Rising like a Phoenix: Emerging from the Pandemic and Reshaping Human Endeavors with Digital Technologies ICIS 2023

IS in Healthcare Addressing the needs of post-pandemic digital healthcare

Dec 11th, 12:00 AM

# The Impact of Stigmatizing Language in EHR Notes on AI Performance and Fairness

Yizhi Liu
*University of Maryland*, yizhiliu@umd.edu

Weiguang Wang
*University of Rochester*, wwang90@simon.rochester.edu

Gordon Gao
*Johns Hopkins University*, gordon.gao@jhu.edu

Ritu Agarwal
*Johns Hopkins University*, ritu.agarwal@jhu.edu

## Recommended Citation

Liu, Yizhi; Wang, Weiguang; Gao, Gordon; and Agarwal, Ritu, "The Impact of Stigmatizing Language in EHR Notes on AI Performance and Fairness" (2023). *Rising like a Phoenix: Emerging from the Pandemic and Reshaping Human Endeavors with Digital Technologies ICIS 2023*. 23.
https://aisel.aisnet.org/icis2023/ishealthcare/ishealthcare/23

# The Impact of Stigmatizing Language in EHR Notes on AI Performance and Fairness

*Completed Research Paper*

**Yizhi Liu**
Robert H. Smith School of Business,
University of Maryland, College Park
7621 Mowatt Ln, College Park,
Maryland, USA
yizhiliu@umd.edu

**Weiguang Wang**
Simon Business School,
University of Rochester
300 Wilson Blvd, Rochester, New
York, USA
wwang90@simon.rochester.edu

**Guodong (Gordon) Gao**
Johns Hopkins Carey Business School,
Center for Digital Health and AI
(CDHAI)
100 International Drive, Baltimore,
Maryland, USA
gordon.gao@jhu.edu

**Ritu Agarwal**
Johns Hopkins Carey Business School,
Center for Digital Health and AI
(CDHAI)
100 International Drive, Baltimore,
Maryland, USA
ritu.agarwal@jhu.edu

## Abstract

*Today, there is significant interest in using electronic health record data to generate new clinical insights for diagnosis and treatment decisions. However, there are concerns that such data may be biased and result in accentuating racial disparities. We study how clinician biases reflected in EHR notes affect the performance and fairness of artificial intelligence models in the context of mortality prediction for intensive care unit patients. We apply a Transformer-based deep learning model and explainable AI techniques to quantify negative impacts on performance and fairness. Our findings demonstrate that stigmatizing language written by clinicians adversely affects AI performance, particularly so for black patients, highlighting SL as a source of racial disparity in AI model development. As an effective mitigation approach, removing SL from EHR notes can significantly improve AI performance and fairness. This study provides actionable insights for responsible AI development and contributes to understanding clinician EHR note writing.*

**Keywords:** artificial intelligence, electronic healthcare records, stigmatizing language, racial disparity, AI fairness

## Introduction

Healthcare is in the midst of an artificial intelligence (AI)-driven transformation and is being reshaped both clinically (Rajpurkar et al. 2022), with high-performing models for detecting diseases (Esteva et al. 2017;

Hannun et al. 2019), and operationally (Lou and Wu 2021), where the important yet largely unfulfilled opportunity of medical AI is attracting substantial investment. The benefits of adopting AI in healthcare are significant, estimated to reduce healthcare spending by 5% to 10%, or $200 billion to $360 billion a year[1].

Natural Language Processing (NLP) AI models are emerging as especially useful for medical AI, with remarkable achievements being documented for large language models such as ChatGPT (Chat Generative Pre-training Transformer) such as passing the medical licensing exam (Kung et al. 2023) and authoring scientific papers (Stokel-Walker 2023). 80% of our healthcare information is recorded in unstructured text and digitized in electronic health records (EHRs) (Negro-Calduch et al. 2021). These data, capturing the physician's expertise and diagnostic and treatment decisions, together with observations about the patient, represent a vast untapped resource and can now be leveraged through the application of NLP (Sarzynska-Wawer et al. 2021; Lee et al. 2020). Models such as ClinicalBERT (Clinical Bidirectional Encoder Representations from Transformers) (Huang et al. 2019) and BEHRT (BERT for EHR) (Li et al. 2020) incorporate all available data tokens (or "words") to create medical predictions, without considering the unwanted non-medical patterns embedded in the data. However, EHR data, especially the free text in clinical notes, do not solely reflect objective clinical and medical facts; rather, behavioral patterns and even their biases in clinical practices may be recorded in the training data (Ganju et al. 2020); these factors may well be critical for AI fairness (Buolamwini and Gebru 2018; Bolukbasi et al. 2016).

The importance of ensuring that AI is fair and equitable and treats all "populations" in an equivalent manner (Schwartz et al. 2022; The White House 2022) is especially critical for health-related decisions, as health inequity can have life-changing consequences. Racial biases and disparities are widely present in healthcare and have the potential to be exacerbated by digital technologies and AI (Agarwal et al. 2022). Numerous studies have documented this phenomenon. Chen et al. (2021) summarize the racial inequalities in various aspects of clinical practice revealed by recent AI models. Tamayo-Sarver et al. (2003) show that black patients were less likely to be provided opioids than whites and Latinos. Dresser (1992) highlights the racial bias in medical research and notes that a majority of medical findings are based on conclusions drawn from white male samples that exclude minority patients. There is a growing concern that EHR data that can embed these biases, when used as the training fuel for AI, may propagate or even magnify implicit biases (Sun et al. 2022). In this study, we focus on a specific type of bias in EHR notes, *stigmatizing language* (SL), which reflects clinicians' implicit bias towards patients. An example of SL in EHR notes is shown in Figure 1, with pejorative words such as "abuser" and "noncompliant" in the text. Previous research has documented the widespread existence of SL in EHR notes, with different racial groups being affected disproportionately. For example, the EHR notes of black patients contain significantly more SL than the EHR notes of white patients (Himmelstein et al. 2022).

We conduct a series of experiments to examine how SL affects AI performance and fairness in a classic clinical prediction task, i.e., mortality prediction, for ICU patients. Specifically, we first train a Transformer-based model on EHR notes and examine how the presence of SL in the testing data alters mortality prediction outcomes. Next, we focus on AI fairness by investigating whether black patients are disadvantaged by the AI's predictions, and if SL in EHR notes is associated with such racial disadvantage. Finally, we study if the removal of SL from the training data can be a potential solution to help reduce its adverse influence.

> "...pt is a long time **abuser** of etoh, s/p CVA in [**2184**] w/ minimal sequelae. pt has been **noncompliant** for years regarding medical care, etoh addiction, etc. pt is estranged from all family, and has a mentally disabled girlfriend as well (the witness of the pt's fall)..."

**Figure 1. An Example of An EHR Note That Contains SL**

[1] Source: https://www.healthcaredive.com/news/artificial-intelligence-healthcare-savings-harvard-mckinsey-report/641163/, accessed on Mar 19, 2023

Our data come from a publicly available archival EHR dataset, Medical Information Mart for Intensive Care III (MIMIC-III), which consists of de-identified EHR records associated with over 60,000 intensive care unit (ICU) admissions over the 2001-2012 time period. The dataset contains a wide range of attributes related to patients, such as free-text clinical notes, demographic characteristics, the ID of the clinician who treated the patient, admission and discharge records, and death time (if applicable). We first use the free-text clinical notes as the textual features to feed a Transformer-based model for mortality prediction, with the label obtained from the death time attribute. Second, we utilize a list of SL keywords identified by previous research to determine if SL exists in each note (Himmelstein et al. 2022), as shown in Table 1. Then, we are able to understand the impact of SL on AI performance by manipulating the existence of SL in the testing and training data, respectively. Third, the recorded ethnicity of patients allows us to evaluate AI fairness. Specifically, if the model performs unequally on the black and white patient instances or subsets, we can conclude the existence of AI racial biases.

'adherence', 'nonadherent', 'compliance', 'unwilling', 'abuse', 'belligerent', 'drug seeking', 'abuser', 'difficult patient', 'refused', 'refuses’, 'noncompliance', 'argumentative', 'cheat', 'abuses', 'malingering', 'user', 'secondary gain', 'in denial', 'refuse', 'compliant', 'substance abuse', 'nonadherence', 'degenerate', 'drug problem', 'combative', 'fake', 'been clean', 'noncompliant', 'addicted', 'narcotics', 'habit', 'adherent’

**Table 1. A Keyword List of SL**

Our key findings are threefold. First, using explainable AI (XAI) methods such as the leave-one-out and input reduction strategies, we find that SL can hinder the performance of a trained AI model for mortality prediction. Second, utilizing adversarial perturbation techniques, we reveal that the trained model exhibits racial biases related to the presence of SL. Specifically, the prediction remains almost unchanged when the model is informed that the patient is Caucasian, while indicating that the patient is black leads to a decrease in the predicted probability by 15.97%. Interestingly, such a racial gap in the model’s predicted probability almost disappears if SL is removed, suggesting that SL is associated with the model’s racial bias. Consistent findings are obtained from global explanations derived from the global leave-one-out strategy and global adversarial perturbations on all testing examples. Third, we explore the training set and find that removing SL helps improve the model performance and narrows the racial gap in mortality prediction. In particular, using the first 24 hours of ICU admission as an observation window, the removal of SL narrows the racial gap between the model’s predicted F1 scores for black and white patients from 2.97% to a mere 0.05%.

This study contributes to the general literature on AI in healthcare, and more specifically, to the areas of health equity and racial disparities. AI models that can harness complex EHR data for clinical outcome prediction are increasingly developed as the digitization of healthcare data continues (Seinen et al. 2022). The breakthroughs in NLP, such as large language models, make unstructured clinical notes more commonly used compared to structured data (Patel and Lam 2023). However, new challenges of racial disparity may emerge if clinical notes are used by AI models, but the exploration of such challenges remains limited. This study identified a striking phenomenon that warrants careful consideration when developing and implementing AI models in clinical decisions that rely on clinical notes. Previous research has expressed concerns about the racial disparity of AI mainly because of the human biases embedded in the clinical text that can be inherited by AI (Posner and Fei-Fei 2020). This study sheds nuanced light on these concerns by showing that a trained AI model could still produce significant racial disparities even if the training data is devoid of systematic discrimination.

## Literature Review and Hypothesis Development

### *Race in Medical Decision Making and AI*

The US Census Bureau defines five distinct races: White, Black or African American, American Indian or Alaska Native, Asian, and Native Hawaiian or Other Pacific Islander.[2] This categorization is based on one's ancestral geographical region and reflects a unique genetic heritage (Risch et al. 2002). "Race" plays a significant role in various aspects of medical decision making (MDM), as population genetic studies have identified race-related genetic variations and ancestral-tree diagrams supporting the grouping within races (Calafell et al. 1998; Bowcock et al. 1994; Bowcock et al. 1991). Since biological differences exist between races, racial information can be used legitimately in clinical settings and is often incorporated in MDM (Merryweather-Clarke et al. 2000; Stephens et al. 1998). Despite the potential benefit, it must be cautiously balanced against potential harm. The presence of race-based health inequities in society is widely recognized, and factoring race into MDM may unintentionally lead to adverse consequences that worsen race-based disparity. For example, using race in MDM may reinforce negative stereotypes and increase the likelihood of stigma and discrimination against certain racial groups. Similarly, using race to identify individuals at higher risk for substance abuse may lead to increased scrutiny and suspicion of individuals from those racial groups, even if they do not have a history of substance abuse.

As AI gains increasing prominence in healthcare, concerns have been raised about its potential to exacerbate health disparities (Chen et al. 2021; Char et al. 2020). Recent studies and the World Health Organization have urged a reassessment of using racial information in AI (World Health Organization 2021; Vyas et al. 2020). Data and algorithms are fundamentally products of human creation and reflect individual, social, and institutional biases. Consequently, AI directly trained on the EHR data could inherit the racial bias from human doctors and perpetuate discriminatory practices, such as referring black patients less to critical treatments that save their lives. Given that a wide variety of human biases are created by social and economic environments, the healthcare system, and individual health professionals, they are widely prevalent and deeply embedded in clinical data. In this case, incorporating racial information could make an AI model inevitably racially biased if directly trained on real clinical data.

In addition, AI's inability to perform human causal reasoning makes it more prone to causing racial harm. Machine learning models directly extract association patterns from a complex combination of various input features. Without sufficient guidance, healthcare AI models may misapply certain patterns. As a vivid example, Obermeyer et al. (2019) documented that racial bias in healthcare machine learning models results from using health costs as a proxy for health needs. Lower spending on healthcare by black people is not because of lower needs but rather socioeconomic factors that constrain the fulfillment of these needs. Relying on historical data, the AI falsely interprets lower spending as a signal of better health than equally sick white patients. Similarly, Zhang et al. (2020) start from the NLP perspective and indicate that contextual language models trained on scientific articles tend to recommend prisons for violent black patients and hospitals for violent white patients in completing clinical note templates.

## *Implicit Bias, Stigmatizing Language, and AI*

MDM studies have uncovered numerous forms of implicit biases, including racial biases that pervade the healthcare system (Chapman et al. 2013), such as prejudice that black people are violent (Holroyd et al. 2016), and such racial biases even exist in Intensive Care Units (ICU) (FitzGerald and Hurst 2017; Martin et al. 2016). Racial biases can be evident in actions, such as being less attentive to a black patient or trivializing patient fears and preferences (Beck et al. 2022), or in language, such as writing discriminatory language against certain groups of patients (Park et al. 2021). Since clinicians' biases can be recorded in EHR data, and healthcare AI models are often developed using such data containing human biases, EHR data has become a core source of racial disparity in healthcare AI (Parikh et al. 2019).

Gaining deeper insights into medical providers' biases and assessing their influence on healthcare AI racial fairness is crucial. Thus, we focus on medical notes where clinicians express their unfiltered opinions. With

[2] Source:
https://www.census.gov/quickfacts/fact/note/US/RHI625221#:~:text=OMB%20requires%20five%20minimum%20categories,report%20more%20than%20one%20race. accessed on Mar 19, 2023

the growing availability of EHR data and the breakthroughs of NLP AI, medical notes have become essential training data for various AI models, such as those predicting mortality, specific diseases, readmission to hospitals, and length of stay (Seinen et al. 2022). The striking success of NLP models, such as the widely-acclaimed performance of ChatGPT, allows medical notes to be used not only for extracting certain medical entities as subsequent models' input. Instead, complex deep learning models have evolved to be applied in making various clinical predictions based on the information contained in medical notes, ranging from traditional NLP models like TFIDF (term frequency-inverse document frequency), to basic artificial neural networks like RNN (recurrent neural networks) and CNN (convolutional neural networks), to the Transformer-based models in recent studies (Wen et al. 2020). However, the advance of deep learning models also comes with diminished interpretability (black box issue), which makes the impact of implicit racial biases in medical notes more opaque and potentially more detrimental to racial fairness (Rudin 2019).

Within clinical notes, biases manifest as SL (Himmelstein et al. 2022). This form of language typically discriminates against an identifiable group of people, a place, or a nation (National Institute on Drug Abuse 2021). SL attributes negative labels, stereotypes, and judgments to certain groups of people, including inaccurate or unfounded beliefs that portray them as dangerous, incapable of managing treatment, or at fault for their condition (Werder et al. 2022). Utilizing SL can amplify clinicians' implicit biases and negatively impact patients' sense of hope (Kelly et al. 2015). As contemporary deep learning-empowered AI enters the field of medical notes-based predictions, it becomes increasingly important to understand the ramifications of SL on cutting-edge deep learning models. Our study addresses this by investigating the effect of including and excluding SL in medical notes for a mortality prediction task within an ICU context, using a Transformer based deep learning model. Previous research often emphasizes the importance of predictor selection in ensuring AI fairness, focusing on clinical and demographic variables (Chen et al. 2021; Obermeyer et al. 2019). Our study uniquely concentrates on SL as a novel source of bias, thereby contributing a new dimension to this crucial discussion.

Contrasting with an AI that decides whether to refer an end-stage renal disease patient to critical treatments (e.g., renal transplantation), the outcome measure for mortality prediction is, in theory, less affected by clinicians' racial bias (McGarvey et al. 2007). While a patient referral for treatment could be biased due to the clinician's decision patterns, mortality is a more objective outcome than clinicians' subjective assessment, i.e., a physician's estimate of mortality risk. However, the presence of SL signals bias, and it is plausible that patients with SL in their notes may receive clinically inferior recommendations of treatments, similar to a clinician making sub-optimal treatment choices based on race. Thus, although mortality is an objective outcome, the endpoint of mortality is not impervious to bias. Whether the presence of SL improves or detracts from the performance of an AI for mortality prediction remains to be an unanswered question.

Echoing the ongoing debate in medicine regarding the use of racial information, incorporating SL into state-of-the-art deep learning models, e.g., Transformer based models, could potentially have both positive and negative consequences. On the one hand, systematic discrimination against certain racial groups in the SL may be present if clinicians discriminate against these groups during treatment, resulting in higher mortality rates than other racial groups (Greenwood et al. 2020). In this case, SL would effectively help predict the medical consequences (i.e., mortality) of biases in health practice. As a result, removing SL from the deep learning model could cause information loss, thereby reducing mortality prediction performance. This is especially true for Transformer based deep learning models which use a self-attention algorithm that learns from the global contexts of all inputs rather than only focusing on local contexts (e.g., words surrounding a focal word). This enables the model to integrate more information (e.g., patterns in SL across different races) to make predictions (Vaswani et al. 2017). Therefore, we test:

**Hypothesis 1a (H1a).** *Removing SL from medical notes can reduce the performance of AI models for mortality prediction.*

On the other hand, despite the pervasiveness of racial disparities, they do not necessarily capture informative patterns in SL embedded in medical notes. Park et al. (2021) conduct an extensive analysis of SL in EHR notes and suggest that SL might merely express clinicians' feelings (positive or negative), which could be not informative but reflective of their biases. If SL is noninformative and lacks a clear systematic racial bias, it essentially adds noise to the information. Prior research has discussed the negative impact of noise in medical notes (Xiao et al. 2018; Miotto et al. 2016), suggesting AI performance could be hindered by the included SL. In other words, removing noise in training data (i.e., the noninformative SL in medical notes) can effectively improve AI performance (Yang and Song 2010). Accordingly, we hypothesize that:

**Hypothesis 1b (H1b).** *Removing the SL from medical notes can improve the performance of AI models on mortality prediction.*

Among all racial groups, black patients experience greater racial disparities than white patients in clinical practice (Bailey et al. 2021). It is therefore reasonable to believe that black patients, compared with white patients, face additional disadvantages in receiving high-performance predictions of a healthcare AI, especially in a critical context like mortality prediction. If SL mirrors racial discrimination in clinicians' healthcare practice, the information value of SL as an indicator decides its effectiveness in improving mortality prediction. As the information value of data correlates with racial biases (Axt and Lai 2019), the same word can be interpreted as reflecting biases toward black patients while indicating personalized care for white patients. To illustrate, SL can be helpful in predicting health outcomes, if a clinician makes an annotation of "non-compliant" for a white patient while specifically referencing a medication or appointment that the patient missed. However, the same SL can lose its informativeness if the same word is utilized with black patients broadly and stereotypically. In this case, white patients benefit more than black patients in the prediction if SL is incorporated.

Even if SL acts as a noise rather than a predictor of the consequences of biases in health practice, the impact of the noise could still be heterogeneous and pose a risk to AI fairness. Since noise hurts the performance of the mortality prediction AI, the level and pattern of using SL in patients of different racial groups determine their level of disruption on AI's performance. Studies have documented that the medical records of black patients include more SL than those of white patients (Sun et al. 2022). Therefore, SL disrupts AI performance more for black patients than for white patients.

**Hypothesis 2 (H2).** *The existence of SL in medical notes is associated with racial disadvantage in AI fairness for black patients.*

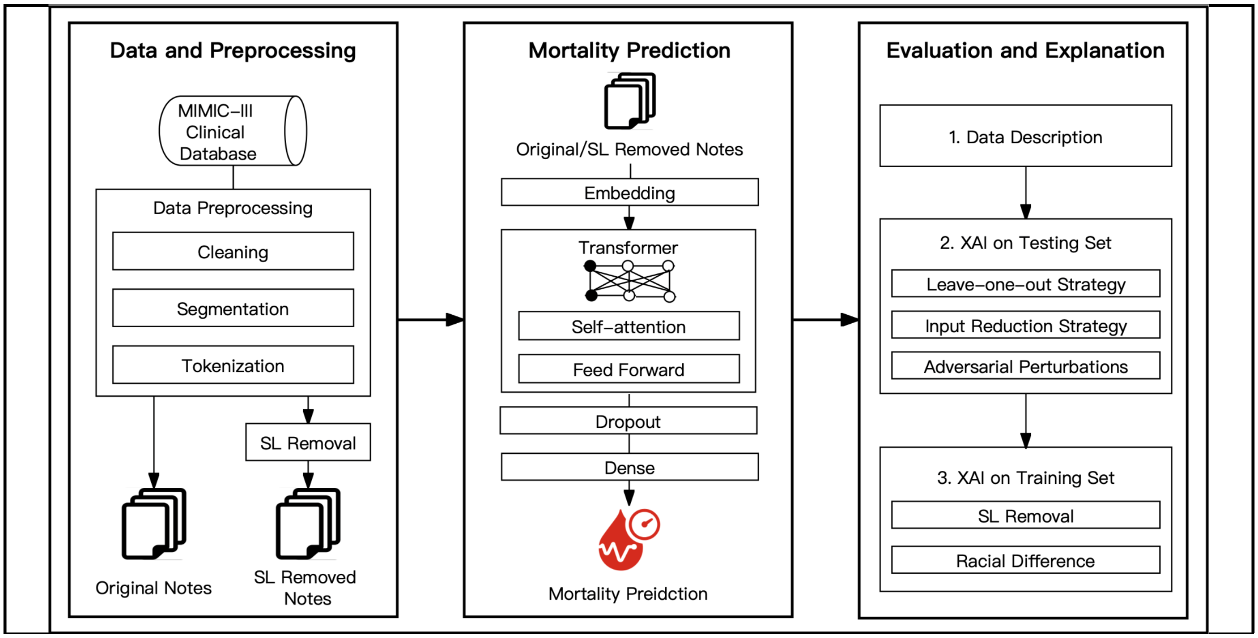

**Figure 2. Our Proposed Research Framework**

## Research Design and Procedure

To investigate our research questions, we develop a Transformer-based deep learning model using the MIMIC-III Clinical Database and apply XAI techniques to understand SL's effects. As illustrated in Figure 2, our research framework comprises three major components: Data and Preprocessing, Mortality Prediction, and Evaluation and Explanation. We explain each component in detail next.

### *Data and Preprocessing*

Large amounts of labeled data are often necessary for deep-learning-based AI models to function effectively. To facilitate the development of clinical predictive models, medical researchers have made extensive EHR datasets available. Various models relying on EHR notes are subject to biases if SL is indeed problematic. Among existing EHR datasets, MIMIC-III (Medical Information Mart for Intensive Care III) is the most widely used for model development in previous studies (Johnson et al. 2016). Therefore, we follow the literature to use MIMIC-III for a well-defined task, mortality prediction, by using either all records of ICU stays (i.e., full dataset) or the first 24 hours of ICU stays (i.e., 24-hour subset) (Seinen et al. 2022).

Using the same preprocessing approach as existing literature, such as the exclusion of discharge notes and token segmentation, we transform the complete data into a clean input to AI models (Chen et al. 2019). Identifying SL in EHR notes is a context-specific task requiring strong domain knowledge. Fortunately, an SL keyword list has been developed by prior research and is widely used to identify (and to further eliminate) the SL in EHR notes (Himmelstein et al. 2022; Association of Diabetes Care and Education Specialists 2021; National Institute on Drug Abuse 2021). As a result, we create another version of data by removing the SL keywords to examine the impact of SL on the subsequent predictive model development.

### *Mortality Prediction*

To perform the mortality prediction task, we utilize a cutting-edge NLP model structure, Transformer, which is the foundation of the emerging GPT models (e.g., ChatGPT and GPT-4[3]). Transformer has been used by prior research and achieved state-of-the-art performance on clinical prediction tasks using EHR notes (Wen et al. 2020). A common practice in deep learning model development is transfer learning, where new models are developed on top of an existing pre-trained model. While previous models (e.g., ClinicalBERT) have achieved state-of-the-art performance on clinical prediction tasks, they were pre-trained on sizable EHR notes, with MIMIC-III as a common pre-training data source (Seinen et al. 2022). Therefore, these pre-trained models are highly relevant to the potential racial bias of SL. However, to fully investigate our research question, we need to develop a Transformer model without transfer learning. This is because using those pre-trained models, which have learned MIMIC-III data through transfer learning, for the mortality prediction on MIMIC-III data is subject to both information leakage and indirect impact of SL. Therefore, we build a Transformer model from scratch without any pre-training. The model takes the EHR notes as input and makes a binary judgment on each patient's mortality risk, expressed as a predicted probability. If this probability exceeds 50%, a standard threshold, the patient is deemed at risk of mortality. The model was built using the TensorFlow framework version 2.10, and the experiments were conducted on a single Microsoft Windows 11 Pro Server with 128GB of RAM, an Nvidia GeForce GTX 3090 GPU, and an 12th Intel i9-12900K CPU at 3.20 Gigahertz (GHz).

### *Evaluation and Explanation*

Since the decision-making process of AI models is usually complex, determining the effect of the model inputs on the decisions of AI is challenging. Therefore, we apply XAI techniques to understand how SL

[3] Due to ethical reasons, ChatGPT (both GPT-3.5 and GPT-4 versions) cannot predict specific clinical outcomes such as mortality. ChatGPT's standard response to such requests involves summarizing key elements in the EHR notes and recommending consultation with a licensed medical professional for a comprehensive assessment.

affects AI performance during the testing and training phases. Information Systems researchers have also proposed similar methods, such as counterfactual explanations (Fernandez et al. 2022; Martens and Provost 2014), which entails considering a scenario: *If a certain input were absent, what would have the AI model predicted?* The XAI techniques used in this study align with this idea.

Our evaluation and explanation focus on two questions. First, how does SL in future prediction (the testing data) affect racial disparity when a Transformer-based deep learning model is trained on EHR notes containing SL? To examine the impact of SL on the testing set, we employ three XAI techniques: leave-one-out strategy (Li et al. 2016), input reduction strategy (Feng et al. 2018), and add-sentence adversarial perturbations (Jia and Liang 2017). Second, is it beneficial to remove SL in the training phase? We proceed to train the same model using EHR notes with SL and without SL. We assess the impact of SL by comparing the model performance using four metrics: Accuracy, Precision, Recall, and F1 Score.

# Results

## *Data Description*

MIMIC-III is a widely used de-identified dataset focusing on critical care units (Johnson et al. 2016). Despite the rich information, the readiness of the raw MIMIC-III data is limited by its noisy nature and complexity in structure. Previous researchers have developed open-source pipelines to convert the raw MIMIC-III data into suitable data structures for clinical prediction tasks (Wang et al. 2020). Such pipelines also allow directly comparing different methods on the same clinical prediction tasks. In this study, we use the open-source pipeline MIMIC-Extract (Wang et al. 2020) to process the raw MIMIC-III data, resulting in our complete data, and we focus on the EHR notes in combination with the mortality results. As shown in Table 2, the full dataset contains 566,597 notes, while the 24-hour subset contains the EHR notes of the first 24 hours, covering 86.37% of all patients.

| Dataset | # of SL Notes | # of Non-SL Notes | # of Notes |
|---|---|---|---|
| Full | 121,765 | 444,832 | 566,597 |
| 24-hour | 11,661 | 123,862 | 135,523 |
| Dataset | Patients | # of Notes with Label = 1 | # of Notes with Label = 0 |
| Full | 32,671 | 76,202 | 490,395 |
| 24-hour | 28,565 | 12,913 | 122,610 |

**Table 2. Statistics of Preprocessed Data**

The number and percentage of EHR notes that contain SL in the data are presented in Table 3. On average, 21.49% of EHR notes contain SL, with black patients having a 1.86% higher likelihood of having SL in their EHR notes compared to white patients, consistent with prior studies (Himmelstein et al. 2022; Sun et al. 2022). The statistics suggest a noticeable presence of SL in the EHR notes, serving as a vehicle of clinicians' bias toward patients.

| Ethnicity | # of SL Notes | Percentage | # of Non-SL Notes | Percentage |
|---|---|---|---|---|
| Black | 9,705 | 24.23% | 30,350 | 75.77% |
| White | 92,598 | 22.37% | 321,423 | 77.63% |
| Hispanic | 4,246 | 22.72% | 14,444 | 77.28% |
| Asian | 2,669 | 19.25% | 11,193 | 80.75% |
| Others | 12,547 | 15.69% | 67,422 | 84.31% |
| Sum | 121,765 | 21.49% | 444,832 | 78.51% |

**Table 3. Number of EHR Notes by SL and Ethnicity**

## *Experimental Results: XAI on the Testing Set*

### XAI: Leave-one-out Strategy

| Input | Probability | Improvement | Word Importance | At Risk? |
|---|---|---|---|---|
| Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. When he aroused he was very ***combative***. | 48.23% | - | - | NO |
| Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. [MASK] he aroused he was very ***combative***. | 44.55% | -3.68% | #2 | NO |
| Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. When [MASK] aroused he was very ***combative***. | 46.34% | -1.89% | #4 | NO |
| Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. When he [MASK] he was very ***combative***. | 27.19% | -21.04% | #1 | NO |
| Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. When he aroused [MASK] was very ***combative***. | 46.16% | -2.07% | #3 | NO |
| Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. When he aroused he [MASK] very ***combative***. | 50.28% | 2.05% | #6 | YES |
| Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. When he aroused he was [MASK] ***combative***. | 49.55% | 1.32% | #5 | NO |
| Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. When he aroused he was very [MASK]. | 59.76% | 11.53% | #7 | YES |

**Table 4. Results of the Leave-one-out Strategy**

We first apply the leave-one-out strategy to examine whether SL in the testing set affects our AI model performance. Using a positive example (i.e., the patient was eventually deceased during the ICU stay) extracted from the test set, we explore how SL affects mortality prediction results. As shown in Table 4, the model originally predicts a 48.23% mortality probability for this patient, which falls below the 50% threshold required for correct prediction. This prediction erroneously suggests that the patient is not at mortality risk. Notably, when we remove the SL word "combative," the predicted probability increases by 11.53%, leading to a correct prediction. No other word removal results in similar improvements. In particular, removing "aroused" reduces the mortality prediction by 21.04%, suggesting that it is an important predictive feature in this case. The results suggest that SL might hinder the performance of an AI model trained on EHR notes in mortality prediction, and removing the SL could be a viable solution.

To address potential concerns that these results are case-specific or coincidental, or there is no difference in the effect of removing SL and random words, we apply the leave-one-out strategy to our testing set globally and compare the average predictive probabilities under three conditions: a) utilizing the original notes, b) removing SL, and c) removing an equivalent number of random non-SL words. We perform random removal 100 times and report their average to further ensure the robustness of the results. Additionally, we focus on samples containing SL to efficiently investigate the impact of SL on prediction results. This is because we evaluate the same trained model in all conditions, and therefore, the predictions for samples without SL would not change with SL removal.

As shown in Table 5, removing SL from the testing set only slightly impacts the precision and F1 score, but it does lead to an increase in the recall rate (0.36%), especially for the recall rate for positive cases (1.07%). Improved recall rates in medical predictions can help models capture more patients at risk, thereby reducing malpractice and moral hazard. In contrast, randomly removing the same number of non-SL words has virtually no impact on the model performance, indicating that SL in medical notes introduces noise rather than informative patterns into AI models. Our findings remain consistent across these multiple tests, supporting Hypothesis 1b and refuting Hypothesis 1a.

| Condition | Precision | Recall | Compared with Recall | Recall for Positive Cases | Compared with Recall for Positive Cases | F1 |
|---|---|---|---|---|---|---|
| Original | 66.64% | 73.46% | - | 57.26% | - | 69.08% |
| SL Removal | 66.52% | 73.82% | 0.36% | 58.33% | 1.07% | 69.14% |
| Random Removal | 66.58% | 73.43% | -0.03% | 57.15% | -0.11% | 69.06% |

**Table 5. Results of Global Leave-one-out Strategy on All SL Samples**

### XAI: Input Reduction Strategy

To further examine the robustness of the leave-one-out strategy results, we use the input reduction strategy to drop the least important word sequentially and observe the changes in the mortality prediction results. Table 6 presents the results, which are consistent with our earlier findings that removing "combative" can help increase the predicted probability, and SL is responsible for the incorrect prediction. In contrast, removing the rest of the sentence can reduce the predicted probability (except for "was," which only leads to a marginal increase of 0.35% in predicted probability). In particular, removing "aroused" leads to a 17.61% decrease in the predicted probability, underscoring its importance. These results further support our findings that SL negatively affects AI models on clinical prediction, leading to inferior performance.

| Input | Probability | Improvement | Change Direction |
|---|---|---|---|
| Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. When he aroused he was very combative. | 48.23% | - | - |
| Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. When he aroused he was very ~~**combative**~~. | 59.56% | 11.33% | ↑ |
| Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. When he aroused he ~~**was**~~ very. | 59.91% | 11.68% | ↑ |
| Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. When he aroused he ~~**very**~~. | 59.02% | 10.79% | ↓ |
| Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. When ~~**he**~~ aroused he. | 53.78% | 5.55% | ↓ |
| Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. When aroused ~~**he**~~. | 49.30% | 1.07% | ↓ |
| Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. ~~**When**~~ aroused. | 43.90% | -4.33% | ↓ |
| Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. ~~**aroused**~~. | 30.62% | -17.61% | ↓ |

**Table 6. Results of the Input Reduction Strategy**

### XAI: Adversarial Perturbations

| Example | Input | Probability | Probability Change |
|---|---|---|---|
| Original | Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. When he aroused he was very combative. | 48.23% | - |
| White + SL | Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. When he aroused he was very combative. ***Pt is a Caucasian***. | 48.18% | -0.05% |
| Black + SL | Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. When he aroused he was very combative. ***Pt is an African American***. | 32.36% | -15.87% |
| Non-SL | Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. When he aroused he was very [MASK]. | 59.76% | 11.53% |
| White + Non-SL | Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. When he aroused he was very [MASK]. ***Pt is a Caucasian***. | 61.07% | 12.84% |
| Black + Non-SL | Transferred from outside hospital via [**Location (un) **]. Pt. apparently fell at home down [**5-4**] steps and had a +loc. When he aroused he was very [MASK]. ***Pt is an African American***. | 63.62% | 15.39% |

**Table 7. Results of Add-Sentence Adversarial Perturbations**

Next, we employ add-sentence adversarial perturbations, by inserting an adversarial sentence into the original input and analyzing how the model responds, to gain insights into the model's racial disparity caused by the use of SL. As shown in Table 7, the model's prediction remains nearly the same (-0.05%) when adding the "Caucasian" sentence (i.e., "Pt is a Caucasian.") but decreases substantially when adding the "African American" sentence (-15.87%). Interestingly, without SL, this racial gap can be significantly narrowed, with the addition of the "Caucasian" sentence and the addition of the "African American" sentence leading to a 12.84% increase and a 15.39% increase in the predicted mortality, respectively. It is striking that, in this example, simply removing SL words not only narrows the racial gap but also makes the predicted probability for the sentence containing "African American" even higher than for the sentence containing "Caucasian." This provides initial evidence to support H2.

As discussed earlier, it is a common concern that the results derived from local XAI methods may only be valid for the selected example. To gain a global understanding of the adversarial perturbations results, we add the adversarial sentence to all samples of the testing set to systematically examine the impact of SL on racial disparity in model performance. As illustrated in Table 8, adding the "Caucasian" sentence results in minimal changes (0% in F1, -0.10% in recall rate, and -0.34% in recall rate for positive cases), while adding the "African American" sentence causes a decline in F1 (-0.09%), as well as a noticeable drop in the recall rate and the recall rate for positive cases, respectively (1.58% and 5.68%). Furthermore, we introduce an extra SL phrase[4] indicating a history of drug abuse to investigate how SL interacts with racial information. This further lowers the recall rate, with black patients experiencing a greater drop in positive case recall than Caucasians (9.76% compared with 4.96%). The results are largely consistent with our previous findings. Echoing our discussion, a decreased recall rate is detrimental to medical predictions, as it increases the risk of malpractice and moral hazard for patients at high mortality risk who are overlooked.

[4] The added phrase is "*and has a history of drug abuse*" which attempts to express that the individual has struggled with substance use in the past. To avoid perpetuating stigma, it is important to use more neutral and person-centered language such as "*and has a history of substance use disorder*."

| Condition | Added Sentence | Precision | Recall | Compared with Recall | Recall for Positive Cases | Compared with Recall for Positive Cases | F1 |
|---|---|---|---|---|---|---|---|
| Original | None | 63.44% | 66.67% | - | 46.31% | - | 64.68% |
| White | *Pt is a Caucasian.* | 63.47% | 66.57% | -0.10% | 45.97% | -0.34% | 64.68% |
| Black | *Pt is an African American.* | 64.14% | 65.09% | -1.58% | 40.63% | -5.68% | 64.59% |
| SL | *Pt has a history of drug abuse.* | 63.64% | 65.57% | -1.10% | 42.68% | -3.63% | 64.47% |
| White + SL | *Pt is a Caucasian and has a history of drug abuse.* | 63.48% | 65.04% | -1.63% | 41.35% | -4.96% | 64.17% |
| Black + SL | *Pt is an African American and has a history of drug abuse.* | 64.09% | 63.65% | -3.02% | 36.55% | -9.76% | 63.86% |

**Table 8. Results of Global Add Sentence Adversarial Perturbations**

## *Experimental Results: XAI on the Training Set*

By training a Transformer model across 12 different settings and comparing the relative performance, as shown in Table 9, we find that removing SL from the full dataset leads to a 1.68% increase in the F1 score (Set 1 and 2). The Transformer model performs better on white patients than on black patients using the original notes (for both full data and 24-hour data), and we observe a more salient racial gap on the 24-hour subset (Set 9 and 10), where the model achieves a 2.97% higher F1 score for white patients than black patients. However, when SL is removed in Set 11 and 12, the racial gap almost vanishes (0.05%).

These results indicate that racial disparities in EHR notes can propagate to AI models, with SL as a key factor contributing to the racial disparities. Removing SL could significantly improve AI performance and fairness. Therefore, our H1b and H2 are supported on the training set as well. As SL is pervasive in public and private EHR notes, our findings are critical to AI practitioners in clinical settings. Despite the limited attention paid to SL in previous model development, existing development and publicly accessible pre-trained models may retain racial disparities caused by SL. Immediate actions are needed to prevent racial disparities from being propagated through hard-to-interpret deep learning models.

| Set ID | Dataset | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| 1 | Full (Original) | 81.57% | 63.44% | 66.67% | 64.68% |
| 2 | Full (SL Removed) | 86.37% | 69.56% | 64.44% | 66.36% |
| | Improvement from SL Removal | 4.80% | 6.12% | -2.23% | 1.68% |
| 3 | Full - White patients (Original) | 81.85% | 62.77% | 66.30% | 64.08% |
| 4 | Full - Black patients (Original) | 82.90% | 61.56% | 67.80% | 63.39% |
| | Racial Difference (Black-White) | 1.05% | -1.21% | 1.50% | -0.69% |
| 5 | Full - White patients (SL Removed) | 86.75% | 68.79% | 64.12% | 65.91% |
| 6 | Full - Black patients (SL Removed) | 88.31% | 67.12% | 65.34% | 66.15% |
| | Racial Difference (Black-White) | 1.56% | -1.67% | 1.22% | 0.24% |
| 7 | 24-hour (Original) | 89.54% | 67.67% | 62.13% | 64.16% |
| 8 | 24-hour (SL Removed) | 90.53% | 71.56% | 61.16% | 64.13% |
| | Improvement from SL Removal | 0.99% | 3.89% | -0.97% | -0.03% |
| 9 | 24-hour - White patients (Original) | 89.83% | 67.14% | 62.25% | 64.11% |
| 10 | 24-hour - Black patients (Original) | 92.17% | 61.56% | 60.76% | 61.14% |
| | Racial Difference (Black-White) | 2.34% | -5.58% | -1.49% | -2.97% |
| 11 | 24-hour - White patients (SL Removed) | 90.80% | 70.69% | 60.92% | 63.78% |
| 12 | 24-hour - Black patients (SL Removed) | 93.67% | 67.31% | 61.56% | 63.73% |
| | Racial Difference (Black-White) | 2.87% | -3.38% | 0.64% | -0.05% |

**Table 9. Results of Removing SL from the Training Set**

### *Robustness Checks*

To assess the robustness of our results with pre-trained large language models, we fine-tune Clinical BERT using the same dataset as the Transformer. As Table 10 shows, removing SL during fine-tuning significantly improves Clinical BERT's accuracy in mortality risk prediction. In addition, to address concerns that our findings may rely on imbalanced data, we create a balanced dataset with equal samples across races. As Table 10 indicates, removing SL continues to enhance model performance even in this balanced context.

| Model | Dataset | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| Transformer | Full - Balanced (Original) | 90.34% | 78.06% | 66.10% | 69.89% |
| | Full - Balanced (SL Removed) | 90.40% | 78.25% | 66.43% | 70.20% |
| | Improvement from SL Removal | 0.06% | 0.19% | 0.33% | 0.31% |
| Clinical BERT | Full (Original) | 88.74% | 85.50% | 59.90% | 63.33% |
| | Full (SL Removed) | 88.96% | 79.96% | 64.52% | 68.46% |
| | Improvement from SL Removal | 0.22% | -5.54% | 4.62% | 5.13% |

**Table 10. Results of Clinical BERT**

To better understand the role of race in model predictions, we included it as a structural feature, concatenated with the Transformer layer's output. This step is taken to allow race to effectively contribute to the final prediction without disrupting the learned textual features. As shown in Table 11, this inclusion improves the model's overall performance. However, it also exacerbates racial disparities when SL is removed. This outcome aligns with existing fairness literature, which advises against using sensitive attributes like race in predictive models to prevent unintended discrimination (Chen et al. 2021).

| Dataset | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Full (Original) + Race | 83.33% | 64.86% | 65.95% | 65.37% |
| Full (SL Removed) + Race | 86.90% | 71.21% | 66.72% | 68.53% |
| Improvement from SL Removal | 3.57% | 6.35% | 0.77% | 3.16% |
| Full - Black patients (Original) + Race | 85.47% | 62.65% | 65.10% | 63.69% |
| Full - White patients (Original) + Race | 83.89% | 64.12% | 64.79% | 64.44% |
| Racial Difference (Black-White) | 1.58% | -1.47% | 0.31% | -0.75% |
| Full - Black patients (SL Removed) + Race | 90.09% | 72.21% | 63.09% | 66.02% |
| Full - White patients (SL Removed) + Race | 87.25% | 70.28% | 65.31% | 67.24% |
| Racial Difference (Black-White) | 2.84% | 1.93% | -2.22% | -1.22% |

**Table 11. Results of Transformer with Race as An Additional Feature**

## Conclusion

In this study, we examined how bias manifested as stigmatizing language in EHR notes affects AI performance and fairness. Our results suggest that SL can hinder an AI model from performing mortality prediction, primarily due to the noisy patterns in SL usage. Furthermore, we found that SL engenders racial disadvantages for black patients during the development of AI models, thereby leading to racial disparities and compromising AI fairness. The differential performance of AI-based mortality predictions has direct implications for patient care, potentially perpetuating biased healthcare decision-making. When employed in treatment planning or resource allocation, these racial biases may compromise the quality of care for specific racial groups, notably black patients, thereby exacerbating existing health disparities. Our findings are important for both the training of AI models using SL EHR notes and the implementation of trained models on SL EHR notes. First, as the racial disparity associated with SL can propagate to AI models, the mitigation of SL's impact can be achieved by eliminating it from the training data, consequently enhancing AI performance and fairness. Second, even if an AI has been trained on SL EHR notes, the racial gap can still be minimized by excluding SL from the testing data. To contextualize, our focus on linguistic markers of racial bias is a part of the complex tapestry of algorithmic bias in healthcare. We acknowledge that factors like data quality and clinical settings are also impactful and represent key areas for future exploration.

Recent advancements in large language models, such as ChatGPT, represent a major engineering achievement in applying state-of-the-art models using high-quality data. The core asset of ChatGPT is undeniably the training dataset (Perrigo 2023), which has emerged as a contentious issue between OpenAI and Google (Hollister 2023). Given the lack of new algorithmic breakthroughs in the deep learning models for processing natural language since the advent of Transformers in 2017 (Vaswani et al. 2017), the emphasis on data quantity and quality has become indispensable l for deep learning model development, encompassing healthcare AI as well. However, various healthcare-specific factors warrant consideration when developing clinical AI models. This is especially true for medical texts such as EHR notes, where factors influencing AI performance and fairness remain understudied.

Our study endeavors to bridge the gap by identifying a striking phenomenon: racial disparities in clinicians' SL in EHR notes can propagate to AI models, resulting in both low performance and racial unfairness in AI. This finding presents numerous opportunities for future work and contributes to both medical AI development and AI fairness more broadly. Although Information Systems researchers have been developing machine learning models utilizing EHR or related clinical data (Samtani et al. 2023; Yu et al. 2021; Bardhan et al. 2020), the negative impact of SL has been largely overlooked in most clinical predictive model development. AI fairness is a popular research topic in general (Chen et al. 2021; Buolamwini and Gebru 2018; Bolukbasi et al. 2016), and it is rising in business research (Kallus et al. 2022; Yue et al. 2022; Bjarnadóttir and Anderson 2020). We adopt a distinctive perspective to examine how human biases (SL) can negatively impact a broad scope of patients through its propagation in AI model development. Finally, research on human biases in clinical practice has a long history in Information Systems (Ganju et al. 2022; Lin et al. 2019). This study connects this stream of literature to the emerging topic of AI development. Practically, our study offers a meaningful mitigation strategy for reducing the harms of SL. Removing clinician-written SL proves effective in curbing racial disparities in AI prediction, lending actionable insights to policymakers and industry practitioners to promote responsible AI development.

## References

Agarwal, R., Bjarnadottir, M., Rhue, L., Dugas, M., Crowley, K., Clark, J., and Gao, G. 2022. "Addressing Algorithmic Bias and the Perpetuation of Health Inequities: An AI Bias Aware Framework," *Health Policy and Technology*, p. 100702.

Association of Diabetes Care and Education Specialists. 2021. "Language Guidance for Diabetes-Related Research, Education and Publications," Retrieved Apr 2, 2023, from https://www.diabeteseducator.org/docs/default-source/practice/educator-tools/HCP-diabetes-language-guidance.pdf?sfvrsn=22.

Axt, J. R., and Lai, C. K. 2019. "Reducing Discrimination: A Bias versus Noise Perspective," *Journal of Personality and Social Psychology* (117:1), pp. 26.

Bailey, Z. D., Feldman, J. M., and Bassett, M. T. 2021. "How Structural Racism Works—Racist Policies as a Root Cause of US Racial Health Inequities," *New England Journal of Medicine* (384:8), pp. 768-773.

Bardhan, I., Chen, H., and Karahanna, E. 2020. "Connecting Systems, Data, and People: A Multidisciplinary Research Roadmap for Chronic Disease Management," *MIS Quarterly* (44:1), pp. 185-200.

Beck, A. S., Svirsky, L., and Howard, D. 2022. "'First Do No Harm': Physician Discretion, Racial Disparities and Opioid Treatment Agreements," *Journal of Medical Ethics* (48:10), pp. 753-758.

Bjarnadóttir, M. V., and Anderson, D. 2020. "Machine Learning in Healthcare: Fairness, Issues, and Challenges," In *Pushing the Boundaries: Frontiers in Impactful OR/OM Research*, pp. 64-83. INFORMS.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., and Kalai, A. T. 2016. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," *Advances in Neural Information Processing Systems* (29).

Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R., and Cavalli-Sforza, L. L. 1994. "High Resolution of Human Evolutionary Trees with Polymorphic Microsatellites," *Nature* (368:6470), pp. 455-457.

Bowcock, A. M., Kidd, J. R., Mountain, J. L., Hebert, J. M., Carotenuto, L., Kidd, K. K., and Cavalli-Sforza, L. L. 1991. "Drift, Admixture, and Selection in Human Evolution: A Study with DNA Polymorphisms," *Proceedings of the National Academy of Sciences* (88:3), pp. 839-843.

Buolamwini, J., and Gebru, T. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," In Conference on Fairness, Accountability and Transparency, pp. 77-91. PMLR.

Calafell, F., Shuster, A., Speed, W. C., Kidd, J. R., and Kidd, K. K. 1998. "Short Tandem Repeat Polymorphism Evolution in Humans," *European Journal of Human Genetics* (6:1).

Chapman, E. N., Kaatz, A., and Carnes, M. 2013. "Physicians and Implicit Bias: How Doctors May Unwittingly Perpetuate Health Care Disparities," *Journal of General Internal Medicine* (28:11), pp. 1504-1510.

Char, D. S., Abràmoff, M. D., and Feudtner, C. 2020. "Identifying Ethical Considerations for Machine Learning Healthcare Applications," *The American Journal of Bioethics* (20:11), pp. 7-17.

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., and Ghassemi, M. 2021. "Ethical Machine Learning in Healthcare," *Annual Review of Biomedical Data Science* (4), pp. 123-144.

Chen, I. Y., Szolovits, P., and Ghassemi, M. 2019. "Can AI Help Reduce Disparities in General Medical and Mental Health Care?," *AMA Journal of Ethics* (21:2), pp. 167-179.

Dresser, R. 1992. "Wanted: Single, White Male for Medical Research," *The Hastings Center Report* (22:1), pp. 24-29.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. 2017. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks," *Nature* (542:7639), pp. 115-118.

Feng, S., Wallace, E., Grissom II, A., Iyyer, M., Rodriguez, P., and Boyd-Graber, J. 2018. "Pathologies of Neural Models Make Interpretations Difficult," In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3719-3728.

Fernandez, C., Provost, F., and Han, X. 2022. "Explaining Data-Driven Decisions Made by AI Systems: The Counter Factual Approach," *MIS Quarterly* (46:1), pp. 105-122.

Huang, K., Altosaar, J., and Ranganath, R. 2019. "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission," *arXiv preprint arXiv:1904.05342.*

Jia, R., and Liang, P. 2017. "Adversarial Examples for Evaluating Reading Comprehension Systems," In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2021-2031.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., ... & Mark, R. G. 2016. "MIMIC-III, a Freely Accessible Critical Care Database," *Scientific Data* (3:1), pp. 1-9.

Kallus, N., Mao, X., & Zhou, A. 2022. "Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination," *Management Science* (68:3), pp. 1959-1981.

Kelly, J. F., Wakeman, S. E., & Saitz, R. 2015. "Stop Talking 'Dirty': Clinicians, Language, and Quality of Care for the Leading Cause of Preventable Death in the United States," *The American Journal of Medicine* (128:1), pp. 8-9.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... & Tseng, V. 2023. "Performance of ChatGPT on USMLE: Potential for AI-assisted Medical Education Using Large Language Models," *PLOS Digital Health* (2:2), e0000198.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. 2020. "BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining," *Bioinformatics* (36:4), pp. 1234-1240.

Li, J., Monroe, W., & Jurafsky, D. 2016. "Understanding Neural Networks Through Representation Erasure," *arXiv preprint arXiv:1612.08220.*

Li, Y., Rao, S., Solares, J. R. A., Hassaine, A., Ramakrishnan, R., Canoy, D., ... & Salimi-Khorshidi, G. 2020. "BEHRT: Transformer for Electronic Health Records," *Scientific Reports* (10:1), pp. 1-12.

Lin, Y. K., Lin, M., & Chen, H. 2019. "Do Electronic Health Records Affect Quality of Care? Evidence from the HITECH Act," *Information Systems Research* (30:1), pp. 306-318.

Lou, B., and Wu, L. 2021. "AI on Drugs: Can Artificial Intelligence Accelerate Drug Development? Evidence from a Large-Scale Examination of Bio-Pharma Firms," *MIS Quarterly* (45:3), pp. 1451-1482.

Martens, D., and Provost, F. 2014. "Explaining Data-Driven Document Classifications," *MIS Quarterly* (38:1), pp. 73-100.

Martin, A. E., D'Agostino, J. A., Passarella, M., & Lorch, S. A. 2016. "Racial Differences in Parental Satisfaction with Neonatal Intensive Care Unit Nursing Care," *Journal of Perinatology* (36:11), pp. 1001-1007.

McGarvey, L. P., John, M., Anderson, J. A., Zvarich, M., & Wise, R. A. 2007. "Ascertainment of Cause-Specific Mortality in COPD: Operations of the TORCH Clinical Endpoint Committee," *Thorax* (62:5), pp. 411-415.

Merryweather-Clarke, A. T., Pointon, J. J., Jouanolle, A. M., Rochette, J., & Robson, K. J. 2000. "Geography of HFE C282Y and H63D Mutations," *Genetic Testing* (4:2), pp. 183-198.

Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. 2016. "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records," *Scientific Reports* (6:1), pp. 1-10.

National Institute on Drug Abuse. 2021. "Words Matter–Terms to Use and Avoid When Talking About Addiction."

Negro-Calduch, E., Azzopardi-Muscat, N., Krishnamurthy, R. S., & Novillo-Ortiz, D. 2021. "Technological Progress in Electronic Health Record System Optimization: Systematic Review of Systematic Literature Reviews," *International Journal of Medical Informatics* (152), 104507.

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), pp. 447-453.

Parikh, R. B., Teeple, S., & Navathe, A. S. 2019. "Addressing Bias in Artificial Intelligence in Health Care," *JAMA* (322:24), pp. 2377-2378.

Park, J., Saha, S., Chee, B., Taylor, J., & Beach, M. C. 2021. "Physician Use of Stigmatizing Language in Patient Medical Records," *JAMA Network Open* (4:7), e2117052-e2117052.

Patel, S. B., and Lam, K. 2023. "ChatGPT: The Future of Discharge Summaries?," *The Lancet Digital Health* (5:3), pp. e107-e108.

Perrigo, B. 2023. "OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic," *Time* (Jan 18), https://time.com/6247678/openai-chatgpt-kenya-workers/.

Posner, T., & Fei-Fei, L. 2020. "AI Will Change the World, So It's Time to Change AI," *Nature* (588:7837), S118-S118.

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. 2022. "AI in Health and Medicine," *Nature Medicine* (28:1), pp. 31-38.

Risch, N., Burchard, E., Ziv, E., & Tang, H. 2002. "Categorization of Humans in Biomedical Research: Genes, Race and Disease," *Genome Biology* (3:7), pp. 1-12.

Rudin, C. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence* (1:5), pp. 206-215.

Samtani, S., Zhu, H., Padmanabhan, B., Chai, Y., Chen, H., & Nunamaker Jr, J. F. 2023. "Deep Learning for Information Systems Research," *Journal of Management Information Systems* (40:1), pp. 271-301.

Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., & Okruszek, L. 2021. "Detecting Formal Thought Disorder by Deep Contextualized Word Representations," *Psychiatry Research* (304), 114135.

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. 2022. "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence," *NIST Special Publication* (1270), pp. 1-77.

Seinen, T. M., Fridgeirsson, E. A., Ioannou, S., Jeannetot, D., John, L. H., Kors, J. A., ... & Rijnbeek, P. R. 2022. "Use of Unstructured Text in Prognostic Clinical Prediction Models: A Systematic Review," *Journal of the American Medical Informatics Association* (29:7), pp. 1292-1302.

Stephens, J. C., Reich, D. E., Goldstein, D. B., Shin, H. D., Smith, M. W., Carrington, M., ... & Dean, M. 1998. "Dating the Origin of the CCR5-Δ32 AIDS-Resistance Allele by the Coalescence of Haplotypes," *The American Journal of Human Genetics* (62:6), pp. 1507-1515.

Stokel-Walker, C. 2023. "ChatGPT Listed as Author on Research Papers: Many Scientists Disapprove," Nature (Jan 18), https://www.nature.com/articles/d41586-023-00107-z.

Sun, M., Oliwa, T., Peek, M. E., & Tung, E. L. 2022. "Negative Patient Descriptors: Documenting Racial Bias in the Electronic Health Record," *Health Affairs* (41:2), pp. 203-211.

Tamayo-Sarver, J. H., Hinze, S. W., Cydulka, R. K., & Baker, D. W. 2003. "Racial and Ethnic Disparities in Emergency Department Analgesic Prescription," *American Journal of Public Health* (93:12), pp. 2067-2073.

The White House. 2022. "A Blueprint for an AI Bill of Rights," Report, The Office of Science and Technology Policy (OSTP), The Executive Office of the President, Washington, DC.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. 2017. "Attention is All You Need," *Advances in Neural Information Processing Systems* (30).

Vyas, D. A., Eisenstein, L. G., & Jones, D. S. 2020. "Hidden in Plain Sight—Reconsidering the Use of Race Correction in Clinical Algorithms," *New England Journal of Medicine* (383:9), pp. 874-882.

Wang, S., McDermott, M. B., Chauhan, G., Ghassemi, M., Hughes, M. C., & Naumann, T. 2020. "Mimic-Extract: A Data Extraction, Preprocessing, and Representation Pipeline for Mimic-III," In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 222-235.

Wen, Z., Lu, X. H., & Reddy, S. 2020. "MeDAL: Medical Abbreviation Disambiguation Dataset for Natural Language Understanding Pretraining," In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pp. 130-135.

Werder, K., Curtis, A., Reynolds, S., & Satterfield, J. 2022. "Addressing Bias and Stigma in the Language We Use with Persons with Opioid Use Disorder: A Narrative Review," *Journal of the American Psychiatric Nurses Association* (28:1), pp. 9-22.

World Health Organization. 2021. "Ethics and Governance of Artificial Intelligence for Health: WHO Guidance."

Xiao, C., Choi, E., & Sun, J. 2018. "Opportunities and Challenges in Developing Deep Learning Models Using Electronic Health Records Data: A Systematic Review," *Journal of the American Medical Informatics Association* (25:10), pp. 1419-1428.

Yang, D., & Song, J. 2010. "Web Content Information Extraction Approach Based on Removing Noise and Content-Features," In *2010 International Conference on Web Information Systems and Mining* (Vol. 1), pp. 246-249.

Yu, S., Chai, Y., Chen, H., Sherman, S., & Brown, R. 2022. "Wearable Sensor-Based Chronic Condition Severity Assessment: An Adversarial Attention-Based Deep Multisource Multitask Learning Approach," *MIS Quarterly* (46:3), pp. 1355-1394.

Yue, X., Nouiehed, M., & Al Kontar, R. 2022. "Gifair-FL: A Framework for Group and Individual Fairness in Federated Learning," *INFORMS Journal on Data Science*.

Zhang, H., Lu, A. X., Abdalla, M., McDermott, M., & Ghassemi, M. 2020. "Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings," In *Proceedings of the ACM Conference on Health*, Inference, and Learning, pp. 110-120.