Ethics of Algorithmic Decision Making

Algorithmic Bias & Racial Disparities in Medicine

Kevin Hoang & Peter Gao (Mentor)

As part of the UW Statistics and Probability Association's Directed Reading Program, I explored how algorithmic bias led to disparate impacts in medicine with my mentor Peter Gao. We were motivated by the COVID-19 pandemic which exposed serious racial disparities in health outcomes. We wanted to see if algorithms could make the healthcare system more equitable. We analyzed many news articles and scientific papers over the ten weeks of the program. Not all of our materials were health related. Many types of biases that plague medical algorithms are also present in other fields such as policing or banking. Some of the most informative papers about algorithmic racial bias were about the COMPAS recidivism risk-scoring algorithm and commercial AI facial-recognition software. While not included in our presentation, we also used R to replicate the statistical analyses done in some of the papers we examined.

The United States has a long and dark history of medical mistreatment of minority groups which has led to their mistrust of our current medical institutions. While such mistreatments are unlikely to occur today, serious medical biases still exist. Studies have shown African-Americans are routinely undertreated for pain while others have shown medical students believe misconceptions such as Black patients are less sensitive to pain. Disadvantaged groups are less likely to be included in vaccine or drug trials which can result in less effective drugs for these groups. Medical devices such as pulse oximeters have been shown to be less effective on dark skin tones. These implicit biases alongside with disadvantaged groups' historical mistrust of medical institutions and socio-economic factors lead to disparate health outcomes.

Medical algorithms are often seen as a silver bullet to these problems by removing human bias from the decision-making process. This is a flawed line of thinking. Algorithms can range from formulas made by humans to AI/ML models. With human formulas, there can be arbitrary unexplainable weights or input variables that can result in different outcomes for different racial groups. This is a common criticism for medical formulas that determine procedural risks for patients. On the other hand, AI/ML models need to be trained from sample datasets. If these datasets aren't fairly representative of the population, it can result in worse performance for the underrepresented groups. In addition, many of these AI/ML models' decision-making processes are often unknown or unintelligible. As a result, physicians cannot explain why a certain decision was made to their patient and it may be difficult to determine if the algorithm is racially-biased as a result.

We focused on two case studies of medical algorithms. The first was the performance of the APACHE and OASIS mortality risk-scoring systems (Sarkar *et al*., 2021). Both systems frequently overpredicted mortality for Hispanic and Black people compared to White and Asian people. While the cause for this scoring disparity is unclear, it is concerning because these risk-scoring systems could be used to determine the allocation of limited medical resources based on

the likelihood of a patient's survival. The second case involved an algorithm which used "sickness" to determine patient eligibility for "high-risk care management" programs (Obermeyer *et al*., 2019). The algorithm resulted in Black patients receiving the same score as White patients who had less chronic illnesses. This disparity occurred because medical expenditure history was used as a measure of "sickness". Black patients typically spend less on their healthcare compared to whites due to socioeconomic inequities. This will result in an inefficient "high-risk care management" program where Black patients who have many chronic conditions may be deemed healthy due to their low medical expenditures.

How can we determine if these algorithms are fair? One of the methods proposed by many of the papers we've read is by using mathematical definitions. It is impossible for an algorithm to satisfy all of these definitions so some metrics must be prioritized over others. An algorithm that determines who gets emergency treatment in an ICU could be judged on equal patient outcomes such mortality rates. This algorithm could also be judged on equal allocation of resources. With COVID-19, this would mean an equal allocation of ventilators between races. The last mathematical metric is equal performance which is more relevant for AI/ML models that predict, diagnose, or prognose diseases. These models can be judge by equal accuracy, false positive, or false negative rates between races.

The other method is by providing enough resources for others to detect discrimination in algorithms themselves. This is either by making the algorithm transparent or interpretable. An algorithm can be transparent if the training dataset and code is provided. This allows plaintiffs or experts who believe algorithmic discrimination exists to replicate the model and prove it by manipulating certain input variables themselves. This allows for accountability because these algorithms can be audited. On the other hand, an interpretable algorithm's decision-making process can be easily examined and its merits can be debated by those who have qualms with the algorithm. However, by using only interpretable algorithms, there may be performance tradeoffs.

I'm very pleased with what I learned over this quarter through my research with my mentor and the skills I have picked up. I explored many different medical algorithms currently in use and potential applications of future ones. I learned about the flaws of algorithms and their potential to widen racial disparities in healthcare. I also learned the different metrics to judge algorithmic fairness which could be used to catch these flaws. Lastly, I got hands-on experience with using R for statistical analyses.

References

Sarkar, R., Martin, C., Mattie, H., Gichoya, J. W., Stone, D. J., & Celi, L. A. (2021). Performance of intensive care unit severity scoring systems across different ethnicities in the USA: a retrospective observational study. The Lancet Digital Health, 3(4), e241-e249.

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.