CHEST

Measuring Implicit Bias in ICU Notes Using Word-Embedding Neural Network Models

Julien Cobert, MD; Hunter Mills, MS; Albert Lee, MS; Oksana Gologorskaya, MA; Edie Espejo, MA; Sun Young Jeon, PhD; W. John Boscardin, PhD; Timothy A. Heintz, BS; Christopher J. Kennedy, PhD; Deepshikha C. Ashana, MD; Allyson Cook Chapman, MD; Karthik Raghunathan, MD; Alex K. Smith, MD; and Sei J. Lee, MD

BACKGROUND: Language in nonmedical data sets is known to transmit human-like biases when used in natural language processing (NLP) algorithms that can reinforce disparities. It is unclear if NLP algorithms of medical notes could lead to similar transmissions of biases.

RESEARCH QUESTION: Can we identify implicit bias in clinical notes, and are biases stable across time and geography?

STUDY DESIGN AND METHODS: To determine whether different racial and ethnic descriptors are similar contextually to stigmatizing language in ICU notes and whether these relationships are stable across time and geography, we identified notes on critically ill adults admitted to the University of California, San Francisco (UCSF), from 2012 through 2022 and to Beth Israel Deaconess Hospital (BIDMC) from 2001 through 2012. Because word meaning is derived largely from context, we trained unsupervised word-embedding algorithms to measure the similarity (cosine similarity) quantitatively of the context between a racial or ethnic descriptor (eg, *African-American*) and a stigmatizing target word (eg, *nonco-operative*) or group of words (*violence*, *passivity*, *noncompliance*, *nonadherence*).

RESULTS: In UCSF notes, Black descriptors were less likely to be similar contextually to violent words compared with White descriptors. Contrastingly, in BIDMC notes, Black descriptors were more likely to be similar contextually to violent words compared with White descriptors. The UCSF data set also showed that Black descriptors were more similar contextually to passivity and noncompliance words compared with Latinx descriptors.

INTERPRETATION: Implicit bias is identifiable in ICU notes. Racial and ethnic group descriptors carry different contextual relationships to stigmatizing words, depending on when and where notes were written. Because NLP models seem able to transmit implicit bias from training data, use of NLP algorithms in clinical prediction could reinforce disparities. Active debiasing strategies may be necessary to achieve algorithmic fairness when using language models in clinical research.

CHEST 2024; 165(6):1481-1490

KEY WORDS: critical care; inequity; linguistics; machine learning; natural language processing

ABBREVIATIONS: BIDMC = Beth Israel Deaconess Hospital; LLM = large language model; MIMIC-III = Medical Information Mart for Intensive Care III; NLP = natural language processing; UCSF = University of California, San Francisco

AFFILIATIONS: From the Anesthesia Service (J. C.), San Francisco VA Health Care System, the Department of Geriatrics, Palliative, and Extended Care (A. K. S.), Veterans Affairs Medical Center, the Department of Anesthesia and Perioperative Care (J. C.), the Bakar Computational Health Sciences Institute (H. M., A. L., and O. G.), the Division of Geriatrics (E. E., S. Y. J., W. J. B., A. K. S., and S. J. L.), Department of Medicine, the Division of Critical Care and Palliative Medicine (A. C. C.), the Department of Surgery (A. C. C.), University of California, San Francisco, San Francisco, the School of Medicine (T. A. H.), University of California, San Diego, San Diego, CA,

Take-home Points

Study Question: Can we identify implicit bias in clinical notes and are biases stable across time and geography?

Results: We found that Black descriptors were less contextually similar to violent language terms, but we found opposite relationships in ICU notes from Massachusetts from 2001 through 2012.

Interpretation: Implicit race bias is identifiable in ICU notes, but relationships between race and ethnicity descriptors and stigmatizing words are not stable across time and geography, suggesting that natural language processing algorithms relying on clinical notes could reinforce disparities.

Unwanted racial stereotypes in large language models (LLMs) are well described in non-health care domains.[1,2] The discovery that LLMs can perpetuate biases from training data led some groups to use these methods to measure the magnitudes of stereotypes from nonmedical data sets.[3-6] Although racial stereotypes in clinical notes are rarely overtly blatant, toxic, or racist, recognition is growing that disparities in written language are present. Sometimes this language is explicit, such as race descriptors occurring more frequently in notes about Black patients.[7] However, stigmatizing language, representing words that dishonor, disgrace, or differentiate individuals in a depersonalizing way,[8] more often occurs in an implicit or unconscious manner.[9,10] Frequency and use of stigmatizing words relative to a patients' race, ethnicity, and diagnosis may provide important insights into implicit bias.[11,12] Such implicit bias affects how clinicians view patients,[13] potentially leading to patient disempowerment[14] and differential treatment decisions.[15-17]

Studies on use of racial, ethnic, and stigmatizing descriptors in notes rely on manual chart review,[7,10] which is burdensome and makes global understanding of bias difficult. Natural language processing (NLP), the computational study of language, provides the opportunity to study linguistic representations at scale. Some groups have used NLP to identify stigmatizing language and the differential use of such words across self-reported race,[11,12] but challenges exist in using well-established social science NLP tools when studying implicit bias.[18]

Word embeddings represent an NLP method that may be well suited for the study of implicit bias in language and how implicit bias may differ across training data. Broad agreement exists in linguistics that word meanings are defined by their context.[19] *Word embeddings* operationalize the importance of context by using neural networks to transform words into vectors within a contextual semantic space.[19] Mapping words in space provides nuanced descriptions of interrelationships between words such as race descriptors and stigmatizing words. To date, few studies have applied emerging NLP methods to study language patterns within medical notes around social constructs. Although it is now clear that such NLP methods can perpetuate biases in nonmedical contexts, it remains unclear whether clinical notes are spared of these biases, given their presumed objective data and less overt toxicity. Understanding these biases could inform future studies regarding how to approach training data and potentially strategies to remove bias from NLP algorithms, as has been carried out in other nonmedical algorithms.[12]

We sought to explore the contextual relationships of racial and ethnic descriptors and stigmatizing language in notes and to determine whether such relationships are consistent across notes from different time and geography. Extending work on counting stigmatizing language[11] in a broad population of patients, we used word embeddings to examine the presence of stigmatizing language specifically in ICU notes and its presence in relationship to documented race and ethnicity. We further examined whether the relationships of racial and ethnic descriptors with stigmatizing language differed across Black, White, and Latinx patients and whether such relationships were consistent across different periods and geographic locations across two US hospital systems. We hypothesized that: (1) stigmatizing words may be more contextually similar to Black or Latinx descriptors compared with White descriptors and, (2) like other studies in nonmedical contexts, these contextual relationships depend on the training data.

the Department of Psychiatry (C. K.), Harvard Medical School, the Center for Precision Psychiatry (C. J. K.), Massachusetts General Hospital, Boston, MA, the Division of Pulmonary, Allergy, and Critical Care Medicine (D. C. A.), and the Department of Anesthesia and Perioperative Care (K. R.), Duke University, Durham, NC.

CORRESPONDENCE TO: Julien Cobert, MD; email: Julien.cobert@ucsf.edu

Published by Elsevier Inc. under license from the American College of Chest Physicians.

DOI: https://doi.org/10.1016/j.chest.2023.12.031

Study Design and Methods

Overall Framework and Underlying Assumptions

Our primary assumption in this study is that co-occurrence of words in a note correlates with similarity in meaning.[19,20] When words co-occur more often, their meanings may be more associated. We determined to what extent an unsupervised NLP model considered a racial or ethnic descriptor as contextually similar to stigmatizing language. When presented with a racial descriptor, if the model predicts the presence of a stigmatizing word more often than chance, then the predicted stigmatizing word and the original racial descriptor could be considered contextually similar (Fig 1). All analyses were performed in two different data sets to determine stability and generalizability of these word relationships (e-Fig 1). Temporal evolution could change in use of racial descriptors and stigmatizing language; it adds additional complexity and was not included in the current study.

Study Design and Population

We conducted a retrospective cohort analysis of notes from patients aged 18 years or older admitted to ICUs across the University of California, San Francisco (UCSF; San Francisco, California), from 2012 through 2022 and Beth Israel Deaconess Medical Center (BIDMC; Boston, Massachusetts; from the Medical Information Mart for Intensive Care III [MIMIC-III] database) from 2001 through 2012. MIMIC-III is a de-identified data set developed by the Massachusetts Institute of Technology and BIDMC.[21] MIMIC-III was chosen because it is publicly available, and all code made for this project is made publicly available. Privacy concerns prevent UCSF's source data from being made publicly available. Also, MIMIC-III is temporally and geographically distinct from UCSF, and similarities in results across both data sets could represent more generalizable relationships between words. This study was approved by UCSF's institutional review board (Identifiers: IRB 20-30590, IRB 20-29918, and IRB 19-29429).

Overview of NLP

The note text was the primary unit of analysis. A complete list of included note types are found in e-Table 1. We used the Natural Language Toolkit version 3.8.1 for note tokenization.[22] Prespecified race and ethnicity words and phrases—called *tokens* (eg, *Black man* and *African American man*)—were combined at the preprocessing level using the Python regular expression library (Python Software Foundation). We performed token replacement to create three broad race and ethnicity descriptor groups (e-Table 2): Black, White, and Latinx. These represent descriptors of race or ethnicity by note writers and not what a patient self-identifies as their identity. Tokenized lines of text were processed further into n-grams using the Phraser package implemented in Gensim version 4.3.2 software.[23]

We inputted preprocessed n-grams into word2vec[24] using a continuous bag of words architecture (e-Appendix 1). Word embedding represents an unsupervised machine learning algorithm whereby tokens are converted into matrix vector representations and then used to explore semantic relationships (see Outcomes section that follows). See Figure 1 for a broad visualization of how word2vec can lead to semantic relationship descriptions.

Outcomes and Study Measurement

The explicit outcome of interest was the vector distance between an inputted base token (ie, race or ethnicity descriptors) and a prespecified target word or phrase (stigmatizing language) using the cosine similarity (cosine θ, whereby θ is the angle between two word vectors). This is used commonly for word embedding outputs and measures co-occurrence of words in a semantic space.[25] Cosine similarity represents a continuous range of numbers between –1 and +1. Base-target distances close to –1 are contextually similar, and distances close to +1 are contextually dissimilar. For a similarity = 0, each word's set of contexts are equally similar and dissimilar, suggesting that the words are contextually unrelated or neutral.

Potentially Stigmatizing Target Words, Phrases, and Thematic Groupings

We adopted previously published potentially stigmatizing terms[11] and expanded them to include alternative parts of speech, misspellings, and negations. We combined stigmatizing words into thematic groups of violence, passivity, nonadherence, and noncompliance (Table 1)[11] to determine if various race or ethnic base groups resulted in differences across thematic groups of words. Themes were constructed using domain knowledge. A subgroup analysis combined noncompliance and nonadherence words, given that both are conceptually similar and similarly lead to loss in patient autonomy.[26]

Statistical Analysis

We fit a word2vec model to estimate contextual word similarity within ICU notes. For each base group (Black, White, and Latinx) and target word pair, average cosine similarity was estimated using 20 stochastic estimates from the model accounting for Monte Carlo error. We used two measures, the mean and the precision-weighted mean, to produce point estimates for the average and for differences in average cosine similarities. Precision-weighted averages were used to combine targets belonging to the same theme. The bootstrap method with resampling was used to quantify the amount of uncertainty about our point estimate using resampled versions of our original notes data set (e-Appendix 1).[27] Because of the resampling nature of this method, some words had incomplete sets of predicted cosine similarities and do not have an associated CI (e-Table 3). Given that cosine similarities are foreign to most readers, we also chose tokens unlikely to be related to base race groups that we believed might demonstrate a neutral control for readers (ie, similarity = 0). We calculated these for the UCSF data set along with bootstrap standard errors and present details of these neutral controls and associated results in e-Table 4. All analyses were performed using R version 4.2.2 software (R Foundation for Statistical Computing). Figures were drawn using ggplot2 version 3.4.0 in R.

Results

We studied 392,982 notes across 8,214 critically ill adults at UCSF and 887,697 notes from 38,512 critically ill adults at BIDMC. Sociodemographic and clinical characteristics of patients and characteristics of notes and note authors are shown in e-Tables 5-6. MIMIC-III has a higher prevalence of White patients and a lower prevalence of Black and Latinx patients than UCSF. Within the UCSF data set, 966 unique patients had a race or ethnicity descriptor compared with 3,406 unique patients in MIMIC-III (specific breakdowns shown in e-Tables 7-9). Across UCSF notes were 9,511 unique note writers, whereas across MIMIC-III notes were 5,851 unique note writers. Across both data sets, most notes were written by physicians or nurses. Race and ethnicity characteristics of note authors were not available.

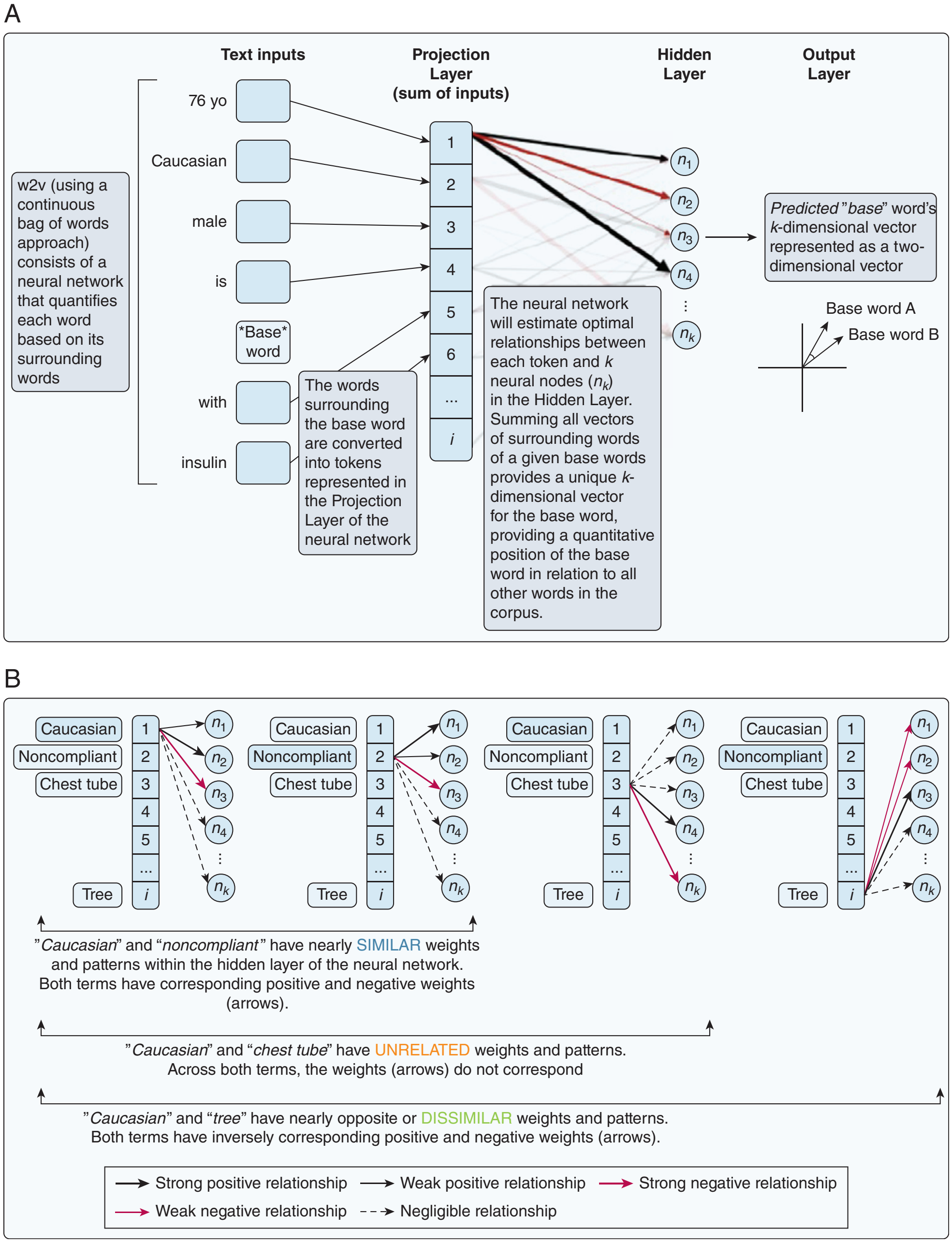
A
Text inputs
Projection Layer (sum of inputs)
Hidden Layer
Output Layer
w2v (using a continuous bag of words approach) consists of a neural network that quantifies each word based on its surrounding words
76 yo
Caucasian
male
is
Base word
with
insulin
1
2
3
4
5
6
...
i
The words surrounding the base word are converted into tokens represented in the Projection Layer of the neural network
n_1
n_2
n_3
n_4
n_k
The neural network will estimate optimal relationships between each token and k neural nodes (n_k) in the Hidden Layer. Summing all vectors of surrounding words of a given base words provides a unique k-dimensional vector for the base word, providing a quantitative position of the base word in relation to all other words in the corpus.
Predicted "base" word's k-dimensional vector represented as a two-dimensional vector
Base word A
Base word B
B
Caucasian
Noncompliant
Chest tube
Tree
1
2
3
4
5
...
i
n_1
n_2
n_3
n_4
n_k
"Caucasian" and "noncompliant" have nearly SIMILAR weights and patterns within the hidden layer of the neural network. Both terms have corresponding positive and negative weights (arrows).
"Caucasian" and "chest tube" have UNRELATED weights and patterns. Across both terms, the weights (arrows) do not correspond
"Caucasian" and "tree" have nearly opposite or DISSIMILAR weights and patterns. Both terms have inversely corresponding positive and negative weights (arrows).
Strong positive relationship
Weak positive relationship
Strong negative relationship
Weak negative relationship
Negligible relationship

Figure 1 – *Diagrams showing the overall framework of the w2v model and outcomes being assessed. A, Diagram showing what is inputted into w2v. All notes are preprocessed and eventually broken down to individual words or phrases (tokens). All tokens then are turned into a string of numbers that allows for imputation into the neural network (w2v). w2v consists of a single layer of neurons or nodes (n_k, within the neural network's so-called hidden layer) that learns the similarity of words from the underlying training note data set. Each node learns and iterates different weights for each word or token based on its context. Similarities of different words represent the probability such that a word is replaceable with another relative to the original word's context. Arrows in the w2v icon represents the positive (black) or negative (red) weight learned by each node in the hidden layer. B, Diagram showing how neural networks work in w2v. Neurons in the hidden layer calculate unique weights (black and red arrows) for each word or text string inputted into the model. Weights can be large and small and positive and negative. The group of neurons carries a unique pattern (represented by black and red arrows) unique to each inputted word. The unique patterns or weights are used to relate the input words to one another in geometric space to determine the relationship of each word to one another. The outputted word vectors are called word embeddings and are unique to the data set used as input and represent how different words that appear in similar contexts are represented as being close together spatially. Word embeddings (outputs of the w2v neural network) provide geometric relationships between two words within a document. C, Diagram showing how w2v outputs contextual similarity. The neural network calculates probabilities such that each word can be interchanged with every other word in the training data, represented by the degree of similarity between the* k*-dimensional vector outputted from the w2v neural network. The contextual similarity or dissimilarity then can be calculated by projecting the* k*-dimensional vectors onto a 2-D plane. Cosine similarity can be used as a continuous measure of how similar or dissimilar a base word (eg, Caucasian) is from a target word (eg, noncompliant) based on their contexts. A cosine similarity of +1 represents two words that are perfectly similar (Caucasian will perfectly match to Caucasian or a misspelling and will match closely to White person), and a cosine similarity of –1 represents two words that are perfectly dissimilar (Caucasian and Asian likely are nearly perfectly dissimilar because they are mutually exclusive). cos = cosine; w2v = word2vec.*

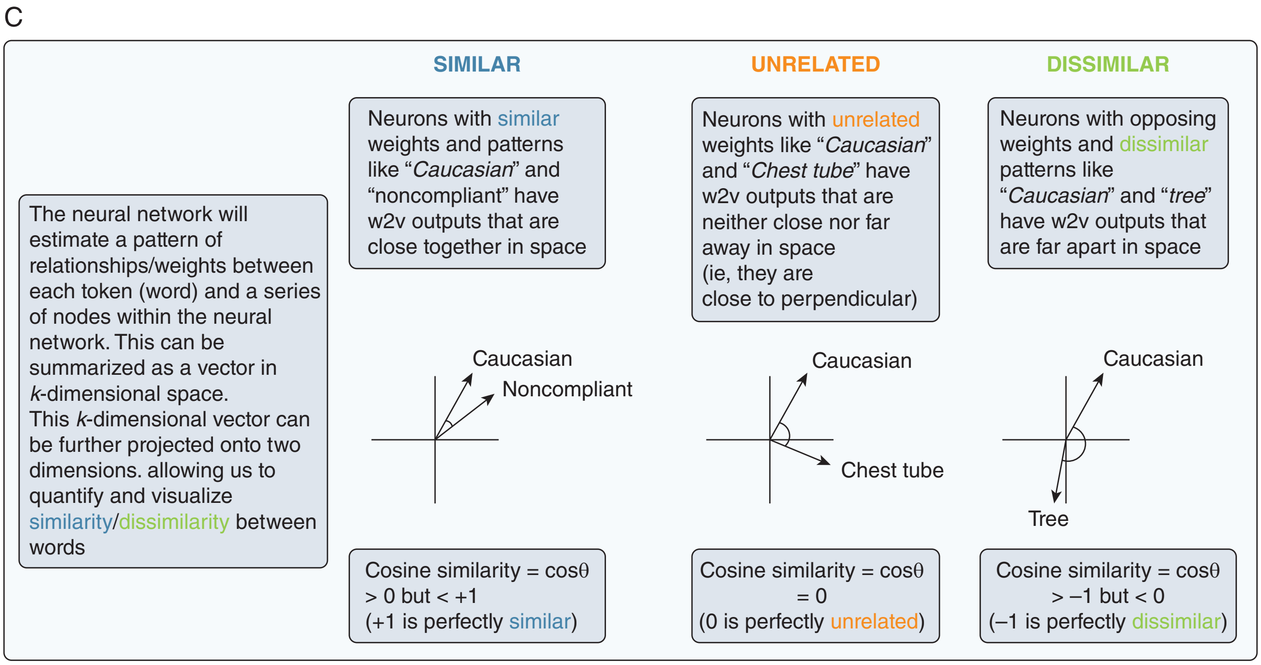
C
SIMILAR
UNRELATED
DISSIMILAR
The neural network will estimate a pattern of relationships/weights between each token (word) and a series of nodes within the neural network. This can be summarized as a vector in k-dimensional space. This k-dimensional vector can be further projected onto two dimensions. allowing us to quantify and visualize similarity/dissimilarity between words
Neurons with similar weights and patterns like "Caucasian" and "noncompliant" have w2v outputs that are close together in space
Neurons with unrelated weights like "Caucasian" and "Chest tube" have w2v outputs that are neither close nor far away in space (ie, they are close to perpendicular)
Neurons with opposing weights and dissimilar patterns like "Caucasian" and "tree" have w2v outputs that are far apart in space
Caucasian
Noncompliant
Caucasian
Chest tube
Caucasian
Tree
Cosine similarity = cosθ > 0 but < +1 (+1 is perfectly similar)
Cosine similarity = cosθ = 0 (0 is perfectly unrelated)
Cosine similarity = cosθ > –1 but < 0 (–1 is perfectly dissimilar)

Figure 1 – *Continued*

Many more mentions of White descriptors were found in the UCSF data set (85.8% of all race or ethnicity mentions at UCSF) compared with MIMIC-III (71.2% of all MIMIC-III mentions). Black mentions were less common in the UCSF data set (8.7%) compared with MIMIC-III (23.9%), and Latinx mentions were the least common, but were higher in the UCSF data set (6.5%) vs MIMIC-III (4.9%) (e-Tables 7-9). Some historical changes in language are apparent when comparing data sets as terms like *Latina* and *Latino* were present in MIMIC-III, but not in UCSF data.

We first determined whether individual base descriptors were contextually similar to individual stigmatizing words. A full list of results for individual base and target words is shown in e-Tables 10-11, and contextual similarity between race or ethnicity descriptors with thematic groups are shown graphically in e-Figures 2-5 across both data sets. In general, White descriptors were similar most frequently to individual stigmatizing words in the UCSF data set, but Latinx descriptors were similar most frequently in the MIMIC-III data set. Contextual dissimilarity was uncommon in general, but most often

TABLE 1] Thematic Groupings of Potentially Stigmatizing Words

Thematic Group	Strings Used
Violence	*combative*, *defensiveness*, *agitation*, *agitated*, *defensive*, *confronting*, *agitate*, *confront*, *confronted*, *combativeness*, *angry*, *angrily*, *aggressiveness*, *aggression*, *confrontation*, *aggressively*, *aggressive*, *confrontational*
Passivity	*nonadherent*, *challenges*, *noncooperative*, *resists*, *challenging*, *resisting*, *resist*, *resisted*, *non_compliant*, *noncompliant*, *resistance*, *unpleasant*, *noncompliance*, *non_adherence*, *non_compliance*, *challenged*, *non_adherent*, *resistances*, *resistant*, *nonadherence*
Nonadherence	*nonadherent*, *non_adherence*, *non_adherent*, *nonadherence*
Noncompliance	*non_compliant*, *noncompliant*, *noncompliance*, *non_compliance*
Noncompliance or nonadherence	Combined two lists (nonadherence and noncompliance) above

This list represents the complete list of words used for stigmatizing language, including alternative forms of speech, hyphenations, and negations. Thematic groups represent words used for precision-weighted averages. Multiple authors agreed on the violence and passivity subclass collation. Each subclass was evaluated using one-sample and two-sample *t* tests for cosine similarities. Underscores ("_") represent multiword tokens (eg, non adherence) inputted as single vectors into word2vec (eg, non_adherence). See e-Table 2 for strings used for Black, White, and Latinx descriptors. (Adapted from Sun et al.[11]).

occurred in the Latinx descriptor group in the UCSF data and in the Latinx and Black groups in MIMIC-III.

Relationships between racial and ethnic descriptors and thematic groups of stigmatizing words are shown in Figure 2 and numerically in e-Table 12. In UCSF notes, Black and White descriptors were contextually similar to violence-, passivity-, and noncompliance-themed words. In MIMIC-III, White descriptors were contextually dissimilar to violence and passivity-themed words. Black descriptors were contextually similar to noncompliance-themed words.

We next determined whether different race or ethnic descriptors were more or less similar to stigmatizing words relative to each other race or ethnic group. Full lists of results for individual target stigmatizing words are shown in e-Tables 13-14 and e-Figures 2-5. Differences between race and ethnic descriptors and thematic groups are shown in e-Figure 2A-2B and e-Table 15. In the UCSF data set, Black descriptors were less similar to violence-themed words relative to White descriptors (difference in cosine similarity, –0.055 [95% CI, –0.084 to –0.025]; $P < .05$). An opposite effect was seen in MIMIC-III, whereby Black descriptors were more similar to violence-themed words relative to White descriptors (difference in cosine similarity, 0.042 [95% CI, 0.016-0.068]; $P < .05$). The UCSF data set also showed that Black descriptors were more similar to passivity-themed groups (difference in cosine similarity, 0.051 [95% CI, 0.017-0.085]; $P < .05$) and noncompliance-themed groups (difference in cosine similarity, 0.110 [95% CI, 0.046-0.175]; $P < .05$) compared with Latinx descriptors. In MIMIC-III, Black descriptors were more similar to passivity-themed groups (difference in cosine similarity, 0.033 [95% CI, 0.002-0.063]; $P < .05$) and noncompliance-themed groups (difference in cosine similarity, 0.068 [95% CI, 0.024-0.113]; $P < .05$) compared with White descriptors. Combined nonadherence and noncompliance groups are shown in e-Figure 6.

Discussion

Word embeddings likely are foreign to most clinicians, but provide rich opportunities to study relationships between words over easily understandable approaches like counting words. However, word counts used by others to assess author bias[7,10,11] do not convey information about context or co-occurrence of neighboring words. Separate groups have studied the presence of racial descriptors and stigmatizing descriptors, but the proximity of these descriptors reveal additional insights. Our primary assumption was that clinical notes contain human biases because they are written by humans carrying their own cognitive frameworks and biases.

Concerningly, we demonstrated that many social descriptors are contextually similar to stigmatizing language and that some social descriptors are more often similar to disempowering words than others. Results are opposite when comparing data across different data sets. In the older MIMIC-III data set, compared with White descriptors, Black descriptors are more contextually similar to violence words, but are more dissimilar in the recent and geographically different UCSF data set. In MIMIC-III, fewer interactions are found, but Black descriptors more often are similar contextually to individual stigmatizing language than White or Latinx descriptors. We believe that the distinct base-target interactions are less important than the fact that they are present at all and, further, rely extensively on the underlying training data. Any unsupervised NLP algorithm trained broadly on notes (including more sophisticated ones like Transformers[28]) learn from these unequal use of descriptors. Our results suggest that LLMs can identify biases in clinical notes, raising concerns that LLMs used in clinical decision-making may reinforce or exacerbate disparities.

Language affects thought (Sapir-Whorf hypothesis[29]) and impacts clinical decisions.[30] In one study, a patient described as a “substance abuser” vs “having a substance use disorder” affected how clinicians judged patients.[31] Language around sickle cell disease affected choices in analgesia.[16] Although direct negative effects on patient outcomes have not been demonstrated robustly outside of experimental conditions, multiple groups highlighted the harm of stigma in clinical communication.[14,32,33] They call for avoidance of pejorative terms, use of person-first language,[34] use of inclusive language, avoidance of biased labels (eg, *nonadherent*), removal of blame,[35] and avoidance of patient quotations that could propagate stigma.[10]

We build on other studies demonstrating how language can lead to patient disempowerment and disparities or inequities in care. Instead of focusing on self-identified race and ethnicity and correlating to note language as others have done,[11] we sought to demonstrate the relationship between terms within notes. The decision to use a racial or ethnic descriptor, which is a distinct choice, in proximity to disempowering language, which is

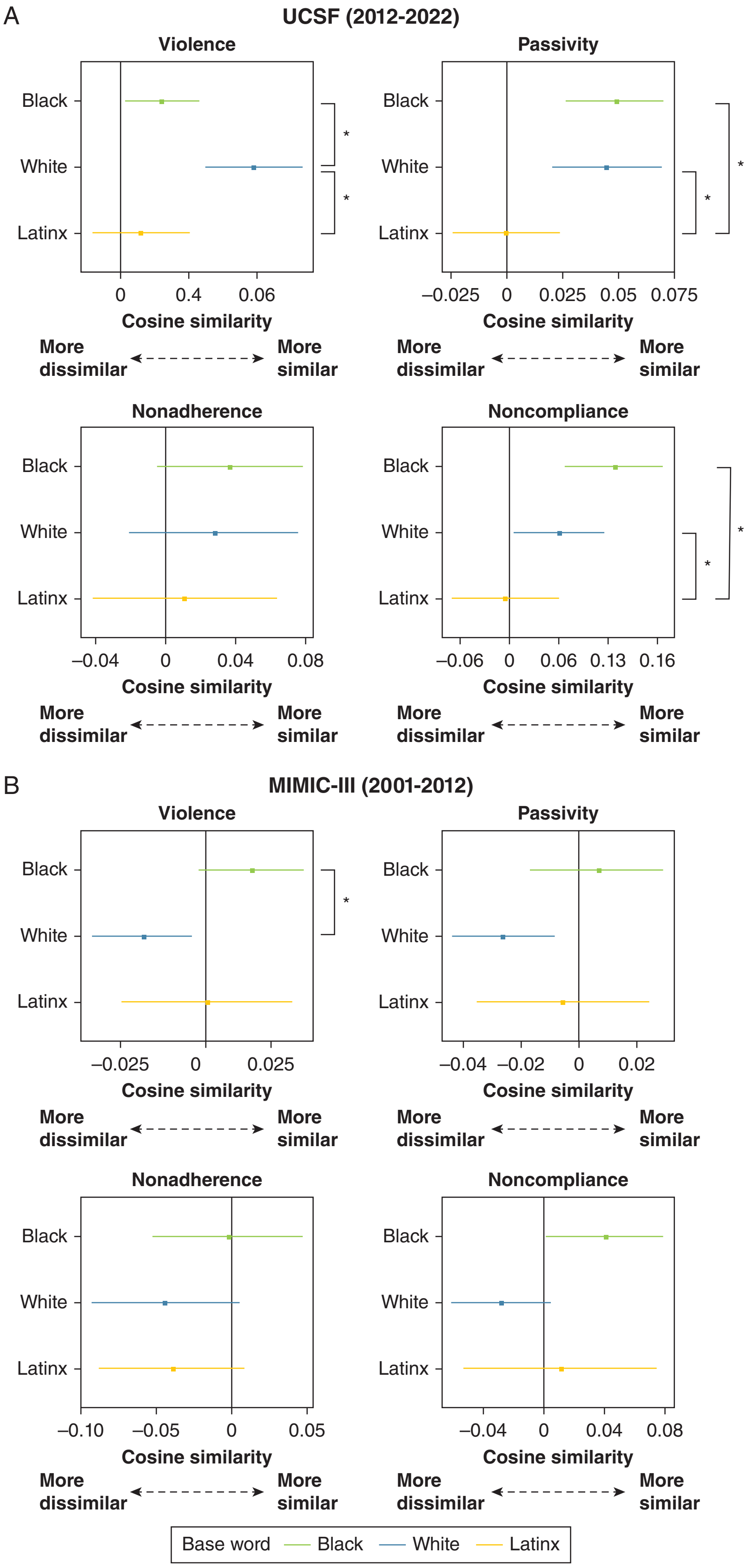

Figure 2 – *A, B, Precision-weighted averages were used to combine words containing to violence, passivity, nonadherence, and noncompliance across the UCSF data set (A) and MIMIC-III data set (B) to show individual and comparative similarity in race and ethnicity base words and target thematic word groups. Each part of the figure demonstrates the association of individual race or ethnicity descriptor groups and the thematic group of words and the differences across the racial or ethnic groups themselves. Individual horizontal bars represent the bootstrap 95% CIs for cosine similarities within each individual race or ethnicity and a thematic group of words. For horizontal CIs, statistical significance (defined as a* $P < .05$) *was met if they did not cross the cosine similarity value of 0. Vertical brackets with asterisks represent whether a statistically different difference exists in the similarity or difference across different races or ethnicities relative to a specific thematic group of words. Asterisks represent statistical significance with* $P < .05$ *for vertical brackets. MIMIC-III = Medical Information Mart for Intensive Care III; UCSF = University of California, San Francisco.*

another choice, differs from other studies describing how disempowering language is used across different racial groups using structured data. This represents a more explicit bias that simultaneously could be more nuanced (an additional decision is made to include a racial descriptor) and potentially more toxic (a direct link between a stigmatizing word and a race term). One group used machine learning models to determine if race information is embedded within note language when using unstructured notes to predict self-reported race, demonstrating the interplay of semantic and cognitive frameworks and how they are influenced by social constructs.[36] Our work supports these conclusions, but emphasizes the structural relationships of stigmatizing language specifically in the presence of racial descriptors (as opposed to when they are masked). Our study supports social science research demonstrating that collocation of words can lead to the emergence of implicit bias in accordance with distributional[20] and causal embedding[37,38] hypotheses (social stereotypes tracking from semantic distributions in language). Our study bridges work in the space of racial descriptors being used unnecessarily[7] and stigmatizing language used differentially across patients with differing self-identified race and ethnicity.[11] It also extends social science studies demonstrating how artificial intelligence algorithms transmit bias from underlying training data sets.[1]

The differences between relationships found in UCSF vs BIDMC are important because problematic relationships transmit in NLP models across different data sets, suggesting that bias may transmit across any medical data sets. Also, heterogeneous training data used for NLP models could lead to nongeneralizable results. Language is dependent on culture, historical time, and geography. It may be reassuring that more recent UCSF data suggest less contextual similarity between Black descriptors and violence-related language, but globally more references to race and associations between bases and targets were found in the UCSF data. Our results demonstrate broadly that temporal and geographic differences in the training data matter and that generalizability of such algorithms may be challenging. Although speculative, it is possible that differences in sociodemographic distributions of patients, note authors, or both or cultural differences based on geography or study periods could contribute to differences seen across the two data sets.

Implications of Study Findings

An important implication of this work is the concern that unsupervised models, including but not limited to those trained on clinical notes, transmit bias from underlying training data sets. This, in turn, could lead to direct negative patient effects when LLMs are applied to clinical settings. One commercial health care use prediction algorithm underestimated illness severity in Black vs White patients because it was trained on cost as the outcome of interest, likely leading to disproportionate care delivery.[39] The transmission of bias from algorithms has been well described and theorized in the literature.[40,41] Our results extend this to the health care NLP domain and critical care, allowing for implicit bias to be quantified. Explicit implications also include using our approach as a possible benchmark of underlying training data implicit bias, using cosine similarity between social descriptors and stigmatizing words as a target for debiasing strategies and the potential for word co-occurrences as a potential target of behavior change in note-writing tools.

Study Limitations

An important limitation related to neural networks is that they can be difficult to understand. They use nonlinear transformations of inputs that are nonintuitive and add complexity. However, studying language at scale necessitates these techniques. Collapsing racial and ethnic descriptors oversimplifies and results in a loss of identity information (inaccuracy of representation). Regarding White and Latinx as separate identities is another oversimplification, because some Latinx individuals identify as being White and others identify as being non-White. We did not study Asian ethnicities because many patients are described by country and language of origin. Thematic groupings do not capture nuanced conceptualizations between words. *Aggression* and *defensiveness*, grouped under *violence*, carry different connotations of culpability and victimization. Copy-and-paste was not addressed in this study[42] because its use still perpetuates unnecessary use of language visible to readers. We acknowledge that many stigmatizing language terms are nonspecific. Further sophisticated NLP studies and validations are needed to delineate which words are more specific to patients and personalities. We did not study to what extent differences in author and patient characteristics contributed to different results across data sets and would be useful for future study. Stigmatizing language use across different self-reported race and ethnicity has been published elsewhere,[10,11] and presence of race descriptors relative to self-reported race is beyond the scope of this study. We did not explore how language changed over time, and it is possible that many of these terms were used differentially during the study period. Each of these represent future areas of research.

Interpretation

Identifying implicit biases in language is possible. Clinical notes are not immune to biases, and the contextual relationships may be as important as the presence of stigmatizing words. Word embeddings provide unique opportunities to quantify these linguistic relationships from electronic health record data sets. LLMs should be used cautiously when incorporated for clinical decision-making because they can transmit subtle inequities found in underlying training data and are reliant on training data characteristics. These have important implications in potential disparities in their use in critical care, but also highlight new sources of disparity in how health care workers document patient encounters.

Funding/Support

J. C. was supported by the UCSF Noyce Initiative for Digital Transformation in Computational Biology & Health, the Hellman Fellows Foundation, UCSF Anesthesia Department Seed Grant, and by the UCSF Claude D. Pepper Older Americans Independence Center funded by the National Institute on Aging, National Institutes of Health [Grant P30 AG044281]. A. L. was supported by the National Institute on Aging, National Institutes of Health [Grants K24AG066998 and R01AG057751]. A. K. Smith was supported by the National Institute on Aging, National Institutes of Health [Grants R01AG057751 and K24AG068312].

Financial/Nonfinancial Disclosures

None declared.

Acknowledgments

Author contributions: J. C., S. J. L., and A. K. S. provided substantial contributions to the conception, design, analysis, and interpretation of data and drafting and final approval of the work. H. M., A. L., E. E., S. Y. J., T. A. H., W. J. B., C. J. K., and O. G. were involved in the conception, design, development of the data pipeline; analysis; programming; statistical methods; and drafting of the manuscript. D. C. A., A. C. C., and T. A. H. specifically provided contributions to the conception, design, interpretation of data and drafting of the manuscript. J. C. agrees to be accountable for all aspects of the work.

Role of sponsors: The funding sources had no role in the design and conduct of the study; collection, management, analysis, or interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Data and code availability statement: Replication code is available at https://github.com/geriatricsresearch/stigmatizing-language.

Other contributions: The authors thank UCSF Information Commons for developing and managing the deidentified data set used for this study.

Additional information: The e-Appendix, e-Figures, and e-Tables are available online under "Supplementary Data."

References

1. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science*. 2017;356(6334):183-186.
2. Bolukbasi T, Chang KW, Zou J, Saligrama V, Kalai A. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings [published online July 21, 2016]. Accessed November 17, 2023. http://arxiv.org/abs/1607.06520
3. Durrheim K, Schuld M, Mafunda M, Mazibuko S. Using word embeddings to investigate cultural biases. *Br J Social Psychol*. 2023;62(1):617-629.
4. Garg N, Schiebinger L, Jurafsky D, Zou J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc Natl Acad Sci U S A*. 2018;115(16):E3635-E3644.
5. Charlesworth TES, Caliskan A, Banaji MR. Historical representations of social groups across 200 years of word embeddings from Google Books. *Proc Natl Acad Sci U S A*. 2022;119(28):e2121798119.
6. Charlesworth TES, Yang V, Mann TC, Kurdi B, Banaji MR. Gender stereotypes in natural language: word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychol Sci*. 2021;32(2):218-240.
7. Balderston JR, Gertz ZM, Seedat R, et al. Differential documentation of race in the first line of the history of present illness. *JAMA Intern Med*. 2021;181(3):386.
8. Broyles LM, Binswanger IA, Jenkins JA, et al. Confronting inadvertent stigma and pejorative language in addiction scholarship: a recognition and response. *Substance Abuse*. 2014;35(3):217-221.
9. FitzGerald C, Hurst S. Implicit bias in healthcare professionals: a systematic review. *BMC Med Ethics*. 2017;18(1):19.
10. Park J, Saha S, Chee B, Taylor J, Beach MC. Physician use of stigmatizing language in patient medical records. *JAMA Netw Open*. 2021;4(7):e2117052.
11. Sun M, Oliwa T, Peek ME, Tung EL. Negative patient descriptors: documenting racial bias in the electronic health record: study examines racial bias in the patient descriptors used in the electronic health record. *Health Affairs*. 2022;41(2):203-211.
12. Himmelstein G, Bates D, Zhou L. Examination of stigmatizing language in the electronic health record. *JAMA Netw Open*. 2022;5(1):e2144967.
13. Kelly JF, Westerhoff CM. Does it matter how we refer to individuals with substance-related conditions? A randomized study of two commonly used terms. *Int J Drug Policy*. 2010;21(3):202-207.
14. Cox C, Fritz Z. Presenting complaint: use of language that disempowers patients. *BMJ*. 2022;377:e066720.
15. Burgess DJ, Crowley-Matoka M, Phelan S, et al. Patient race and physicians' decisions to prescribe opioids for chronic low back pain. *Soc Sci Med*. 2008;67(11):1852-1860.
16. Goddu AP, O'Conor KJ, Lanzkron S, et al. Do words matter? Stigmatizing language and the transmission of bias in the medical record. *J Gen Intern Med*. 2018;33(5):685-691.
17. Green AR, Carney DR, Pallin DJ, et al. Implicit bias among physicians and its prediction of thrombolysis decisions for Black and White patients. *J Gen Intern Med*. 2007;22(9):1231-1238.
18. Penn JA, Newman-Griffis D. Half the picture: word frequencies reveal racial differences in clinical documentation, but not their causes. *AMIA Jt Summits Transl Sci Proc*. 2022;2022:386-395.
19. Langendoen DT. Studies in linguistic analysis. *Language*. 1964;40(2):305.

20. Boleda G. Distributional semantics and linguistic theory. *Annu Rev Linguist*. 2020;6(1):213-234.

21. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3(1):160035.

22. Bird S, Klein E, Loper E. *Natural Language Processing with Python*. 1st ed. O'Reilly; 2009.

23. Řehůřek R, Sojka P. Software framework for topic modelling with large corpora. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA; 2010:45-50.

24. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality [published online October 16, 2013]. Accessed November 17, 2023. http://arxiv.org/abs/1310.4546

25. Garcia-Rudolph A, Saurí J, Cegarra B, Bernabeu Guitart M. Discovering the context of people with disabilities: semantic categorization test and environmental factors mapping of word embeddings from Reddit. *JMIR Med Inform*. 2020;8(11):e17903.

26. Steiner JF, Earnest MA. The language of medication-taking. *Ann Intern Med*. 2000;132(11):926.

27. James G, Witten D, Hastie T, Tibshirani R, eds. *An Introduction to Statistical Learning: With Applications in R*. Springer; 2013.

28. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [published online August 1, 2023]. Accessed November 17, 2023. http://arxiv.org/abs/1706.03762

29. Kay P, Kempton W. What is the Sapir-Whorf hypothesis? *Am Anthropol*. 1984;86(1):65-79.

30. Thierry G. Neurolinguistic relativity: how language flexes human perception and cognition. *Lang Learn*. 2016;66(3):690-713.

31. Ashford RD, Brown AM, McDaniel J, Curtis B. Biased labels: an experimental study of language and stigma among individuals in recovery and health professionals. *Subst Use Misuse*. 2019;54(8):1376-1384.

32. Cooper A, Kanumilli N, Hill J, et al. Language matters. Addressing the use of language in the care of people with diabetes: position statement of the English Advisory Group. *Diabet Med*. 2018;35(12):1630-1634.

33. Healy M, Richard A, Kidia K. How to reduce stigma and bias in clinical communication: a narrative review. *J Gen Intern Med*. 2022;37(10):2533-2540.

34. Dickinson JK, Maryniuk MD. Building therapeutic relationships: choosing words that put people first. *Clin Diabetes*. 2017;35(1):51-54.

35. Browne JL, Ventura A, Mosely K, Speight J. 'I call it the blame and shame disease': a qualitative study about perceptions of social stigma surrounding type 2 diabetes. *BMJ Open*. 2013;3(11):e003384.

36. Adam H, Yang MY, Cato K, et al. 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*. ACM; July 2022:7-21.

37. Hauser DJ, Schwarz N. Implicit bias reflects the company that words keep. *Front Psychol*. 2022;13:871221.

38. Caliskan A, Lewis M. Social biases in word embeddings and their relation to human cognition. In: Dehghani M, Boyd RL, eds. *Handbook of Language Analysis in Psychology*. The Guilford Press; 2020:478-493.

39. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453.

40. Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. *J Glob Health*. 2019;9(2):010318.

41. Shah H. Algorithmic accountability. *Phil Trans R Soc A*. 2018;376(2128):20170362.

42. Liu J, Capurro D, Nguyen A, Verspoor K. "Note bloat" impacts deep learning-based NLP models for clinical prediction tasks. *J Biomed Inform*. 2022;133:104149.