Proposed Policy for Mitigating Bias in AI Healthcare applications (hybrid class sections)

- One focus point that should be taken account of is to not care for the patient cost when looking for these treatments. That is something that should be decided in a direct conversation with the patient rather than excluding them based on a premise that they may not be able to afford the care or treatment.
- The public having access to information regarding the data being used to train these algorithms could have similarly played an important part in catching situations like this before it got as bad as it did. In general, more transparency with the public from tech companies around how (practices + data) they train their algorithms would be good, but companies that use this data in health care settings should be held even more so accountable.
    - If this falls under a domain other than the FDA (the FDA claimed it does), create a body to enforce this.
    - Using the above governing body to make sure that the data being used to train these models accurately represents the population instead of it being skewed heavily against patients of color. An AI based on training data can't accurately detect medical problems if it has not seen enough data about other medical diagnoses from patients of the same demographic.
    - Ensure that while data being used to train algorithms is public, that data does not compromise the right to privacy of the patients.
    - Implement ways that an algorithm can show how it uses the training data to make decisions along with releasing the training data. This can be used to ensure that even if the training sets are made to be more diverse, it will also have the benefit of positively affecting the algorithm to care more for a more diverse population.
- These machines need to be trained on a diverse population. Almost all the data that is gathered should be close to the same number of individuals in each race when training the AI. This should be in a database that is accessible to the public and monitored by more than one third party association, to avoid corruption.
- Any data collected needs to be in compliance with HIPAA, and that data should be regulated to ensure that no personal/medical information is shared over insecure networks that are susceptible to security breaches (which in today's world are all-too-common). That data should also be strictly used to improve the AI's accuracy in diagnosing medical conditions, organizing medical data, or improving patient quality of life.
    - The data should be heavily encrypted so as patient records can be kept confidential. There should be no risk of a data leak. Additionally, fail-safes should be set up so that any leaks could be quickly fixed.
- The algorithms used should not be purchased from a third-party commercial entity exclusively focused on profits and revenue, i.e. Optum or other insurance companies. As it is against their best corporate interest to create a genuinely neutral and unbiased algorithm that only highlights medical risk and does not boost their bottom line. Thus, using a third-party supplied, commercial algorithm will only further demonstrate the inherent flaws and imbalances in our society regarding race and poverty. Instead,

looking to non-profit organizations or requiring additional scrutiny and approval regarding these algorithms would be a better option. A series of standards should be set out to ensure algorithms maintain a health first focus.

    - Moreover, the algorithms should also be designed in such a way that when biases are detected the algorithm can trigger its retraining process again and adjust its parameters.

- In addition to being open and transparent about the use of AI in the healthcare space, it is also crucial that the stakeholders regarding the situation are aware (at least at a fundamental level) of how these AI systems work. It is not enough to simply relinquish data, the people involved have to have an understanding of how that data is processed.
- How does the algorithm tell if a patient is in a minority group? The three articles do not say anything about facial recognition, so it is implied that a person inputs their ethnicity on a form. What happens if we eliminate ethnicity as a factor from the algorithm? It would be a quick way to reduce racial bias by ignoring race as a factor.
    - Additionally, data such as income and ability to pay for treatment should not be collected or used for training the algorithm. An algorithm trained on such data such as the ones in the articles prioritize profit over helping people.
- There should be a way to normalize the data. The program should try its best to find a commonality with all the data sets. If this is not possible then, it should use a different metric to give help to people. It should not be based on race and how much money they have. It could be based on something that has nothing to do with race, like how many times this patient has paid their bills, or how frequently they get medication. Lastly, this AI should only be a tool that is used alongside with the human's decision process.
- There should be a committee formed prior to training the algorithm which determines factors which should NOT have weight on a patient's care (e.g. race, income, etc.). Correlation analysis should then be performed on the features fed to the algorithm to determine whether those forbidden factors can be inferred. If so, transform the data so that those correlations are addressed.
- The algorithms used should be interpretable. In other words, we should be able to understand how an algorithm makes the predictions that it does.
- Utilizing deep learning models, which are inherently non-interpretable, makes it challenging to identify errors and biases. Consequently, such models are not advisable for high-risk applications. A shift towards more interpretable machine learning methodologies is recommended. If pursued, the following criteria should to be met:
    - Implementation of data protection laws is essential.
    - Patients should not be evaluated based on their behavior, socio-economic status, or personal traits, as this could serve as a profiling method, potentially leading to increased discrimination.
    - Balanced data is recommended.
    - A public agency should overlook these applications to make sure it meets the criteria.
    - The same agency could be charged with looking into major decisions made by insurance agencies, healthcare providers, hospitals etc. to ensure the outcomes of said decisions are not unfairly biased. Similarly they could audit/analyze the

    outcomes from a given provider over the course of a year to see if any bias is present.

- The algorithms should have bias detection and mitigation techniques. The end-user and programmer of the algorithm should be connected and share feedback in order to best improve the algorithms.

- These algorithms need to be reviewed consistently, and maybe a way to do this is by publishing the outcomes of the algorithm every set period of time. This would allow for better analyses on any biases caught with the algorithms and tweaked to better target the racial inequalities that arise.
- Prior to being released on a large scale, certain information on algorithms such as the data it is being fed should be made public or be accessible to people with the knowledge to dissect and give a stamp of approval.
- The AI should also be able to develop a treatment plan given incomplete data about a patient's race, skin color, gender, or any other non-symptomatic information and 2 treatment plans should be made one accounting for the extraneous data and one without
- Some modeling techniques emphasize the simplicity/sparsity of the model. Using this sort of approach, we could make obtaining more secure data more penalized for the research group, which can then reduce the amount of data they may get in the future, or their potential funding for future project. This way, if secure data is needed, then the data can still be obtained, but if the result did not warrant the data used, then the group is penalized for the requested access.
-