Biased Healthcare Algorithms Built Using Discriminatory Medical Data

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

Matthew Burkher
Spring, 2021

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

Signature _____ Date _____
Matthew Burkher

Approved __ H Rogers _____ Date _5/4/2021___
Hannah Rogers, Department of Engineering and Society

Abstract

Computer algorithms in all fields have been shown to be biased, but none are more dangerous than those in healthcare. Thus, algorithms in the healthcare system perpetuate systematic discrimination because they are created using historically biased data, and the groups creating these algorithms lack an understanding of these biases. In order to present this point, this research paper utilizes methods of documentary research and Wicked problem framing. This paper studies past research that analyzes bias in various healthcare algorithms and medical frameworks that many are built upon. A Wicked Problem can be described as a symptom of another problem. In this case, analysis of biased medical algorithms shows connections to a long history of racism and discrimination, which is further amplified in these technologies. The algorithms studied are created to be equal and non-discriminatory, but they often overlook the bias that already exists in the data, and since these large data sets are used again and again, these algorithms seem to be the discriminatory part, but in fact are only the part that show the output of discriminatory data. This seems like a problem that can be directly addressed, but the solution is complicated by the lack of social justice expertise in the creation of these algorithms and missing actors that would be able to understand the biased data and the reasons behind these biases.

Biased Healthcare Algorithms Built Using Discriminatory Medical Data

Introduction

Many computer algorithms perpetuate systematic racism and discrimination, from Amazon's AI hiring tool discriminating against women, or Megvii's gender-recognition software, which was 99% accurate for white men, but only 35% accurate for dark-skinned women (Hamilton, 2018; Cossins, 2018). However, there are few discriminatory algorithms that are more life-threatening than those in healthcare. Medical algorithms serve purposes such as diagnosing illnesses, recommending further and specified treatments, and many others. As doctors are relying more and more on these technologies, it is increasingly important that these technologies are not only accurate, but also fair. One study on a popular algorithm used to allocate resources and additional medical care to patients shows that it is much more likely to recommend white patients over black patients, despite having the same level of severity of illnesses (Obermeyer et al., 2019). The algorithm Obermeyer studied is even race-blind, like many others in order to treat all patients equally, but eliminating racial or socioeconomic factors is not enough and is even dangerous at times. Medical algorithms continue to discriminate against race, and the reason lies within the history and the context of the data being fed into healthcare algorithms. Understanding this already biased data is crucial in fixing existing discriminatory and biased technologies, and preventing the creation of more racist and discriminatory algorithms. Thus, algorithms in the healthcare system perpetuate systematic discrimination because they are created using historically biased data, and the groups creating these algorithms lack an understanding of these biases.

Frameworks and Methods

This research paper will be looked at through the theory of the Social Construction of Technology (SCOT) to further understand how society and human interactions shaped and led to discriminatory algorithms within the healthcare system. Proponents of SCOT thought "it was necessary to show that the development, stabilization, and even working of technology are socially constructed" (Bijker, 2015, Pinch & Bijker, 1984). SCOT will not only be used to analyze how the technology has gotten to this point, but also how the role of underrepresented social groups plays in preserving medical discrimination.

To gather a solid understanding of the problem, this research paper will utilize documentary research and Wicked Problem Framing methods. Documentary research will provide evidence of a multitude of discriminatory algorithms. These case studies will help formulate a perception of why society created discriminatory healthcare technologies. Additionally, this research will give insight to historic discrimination that led to the creation of these biased technologies. From the evidence collected through documentary research, Wicked Problem Framing will connect biased healthcare algorithms to biased health data and reasons the data is biased in the first place. Design theorist Horst Rittel describes every wicked problem as a "a symptom of another problem" (Rittel, 1973). In this case, Wicked Problem Framing will exhibit the misunderstanding of connections between medical data and societal injustices and how discriminatory algorithms are a subset of systematic discrimination as a whole.

Cases of Discriminatory Healthcare Algorithms

Algorithms being blind to race and socioeconomic factors are not enough to make an algorithm non-discriminatory. In fact, it will often worsen the problem. There have been many

studies looking at racial and socioeconomic bias in medical algorithms and the healthcare system as a whole. In a previously mentioned study, Obermeyer studied an algorithm that uses a database of about 200 million people in the United States each year to preference patients to "'high-risk care management' programs". The algorithm uses "a large set of raw insurance claims" to assign a risk score to patients based on the health costs it believes a patient will incur. The algorithm does a good job in this respect, however there is a disparity between who needs medical treatment and who receives treatment. This was found to be correlated with race. "At the 97th percentile of risk score… Black [patients] have 26.3% more chronic illnesses than Whites", but Obermeyer et al. simulated the algorithm without biases, which showed "17.7% of patients that the algorithm assigned to receive extra care were black" whereas the proportion would be 46.5% otherwise (Obermeyer et al., 2019). As shown, there is a huge disparity between who should receive treatment and who ends up receiving it. Although Obermeyer studied one specific algorithm, it uses the same generic approach that many widely-adopted, risk-predicting software use in the health sector, so it can be assumed that most of these other algorithms carry the same problems. Since this method is so popular and the biases are not so obvious, it continues to go unchecked and be recreated.

Another software that is gaining traction in healthcare is artificial intelligence that "can assist with targeted overbooking by predicting which patients are most likely to no-show", allowing practices to try to maximize the patients seen in a given day. The trained model only uses "prior patterns of health care use and features of the appointment", leaving out all personal information such as ethnicity or socioeconomic status. Nevertheless, this does "not eliminate potential to propagate societal inequity, … [since] prior no-shows… [are] likely to correlate with socioeconomic status" and factors such as transportation, childcare, or work-life balance

(Murray, Wachter, & Cucina, n.d.). At first glance, this AI software might seem like an easy solution, since it is built on top of already existing electronic health record software, but the ease of implementation makes it even easier to overlook possible consequences of bias.

Similarly, a new "policy framework for distributing scarce critical care resources" during the COVID-19 pandemic uses a point system to determine which patients are in need of these resources. This framework is "color blind", and is based on a "Sequential Organ System Assessment score (SOFA)", which scores different vital organ systems based on their health and functionality, and the patients' life expectancy. What this framework "fails to address is the contribution of racism and health inequities to its scoring system… [and how] these health disparities, compounded by effects of poverty and social disenfranchisement, have resulted in lower life expectancies for Black [people] compared to U.S. Whites" (Williams et al., 2020). "Mortality data for the United States reveal that, compared to the white population, African Americans/Black [people] have an elevated death rate for 8 of the 10 leading causes of death" (Williams, D. R., & Collins, C., 2001). Statistics like these aren't taken into account and data are often fed into the training of machine learning algorithms without these checks. Despite it not being a computer algorithm, similar frameworks are often encoded into computer software to create these healthcare algorithms. There must be an emphasis on spotting biases long before an idea is turned into code and before the process is automated and even harder to detect.

Computer vision and extracting features from images is a widely used and advanced form of machine learning, which makes dermatology a prime field for machine learning technologies. However, accurate machine learning labeling requires a huge input of data to be trained on and if the data is not diversified enough, different groups may be less accurately labeled. Diversity in the data sets is extremely important in the creation of these software. "A 2009 analysis revealed

that 96% of participants in genome-wide association studies (GWAS) were of European descent". This research is extremely important in identifying genetic related diseases, and will likely be used in training future healthcare algorithms. "Tens of thousands of significant associations between genetic variants and biological traits have now been found, and many of these associations have helped geneticists to uncover biological mechanisms underpinning conditions from diabetes to schizophrenia". However, within the GWAS Catalog, people of African, Latin American, and Native American descent made up less than 4% of analyzed samples as of August, 2016 (Popejoy, 2016). The lack of diversity in medical data is a major concern in Dermatology, as well as other fields, since doctors are moving towards machine learning to detect melanoma. It is a major concern that dark skin types will be underrepresented, and therefore people with darker skin tones will be much more prone to misdiagnoses. Within the "Melanoma Project,… much of the patient data are heavily collected from fair-skinned populations in the United States, Europe, and Australia" (Adamson & Smith, 2018).

Attempts to Fix Discriminatory Algorithms

In addition to discriminatory algorithms that are blind to racial and socioeconomic factors, there are many that try to adjust their algorithm to create equity, but fail to do so. The American Heart Association (AHA) created a score-generator to predict a hospitalized patient's risk of death. The algorithm "assigns three additional points to any patient identified as 'nonblack', thereby categorizing all black patients as being at lower risk". (Vyas, 2020). A 2019 case study of this system provided evidence that "black and Latinx patients who presented to a Boston emergency department with heart failure were less likely than white patients to be admitted to the cardiology service" (Eberly et al., 2019). Additionally, in this case, "the AHA does not provide a rationale for this adjustment" (Vyas, 2020).

The Society of Thoracic Surgeons, a not-for-profit organization focused on surgical procedures within the chest, uses specialized calculators to estimate the risk of complications during surgery, which "include[s] race and ethnicity because of observed differences in surgical outcomes among racial and ethnic groups" (*About STS*). However, the authors have stated that underlying reasons for these differences are unknown. "An isolated coronary artery bypass in a low-risk white patient carries an estimated risk of death of 0.492%. Changing the race to 'black/African American' increases the risk… to 0.586%." (Vyas, 2020). These calculations steer minority patients away from life-saving surgeries.

Reasons for Healthcare Discrimination

Data used in training various healthcare algorithms are already discriminatory as people of color and poor socioeconomic status historically have had less access to healthcare, which has also often been lower quality, and as a result produces biased algorithms. Therefore, it is of utmost importance to understand the data that is fed into these technologies. Eli Caham, a researcher in health policy, describes generic healthcare algorithms as "the terminal node in the big data value-chain: the generation, sanitization, transmission, and storage of data all precede its final predictions… Those outputs should be viewed as inevitable byproducts of preceding inputs" (Cahan et al., 2019). Any algorithm that is trained on biased data will always produce biased results, especially with no understanding of that training data.

Biased data comes from a long history of discrimination in the United States. The effects of residential segregation can still be seen today and is a primary cause of racial disparities in health, at individual, household, and community levels. "It was imposed by legislation" from federal, state, and local governments, "enforced by the judicial system", and "legitimized" by

white supremacy in “major economic… [and] cultural institutions" of the time. Although research shows that the black-white gap in economic status and health has narrowed in the last half-century, it still affects the data used today and is “still a fundamental cause of disparities in diseases" (Williams & Collins, 2001). In a study focused on the correlation of residential segregation, disparities in access to health care facilities, and late-stage breast cancer diagnosis in Detroit, researchers noted that “access to health care for persons solely depending on mass transit diminishes greatly”.

Additionally, ‘late-stage cancer rates are significantly greater in areas with poorer mammography access” and “significantly increased rates of late-stage breast cancer can be seen in neighborhoods with greater black [residential] segregation or lower socioeconomic status” (Dai, 2010). In the previously mentioned study on GWAS, geographic location was a major cause in the lack of diversity. Geneticists tend to use existing data sets from “well-established medical centres”, and this data is often collected from the same geographic locations (Popejoy, 2016). Due to the aforementioned reasons, people may have less access to these major medical centers and although the need for a large abundance of data is important, it cannot overshadow the need for diversity. As a whole, “the research enterprise itself tends to collect more data and advance more quickly on problems that disproportionately affect those in society with more power and resources” (Robinson, 2020).

More disparities can be seen when studying gaps in different demographics when it comes to health insurance coverage. As of 2018, 9.7 percent of Black Americans were uninsured, while that number was only 5.4 percent for whites. “41.2 percent [of African Americans] were enrolled in Medicaid or some other… public health insurance”. (Taylor, 2020). Even within these government-aided healthcare programs there are imbalances. “A recent survey highlighted

that a third of very ill Medicare beneficiaries had trouble paying for prescriptions drugs and medical bills, and 40% of these patients said that they had exhausted all their savings" (*The Lancet Digital Health*, 2019). Additionally, states that "have not expanded Medicaid under the Affordable Care Act (ACA), African Americans and other people of color are most likely to earn too much to qualify for the traditional Medicaid program, yet not enough to be eligible for premium tax credits under marketplace plans" (Taylor, 2020). There are huge racial and socioeconomic disparities when looking at health insurance alone, and as seen before, Obermeyer's studied algorithm, among others, pulled data straight from raw insurance claims and thus it was trained on extremely biased data, which showed in the skewed outputs of the algorithm (Obermeyer, 2016).

Analysis

Discriminatory healthcare algorithms are not only a result of a long history racial and socioeconomic discrimination, but is a problem of general bias in machine learning and artificial intelligence as well. What ties these problems together is the lack of understanding of the data that these algorithms are trained on, and how the data is already biased. From a first glance, researchers generally follow seemingly sound logic. In the development of many of the earlier examples this paper addressed, developers "performed regression analyses to identify which patient factors correlated significantly with the relevant outcomes". However, "since minorities routinely have different health outcomes from white patients, race and ethnicity often correlated with the outcome of interest" (Vyas et al., 2020). The data generation process "[is] an inherently subjective enterprise in which a discipline's norms and conventions help to reinforce existing racial (and other) hierarchies" (Ford and Airhihenbuwa, 2010).

Dr. Whitney R. Robinson asserts that "structural racism is a critical body of knowledge needed for generalizability in almost all domains of health research", which "is especially true when inference relies on algorithms… to choose statistical models" (Robinson, 2020). Researchers are overly focused on the algorithms themselves, thus circumventing the crucial step of understanding the data fed into these algorithms. Many developers "offer no explanation of why racial or ethnic differences might exist." The times developers do offer rationales, they are often "traced to their origins… to outdated suspect racial science or to biased data" (Vyas, 2020). Overlooking core, systematic discrimination "is a decision to ignore structures that give rise to many and varied associations with health" (Robinson, 2020). With no preemptive focus on equity, research is centered around the majority and the minority groups are left behind. Even when biases are discovered, such as in the Society of Thoracic Surgeons' calculator, researchers attempt to modify their algorithm rather than fix the root of the problem (Vyas, 2020). It should be of utmost priority to discover the reasons for these differences, rather than to only take note of them and continue building medical algorithms around these differences.

Evidence has shown that many racial correlations within healthcare data are a result socioeconomic inequalities and racial and residential segregation. In addition, researchers do not fully understand the correlation between race and biological factors that are so important in personalized medicine and medical algorithms. "Relationships between race and health reflect enmeshed social and biological pathways… Most race corrections… operate on the assumption that genetic difference tracks reliably with race" (Vyas, 2020). As seen over and over again, there is no real understanding of the problem's roots. Oftentimes, researchers will try to correct for race, when the problem may lie in genetics, or vice versa. Racist medical algorithms are a result of systematic discrimination and a lack of diversity those researching these algorithms and

within the data collected. Attempts to prevent and fix algorithmic discrimination often fail or worsen the problem since researchers often do not recognize the exact cause of the bias within the data. Even if they are aware of biases, researchers and developers try to create a work-around within their algorithms. However, the real reasons for algorithmic healthcare disparities go unresolved.

Counterargument

Many software and algorithms in healthcare have proven to continue historical trends of discrimination and racism. However, many of these technologies are studied individually, rather than how they interact with physicians and healthcare workers. In an interview with Dr. Cathy Fang, the Vice President of Yitu Healthcare in China, Dr. Fang discusses the role of artificial intelligence in healthcare. "Ultimately, our AI system is designed as an assistive tool for doctors – doctors still make the final decision concerning patients' course of treatment or medication". For example, "lung cancer… presents a challenge to China's healthcare sector… [and] with the number of radiologists that are available in China, every radiologist would need to analyze at least 20,000 CT images a day to provide a diagnosis for every patient. This presents a risk of missed diagnosis or misdiagnosis due to fatigue or human error" (Bajpai, 2018).

While this paper has mostly focused on the United States healthcare system, healthcare algorithms have the ability to discriminate against any factor and will be dependent on the situation they are used in. Here and in many cases, AI is designed to be complementary to physicians' own training rather than replacing these professionals. Therefore, doctors can focus on face-to-face interaction or research, rather than focusing on much of the day-to-day busy

work. Of course, this system will only work equitably if those using the software are aware of the biases built within the technology.

Obermeyer's researched algorithm, which was meant to recommend patients to "'high-risk care management' programs", was also meant to be a complementary algorithm. Here, too, doctors and healthcare workers interpret the output of the algorithm, and have the final say whether the patient will receive further and more advanced treatment (Obermeyer et al., 2019). Similarly, the policy framework for distributing resources during the COVID-19 pandemic are only recommendations, and the doctors have the final say (Williams et al., 2020). However, since healthcare professionals are already extremely overburdened, and as they get acquainted with this software, they will begin to trust the outputs of the algorithms more and more and will use it less as a suggestion. Once healthcare workers are trusting of these technologies, doctors do not have the training to recognize biases that exist, nor the skills to circumvent them.

Conclusion

This paper has given many instances where algorithms and software in healthcare perpetuate systematic discrimination, however they are not developed with malicious intentions. The problem lies in the biased medical data the machine learning and artificial intelligence algorithms are trained on, and therefore society's actions have shaped the way these new medical algorithms and technologies were created. A long history of systematic racism and discrimination has forced minorities groups to have less access to quality healthcare, both in treatment and insurance, and this is reflected in healthcare computer algorithm outputs. And although history cannot be reversed, a better understanding of how the data came to be biased, and a prioritization of preventing the biases can help prevent further harm.

Since this is an extremely complex problem, and is a sub problem on other complex problems within society, there is no universal solution to biased healthcare algorithms. Sometimes it is crucial to incorporate possible discriminatory factors into algorithms to prevent bias, while other times it's harmful. Each algorithm is situational; however, solutions will start with a better understanding of the problems that algorithmic bias stems from, those being systematic discrimination and general bias in machine learning. Data must be checked for biases, inconsistencies, and all around diversity. Not only must developers of these technologies be learned in fields surrounding social justice and equity, but they must also bring in people with expertise in these fields. While this won't be a solution, it will help guide developers in a direction to where they will not perpetuate systematic discrimination any longer.

References

About STS. (n.d.). Retrieved April 05, 2021, from https://www.sts.org/about-sts

Adamson, A. S., & Smith, A. (2018). Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatology, 154*(11), 1247. doi:10.1001/jamadermatol.2018.2348

Bajpai, P. (2018, December 11). "AI cannot replace doctors in making the final decision on Patient's course of Medication TREATMENT": Dr Cathy Fang. Retrieved March 15, 2021, from https://www.biospectrumasia.com/opinion/27/12270/ai-cannot-replace-doctors-in-making-the-final-decision-on-patients-course-of-medication-treatment-dr-cathy-fang.html

Bijker, W. E. (2015). Technology, social construction of. *International Encyclopedia of the Social & Behavioral Sciences,* 135-140. doi:10.1016/b978-0-08-097086-8.85038-2

Cahan, E. M., Hernandez-Boussard, T., Thadaney-Israni, S., & Rubin, D. L. (2019). Putting the data before the algorithm in big data addressing personalized healthcare. *Npj Digital Medicine, 2*(1). doi:10.1038/s41746-019-0157-2

Cossins, D. (2018, April 12). Discriminating algorithms: 5 times AI showed prejudice. Retrieved October 01, 2020, from https://www.newscientist.com/article/2166207-discriminating-algorithms-5-times-ai-showed-prejudice/

Dai, D. (2010). Black residential segregation, disparities in SPATIAL access to health care facilities, and late-stage breast cancer diagnosis in metropolitan Detroit. *Health & Place, 16*(5), 1038-1052. doi:10.1016/j.healthplace.2010.06.012

Eberly, L. A., Richterman, A., Beckett, A. G., Wispelwey, B., Marsh, R. H., Cleveland Manchanda, E. C., Chang, C. Y., Glynn, R. J., Brooks, K. C., Boxer, R., Kakoza, R., Goldsmith, J., Loscalzo, J., Morse, M., Lewis, E. F., Abel, S., Adams, A., Anaya, J., Andrews, E. H., Atkinson, B., … Zon, R. (2019). Identification of Racial Inequities in Access to Specialized Inpatient Heart Failure Care at an Academic Medical Center. Circulation. Heart failure, 12(11), e006214. https://doi.org/10.1161/CIRCHEARTFAILURE.119.006214

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

Hamilton, I. (2018, October 10). Amazon built an AI tool to hire people but had to shut it down because it was discriminating against women. Retrieved October 01, 2020, from https://www.businessinsider.com/amazon-built-ai-to-hire-people-discriminated-against-women-2018-10

Murray, S. G., Wachter, R. M., & Cucina, R. J. (n.d.). Discrimination By Artificial Intelligence In A Commercial Electronic Health Record—A Case Study. *Health Affairs Blog*. doi:10.1377/hblog20200128.626576

Pinch, T. J., & Bijker, W. E. (1984). The Social Construction of Facts and Artefacts: or How the Sociology of Science and the Sociology of Technology might Benefit Each Other. *Social Studies of Science*, *14*(3), 399–441. https://doi.org/10.1177/030631284014003004

Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature, 538*(7624), 161-164. doi:10.1038/538161a

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. doi:10.1126/science.aax2342

Rittel, H. W., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences, 4*(2), 155-169. doi:10.1007/bf01405730

Robinson, W. R. (2020). Teaching yourself about structural racism will improve your machine learning. *Biostatistics, 21*(2), 339-344. doi:10.1093/biostatistics/kxz040

Taylor, J. (2020, May 07). Racism, inequality, and health care for African Americans. Retrieved March 01, 2021, from https://tcf.org/content/report/racism-inequality-health-care-african-americans/?agreed=1#easy-footnote-bottom-14

Vyas, D. A., Eisenstein, L. G., & Jones, D. S. (2020). Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. *New England Journal of Medicine, 383*(9), 874-882. doi:10.1056/nejmms2004740

Williams, D. R., & Collins, C. (2001). Racial residential segregation: a fundamental cause of racial disparities in health. *Public health reports (Washington, D.C. : 1974)*, *116*(5), 404–416. https://doi.org/10.1093/phr/116.5.404

Williams, J. C., Anderson, N., Mathis, M., Sanford, E., Eugene, J., & Isom, J. (2020). Colorblind Algorithms: Racism in the Era of COVID-19. *Journal of the National Medical Association*. doi:10.1016/j.jnma.2020.05.010