Andrew DiBiasio

Amanda Girard

Computing, Society, and Professionalism

27 November 2023

Racial Bias in Healthcare Algorithms

1. **Introduction: Defining Racial Bias and its Significance.**

Bias emerges when prejudices in data influence the decision-making process of an algorithm [6]. Although biases come in various forms, a machine learning model exhibiting racial bias will render disparate decisions and predictions based on the input's racial characteristics. Given that machine learning algorithms rely on data, algorithmic bias frequently originates from prejudices embedded in training datasets. For instance, if a training dataset lacks representativeness of the sampled population or mirrors historical inequalities, the resulting machine learning model will capture these patterns and extrapolate them to the entire population [6]. Consequently, this significantly impacts the fairness of its decision-making process. While these instances highlight certain origins of bias in healthcare algorithms, numerous others exist.

Now that it is clear what bias is, it is natural to question its impact. The primary concern arises from the potential harm that biased algorithms can inflict on the people against which they discriminate. For example, suppose a health insurance company opts to utilize cutting-edge machine-learning technology to compute their client’s monthly deductibles. In the presence of racial biases, this might unjustly imply that a specific demographic faces a higher risk of injury [3]. Consequently, this could result in elevated monthly payments for that entire demographic, leading to an inequitable surge in health insurance expenses. The integration of biased algorithms like these into our societal frameworks essentially reinforces that bias whenever it is used.

In our contemporary world, artificial intelligence and machine learning are pivotal components. Their pervasive presence in data-driven scenarios necessitates a comprehensive understanding of mitigating biases. This article aims to assess our ability to mitigate racial bias within healthcare algorithms by examining case studies, exploring strategies for bias mitigation, and hindrances to widespread fair model use. As such, the essential question posed is whether machine learning algorithms can make fair decisions or if the nature of interpolation is inherently linked to biases present in training data.

1. **Case Studies.**
   1. Case Study 1: Optum Patient Prioritization

In 2019, Optum, a healthcare company entrusted with critical health system decisions, such as patient prioritization, undertook measures to mitigate racial bias in their algorithms by eliminating variables related to race from their data [8]. Despite these efforts, unexpected racial biases surfaced within ostensibly race-blind variables. An illustrative example is the prioritization variable based on the hospital charges incurred by patients, where higher bills indicated higher priority [8]. However, it was revealed that white patients, on average, paid an additional $1,800 compared to their African American counterparts [8]. This phenomenon, known as a proxy variable in statistics, occurs when associated attributes substitute hidden explanatory variables, like race [10]. Regardless of Optum's attempts to mitigate racial bias, their algorithm disproportionately deprioritized black patients in need of complex medical attention. Consequently, many white patients with less severe medical issues gained precedence in healthcare access compared to their African American counterparts who were more heavily injured. Optum's experience is not isolated, as evidenced by similar issues faced by UnitedHealth, an American healthcare insurance company, as presented in the second case study. Given the pervasive nature of racial bias in healthcare algorithms, it prompts the question of why comprehensive actions have not been taken to address this issue.

* 1. Case Study 2: United Health Patient Cost Evaluation

UnitedHealth, a significant player in the American healthcare insurance sector, also faces challenges related to racial bias in patient cost evaluation, mirroring concerns observed in its subsidiary, Optum. Despite earnest attempts to eradicate bias, UnitedHealth's algorithms inadvertently expose biases within seemingly race-neutral variables. For example, a metric influencing patient priority hinges on the duration of hospital stays, with longer stays indicating higher priority [9]. However, an unexpected revelation surfaces — black patients, on average, experience shorter hospital stays compared to their white counterparts, introducing further complexities associated with proxy variables [9]. One possible explanation could be that since the average income for black patients is lower than that for white patients, they may tend to stay in the hospital for less time, making the length of the hospital stay a proxy variable. These variables, intricately linked to race, substitute hidden explanatory factors and persist despite efforts to rectify bias. UnitedHealth's algorithms, akin to Optum's, manifest disparities that may lead to the inadvertent under prioritization of black patients requiring extensive medical attention. This perpetuates a cycle where white patients with less severe medical issues receive priority over black patients with more critical injuries, underscoring the systemic nature of racial bias in healthcare algorithms within the broader UnitedHealth framework. These challenges prompt a critical examination of the healthcare industry's approach to addressing and rectifying racial bias issues.

1. **Mitigating Bias in Machine Learning Algorithms.**
   1. Preface

Before reading this section, it is important to highlight that while the ensuing methods are explored in the specific context of racial bias within healthcare algorithms, their application can be extended to tackle different forms of bias across various algorithmic domains.

* 1. Common Mitigation Methods

Mitigating bias in machine learning algorithms involves three primary methods: preprocessing, in-processing, and postprocessing [1]. Preprocessing mitigation works to alleviate present biases in the dataset before model training, effectively handling the disproportionate data representing the various classes of data. Some examples of common techniques are the resampling of the dataset, adding newly sampled data to the pre-existing data, and relabeling the data [1]. In-processing mitigation strategies focus on mitigating biases by incorporating metrics to monitor biases and optimizing them during training. Methods may involve data regulation, imposing fairness constraints, and adversarial debiasing [1]. Adversarial debiasing employs an adversary model alongside the predictor model, predicting attributes not to be used in predicting the target variable [2]. Postprocessing mitigation methods mitigate biases by altering how model outputs are used and interpreted, such as modifying decision thresholds and adjusting model outputs [1].

These are just a few examples of the different types of methods that can be used for mitigating racial biases in healthcare algorithms. Naturally, these techniques can be used regardless of where the data was collected from, whether it be from healthcare systems or a financial firm. But nevertheless, they prove to be effective in mitigating the various sources of bias mentioned earlier in this article.

* 1. Benefits of Mitigation

Bias mitigation in machine learning brings about several benefits that contribute to the development of fair and equitable models. Firstly, it promotes inclusivity by addressing and rectifying historical biases present in training data [7]. As models are trained on more diverse and representative datasets, they become better equipped to make informed predictions across different demographic groups, reducing the risk of perpetuating existing inequalities [7].

Moreover, bias mitigation enhances the accuracy and reliability of machine learning models. By identifying and correcting biases during the training process, models are less likely to make unfair predictions based on irrelevant or discriminatory factors. This results in more precise and trustworthy outcomes, fostering user confidence and trust in the technology. Additionally, mitigating bias helps to comply with ethical standards and legal regulations. Organizations and developers can demonstrate a commitment to fairness and accountability, mitigating the potential negative impacts associated with biased decision-making in sensitive areas such as finance, healthcare, and criminal justice. Overall, bias mitigation not only aligns machine learning models with ethical principles but also ensures that the benefits of AI technologies are distributed more equitably across diverse populations.

* 1. Downsides of Mitigation

While bias mitigation in machine learning is crucial for promoting fairness, it introduces a nuanced challenge known as the bias-variance tradeoff. This tradeoff involves finding a delicate balance between mitigating bias and maintaining the model's ability to generalize well to diverse data [4]. Intensive bias mitigation measures, such as aggressively removing certain features or oversampling underrepresented groups, may inadvertently lead to increased model variance. In other words, the model becomes overly sensitive to the peculiarities of the training data, making it less adaptable to unseen instances. This heightened variance can result in overfitting, where the model performs exceptionally well on the training data but struggles to generalize effectively to new, diverse inputs.

Furthermore, bias mitigation techniques may inadvertently introduce new sources of bias or create unintended consequences. For instance, in an effort to address biases related to one demographic group, the model might inadvertently amplify biases against another [4]. This phenomenon can occur when the mitigation strategies disproportionately affect certain groups or when the data used for mitigation is itself biased. Striking the right balance between mitigating existing biases and avoiding the introduction of new ones is a complex task that requires careful consideration and continual evaluation throughout the model development process. The challenge lies in navigating the intricate tradeoff between reducing bias and maintaining the model's ability to generalize accurately across diverse and representative datasets.

* 1. Does this Really Make the Model Fair?

Incorporating methods to alleviate bias is a common practice in machine learning algorithms, but the effectiveness of such approaches prompts scrutiny. Can the integration of pre/in/postprocessing procedures genuinely transform a machine learning algorithm into a fair entity? Unfortunately, the answer is not straightforward. While these approaches may mitigate racial biases to some extent, this doesn’t equate to a completely fair algorithm.

The underlying challenge stems from the interdependence of attributes like race, gender, and ethnicity. Striving to eliminate racial bias in a model may inadvertently compromise the fairness of other factors, such as gender or income status [5]. Consequently, even if the model ceases to discriminate based on race, it may instead start to exhibit bias against individuals with lower incomes, for instance.

These complexities draw attention to a broader societal barrier that impedes the creation of fair models: the lack of a clear definition of fairness. The concept of fairness has sparked considerable debate among researchers, with ongoing discussions persisting to this day. The fundamental question emerges — if we struggle to articulate what fairness is, how can we anticipate accurately simulating it through our algorithms?

Machine learning algorithms inherently make decisions based off of data they are provided, so any racial biases in their decisions are a direct result of biases in their training data. However, if data sets are representative, then the biases present in the data are simply the observed biases present in the world. So, machine learning can be used to model situations based on how the world currently is, but not on how it should be. They don't take into account that there shouldn't be certain prejudices in the world, it simply repeats what it sees. As such, even the model that has been trained on the most carefully processed training dataset isn't truly considered "fair," in the sense that people shouldn't be defined or judged based solely on certain traits. However, this is exactly what machine learning does and since it doesn't account well for special cases that go against the majority of the derived population, it doesn’t make sense to call it fair in these cases.

1. **Conclusion.**

Overall, the main takeaway from this article is that while mitigated machine learning models can serve as useful tools, they aren’t always fair, and so they shouldn’t be used as lone substitutes for human decisions that aren’t subjected to human review.

Racial biases exhibit substantial real-world consequences for targeted groups, as illustrated in case studies involving Optum and UnitedHealth. The exploration of three primary bias mitigation methods—preprocessing, in-preprocessing, and postprocessing—not only shed light on the mechanics of mitigation but also provided insights into enhancing the equity of models. These approaches are designed to alleviate or eradicate bias sources at various stages of the process, contributing to a more equitable outcome. Nevertheless, despite the implementation of these mitigation techniques, inherent complexities persist, hindering the attainment of genuine fairness in models.

While artificial intelligence and machine learning algorithms are still in the process of evolving to effectively eradicate bias, their significance lies in their pivotal role in expanding access to high-quality healthcare. This impact is particularly noteworthy for populations lacking robust social infrastructures, as these technologies can serve as crucial facilitators in bridging gaps and providing access to appropriate healthcare systems.

1. **Afterword: Why Don’t All Companies Mitigate Bias?**

Putting the matter of whether machine learning models can be made truly fair aside, the main concern is that more equitable care is provided to the population. Although bias mitigation isn’t perfect, it does help advance this cause. Thus, this raises the question: why don’t all companies mitigate bias? The response lies in the inherent difficulty of identifying bias. Examining the case of Optum reveals that despite developers' efforts to eliminate bias in their healthcare algorithms, they overlooked underlying biases concealed in proxy variables. Many companies encounter similar challenges; despite implementing precautions, they often fail to recognize biases embedded in proxy variables.

During a discussion of the fairness of AI, machine learning expert Sharad Goel summarizes this issue best, “[F]undamentally, these are hard problems. It’s not particularly surprising that we don’t have an algorithm to help us make all of these algorithms fair. … What is most important is that we really interrogate the data” [11].

Addressing these issues requires significant financial and temporal investments—resources that numerous companies may be unwilling to expend unless mandated. This economic barrier serves as a deterrent, hindering companies from adopting bias mitigation techniques. Whether due to inadequate resources or an unwillingness to allocate funds and efforts, companies face challenges in implementing comprehensive bias mitigation measures. To overcome this, advocating for awareness of racial and other biases in algorithms can exert pressure on companies, compelling them to take proactive steps.

Works Referenced

[1] J. Huang, G. Galal, M. Etemadi, and M. Vaidyanathan, “Evaluation and mitigation of racial bias in clinical machine learning models: Scoping review,” *JMIR Medical Informatics*, vol. 10, no. 5, 2022. doi:10.2196/36388

[2] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara, “Addressing bias in big data and AI for health care: A call for open science,” *Patterns*, vol. 2, no. 10, p. 100347, 2021. doi:10.1016/j.patter.2021.100347

[3] T. Panch, H. Mattie, and R. Atun, “Artificial Intelligence and algorithmic bias: Implications for health systems,” *Journal of Global Health*, vol. 9, no. 2, 2019. doi:10.7189/jogh.09.020318

[4] E. Briscoe and J. Feldman, “Conceptual Complexity and the Bias/Variance Tradeoff,” *Cognition*, vol. 118, no. 1, pp. 2–16, 2011. doi:10.1016/j.cognition.2010.10.004

[5] D. T. Newman, N. J. Fast, and D. J. Harmon, “When Eliminating Bias isn’t Fair: Algorithmic Reductionism and Procedural Justice in Human Resource Decisions,” *Organizational Behavior and Human Decision Processes*, vol. 160, pp. 149–167, 2020. doi:10.1016/j.obhdp.2020.03.008

[6] N. Turner Lee, “Detecting Racial Bias in Algorithms and Machine Learning,” *Journal of Information, Communication and Ethics in Society*, vol. 16, no. 3, pp. 252–260, 2018. doi:10.1108/jices-06-2018-0056

[7] A. Jain *et al.*, “Awareness of Racial and Ethnic Bias and Potential Solutions to Address Bias with Use of Healthcare Algorithms,” *JAMA Health Forum*, vol. 4, no. 6, 2023. doi:10.1001/jamahealthforum.2023.1197

[8] C. Y. Johnson, “Racial Bias in a Medical Algorithm Favors White Patients Over Sicker Black Patients,” The Washington Post, https://www.washingtonpost.com/health/2019/10/24/racial-bias-medical-algorithm-favors-white-patients-over-sicker-black-patients/ (accessed Nov. 7, 2023).

[9] A. Akhtar, “New York is Investigating UnitedHealth’s use of a Medical Algorithm that Steered Black Patients away from Getting Higher-Quality Care,” Business Insider, https://www.businessinsider.com/an-algorithm-treatment-to-white-patients-over-sicker-black-ones-2019-10 (accessed Nov. 7, 2023).

[10] I. Grabski, “Fairness in Machine Learning,” Science in the News, https://sitn.hms.harvard.edu/uncategorized/2020/fairness-machine-learning/ (accessed Nov. 7, 2023).

[11] B. Lorica, “Why it’s Hard to Design Fair Machine Learning Models,” O’Reilly Media, https://www.oreilly.com/radar/podcast/why-its-hard-to-design-fair-machine-learning-models/ (accessed Nov. 27, 2023).