The Perpetuation and Exacerbation of Racial Bias in Healthcare Through the Use of Machine Learning

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Zachary Mills

Spring 2024

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisors

William F. Stafford, Jr., Department of Engineering and Society

Machine learning (ML) and artificial intelligence (AI) are becoming increasingly popular, and they are being used in almost every industry today. These technologies are seemingly ubiquitous now, but most people do not understand how they work. Machine learning and AI use high level math and linear algebra to analyze data and identify patterns in the data at the same or even higher level than humans. These patterns and insights can be used to make predictions or classify previously unseen data. Because of this, the performance and accuracy of these models largely depends on the quality of the data being used. Nonrepresentative, small, or inaccurate data can not only make these technologies ineffective, but perhaps more dangerous, it can introduce bias into a seemingly accurate model.

The medical industry is no exception to the development of ML and AI technologies. In recent years the amount of research into machine learning for medical applications has dramatically increased. There are now technologies that exist that can outperform physicians such as algorithms that detect eye diseases (Abràmoff et al., 2018) and apps that can diagnose skin cancer using only a cell phone camera (Freeman et al., 2020). These technologies have the potential to increase the quality of care experienced by millions as well as increase the efficiency with which the industry operates. As these technologies begin to be widely adopted into clinical use with real patients, the danger of bias in the models becomes ever more pressing. If models with undiscovered biases are being extensively utilized, then the biases will become engrained in the system and cause millions of patients across the country to be affected. Because most of the people using these new technologies do not fully understand how they work, they will just blindly trust the results and use them without doubt. This will cause racial bias to become a systematic piece of the healthcare process.

Machine learning models are highly dependent on the data that is used to train them; the most important determining factor in the performance and effects of a machine learning model is the data. If low quality data is used, then it is likely that low quality results will be given by the algorithm. In the same way, if data is biased, then the results from the model will be biased as well. But how does bias appear in data that, in theory, should be impartial? One way that bias can enter into data is through poor data collection techniques. Sometimes researchers and others involved may intentionally or unintentionally collect data in ways that do not fully or correctly capture the information. This can come from poor sample selection, missing data, unnecessary features, and many more.

While poor data collection is one way bias can enter data, it is easy to identify and correct, and it is not common in professional research. Another potential explanation is that social patterns that appear in human behavior can be reflected in data. This causes bias to implicitly appear in these technologies without the intent of the developer (Johnson, 2021). A major part of this is the use of proxies for socially sensitive attributes that may be hard to quantify on their own. For example, while race may be excluded from the data, zip code may be used as a proxy to a similar effect (Johnson, 2021). This leads to preexisting societal biases being unintentionally adopted by ML algorithms. This problem arises when bias is ubiquitous in the environment, as it has historically been in the healthcare industry. This presents a problem that even when good data collection techniques are used, bias will still exist in the data. Even when trying to remove sensitive features from data, people and algorithms will often utilize a proxy that mimics the original attribute. For example, if a philosophy company wanted to not hire based on gender, they could remove gender from their consideration and utilize publication rate to make their decisions. While this seems like an objective fix to the problem, it maintains the

stereotype because the percentage of papers published in philosophy by women is low (Johnson, 2021). This raises the issue of how to deal with proxies, because as you try to limit proxy attributes the less effective models become. All of this shows that cultural biases are often mimicked in algorithms due to their ubiquitous nature and are often difficult to correct without harming effectiveness. It is highly likely that this issue would be present in healthcare where there already exists a large amount of implicit bias.

Another method for how bias enters data, is through misrepresentative or under representative datasets. When data doesn't represent all conditions equally, ML algorithms trained on it will often favor the conditions that are most present. An example of this phenomenon can be seen in a case where a machine learning approach was used to predict outcomes for an anticoagulant drug by predicting the international normalized ratio (INR), a measure of the clotting time, in patients. An ANN model was trained to predict this value and was found to be 30% more accurate than physicians on average (Mac Namee et al., 2002). However, the model performs so highly because it mostly predicted the most commonly occurring outcomes. This happens because most of the training data is composed of these types of samples and the less common cases do not have as much data (Mac Namee et al., 2002). While the overall accuracy of the system is good, it is often the case in healthcare situations that it is more important to identify the rare outcomes rather than the common ones. This makes this tool less clinically useful than the surface level analysis would lead to believe. It is possible that because minorities make up a small percentage of data points, partially due to historical lack of access, that algorithms will weigh them as less important. This is a potential area to look for and identify in specific cases.

A specific example of how the implicit biases of physicians and other health care professionals can enter data can be seen in a study on the content of notes from these professions. This study performed a natural language processing analysis on over 1.8 million caregiver notes from Beth Israel Deaconess Medical Center in Boston from 2001 to 2012 (Markowitz, 2022). The study measured positive emotion terms, negative emotion terms, body terms, impersonal pronouns, analytic thinking, and cognitive processing terms. The analysis found that physicians focused more on emotion for women, while writing more about the body for men. This supports the hypothesis that women are often seen as more emotional and hysterical in medical settings (Markowitz, 2022). It was also found that more impersonal pronouns were used for women compared to men suggesting that physicians tend to psychologically distance themselves more from their female patients. Physicians attending to Black/African patients used fewer emotion terms compared to white patients and focused less on pain and negative experiences of these patients. This aligns with the already observed stereotypical belief in physicians that black patients experience less pain than white patients (Markowitz, 2022). This specific example shows that implicit bias in health care workers does exist and that it can be seen in data.

These all show the different ways that bias can enter data and what needs to be looked for when determining the core cause of bias in a machine learning technology because these data issues are the root cause of bias in the algorithms. Looking at some specific examples of these technologies that were used in practice can give even more insight into the problem. These cases will illustrate how a bias came to be in a machine learning model, and how the use of the model directly led to a lower level of care for racial minorities.

In the United States, there exists vast racial disparities in maternal and gynecological care. Black women are at three to four times the risk of dying due to childbirth-related causes

and are twice as likely to experience sever maternal morbidity (SMM) which causes extreme consequences to the mother's health (*Pregnancy Mortality Surveillance System | Maternal and Infant Health | CDC*, 2023). A particular case where this is present is in the recommendation of vaginal birth after a previous Cesarean procedure. Historically, all patients who have undergone a Cesarean have been recommended to avoid natural birth due to increased risk for infection or rupture of the scar from the procedure. However, recently some patients who meet certain criteria that reduce the risk of adverse effects have been cleared for vaginal birth after Cesarean (VBAC). VBACs are often seen as preferable to another Cesarean because they avoid surgery, have shorter recovery times, and have lower risk of complications for later pregnancies (Curtin et al., 2015).

The VBAC calculator is a tool endorsed by the National Institute of Child Health and Human Development used to estimate the chance of success for VBAC depending on patient-specific factors such as age and body mass index (BMI). The algorithm also directly uses race as a factor in the prediction; African American and Hispanic women are automatically given a lower predicted success rate than white women who have otherwise the same factors (Vyas et al., 2019). When calculating the probability of success, the algorithm directly lowers the prediction if the patient is African American or Hispanic. In the case of a 30-year-old woman who has a BMI of 35 and has had 1 prior cesarean, they are given a 46% chance of success if they are identified as white compared to a 31% chance if they are identified as African American or Hispanic (Vyas et al., 2019). This means that this technology systematically recommends black and Hispanic women for C-sections at a much higher rate.

The reasoning behind including race in the calculation was due to a study that found a correlation between race, marital status, insurance type, and other societal factors to VBAC

success (Landon et al., 2005). However, all these factors except for race were removed from the final calculator. All other factors in the final product had some form of biological relevance, except race, which only had only a loose connection to pelvic shape largely rooted in racism and eugenics. Multiple studies have found that similar calculators that do not utilize race are effective in predicting the risk, yet this widely used calculator choose to still utilize race as a factor (Vyas et al., 2019). Thus, this technology sets a dangerous precedent of including race in biological calculators and systematizing the racial inequities that have long plagued the medical system.

Access to high quality health care regardless of race is an important piece to building an equitable health system. This means it is even more important to investigate technologies that act as gatekeepers to care. Because high level health care is expensive, providers want to ensure that only the patients that need the care are recommended for it. To achieve this, they utilize a risk-prediction algorithm. One in particular is widely used and affects an estimated 200 million Americans (Obermeyer et al., 2019). This algorithm assigns patients a 'risk score' which is used to recommend patients to more care including increased resources and greater attention. It has been found that this algorithm has been assigning black patients lower risk scores than white patients who are experiencing the same level of symptoms. This has lowered the proportion of black patients receiving additional care from a predicted 46.5% to a meager 17.7% (Obermeyer et al., 2019). To understand why this is occurring, a deeper look must be taken at the data and specific algorithm being used.

This algorithm is race blind, meaning that it does not account for race in any manner. The introduction of bias is more subtle than in the previous case. The problem is because the algorithm is not actually predicting health risk, it is predicting health care costs. While this seems like a viable substitution, as more harmful illnesses require more expensive treatment, it actually

introduces bias. Historically black patients have had unequal access to healthcare which resulted in lower costs compared to white patients (Obermeyer et al., 2019). This trend was evident in the data and as such it predicted lower costs for black patients even if their symptoms were severe. In this case, the health care costs served not only as the labels for the data, but as a proxy for race to be used in decision making. This proxy caused the algorithm to still make biased predictions that harmed black patients.

In order to address this bias, new labels were tried in replacement of cost of care. Ultimately, a combination label that used both health and cost factors to determine risk scores was decided upon. This new label reduced bias by a reported 84% suggesting that biases that arise from label selection can be corrected when identified (Obermeyer et al., 2019). This illustrates the importance of identifying labels that are not only effective, but do not achieve their goal at the expense of a minority group. It is likely that there are many technologies in use today where these issues still exist and could be corrected.

Another example of a gatekeeper technology is a scheduling algorithm. These programs seek to reduce average patient wait time and provider down time by predicting no shows. Once patients are identified as a potential no show, they are scheduled into overbooked times with the assumption that some patients will not show up for their appointment. Patients scheduled for these slots often experience longer wait times and less time interacting with a provider causing an overall lower quality experience than others. The dataset used to train the algorithm was acquired from a clinic whose black patients no showed at a higher rate than nonblack patients. Because of this trend in the data, the algorithm predicted higher probability of no showing for black patients. This resulted in a 30% increase in wait times for black patients (Samorani et al., 2022).

Like the first example, race is being directly used as part of the decision-making process in a way that negatively affects black patients. The solution, however, is not as simple as removing race from the data. This approach was attempted, as the researchers tried a race-blind solution to the problem, but it was unsuccessful in both reducing scheduling cost, calculated with patient wait time, provider down time, and provider over time, and racial fairness (Samorani et al., 2022). This is likely due to the appearance of proxies in the data. After this method failed, they tried a race-aware method. This approach did not try to reduce the average wait time of all patients, but instead it focused on reducing the wait times of the largest affected racial group. This method resulted in an equivalent reduction in schedule cost of the original model, while also eliminating any racial disparities that had been present (Samorani et al., 2022). This case shows how race-blind approaches are not always the solution because of bias entrenched in data; rather, more proactive race-aware models that purposefully seek to reduce bias are sometimes needed.

Looking at these cases shows that there are multiple ways that bias emerges in ML technologies including direct negative use of race in the algorithm, bad data collection techniques, and proxies. It can be seen that bias emerges in the data even when race is removed. This is due to the proxy factors that exist due to large amounts of historical bias that is present in the data. This means that a large amount of data collected in medical institutions, especially those that back track later in time, may contain racial bias trends. This problem in the data is not likely to be able to be corrected with standard data cleaning or even be identified before algorithm development. This makes this problem especially dangerous because it is difficult to identify, requires diligent work to correct for, and may even require a large-scale change to the algorithm after development and training.

There is no one perfect solution to fix this problem, sometimes race-blind solutions work, but other problems will require proactive race aware solutions like the one seen in the scheduling algorithm example. Developers will likely need to go out of their way to make solutions that acknowledge and correct for the bias. It is important to test algorithms before deployment to ensure that not only do they work, but that they do not include any kind of bias. This kind of testing is likely to look different for different kinds of algorithms, but it is important that a standard is set that there is some statistical evidence that a model is not biased before it is used in clinical settings. If successful, these technologies can correct for bias and transform the health care industry in a way that is more accurate and more efficient. If not, they can perpetuate and even exacerbate the bias in the industry causing systematic bias to become entrenched in our healthcare system.

However, acknowledging that this problem exists is not enough to remedy it. In order to accomplish that it is likely that some kind of national mandate would be necessary. All new machine learning algorithms attempting to enter the health care space should be subject to mandatory bias testing on top of the typical effectiveness testing to ensure that they are both fair and accurate. It is paramount that research is done to produce standardized testing mechanisms that utilize empirical and statistical methods that ensure people of all races are treated equally. If this is done a regulatory agency such as the FDA would be able to require these technologies to present statistical evidence of a lack of bias before granting approval. This would help to ensure that bias is not being made worse by new ML algorithms.

Machine learning has the potential to revolutionize the way health care is managed and performed, but it also poses the danger of systematizing the racial disparities suffered by many. It is important that action is taken now in the early stages of this technology before more damage

occurs. With proper oversight, machine learning can make health care better and more accessible for all.

References

Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *Npj Digital Medicine*, *1*(1), 39. https://doi.org/10.1038/s41746-018-0040-6

Curtin, S. C., Gregory, K. D., Korst, L. M., & Uddin, S. F. (2015). Maternal Morbidity for Vaginal and Cesarean Deliveries, According to Previous Cesarean History: New Data From the Birth Certificate, 2013. *National Vital Statistics Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, *64*(4), 1–13, back cover.

Freeman, K., Dinnes, J., Chuchu, N., Takwoingi, Y., Bayliss, S. E., Matin, R. N., Jain, A., Walter, F. M., Williams, H. C., & Deeks, J. J. (2020). Algorithm based smartphone apps to assess risk of skin cancer in adults: Systematic review of diagnostic accuracy studies. *BMJ*, m127. https://doi.org/10.1136/bmj.m127

Johnson, G. M. (2021). Algorithmic bias: On the implicit biases of social technology. *Synthese*, *198*(10), 9941–9961. https://doi.org/10.1007/s11229-020-02696-y

Landon, M. B., Leindecker, S., Spong, C. Y., Hauth, J. C., Bloom, S., Varner, M. W., Moawad, A. H., Caritis, S. N., Harper, M., Wapner, R. J., Sorokin, Y., Miodovnik, M., Carpenter, M., Peaceman, A. M., O'Sullivan, M. J., Sibai, B. M., Langer, O., Thorp, J. M., Ramin, S. M., … Gabbe, S. G. (2005). The MFMU Cesarean Registry: Factors affecting the success of trial of labor after previous cesarean delivery. *American Journal of Obstetrics and Gynecology*, *193*(3), 1016–1023. https://doi.org/10.1016/j.ajog.2005.05.066

Mac Namee, B., Cunningham, P., Byrne, S., & Corrigan, O. I. (2002). The problem of bias in training data in regression problems in medical decision support. *Artificial Intelligence in Medicine*, *24*(1), 51–70. https://doi.org/10.1016/S0933-3657(01)00092-6

Markowitz, D. M. (2022). Gender and ethnicity bias in medicine: A text analysis of 1.8 million critical care records. *PNAS Nexus*, *1*(4), pgac157. https://doi.org/10.1093/pnasnexus/pgac157

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Pregnancy Mortality Surveillance System | Maternal and Infant Health | CDC. (2023, March 31). https://www.cdc.gov/reproductivehealth/maternal-mortality/pregnancy-mortality-surveillance-system.htm

Samorani, M., Harris, S. L., Blount, L. G., Lu, H., & Santoro, M. A. (2022). Overbooked and Overlooked: Machine Learning and Racial Bias in Medical Appointment Scheduling. *Manufacturing & Service Operations Management*, *24*(6), 2825–2842. https://doi.org/10.1287/msom.2021.0999

Vyas, D. A., Jones, D. S., Meadows, A. R., Diouf, K., Nour, N. M., & Schantz-Dunn, J. (2019). Challenging the Use of Race in the Vaginal Birth after Cesarean Section Calculator. *Women's Health Issues*, *29*(3), 201–204. https://doi.org/10.1016/j.whi.2019.04.007