018) suggest that representative data is most critical using this approach as the addition and use of protected attributes is justified.

The second commonly used definition is the *classification parity*, where common measures of performance are equal across all groups. Models are considered fair when error rates are distributed similarly across groups (Chen et al., 2018). This approach is known to be a family of criteria that involves most mathematical definitions of fairness and readily used by researchers (Hardt et al., 2016; Argarwal et al., 2018; Chouldechova, 2017). Popular forms include "equalized odds criterion", where a model is fair when false negative rates and false positive rates are equal across groups and "equal opportunity", where false positives must be equal (Hardt et al., 2016). One challenge with this fairness definition is that the distribution of risk will vary across groups and lead to differences in error rates, a phenomenon known as infra-marginality (Simoiu et al., 2017). Pierson (2020) explains that we may incorrectly conclude from high positive rates because we are measuring the variables at face value, rather than assessing the probability of the risk being measured. Davies & Goel (2018) also explains that this approach is linked to the legal and economic understanding of margins and its effects on social welfare, which will then become poorer measures for equity and well-being.

The third widely considered definition is *calibration*, where the condition is adjusted to hold simultaneously for individuals within each group and the predicted probability is compared (Liu et al., 2017). A challenge with this approach is that individuals can be misclassified to a false outcome that does not reflect true reality (Chouldechova, 2018). Therefore, there can be discrepancies with those who are predicted to exhibit the condition tested and whether that algorithm made an effective prediction. Davies and Goel (2018) describe this approach as weak guarantee of equity because a positive or negative classification is dependent upon the risk distribution attributed to the low-risk and high-risk groups. Often predictions are based on one or very few variables such as postal code that do not allow for other important factors to be accounted and can be easily affected by changing the risk distribution to be concentrated around the average (Davies & Goel, 2018). This approach demonstrates the importance of acknowledging all available information when constructing statistical predictions, as such assessments could mistakenly ignore data that may facilitate biased outcomes while staying true to calibration (Davies & Goel, 2018). Obermeyer et al. (2019) demonstrates this concept through their research further explored as a case study in the following section.

Limitations of Fairness Measurements

These commonly used mathematical definitions of fairness are an attempt to pave a way towards equitable and fair algorithmic decision, however they all possess shortcomings from their design and framing of risk measurements. Many data sets already hold biases due to historical practices and outcomes among marginalized patients and can elicit feedback loops that will inevitably have long-term health and generational impacts. The data quality issues with the anti-classification approach speaks to representation or selection bias where the data sets do not accurately reflect or appropriately capture the experience of diverse people. By grouping the different subgroups into one classification of a health condition, there is a risk of maintaining aggregation bias as differences in other correlated variables are not accounted. Thirdly, calibration processes points to the issues of measurement bias, where labels and annotations may capture the interest of the model's design purpose rather than the actual needs of the population assessed. Across these three definitions, historical biases arise as the quality of the data sets used is influenced by previous collection, training, and interpretation practices and the worldviews of scientists who develop and monitor the algorithm. As such current definitions of algorithmic fairness do not readily detect discriminatory processes or features and can therefore perpetuate biases unknowingly (Davies & Goel, 2018). Although it may be difficult to quantify, considerations for the broader socio-political and socio-technical context may improve the overall value and efficacy of measuring fair algorithms. Table 1 below summaries these fairness definitions, strengths, weaknesses and provide an example of real-life applications.

Fairness Measurement	Definition	Strengths	Weaknesses	Biases Implications	Real-life example
Anti-Classification	Two groups are given the same values of variables, have the same decision and these features do not change the decision	Social identities such as race and gender are often not explicitly used to enable decision making	Leads to the exclusion of critical information and discriminatory decisions	Representation bias Historical bias	The lack of inclusion of skin of colour in ML algorithms can lead to suboptimal data for diagnostic tools

Classification Parity	The proportion of positive predictions should be the same across all groups	Provides same level of risk for all groups and allows more equal assessment of harm across each individual	True underlying distribution of risk varies across groups and can lead error rates differences when individual risks are captured	Aggregation bias Historical bias	The treatment of diabetes is generally the same across all groups, yet biomarkers appear differently and have various effects
Calibration	Comparison of the actual prediction and the expected prediction	The model represents the true probability of the occurrence of the actual outcome	Provides weak guarantee of equity; it often teaches a model to misclassifies individuals to a false outcome that is not true to reality	Measurement bias Historical bias	For any given prediction for enrolment into specialized care, Black and White patients experienced similar levels of pain

Table 1: Summary of fairness definitions, strengths, weaknesses, and real-life examples for each approach

Case Study: Promising strategies to accurate measurements of bias

This section will further explore one case study to exemplify a promising approach to ensuring fair algorithms. Obermeyer et al. (2019) revealed that a widely used algorithms that identifies high-risk patients for additional health care services prioritizes healthier, White patients ahead of sicker Black patients. The scientists addressed the measurement bias that implicates the algorithm used to better understand how it may compute more accurate predictions that meet the needs of Black patients.

Background: The researchers use a comprehensive data set to gain insight into an algorithm used as a screening tool for high-risk care management programs. These programs are considered effective at improving the care of patients with complex needs while reducing costs of care (McCall, 2010). Often, doctors time and resources are limited, and these algorithms assist with accelerating the process of detecting individuals who require specialized or greater care and preventing the deterioration of health. Although there is an assumption that those with the greatest needs will benefit most from the program, the researchers point to a targeting issue where this decision-making algorithm relies on pre-existing data and misidentified labels to establish health needs predictions.

Methods and fairness analysis: The algorithm studied uses insurance claim and laboratory data to produce predictions on health needs. The data sample identified 6079 Black and 43,539 White primary care patients from 2013 to 2015, where 71.2% were enrolled in commercial insurance and 28.8% in Medicare. The researchers focus their analysis on calibration bias where Black and White patients' scores are equalized to indicate no bias has occurred. They compared the predictions based on insurance data and patient's health data to assess the algorithms efficiency across health outcomes. They also evaluated whether the algorithms predictions were grounded in costs by linking predictions with insurance data to actual health care utilization.

Findings: The researchers found that Black patients are significantly more ill than White patients, with 26.3% more chronic illnesses than White patients, including hypertension, diabetes, renal failure, and high cholesterol. The researchers also found evidence of program screening disparities, where Black patients were least likely to be approved into the program. They recalibrated the algorithm to ensure no predictive gap between Black and White patients and found that the percentage of Black patients substantially increased from 17.7% to 46.5%. Although significant health disparities are apparent, the researchers found very little difference in costs showing that a substantial number of Black patients are likely not receiving the care they actually need. The researchers suggested that the algorithm used to predict health needs, is in actuality, a prediction of health costs. This rationale demonstrates how ML systems are designed to prioritize the needs of the health care institutions, rather than patients health.

Conclusions: The researchers attributed the causes of bias to the labels selected to assess these outcomes. This study provided a socially conscious approach to machine learning algorithms demonstrating the myriad of influences on predictions of health needs and status.

Discussion

In this literature review, the aim was to understand ways that bias is revealed through algorithms and whether current approaches to measure fair algorithms are sufficient. After acknowledging prominent biases that mediate machine learning algorithm processes, fairness measurements such as anti-classification, classification parity and calibration are limited in that they do not actively detect discriminatory decisions. As demonstrated above, there are many studies that have documented the disparities that exist as a result of biased data and deployment processes, however very few that addressed the mechanisms in which they arise. There are also very few studies that interrogate systemic biases within the data sets and collection policies, which can be argued as the root causes of algorithmic biases. These challenges could be attributed to barriers to accessing critical data sets and relevant information that inform the development and ideation processes. Several researchers agree that there is a large focus on individuals rather than the structural factors that shape racial disparities. It is critical to continuously uncover these factors and address them directly (Crawford, 2019; Mehrabi et al., 2019; Obermeryer et. al, 2019).

In the case study explored, the researchers were able to create a simulated algorithm that specifically addressed the health needs of an individual. Obermeyer et al. (2019) had the unique opportunity to assess the quality of the algorithms' design and inputs as it relates to the program that it informs. The collaborative nature of this work gave the researchers ability to access the data type and work with the engineers to address the root issue of the apparent racial bias. The researchers draw upon two potential causes of these racial biases. For Black communities specifically, many have low levels of trust in the health care system due to its legacy of anti-Black racism. There are also patterns of mistreatment and negative interactions among Black patients and physicians. Boag et al. (2019) used trust scores to quantify doctor-patient relationships and found racial associations among patients with higher levels of mistrust have worse experiences with health care providers. These findings demonstrate that there are strong connections between the structural and interpersonal mechanisms of bias that shape a person's experience with the health care system. Therefore, the ways in which these factors influence, and shape health care data sets, algorithmic designs and evaluation procedure must be recognized and questioned.

Machine learning methods in health should be intentionally created and critiqued with a deep awareness of multiple perspectives and methodologies to ensure the inequalities are addressed (Ciston, 2019). An intersectional approach can be used to analyze algorithms, uncover the various elements influencing a health needs, and create unique, tailored approaches to help shift current designs and interpretation models. Intersectional approaches are reflexive in nature and can help rework stereotypes and reprogram into technologies of agency, rather than inaction on reducing inequities (Ciston, 2019). Although this may be a costly and timely approach, it ensures that technologies are developed account for the diverse perspectives, experiences, and histories of many people and accurately assessed nuanced conditions.

As the field continues to emerge, it is essential to ensure that the downstream effects do not overshadow discriminatory and political contexts that makes automated tools an area of concern to begin with (Benjamin, 2020). Evidence has shown that these biases create a cycle that propels the lack of research and investment into the realities of vulnerable populations. The underinvestment into equitable ML demonstrates the poor measures of the needs of those historically marginalized, which can impede efforts to advancing racial justice in health care through algorithms. The current definitions of fairness should therefore move from assessing individual risks to evaluating the production of risks by institutions to prevent the great yet subtle harms of technology. This can help to ensure medical institutions and health care providers are being held accountable for inequities they perpetuate. Greater levels of transparency in the collection of race-based data and algorithmic labels are also necessary to better assess the needs of racialized patients and identify those falling between the cracks.

Conclusion

Moving forward, approaches to machine learning experiments and interventions should produce new possibilities for more thoughtful and inclusive algorithms and technologies. As researchers, physicians and health care institutions become more aware of the structural and cultural biases within the field, there will are greater opportunities to ensure fairness and reduce barriers to accessing care among historically marginalized communities. It is essential to acknowledge the difference between biases that are inevitable and discrimination that can be prevented. Knowing this can help steer clear of a future where algorithms are used in harmful ways and become more proactive to address the foreseeable disparities. Ongoing efforts should better document the inputs, outputs, and clear objectives of the algorithm's particular functions. The collection of data that is representative of the diverse patient population is also critical to enable predictions that accurately reflect all people's experiences and phenotypes. With transparency and reliable data information, the health care industry as a whole can eliminate disparities and advance more equitable machine learning algorithms for future generations.

Acknowledgements

I would like to express my gratitude to Dr. Jesse Gronsbell and Dr. Monica Alexander for their continued guidance, support and invaluable feedback in completing this paper. I would also like to thank the Toronto Data Workshop at the University of Toronto for the opportunity to explore this work.

References

Adamson, A. & Smith, A. (2018). Machine Learning and Health Care Disparities in Dermatology. *American Medical Association, 154:11, 1247-1248.*

Agarwal, A., Beygelzimer, A., Dud´ık, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In International Conference on Machine Learning.

Benjamin, R. (2019). Assessing risk, automating racism. *Social Science, 336:6464, 421-422.*

Boag, W., Suresh, H., Celi, L.A., Szolovits, P., & Ghassemi, M. (2019). Modeling Mistrust in End-of-Life Care.

Brandon, D.T., Issac, L.A, LaVeist, T.A. (2005). The legacy of Tuskegee and trust in medical care: is Tuskegee responsible for race differences in mistrust of medical care? *National Library of Medicine Association, 97(7): 951-956.*

Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. 2018. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In Proceedings ofthe 12th ACM Conference on Recommender Systems. ACM, 224–232.

J. Angwin, J. Larson, S. Mattu, L. Kirchner. (2016). Machine Bias. *ProPublica*

Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research), Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, New York, NY, USA, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

Canadian Medical Association (2013). *Health care in Canada: What makes us sick?: Canadian Medical Association Town Hall Report.* Ottawa (ON): Canadian Medical Association.

Chen, I.Y., Johansson, F.D., & Sontag, D. (2018). Why Is My Classifier Discriminatory?

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data 5(2):153–163.

Colin R Blyth. 1972. On Simpson’s paradox and the sure-thing principle. J. Amer. Statist. Assoc. 67, 338 (1972), 364–366.

Crawford, K. (2016). Artifical Intelligence’s White Guy Problem – Opinion in New York Times

Cruz, J.A & Wishart, D.S. (2007). Applications of machine learning in cancer prediction and prognosis. *Cancer Information, 11, 59-77.*

Davies, S.M. & Goel, S. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning.

Drysdale, E. (2019). Implementing AI in health care. *Vector-SickKids Health AI Deployment Symposium*

Fabbri, A., Lai, A., Grundy, Q., & Bero, L. A. (2018). The Influence of Industry Sponsorship on the Research Agenda: A Scoping Review. *American journal of public health*, *108*(11), e9–e16. https://doi.org/10.2105/AJPH.2018.304677

George Hripcsak, Charles Knirsch, Li Zhou, Adam Wilcox, and Genevieve Melton. 2011. Bias Associated with Mining Electronic Health Records. Journal ofBiomedical Discovery and Collaboration 6 (2011), 48–52. https://www.ncbi.nlm.nih.gov/ pmc/articles/PMC3149555/

Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA internal medicine*, *178*(11), 1544–1547. https://doi.org/10.1001/jamainternmed.2018.3763

Herman, W.H., & Cohen, R. (2012). Racial and Ethnic Differences in the Relationship Between HbA1c and Blood Glucose. Obstetrical Gynecological Survey 67 (08 2012), 468–469.

Jiang, F., Jiang, Y. & Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke Vascular Neurology 2, pp. 230-243*

Jipguep-Akhtar, M. (2020). Book Review of Race After Technology: Abolitionist Tools for the New Jim Code.

Johnson, K. D., Foster, D. P., and Stine, R. A. (2016). Impartial predictive modeling: Ensuring fairness in arbitrary models. arXiv preprint arXiv:1608.00528

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances In Neural Information Processing Systems, pages 3315–3323.

Holland, S., Hosny, A., Newman, S., Joseph, J. & Chmielinski, K. (2018). The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *Cornell University Computer Science Databases.*

Kaiser Family Foundation. 2019. Black Americans and HIV/AIDS: The Basics. https://www.kff.org/

Kusner, M., Loftus, J., Russell, C., Silva, R. (2017). Counterfactual Fairness. Proceedings from the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA.

Lashbrook, A. (2018) AI-Driven Dermatology Could Leave Dark Skinned Patients Behind. *The Atlantic.*

McCradden, M.D., Joshi, S., Mazwi, M., & Anderson, J. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet*

McDonald, N. & Pan, S. (2020). Intersectional AI: A Study of How Information Science Student Think about Ethics and Their Impact. *AMC Journal Vol 4 (CSCW2).*

Mehrabi et al. (2019) A Survey on Bias and Fairness in Machine Learning. *Cornell University*

N. McCall, J. Cromwell, C. Urato, “Evaluation of Medicare Care Management for High Cost Beneficiaries (CMHCB) Demonstration: Massachusetts General Hospital and Massachusetts General Physicians Organization (MGH)” (RTI International, 2010).

Noorbakhsh-Sabet, N., Zand, R., Zhang, Y., Abedi, V. (2019). Artificial Intelligence Transforms the Future of Health Care, *132(7), 795-801.*

Pierson, E. (2020). Assessing racial inequality in COVID-19 testing with Bayesian threshold tests. *Machine Learning for Health.*
Pletcher, M.J., Kertesz, S.G., Kohn, M.A., Gonzales, R. (2008). Journal of American Medical Association, 299(1), 70-78.

Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G., Chin, M.H. (2018). Ensuring Fairness in Machine Learning to Advance Health Equity. *Ann International Medicine, 169(12): 866-872.* doi: 10.7326/M18-1990.

Razavian, N. & Tsirigos, A. (2018). Pathologists meet their match in tumour-spotting algorithm
Nature, 561, 436-437

Harini Suresh and John Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. https://arxiv.org/abs/1901.10002

Schnyer, D.M., Clasen, P.C., Gonzalez, C., Beevers, C.G. (2017). Evaluating the diagnostic utility of applying a machine learning algorithm to diffusion tensor MRI measures in individuals with major depressive disorder *Psychiatry Res Neuroimaging, 264, pp. 1-9*

Sjoding, M.W., Dickson, R.P., Iwashyna, T.J., Gay, S.E., Valley, T.S. (2020). Racial Bias in Pulse Oximetry Measurement. *New England Journal of Medicine.*

Solar O, Irwin A. A conceptual framework for action on the social determinants of health. Social Determinants of Health Discussion Paper 2 (Policy and Practice).

Vaughn, J., Vadari, M., Baral, A., & Boag, W. (2020). Dataset Bias in Diagnostic AI system: Guidelines for Dataset Collection and Usage. CHIL, Toronto, ON.

Zemel, R., Wu, Y., Swersky, K., & Pitassi, T, Dwork, C. (2013). Learning fair representations. Proceedings of the 30th International Conference on Machine Learning, PMLR 28(3):325-333, 2013.
Racial bias in pulse oximetry measurement,” New England Journal of Medicine. DOI: 10.1056/NEJMc2029240