Exploring machine learning in health and its impacts on racialized populations

Irene Duah-Kessie

Introduction

Artificial intelligence (AI) in health care is increasingly becoming adopted and brings a paradigm shift to the field as the availability of data and analytic techniques progress (Jiang et al., 2017). AI can be applied within various health care data sets and is known to perform human-like activities and increase its capacity such as learning, perception and problem solving (Noorbakhsh-Sabet et al., 2019). A popular AI technique used in health includes machine learning (ML), which can be used to identify patterns and relationships in the data, produce new hypotheses and direct researchers and health care providers towards opportunities that inform and advance their work (Noorbakhsh-Sabet et al., 2019).

Machine learning systems can improve their performance without any explicitly programming on how to perform the task (Marr, 2016). This assists in developing automated clinical decision-making processes and enables machine learning algorithms to carry out activities such as self-diagnosis or prescription auditing (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 and prognosis, 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 algorithms 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 algorithms deployed at scale are only as good as the data set used. For instance, there is growing concern that algorithms can replicate or exacerbate health disparities (Angwin et al., 2016). The links between health technology and social justice should be a focal point for reducing health inequities and developing AI and ML algorithms that function to produce equitable predictions.

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). They enable themselves to carry out predictions while simultaneously learning and improving from over trillions of data observations (Kumar, 2018). An algorithm is an explicit set of instructions designed to perform a specific task, allowing machines to work smarter and faster. The algorithm is trained using sets of relevant data and repeatedly tested until data points are ranked, aggregated or newly generated (Kumar, 2018). With time, algorithms develop their accuracy as thousands to billions of data points are continuously introduced to the algorithm (Costa, 2020). This process of consistently exposing an algorithm to new data sets and experiences improves efficiency of the machine being used. The output of the algorithm is

referred to as a prediction or decision depending on the context of the situation – this could be either a medical diagnosis or enrolment into specialized care. The predictions are often 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 computer algorithmic, which then leads to unfair outcomes and gives privileges to one group over another. (Mehrabi et al., 2019). Algorithmic biases tend produce predictions that replicate hidden stereotypes and forms of discrimination that live within digital platforms and exacerbates unfair outcomes among those who are marginalized. (Mehrabi et al., 2019). The challenge is that these predictions are used for future training and the model becomes amplified by inaccurate results, and therefore normalizes that process. This phenomenon is known as a feedback loop and can occur between the data and algorithm, as well as the algorithms and user interactions (33;28). Although some biases may seem harmless, or mildly inconvenient, if left unchecked they become a part of the rationale for simple algorithms.

Types of Algorithmic Bias in Health Care Systems and Data Sets

A wide range of biases is discussed in the literature that emerge within datasets, data collection, annotated and deployment processes. A common type of bias is *representation 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) that 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). Another example are datasets on HIV/AIDS infection rates initially characterized white men as the disease's primary target, although the condition disproportionately impacted Black and people of color. (Kaiser Family Foundation, 2019).

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 computing equitable predictions. Electronic health records (EHR) are used routinely by health care practitioners to document a patient's condition and relevant information. However, functionalities within EHR systems affect the completeness or precision of the recorded data. Gianfrancesco et al. (2019) highlights that a source of bias within EHR data is missing data, where there are health records may only contain information for severe cases of a particular population compared to another. 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, certain experiences and conditions of various patients may not exist in sufficient number for a predictive algorithm (Gianfrancesco et al., 2019). The underestimation or overestimates can lead to uninformative predictions that can have influence decisions around clinical support.

Aggregation bias is another form of biases 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).

Historical bias is a form of bias that refers to the existing biases and socio-technical issues that seep into the data generation process (Suresh et al, 2019; Crawford, 2016; Anonymous Authors, 2020). Studies have shown that Black, Hispanic, and Asian patients are often given less pain medication for similar injuries and reported pain level based on stereotypical and cultural myths that Black people have fewer sensitive nerves or thicker skin (Goyal et al., 2015; Pletcher et al., 2008). Using the National Ambulatory Medical Care Survey, an analysis was conducted on patients aged 21 years or younger who received a diagnosis of appendicitis in the emergency department to calculate how often opioid and non-opioid medications were administered (Goyal et al., 2015). The data indicated that race was associated with pain medication, where Black patients with severe pain were less likely to receive opioid that White patients. However, there were not differences in overall analegesia administration. The findings suggest that pain may be equally recognized across racial groups, however physicians are reacting differently to treating similar levels of pain. The researchers demonstrate that may be a higher threshold of pain predictions for administering such medications to Black and racialized patients. It highlights the role of implicit and explicit personally mediated biases as a factor influencing health care delivery and health outcomes among racialized populations.

The quality of the health services does not only impact health, but also lack of or limited access to the health services greatly impacts an individual's health status. For example, when individuals do not have health insurance or may not be able to afford certain services, despite having health insurance. Chen, Szolovits & Ghassemi (2018) used MIMIC-III data to assess differences in error rates for insurance type to indicate ways public insurance versus private insurance affects patient care in intensive care unit and psychiatric settings. Their study demonstrates an association between insurance type and access to care, as they found significant differences in ICU mortality for race, gender, and insurance type, as well as 30-day psychiatric readmission for insurance type (Chen et al., 2018). With the rising health care costs, many may struggle in paying for needed health care services or keeping up with their physician advice or treatment plan, leading to increased rates of poor health and mortality. It is therefore important that policies around health care insurance and coverages confront the nonmedical factors that tend to affect people's health and access to care so dramatically. These factors are known as the social determinants of health and include whether a person lives in a safe and high-quality housing, live is close proximity to fresh produce and healthy food, or how easily one can access transportation. Studies have shown that SDOH account for 60% of people's health outcomes (Canadian Medical Association, 2013). Schmutte, Dunn, & Slegde found that housing and employment were considered more accurate measures for psychiatric readmission rather than mental illness itself, demonstrating ways social factors mediate one's ability to seek and access care. Therefore, further research and policies around how these determinants of health intertwine with data collection process and analysis to ensure machine learning algorithms can produce predictions that account for nonmedical factors and shed light on structural factors such as insurance or hospital policies, occupation type, working condition as a hinderance to care for patients.

Evaluation bias, the use of inappropriate benchmarks, has been well-documented within facial recognition systems that are often biased towards skin color and gender (Buolamwini & Gebru, 2018). Adamson & Smith (2018) conducted multiple studies and found demographic imbalances in dermatology. Black Americans had the highest mortality rate for skin cancer, and screening systems and doctors are not trained on darker skin types. Researchers argue that evaluation bias closely aligns with *funding bias*, when biased results are reports to support or satisfy the financial supporter of the study (Mehrabi et al., 2019). This also has implication on the data selection and analysis approach, which ties back to the representation and the other biases discussed above. As such several forms of biases tend to work simultaneously at various stages of the machine learning 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 including computer scientists and physicians (Chaney, Stewart, Engelhardt, 2018)

Measurements of Fairness in Health Care Data Sets

Within the computer science literature, fairness is broadly defined as "the absence of any bias based on an individual's inherent or acquired characteristics that are irrelevant in the context of decision making" (Chouldechova, 2017). These characteristics tend to be protected identities such as gender, race, religion, skin color, age, or ethnicity in many contexts. Mehrabi et al. (2019) categorizes the fairness definitions used across the literature into individual, group and sub-group fairness. Notions of "individual fairness" refers to treating similar individuals similarly, whereas "group fairness" is defined as when different groups are treated equally (Dwork et al., 2012; Kusner et al., 2017). Liu et al. (2017) use the definition of calibrated fairness, where one is selected in proportion to their merit. In the sections below, this paper provides more detail as to how these definitions of fairness are measured and its implications.

Across the literature, there are several definitions of fairness. McDonald & Pan (2020) discuss that the absence of intentional bias does not equate fairness, rather unfairness is a characteristic 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, however, 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 could be correlated with race and other socially oppressed identities. Zemel et al. (2013) discuss representation learning, where one learns latent representation to minimize the groups information, however, challenges arise where too much information and misrepresentation may occur. 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 diverse groups and defined by the protected attributes. Models are considered fair when its 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, where the false positive rate and the proportion of positive decisions are considered the two most popular, and readily used by researchers in the machine learning field (Hardt et al., 2016; Argarwal et al., 2018; Chouldechova, 2017).

A popular form of classification parity is defined as the *equalized odds criterion*, which defines a model as fair if its false negative rates and false positive rates are equal across groups (Hardt et al., 2016). Predictions where only false positive must be equal are referred to as the *equal opportunity* (Hardt et al., 2016; Agarwal et al., 2018). The problems with classification parity are that the risk distribution measures of protected groups will likely differ and yield error rates that will also differ across groups, a phenomenon known as infra-marginality (Simoiu et al., 2017). Pierson (2020) highlights this concept using positive COVID-19 rates to assess racial disparities in undertesting revealing higher positive rates for non-White populations. The researchers explain that we incorrectly conclude from high positive rates because we are measuring the variables at face value, rather than computing the probability

threshold of patient's likelihood to be tested. This issue also speaks to the broader social determinants of health that shape our living and working conditions, hence one's chances to access resources such as COVID-19 testing would greatly depend on these factors. 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.

Chen et al. (2018) have described challenges to assessing parity measures in practice and propose to examine the differences in the cost function between the groups to demonstrate the expected discrimination level. They first define fairness in the context of loss, then formalize unfairness as group differences (Chen et al., 2018). The computed discrimination level can be decomposed into differences in bias, variance among groups, and differences in noise, which is often referred to as Bayes error – the lowest possible error rate of any outcome.

The third widely considered definition is *calibration*, meaning that risk estimates are not dependent upon individual attributes conditional on the algorithm's predictions (Davies & Goel, 2018). Chouldechova (2018) refers to calibration as the assessment that reveals the same positive predictive values across group, however, these values may be in conflict with error rate balances. Therefore, there are discrepancies with those who are predicted to exhibit the condition tested and whether that algorithm made an effective prediction.

Kusner et al. (2017) studied fairness under the framework of casual inference and shows a prediction is fair towards a person if it is same in the actual world and a counterfactual world if they we apart of a different demographic group. 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) demonstrate this concept through their research further discussed below.

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. As mentioned earlier, many algorithm predictions already hold biases, whether from the data collection phase to the annotation, produce feedback loops that will inevitably create unintended consequences that can have long-term and generational impacts. The data quality issues with the anti-classification approach speaks to representation or selection bias where the data sets may not accurately reflect or appropriately capture the experience of diverse people. By grouping the different subgroups into one classification, there is a risk of maintaining aggregation bias, where it may seem as though each group is assessed equally, however, there are drastic differences across other variables that influence health or social status. Thirdly, calibration processes points to the issues of measurement bias, where labels and annotations capture the interest of the model's design purpose rather than the actual needs of the population. Across these three definitions, historical and social biases can arise by the quality of the data sets used and the worldviews of scientists developed and monitoring the algorithm. As such these well-known definitions of algorithmic fairness cannot readily detect discriminatory biases and can therefore be used to perpetuate biases when these algorithms are designed to satisfy them (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	Implications of Biases	Real-life example
Anti-Classification	Two groups are given the same values of social identities such as race, have the same decision and these features do not change the decision	Social identities such as race and gender are 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
Classification Parity	The proportion of positive predictions should be the same across all groups	Providing the same level of risk for all groups can allow more equal assessment of harm across each individual	True underlying distribution of risk varies across groups and can lead to difference in error rates 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

The first is an *anti-classification*, where social identities such as race, gender are not explicitly referenced, however, 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 could be correlated with race and other socially oppressed identities. Zemel et al. (2013) discuss representation learning, where one learns latent representation to minimize the groups information, however, challenges arise where too much information and misrepresentation may occur. Davies & Goel (2018) suggest that representative data is most critical using this approach as the addition and use of protected attributes is justified.

Case Study: Promising strategies to accurate measurements of bias

This section will further explore one case study to exemplify a promising approach to developing algorithms and measuring fairness. Last year, Obermeyer et al. (2019) revealed that a widely used algorithms that identifies patients that get additional health care services prioritizes healthier, White patients ahead of Black patients who were sicker and needed the services more. The scientists addressed the label of the bias used to make this prediction to compute a more accurate number of Black patients that would qualify for such services based on their health needs.

Background of the study: 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). 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 to establish health needs predictions

Methods and fairness analysis: The algorithm studied utilized 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 equal 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 connecting 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 Black patients are likely not receiving the care they need. The researchers that the algorithm used to predict health needs, is in actuality, a prediction of health costs.

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 the machine learning algorithm process, 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. This challenge could be attributed to the access barriers

to critical data sets and processes that inform the algorithmic prediction. Several researchers agree that there is a large focus on individuals rather than the structural factors that shape health disparities, and it is critical it uncover these factors and address them directly (Crawford, 2019; Mehrabi et al., 2019; Obermeryer et. al, 2019).

In the case study explore, the researchers were able to create an alternative algorithmic that specifically addressed the health needs of an individual and accounted for social needs as well. 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 algorithm developer to address the root issue of the apparent racial bias. For the Black population particularly, the researchers emphasize that many have reduced trust in the health care system, which can be seen has a compounded factor in reducing the use of care and therefore the generation of diverse data sets. 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 structural and interpersonal mechanisms of bias that shape a person's experience with the health care system, and those of Black and racialized population are not consistently captured or assessed within algorithmic designs, data sets, and evaluation procedures.

The biases that exist within the data sets and among those that develop the labels within the algorithms can be a direct causal pathway to exacerbating structural inequalities. ML methods in health should therefore be created and critiqued with a deep awareness of multiple perspectives and methodologies to address the inequalities reinforced (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 design and interpretation models. Intersectional approaches are reflexive in nature, which can help rework stereotypes and reprogram technologies of agency, rather than inaction on reducing inequities (Ciston, 2019). Although this may be a costly and timely approach, it ensures that communities and technologies are developed many people feel valued and appropriated assess and treated.

As the field continues to emerge and decipher the ways bias is perpetuated within data sets and model analyses, it is critical that the downstream effects do not overshadow discriminatory and political contexts that makes automated tools an area of concern to begin with (Benjamin, 2020). It is evident that these biases create a cycle that propels the lack of research and investment into the realities of vulnerable populations, which can be linked to the poor measures of their needs within machine learning algorithms. Therefore, definitions of fairness should move from assessing individual risks to evaluating the production of risks by institutions to prevent greater yet subtle influences of harm. This can help to ensure medical institutions and health care providers are being held accountable for inequitable outcomes they perpetuate. In addition, a greater level of transparency in data inputs and algorithmic labels can enable researchers and health care providers accurately assess fairness and better assess the health needs of vulnerable patients.

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 be more 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. 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 intentionally 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 to complete this paper and prepare of the final presentation and web page. I would also like to thank the Toronto Data Workshop 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*

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