Automation in Medicine: Bias in Medical AI

A Sociotechnical Research Paper
presented to the faculty of the
School of Engineering and Applied Science
University of Virginia

by

Kelly Chen

May 11, 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.

Kelly Chen

Sociotechnical advisor: Peter Norton, Department of Engineering and Society

Automation in Medicine: Bias in Medical AI

Hospitals in the United States are adopting increasingly technological practices and seeing new benefits. The artificial intelligence (AI) healthcare market is especially expanding, and is expected to grow from 4.9 billion dollars in 2020 to 45.2 billion dollars by 2026. Physicians utilize medical AI for many usages, including risk analysis, inpatient care, medical imaging, and drug testing. The rapid growth of this market is attributed to new technologies and increased datasets (Markets and Markets, 2020). Despite the benefits introduced by medical AI in the United States, predictive algorithms also reflect common biases in society to the detriment of those receiving healthcare.

Among physicians, hospitals, insurers, patient advocacies, and medical technology vendors, some promote medical AI, but others caution that it can exacerbate social biases (Obermeyer et al., 2019). Obermeyer et al. (2019) found that racial biases encoded in predictive algorithms reduced the chances a Black patient will receive additional patient care from 46.5 to 17.7 percent. Dr. Danton Char, assistant professor of anesthesiology, perioperative, and pain medicine at Stanford University Medical Center, contends that "AI and machine learning may worsen the economic and racial disparities already inherent in U.S. healthcare" (Char qtd. in Ward, 2019).

The medical community faces a major challenge: How can the healthcare community effectively use medical AI in clinical care while reducing the systematic bias in its algorithms? The healthcare community can attempt to reduce bias in medical AI by providing reliable datasets, reducing human bias, and improving the doctor-patient relationship. Acknowledging the deep-rooted obstacles provides a better understanding of where medical AI is failing patients and will ultimately serve as a blueprint for improved algorithms.

Review of Research

In a widely used algorithm, racial bias was detected that decreased the number of Black patients identified for extra care by more than half because historically, less money were spent on Black patients than on white patients. The algorithm used past health costs as data, and falsely assumed that African American patients were healthier than Caucasian patients, even when the two racial groups were equally sick (Obermeyer et al., 2019). Bias in medical AI also leads to societal, legal, and monetary concerns for healthcare companies as well. A study of one of UnitedHealth's algorithms, Optum, showed that the medical AI was racially biased. As a consequence, New York regulators have been called upon to "either prove that the algorithm is not discriminatory or stop using it." If the company is found to be violating New York's antidiscrimination laws, then it may be financially penalized by the state (Watson & Marsh, n.d.).

In 2016, a team of Harvard researchers developed an algorithm to identify melanomas. The model was able to identify a skin lesion as "malignant" or "benign" based on images. The team trained the algorithm with more than 100,000 photographs of skin lesions, and the algorithm was able to detect 95% of the melanomas and 82.5% of the moles. However, the team was unable to account for skin lesions on dark-skinned individuals, as more than 95% of the images used to train the model was of light-skinned individuals. The possibilities of false positive and false negative results on dark-skinned individuals were too high for the team to use the algorithm reliably (Dutchen, n.d.). If this tool was used in a hospital setting, incorrect results may cause a patient to be misdiagnosed, which could lead to lethal consequences due to the severity of skin cancer and how quickly it progresses.

A group of researchers also examined the types of datasets many AI models were using. In this event, researchers from Stanford University examined several studies that used deep learning algorithms to perform image-based diagnostic tasks within different clinical categories. The researchers examined a total of 74 studies that met their criteria and were published between the years of 2015 to 2019. They found that many AI algorithms only used datasets from limited states. After their examination, the researchers discovered that 71% of the studies used patient data from mainly California, Massachusetts, or New York to train their algorithms. Around 60% used data only from the three states. Thirty-four states were not represented at all, while other states only had minimal data (Kaushal, Altman, & Langlotz, 2020). As an effect, some medical AI used in the examined studies resulted in skewed data that favored populations from the three states.

Human Bias

Bias in medical AI is largely attributed to human bias. Many physicians and hospital employees may have an implicit bias towards a certain group. According to Dr. Carmen Green, a professor of anesthesiology at the University of Michigan, "Emergency room clinicians may be choosing which patients get pain relief based on conscious, unconscious, and implicit bias as well as negative stereotypes based upon race, ethnicity, and class" (Green qtd. in Watson & Marsh, n.d.). Such biases are reproduced in medical AI, leading to skewed data produced by the algorithms. In order to train an algorithm, the developer must provide data that the algorithm can learn from. If the data are not representative of the true population, then the algorithm will learn incorrectly. The provided data may be taken disproportionally from certain groups, leading the algorithm to provide data that caters towards one group over another. Recent studies have shown that machine learning algorithms are better at detecting skin cancer in light-skinned individuals as opposed to dark-skinned. This is because the algorithm had been trained mostly with data from light-skinned individuals. Adewole Adamson, MD, MPP, professor at UT Austin Dell Medical School explains the study: “The algorithm is only as good as what you’ve taught it to do. If you’ve not taught it to diagnose melanoma in skin of color, then you’re at risk of not being able to do it when the algorithm is complete” (Adamson qtd. in Thakar, 2019). Thus, if an algorithm isn’t trained with representative data in the beginning, then it may not be able to perform accurately in the future (Thakar, 2019).

Implicit bias is especially common within the healthcare system, and is demonstrated across many studies. In one such study, Hall et al. (2015) conducted fifteen studies on health care professionals and medical students with the Implicit Association Test (IAT). This test was used to measure implicit bias against different ethnic groups. The team of researchers found that nine of the studies reported bias against Black individuals compared with white individuals, three of the studies showed bias against both Black and Hispanic people compared with white individuals, and one study examined bias against Hispanic individuals compared with white individuals. Further research reveals implicit bias exists in maternal care. Organizations such as the American College of Obstetricians and Gynecologists acknowledge that “racial bias within the healthcare system is contributing to the disproportionate number of pregnancy-related deaths among women of color. Providers spend less time with Black patients, ignore their symptoms, dismiss their complaints, and undertreat their pain” (Zephyrin, 2019). Such implicit bias is then reflected on medical AI algorithm results.

Another report detailed equity concerns with specific input variables programmed by those artificial intelligence which govern risk scores for different clinical utilities. In one example, a urinary tract infection (UTI) calculator was developed. This calculator gauges the risk of a UTI in children of two months to 23 months of age. The calculator assigns a likelihood to guide judgments for when to pursue further, more invasive urine testing. The calculator takes in the inputs of age, temperature, race, and sex. The algorithm was shown to assign a lower likelihood of a UTI if the infant is Black. Studies have shown that Black infants are less likely to suffer from UTIs. However, because the studies weren't intersectional, there is no proof that clinicians should not pursue diagnostic testing for Black infants with UTI symptoms. Caregivers don't take into account specific circumstances. The same research paper noted that "uncircumcised male infants less than 3 months of age... had the highest baseline prevalence of UTI." If this algorithm was tested on a Black male infant under three months of age with symptoms of UTI, it would tell the healthcare provider that the infant is low-risk, despite all other indicators. Another tool that uses an AI algorithm, Kidney Donor Risk Index, estimates the likelihood of a kidney graft failure. The algorithm increases predicted risks of kidney graft failure if the donor is identified as African American, has hypertension, or diabetes. Because this algorithm is more likely to strike out potential donors from African Americans, since African American patients are increasingly likely to get a kidney donation from another African American, the tool skews the racial inequity in equal access kidney transplants (Chen & Baker, 2006).

Complexities of the Data Collection Process

Defective data is one of the reasons why biases exist in medical AI. However, reliable data may be difficult to collect. A patient of Drs. Amit Kaushel, Russ Altman, and Curt Langlotz once asked, "Why is it that we can see a specific car in a moving convoy on the other side of the world, but we can’t see my CT scan from the hospital across the street?” (Kaushal, Altman, & Langlotz, 2020). Indeed, obtaining and sharing data for even a single patient is a difficult and lengthy process. Medical AI would need data from a vast number of patients to train. In order to gather less biased data, patient data would have to be easier to access. However, many privacy laws protect the data and patients expect confidentiality from their physicians. Many hospitals additionally hide data to prevent access from competitors.

Privacy is also an increasing concern within the public. If medical data sharing increased, hospitals would likely have to address public backlash. Many patients may be hesitant to release their information, and thus data may be missing due to the simple refusal of surveying. In fact, a cycle of mistrust leads to an inescapable lack of information. Groups of people that have been marginalized in the past are wary of medical professionals and healthcare, and thus refuse to release their information because of this distrust. The lack of information from these groups of people makes it inapplicable to them, and then when they get misdiagnosed, they distrust the system even more. Thus, collecting reliable data is more challenging with the current privacy conducts (Kaushal, Altman, & Langlotz, 2020).

Collecting consistent data for medical AI may be difficult as well. Different data standards and data collection policies may apply to several datasets. Two AI programs may have different data collection practices as well, and may be impossible to compare. Thus, two medical AI algorithms may yield dissimilar results while running similar services. Take collecting data on disabilities for example. According to the Victorian Government (n.d.), “medical services may be more likely to collect disability information by way of diagnoses and medical history, while non-disability specific services may be more interested in collecting information concerning support needs or a need for reasonable adjustments. As a result, the scope and detail of information collection may not be consistent across services, making it complex to compare data between services, or to population level data sets." Algorithms will thus be more difficult to test and verify.

Representation of Data

Even within voluntary participants, many groups are being underrepresented. In the recent past, women and minority groups have been underrepresented, and thus did not receive the same amount of care compared to other groups. The NIH, FDA, researchers, and an act from the Congress attempted to address this in 1993, but have not yet been completely successful (Kaushal, Altman, & Langlotz, 2020). Attempting to collect data from a more diverse pool may also cause delays. Moderna, a biotechnology company that developed a coronavirus vaccine, attempted to enroll a more diverse group for its clinical trials. The company especially tried to cater towards the elderly and minorities. When the trials first began, Moderna CEO Stephane Bancel predicted that the data could be available in October in "a really optimistic scenario; maybe November." By August 28, 2020, the company had enrolled a total of 17,458 participants. 24% of those participants were from minority groups. In late August, Bancel noted a delay due to the attempt to diversify the participant pool, stating that "a small delay to have a better quality is the right decision in the long run" (Bancel qtd. in Tirrell, & Miller, 2020). Many other medical datasets, however, disregard the diversity of participants to avoid delay. These datasets are then used in AI algorithms, which will lead to bias.

Developers may also select from a collect data from a very limited pool of participants due to geographical convenience. In one case, the Toronto-based company Winterlight Labs developed a deep learning algorithm that detects Alzheimer's disease by speech. It was soon reported that the company’s technology only worked for English speakers with a specific Canadian dialect. The developers from the company had only selected training data from speakers native to their area. Many others evaluated by the model resulted in false positive diagnoses for Alzheimer’s, as any grammatical inconsistencies were considered a sign of the disease based on the training data (Gershgorn, n.d.). Artificial intelligence reporter Dave Gershgorn (n.d.) investigated Winterlight Labs’ technology, and pointed out that: “It doesn’t work […] because you’re from Africa, from India, or China, or Michigan. Imagine most of the world is getting healthier because of some new technology, but you’re getting left behind.” As demonstrated by Winterlight Labs, geographical disparity would favor one population over another.

The outliers are additionally often not represented in datasets. Unreliability can be what is called a “grey rhino” in the study of healthcare AI: a problem that is highly impactful and would very likely be a neglected threat. Misdiagnoses in the medical field, such as those in mental illnesses, serve to increase the many disparities that certain groups already face. Since certain aspects of the medical industry rely on preciseness, AI certainly seems to exacerbate problems that need more nuance. According to Walsh et al. (2020): “While AI might be well-suited for the diagnosis and prognosis in complex, nuanced phenotypes like these, we risk producing models that incorporate bias in underlying data (e.g. lack of nonbinary gender categories in EHRs) and algorithmic bias in model specification.” This means that these algorithms may go the opposite way, often over-diagnosing based on phenotypes. This is important because it may mean that specific outliers of patients could be overruled. Algorithms that are extremely tailored to accessible current data may miss important prognostic risk factors. Illnesses that cannot be measured quantitatively, such as mental illnesses, are diagnosed from medical history and a medical provider’s reception to the patient’s thoughts. Current technological methods of identification, such as natural language processing (NLP), is woefully incompetent to handle these types of diagnoses. To those who disproportionately suffer from mental illnesses, the AI is not remotely able to handle feedback, much less diagnoses. For example, women have a higher pervasiveness to anxiety disorders, and an unprepared and unreliable AI algorithm would be more likely to misdiagnose women (Walsh et al., 2020).

The Doctor-Patient Relationship

Even with improvements to medical AI, a machine cannot be the replacement for a doctor. If a doctor is able to recognize discrepancies in an algorithm, then he or she may be able to make the correct diagnosis and eliminate any algorithmic bias. However, this involves a healthy relationship between the physician and the patient. In a strong doctor-patient relationship, the patient will feel more comfortable with the doctor. A healthy bond will in fact, “promote recovery, reduce relapse, and enhance treatment adherence” among the patients (Harbishettar, Krishna, & Srinivasa, 2019). The doctor, in turn, will better understand the patient and be able to provide a more accurate diagnosis. According to Dr. John Kelley, a psychologist at Harvard-affiliated Massachusetts General Hospital, “A good relationship fosters better communication, which improves diagnosis. It also encourages people to tell their doctors about symptoms they might not otherwise disclose” (Kelley qtd. in Harvard Health, 2014). Based on what the patient tells the doctor, the doctor may be able to make a better decision for the patient than an algorithm.

With the influx of technologies into the healthcare system, doctors are becoming more dependent on scans and machinery. This not only dehumanizes the patient, but it also causes physicians to rely more on technologies than their medical training. According to Dr. Abraham Verghese (2011), a professor at the Stanford University Medical School, the "ease of ordering a [CT] scan has caused doctors' most basic skills in examining the body to atrophy." Many physicians enjoy the simplicity of ordering tests and scans, and choose to communicate only to a computer, not to their patients. In one case noted by Dr. Verghese, a patient had developed metastatic breast cancer, and yet was not treated until her cancer became an emergency. The patient had been to several different hospitals, but was not examined thoroughly. The physicians did not develop a strong doctor-patient relationship with her, and chose to rely on a computer instead. The physicians missed "simple things" during the examination; the result was "the consequence of losing both faith and skill" (Verghese, 2011). Thus, physicians cannot be overly reliant on medical AI to provide a diagnosis.

Conclusion

There is no doubt that bias exists in AI. However, the depth of bias is more complicated than originally though, and deeply rooted in systemic dilemmas that often seem cyclical. How do one begin to solve a problem that has been accumulating ever since AI has been used in healthcare? The answer cannot be answered concretely since the problem of AI bias is more pervasive than quantifiable methods of surveying. Simply taking a percentage of those who have been misinformed or have suffered harm from medical AI bias veers from the true intent of fostering better connections between the healthcare system and its consumers into a number game. It is not the question of how many people have been affected by bias in AI, but rather how they have been affected. Focusing on the quality of care provided will lead to a better understanding of where AI has fallen short: human bias, unreliability, data representation, and the declining doctor-patient relationship.

References

Chen, L., & Baker, M. D. (2006). Racial and ethnic differences in the rates of urinary tract infections in febrile infants in the emergency department. *Pediatric emergency care, 22*(7), 485–487. https://doi.org/10.1097/01.pec.0000226872.31501.d0

Dutchen, S. (n.d.). The Importance of Nuance. *Harvard Medical School*. https://hms.harvard.edu/magazine/artificial-intelligence/importance-nuance.

Gershgorn, D. (n.d.). If AI is going to be the world's doctor, it needs better textbooks. *Quartz*. https://qz.com/1367177/if-ai-is-going-to-be-the-worlds-doctor-it-needs-better-textbooks/.

Hall, W. J., Chapman, M. V., Lee, K. M., Merino, Y. M., Thomas, T. W., Payne, B. K., Eng, E., Day, S. H., & Coyne-Beasley, T. (2015). Implicit Racial/Ethnic Bias Among Health Care Professionals and Its Influence on Health Care Outcomes: A Systematic Review. *American journal of public health, 105*(12), e60–e76. https://doi.org/10.2105/AJPH.2015.302903

Harbishettar, V., Krishna, K. R., Srinivasa, P., & Gowda, M. (2019). The enigma of doctor-patient relationship. *Indian journal of psychiatry, 61*(Suppl 4), S776–S781. https://doi.org/10.4103/psychiatry.IndianJPsychiatry_96_19

Harvard Health. (2014, June). Doctor-patient relationship improves your health. *Harvard Health.* https://www.health.harvard.edu/staying-healthy/doctor-patient-relationship-improves-your-health#:~:text=%22A%20good%20relationship%20fosters%20better,Harvard%2Daffiliated%20Massachusetts%20General%20Hospital.

Kaushal, A., Altman, R., & Langlotz, C. (2020, September 22). Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms. *JAMA.* https://jamanetwork.com/journals/jama/fullarticle/2770833?guestAccessKey=ad8f72ad-8b98-42fa-87c3-58c7112d923f&utm_source=For_The_Media&utm_medium=referral&utm_campaign=ftm_links&utm_content=tfl&utm_term=092220.

Kaushal, A., Altman, R., & Langlotz, C. (2020, November 17). Health care AI systems are biased. *Scientific American*. https://www.scientificamerican.com/article/health-care-ai-systems-are-biased/

Markets and Markets. (2020, June). Artificial intelligence in healthcare market. *Markets and Markets.* https://www.marketsandmarkets.com/Market-Reports/artificial-intelligence-healthcare-market-54679303.html.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019, October 25). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. https://science.sciencemag.org/content/366/6464/447

Thakar, S. (2019, March 01). Can ai diagnose skin cancer in all races? *Innovate Healthcare.* https://www.aiin.healthcare/topics/diagnostics/can-ai-diagnose-skin-cancer-all-races

Tirrell, M., & Miller, L. (2020, September 04). Moderna slows Coronavirus vaccine TRIAL enrollment to ensure MINORITY REPRESENTATION, CEO says. *CNBC.* https://www.cnbc.com/2020/09/04/moderna-slows-coronavirus-vaccine-trial-t-to-ensure-minority-representation-ceo-says.html

Verghese, A. (2011, February 26). Treat the Patient, Not the CT Scan. *New York Times.* https://www.nytimes.com/2011/02/27/opinion/27verghese.html

Victorian Government. (n.d.). Data collection challenges and improvements. *Vic.gov*. https://www.vic.gov.au/victorian-family-violence-data-collection-framework/data-collection-challenges-and-improvements.

Walsh, C. G., Chaudhry, B., Dua, P., Goodman, K. W., Kaplan, B., Kavuluru, R., Solomonides, A., Subbian, V. (2020, January 22). Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence. *OUP Academic.* https://academic.oup.com/jamiaopen/article/3/1/9/5714181.

Ward, L. (2019, October 14). The Ethical Dilemmas AI Poses for Health Care. *Wall Street Journal.* https://www.wsj.com/articles/the-ethical-dilemmas-ai-poses-for-health-care-11571018400

Watson, W., & Marsh, C. (n.d.). Artificial intelligence bias in healthcare. *Booz Allen Hamilton.* https://www.boozallen.com/c/insight/blog/ai-bias-in-healthcare.html

Zephyrin, L. (2019, July 9). Pregnancy-related deaths reflect how implicit bias harms women. We need to fix that. *STAT*. https://www.statnews.com/2019/07/10/pregnancy-related-deaths-implicit-bias/.