

See Me, Believe Me: Causality, Intersectionality, and Interventions Improving the Appearance of Patients

Kenya S. Andrews¹, Mesrob I. Ohannessian², Elena Zheleva²

¹Brown University

²University of Illinois Chicago

kenya_andrews@brown.edu, mesrob@uic.edu, ezheleva@uic.edu

Abstract

In the context of medical records, patients often experience testimonial injustice, where the textual account undermines the validity of their experiences. Past work has demonstrated that intersectionality of demographic features is crucial to *detect* such injustice. We use causal discovery to study the degree to which certain demographic features tied to marginalization (namely age, gender, and race), together lead to specific types of testimonial injustice terms. This offers us an insightful Structural Causal Model (SCM) relating those demographic features to the different ways a patients' reality may be undermined. We then move toward *addressing* such injustice, by very selectively (based on insights from the SCM) editing physicians' notes. For comparison, we contrast these rule-based edits with blanket context-based modifications using an LLM. We assess the impact of these changes on the perception of patients' experiences, using human experts and an LLM. We find that edits, in general, enhance clarity regarding the urgency and causes of patients' conditions. Additionally, for human experts, rule-based modifications show a tendency to shift blame away from patients, and toward more objective external factors. These findings (1) underscore the importance of quantifying sources of injustice in how patients' testimonies are recorded, (2) reveal that making minimal intentional changes accordingly can effect improved patient perception, and (3) call for larger efforts to assess health outcomes under such more just representation.

Introduction

Patients seeking medical treatment are not only vulnerable but are simultaneously dependent upon whomever is giving them care at the time. This fact is particularly concerning for those who are not believed or appropriately perceived because of prejudices about them, an experience known as testimonial injustice (Fricker 2019). It has been proven that clinicians are more likely to ignore and make light of the concerns of Black and female patients than White and male patients (Beach et al. 2021; L. 2020). Yet, very little work has been done to show the nuances of the experiences for younger Black females, younger Black males, senior White males, senior Latina females, and those of other intersections. *Intersectionality* informs on how people experience

the world due to their attributes, such as demographic features (e.g., race, age, gender) (Crenshaw 2013; Marques 2018). We hypothesize that a key measurable contributor to these nuanced experiences in healthcare might be testimonial injustice.

In such critical circumstances, the trust that patients place in clinical staff to be heard and understood becomes ever more pertinent. However, this trust can be compromised by deeply rooted biases that may permeate the staff's language (i.e., word choices) in patient records (i.e., physicians' notes). The choice of words in clinical notes often reflects biases, which can skew understanding and lead to disparate patient outcomes for readers of the notes (including Machine Learning (ML) informed tools), leading to injustices. Further, as Large Language Models (LLMs) are more heavily incorporated into clinical settings, it becomes imperative that we scrutinize the linguistic choices that will inform these models and shape their perceptions of patients. Both clinicians and LLMs risk absorbing these biases, potentially influencing their clinical judgments, which can compromise accuracy and effectiveness in patient outcomes.

Though it is known that intersectionality of features is necessary to uncover certain cases of testimonial injustice (Andrews, Shah, and Cheng 2023), we lack a comprehensive understanding of *which* features causally lead to testimonial injustice, *how* they do so, and their levels of impact. This gap in knowledge prevents exploring if *intentional interventions* can lead to changes within patient experiences and outcomes.

Here, we aim to address this by using causal discovery to not only show how attributes of a person can come together — i.e., intersectionality — to contribute to them experiencing testimonial injustice, but also to understand the specific ways injustice is revealed, the intensity with which attributes are contributors, test strategic interventional methods along the intersectional paths of injustice, and understand the impact of realigning descriptions of conditions for patients who fall along such paths.

Thus, the research questions we explore here are: (RQ1) **Can we identify how individual demographic features influence language in medical settings, leading collectively to testimonial injustice?** And, once we do, (RQ2) **can we quantify the degree to which these interactions contribute to this experience to make intentional inter-**

ventions?

The contributions of this paper are: (1) using causal discovery methods and the resulting Structural Causal Model (SCM) to identify and describe how demographic features of concern (i.e., race, age, and gender) contribute to a patient experiencing testimonial injustice in physicians' notes, (2) discussing the revealed interactions and quantifying the extent of their influence on testimonial injustice, (3) investigating a use-case study for these insights through intentional intervention methods (i.e., modifications to physicians' notes) to address injustices in physicians' notes based on the likelihood of such experiences as indicated by the SCM and context-specific adjustments, and (4) assessing such interventions by comparing the downstream effects of unmodified vs. modified notes, specifically in how patients and their conditions are perceived both by human experts and LLMs.

Related Works

An empirical study conducted by Andrews, Shah, and Cheng (2023) showed that there is differential treatment between subgroups experiencing testimonial injustice in medical settings, noting that these nuances can only be revealed through the lens of intersectionality. However, the causal nature of this intersectionality was left open, namely, regarding which attributes contribute to testimonial injustice and the degree to which these attributes influence someone experiencing such injustice. Thus, this work shifts the focus to how one's demographic features—namely race, gender, and age—can contribute to them experiencing testimonial injustice, using causal discovery. To the best of our knowledge, ours is the first use of causal discovery to understand these nuanced experiences of testimonial injustice in medical settings through an intersectional lens.

Amemiya et al. (2023) developed a framework which tells how people attribute inequality to structural causes, namely instances in society that systematically advantage some and marginalize others. This is rampant in the medical field (Hall et al. 2015), with various instances of preferential treatment to those with specific insurance policies, race, income, etc. (Stepanikova and Cook 2008; Yearby, Clark, and Figueroa 2022). They showed that when 2 groups have the same abilities, but are systematically treated differently, there is a case for using between-group comparisons to build causal models. In medical settings, abilities are similar—however, features can vary vastly, i.e., among those who do/do not have insurance, gender, race, age, education level, etc. Though this work focuses on observably inferable features (i.e., race, gender, and age), we acknowledge here that there are many demographic features—both externally observable and latent—that can also contribute to someone experiencing testimonial injustice, particularly in medical settings. Further, the degree to which the presence of testimonially unjust terms affects care can vary due to various factors like experience. In (Lane and Rae 1983), registered nurses (RNs) were shown to be more susceptible to stigmatizing and judging patients when such words were present, in contrast to nursing students. The presence of such words led both students and RNs to assumptions of *psychological* pain and labeling the patient as *difficult*. For students, they

reported less desire to work with such patients. These perceptions are likely to lead to poor outcomes for patients. Inspired by these studies, the perspective we adopt here is that, as long as there are even some clinicians whose perceptions are going to negatively affect patients, it is worthwhile to intervene and prevent it from happening.

Recent studies indicate that LLMs can significantly enhance diagnostic accuracy, nearly doubling clinicians' performance in difficult cases from the U.S. Medical Licensing Examination Step 1 Sample Exam (USMLE; Lee, Bubeck, and Petro 2023). This suggests that even in complex cases, LLMs are strong evaluators of a patient's condition. Consequently, there has been an increase in doctors using publicly available LLMs to assist with clinical decision-making (Gliadkovskaya 2024). Further, researchers have shown that even though demographic features, specifically race, are often absent when provided to LLMs, the models are still able to infer them due to the commonality of unjust word choices in records of racial minorities (Adam et al. 2022). This has led to the perpetuation of injustices in clinical treatment decisions. The concern then shifts from note readability to understanding what improvements are necessary to help LLMs interpret physicians' notes despite human bias and perceive patients more accurately.

Numerous strategies have been proposed for editing text to enhance clarity and standardization. One common approach involves neutralizing text based on individual characteristics (Pryzant et al. 2020). However, blindly neutralizing text is not always ideal, as it can diminish personal experiences rather than improve perception (Boudana 2016; Gonen and Goldberg 2019). In many cases, it can diminish the nuances of an experience or erase culturally significant aspects of an experience. Therefore, this work prioritizes building a nuanced understanding of textual influence through causal modeling, and selectively and judiciously using this insight for potential interventions to improve how patients appear to both human experts and LLMs.

Data

MIMIC-III

We use the MIMIC-III dataset as referenced in the Data Availability Statement. These records contain features of interest to the experiments conducted here, including: ethnicity/race, gender, age, patient id, diagnosis, physicians' notes, etc. (e.g., Black, female, 47, 5432, Bronchitis, "patient claims to be experiencing...").

Pre-processing We identify race, gender, and age as features present that can be *inferred* by people, based on visible observations. The proportion of racial groups represented in the dataset are highly imbalanced (see Table 1), which is likely due to the region the hospital is located in. The MIMIC-III feature 'ethnicity' often included both race and region of origin (e.g., 'Asian – VIETNAMESE'). For our analysis, we focused on race rather than region, recoding entries such as 'Asian – VIETNAMESE' simply as 'Asian,' following the approach of Andrews, Shah, and Cheng (2023). We removed ethnicities that were listed as "unknown/not specified", "multi-race ethnicity", "other",

“unable to obtain”, and “patient declined to answer” since they cannot be clearly associated with any race.

To address the presence of multiple records for patients, we combine the patients’ records based on their patient id, gender, race, and diagnosis (e.g., 2213, male, Latino, Pancreatic Cancer). We do not combine records based on age since, in a single year, many of the patients returned for multiple visits to the ICU— many for the same condition/diagnosis. When we run an analysis on the physicians’ notes to quantify terms that are testimonially unjust for each category of linguistic term, we aggregate counts over records for each patient and divide by the number of their records, thus obtaining the *(average) number of unjust terms per record*. This allows us to ensure that no patient is more heavily weighted than another, depending on the duration of their stay or the number of doctor visits. After pre-processing, we maintain 41,886 unique patients.

Race	Gender	Age	Count
Asian	Female	Senior	212
Asian	Female	Adult	198
Asian	Female	Child	102
Asian	Male	Senior	304
Asian	Male	Adult	267
Asian	Male	Child	119
Black	Female	Senior	945
Black	Female	Adult	1095
Black	Female	Child	482
Black	Male	Senior	776
Black	Male	Adult	875
Black	Male	Child	390
Latino	Female	Senior	81
Latino	Female	Adult	56
Latino	Female	Child	27
Latino	Male	Senior	87
Latino	Male	Adult	109
Latino	Male	Child	45
White	Female	Senior	6076
White	Female	Adult	6452
White	Female	Child	2871
White	Male	Senior	8106
White	Male	Adult	8496
White	Male	Child	3715

Table 1: Counts of patients by race, age, and gender

Testimonial Injustice Lexicon

To assess testimonial injustice in physicians’ notes, we focus on 4 categories of unjust terms: **evidential**, **judgmental**, **negative**, and **stigmatizing**. **Evidential** terms report a claim without asserting its truth (e.g., “complains”, “says”), making the patient’s experience easier to dismiss (Beach et al. 2021). **Judgmental** terms convey physician skepticism (e.g., “apparently”, “claims”). **Negative** terms, often associated with racial and ethnic disparities (Sun et al. 2022), signal denial or rejection (e.g., “combative”, “exaggerate”) using the lexicon from Zhang (2022). **Stigmatizing** terms label patients via stereotypes (e.g., “user”, “faking”) and may influence treatment, transmit bias, and alienate patients (Himmelstein, Bates, and Zhou 2022; Link and Phelan 2001). Our lexicon includes conditions disproportionately affecting racial minorities, such as diabetes, substance use disorder,

and chronic pain. The full based lexicon is provided in the Appendix¹.

Causal Discovery

To describe how inferrable demographic features (i.e., race, age, gender) can collectively lead to a patient experiencing testimonial injustice (RQ1), we perform causal discovery to build a Structural Causal Model (SCM). For this, we rely on the Fast Causal Inference (FCI) algorithm, which we find most suited to the present context as explained below. Before conducting our analysis, however, we first outline the assumptions that contextualize it.

Do you see me? We binarize our exogenous variables (i.e., gender, race, and age) according to whether an individual belongs to a marginalized group, as defined by prior empirical and sociotechnical literature.

Experiments from Beach et al. (2021) show that those who are women and/or those who are Black are more likely to experience testimonial injustice compared to their White or male counterparts. Other studies have demonstrated that patients who are Black or Latino are more likely to encounter testimonial injustice in medical settings (Howell 2018). Andrews, Shah, and Cheng (2023) show that the experience is much more nuanced when looking at race and gender, asserting from their experiments that Black men, Black women, Latino men, Latino women, and in some instances White women patients are more likely to experience testimonial injustice. Asian patients in ICU settings have experiences closer to that of their White counterparts and even better in some cases, (Zhang et al. 2020) — so we do not consider them in this particular context to be a part of the marginalized races. Studies have also shown that ageism is a strong contributor to lack of proper healthcare amongst senior adults (Ben-Harush et al. 2017), but also for children (Goyal et al. 2015). In the United States, how they define the age at which minors can independently consent to medical care varies. Many states allow minors to consent to certain types of care, such as sexual, mental health, or substance-related treatment, starting around age 14. Some states also permit minors as young as 14 to consent to general medical care without parental involvement, although several require minors to be 16 or older for emergency care consent without an adult present. Given these variations, we chose age 15 to capture a developmental stage where many states grant increased autonomy for medical decision-making, yet patients may still receive pediatric care.

Combining the results of these works, the demographic features that we adopt are:

- **Gender:** `is_marginalized_gender` = 1 if a person is female,
- **Race:** `is_marginalized_race` = 1 if a person is Black or Latino, and
- **Age:** `is_marginalized_age` = 1 if a person is a child (age ≤ 15) or a senior (age ≥ 65).

To study the importance of intersectionality to nuanced experiences that cannot be observed otherwise, our experiments vary the granularity of these features. In addition to

¹Appendix can be found at: <https://arxiv.org/pdf/2410.01227>

the *fine* features above, we also introduce a single *coarse* feature that collapses the above to `is_marginalized = 1`, if a person is Black male child, Black female child, adult Black man, adult Black woman, senior Black man, senior Black woman, Latino male child, Latino female child, adult Latino man, adult Latino woman, senior Latino man, senior Latino woman, White female child, adult White woman, or senior White woman, as these groups are likely to experience testimonial injustice as shown in prior work.

Do you believe me? The degree to which someone can experience testimonial injustice can vary from instance to instance (e.g., education of the listener, temperament of the listener). However, we simplify the complexity of this problem by employing a single binary outcome indicating this: **Injustice:** `is_testinj = 1` if any of the categories of terms is present in “high” numbers. This binarization reflects the presence or absence of testimonial injustice in a given instance and does not imply that all experiences are equivalent in severity.

Methods

Lexicon Lookup Since word choice reveals attitudes one may have about a subject (Von Hippel, Sakaquaptewa, and Vargas 2008), we analyze the word choices of physicians in their notes to quantify patient experiences with testimonial injustice. We combine the testimonially unjust terms, into a lexicon to be used for exact-matching lookup. We *expand* this lexicon by adding the stem of these words and 5 synonyms associated with each unjust term in its respective category using nltk’s WordNet corpus (Bird, Klein, and Loper 2009). This expansion is necessary since exact-matching is limited in reach — there can be many variations of the same word, improperly used tenses of words, or words which are similar in meaning. This helps to more accurately cover occurrences of testimonially unjust terms.

To address the potential limitations of keyword matching, we also investigated accounting for linguistic context, namely to discount the potential uses of negation or quotation marks. Specifically, we tested for negations preceding testimonially unjust terms using the following list: [“did not”, “don’t”, “will not”, “won’t”, “has not”, “hasn’t”, “are not”, “aren’t”, “have not”, “haven’t”, “do not”, “does not”, “doesn’t”]. We found that out of a total of 9,353,867 testimonially unjust terms in our dataset, only 35,985 (0.38%) have instances of negations that occur prior to them. We also found that only 5.5% of quoted text contains these words. Thus negations prior to testimonially unjust words and their occurrence in quotations are relatively rare and unlikely to significantly affect our causal analysis.

Instances of Injustice For each of the categories of unjust terms (**Evidential**, **Judgemental**, **Negative**, **Stigmatizing**) and for each patient, we calculate the average number of unjust terms per record. Instead of representing these with their absolute numerical values, we discretize them to numbers in the range 0–9, based on which tenth quantile they belong to (i.e., 0–10th, 10th–20th, ..., 90th–100th). Such a discretization acts as a quantization/kernel and makes the causal discovery more robust.

Ideally, our stance is that a patient is experiencing fair and just testimony when there are *no* instances of unjust terms found in their records. This agrees with the perspective of Andrews, Shah, and Cheng (2023). Equivalently, this means testimonial injustice occurs whenever there is presence from *any* of the unjust term categories. However, to generously account for variability in context behind word choice, we loosen the notion of presence to presence with a “high” number of occurrences.

We choose the 90th percentile of unjust terms per record for any single patient as such a “high”, which, based on the above, corresponds to a numerical value of 9. Thus, we can more precisely write: **Injustice:** `is_testinj = 1` if **Evidential**=9 or **Judgemental**=9 or **Negative**=9 or **Stigmatizing**=9.

Experiments and Results

We conducted experiments to inform us of how demographic features may contribute to someone experiencing testimonial injustice, using causal discovery with the FCI algorithm. We constrain Age, Gender, and Race as exogenous variables, i.e., root causes in the graph, as we know that none of the other variables in the graph cause them. We also constrain Injustice as a leaf node, i.e., it must be an outcome of the SCM, because it can only be caused by the presence of unjust terms. We set $\alpha = 0.05$ as the significance level threshold for the conditional independence tests used by the FCI algorithm, see Fig. 1. Our α functions as a decision cut-off controlling edge removal during skeleton discovery and orientation steps. We conducted additional runs at lower and higher α -values: {0.01, 0.1}, see Fig. 8 in the Appendix.

The resulting SCM shows a direct causal edge from both race and age to judgmental terms, indicating that these factors jointly influence the outcome in an intersectional manner. The graph also reveals more subtle instances of intersectionality, mediated through the categories of testimonial injustice terms. Experiencing stigmatizing terms is affected both directly by race and indirectly by age via the mediator of judgmental terms. Similarly, experiencing evidential terms is affected both directly by age and indirectly by gender via negative terms, as well as by race via stigmatizing terms. Gender is the only feature that does not influence each and every testimonially unjust term, with judgmental terms being the only variable not connected to it. Despite the under-representation of marginalized racial and age groups in this dataset (see Table 1), there is a clear indication of injustice along these demographic features across all identified pathways of injustice. Even more so, since all of the features are consistently 1-hop away from experiencing testimonial injustice, we can rationalize that they are important variables in detecting if someone is likely to experience injustice. However, due to having overlapping mediators, we hypothesize that race and age particularly need to work in tandem to achieve this effect, especially when giving rise to judgmental terms.

We can also see that there exists an association between stigmatizing and negative terms that reflects a shared generative mechanism in these clinical notes, without implying a direct causal relationship between them. The same applies

between judgmental and negative terms.

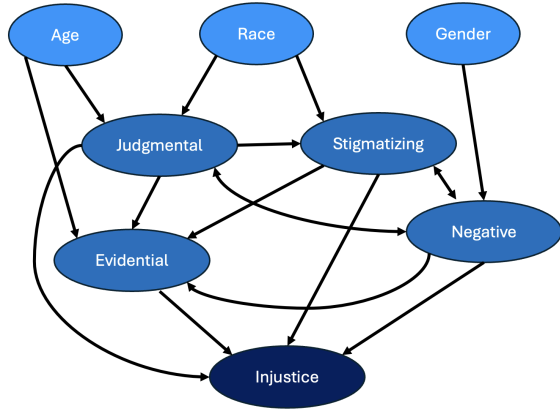


Figure 1: FCI SCMs with $\alpha = 0.05$

Though we focus on the presence of inferrable, observable demographic features, one may wonder how other features, such as insurance type (which can act as a proxy for socioeconomic status even in ICU settings (Monuteaux, Du, and Neuman 2024; Cantiello et al. 2015)), might affect the structure of our SCM. To assess potential confounders that could significantly alter our SCM, we included all relevant factors available in MIMIC III, namely marital status, religion, and insurance type. We marked patients as likely to experience marginalization by marital status if they were single, divorced, widowed, or separated (i.e., `is_marginalized_marital` = 1), given evidence that married patients often experience lower hospital mortality, shorter length of stay, and reduced readmission rates (Metersky, Fine, and Mortensen 2012); those likely to be marginalized by insurance type as those with non-private insurance (i.e., Medicare, Medicaid, other government, or self-pay) (i.e., `is_marginalized_insur` = 1), reflecting structural differences in access to care, as patients with government insurance may experience longer wait times and uninsured patients may have to leave without being seen (Sangal et al. 2023; Tolbert et al. 2024); and likely to experience marginalization through religious affiliation if they identify as Muslim, Jehovah’s Witness, or Hebrew/Arabic background) (i.e., `is_marginalized_religion` = 1), based on prior studies showing that Muslim patients frequently report that providers do not listen to them or question their competence, face challenges related to Islamic dress, and that providers may be reluctant to care for Jehovah’s Witnesses (Eriksson et al. 2023; Gouezec et al. 2016; Keshet and Popper-Giveon 2019). We note here, that though these variables capture sources of marginalization, they are not exhaustive of all potential factors affecting patient experiences. Upon running the causal analysis, we removed any patients that had missing data for these features, which brought our patient count to 22,624. To err on the side of detecting more links from confounders, especially due to additional features and smaller number of data points, we increased the significance level to $\alpha = 0.15$, see Fig. 2. De-

spite this, compared to Fig. 1, only Gender became completely mediated through the new variables, while the direct connections from Age to Judgmental and Evidential, and the direct connection from Race to Stigmatizing persisted. Religion, interestingly, did not show direct influence on testimonial injustice. Marital status could potentially mediate Gender’s influence and insurance type appears to mediate part of Race’s and Gender’s influence. However, the latter’s influence on Injustice is complicated enough that the discovered causal link is direct rather than through any of the terms (despite the fact that Injustice itself is a function of those terms). There are possibly other mediators not measured in MIMIC-III that could affect the SCM. We conclude that although confounders may be present in general, the evidence MIMIC-III provides points to the causal influence of our chosen demographic variables on unjust terms and testimonial injustice to be rather direct.

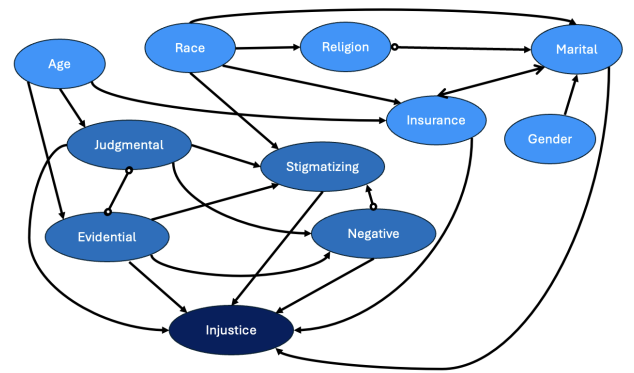


Figure 2: FCI SCMs with additional features in MIMIC-III that could act as mediators $\alpha = 0.15$

Intentional Interventions: A Use Case Study

We conduct a study to understand how intentional interventions along the causal paths of our SCM (Fig 1) can affect patient perception and ability to gauge their condition by human experts or an LLM (RQ2). We employ modifications to physicians notes by word removal, empathy reframing, and using preferred phrasing, with the goal of refining how patients are represented in such human and LLM interpretations. By clarifying marginalized patient experiences through precise linguistic adjustments, we take a step toward justice and better understand the potential of LLMs as collaborative partners in clinical practice. The development of patient-centered care heavily depends on the ability of sociotechnical systems to address these biases and ensure that every patient is truly understood.

To further motivate edits, note that neutral text is the suggested way to discuss patient experiences in medical settings (Harrigian et al. 2023). To address negative and stigmatizing language in text, alternative terms can be used to convey the same information in a more neutral manner. Additionally, to reduce evidential and judgmental terms, empathy can be injected through alternative empathic rephrasing

Words	Alternatives	Term Type
complains	expressed concerns about	Evidential
addict	has a use disorder	Stigmatizing
blatantly		Negative
apparently		Judgmental

Table 2: Example of Alternative Words/Phrases to Testimonially Unjust Terms

or judgmental words can be removed entirely. We collect commonly used stigmatizing, judgmental, evidential, and negative terms in medical settings and preferred alternative phrasing for those terms (Beach et al. 2021; Harrigian et al. 2023; Ho et al. 2022; on Drug Abuse 2021; for Environmental Conflict Resolution 2010). The suggestions from these works, primarily from experts in the fields of substance use disorders and mental health, emphasize the following principles: avoiding overly positive reinterpretations of patient experiences, refraining from demonizing patients based on their condition, using empathy to describe the patient, adopting a neutral tone, and using first-person language such that a person’s identity is detached from their condition.

With these interventions we see interesting insights which are discussed in the Results. This shows how even minimal modifications can lead to measurable changes in downstream outcomes.

Editing Approaches. We use the following editing approaches when making modifications to physicians’ notes based on expert recommendations, as shown in Table 2 (full list in Table 5 of the Appendix):

- **Word Removal** Adverbs modify verbs, adjectives, or other adverbs by conveying manner, conditions, or degree. In physicians’ notes, they frequently add emphasis (e.g., blatantly), which can reinforce testimonial injustice. Accordingly, unjust terms ending in “-ly” were removed, as they were unnecessary and did not affect the meaning of the notes.
- **Empathy Reframing** Empathy, “acknowledging another’s emotional state without experiencing it” (Markakis et al. 1999), is critical in medicine for ensuring patients feel heard. Clinical empathy (Halpern 2003) uses reasoning to communicate effectively, unlike sympathy, which risks physicians assuming patients’ burdens. We applied clinical empathy to physicians’ notes by rephrasing content to detach patients from their conditions, a method effective in law (Lee 2014) and medicine, while avoiding emotional language that could bias clinicians.
- **Preferred Rephrasing and Word Replacement** Synonyms are words or phrases that carry the same or similar meaning as another word. Many alternative wordings to judgmental and stigmatizing words have been proposed in clinical settings (Beach et al. 2021; Harrigian et al. 2023; Ho et al. 2022; on Drug Abuse 2021; for Environmental Conflict Resolution 2010). Here, we simply replace the words and/or phrasing in the text with select proposed terms by substance disorder and mental health experts.

Logistic Regression Our goal is to apply these edits *only when they are needed*, in order to minimize the scale of changes. We do this based on the insights provided by the SCM. However, we need to distinguish instances of positive vs. negative causal influence. To achieve this, we train logistic regression models, 1 for each category of unjust term,

to predict based on the 3 demographics and their conjunction whether they will incur a “high” occurrence (a value of 9). We employ `scikit-learn` logistic regression models (Pedregosa et al. 2011) to evaluate the causal relationships identified by our SCM (recall Fig. 1). The coefficients of the resulting models are given in Table 3.

The first thing to highlight is that the highest magnitude coefficients in each demographic column are indeed at the causal links identified by the SCM. Next, for each category of terms, we would like to build a “simple” rules based on the demographics, to characterize conditions under which we will experience more of those terms. Simplicity in this case is about using a single disjunction and/or conjunction. In some instances, such rules are easy to deduce. For example, **Race** is the most significant factor positively influencing evidential terms, with **Gender** as the second most substantial contributor. Accordingly, we consider patients marginalized by **Race or Gender** for modifications to evidential terms. In other cases, it is less obvious how to distill such simple rules. So, instead, we built truth tables with the thresholded logistic regression predictions and identified the following rules, which disagree in at most 2 places with these truth tables:

- **Evidential:** Race or Gender
- **Judgmental:** Race or Age or Gender (Conjunction)
- **Negative:** Gender or Age
- **Stigmatizing:** Race and (Gender or Age)

Term Type	Race	Age	Gender	Conjunction
evidential terms (89.326% accuracy)	0.553	-0.522	-0.0347	-0.2852
judgmental terms (90.183% accuracy)	0.621	-0.383	0.0193	0.0184
negative terms (87.150% accuracy)	0.445	-0.221	-0.170	-0.1999
stigmatizing terms (89.650% accuracy)	0.638	-0.222	0.0269	0.0645

Table 3: This table presents the logistic regression coefficients from experiments conducted on each term type (i.e., evidential, judgmental, negative, and stigmatizing) predicting whether a patient belongs to a marginalized demographic group (i.e., race, age, gender, or their conjunction, defined as having at least 1 marginalized feature (*is_marginalized* = 1))

Rule-based: Causality-Informed Modifications In the rule-based approach, we perform string replacement using pattern matching to reframe and replace language (i.e., words and phrases) that leads to testimonial injustice. This is done with each of the aforementioned editing approaches: word removal, empathy reframing, and using preferred rephrasing. Our modifications were guided by the insights we gleaned from the SCM and logistic regression models. The causal graph analysis identified how patterns of testimonial injustice appear in the physicians’ notes based on demographic features and their intersections (i.e., intersectionality), while logistic regression applied to the causal graph connections quantifies the strength of these relationships.

To trigger the word removal, empathy reframing, and preferred rephrasing modifications, we find if a patient is likely to be marginalized based on race, gender, age, or any combination of these features, as indicated by the variables: **is_marginalized_race = 1**, **is_marginalized_gender =**

1, is_marginalized_age = 1, and is_marginalized = 1. The specific ways of the modifications are described in the Logistic Regression Section. An example of how a patient’s notes are modified from the unedited version to the rule-based edits is illustrated in Fig. 3, the patient is an adult, Female Latina. Here we see how, for the rule-based approach, “complains” is updated to “expressed concerns about”.

Context-based: LLM Modifications In our context-based approach, we utilized OpenAI’s GPT-4o-mini (paid tier 1) (OpenAI 2024) to edit the EHR excerpts with specific instructions to act as “an expert editor of clinician notes who wants to reduce negative and stigmatizing language in text using non-negative and non-stigmatizing language and reduce evidential and judgment, one can inject empathy/use alternative empathic rephrasing or remove words when possible.” To achieve optimal performance, we carefully selected and fine-tuned the model’s parameters: `top_p = 1.0`, `max_tokens = 700`, and `temperature = 0.5`. We did not specify demographic features for the LLM; however, gender was frequently mentioned in the notes, and many conditions are age-related, making age easily inferrable by the model. More critically, the word choices in physicians’ notes embed race-specific biases, enabling models to readily detect them (Gabriel et al. 2024) (e.g., “non-compliant”, “resistant” for Black patients). An example of how a patient’s notes are modified from the unedited version to context-based edits is shown in Fig. 3, the patient is an adult, Female Latina. Here we see that, for the context-based approach, “complains” is updated to “is experiencing”. Ironically, but not unexpectedly, we observed that many of the word replacements generated by the LLM during editing were also present in the lexicons and classified as testimonially unjust. The replacement choice “reports”, or lack of replacement, in some instances, was among the most common.

Unedited	Rule-Based Edits	Context-Based Edits
She primarily complains of her headache, which she describes as the worst headache of her life. She reports taking an unknown antibiotic for a week approximately two weeks ago for an oral infection after having a tooth pulled.	She expressed concerns of her headache, which she describes as the worst headache of her life. She has been taking an unknown antibiotic for a week approximately two weeks ago for an oral infection after having a tooth pulled.	Currently, she is experiencing significant discomfort, primarily noting her headache, which she describes as the most severe she has ever experienced. She mentioned taking an unspecified antibiotic for about a week two weeks ago for an oral infection following a tooth extraction.

Figure 3: Example of Edited Notes Across Modification Types for an Adult, Female Latina

LLM Analysis of Notes Post-Modification To mitigate the risk of the model being biased against itself, especially after the context-based edits, we used a different LLM to analyze the notes. We asked Google DeepMind’s Gemini 2.0 Experimental Flash (DeepMind 2024) (`top_p = 0.95` and `temperature = 0.5`) using the free-tier, the following questions to analyze each of the 24 physicians’ notes. We instructed the model to “answer the questions in a professional manner”. Here, we explore whether improved note-taking

— i.e., reduced testimonial injustice in the notes — would enable an LLM to better perceive the patient or more accurately assess their condition.

- How do you perceive the patient’s current condition? Select one of the following:
 - Preventable due to better preventative care
 - Preventable but caused by their own doing
 - Likely environmental cause
 - Unfounded cause
- How do you perceive the patient? List four adjectives.
- What are the most critical symptoms of the patient?
- What are the least critical symptoms of the patient?
- What are the most likely diagnoses/conditions of the patient given what you have read?
- Should the patient receive immediate treatment or are they able to handle it at a later time? Select one: (Immediate, Later)

We ran the physicians’ notes through Gemini asking the aforementioned questions without edits (i.e., as is directly from the MIMIC-III dataset), with edits as informed by the causal model, and with edits considering context by GPT-4o-mini.

Survey: Human Expert Analysis of Notes Post-Modifications We surveyed US Board-Certified medical experts who regularly read physicians’ notes and interact with patients — specifically Physicians, Nurse Practitioners, and Physician Assistants to see if improved note-taking — i.e., reduced testimonial injustice in the notes — would enable them to better perceive the patient or more accurately assess their condition. They reviewed both edited and unedited versions of the notes and answered the same aforementioned questions. Additionally, we collected demographic information from them as listed in the Appendix.

We presented our human experts with 2 versions of the survey, determined by a coin-flip. If the participant reported they flipped “Heads” in their coin flip, they were presented with the first version of the survey, referred to hitherto as **VA**. Otherwise, they were presented with the second version of the survey, **VB**. VA contained the original, unedited notes from MIMIC-III, while VB included edited notes, evenly split between rule-based and context-based.

Results

We hypothesize that testimonial injustice terms in ICU medical records can bias perceptions of patients and compromise fair assessment. Guided by our SCM, we adapt note excerpts to reduce injustice through word removal, alternative phrasing, and empathy reframing, aiming to promote more just and accurate patient evaluations.

Semantic Similarity Post-Modifications One may wonder if such changes are actually changing the meaning of the patient summaries. Upon performing semantic similarity analysis (cosine similarity (Reimers and Gurevych 2019)) on the embeddings of the experts of the patient summaries for rule-based and context-based edits, we found the minimum similarity values were high across both types with senior Latino males having 97.63% and adult Latino males

having 86.11% respectively. Maximum similarity for child White males at 100% for rule-based edits and adult White males at 99.1% for context-based edits.

LLM Analysis We sent the 3 versions of the 24 notes to Gemini: (1) unedited, (2) rule-based edits, and (3) context-based edits. The general trend changes in Gemini’s responses across patient perception and gauging the condition of the patient, post-modifications, are summarized in Fig. 5. Both approaches show occasional changes in patient descriptors, but no change in blame or diagnosis. However, urgency & causal reasoning increased in the rule-based approach and even more significantly in the context-based approach.

For the unedited notes, the model’s responses were efficient and generally neutral. However, after editing the notes, we observed the greatest variability in the model’s justification when answering the question of whether a patient’s symptoms required immediate or delayed attention. Notably, urgency increased with rule-based edits and, at times, even more so with context-based edits. Additionally, for the same question, both rule-based and context-based edits produced more detailed justifications, linking symptoms to potential conditions or diseases in a causal manner. Context-based edits further emphasized the urgency of the issue, whereas unedited text often led to more direct recommendations. Lastly, we saw after the edits only 1 instance of variability when the model was asked to give adjectives of the patient whereas in the unedited version the list of adjectives changed from Critical, *Unstable*, Compromised, High-Risk to Critical, *Vulnerable*, Compromised, High-Risk post edits. By observation, in this medical setting, “unstable” was mostly used to describe patients whose conditions could change at any moment. While “vulnerable” was most used to describe patients who were likely to contract another disease or have worsening conditions without immediate treatment. It follows logically that vulnerability can be seen as a more specific characterization of unstable in this context. This change only occurred once in this very small sample set. However, with a larger dataset, it is reasonable to expect that changes in how patients are described (through adjectives) could vary with edits—just as the perceived urgency of their conditions changed. It is important to recall that these physicians’ notes originate from ICU settings, where the needs of every patient are always immediate. However, adding context, detailed explanations, and a sense of urgency can enhance understanding and improve the perception of the patients’ conditions. Therefore, when LLMs are fed physicians’ notes, physicians can aid the perception of their patients and clarity in the actual condition by being mindful of their word choice and, when appropriate, editing their notes. An example of the feedback provided by Gemini 2.0 Flash is shown in Fig. 4.

A limitation of the model’s responses was its tendency to interpret several aspects of a patient’s experiences as symptoms, despite categorizing them as the least critical. For instance, in the case of a adult Black woman, the original medical record stated: “The patient lives alone in the hospital. She is currently unemployed.” GPT-4.o-mini revised this to “The patient is currently hospitalized and lives indepen-

dently”, whereas the rule-based edits left the text unchanged due to none of the terms being present in the lexicons of concern. Here, Gemini identified “Lives Alone and Unemployed” as 1 of the least critical symptoms. Additionally, the model’s responses lacked diversity in the adjectives used to describe patients. This is likely due to the tendency of LLMs to over-generalize (Allaway et al. 2024).

Human Expert Analysis We presented our human experts with 2 versions of the survey, VA and VB, as described previously. An example of the feedback provided by the experts is shown in Fig. 4. We had 17 complete expert surveys, with each expert reviewing 5 distinct patient cases, resulting in 85 expert evaluations. Thus, this experimental design provided multiple independent judgments per condition, increasing the robustness and reliability of our findings.

To give more insight, surveys that were not completed typically ended at the coin-flip question. Since we provided a link to an online coin flipper, the lack of a physical coin was not a limiting factor. However, the survey length may have been a challenge, as it took participants an average of 25 minutes to complete. Also, there were slightly more respondents who flipped heads (3) than tails — well within the typical statistics for a Bernoulli($\frac{1}{2}$) random variable.

Each participant reviewed 5 physicians’ notes in VA or VB, randomly selected from the set of 24 patients to improve response quality and manage survey length. All but 3 patient profiles were randomly selected for review and none were reviewed more than 3 times. The 3 profiles that were not randomly selected were: adult White female, child Black female, and child Asian female. With a larger sample size and extended survey time, all patient profiles would likely have been reviewed at least once. The general trend changes in human experts’ opinions across patient perception and gauging the condition of the patient are summarized in Fig. 6. There were frequent changes in patient descriptors across both methods, as the clinicians had a number of different descriptors chose to use (since this was an open ended question in the survey). It primarily reflects the richness of their vocabularies. The blame for the patient’s condition tended to shift from the patient to poor medical care in the rule-based approach and to unfounded causes in the context-based approach. Urgency & causal reasoning increased in the rule-based approach and even more significantly in the context-based approach, just as with the LLM analysis. Diagnosis changed occasionally in the rule-based approach and more frequently in the context-based approach.

When asked, “How do you perceive the patient’s current condition?”, human experts at times attributed conditions to “unfounded causes” even in unedited notes (VA), which did not occur for the notes analyzed with Gemini. Further, they tended to shift blame away from patients and toward poor medical care. In cases where experts blamed the patient for their condition in unedited notes, rule-based edits most often led them to shift blame toward poor medical care. Overall, rule-based edits resulted in the least blame attributed to patients, with more blame assigned to unfounded or environmental causes. However, context-based edits reversed this trend, leading to an increase in patient blame and a slight

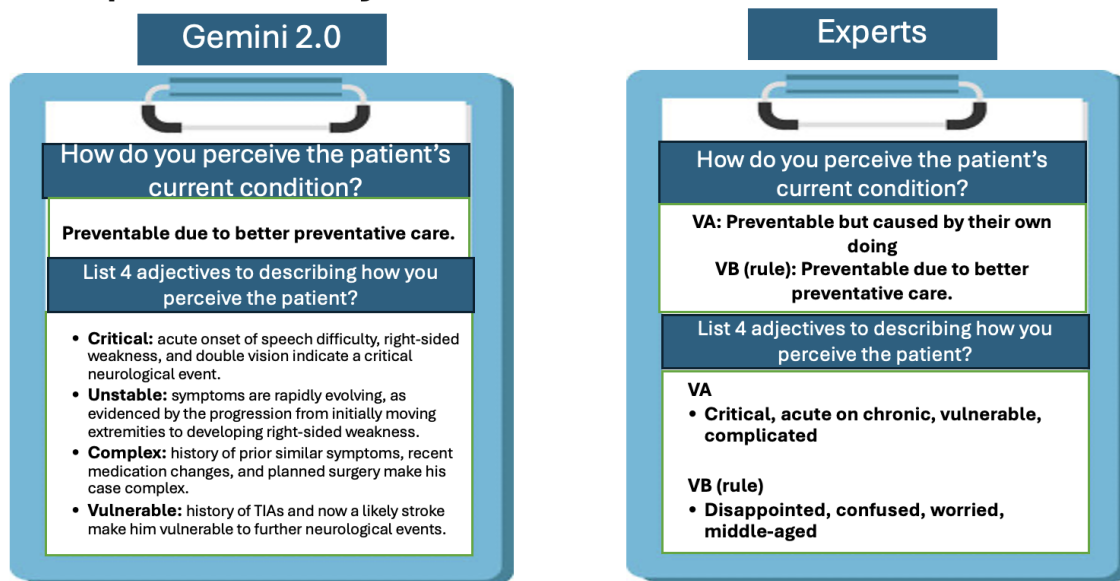


Figure 4: Example of Expert and Gemini 2.0 Feedback on Patient Notes Post-Modifications

	Perception of Patient		Gauge Patient Condition	
	Descriptors*	Blame	Urgency & Causal Reasoning	Diagnosis
Rule Based	Occasional Changes	No change	Moderate Increase	No change
Context Based	Occasional Changes	No change	High Increase	No change

Figure 5: Trends Post-Modifications of Notes in Rule-Based and Context-Based Approaches Across Patient Perception and Condition by Gemini 2.0 Flash. *LLMs have little variety in describing patients and their experiences

	Perception of Patient		Gauge Patient Condition	
	Descriptors*	Blame	Urgency & Causal Reasoning	Diagnosis
Rule Based	Moderate Changes	Patient to medical care	Moderate Increase	Occasional Changes
Context Based	Moderate Changes	Medical care to unfounded causes	High Increase	Moderate Changes

Figure 6: Trends Post-Modifications of Notes in Rule-Based and Context-Based Approaches Across Patient Perception and Condition by Experts. *Humans had various ways of describing patient experiences due to their vast vocabulary.

increase in unfounded causes (just 1 additional case).

We also observed high variability in expert responses to the same question, particularly regarding the cause of the patient's condition. Across the 21 profiles, respondents disagreed on 16 cases but had alignment with 5 of the VA excerpts and 4 with VB. Additionally, expert analysis revealed that some patients were perceived as eligible for delayed treatment rather than immediate care. This occurred only once, with responses remaining consistent across the 2 participants who answered this question — 1 reviewing VA and the other reviewing VB with context-based edits.

We do see that physicians gauge conditions differently for 3 patients under context-based edits and 1 patient under rule-based edits. Specifically, under context-based, pneumonia was newly perceived as hypertension, heart failure or edema as an infected AV site, which can lead to edema, and gastrointestinal issues and ulcers as COPD. After rule-based edits, a urinary tract infection was reclassified as a stroke.

Finally, human experts used a wider variety of adjectives to describe patients and their experiences compared to the LLM. Unlike the LLM, experts did not misinterpret life experiences as symptoms.

Sentiment Analysis We use the nltk library's SentimentIntensityAnalyzer (Loper and Bird 2002) to perform sentiment analysis on the excerpts of physicians' notes and feedback from Gemini and the human experts.

Patient Adjectives by LLM: Across all 24 patients, each has at least 2 negative adjectives, and 22 have at least 1 neutral adjective, while 2 have a positive adjective. The most common negative adjectives are "Unstable", "Vulnerable", and "Critical", with "Critical" reflecting life-threatening ICU conditions in the MIMIC-III dataset. Neutral adjectives, such as "Complex" and "Monitored", describe multifaceted conditions or observation status. The only positive adjective, "Challenging", likely carries negative connotations due to patient apprehension. Sentiment analysis across original, rule-based, and LLM-edited responses shows negative sentiment dominates in all categories, with 24 negative and 4 neutral responses for original and LLM-edited responses, and 23 negative and 5 neutral for rule-based edits. Overall, LLM-generated descriptions predominantly use negative terms, with few neutral and almost no positive descriptors.

Patient Adjectives by Experts: Experts generally provided neutral responses overall, though their adjectives reveal

more nuance. For VA, all but 2 patients had at least 2 neutral adjectives, with the remaining patients described solely with negative terms. Common negative adjectives were Ill”, Vulnerable”, and Confused”; neutral adjectives included Complex”, Guarded”, and Treatable”. Positive adjectives appeared once (Determined”). For VB (context-based), frequent negative terms were Anxious”, Ill”, Grimacing”, and Frustrated”, while neutral adjectives included Tachypneic”, Edematous”, and Resilient”; positive terms appeared once (“Hopeful”). Overall, neutral terms were most frequent, consistent with recommended documentation practices. Term counts were: VA—31 neutral, 27 negative, 6 positive; VB (context)—32 neutral, 24 negative, 3 positive; VB (rule)—14 neutral, 11 negative, 2 positive.

Edits The sentiment analysis of the unedited excerpts, rule-based edits, and LLM-based edits of the EHRs revealed some differences in how sentiment was altered across these versions. The original excerpts showed a predominance of negative sentiment, with 21 instances categorized as negative, 5 neutral, and 2 positive. Upon editing EHRs according to the causal-informed rule-based edits, the number of negative excerpts decreased slightly to 19, while the neutral excerpts increased to 6, and the positive rose to 3. This indicates that the rule-based edits softened the negative tone slightly but did not drastically change the overall sentiment distribution. The LLM-based edits showed a similar decrease in negative sentiment, with 19 instances, however, the number of positive excerpts increased more significantly, from 3 in the rule-based edits to 5, and the neutral category slightly decreased from 6 in the rule-based edits to 4. This suggests that the LLM-based edits introduced a more positive tone, compared to both the original and rule-based edits. Overall, both rule-based edits and LLM-based edits reduced the negative sentiment across the notes compared to the unedited versions. The LLM-based edits led to a more pronounced shift towards positive sentiment, while the rule-based edits had a more modest effect which is a much more desirable trait in medical writing and the preferred style of speech in medicine (Healy, Richard, and Kidia 2022).

Discussion and Conclusion

Our findings demonstrate that age, gender, and race each contribute to testimonial injustice in distinct and interacting ways, with evidential, judgmental, and stigmatizing terms driving most unjust documentation. Race is associated with stigmatizing and judgmental language, suggesting that patients’ experiences—particularly in high-stakes contexts like the ICU—are often diminished or vilified. Age influences the use of evidential and judgmental terms, indicating that even when patients are acknowledged, physicians may remain skeptical or critical. Gender is linked to negative terms, contributing to experiences being dismissed. Negative terms rarely appear in isolation, instead emerging alongside other aspects of injustice. These findings highlight the importance of intersectionality: no single factor explains unjust language, and understanding their interactions is essential to target interventions that can mitigate these harms. Ethically, these disparities represent not only misperception but a failure to uphold equitable, just care, with implications for pa-

tient trust and wellbeing.

Our causal analysis situates these patterns within a broader institutional and societal context. Although our data focus on ICU notes, similar cognitive and systemic biases are pervasive across healthcare, as noted by the American Medical Association. The disparities we observe are not merely descriptive; they reveal mechanisms through which structural and individual biases are encoded in clinical documentation. Interventions that address these patterns—through careful language choices, rule-based or context-informed edits, or awareness training—can influence how both human experts and AI systems perceive patients, helping to reduce misdiagnosis, unjust blame, and inequitable outcomes. At the same time, these interventions must be implemented thoughtfully, acknowledging potential limitations, such as over-correction, clinician resistance, or misinterpretation by downstream AI tools.

Importantly, both human experts and LLMs responded positively to edited notes: LLMs expressed greater urgency, provided more contextualized reasoning, and linked symptoms to severity more accurately, while human experts shifted blame away from patients toward systemic factors, particularly under rule-based edits. These results underscore that small, targeted interventions in documentation can have measurable downstream effects on perception and decision-making. Moving forward, encouraging awareness of testimonial injustice, promoting unbiased documentation, and integrating these insights into clinical AI systems are critical steps toward equitable, ethical, and effective healthcare assessment.

Limitations & Future Work

Better data: Boston has a median age of approximately 33 for both males and females, with a racial composition of White (50.13%), Black (21.7%), Asian (9.59%), and Latino (5.92%) (USA 2022). While race emerges as a frequent contributor to testimonial injustice, age and gender also play important roles, reflecting nuanced intersectional effects. More representative datasets of marginalized groups are critical to advance this work, yet such data are rare due to limited access to care and historical distrust in healthcare systems. Expanding datasets to include balanced race, age, and gender distributions would improve causal inference and allow incorporation of features known to clinicians but not otherwise observable. Examination of our SCMs further reveals potential unknown confounders; for instance, stigmatizing and negative terms may co-occur due to implicit biases, potentially linked to unobserved, inferrable characteristics such as experiences specific to women (e.g., perceived obesity). These findings underscore the need to account for intersectional identities and systemic factors to fully understand and mitigate testimonial injustice.

Analysis: In conducting further analysis on the expert perceptions, we would like to understand the nuances that occur across them, given their own features. Specifically, we believe it worth accounting for personal features such as length of practice, sex, and the primary medical institution in terms of how they influence their assessments.

Ethical Statement

In this work, we examine whether mitigation of harms in clinical language can meaningfully benefit populations historically marginalized within ICU and broader healthcare settings, as defined by prior literature. While improvements in language and clinical modeling may provide broader benefits across patient populations, our primary objective in this work is to investigate whether such approaches can improve representation, accessibility, and care-related outcomes for groups that have historically experienced disproportionate barriers to equitable and accurate treatment and participation in healthcare systems. This focus is intentional rather than exclusionary: we do not regard harms experienced by other populations as unimportant, but instead prioritize evaluating if these methods can address disparities affecting communities that are disproportionately impacted by structural inequities, experience systemic barriers to care, or are historically excluded not only within clinical settings, but across other high-stakes contexts as well. These considerations highlight the importance of design choices in such settings, often sensitive to structural inequities, and their downstream effects on patients of marginalized backgrounds.

Acknowledgments

This work was partially supported by the NSF Program on Fairness in AI in Collaboration with Amazon under Award No. IIS-1939743, titled FAI: Addressing the 3D Challenges for Data-Driven Fairness: Deficiency, Dynamics, and Disagreement. Any opinion, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation or Amazon.

References

- Adam, H.; Yang, M. Y.; Cato, K.; Baldini, I.; Senteio, C.; Celi, L. A.; Zeng, J.; Singh, M.; and Ghassemi, M. 2022. Write It Like You See It: Detectable Differences in Clinical Notes by Race Lead to Differential Model Recommendations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 7–21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392471.
- Allaway, E.; Bhagavatula, C.; Hwang, J. D.; McKeown, K.; and Leslie, S.-J. 2024. Exceptions, instantiations, and overgeneralization: Insights into how language models process generics. *Computational Linguistics*, 50(4): 1211–1275.
- Amemiya, J.; Mortenson, E.; Heyman, G. D.; and Walker, C. M. 2023. Thinking structurally: A cognitive framework for understanding how people attribute inequality to structural causes. *Perspectives on Psychological Science*, 18(2): 259–274.
- Andrews, K.; Shah, B.; and Cheng, L. 2023. Intersectionality and Testimonial Injustice in Medical Records. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, 358–372. Toronto, Canada: Association for Computational Linguistics.
- Beach, M. C.; Saha, S.; Park, J.; Taylor, J.; Drew, P.; Plank, E.; Cooper, L. A.; and Chee, B. 2021. Testimonial injustice: linguistic bias in the medical records of Black patients and women. *Journal of general internal medicine*, 36(6): 1708–1714.
- Ben-Harush, A.; Shiovitz-Ezra, S.; Doron, I.; Alon, S.; Leibovitz, A.; Golander, H.; Haron, Y.; and Ayalon, L. 2017. Ageism among physicians, nurses, and social workers: Findings from a qualitative study. *European journal of ageing*, 14: 39–48.
- Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc."
- Boudana, S. 2016. Impartiality is not fair: Toward an alternative approach to the evaluation of content bias in news stories. *Journalism*, 17(5): 600–618.
- Cantiello, J.; Fottler, M. D.; Oetjen, D.; and Zhang, N. J. 2015. The impact of demographic and perceptual variables on a young adult's decision to be covered by private health insurance. *BMC Health Services Research*, 15(1): 195.
- Crenshaw, K. W. 2013. Mapping the margins: Intersectionality, identity politics, and violence against women of color. In *The public nature of private violence*, 93–118. Routledge.
- DeepMind, G. 2024. Gemini 2.0 Flash Experimental.
- Eriksson, L.; Dudas, V.; Carlbon, A.; and Essén, B. 2023. Perceived religious discrimination in healthcare: A qualitative study of formal complaints. In *Culture, Spirituality and Religious Literacy in Healthcare*, 90–107. Routledge.
- for Environmental Conflict Resolution, U. I. 2010. The University of Wisconsin-Stevens Point. *Training Workshop on Introduction to Managing Environmental Conflict*.
- Fricker, M. 2019. Testimonial injustice. *Contemporary Epistemology: An Anthology*, 149–163.
- Gabriel, S.; Puri, I.; Xu, X.; Malgaroli, M.; and Ghassemi, M. 2024. Can AI Relate: Testing Large Language Model Response for Mental Health Support. *arXiv preprint arXiv:2405.12021*.
- Gliadkovskaya, A. 2024. Some doctors are using public AI chatbots like CHATGPT in clinical decisions. is it safe?
- Gonen, H.; and Goldberg, Y. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.
- Gouezec, H.; Lerenard, I.; Jan, S.; Bajeux, E.; Renaudier, P.; Mertes, P.; et al. 2016. Physicians' perception on taking care of Jehovah's witnesses. *Transfusion Clinique et Biologique: Journal de la Societe Francaise de Transfusion Sanguine*, 23(4): 196–201.
- Goyal, M. K.; Kuppermann, N.; Cleary, S. D.; Teach, S. J.; and Chamberlain, J. M. 2015. Racial disparities in pain management of children with appendicitis in emergency departments. *JAMA pediatrics*, 169(11): 996–1002.
- Hall, W. J.; Chapman, M. V.; Lee, K. M.; Merino, Y. M.; Thomas, T. W.; Payne, B. K.; Eng, E.; Day, S. H.; and Coyne-Beasley, T. 2015. Implicit racial/ethnic bias among health care professionals and its influence on health care

- outcomes: a systematic review. *American journal of public health*, 105(12): e60–e76.
- Halpern, J. 2003. What is clinical empathy? *Journal of general internal medicine*, 18: 670–674.
- Harrigian, K.; Zirikly, A.; Chee, B.; Ahmad, A.; Links, A.; Saha, S.; Beach, M. C.; and Dredze, M. 2023. Characterization of Stigmatizing Language in Medical Records. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 312–329. Toronto, Canada: Association for Computational Linguistics.
- Healy, M.; Richard, A.; and Kidia, K. 2022. How to reduce stigma and bias in clinical communication: a narrative review. *Journal of General Internal Medicine*, 37(10): 2533–2540.
- Himmelstein, G.; Bates, D.; and Zhou, L. 2022. Examination of Stigmatizing Language in the Electronic Health Record. *JAMA Network Open*, 5(1): e2144967–e2144967.
- Ho, J. J.; Jones, K. F.; Sager, Z.; Wakeman, S.; and Merlin, J. S. 2022. De-Stigmatizing the Language of Addiction# 429. *Journal of Palliative Medicine*, 25(1): 155–157.
- Howell, E. A. 2018. Reducing disparities in severe maternal morbidity and mortality. *Clinical obstetrics and gynecology*, 61(2): 387.
- Keshet, Y.; and Popper-Giveon, A. 2019. Language practice and policy in Israeli hospitals: The case of the Hebrew and Arabic languages. *Israel Journal of Health Policy Research*, 8(1): 58.
- L., S. 2020. Nine times out of 10, I was completely brushed off’: Black Chicagoans confront bias in health care, hope for change. *Chicago Tribune*.
- Lane, B. J.; and Rae, D. I. 1983. The prejudice of language: Effects of word choice on impressions formed by nurses. *Canadian Journal of Nursing Research Archive*, 21–33.
- Lee, P.; Bubeck, S.; and Petro, J. 2023. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine*, 388(13): 1233–1239.
- Lee, R. K. 2014. Judging Judges: Empathy as the Litmus Test for Impartiality. *University of Cincinnati Law Review*, 82.
- Link, B. G.; and Phelan, J. C. 2001. Conceptualizing stigma. *Annual review of Sociology*, 27(1): 363–385.
- Loper, E.; and Bird, S. 2002. NLTK: The Natural Language Toolkit.
- Markakis, K.; Frankel, R.; Beckman, H.; and Suchman, A. 1999. Teaching empathy: it can be done. In *Working paper presented at the Annual Meeting of the Society of General Internal Medicine*. San Francisco, Calif.
- Marques, A. C. 2018. Patricia Hill Collins and Sirma Bilge, Intersectionality.
- Metersky, M. L.; Fine, M. J.; and Mortensen, E. M. 2012. The effect of marital status on the presentation and outcomes of elderly male veterans hospitalized for pneumonia. *Chest*, 142(4): 982–987.
- Monuteaux, M. C.; Du, M.; and Neuman, M. I. 2024. Evaluation of insurance type as a proxy for socioeconomic status in the pediatric emergency department: a pilot study. *Annals of Emergency Medicine*, 83(6): 562–567.
- on Drug Abuse, N. I. 2021. Words matter-terms to use and avoid when talking about addiction. *U.S. Department of Health and Human Services*.
- OpenAI. 2024. GPT-4o mini: Advancing Cost-Efficient Intelligence.
- P Goddu, A.; O’Conor, K. J.; Lanzkron, S.; Saheed, M. O.; Saha, S.; Peek, M. E.; Haywood, C.; and Beach, M. C. 2018. Do words matter? Stigmatizing language and the transmission of bias in the medical record. *Journal of general internal medicine*, 33: 685–691.
- Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.
- Pryzant, R.; Martinez, R. D.; Dass, N.; Kurohashi, S.; Jurafsky, D.; and Yang, D. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, 480–489.
- Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Sangal, R. B.; Su, H.; Khidir, H.; Parwani, V.; Liebhardt, B.; Pinker, E. J.; Meng, L.; Venkatesh, A. K.; and Ulrich, A. 2023. Sociodemographic Disparities in Queue Jumping for Emergency Department Care. *JAMA Network Open*, 6(7): e2326338–e2326338.
- Spirtes, P.; Glymour, C. N.; and Scheines, R. 2000. *Causation, prediction, and search*. MIT press.
- Stepanikova, I.; and Cook, K. S. 2008. Effects of poverty and lack of insurance on perceptions of racial and ethnic bias in health care. *Health services research*, 43(3): 915–930.
- Sun, M.; Oliwa, T.; Peek, M. E.; and Tung, E. L. 2022. Negative Patient Descriptors: Documenting Racial Bias In The Electronic Health Record. *Health Affairs*, 41(2): 203–211.
- Tolbert, J.; Cervantes, S.; Bell, C.; and Damico, A. 2024. Key Facts about the Uninsured Population. <https://www.kff.org/uninsured/key-facts-about-the-uninsured-population/>.
- USA, D. 2022. Boston, MA DataUSA.
- (USMLE), U. S. M. L. E. 2023. Step 1 Sample Test Questions.
- Von Hippel, W.; Sekaquaptewa, D.; and Vargas, P. T. 2008. Linguistic markers of implicit attitudes. In *Attitudes*, 449–478. Psychology Press.
- Yearby, R.; Clark, B.; and Figueroa, J. F. 2022. Structural Racism In Historical And Modern US Health Care Policy: Study examines structural racism in historical and modern US health care policy. *Health Affairs*, 41(2): 187–194.
- Zhang, L. 2022. Negative Adverb List. 1–12.

Zhang, X.; Carabello, M.; Hill, T.; Bell, S. A.; Stephenson, R.; and Mahajan, P. 2020. Trends of racial/ethnic differences in emergency department care outcomes among adults in the United States from 2005 to 2016. *Frontiers in medicine*, 7: 300.

Zheng, Y.; Huang, B.; Chen, W.; Ramsey, J.; Gong, M.; Cai, R.; Shimizu, S.; Spirtes, P.; and Zhang, K. 2023. Causal-learn: Causal Discovery in Python. *arXiv preprint arXiv:2307.16405*.