Modernizing Regulatory Practices for Artificial Intelligence Driven Medical Tools

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

Rishub Handa

Spring 2022

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

Advisor

Kent Wayland, Assistant Professor, Department of Engineering and Society

Introduction

Artificial Intelligence (AI) and Machine Learning (ML) have gained immense traction in the past decade, allowing systems to autonomously make decisions on the behalf of human operators. We have seen success with this approach in fraud detection, autonomous vehicles, and language translation. We are now trusting AI systems with supporting physicians in life-saving medical decisions. Predictive models are able to forecast patient outcomes in kidney failure, heart disease, and many other chronic conditions (Jiang et al., 2017).

However, these models can suffer from algorithmic bias embedded in how it makes decisions. Specifically, this means that the AI software finds hidden trends in the data it learns from that do not accurately reflect reality. When physicians employ the software to create predictions for new data, it can reinforce the hidden biases from its training. For example, a model developed to identify lung disease trained on mostly X-rays from male subjects performed poorly for female patients, since it created an association between the male chest anatomy and the disease (Larrazabal et al., 2020).

The problem with the current understanding of bias in medical AI is that scholars often point to the technical shortcomings of the model when it performs poorly (Obermeyer et al., 2019). These critiques relate to design choices of the model designers or system biases in how the data was generated or collected.

While it is important to consider the technical aspects of mitigating bias in AI, this dominating view neglects the full sociotechnical nature of this problem. It is equally important to create social systems that allow for the safe deployment and utilization of these algorithms, even in the presence of technical flaws. Given the pace of innovation in AI, researchers are unlikely to account for every source of bias and identify issues with newer models and datasets. Cutting-

edge medical technologies often pose risks to patients, and it is the Food and Drug Administration's (FDA) responsibility to mitigate these risks by evaluating the safety and efficacy of new tools and therapies. Though the FDA has a robust procedure for classifying and regulating hardware medical devices, its traditional practices are not suited for AI diagnostic software tools.

If algorithmic bias continues to be improperly addressed, it can have drastic consequences. AI is often used in healthcare for early detection for many types of cancer, genetic screening, etc. (Jiang et al., 2017). Missing patients that actually have the illness can lead to fatal outcomes, and wrongly classifying healthy patients as diseased can create high healthcare costs for patients and hospitals.

To avoid further devastating patients with biased decisions, the FDA must implement more relevant regulations for software-based tools. My research serves to uncover shortcomings in the current FDA practices for evaluating the safety and efficacy of AI algorithms. Specifically, I will use a case study of the Optum Future Cost Algorithm (OFCA) to demonstrate how the current system is failing in practice. An improved understanding of this gap will reveal how we can improve our regulatory systems to ensure safer and equitable treatment for patients.

Background

Current Approaches for Mitigating Bias

Many model developers will assign data as the root of the problem – "garbage in, garbage out". They point to the demographic breakdown of the training data, and they often advocate for better control of race, sex, socioeconomic background, etc. (Centola et al., 2021) to ensure the model is exposed to a diverse set of conditions. Unfortunately, how the data is sourced is

controlled by the healthcare organization contracting these engineers, so the model developers can only suggest this as a technical fix; there is no oversight ensuring that the training data is diverse and representative of all relevant patient demographics (Panch et al., 2019).

Another frequently discussed technical issue is unrepresentative data labeling. The "label" is the output or class the model is trying to predict, given a set of inputs or "features". The label should accurately reflect what the physicians want to predict. Using a proxy to approximate the true outcome can lead the algorithm to identify hidden trends that optimize this metric without achieving the desired results. For example, an algorithm designed with the intent to screen which patients will have a deteriorating chronic illness used the label of "future healthcare costs" to proxy for a worsening condition; this algorithm optimized for more expensive patients, rather than those who truly needed the care (Obermeyer et al., 2019).

In machine learning academia, new models and datasets are subject to interrogation by peer review. Having multiple teams validate the results and calculate their own statistical measures for bias helps to iteratively improve the model/data (Le Sueur et al., 2020). This collaborative culture does not exist in healthcare. Hospitals and insurance companies will often contract external data analytics companies to design their model. These teams often work in isolation to protect their proprietary model, so the extent of bias mitigation is based on the competency and diversity of the engineering team (Shastri, 2020).

While these technical considerations are important for creating more equitable algorithms, professionals in AI and healthcare need to open the discussion to enhancing the social systems patients rely on to provide oversight. AI research is advancing at an unprecedented pace; the number of journal publications increased 34.5% from 2019 to 2020 (Entwood, 2021). The next model is here before current ones can be vetted in deployment. Given

this pace, the public sector, i.e. the FDA, must develop new regulation relevant for companies employing these technologies.

FDA Software as a Medical Device

Under the Federal Food, Drug, and Cosmetics Act, the FDA regulates products used for medical purposes; specifically, this includes technologies that would be used to treat, diagnose, cure, and/or prevent disease. These practices were initially designed for therapeutic drugs and medical devices. After classifying the technology based on its potential risk to the patient (Class I – III), the FDA requires the company to perform multiple stages of clinical trials and document every technical design decision in the development process. If the trials prove the technology is safe for the patient and produces clinically significant results, the FDA approves commercialization, via a Premarket Approval (PMA) document. If the product poses minimal risk for the patient, like a Band-Aid, they can apply for a 510K exemption from the traditional 3 Phase Clinical Trial process, but they still need to provide evidence from IRB approved trials, or equivalent testing, that the product is not harmful. The company must then demonstrate adherence to Current Good Manufacturing Practice Regulations (CGMP); this entails freezing the product design and creating safeguards for manufacturing and quality control. Any updates to the device must prove their safety and efficacy once again (FDA, 2020).

This procedure is intuitive for a surgical tool or diagnostic test, but fundamentally misaligned with how software tools are developed and deployed. To adapt these practices, the FDA released a draft for new guidelines relating to Software as a Medical Device (SaMD) in November 2021; this was the first update to the Premarket Approval/Exemption process since 2005 and has yet to be finalized (FDA, 2021). This draft outlines what the FDA would classify as SaMD and the documentation the FDA requires from the design process to ensure patient

safety before 510K exemption or PMA. It is still modeled after the traditional approval process, where software is categorized based on its potential harm to the patient and must conduct studies to prove safety and clinical significance. Once approved, the company must outline how they will ensure quality in deployment and must request further approvals before any updates. The shortcomings of SaMD regulations in practice are discussed below.

Methods

Literature Review

The existing literature on best practices to mitigate bias in AI for healthcare applications tend to focus on technical improvements to the model design and data sourcing/cleaning. A few case studies on large scale deployments of biased AI models thoroughly discussed these issues. For example, researchers at UC Berkeley found how using future health costs as an improper approximation of patient's disease severity can lead to a racially bias algorithm when choosing to admit patients for an advanced care program (Obermeyer et al., 2019). Another study showed how differences in patient demographics across hospitals create unbalanced datasets; this led to a pneumothorax detection algorithm performing poorly for Black patients while accurately detecting the condition in chest X-rays of White patients (Wu et al., 2021).

Another set of research attempts to codify best practices in ML research and healthcare applications. These researchers attempt to set a standard for countering algorithmic bias in healthcare settings. They concur that bias is typically revealed in public settings once the model is deployed, and that we need systems to address systemic issues exposed by these models. These systemic issues often embed themselves in how the data is collected, so the bias can go unnoticed (Jiang et al., 2017; Panch et al., 2019). These principles typically apply to how

researchers and healthcare organizations can improve their practices, without considering the ecosystem and stakeholders involved after the model is employed in practice.

Norori et al. take an interesting stance on robust model development by advocating for "open science," a framework similar to how computer scientists typically vet models through open sourced code. This system allows model developers to continuously iterate their algorithm while it is being deployed in the field, which also promotes open sourced data collection. This collective scrutiny tends to allow for quick identification for bias and fast methods to address the issue. For example, bias can arise from unexpected correlations between input variables (often demographic attributes such as race, sex, etc.) and the label(s); a peer not directly involved in the research question can bring quantitative measure to identify these confounding relationships (Norori et al., 2021). However, this ideology is incompatible with how healthcare AI companies currently protect their proprietary models as trade secrets.

Approach

My research serves to provide a sociotechnical view of mitigating bias in AI driven healthcare applications. Specifically, I demonstrate why technical improvements to AI alone will not fully mitigate bias and why improved FDA oversight is required to ensure patient safety when employing these algorithms. In the discussion below, I perform a case study analysis of the Optum Future Cost Algorithm to illustrate how racial bias can lead to drastic differences in health outcomes for patients. I then look to the FDA's official documentation to identify shortcomings in the existing regulatory framework.

To gain a thorough understanding of the existing research landscape, I investigated other case studies, regarding diagnosis of kidney disease, lung disease, and substance abuse, where bias in AI software tools or poor regulatory practices led to poor health outcomes. I also

performed a literature search of popular review articles relating to AI bias in healthcare. When analyzing the Optum Future Cost Algorithm, I researched the work of the original author who conducted the investigation into Optum and assisted them in reforming their algorithm. Finally, I found gaps that enabled the flaws in Optum by directly referencing the FDA's statements on regulating SaMD, 510K exemption, and AI/ML, as well as heavily cited reviews.

Discussion

The Optum Future Cost Algorithm

Hospital systems, insurance companies, and pharmacy benefit managers frequently utilize prediction algorithms and screening tools to identify patients with complex and deteriorating health conditions. By identifying patients that will likely be costly to the healthcare system in the future, they can direct physicians to take additional preventative measures that improve patients' health and avoid future costs. Such forecast models often predict future health costs as a proxy for disease burden to enroll patients in more intensive care programs (Bates et al., 2014). Whether this label is chosen out of the corporate incentive to minimize costs or out of convenience in gathering data, neglecting to directly classify patients on their risk for serious illness can prevent sick patients from accessing the care they need.

Optum, a large pharmacy benefit manager company, employed this approach of predicting future healthcare costs for an estimated 200 million patients each year; the patients above the 97th percentile were enrolled in an advanced care program. The OFCA successfully achieved its task of predicting a patient's future healthcare resource utilization. However, it inadvertently introduced a strong racial bias to enroll White patients, preventing many Black patients from receiving critical healthcare resources. I refer to Black patients as those with

African ancestry receiving healthcare in the US without necessarily a US citizenship. More specifically, a Black patient assigned the same cost label as a White patient tended to have 26.3% more severe chronic illness than their racial counterpart, when controlled for other demographic attributes. Accounting for this bias would increase the representation of Black patients in the program from 17.7% to 46.5% (Obermeyer et al., 2019).

Upon analysis, researchers investigating bias in AI healthcare systems claimed that this bias can be attributed to the unrepresentative label choice of future healthcare cost instead of disease burden. The training data showed that White patients costed more than Black patients, which controlled for disease burden and demographic factors. The model learned a hidden association between race and cost, leading to limiting access to critical care for many Black patients. This disparity has an underlying social cause. Surveys show that Black patients tend distrust physicians and healthcare institutions more than White patients, so they are less likely to utilize healthcare resources (Armstrong et al., 2007). After identifying this poor design choice, Obermeyer et al. experimented with other label choices that represented disease progression, de-correlated from race, to create a more equitable algorithm. His team was then able to work with the original model designers to remedy this technical flaw.

This collaboration is a rare occurrence. While the contracted designers of the OFCA were willing to collaborate with the researchers, most proprietary algorithm developers tightly safeguard their codebase (Norori et al., 2021). Obermeyer's team also faced many difficulties trying to approximate the model's behavior, since his team did not have access to the full model while analyzing its results. In the private sector, where trade secrets are the intellectual assets of software companies, it is the FDA's role to provide the oversight to ensure these algorithms do not lead to harmful outcomes for patients

Regulatory Shortcomings

As aforementioned, the FDA has only recently formalized the requirements to commercialize medical software used for diagnosing disease, predicting patient health, and recommending treatment. However, these SaMD guidelines merely adapted the existing 510K fast-track to software. This system was initially designed for medical tools that do not directly pose a risk to the patient, such as surgical gloves, but still need to be validated for safety (FDA, 2021). It is now evident that AI tools have a significant influence on patients' health outcomes, so there are several shortcomings with the 510K exemption process that the FDA must address.

The 21st Century Cures Act of 2016 outlines categories of software that can qualify for exemption from any regulation at all; the definition of this software is very vague and often exploited. Clinical Decision Support tools (CDS) include software that inform healthcare providers without driving decisions related to patient care (PEW, 2021). This ambiguous distinction allows many AI tools to be deployed in practice without ever consulting the FDA, including the OFCA. The results of this algorithm had tangible impacts on the health of patients, but it was framed as a "support" tool, even though it was screening for qualifying patients. In practice, however, it is very difficult for the FDA to evaluate whether doctors, nurses, and administrators use the tool with oversight. In the case of OFCA, doctors could not feasibly screen all the patients themselves and relied on the tool to narrow candidates for the advanced care program, which led to excluding many Black patients from appropriate care. Prescription Drug Monitoring Programs (PDMP), which track how patients consume controlled substances for medical use to predict abuse, also exploited this legal ambiguity to avoid regulation (Oliva, 2021). Given the widespread misuse of critical, unregulated CDS/AI tools like OFCA and PDMP, the FDA must expand the scope of what types of software must be regulated.

For the software tools that cannot avoid regulation, the 510K pathway is not fit to adequately evaluate the safety and efficacy of AI applications. The 510K exemption allows companies to avoid length clinical trials; however, the developers simply need to provide documentation or minimal evidence that the data will be handled securely to prove it does not cause harm to the patient, since the AI itself cannot directly cause harm (excluding surgical robotics and similar software) (FDA, 2019). This system is ineffective because AI developers often cannot gauge the accuracy of the model beforehand, since it learns as it is exposed to more data. Thus, it is incredibly difficult to uncover biases in the data and its indirect impact on patient outcomes before the tool is operating in production. As the software learns and improves iteratively, it requires regular reevaluations to identify downstream effects. The OFCA was forming cost predictions for 200 million patients per year for several years before its flaws were uncovered (Obermeyer et al., 2019). Furthermore, while the FDA has clear guidelines on regulating "locked" algorithms, which predict the same output for a given input, there are no evaluation frameworks for "adaptive" algorithms which learn over time, as is the case with AI/ML (Benjamens et al., 2020). This lack of ongoing evaluation is especially problematic for adaptive algorithms because if systemic biases are reflected in the data, the algorithm will reinforce those biases as it is exposed to more training examples. In the case of the OFCA, this might have led to a feedback loop. Black patients underutilize healthcare resources, so the algorithm's predictions cause them to receive less care; this means they continue to underutilize resources, so they continue to receive less care.

Conclusion

The complexity of bias in AI driven healthcare applications requires a sociotechnical approach to mitigating the bias and ensuring patient safety. In addition to technical improvements, modernized federal regulatory practices enforced by the FDA would provide the necessary oversight in this rapidly changing field. The current implementation of SaMD merely adapts the procedures for traditional hardware medical devices to software. Given the strict documentation and testing required for these regulatory pathways, Congress created exemptions for tools perceived to have low risk to patients; this enables companies like Optum to exploit the system and avoid regulation altogether. Further, for the companies that cannot take advantage of the legal ambiguity, the lack of ongoing assessment is ineffective for evaluating the safety of adaptive algorithms that learn and change over time.

To address these issues, the FDA must broaden the scope of software subject to regulation, while allowing for a faster path to initial deployment. However, it is essential to regularly analyze the effects of these algorithms on patients while operating commercially to ensure they are not reinforcing existing systemic biases. These improvements could have provided the OFCA with sufficient oversight to prevent its racially biased predictions from scaling to hundreds of millions of patients. Implementing these policies would require coordination from scientists, engineers, doctors, and lawmakers. Though completely eradicating bias is unlikely, taking these steps to quickly identify and mitigate it would decrease the risk of harm to the patient while encouraging the pace of innovation.

References

Armstrong, K., Ravenell, K. L., McMurphy, S., & Putt, M. (2007). Racial/ethnic differences in physician distrust in the United States. *American Journal of Public Health*, *97*(7), 1283–1289. https://doi.org/10.2105/AJPH.2005.080762

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs (Project Hope)*, *33*(7), 1123–1131. https://doi.org/10.1377/hlthaff.2014.0041

Benjamens, S., Pranavasingh, D., & Bertalan, M. (2020, September 11). *The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database | npj Digital Medicine*. https://www.nature.com/articles/s41746-020-00324-0

Centola, D., Guilbeault, D., Sarkar, U., Khoong, E., & Zhang, J. (2021). The reduction of race and gender bias in clinical treatment recommendations using clinician peer networks in an experimental setting. *Nature Communications*, *12*(1), 6585. https://doi.org/10.1038/s41467-021-26905-5

Entwood, J. (2021). *2021 Stanford AI Index Report*. https://aiindex.stanford.edu/wp-content/uploads/2021/03/2021-AI-Index-Report-_Chapter-1.pdf

FDA. (2019, March 28). *The 510(k) Program: Evaluating Substantial Equivalence in Premarket Notifications [510(k)]*. U.S. Food and Drug Administration; FDA. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/510k-program-evaluating-substantial-equivalence-premarket-notifications-510k

FDA. (2020, September 4). *Overview of Device Regulation*. FDA; FDA. https://www.fda.gov/medical-devices/device-advice-comprehensive-regulatory-assistance/overview-device-regulation

FDA. (2021). *510(k) Clearances*. FDA; FDA. https://www.fda.gov/medical-devices/device-approvals-denials-and-clearances/510k-clearances

FDA. (2021, November 3). *FDA In Brief: FDA Provides New Draft Guidance on Premarket Submissions for Device Software Functions*. FDA; FDA. https://www.fda.gov/news-events/press-announcements/fda-brief-fda-provides-new-draft-guidance-premarket-submissions-device-software-functions

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 and Vascular Neurology*, *2*(4), 230–243. https://doi.org/10.1136/svn-2017-000101

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., & Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, *117*(23), 12592–12594. https://doi.org/10.1073/pnas.1919012117

Le Sueur, H., Dagliati, A., Buchan, I., Whetton, A. D., Martin, G. P., Dornan, T., & Geifman, N. (2020). Pride and prejudice – What can we learn from peer review? *Medical Teacher*, *42*(9), 1012–1018. https://doi.org/10.1080/0142159X.2020.1774527

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns*, *2*(10), 100347. https://doi.org/10.1016/j.patter.2021.100347

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

Oliva, J. D. (2021). *Dosing Discrimination: Regulating PDMP Risk Scores* (SSRN Scholarly Paper No. 3768774). Social Science Research Network. https://doi.org/10.2139/ssrn.3768774

Panch, T., Mattie, H., & Atun, R. (2019). Artificial intelligence and algorithmic bias: Implications for health systems. *Journal of Global Health*, *9*(2), 020318. https://doi.org/10.7189/jogh.09.020318

PEW. (2021). *How FDA Regulates Artificial Intelligence in Medical Products*. https://pew.org/3yglbCS

Shastri, A. (2020). *Diverse Teams Build Better AI. Here's Why*. Forbes. https://www.forbes.com/sites/arunshastri/2020/07/01/diverse-teams-build-better-ai-heres-why/

Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, D. E., & Zou, J. (2021). How medical AI devices are evaluated: Limitations and recommendations from an analysis of FDA approvals. *Nature Medicine*, *27*(4), 582–584. https://doi.org/10.1038/s41591-021-01312-x