Proposed Machine Learning Framework for Ethical Clinical Note-Taking

Khyati Singh

Abstract:

Ethnic and gender bias can creep into medical notes by physicians within a clinical setting. Physician bias as seen through verbiage of clinical notes and electronic health records (EHR) for certain demographics (e.g. African-Americans, Hispanic/Latinx, and women) can affect the quality of the health care these patients receive. This is important for future disease diagnosis and solutions for medical issues toward certain demographics, especially when human bias can get in the way. This study first reviews the literature on the extent and types of biases and impacts on certain demographics in the clinical setting. By looking into these issues, we propose a ML question framework to mitigate these biases when a clinician is taking notes in their pre-processing phase of data. Then, we assess the relevance of this framework within the context of the MIMIC-III dataset, where we evaluate a comparison of negative descriptors in the post-processing phase. Finally, we provide a set of conditions with a patient's background to guide a physician to record and evaluate medical needs in a holistic fashion to treat all patients fairly.

Introduction:

Lack of health equity has become a silent killer in healthcare. A computer program that combed through over 18,500 patient notes from January 2019 to October 2020 showed that Black patients were 2.5 times more likely to have negative patient descriptors within their records using language such as “not compliant,” “not adherent,” and “refused” (Sun 2022). This example of bias within physician notes has become a serious issue that has only begun to be recognized in the machine learning world. Bias can paint an incorrect picture of the patient to a physician, affecting treatment and contributing to health inequities across demographics. The rise of data science in health care has become invaluable because of its potential to mitigate bias to create fairness across demographics. The importance of variability in training data and having the framework create a space to mitigate bias is imperative to the development of ethical fairness. This is significant in looking at verbiage in medical notes that can affect the health care that patients receive.

The use of artificial intelligence and machine learning may improve bias by helping facilitate clinical decisions through personalized medicine, disease detection, and drug discovery and development. These new developments, while showing promise for the field of AI and machine learning, may have negative implications if AI processes are not reliable or data is not diverse. This drawback in models may instead exacerbate existing health inequalities rather than easing them (Igoe 2021).

In the context of clinical notes, existing health inequalities between demographics can be further driven by bias from healthcare physicians. Labeling patients of certain demographics with negative descriptors will deepen misunderstandings between patients and clinicians, furthering their bias. Bias is defined as prejudice in favor or against a person, thing, or group. In machine learning, it is seen as a systematic error when an algorithm produces results that are systematically prejudiced. When looking at healthcare notes, we see the impact of bias between demographics. Furthermore, implicit biases are associated with lower adherence to treatment plans from the patient due to more negative interactions that often occur between patient and provider. This can also negatively impact the quality of care those patients receive in the future (Pifer, 2022). It also leads to more discriminatory experiences that return worse psychosocial health outcomes for those demographics. Assessing how ethical and gender bias in clinical notes can result in unfavorable results for the patient is important to provide fairer outcomes that overlook prejudice and avoid systematic issues seen within electronic health records (EHR).

This study begins with a literature review aiming to identify different types of bias within healthcare settings. We examine different kinds of ethnic, gender, and socioeconomic biases that patients go through and the narrative beyond those statistics. By identifying the kinds of bias we present a technical machine learning framework surrounding how we can mitigate the bias within patient clinical notes. The framework will provide a guide to define a process of questions based on the patient to make the user more aware of bias within the clinician's notes. We present the thought framework for a machine learning model when certain descriptors or language are used within EHR to introduce the concept of fairness within the clinician notes. Fairness is seen as the ethical concept of treating people right from all demographics, and machine learning plays a role in correcting algorithmic bias within a model. In our ML framework model, we propose a series of considerations that helps identify biases within clinician notes and aids the propagation of fairer patient assessments.

Categories of Bias:

Ethnic Bias

Different ethnic groups such as African-Americans and Hispanic/Latinx groups tend to face bias within healthcare, have more unfavorable experiences, and are often stigmatized within EHR. One study utilizing machine learning methods and natural processing to analyze health care records found the possible transmission of racial bias within medical records. Black patients were found to have 2.5 times as many negative descriptors as their White counterparts, portraying the lack of understanding between providers and patients. Furthermore, White healthcare providers have been shown to view Hispanic and Latinx patients as unlikely to

accept responsibility for their care and more likely to be noncompliant with treatment recommendations (Hall 2015). This is often an implicit bias towards people of color that has influenced the way care is received and given in healthcare(Sun 2022).

Due to the type of treatment these demographics have previously received, machine learning models end up perpetuating a cycle of bias. A 2019 study found that risk-prediction algorithms led to Black patients receiving lower quality care than their non-Black counterparts. This occurred due to the algorithm using the patient's previous health care spending to determine future risks and thus the need for extra care. Thus, when machine learning algorithms use prior data that already contains bias, this ends up creating a cycle of systematically prejudiced results (Christensen 2021). This is why that bias is important to break down and challenge within ML models.

Gender Bias

Although bias tends to have the largest negative effect racial groups, it is important to note that women regardless of ethnicity are also more likely to experience bias in healthcare and treatment. A person’s gender can affect the relationship between a provider's implicit racial/ethnic bias and quality of care, especially in certain areas or specialties within the healthcare field such as emergency medicine and pediatrics (Hall, 2015). Often women also experience discrimination, with 17% of all women saying they feel they have been treated differently because of their gender compared to 6% of all men, respectively. Studies show that women’s perception of gender bias is correct. In several key areas, such as cardiac care and pain management, women may get different treatment, leading to poorer outcomes (Paulsen 2020). This example of worse health outcomes statistically shows how clinician bias can translate directly to patient harm.

Although this clear bias hasn’t been significantly reflected within clinical notes in terms of descriptors, women are more likely to have at least one quote within their clinical notes than men are. These quotes are not necessarily negative or positive, but women are more likely to have them (Beach 2021). More studies must be done about the linguistic bias of female clinical notes if they often face discriminatory experiences within healthcare as well.

Socioeconomic Status

Other factors affect bias as well. Patients with Medicaid or Medicare have higher adjusted odds of a negative descriptor compared to those with private or employer-based insurance. Unmarried patients also have higher adjusted odds of a negative descriptor compared to married patients. Poorer patients also tended to be blamed for their chronically unhealthy lifestyles (Sun, 2022). This often correlates with demographics that have been systematically

oppressed and confined in poorer communities.

These demographics tend to get caught in a cycle of health inequalities due to biases against them. Often these factors are out of their control, and solutions must be sought to mitigate these provider biases.

Metrics of Existing ML Frameworks Mitigating Bias in Healthcare:

Different solutions have been presented to mitigate the bias within healthcare, and not all are machine learning models. For machine learning models, however, there are different steps where bias can be introduced. Data collection and cleaning is the first and often most canonical example due to sampling bias. When looking at clinical notes, this bias can come with certain demographics being aggravated more than others due to the clinician's treatment of the patient. Sometimes bias exists in the model itself so we can also provide a checklist or guideline for the physician to follow so that greater awareness of how clinical notes are written could be a first step towards mitigating the problem.

Different models will set up different standards of metrics. IBM developed the AI Fairness 360 (AIF320) model, which has a clear set of metrics on the boundaries of modifying training data, learning algorithms, and predictions. These categories are broken down by pre-processing, in-processing, and post-processing (IBM N/A). If the user can modify training data, then pre-processing should be used. If they are allowed to change the learning algorithm, then in-processing could be used. The limitation would be that the user would need to be trained in manipulating the learning algorithm in this stage. If they can only use the learning model without the ability to modify the training data or learning algorithm, then they can use post-processing. AIF360 recommends that the user have permission to apply their pipeline in the earliest category to correct as much bias as possible. In our machine learning framework, the user can learn and use the model in post-processing after clinical notes, and the learning algorithm has been designed to check for bias. To apply our framework earlier within the pipeline, the thought framework could help in guiding pre-processing, a metric learned from the AIF360. This is essential for when the user or physician is creating the clinical notes for the patient. The earlier the guide can be used, the more bias can be avoided in clinical notes, which is why pre-processing is the ideal framework that we present.

In another objective framework for evaluating unrecognized bias in medical AI models predicting COVID-19 outcomes, the authors provided a standard level of model-level metrics for discrimination, accuracy, and reliability. The paper report that reliability is a key factor in AI/ML models' utility in clinical care. Many medical AI/ML models developed in healthcare settings ignore reliability and only report discrimination power (Estiri 2022). To ensure that a medical AI model is reliable the ML model must be closely partnered with the physician using it to validate

that the clinical implementation is correct (Balagurunathan 2021). Our machine learning thought process can be integrated into EHR to create a guide for physicians when taking notes and also be used to analyze the notes after the process.

When tackling bias in AI it is important to assess things with a set of guidelines for the model. By establishing the guideline of how one can create their framework then we get into the kinds of questions that the machine learning program should be asking and what kind of descriptors may have negative implications within clinical notes. Accountability, which must be understood by the AI designers to make sure the model is fair in its questions to the physician, is the first step to mitigating implicit biases. They must be held accountable for the creation of the model. Value alignment ensures that the model is aligned with the values of the user. For a physician or doctor to avoid bias, the model's purpose should fall in line with helping them be more ethical when writing their clinical notes by giving them more holistic questions for their patient. With explainability, the physician should easily be able to use the model to help ask their patient questions while the model can guide them with avoiding certain descriptors or biased writings within their clinical notes. It is important that the user, in this case, the physician, understand the reasoning of the machine learning decision process when asking questions and understand the negative implications of certain biased writing that the AI can notice. Reliability can be tested with the experience of physicians but also by using existing verbiage of clinical note databases to see how the thought framework stands against it. Fairness is significant because it will help minimize bias and promote an inclusive space. This can be shown by the kinds of questions that the model asks and the variability in training that it has had seeing different patient cases and questions to try to minimize the idea of bias for the physician or doctor. By setting guideline metrics for the kinds of things to be considered when creating a framework we can delve into the inner-making of it.

Model Proposal Possibilities:

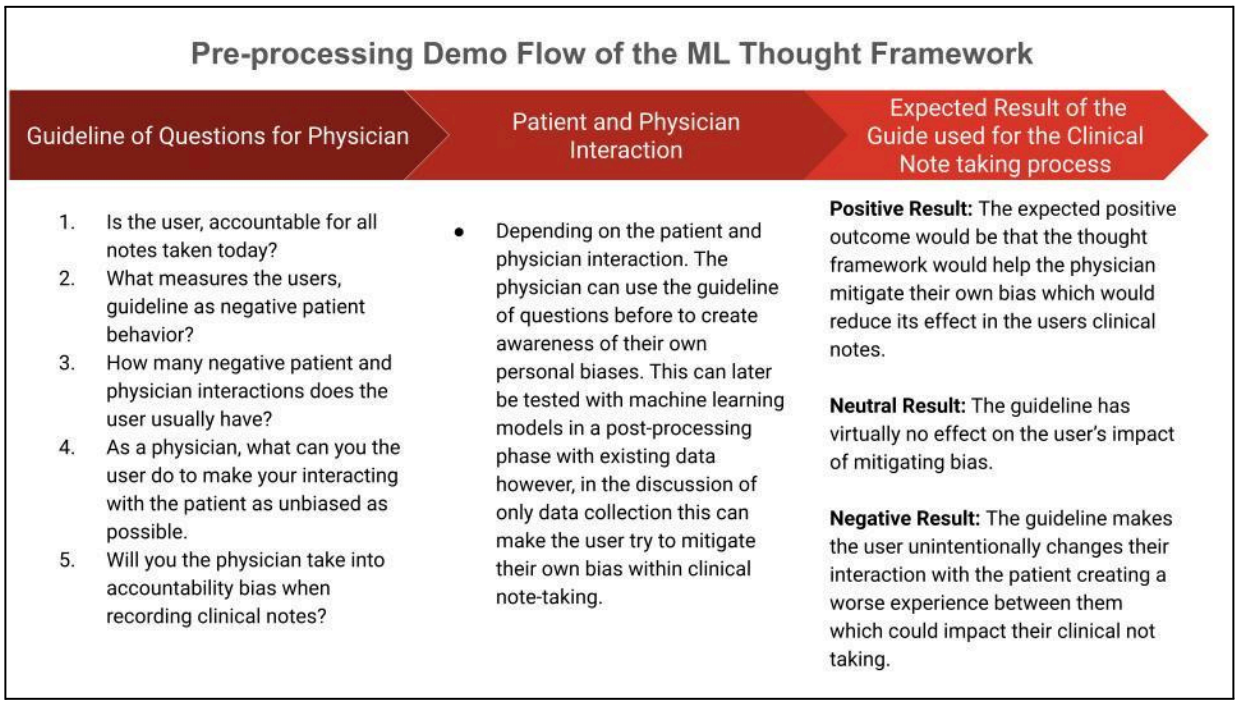
Here we propose a technical machine-learning framework that physicians and doctors can use when creating and assessing the EHR of patients to avoid compromising the quality of the patient's care. When proposing this machine learning idea we can use this thought framework for a physician to have implemented in EHR records so that in the clinical setting they have a checklist to guide their physician note-taking. Based on how the patient is described, the usage of clinical notes offers insight into the fact that physicians can analyze the amount of fairness they keep in their clinical notes with our thought framework. A series of demo questions can be raised.

If Key-Descriptor is Used:

1. Is the user sure that this word must have been used?
2. What out of the patient and physician interaction resulted in this verbiage?

3. How common is it for the physician to use this word?
4. Does the user believe another word could be used?
5. Does the user believe that a level of bias could have resulted in that verbiage?
6. What facets within this exchange between the involved parties prompted the utilization of this particular language?
7. To what degree could personal biases have contributed to the selection of this terminology?
8. Can the user affirm the precision of the selected term in this context?
9. In what ways do the dynamics of the physician-patient relationship impact the linguistic choices employed by both parties?
10. In what ways do the dynamics of the physician-patient relationship impact the linguistic choices employed by both parties?
11. Are there any pertinent cultural or contextual influences that may have shaped the adoption of this linguistic expression?
12. Has the user encountered comparable terminology in other healthcare dialogues?
13. Can the user identify any historical or societal factors contributing to the prevalence of this term?
14. How might the urgency of the medical situation influence the linguistic choices made by the physician?
15. Does the user believe that the usage of this term aligns with contemporary medical guidelines and standards?
16. Could regional or institutional practices contribute to variations in the frequency of this term's usage?
17. In what manner does the educational background of the physician impact their language selection in patient interactions?
18. Does the user anticipate that advancements in medical research could prompt alterations in terminology over time?
19. In what ways do the dynamics of the physician-patient relationship impact the linguistic choices employed by both parties?
20. Are there specific patient demographics that may elicit varied responses to the use of this term?

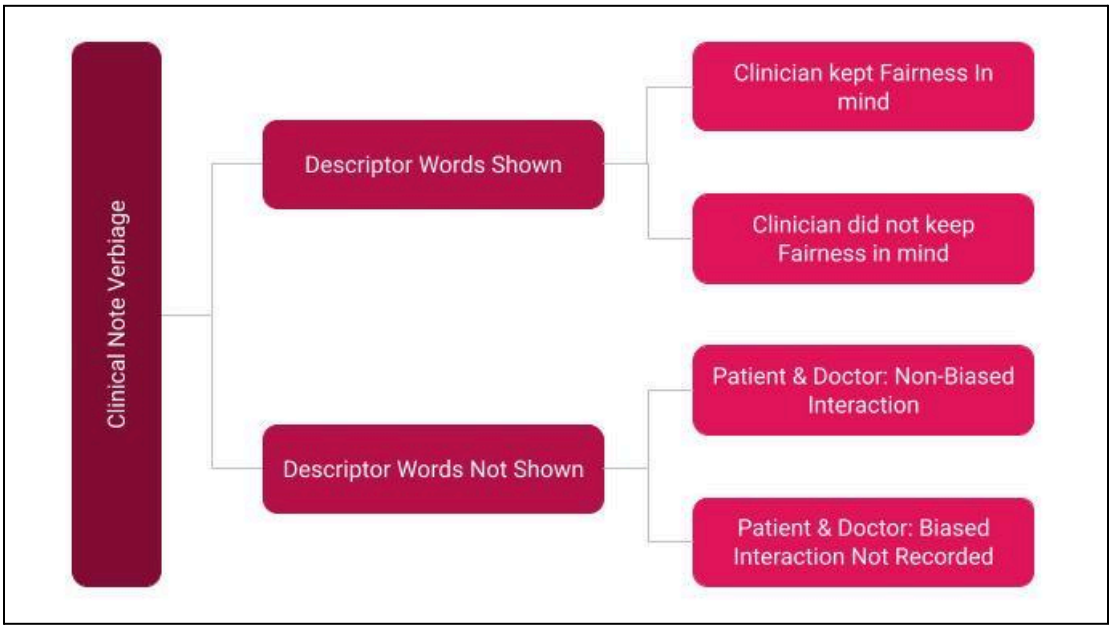
These questions can be raised when certain descriptors are flagged for the user to see fairness within their notes. If so, they have the power to change their language usage to correct and mitigate the bias from their physician's notes. Following the set of standard questions, the model can be used in a pre-processing setting where the data collection being the physician notes would be affected by the thought framework of an ML model. This would be directly involved in the physician's data collection to mitigate the most amount of bias.



An example of how we could test this framework is in a code of a quick demo set of questions within a program that a physician could use. In an in-processing model, we would have the user involved in the creation of the learning model and how it would react to the clinical notes. With the assumption that most users of this ML model would be physicians, they would be less involved in the creation of the algorithmic process. We would see this model more in the pre-processing or post-processing phases, data collection or if they are not modifying the data or learning algorithm would best be seen after when checking for fairness. This change is crucial to end the continuous cycle of bias that exists between physicians and their patients to close gaps in health inequity.

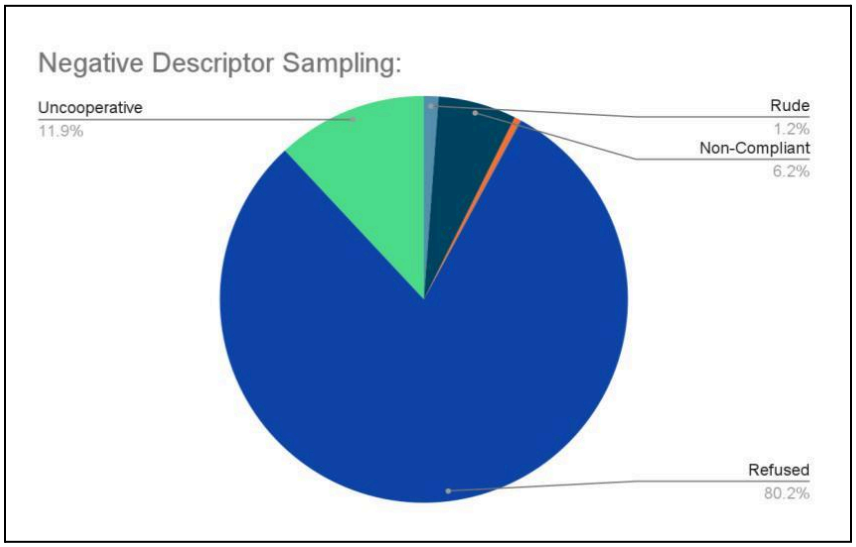
Exploring Implementation through Data:

In the post-processing phase, we would be able to evaluate the effects of the thought framework with verbiage that has already been written. Here we would see a clear process of the input of the user's clinical notes, and based on that data input, the model's job would be to assess how bias creeps in with the question framework designed to challenge that bias.



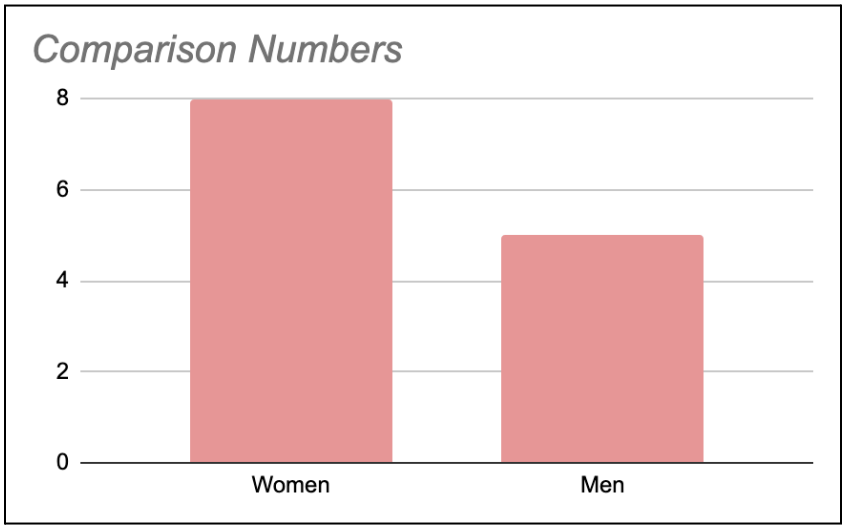
With existing clinical note verbiage in the post-processing phase, the model can identify and question whether descriptive words, such as “non-compliance” or “adherence,” are recorded within the clinical notes. From there it can question the clinician for the possibility that fairness was kept or not kept in mind. If descriptor words aren’t shown, the chance of a non-biased interaction between the patient and doctor is lower, but is still a possibility with the interaction not being recorded. This series of questions gives the user, the clinician, a chance to evaluate the bias of their notes based on those questions with certain verbiage emitting bias.

To see how this machine learning model works against a database in the post-processing phase of implementation, we gained access to MIMIC-III, a database of critically ill patients admitted to an intensive care unit (ICU) at the Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA. After processing the database, we were able to identify key descriptor words and the amount of times they appeared. By analyzing how many key-descriptor words show up, we can see how a physician can use our thought framework of questions. We can use our guideline set for our thought framework to challenge bias within the dataset.



Here we took a sampling of the database from ROW_ID 174 to ROW_ID 1003575 being the first 1,000,000 rows of data. We discovered the word “rude” was mentioned 166 times, “non-compliant” 891 times, and “non-adherent” 72 times. We can see how a vast majority, over 80%, of the negative descriptors that we looked at were “refused,” and the next with 11.9% was “uncooperative.” This is especially remarkable when seeing studies that explore racial and ethnic bias within physician disease diagnosis and clinical note-taking having an overwhelming amount of negative descriptors revolving around the patient’s compliance status.

Similarly, we can evaluate the amount of descriptors using Python to identify key descriptors comparing the percent difference between a sampling of women and men. Essentially, when looking at a sampling of data searching for negative descriptors for both women compared to men we see a higher estimation for women in this case.



This graph shows a higher ratio of women having more negative descriptors: “rude,” “non-compliant,” “non-adherent,” “uncooperative,” and “refused” when we search for the same descriptors in both genders. We looked at the rates of these words for women compared to men and saw an estimated 8 to 5 ratio within the sampling we examined. It is important to acknowledge that the code was limited and could have not been able to identify the gender of every row within the program however we used the same keywords such as “female,” “she,” and “her” to identify women within the physician notes under the TEXT column along with “male,” “he,” and “him” within the same sampling we identified to be accurate.

Comparing the usage of negative descriptors within our graph sampling between men and women, we see the prominence of women emerging in physician notes. This may portray the idea that female patients are harder to deal with not because of their actions but because of their being seen as “over-dramatic” or uncooperative as a result of physician bias. This will, in turn, result in worse quality healthcare for women when they are not taken as seriously with their symptoms and health. It’s important to acknowledge that gender is not the only factor that results in bias as previously stated and race also plays a monumental role in the quality healthcare one receives. We decided to evaluate sex rather than race due to less conclusive research about gender, and limitations of the MIMIC-111 dataset’s physician note structure. In further studies, results may come out differently using different ways to identify gender within the physician notes column and deciding to identify different negative key descriptors.

Conclusion:

Bias seen in both our research and countless other negative key descriptors can impact the way future physicians with access to these EHRs may view and treat a patient. These descriptors may be necessary especially when they properly can describe the patient and physician interaction; however, physicians must do their best to prevent unnecessary bias when it comes to the descriptors going into clinical note-taking. This can affect how positive the interaction ends and even how treatment is viewed. By examining existing solutions to bias within healthcare and proposing our machine-learning framework, we have provided an existing guideline for how we can mitigate bias within clinician notes. The integration of this innovative framework holds promise for evolving clinical practices to foster an unbiased healthcare environment and advance the goal of providing equitable healthcare for diverse patient populations. By leveraging technology to mitigate biases, we aim to contribute to a future where medical professionals make decisions based on objective, data-driven considerations, ultimately enhancing the overall quality and fairness of healthcare delivery to patients of all demographics.

References:

1. Sun, Michael, et al. “Negative Patient Descriptors: Documenting Racial Bias In The Electronic Health Record.” Health Affairs, University of Chicago, 19 Jan. 2022, www.healthaffairs.org/doi/10.1377/hlthaff.2021.01423.

2. Igoe, Katherine J. “Algorithmic Bias in Health Care Exacerbates Social Inequities - How to Prevent It.” Executive and Continuing Professional Education, Harvard, 3 Oct. 2023, www.hsph.harvard.edu/ecpe/how-to-prevent-algorithmic-bias-in-health-care/.

3. Hall, William J, et al. “Implicit Racial/Ethnic Bias among Health Care Professionals and Its Influence on Health Care Outcomes: A Systematic Review.” American Journal of Public Health, U.S. National Library of Medicine, Dec. 2015, www.ncbi.nlm.nih.gov/pmc/articles/PMC4638275/.

4. Paulsen, Emily, and Emily Paulsen. “Recognizing, Addressing Unintended Gender Bias in Patient Care.” Duke Health Referring Physicians, Duke Health, 14 Jan. 2020, physicians.dukehealth.org/articles/recognizing-addressing-unintended-gender-bias-patient-care#:~:text=One%20in%20five%20women%20say,of%20gender%20bias%20are%20correct.

5. Beach, Mary Catherine, et al. “Testimonial Injustice: Linguistic Bias in the Medical Records of Black Patients and Women - Journal of General Internal Medicine.” SpringerLink, Springer International Publishing, 22 Mar. 2021, link.springer.com/article/10.1007/s11606-021-06682-z.

6. Fairness, AI. “Guidance on Choosing Metrics and Mitigation.” Ai Fairness 360 - Resources, IBM Research Trusted AI, aif360.res.ibm.com/resources#guidance.

7. Estiri, Hossein, et al. “An Objective Framework for Evaluating Unrecognized Bias in Medical AI Models Predicting COVID-19 Outcomes.” OUP Academic, Oxford University Press, 12 May 2022, academic.oup.com/jamia/article/29/8/1334/6576634.

8. Balagurunathan, Yoganand, et al. “Requirements and Reliability of AI in the Medical Context.” Physica Medica : PM : An International Journal Devoted to the Applications of Physics to Medicine and Biology : Official Journal of the Italian Association of Biomedical Physics (AIFB), U.S. National Library of Medicine, Mar. 2021, www.ncbi.nlm.nih.gov/pmc/articles/PMC8915137/.

9. Mahmood, Anam. “Tackling Bias in Machine Learning Models.” IBM Developer, IBM, 22

Mar. 2021, developer.ibm.com/articles/tackling-bias-in-machine-learning-models/.

10. Christensen, Donna M, et al. “Medical Algorithms Are Failing Communities of Color.” Health Affairs, Health Affairs, 2021, www.healthaffairs.org/content/forefront/medical-algorithms-failing-communities-color.

11. Pifer, Rebecca. “Study Finds Racial Bias in How Clinicians Describe Patients in Medical Records.” Healthcare Dive, Healthcare Dive, 20 Jan. 2022, www.healthcaredive.com/news/racial-bias-patient-descriptors-medical-records-health-affairs/617422/.