CS 224C Final Project: Artefacts of Bias in Clinical Notes

David F. Castro Peña
dcastrop@stanford.edu

Natalie Dullerud
ndulleru@stanford.edu

Lukas D. Lopez-Jensen
lukaslj@stanford.edu

Abstract

People from marginalized groups often receive inferior healthcare due to human medical bias. This study investigates bias in clinical notes by using sentiments detected within them to predict differential health outcomes. We investigated this topic using the MIMIC-IV-NOTE deidentified free-text clinical notes dataset. Our language transformation was carried out with the NLTK corpus opinion lexicon, a Bag of Words model, BioClinicalBERT, and BioDischargeSummaryBERT. Our analysis included correlation measures such as PCC, Jaccard, and Chi-Squared; Principal Component Analysis; Logistic Regression; Latent Dirichlet Allocation; K-Means Clustering; and Fuzzy Regression Discontinuity. We found that health outcomes and patient groups including race and language were identifiable from attribute-redacted clinical notes; negative sentiment in clinical notes showed a greater association with minoritized groups across race, gender, language, and insurance type; and negative sentiment in clinical notes was associated with worse health outcomes including mortality and pain for minoritized groups in race, gender, and age cohorts.

1 Introduction

Our study examines the widely-documented phenomenon of disparate healthcare access and outcomes. We utilize a vast database and a range of methods to infer differential health outcomes by analyzing textual notes from doctors. Though similar examinations of bias in clinical notes have been carried out by Zhang et al. (2020) and Adam et al. (2022), our work uniquely establishes a correlative relationship between patient demographic factors and diagnostic outcomes. Our study is also novel in that it examines causality in patient identity identifiable through clinical notes and health outcomes.

In Section 2, we provide a brief discussion of related work. Section 3 outlines our methodology, which involved data pre-processing, language transformation, selecting an opinion lexicon, high-dimensional text representation, differential analysis, correlative measures, logistic regression, and fuzzy regression discontinuity. Section 4 details our results, highlighting differential word distributions across group identities, the predictive performance of our Bag-of-Words and BERT embeddings, and pain and health awareness over varying age and gender. We found that health outcomes and patient groups including race and language were identifiable from attribute-redacted clinical notes; negative sentiment in clinical notes showed a greater association with minoritized groups across race, gender, language, and insurance type; and negative sentiment in clinical notes was associated with worse health outcomes including mortality and pain for minoritized groups in race, gender, and age cohorts. We also provide an ethical consideration and authorship statements before listing our references. Several additional figures are outlined in the Appendix.

2 Related Work

Very few existing research works examine textual bias in clinical notes. In Zhang et al. (2020), the authors uncover disparities in machine learning models trained on medical notes from the MIMIC-III dataset via multiple approaches. In particular, they demonstrate differences in the distribution of completion responses from language models in prompting tasks across a variety of identities after fine-tuning BERT on clinical notes. Next, they observe statistically significant gaps in the performance of downstream classification models trained on the BERT text representations of the clinical notes. In general, they find that downstream clinical predictors typically favor the majority group for several attributes, including gender, language, ethnicity, and insurance status. Lastly, the paper outlines a

rudimentary audit of adversarial debiasing strategies that target the obfuscation of protected group information in clinical text embeddings.

Building on the preceding article, Adam et al. (2022) examine characteristics of nurses' clinical notes in MIMIC-IV and Columbia datasets and offer several interesting findings. The authors find that machine learning models can identify self-reported race in patients from clinical notes, even when the text is stripped of explicit indicators of race. Analysis of the textual data and of the language models trained to predict patient race based on clinical notes demonstrates that there is a higher correlation between negative descriptors for Black patients, such as "difficult" and "demanding", in the notes. Thus, this correlation, amongst others, is predictive of patient race. The study shows that models trained on the race-redacted clinical notes to determine treatment paths result in higher errors for Black patients in suggested treatment. Interestingly, through a physician survey, the authors report that humans are unable to identify patient race based on the clinical notes stripped of race indicators.

In our study, we further the investigation by Adam et al. (2022). Our work differs from theirs in that we establish at least a correlative relationship between patient race, gender identity, and diagnostic and mortality outcomes, while they examine race identifiability in clinical notes and learning-facilitated suggested treatment paths. Additionally, our analysis examines causality in patient identity identifiable through clinical notes and health outcomes. We produce a more rigorous exploration of the textual embedding space features of multiple language models that may encode latent patient information. Finally, we look at associations in clinical notes with intersecting identities, such as patient age and gender identity.

3 Methodology

3.1 Data Pre-processing

Owing to computational overhead involved in learning and analysing high-dimensional data over $300,000$ patient records and hospital visits, we randomly sample 10% of the dataset, and proceed with our analysis on our sample. We ran sanity checks to ensure that, at least in relevant identity attributes, our sampled dataset maintained the distribution of the original dataset.

Significant variability exists in the discharge summary notes in the MIMIC-IV dataset. Lab results are often copy-pasted directly into the summary, introducing heterogeneous data, and formatting varies due to the diversity of professionals who compose discharge summaries. Finally, the medical domain uses a plethora of abbreviations not found otherwise in English, and clinical notes are frequently prone to typos due to patient overload. Thus, pre-processing on the clinical notes is crucial to further linguistic analysis in the healthcare domain. As previously used by Alsentzer et al. (2019) in order to fine-tune base BERT on discharge summary notes to create BioDischargeSummaryBERT, we leverage the model "en_core_sci_md" from SciSpacy (Neumann et al., 2019) in order to condense and consistently format our clinical notes. To circumvent the problem of dealing with numeric data through text, we remove all sentences and associated empty section (namely, lab results) that contain non-plain-text numbers, i.e. numeric characters. We then split into three distinct characterizations of the text: (1) directly use the output of "en_core_sci_md", (2) remove newline characters from the output of "en_core_sci_md", and (3) mimic a bag-of-words approach in which all stopwords (characterized by NLTK (Bird et al., 2009)), and words containing non-alpha-numeric characters. In the last case, the linguistic sequential information is largely obfuscated.

Diagnosis labels are often highly specific in MIMIC-IV. For each hospital visit, a patient on average receives 11 diagnosis codes, though not all will pertain to the reason for the visit. MIMIC-IV orders diagnosis codes for each admission from most to least relevant. Due to the great number of highly specific diagnosis codes, we select only the top-3 most relevant diagnoses for a given admission, and then only investigate the 100 most common diagnoses amongst these over all patient admissions in our sample. Thus, our diagnosis information is limited to these 100 diagnoses.

3.2 Language Transformation

After pre-processing the clinical notes, we applied further language transformations in order to operationalize aspects of the text. In particular, we leveraged existing lexicons to determine sentiment in each clinical note, and utilized existing methods that map textual data into the numeric space.

3.2.1 Opinion Lexicon

While there exists an abundance of pre-trained models for sentiment analysis in generic text, clinical text presents a difficulty due to significant difference in word usage distribution. In particular, in our targeting of sentiment analysis, we want to understand descriptors used for the *patient* that relate more to perceived personality and disposition, as opposed to health. Thus, we leveraged the existing opinion lexicon from NLTK (Bird et al., 2009) to identify a list of positive and negative words. We filtered these words by POS tagging, selecting only adjectives, and further refined the list by removing clinically relevant words, classified by the "en_core_sci_md" textual analysis model (Neumann et al., 2019) for biomedical entities. Finally, we removed words not used anywhere in the corpus.

To assess sentiment of each discharge summary, we employed three methods using the filtered opinion lexicon to classify the text: (1) existence of *any* negative (or positive) sentiment words; (2) number of negative vs. positive sentiment words, and (3) existence of the words "difficult" and "demanding", shown in Adam et al. (2022) to be patient descriptors of particular interest w.r.t. racial identity.

3.2.2 High-dimensional text representation

In order to map the text into Euclidean and boolean space, we employ two popular methods: Bag-of-Words vector representation, and frozen BERT embeddings. In the bag-of-words representation, we filter out clinical-specific stop-words (used in over 90% of admissions) and low-frequency words (below 1% of hospital admissions). For fair comparison, the bag-of-words representation is limited to 768 components to match the dimension of the BERT embeddings. The BERT embeddings are obtained from the BioDischargeSummaryBERT model, which is fine-tuned from base BERT on discharge summaries from a previous version of MIMIC. We vary the maximum sequence length of the input tokens in our BERT representations to 256 and 512, and provide corresponding analysis for each chosen cardinality. Tokens are truncated beyond the maximum sequence length.

3.3 Differential Analysis

To understand how clinical notes encode bias, and influence health outcomes, we assess correlation between and predictive index for protected attributes, including race, gender, insurance type, primary language; negative sentiment as measured by our opinion lexicon approaches; and health outcomes, primarily mortality.

3.3.1 Correlative Measures

Using sentiment as labeled in Section 3.2.1, we analyze correlation with race (only White and Black patients due to data availability), gender, insurance (Private vs. Medicaid/Medicare), and language (English or Non-English). We additionally analyze the correlation between sentiment and mortality *conditioned* on each of these attributes. We use the following measures, in addition to simply the proportion of the data with the given covariate (labeled "mean"): normalized mutual information (NMI), Jaccard index, and Matthew's correlation coefficient. We provide χ^2 correlation test results as well. These metrics were chosen due to their utility with binary vectors. However, due to high dataset imbalance (most patients do not die), the measures are not always particularly useful.

3.3.2 Logistic Regression Models on Varying Covariates

We train logistic regression models on the Bag-of-Word vector and the frozen BERT embedding to predict the following binary attributes: race, gender, insurance, language, and mortality. Additionally, we train models on the embeddings, along with diagnosis information, and subsets of the other attributes to predict mortality, as an ablative analysis. We assess the performance via area-under-receiver-operator-curve (AUC), commonly used in medical machine learning due to label imbalance, and accuracy. The performance of the logistic regression models on BERT and BoW representations is intended to demonstrate the predictive capacity of various covariates (including textual information) for the set of labels. The BoW representation models further provide information about words that are more predictive for certain attributes, and covariates, through their learned weights.

3.4 Assessing gender disparities in medical attention drawing from medical notes at the patient level

We further our efforts by correcting the potential confounding effect of comparing different cohorts of patients by age and baseline characteristics. In the following paragraphs, we document our Fuzzy Regression Discontinuity was the most appropriate method for exploring healthcare disparities by

gender using indexes for pain awareness and patient's health. The following analysis draws from our previous findings using MIMIC-IV Deidentified Free-Text Clinical Notes dataset where we found an association between gender, mortality, and notes' sentiments: For middle-aged (46-60 years) and elderly males (>60 years), both living and deceased subgroups show higher instances of negative sentiments compared to their living counterparts. Elderly females, regardless of mortality status, display a high occurrence of negative and neutral sentiments. A Chi-squared test with a statistic of 901 and a P-value of approximately 3.33e-17 confirms a statistically significant association between age, gender, mortality status, and the sentiment of clinical notes, suggesting that these factors significantly influence note sentiment (See Experiment protocol).

3.4.1 First Approach: can we use OLS to assess if detected disparities persist?

Initially, we considered using regression analysis to assess whether gender disparities in medical attention persisted after adjusting for covariates. However, upon analyzing covariate balance using Principal Component Analysis (PCA), we found significant imbalances between men and women. PCA was employed to reduce the dimensionality of the embeddings extracted from patient notes using ClinicalBERT. These embeddings (1 to 10) represent condensed information from the clinical notes, capturing various semantic aspects such as medical conditions, treatments, symptoms, and patient descriptions. Specifically, embeddings 1, 3, 4, 5, 6, 7, 8, 9, 10, and the health index showed statistically significant differences (p-values < 0.05), indicating substantial gender differences in these covariates (See Table 1 in Annex).

The potential for non-compliance was evident, as individuals around the age of 60 might have higher levels of medical notes, suggesting varying degrees of engagement with healthcare services (. This complexity indicated that a more nuanced approach was necessary to address the assignment to treatment (age 60 and above) and the resulting outcomes. Therefore, we opted for a fuzzy Regression Discontinuity Design (RDD), which is better suited to handle such scenarios where the treatment assignment is not perfectly adhered to and allows for a more accurate estimation of the causal effect.

3.4.2 Second Approach: using pain index and health for a Fuzzy Regression Discontinuity

Given the significant covariate imbalances between men and women, particularly in the embeddings and health index derived from patient notes, we decided to use fuzzy Regression Discontinuity Design (FRDD) to analyze whether older women receive differential pain awareness relative to their male peers. FRDD focuses on individuals near the threshold age, thus naturally controlling for unobserved heterogeneity by comparing individuals who are similar in many respects except for their treatment assignment determined by age. This method mitigates the impact of covariate imbalance because it leverages the quasi-random assignment near the cutoff, reducing the confounding effects present in the data. By focusing on local comparisons around the age threshold, FRDD provides a more robust and credible estimate of the treatment effect, isolating the impact of age and gender on pain awareness while accounting for inherent differences in the underlying covariates.

An essential part to build our FRDD was obtaining the pain index and health index. The pain index was developed using a pre-trained ClinicalBERT model to analyze the sentiment expressed in the text, focusing specifically on mentions of pain and related terms. Each mention was quantified based on the intensity and context, providing a composite score that reflects the perceived severity of pain as documented by healthcare providers. Conversely, the health index was formulated by extracting key health metrics such as BMI, blood pressure, weight, and height from the clinical notes. This involved parsing the text for specific health-related terms and their numeric values, which were then standardized and combined into a single composite score. This score represents an overall health status, capturing both the presence and severity of underlying health conditions that could influence medical treatment and documentation practices. By integrating these indices, the study offers a nuanced view of how health status and pain perception are represented in clinical records, aiding in the investigation of potential biases in medical documentation across different patient groups. The following graphs provide an intuition of the distribution of indexes:

The pain index distribution, derived from the frequency of pain-related keywords in clinical notes, typically exhibits right skewness, indicating that

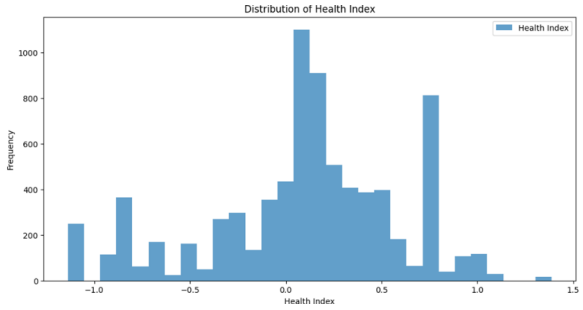

Figure 1: Distribution of Health Index

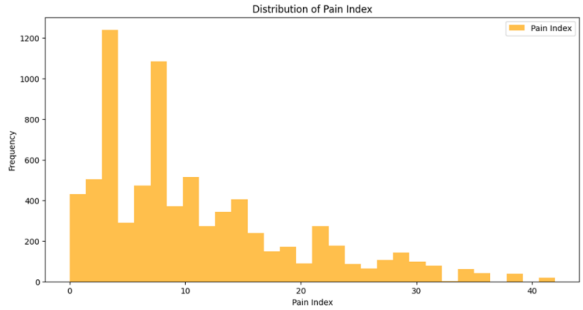

Figure 2: Distribution of Pain Index

most patients report lower pain levels, with fewer reporting high pain levels. Peaks in the distribution may reveal common pain levels among patients, while outliers could highlight individuals with exceptionally high pain experiences. In contrast, the health index distribution, a composite measure of overall health status from various health metrics, often approximates a normal distribution, reflecting a balanced health status across the population. However, it can also exhibit skewness depending on the population's health conditions, with peaks indicating prevalent health levels and outliers representing patients with particularly good or poor health.

The identification strategy for the Fuzzy RDD hinges on age as the running variable, positing that biases in pain documentation become more pronounced as patients cross the threshold of 60 years. This approach allows us to infer causal relationships by comparing the documentation practices just below and above this age threshold, controlling for other covariates captured in the health index. The Fuzzy Regression Discontinuity Design (RDD) equation tailored for your health and pain index study involves two stages.

First Stage (Instrumental Variable Stage): Estimate the probability of receiving the treatment based on the running variable (e.g., age) and the threshold:

$$D_i = \gamma_0 + \gamma_1 Threshold_i + \gamma_2(X_i - c) + \gamma_3 Threshold_i \cdot (X_i - c) + \epsilon_i$$

Where:
D_i is an indicator variable for receiving the treatment (e.g., higher level of medical notes)
$Threshold_i$ is an indicator variable for being above the threshold (1 if $X_i \geq C$, 0 otherwise)
X_i is a running variable (e.g., age)
c is a threshold (e.g., age 60)
ϵ_i is the error term

Second Stage: Use the predicted treatment probability from the first stage to estimate the causal effect on the outcome (e.g., pain index).

$$Y_i = \beta_0 + \beta_1 \hat{D}_i + \beta_2(X_i - c) + \beta_3 Threshold_i \cdot (X_i - c) + \nu_i$$

Where:
Y_i is the outcome variable (e.g., pain index)
$\hat{D}_i$ is the predicted treatment probability from the first stage
ν_i is the error term

4 Results

4.1 Correlation and frequency-based measures over clinical notes: sentiment, group identity and mortality

For each method measuring clinical notes sentiment, explained in Section 3.2.1, we quantify the relationship between sentiment, group identity and health outcomes through a diverse catalogue of correlation measures (discussed in previous Section 3.3.1). As shown in Figures in the Appendix, negative sentiment in clinical notes is proportionally higher in the minoritized group for *all* methods of measuring negative sentiment. Interestingly, the results differ slightly for measures of negative sentiment involving normalized death incidence across groups (see race) but generally, the results show that negative sentiment in clinical notes is associated with higher mortality in the minoritized group across several identities, even when normalizing for differing base rate of death in different groups. To the greatest extent, non-English speakers appear to experience death at significantly higher rates than non-English speakers when they are described negatively by providers in clinical notes, even when normalizing for differing base rates. This could be due to misunderstandings (e.g. being

"difficult") that result in substandard care for non-English speakers. Similarly, insurance type appears to show a milder trend. We hypothesize that this is possibly due to severity of disease in patients with worse insurance – i.e. a patient is more likely to postpone going to the hospital, or only go in sever cases, if a patient knows they have poor insurance coverage. Further and more detailed causal analysis is needed to understand both of these trends.

Correlative measures are rather low in magnitude, thus we get a weak signal in correlation between negative sentiment and mortality amongst any group. We hypothesize that this is largely due to the imbalance in the dataset with respect to mortality – fewer than 3% of the admissions result in mortality.

4.2 Differential word distribution across group identities

We examined the distributions of word usage in clinical notes for each group identity – race, gender, language and insurance – and examined the differential in word appearance distribution in the discharge summaries between majoritized and minoritized groups. The plots in Appendix display the top 15 biggest word differentials over all words used and top 15 biggest word differentials over sentiment specific words, for each group. We additionally conditioned on presence of fatality, and analogously investigated the distribution differentials (see Appendix). In each plot, the direction of differential is positive for the majoritized group and negative for the minoritized group (i.e. negative values mean the word is used *more* in the minoritized group). We note a few generally interesting trends, specifically, in the sentiment word differential distribution, in Figures, the minoritized group nearly always sees significantly higher use of negative words (particularly *difficult*), while the majoritized group see higher use of positive sentiment words. However, conditioning on fatality does not appear to increase the trend by a substantial in gender. The trend does increase precipitously in language, for some words in particular in race, and mildly in insurance, which mirrors the results from the correlation studies. Thus, we conclude that there is a differential in word usage in discharge summaries over race, gender, language and insurance, but these only seem to impact health outcomes significantly for non-English speakers and those on publicly funded insurance. The differential distribution on *all* terms largely demonstrates shifts in morbidities between , and does not change significantly for any . Nonetheless, some of these differentials are clinically important from a healthcare bias perspective – for instance, “mild" is used in clinical notes significantly more for non-English speaking patients, begging the question whether pain (or other conditions) is underestimated in non-English speakers.

4.3 Predictive performance of Bag-of-Words and BERT embeddings

The results of the logistic regression models trained variably on BoW and varying BERT embeddings are shown in Appendix. The results demonstrate relatively high AUC (better than random) on the majority of labels, including race, which is in opposition of what MIMIC-IV claims, which is that racial identity is not identifiable from the clinical notes. Some labels, such as gender, are directly inferable from the data, as sex is explicitly mentioned in the clinical notes. Surprisingly, the BoW model in general does just as well as the BERT embedding model for most predictions, which makes sense as the clinical notes are highly diverse. Despite BioDischargeSummaryBERT being trained on discharge summary notes from MIMIC-III, MIMIC-III discharge notes are significantly different, in content and in length than MIMIC-IV, which are much longer and more varied in media (includes lab notes and numeric data, as well as a parting message for the patient, etc.). Additionally, due to limits on the input dimension of BERT models, the input sequence is truncated to at most 512 tokens, possibly limiting the utility of the embeddings, as the average sequence length of the clinical notes is over 1000. We also see that mortality prediction is enhanced by protected attributes, but predominantly by the diagnosis information we gathered (see Tables in Appendix).

4.4 Pain and Health Awareness over Varying Age and Gender

Results from the Fuzzy Regression Discontinuity for males and females (Tables 2 and 3) show significant differences in how the health index impacts the awareness of medical practitioners to patients' pain, as reflected in the clinical notes. The overall analysis revealed that the health index significantly reduces the pain awareness index for both genders, but the effect is more pronounced in males. For females, the pain awareness index decreases by approximately 5.31 units with each unit increase in

the health index (coefficient = -5.3074, $p < 0.001$), explaining 10.7% of the variance (R-squared = 0.107). In contrast, for males, the pain awareness index decreases by about 7.57 units with each unit increase in the health index (coefficient = -7.5659, $p < 0.001$), explaining 24.1% of the variance (R-squared = 0.241). These findings, illustrated in Graphs 1 and 2, suggest that health improvements have a stronger impact on reducing the recorded pain awareness among medical practitioners for males compared to females.

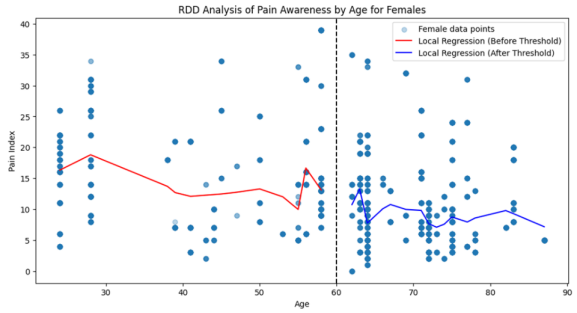

Figure 3: Fuzzy RD for Women: Pain Index Vs. Age

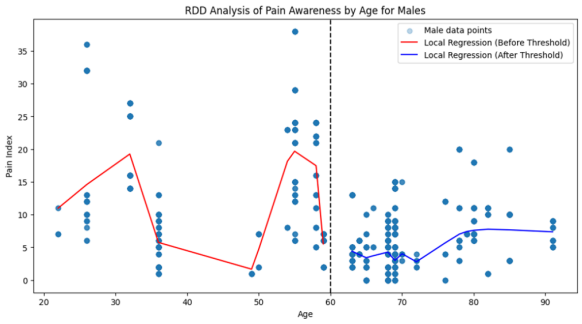

Figure 4: Fuzzy RD for Men: Pain Index Vs. Age

Subgroup analyses further elucidate these findings. For patients below the age of 60, a higher health index is associated with a significant increase in the pain awareness index for both genders. For males, the pain awareness index increases by approximately 3.87 units with each unit increase in the health index (coefficient = 3.8710, $p < 0.001$), explaining 4.6% of the variance (R-squared = 0.046).

For females, the health index is not significantly related to the pain awareness index in the younger cohort (see Table 2 and Graph 3). For patients above the age of 60, the relationship between health index and pain awareness index is marginally significant for males (coefficient = 0.6428, $p = 0.003$) and insignificant for females (see Table 3 and

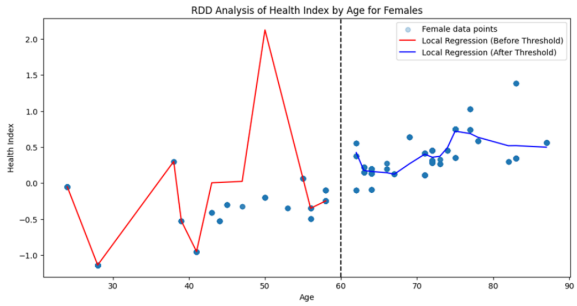

Figure 5: Fuzzy RD for Women: Health Index Vs. Age

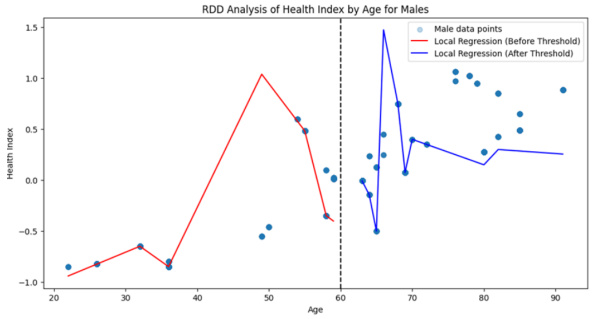

Figure 6: Fuzzy RD for Men: Health Index Vs. Age

Graph 4). These results suggest that older patients might report slightly higher pain levels despite better health due to chronic conditions. The visualizations in Graphs 5 and 6 further support these findings by showing the local regression lines for pain and health indexes across different age groups and genders.

Negative Sentiment in Clinical Notes We observe that negative sentiment, and negative patient descriptors are observed more frequently in clinical notes for minoritized groups than majoritized ones. In our analysis, we demonstrate substantive evidence for bias of medical practitioners in their textual treatment of minoritized groups with respect to racial identity, gender, type of insurance and primary language and provide correlative and predictive analysis that shows minoritized groups experience worse health outcomes when negative sentiment is present in their clinical notes. This is extremely important for further investigation, in order to determine additional routes to train physicians to mitigate bias and prevent worsening healthcare gaps and worse care for already-marginalized groups. Further work needs to be done to tease out a causal relationship between physician treatment (outside of discharge summaries), physician perception of the patient, and health outcomes, in order to further improve healthcare for marginal-

ized populations. We invite additional rigorous research to improve on our patient descriptor sentiment analysis, as ours was relatively rudimentary. We also would like to see future work target health outcomes outside of mortality. Complex and varied diagnoses are challenging to deal with in the MIMIC dataset and in prediction tasks but we urge researchers in this topic to examine diagnostic outcomes through a social and clinical bias lens.

Predictive Ability of Simple Models on Clinical Notes We demonstrate that racial identity, despite claims to the contrary by MIMIC-IV, is identifiable from clinical notes, and *is* useful, along with textual information, in mortality prediction. We encourage future work to iterate on our work through more sophisticated analysis of predictiveness of our studied covariates for certain health outcomes, along with clinical notes and sentiment analyses. In the future, we would use higher n-gram model instead of solely unigram (BoW) model, and try multi-label diagnosis prediction, which presents significant difficulty. In addition, we would like to see treatment trajectories considered and how these are impacted.

Pain Awareness and Health Index Surprisingly, the Fuzzy Regression Discontinuity analysis indicates that becoming older reduces the pain awareness perceived by medical practitioners, as evidenced by the clinical notes. This finding is counter-intuitive since senility typically increases health complications, which would logically lead to more frequent mentions of pain-related terms in medical documentation. However, the results show that for both males and females, pain awareness decreases with better health, particularly more so for males. Specifically, the pain awareness index decreases by 7.57 units for males and 5.31 units for females per unit increase in the health index. These findings suggest that medical practitioners might be less attentive to documenting pain in older adults, possibly because they prioritize other health issues they perceive as more critical in this demographic.

The implications of these findings are significant. They suggest that medical practitioners might be subconsciously reducing the expected life expectancy of older patients, leading to a possible under-documentation of pain. This under-documentation could stem from a bias where practitioners might not expect older adults to benefit as much from pain management interventions. Consequently, older patients might receive less attention for pain-related issues, which starts with the lack of acknowledgment in medical notes. This trend necessitates a closer examination of documentation practices to ensure that pain management is equitably prioritized across all age groups, especially considering the higher risk of chronic pain in older adults. These results underscore the importance of training and awareness programs for medical practitioners to mitigate biases and ensure comprehensive care for all patients.

5 Ethical Consideration

5.1 Ethical Implication of our Project

We recognize that our paper utilizes data containing personal protected information, and we have taken measures (outlined below) to guarantee that we use this information properly and do not enable malicious actors to abuse it. We hope our findings may galvanize those in the medical industry to incorporate sentiment analysis into their clinical notes and to take measures towards diminishing racial, sexual, and all other types of bias in clinical notes, ensuring that every patient receives the care they need.

5.2 Use of Data

When working with actual medical data as was done for this study, it is crucial to protect patient privacy and to acknowledge the sensitive nature of the information utilized. Although our dataset has been thoroughly anonymized by its original curators, malicious parties may be able to infer patient identities with the information preserved in the dataset. To protect patient privacy, our team has decided not to attach our dataset to this paper. We provide citations and a link to our source, but the dataset cannot be downloaded without completing related training on PhysioNet, the website that provides it. To acknowledge the sensitive nature of our data, each member of our team has completed online training provided by PhysioNet, informing us on how to properly handle sensitive data and identify conflicts of interest in our work.

5.3 Potential Applications of our Work

Our results can be used to justify the integration of sentiment analysis into electronic health records systems. If the sentiment of clinical notes is tracked in real time, an automated alert system could flag negative sentiment and prompt the author to rephrase their notes to characterize the situation more neutrally, potentially affording the patient

fairer care in the future. Our results may also motivate further anti-bias and implicit bias training in the medical field. Providing evidence that negative sentiment in clinical notes is correlated with adverse health outcomes may compel new entrants to the medical field to write clinical notes from the most objective perspective possible.

6 Authorship Statement

David F. Castro Peña authored the components of the experiments regarding pain and health index and analyses on age and gender attributes.
Natalie Dullerud authored the experiments and associated sections regarding sentiment, mortality and protected attribute information, in addition to predictive models.
Lukas D. Lopez-Jensen authored this paper's Abstract, Introduction, and Ethical Consideration, and adapted the Related Work section from a jointly-authored literature review.

References

Hammaad Adam, Ming Ying Yang, Kenrick Cato, Ioana Baldini, Charles Senteio, Leo Anthony Celi, Jiaming Zeng, Moninder Singh, and Marzyeh Ghassemi. 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, page 7–21, New York, NY, USA. Association for Computing Machinery.

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical bert embeddings. *Preprint*, arXiv:1904.03323.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics.

Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: Quantifying biases in clinical contextual word embeddings. *Preprint*, arXiv:2003.11515.

A Additional Experimental Results and Figures

A.1 Correlative Measures between Negative Sentiment, Health Outcomes and Group Identity

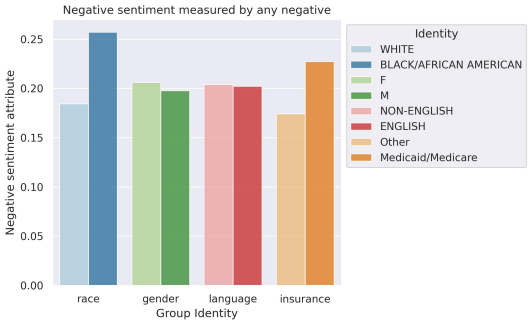

Figure 7: Proportion of group (for each group identity) with negative sentiment determined by presence of any negative word in discharge summary.

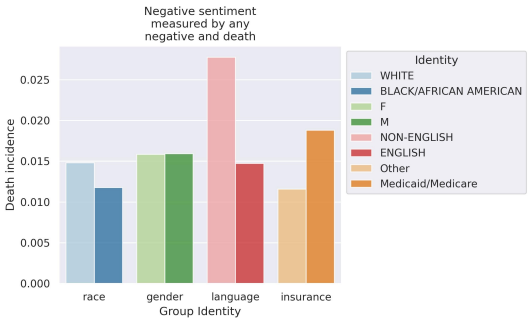

Figure 8: Death incidence in group (for each group identity) with negative sentiment determined by presence of any negative word in discharge summary.

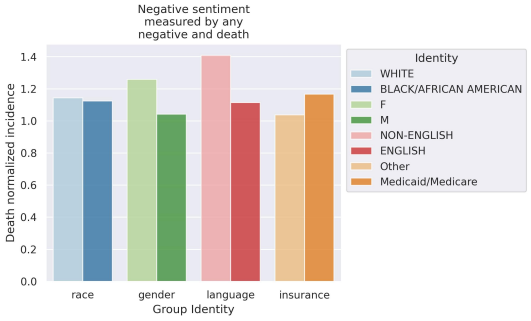

Figure 9: Normalized death incidence in group (for each group identity) with negative sentiment determined by presence of any negative word in discharge summary.

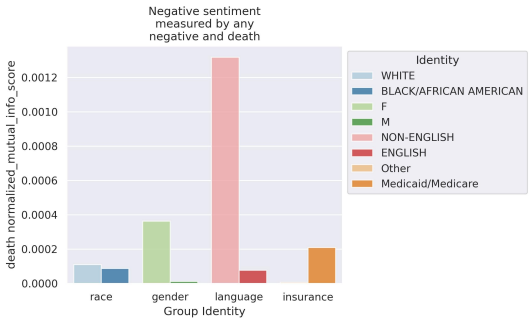

Figure 10: Normalized information score (NMI) in group (for each group identity) of negative sentiment and death determined by presence of any negative word in discharge summary.

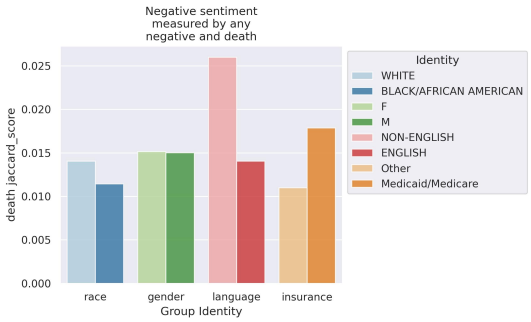

Figure 11: Jaccard index in group (for each group identity) of negative sentiment and death determined by presence of any negative word in discharge summary.

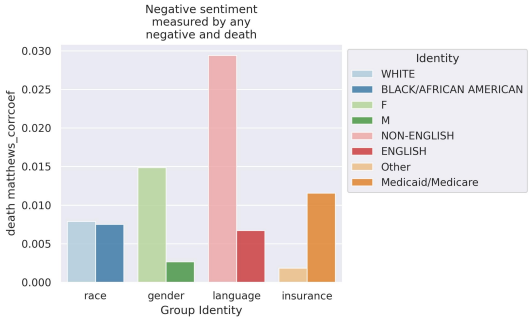

Figure 12: Matthew's correlation coefficient in group (for each group identity) of negative sentiment and death determined by presence of any negative word in discharge summary.

A.2 Differential Word Distributions in Clinical Notes per Group Identity

A.2.1 Sentiment Word Distributions

A.2.2 All Valid Word Distributions

A.3 t-SNE dimensionality reduction of BERT embeddings

B Predictive Performance Results

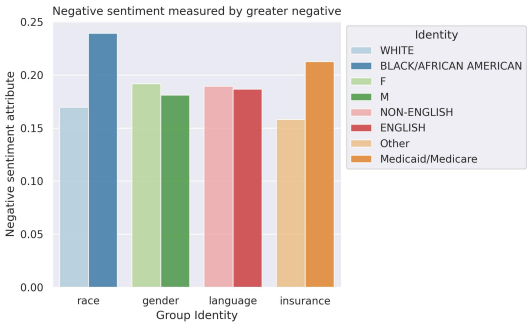

Figure 13: Proportion of group (for each group identity) with negative sentiment determined by greater number of negative than positive words in discharge summary.

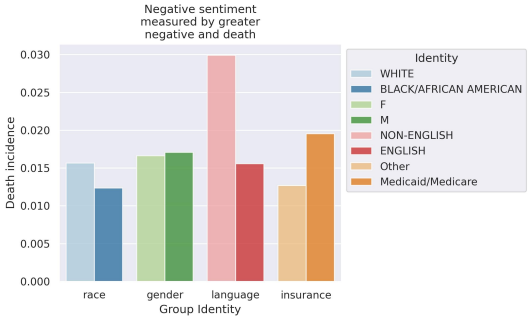

Figure 14: Death incidence in group (for each group identity) with negative sentiment determined by greater number of negative than positive words in discharge summary.

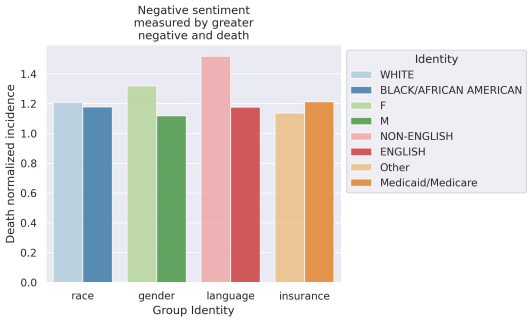

Figure 15

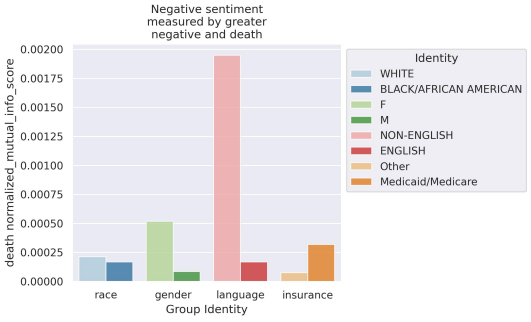

Figure 16

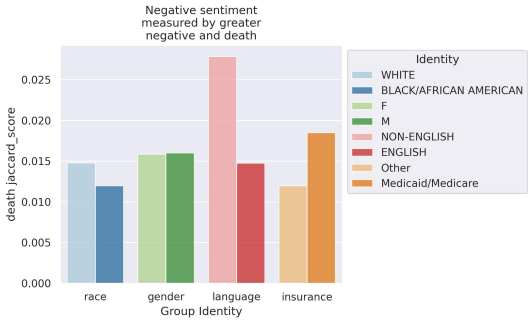

Figure 17

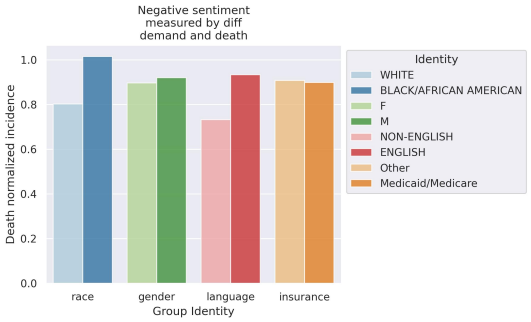

Figure 21

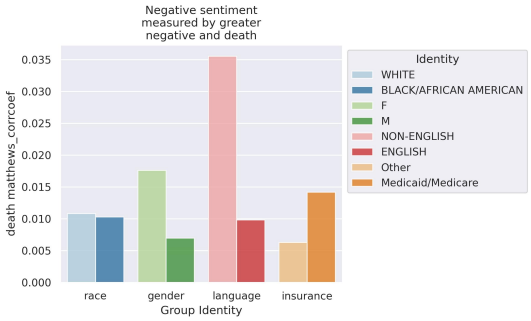

Figure 18

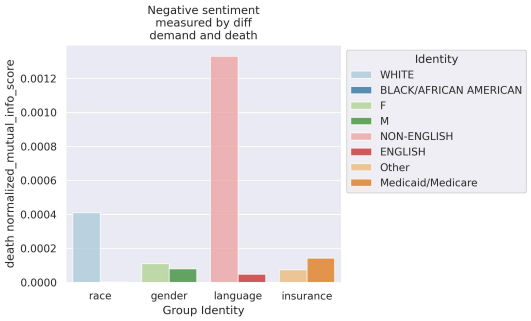

Figure 22

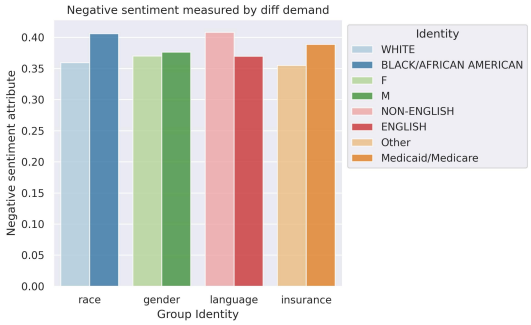

Figure 19

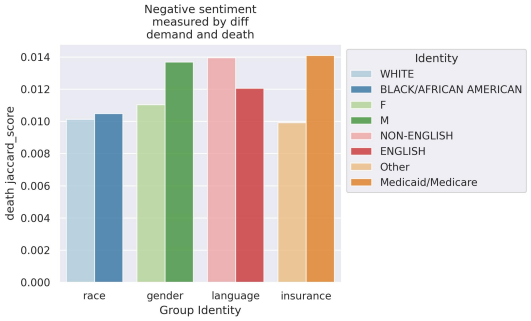

Figure 23

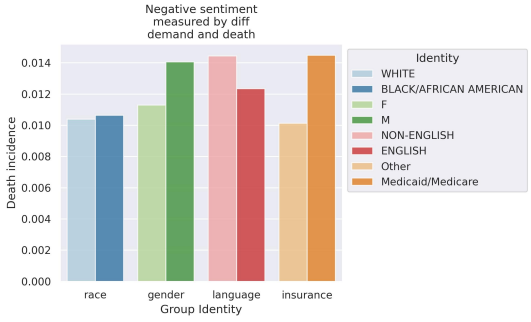

Figure 20

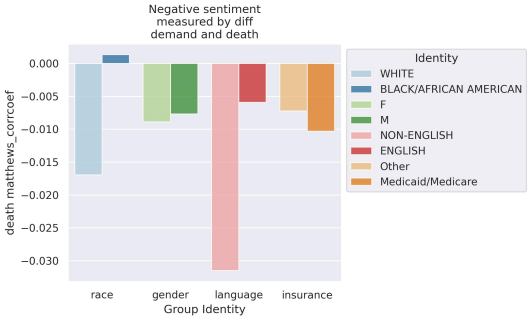

Figure 24

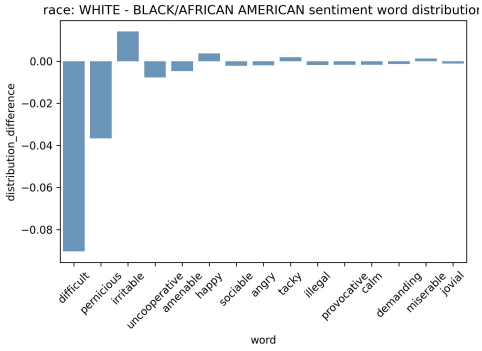

Figure 25

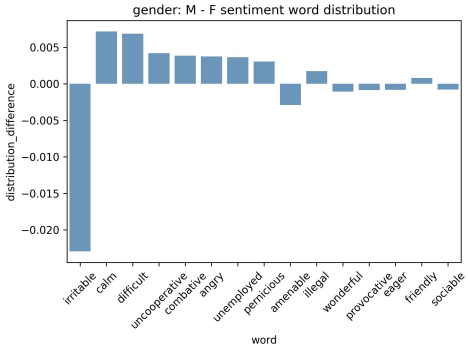

Figure 26

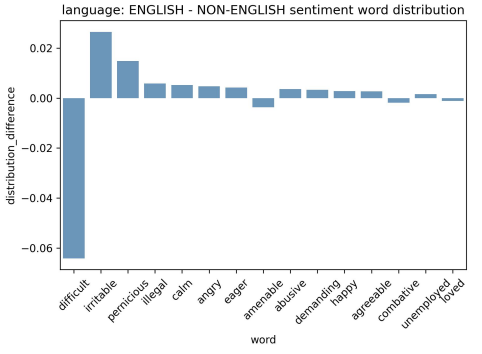

Figure 27

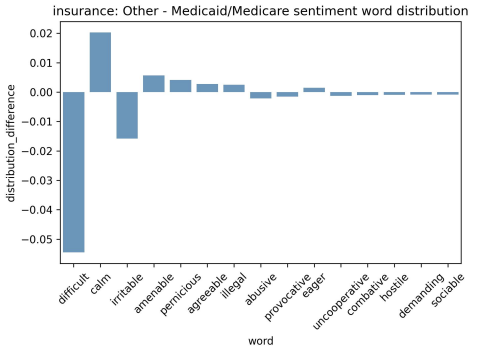

Figure 28

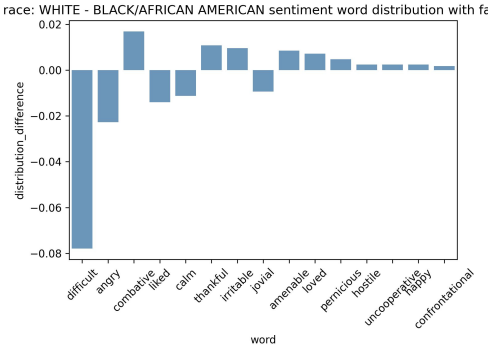

Figure 29

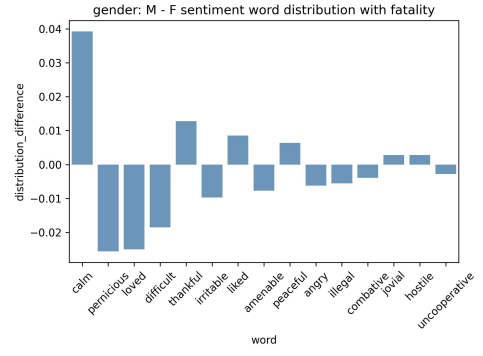

Figure 30

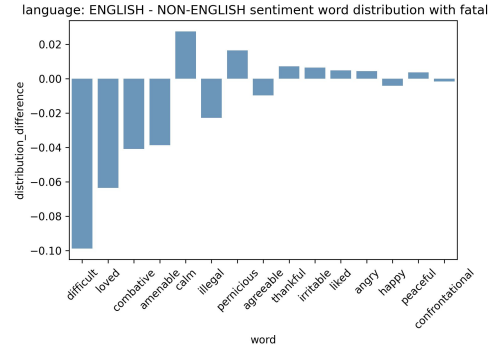

Figure 31

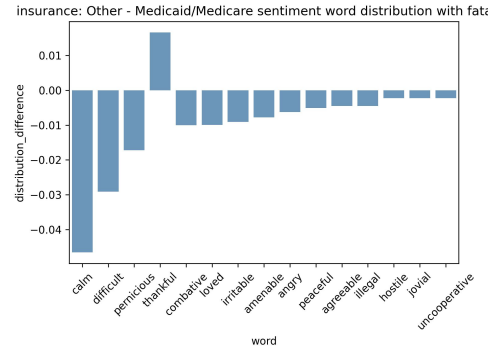

Figure 32

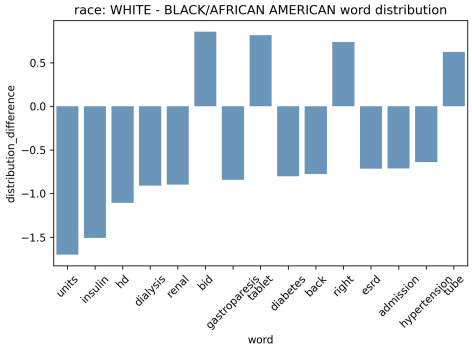

Figure 33

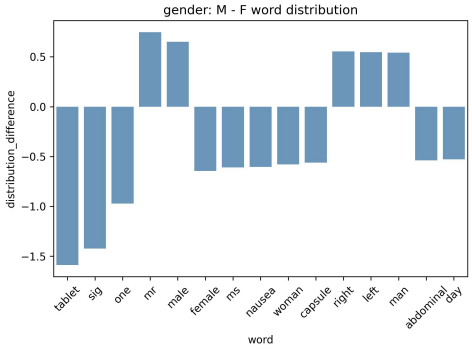

Figure 34

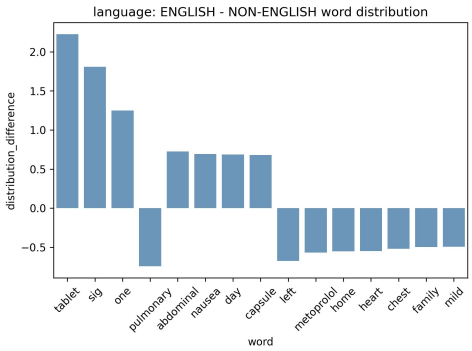

Figure 35

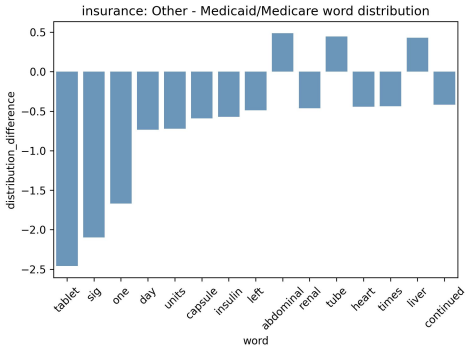

Figure 36

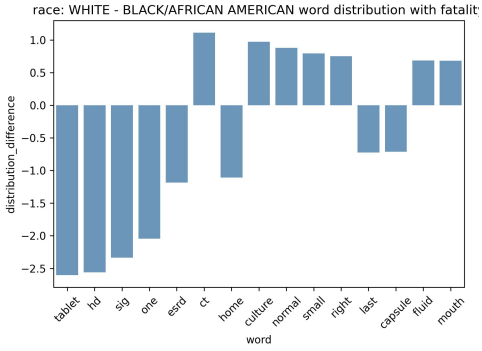

Figure 37

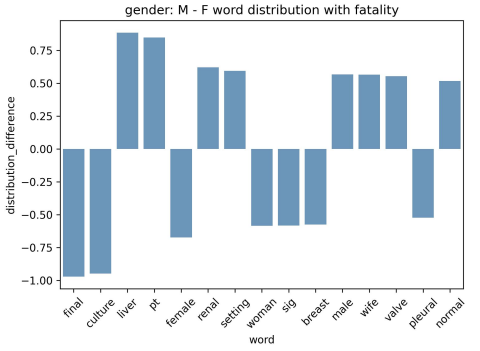

Figure 38

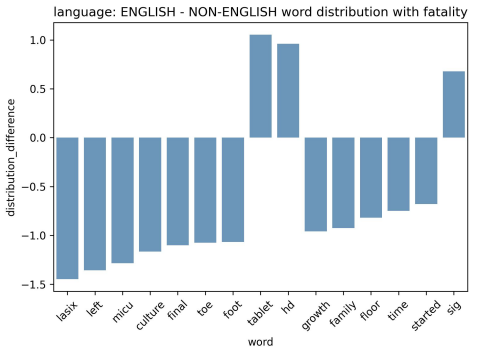

Figure 39

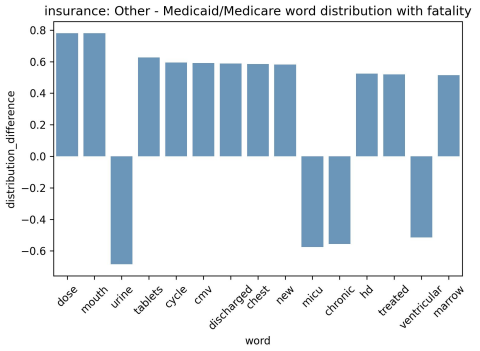

Figure 40

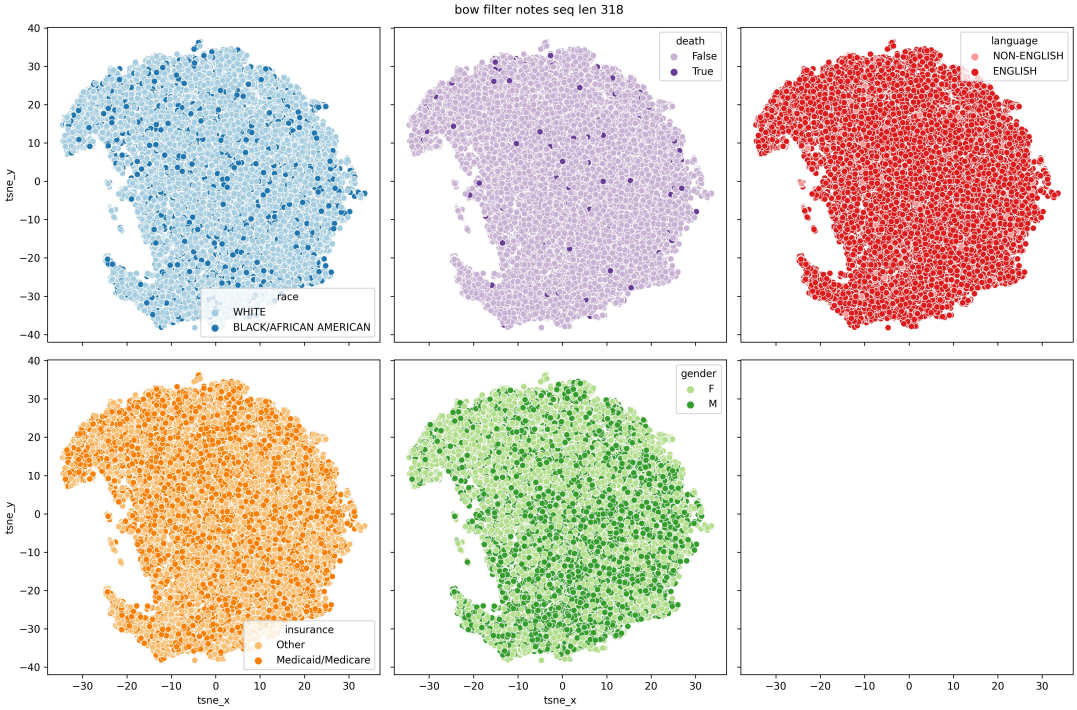

Figure 41

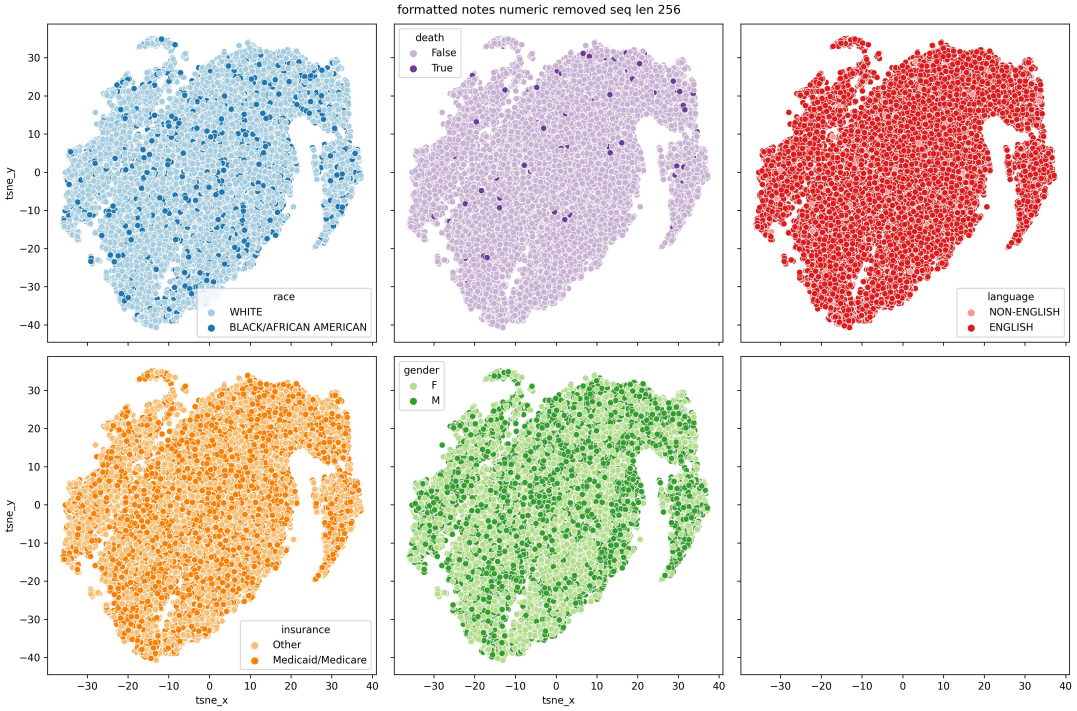

Figure 42

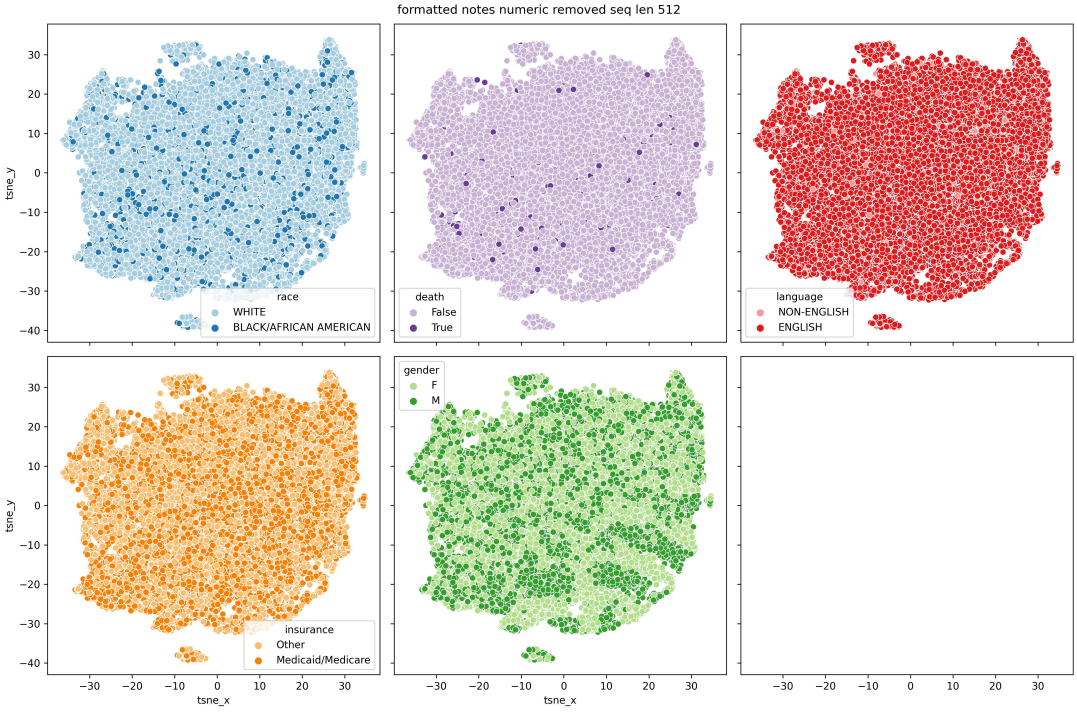

Figure 43

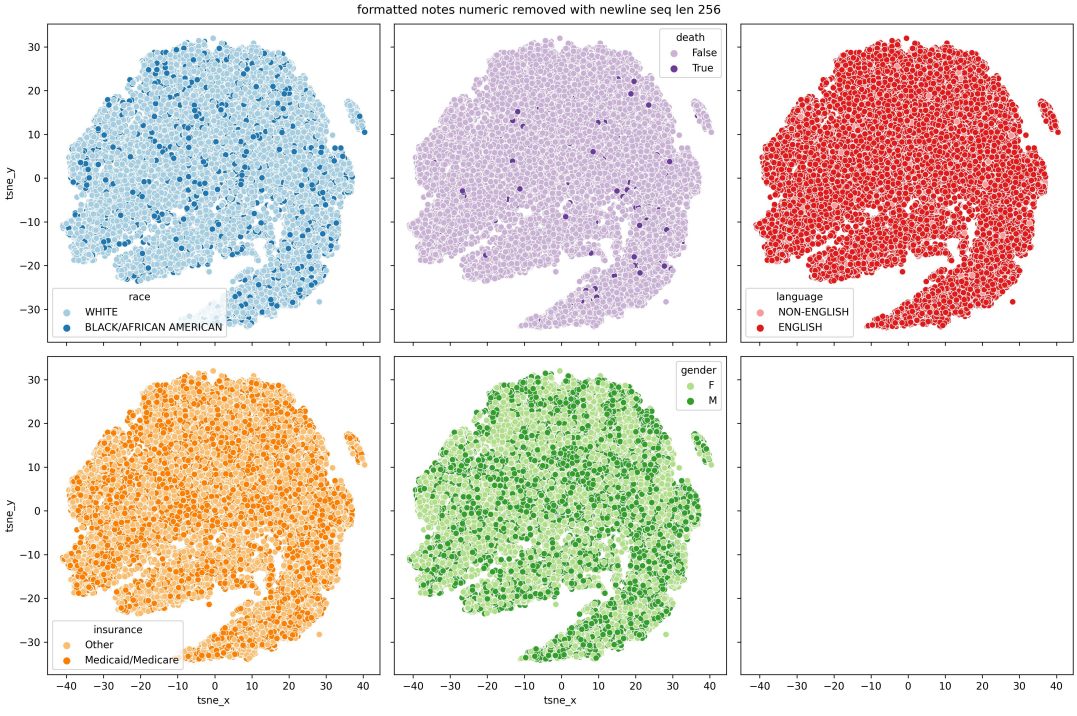

Figure 44

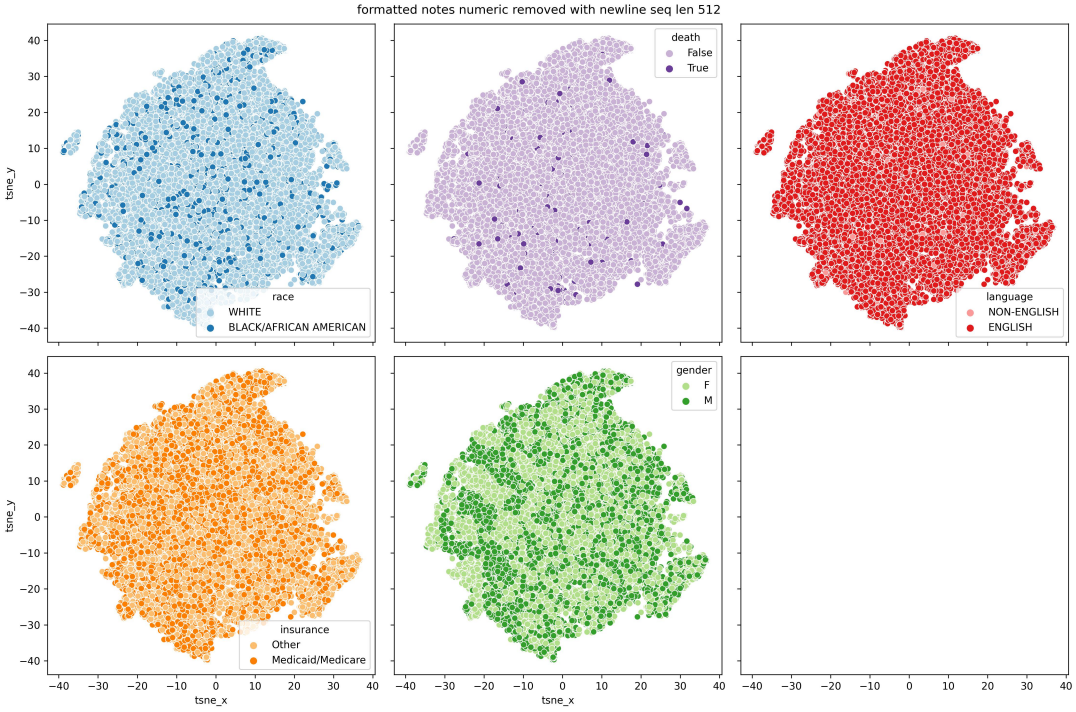
formatted notes numeric removed with newline seq len 512
race
WHITE
BLACK/AFRICAN AMERICAN
death
False
True
language
NON-ENGLISH
ENGLISH
insurance
Other
Medicaid/Medicare
gender
F
M
tsne_x
tsne_y

Figure 45

[h!]

text filter	remove newline	max seq length	normalize	label	AUC
BoW	True	NaN	True	race	0.712
BoW	True	NaN	False	race	0.651
BERT	True	256.000	True	race	0.687
BERT	True	256.000	False	race	0.611
BERT	False	256.000	True	race	0.687
BERT	False	256.000	False	race	0.611
BERT	True	512.000	True	race	0.694
BERT	True	512.000	False	race	0.635
BERT	False	512.000	True	race	0.694
BERT	False	512.000	False	race	0.635

Table 1

[h!]

text filter	remove newline	max seq length	normalize	label	AUC
BoW	True	NaN	True	gender	0.992
BoW	True	NaN	False	gender	0.964
BERT	True	256.000	True	gender	0.990
BERT	True	256.000	False	gender	0.974
BERT	False	256.000	True	gender	0.990
BERT	False	256.000	False	gender	0.974
BERT	True	512.000	True	gender	0.992
BERT	True	512.000	False	gender	0.977
BERT	False	512.000	True	gender	0.992
BERT	False	512.000	False	gender	0.977

Table 2

[h!]

text filter	remove newline	max seq length	normalize	label	AUC
BoW	True	NaN	True	language	0.528
BoW	True	NaN	False	language	0.500
BERT	True	256.000	True	language	0.527
BERT	True	256.000	False	language	0.503
BERT	False	256.000	True	language	0.527
BERT	False	256.000	False	language	0.503
BERT	True	512.000	True	language	0.535
BERT	True	512.000	False	language	0.502
BERT	False	512.000	True	language	0.535
BERT	False	512.000	False	language	0.502

Table 3

[h!]

text filter	remove newline	max seq length	normalize	la
BoW	True	NaN	True	insu
BoW	True	NaN	False	insu
BERT	True	256.000	True	insu
BERT	True	256.000	False	insu
BERT	False	256.000	True	insu
BERT	False	256.000	False	insu
BERT	True	512.000	True	insu
BERT	True	512.000	False	insu
BERT	False	512.000	True	insu
BERT	False	512.000	False	insu

Table 4