MIT Open Access Articles

Write It Like You See It: Detectable Differences in Clinical Notes By Race Lead To Differential Model Recommendations

Massachusetts Institute of Technology

Write It Like You See It: Detectable Differences in Clinical Notes by Race Lead To Differential Model Recommendations

Hammaad Adam
hadam@mit.edu
Massachusetts Institute of Technology
Cambridge, MA, USA

Ming Ying Yang
ming1022@mit.edu
Massachusetts Institute of Technology
Cambridge, MA, USA

Kenrick Cato
kdc2110@cumc.columbia.edu
Columbia University
New York, NY, USA

Ioana Baldini
ioana@us.ibm.com
IBM Research
Yorktown Heights, NY, USA

Charles Senteio
charles.senteio@rutgers.edu
Rutgers University
New Brunswick, NJ, USA

Leo Anthony Celi
lceli@mit.edu
Massachusetts Institute of
Technology, Beth Israel Deaconess
Medical Center & Harvard University
Cambridge, MA, USA

Jiaming Zeng
jiaming@ibm.com
IBM Research
Cambridge, MA, USA

Moninder Singh
moninder@us.ibm.com
IBM Research
Yorktown Heights, NY, USA

Marzyeh Ghassemi
mghassem@mit.edu
Massachusetts Institute of Technology
Cambridge, MA, USA
Vector Institute
Toronto, Ontario, Canada

ABSTRACT

Clinical notes are becoming an increasingly important data source for machine learning (ML) applications in healthcare. Prior research has shown that deploying ML models can perpetuate existing biases against racial minorities, as bias can be implicitly embedded in data. In this study, we investigate the level of implicit race information available to ML models and human experts and the implications of model-detectable differences in clinical notes. Our work makes three key contributions. First, we find that models can identify patient self-reported race from clinical notes even when the notes are stripped of explicit indicators of race. Second, we determine that human experts are not able to accurately predict patient race from the same redacted clinical notes. Finally, we demonstrate the potential harm of this implicit information in a simulation study, and show that models trained on these race-redacted clinical notes can still perpetuate existing biases in clinical treatment decisions.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; **Machine learning**; • **Applied computing → Health informatics**.

KEYWORDS

Clinical notes, health equity, natural language processing

1 INTRODUCTION

There are a number of well-established inequities in hospital-based healthcare delivery that affect patients from racial minority groups. Minority patients have worse hospital outcomes across various medical conditions [27], including congestive heart failure [4], myocardial infarction [53], and perinatal care [29], as well as for various surgical procedures [52]. Minority patients tend to receive care from different physicians than non minority patients; the physicians they see have less clinical training [53], are less likely to be board certified, and are more likely to report that they are unable to provide high-quality care to all their patients [8]. These structural factors along with physician implicit biases create inequitable treatment decisions [25]; for example, physicians are less likely to provide Black patients with analgesia for acute pain in the emergency room [33] or thrombolysis for acute coronary syndromes [25].

These disparities are coming into increasing focus as the use of machine learning (ML) and artificial intelligence (AI) proliferates in healthcare. Deploying ML models in clinical settings has been proposed to improve diagnostic accuracy, treatment decisions, and operational efficiency [55]. However, the use of models also risks replicating and exacerbating implicit biases present in the data used to train them. Prior research found that health systems

were far less likely to refer Black patients to high-risk care management programs than similar White patients, because they relied on algorithms that used healthcare costs as a proxy for health [42]. Similar inequities have been described when ML models have been deployed in other high-stakes domains, including criminal justice [6] and financial lending [39].

Notably, ML models can perpetuate existing biases even if they do not have explicit access to race. Models that were less likely to recommend Black patients to high-risk care management programs [42], more likely to identify Black defendants as high risk [6], and less likely to approve Black mortgage applicants [39] all did not explicitly use race as a variable in making their predictions. However, the models were able to infer race from other correlated attributes, and could thus propagate existing human biases in these decisions. To construct fair models, it is thus vital for machine learning practitioners to understand and account for the implicit racial information present in the data they train their models on. Such information is not always obvious, and can exist in data sources that seem race-redacted to humans [9]. However, little work has focused on investigating the level of race information contained in clinical notes [13], despite these being an increasingly common source of data for ML models [13, 20, 34, 63].

In this study, we investigate the presence of implicit racial information in clinical nursing notes that have had direct race information redacted. Using data from two large hospitals, we determine that models are able to accurately predict a patient's self-reported race from notes written during their hospital stay, even after explicit indicators of race are removed. We conduct a detailed audit of the drivers of predictive accuracy to identify potential disparities in clinical care, and find that while some variations are clinically justifiable (e.g. comorbidities that are more common in Black patients), others may reflect potential avenues of missed care (e.g. references to bruising and rashes being extremely predictive of White race, even though there is no clinical reason for these symptoms to be less common in Black patients). Notably, human experts do not share the ability to identify race: a group of 42 surveyed physicians were unable to accurately determine patient race using the same redacted clinical notes. While the ability of models to predict race is not intrinsically harmful, it implies that ML models trained on clinical notes have access to patient race, even if this information is not explicitly provided and undetectable by clinicians. We illustrate the potential harm of this information in a simulation study, demonstrating that models trained on these race-redacted clinical notes perpetuate existing biases in clinical treatment decisions.

2 RELATED WORK

We discuss three categories of prior research that are directly related to and impacted by our findings: studies that attempt to predict race from clinical data, audits of racial biases in clinical notes, and clinical prediction tasks.

Race Prediction From Clinical Data. Relatively little research has focused on characterizing the implicit racial information present in clinical data by attempting to predict patient race. One such study focused on medical images, and revealed that deep learning algorithms were able to accurately predict self-reported patient race exclusively from chest X-rays, though expert radiologists were completely unable to [9]. To our knowledge, only one other study has attempted such a task with clinical notes. This paper used the publicly available MIMIC dataset [30], and found that while age and gender were easy to predict from clinical notes, race posed a harder challenge [13]. Their final model was only able to distinguish between White and non-White patients with an area under the receiver operator curve (AUC) of 0.62. However, their analysis grouped all non-white minorities into a single combined category, which severely limited predictive accuracy. In contrast, we focus on differentiating between Black and White patients; this approach avoids grouping heterogeneous populations [60] and allows us to specifically focus on disparities faced by Black patients. We have not found any work that attempts to assess the ability of human experts to predict race from clinical notes.

Racial Biases in Clinical Notes. Recent work has established that clinical notes may not be written in the same way for all types of patients, and may reflect racial disparities in clinical care. For instance, notes written for Black patients are much more likely to contain indicators of physician mistrust [10], negative descriptors [54], and other stigmatizing language [45] than those written for White patients. Clinical notes have revealed that Black patients have lower levels of trust in their physicians during end of life care [14]. These studies all audit a specific bias in clinical notes by identifying language that varies significantly by patient race. Our work adopts a broader approach, and characterizes all differences in content and language that are predictive of race.

Clinical Prediction Tasks. Clinical notes are becoming an increasingly common data source for machine learning applications in healthcare. Several studies have used ML models to predict patient outcomes such as in-hospital mortality [20, 26, 34, 62], 30-day mortality [36], and readmission [24, 34, 51]. Language models like MedBERT [48] and Clinical BERT [5] have also demonstrated excellent performance on tasks like disease prediction, de-identification, and named entity recognition. However, recent work has demonstrated that such models can exhibit performance gaps for racial and gender subgroups [64]. Our work emphasizes the dangers of naively using such models for clinical prediction. We demonstrate that even if ML models are trained on seemingly race-redacted data, they may still propagate existing biases in clinical decisions.

3 DATA

Our dataset consists of clinical notes from two sites: Beth Israel Deaconess Medical Center in Boston and Columbia University Medical Center In New York. De-identified clinical notes from Beth Israel are publicly available through the Medical Information Mart for Intensive Care III (MIMIC-III) version 1.4 database [30]. The notes from Columbia are private data available from electronic health records (EHR), and contain protected health information (PHI). The use of this data was approved by Columbia's Institutional Review Board (IRB).

In our analyses, we focused on progress updates and other clinical notes written by nurses. In addition to these nursing notes, we extracted a patient's self-reported race and other demographic information from their EHR. We only included patients who self-report their race as White/Caucasian and Black/African-American due to smaller numbers of other self-reported race categories. To

ensure similarity between the two datasets, our analysis only included adult patients (i.e. over 18) admitted to non-pediatric units, as well as infants admitted to the neonatal ICU (NICU). For patients admitted multiple times, we only considered their first stay.

Our final dataset contained 668,768 notes written for 28,032 patients from MIMIC and 3,554,802 notes for 29,807 patients from Columbia. Table S1 in the Appendix presents a summary of the resulting cohort's demographics. We note that many existing works in clinical natural language processing (NLP) rely exclusively on the publicly available MIMIC dataset. Our ability to analyze data from both MIMIC and the Columbia datasets is important, especially as the patient populations differ significantly across the two sites. Most notably, Columbia sees a much higher proportion of Black patients (∼20% vs. ∼10%).

Before conducting our analyses, we redacted any explicit mentions of patient race from the nursing notes in both datasets. We compiled a list of terms that were used as identifiers of race, and removed these from the two corpora of notes using regular expression operations. These terms were identified by training a logistic regression classifier to predict race using a unigram bag-of-words (BoW) representation of the notes in both datasets. We manually inspected the most predictive terms for each race (according to the model's coefficients), and identified all terms that could be used as an explicit indicator of patient race. The final list contained the terms African-American, African, Black, and Creole as identifiers of Black race, and the words Caucasian and White as identifiers of White race. These terms were removed regardless of capitalization (e.g. black vs Black), case (e.g. AFRICAN vs African), and hyphenation (e.g. African-American vs. African American). As the Columbia dataset contains PHI, we also removed mentions of area codes, neighborhoods, and hospitals that served as proxies for race. A full list of these terms is provided in the appendix.

4 MODEL-BASED RACE DETECTION

The goal of our primary analysis is to demonstrate that ML models are able to identify a patient's self-reported race from nursing notes that describe their condition and progress. We find that even after nursing notes are stripped of explicit racial identifiers, models are able to accurately predict patient race. This accurate prediction was found across different sites, units, and patient types. Investigating the drivers of predictive performance, we determined that clinical notes written for White and Black patients vary greatly in content. While some of the identified differences are clinically justifiable, others might suggest disparities in clinical care and warrant further investigation.

4.1 Methods

We trained four machine learning models to predict a patient's self-reported race from nursing notes written during their stay. Crucially, these models were trained on notes that were stripped of any explicit indicators of patient race. We used a unigram[1] BoW representation of the nursing notes to train an L1-penalized logistic regression [46] and an xgBoost [17] classifier, as well as a stacking ensemble of the two methods. These three models performed

---

[1]We also tried bigram representations, but found that these did not boost performance. See Appendix C

their prediction at the visit level: all notes written for the same patient were aggregated into one large note that was then used for prediction. This approach allowed us to easily pool all the available information on each patient, and yielded higher predictive accuracy.

In addition to the bag-of-words models, we also fine-tuned SciB-ERT [12]–a language model trained on scientific abstracts–to classify patient self-reported race (note that we did not use clinical note-specific models like Clinical BERT [5], as some of these have already been trained on the MIMIC data, creating possible information leakage). For SciBERT, we first fine-tuned the model to predict patient self-reported race from individual nursing notes (as opposed to one combined note), then aggregated the predictions by patient. In the aggregation step, we considered the model to have predicted a patient's race as Black if it did so for any of their individual notes. This approach was necessitated by the fact that the combined notes were usually much longer than SciBERT's 512 token limit. Note that we were unable to test the SciBERT classifier on the Columbia dataset due to privacy and computational constraints (the data contains PHI, and must stay on a private server that does not have access to a GPU).

We trained and evaluated these models on ten random 7:3 train-test splits, evaluating performance by the area under the receiver operator curve (AUC) on the test set of patients after models are

Table 1: Detailed classification accuracy achieved by the ensemble method in predicting race from nursing notes. We evaluated the classifier on ten random train-test splits, and report the mean and standard deviation of test set AUCs across splits. We report accuracy on the whole test set population, as well as patient subgroups based on specific comorbidities, VW comorbidity score [57] decile, and unit type.

Subgroup	AUC	
	MIMIC	Columbia
Overall	0.83 (0.00)	0.78 (0.02)
Comorbidity		
Diabetes	0.80 (0.00)	0.76 (0.01)
Hypertension	0.82 (0.01)	0.80 (0.02)
COPD	0.82 (0.02)	0.79 (0.00)
Obesity	0.77 (0.00)	0.76 (0.02)
Comorbidity Score		
Top decile	0.83 (0.00)	0.73 (0.01)
Bottom decile	0.81 (0.01)	0.80 (0.02)
Unit		
MICU	0.81 (0.00)	0.75 (0.01)
CCU	0.83 (0.03)	0.82 (0.01)
NICU	0.81 (0.01)	0.71 (0.00)
SICU	0.80 (0.01)	0.76 (0.01)
TSICU	0.80 (0.02)	-
NUICU	-	0.78 (0.00)
CSRU	-	0.75 (0.00)
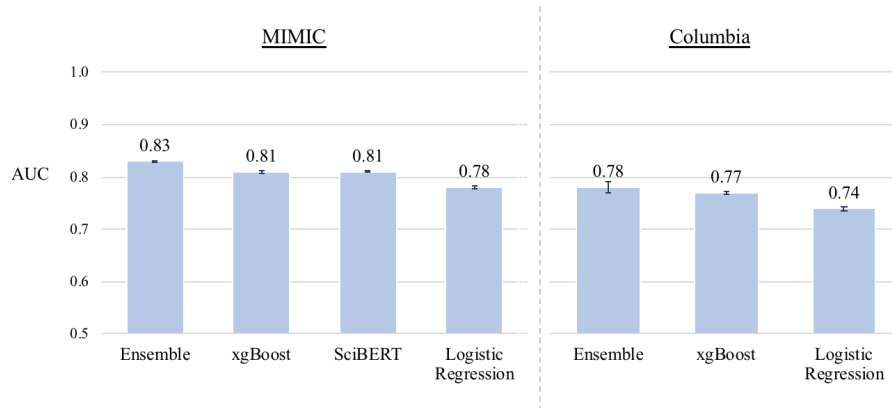
Figure 1: Model classification performance for patient self-reported race from nursing notes. The chart displays the mean AUC (across 10 random train-test splits) and error bars that signify 95% confidence intervals. The ensemble of xgBoost and logistic regression classifiers demonstrate the highest accuracy in both datasets.

Figure 2: An adversarial example demonstrating the association between White race and positive descriptors in the SciBERT model's predictions. We take an excerpt of a note written for a Black patient (in black text), and an excerpt of a note written for a White patient (in red text). Both excerpts were taken from notes from the test set (i.e. not the data the model was trained on) that the model predicted correctly. Adding a positive descriptor of the patient's family led the model to change its prediction of the patient's race from Black to White.

trained to convergence on training data only. We then inspected models in two ways to uncover the differences in the content of notes written for Black and White patients. First, we examined which coefficients of the logistic regression were most predictive of patient race, specifically working to identify how the words predictive of Black race differ from those predictive of White race. Second, we ran a structural topic model (STM) [48] to identify more nuanced differences between notes written for Black and White patients in an unsupervised manner. Further details on the STM are provided in Appendix D.

4.2 Results

4.2.1 Self-Reported Race is Predictable from Nursing Notes.
We found that a patient's self-reported race is predictable from nursing notes written during their hospital stay, even after explicit indicators of race are removed. All models were able to distinguish between Black and White patients (Figure 1), with the best model achieving 0.83 AUC on the MIMIC dataset and 0.78 AUC on the

Columbia dataset. Crucially, predictive performance is not driven by a specific characteristic or patient type: the results are consistent across various subgroups of the held out set, including patients with diabetes, hypertension, chronic pulmonary disease, and obesity, as well as patients admitted to different units (Table 1). While there is some variation in performance, the models all perform above 0.7 AUC on all categories in both datasets. The fact that race is predictable in two large hospitals in different cities with very different patient populations is notable, and speaks to the generalizability of our primary result.

4.2.2 Notes Written for Black and White Patients Differ Significantly.
The finding that an algorithm can distinguish between White and Black patients is not troubling on its own. For example, Black populations have higher rates of comorbidities like diabetes, asthma, and obesity [19, 31]; their notes are likely to mention these conditions more often, creating a pattern that an algorithm will be able to pick up on. However, a similar pattern could also be created by a

different standard of care for White and Black patients [54], which is more concerning. For instance, if stigmatizing language and negative descriptors are used more frequently for Black patients [45, 54], models would also be able to rely on such associations to identify patient race.

The SciBERT model trained to predict self-reported race exhibits such a trend, as it associated positive descriptors of patient family with White race. In several instances, adding a positive description of the patient's family to the note led SciBERT to change its prediction of the patient's race from Black to White (Figure 2). This suggests that the model may have learnt an association between "loving and caring" and White race.

We investigated the 25 words most predictive of each racial group (on average across train-test splits), classifying them into five clinically motivated categories: skin-related, personal, comorbidity, clinical care, and patient condition (Figure 3). We find that Black patients are often identified by comorbid conditions like sickle cell anemia, asthma, and diabetes, which are more common in Black patients [23, 31, 56]. However, references to skin like bruising, redness, or paleness are strong predictors of White self-reported race, but these words don't necessarily reflect conditions that should be more common for White patients. Paleness, redness, and bruising are all clinical symptoms that should be noted for both White and Black patients. The fact that they are strongly associated with White skin is troubling in the context of previous work that suggests that healthcare providers are less equipped to diagnose skin conditions in patients with darker skin. A number of reviews have found that only a small fraction of examples provided in dermatology textbooks are on non-white skin, which can lead to serious underdiagnosis [2, 35].

We also find differences in words that may be subjective rather than clinical. For instance, the phrase "family members" is associated with Black patients in MIMIC, while "husband" and "father" are associated with White patients. Some of this trend can be explained by population differences: more White women in the sample are married. However, even if we only consider married female patients, husband is still referred to more often for White patients than Black (Table 2). Another example is that the word "difficult" is predictive of Black race in the MIMIC dataset, while "demanding" is predictive of Black race in the Columbia dataset. As Figure 4 demonstrates, this word can be used in many contexts: saying a patient is difficult is very different from saying that they are a "difficult stick" (i.e. a

patient whose veins are hard to insert a needle into). While the latter is a more objective claim, the first is subjective, and may hint at differential treatment. Statements implying that the patient was "very difficult" or "very demanding" were more frequent for Black patients, which is a concerning trend. (Figure 4)

We summarize these more nuanced differences in content between notes written for White and Black patients using a structural topic model (Figure 5). This analysis largely supports existing findings around comorbidities and skin. However, it yields a few additional insights: for example, discussion of mental health conditions like anxiety is much more common in the clinical notes of White patients. This may again reveal a systemic issue, as anxiety disorders are understudied, underdiagnosed, and undertreated in Black populations [58, 61]. Overall, the STM establishes that there are several implicit indicators of race in nursing notes, which makes redacting patient race a challenging task.

4.2.3 Predictors of Race Are Deep-rooted in the Text. While we have focused on the top predictors of patient self-reported race, we also find that the ensemble model is able to perform well over chance even after removing the strongest predictors of race from the MIMIC notes (Table 3). This finding indicates that the signals of race are deeply rooted in the text, and simply removing some words will not address this issue.

5 RACE DETECTION BY HUMAN EXPERTS

In the previous section, we established that ML models can infer patient race from nursing notes that are stripped of explicit racial identifiers. We further identified a number of race-based differences in clinical notes that drive this predictive performance. In this section, we evaluate whether humans are also able to identify patient self-reported race from redacted clinical notes. In a survey of 42 physicians, we found this is not true. The surveyed physicians not only struggled to accurately predict patient race, but often admitted that their predictions were no better than complete guesses. This finding speaks to the limits of human supervision of ML models: if a model were relying on its covertly inferred estimate of patient race, human experts would likely not be able to tell.

5.1 Methods

We recruited 42 physicians via email to participate in a short web based experiment. This study was exempt from a full IRB ethical review, as it met the criteria for exemption defined in Federal regulation 45 CFR 46. We chose physicians as experts in this setting

Table 2: References to family for married, female patients in MIMIC. The table displays the percentage of patients by race whose notes contain at least one mention of the given word. Personal descriptors like "husband" and "father" are more common in notes written for White patients, while the group descriptor "family members" is more common in notes written for Black patients.

Word	% Black Patients	% White Patients
husband	50%	64%
family members	22%	14%
father	2%	5%

Table 3: Ablation results for the ensemble model in the MIMIC dataset. Removing the top 25 most predictive words for each race (according to logistic regression coefficients) impacts performance, but the model is still able to detect race.

	AUC
With all features	0.83
Removing common skin-related features	0.76
Removing top 25 features	0.73

(a) Predictive of Black Patients

MIMIC			Columbia		
Word	Category	Coefficient	Word	Category	Coefficient
mongolian	Skin-related	1.00	baptist	Personal	0.31
members	Personal	0.15	sickle	Comorbidity	0.26
stick	Comorbidity	0.08	hypertension	Comorbidity	0.23
sickle	Comorbidity	0.07	mongolian	Skin-related	0.21
htn	Comorbidity	0.07	er	Clinical Care	0.18
anuric	Comorbidity	0.07	htn	Comorbidity	0.16
obese	Comorbidity	0.07	asthma	Comorbidity	0.16
asthma	Comorbidity	0.07	bible	Personal	0.16
another	Clinical Care	0.06	pregnant	Patient Condition	0.12
hugger	Clinical Care	0.06	activities	Clinical Care	0.12

(b) Predictive of White Patients

MIMIC			Columbia		
Word	Category	Coefficient	Word	Category	Coefficient
reddened	Skin-related	-0.44	russian	Personal	-0.43
ecchymotic	Skin-related	-0.24	hypothyroidism	Comorbidity	-0.40
russian	Personal	-0.15	kosher	Personal	-0.30
ruddy	Skin-related	-0.13	hearing	Patient Condition	-0.22
pale	Skin-related	-0.11	pale	Skin-related	-0.20
bruising	Skin-related	-0.11	hebrew	Personal	-0.20
osh	Other	-0.09	spouse	Personal	-0.19
husband	Personal	-0.09	ecchymotic	Skin-related	-0.19
grunting	Patient Condition	-0.09	anxiety	Patient Condition	-0.18
rash	Skin-related	-0.08	reddened	Skin-related	-0.18

Figure 3: Words that are most predictive of race in nursing notes, sorted by the word's logistic regression coefficient. We categorized the predictive words into five clinically motivated categories: skin-related, personal, comorbidity, clinical care, and patient condition.

as they are both experienced in reading nursing notes, and a step removed from actually writing them.

Consenting participants were shown ten notes from the MIMIC dataset. The selected notes were racially balanced (five White patients, five Black patients), and conveyed different levels of model accuracy: four notes that were predicted correctly, four that were predicted incorrectly, and two that the model was unsure about (i.e. ∼ 50% predicted probability of the patient being Black). For each note, participants were asked to indicate (1) whether they believed the patient was White/Caucasian or Black/African-American, and (2) how sure they were in their belief on a scale of 1-5, where 1 reflects a complete guess and 5 a strong belief. Participants were also given the option to highlight parts of the text that informed their belief.

We evaluated physicians on overall accuracy (i.e. the percentage of patients whose race they identified correctly), sensitivity for Black patients (i.e. the number of Black patients who were correctly identified as being Black), and the positive predictive value (PPV) for White patients (i.e. the number of predicted White patients who were actually White). We also evaluated the agreement between physician predictions using Fleiss' kappa measure [18].
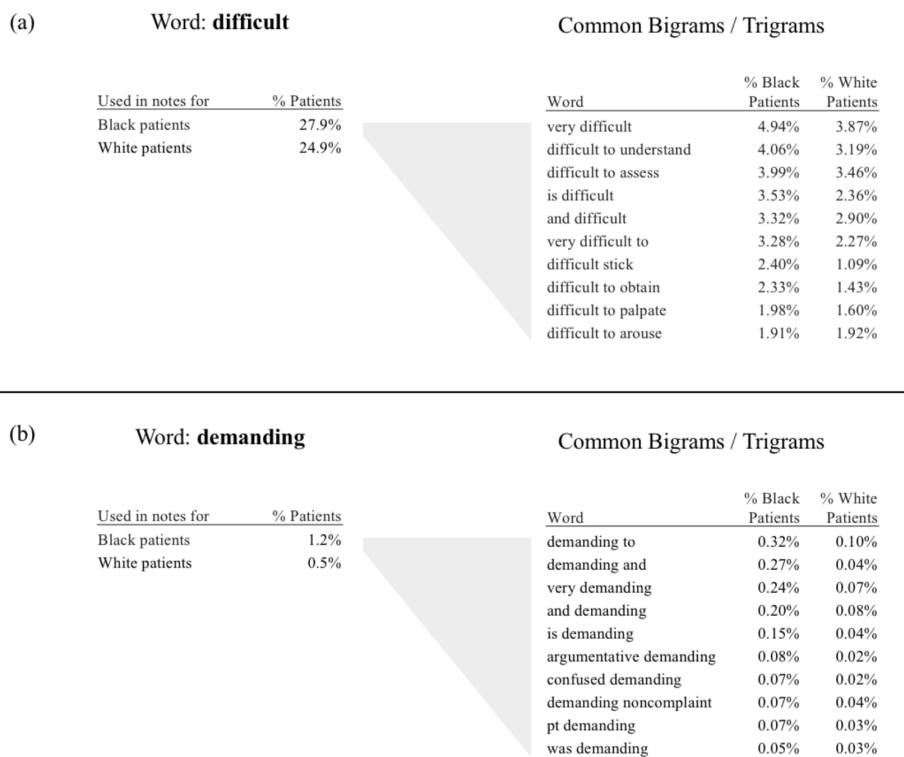
5.2 Results

We found that the physicians in our study were unable to predict patient race, with an average accuracy of 54% ($n = 420$ responses), only slightly better than chance (Figure 6). While they were more accurate for White patients (70% vs 37% for Black patients), this is likely as they defaulted to guessing a patient was White - the positive predictive value for White patients is just 53%. There was also only slight agreement between physicians, with their predictions exhibiting a Kappa statistic of 0.05 (rejected null hypothesis of no agreement with z-value=4.84, p-value< 0.001).

Crucially, in a vast majority of cases (∼ 75%), physicians indicated that their prediction was a complete guess. Moreover, accuracy did not increase with self-reported certainty; physicians who said they had "some idea" of a patient's race were less accurate than the average respondent (40%, $n = 43$ responses).

6 SIMULATION EXPERIMENT ON BIAS PROPAGATION

While it is not inherently problematic for a model to be able to predict a patient's race from nursing notes, it is concerning if these differences lead to poorer performance. We perform an experiment

(a) Word: **difficult**

Common Bigrams / Trigrams

Used in notes for	% Patients
Black patients	27.9%
White patients	24.9%

Word	% Black Patients	% White Patients
very difficult	4.94%	3.87%
difficult to understand	4.06%	3.19%
difficult to assess	3.99%	3.46%
is difficult	3.53%	2.36%
and difficult	3.32%	2.90%
very difficult to	3.28%	2.27%
difficult stick	2.40%	1.09%
difficult to obtain	2.33%	1.43%
difficult to palpate	1.98%	1.60%
difficult to arouse	1.91%	1.92%

(b) Word: **demanding**

Common Bigrams / Trigrams

Used in notes for	% Patients
Black patients	1.2%
White patients	0.5%

Word	% Black Patients	% White Patients
demanding to	0.32%	0.10%
demanding and	0.27%	0.04%
very demanding	0.24%	0.07%
and demanding	0.20%	0.08%
is demanding	0.15%	0.04%
argumentative demanding	0.08%	0.02%
confused demanding	0.07%	0.02%
demanding noncomplaint	0.07%	0.04%
pt demanding	0.07%	0.03%
was demanding	0.05%	0.03%

Figure 4: Common bigrams and trigrams for the word difficult in the MIMIC dataset and demanding in the Columbia dataset, both of which are predictive of Black race. We measure the frequency of each term as the percentage of patients of a given race whose notes contained the term at least once.

using actual clinical notes with a synthetically generated biased treatment decision. We show that if ML models are trained on biased decisions, they make biased recommendations even without explicit access to patient race.

6.1 Methods

We demonstrate the dangers of race-inferring models through a simulation experiment. Prior work has established the existence of several racial disparities in clinical treatment decisions. For example, Black patients are ∼ 30% less likely to be prescribed analgesia for acute pain in emergency settings than White patients [33]. Black patients are also less likely than White patients to be given appropriate cardiac care [7], to receive kidney dialysis or transplants [41], and to receive the best treatments for stroke, cancer [28], and AIDS [40]. Our simulation evaluates whether ML models can perpetuate such biases in treatment even if they are trained on race-redacted clinical notes.

Our experiment uses real clinical notes from MIMIC with a synthetically generated, biased treatment decision. Because White patients far outnumber Black patients in MIMIC (Table S1), we created a balanced dataset of 2,014 adult Black patients and 2,014 adult White patients with random undersampling. Note that this approach captures all the adult Black patients in our cohort. We

assumed that 50% of these patients had a clinical condition (e.g. acute pain) that made them eligible for a specific treatment (e.g. analgesia). The presence of this condition was randomly assigned so that it was equally prevalent in Black and White patients. However, the decision to administer treatment to a patient with this condition was racially biased, that is, the treatment was assigned at a higher rate to White patients than Black patients. This biased decision resembles previously discussed disparities in analgesia prescription and other clinical care [33, 40].

We then evaluated whether a model trained to make this decision would perpetuate the treatment gaps in the data. We trained an L2-penalized logistic regression [46] to predict the treatment from a patient's nursing notes, using an 8:2 train-test split. As before, we used a unigram BoW representation of the nursing notes. These notes were race-redacted, that is, contained no explicit identifiers of patient race. After removing stop words, tokenizing, and lemmatizing, the final vocabulary consists of 54,432 words. The model also received the presence of the clinical condition as an additional variable. We evaluated our model on the test set, and assessed whether the trained model was significantly less likely to recommend the treatment to Black patients than White patients. If such a gap exists, then the bias from the training data has propagated to the model, as differential treatment rates by race create differential model recommendations. We assessed this bias propagation for

Regression of topic prevalence on race
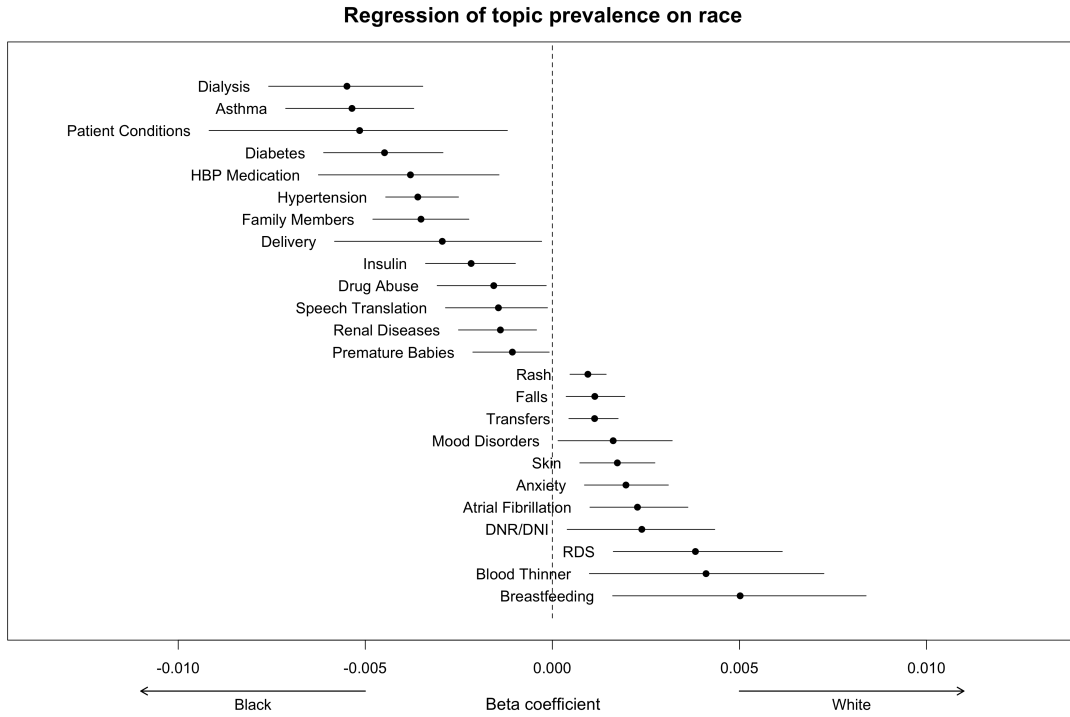
Figure 5: Differences in the topics discussed in notes written for Black and White patients in MIMIC. Topics were algorithmically identified by running a structural topic model (STM) on the MIMIC notes with k=200 topics and race as a covariate. Topics to the left of the dashed vertical line are significantly more common in notes written for Black patients, while those to the right are significantly more common in notes written for White patients. The chart plots the mean effect of race on topic prevalence with an error bar signifying the 95% confidence interval (note that we only display the subset of topics with significant effects). Topics were manually labeled using the high-probability words identified by the STM.

various magnitudes of training bias (10-50%), and report average results and 95% confidence intervals across 100 simulations.

6.2 Results

We found that even without access to patient race, the model propagates the bias in the training data, and is significantly less likely to recommend the treatment to Black patients (Figure 7). This trend is observed for both small ($\leq 20\%$) and large ($\geq 30\%$) levels of training bias. We find that the level of propagation scales with the level of induced disparity, i.e, a training set disparity of 10% results in a 3% gap in model recommendations, while a 30% disparity creates nearly a 10% gap. The magnitude of the training set bias may be generally reduced in the model recommendation gaps because models do not predict race perfectly. However, the bias is replicated here with only the redacted notes as data, and no direct access to patient self-reported race or any other correlated demographics. While this finding is perhaps not surprising, it has not been noted before in prior work using clinical notes, and is important to highlight given the severe consequences of undetected bias propagation.

Overall, our treatment simulation experiment demonstrates that even if a note contains no explicit information on patient race, the implicit racial information provides a signal that ML models

could use to propagate existing biases in clinical care. We know from prior work that the absence of racial information in data is a sufficient condition for achieving fairness in machine learning recommendations [38]. However, other work has shown that learning race-blind representations is challenging, and biases are hard to remove through standard adversarial techniques [64]. Our work further emphasizes this fact: racial information is deeply rooted in clinical notes, and implicit signals provide a potential vector for bias propagation.

7 DISCUSSION AND CONCLUSION

Our work demonstrates that models are able to accurately predict patient self-reported race in the redacted notes of Black and White patients, where human experts are not. We also simulate the implications of this difference in practice, given a biased treatment setting. Our work has several key implications for clinical practice and the deployment of ML tools in healthcare settings.

First, our work highlights potential areas of missed care. The investigation of our model's performance revealed some differences in clinical notes that were hard to explain; for example, references to bruising or rashes are extremely predictive of White race, even though there is no clinical reason for these symptoms to be less
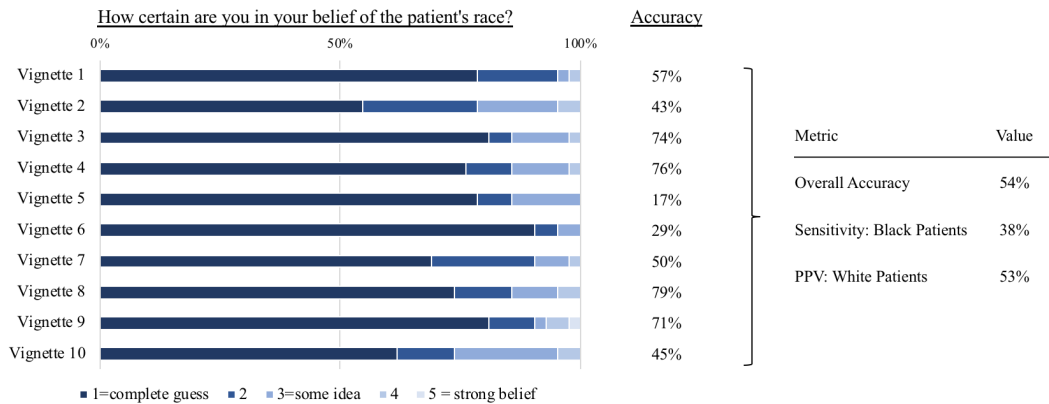
Figure 6: Assessing the ability of human experts to detect race from nursing notes. For all ten notes presented, the majority of the 42 surveyed physicians indicated that their prediction of the patient's race was a complete guess. This lack of surety is borne out in their predictions, which are barely better than chance (average accuracy of 54% across vignettes). The physicians have low sensitivity in identifying Black patients (38%) and low positive predictive value (PPV) in identifying White patients (53%), indicating that they may be defaulting to guessing a patient is White.
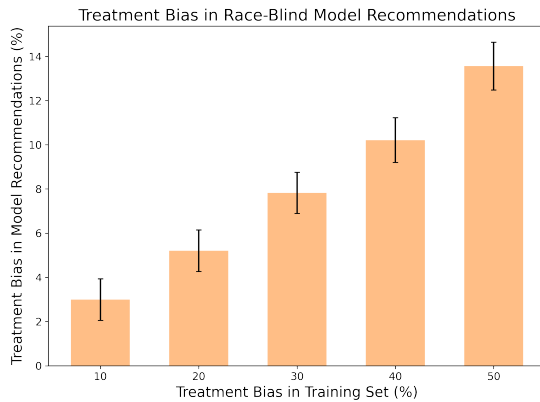
Figure 7: The mean recommendation bias at each level of treatment bias, along with an error bar that signifies the 95% confidence interval. The x-axis plots the racial bias in the treatment decision. For example, a 10% treatment bias describes a situation in which 80% of eligible White patients were administered the treatment, but only 70% of eligible Black patients were. The y-axis plots the corresponding bias in model predictions: how much less likely was the model to recommend the treatment for a Black patient than a White patient?

common in Black patients. While our analysis is not sufficient to establish that nurses are missing skin symptoms in Black patients, the strong association between these terms and White race does suggest the possibility of missed care. Our findings are very concerning from a clinical perspective since patients in the ICU setting are often non-mobile and are at greater risk for skin damage and underlying

soft tissue breakdown. These preventable injuries often cause pain, infection and patient harm [44], and other clinicians have noted the need for increased education to identify skin damage in darker skin to avoid harmful consequences [37, 43]. Another concerning observation is the more frequent use of words like "demanding" and "difficult" for Black patients, which may hint at differential treatment. Investigating these trends further and causally establishing the presence of disparities in clinical care based on differences in documentation is an important avenue for future work.

Second, the risk of bias propagation is compounded by the fact that human experts do not share the ability to identify race from clinical notes. This finding establishes the limits of human oversight of ML systems [32, 50]. Standard machine learning interpretability techniques highlight important features used by models in making predictions [22]. Even if these techniques worked perfectly, human experts would not be able to judge whether highlighted predictors were implicitly conveying racial information. Thus, if a model inferred race in making clinical predictions, human experts may not be able to detect this racial bias. This emphasizes the need to explicitly incorporate fairness considerations when designing ML systems in healthcare. It is vital to embed automated fairness checks and constraints [1, 3, 11] at every stage in the ML pipeline, from data collection [15] to algorithm development [38] to deployment [16].

Finally, we emphasize that removing explicit racial identifiers from clinical notes is not sufficient to obscure patient race. This finding is vital to consider when designing algorithms to support clinical decision making. As our simulation experiment demonstrates, algorithms can still propagate existing biases in clinical care even if trained in a seemingly race-blind fashion. The combination of existing health disparities and potentially race-inferring algorithms makes it incredibly easy to unintentionally encode disparate treatment. Any ML model trained on clinical notes must thus be

thoroughly and continuously audited for racial bias both before and after deployment [21, 47, 59].

ACKNOWLEDGMENTS

REFERENCES

[1] Julius A Adebayo. 2016. *FairML : ToolBox for diagnosing bias in predictive modeling.* Ph. D. Dissertation. Massachusetts Institute of Technology.

[2] Ademide Adelekun, Ginikanwa Onyekaba, and Jules B Lipoff. 2021. Skin color in dermatology textbooks: An updated evaluation and analysis. *J. Am. Acad. Dermatol.* 84, 1 (Jan. 2021), 194–196.

[3] Aniya Aggarwal, Pranay Lohia, Seema Nagar, Kuntal Dey, and Diptikalyan Saha. 2019. Black box fairness testing of machine learning models. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (Tallinn, Estonia) *(ESEC/FSE 2019).* Association for Computing Machinery, New York, NY, USA, 625–635.

[4] M Alexander, K Grumbach, J Selby, A F Brown, and E Washington. 1995. Hospitalization for congestive heart failure. Explaining racial differences. *JAMA* 274, 13 (Oct. 1995), 1037–1042.

[5] 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. (April 2019). arXiv:1904.03323 [cs.CL]

[6] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. ProPublica, May 23, 2016.

[7] Sameer Arora, George A Stouffer, Anna Kucharska-Newton, Muthiah Vaduganathan, Arman Qamar, Kunihiro Matsushita, Dhaval Kolte, Harmony R Reynolds, Sripal Bangalore, Wayne D Rosamond, Deepak L Bhatt, and Melissa C Caughey. 2018. Fifteen-year trends in management and outcomes of non-ST-segment-elevation myocardial infarction among black and white patients: The ARIC Community Surveillance study, 2000-2014. *J. Am. Heart Assoc.* 7, 19 (Oct. 2018), e010203.

[8] Peter B Bach, Hoangmai H Pham, Deborah Schrag, Ramsey C Tate, and J Lee Hargraves. 2004. Primary care physicians who treat blacks and whites. *N. Engl. J. Med.* 351, 6 (Aug. 2004), 575–584.

[9] Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P Lungren, Lyle Palmer, Brandon J Price, Saptarshi Purkayastha, Ayis Pyrros, Luke Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman, Haoran Zhang, and Judy W Gichoya. 2021. Reading Race: AI Recognises Patient's Racial Identity In Medical Images. (July 2021). arXiv:2107.10356 [cs.CV]

[10] Mary Catherine Beach, Somnath Saha, Jenny Park, Janiece Taylor, Paul Drew, Eve Plank, Lisa A Cooper, and Brant Chee. 2021. Testimonial Injustice: Linguistic Bias in the Medical Records of Black Patients and Women. *J. Gen. Intern. Med.* 36, 6 (June 2021), 1708–1714.

[11] R K E Bellamy, K Dey, M Hind, S C Hoffman, S Houde, K Kannan, P Lohia, J Martino, S Mehta, A Mojsilović, S Nagar, K Natesan Ramamurthy, J Richards, D Saha, P Sattigeri, M Singh, K R Varshney, and Y Zhang. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Res. Dev.* 63, 4/5 (July 2019), 4:1–4:15.

[12] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. (March 2019). arXiv:1903.10676 [cs.CL]

[13] Willie Boag, Dustin Doss, Tristan Naumann, and Peter Szolovits. 2018. What's in a Note? Unpacking Predictive Value in Clinical Note Representations. *AMIA Jt Summits Transl Sci Proc* 2017 (May 2018), 26–34.

[14] Willie Boag, Harini Suresh, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. 2018. Racial Disparities and Mistrust in End-of-Life Care. In *Proceedings of the 3rd Machine Learning for Healthcare Conference (Proceedings of Machine Learning Research, Vol. 85)*, Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens (Eds.). PMLR, 587–602.

[15] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why Is My Classifier Discriminatory? (May 2018). arXiv:1805.12002 [stat.ML]

[16] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2021. Ethical Machine Learning in Healthcare. *Annu Rev Biomed Data Sci* 4 (July 2021), 123–144.

[17] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16).* Association for Computing Machinery, New York, NY, USA, 785–794.

[18] Joseph L Fleiss and Jacob Cohen. 1973. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educ. Psychol. Meas.* 33, 3 (Oct. 1973), 613–619.

[19] Erick Forno and Juan C Celedon. 2009. Asthma and ethnic minorities: socioeconomic status and beyond. *Curr. Opin. Allergy Clin. Immunol.* 9, 2 (April 2009), 154–160.

[20] Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. 2014. Unfolding Physiological State: Mortality Modelling in Intensive Care Units. *KDD* 2014 (Aug. 2014), 75–84.

[21] Marzyeh Ghassemi and Elaine Okanyene Nsoesie. 2022. In medicine, how do we machine learn anything real? *Patterns (N Y)* 3, 1 (Jan. 2022), 100392.

[22] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA).* 80–89.

[23] Jeffrey Glassberg, Paula Tanabe, Lynne Richardson, and Michael Debaun. 2013. Among emergency physicians, use of the term "Sickler" is associated with negative attitudes toward people with sickle cell disease. *Am. J. Hematol.* 88, 6 (June 2013), 532–533.

[24] Sara Nouri Golmaei and Xiao Luo. 2021. DeepNote-GNN: predicting hospital readmission using clinical notes and patient network. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* (Gainesville, Florida) *(BCB '21, Article 19).* Association for Computing Machinery, New York, NY, USA, 1–9.

[25] Alexander R Green, Dana R Carney, Daniel J Pallin, Long H Ngo, Kristal L Raymond, Lisa I Iezzoni, and Mahzarin R Banaji. 2007. Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients. *J. Gen. Intern. Med.* 22, 9 (Sept. 2007), 1231–1238.

[26] Paulina Grnarova, Florian Schmidt, Stephanie L Hyland, and Carsten Eickhoff. 2016. Neural Document Embeddings for Intensive Care Patient Mortality Prediction. (Dec. 2016). arXiv:1612.00467 [cs.CL]

[27] Paul L Hebert, Elizabeth A Howell, Edwin S Wong, Susan E Hernandez, Seppo T Rinne, Christine A Sulc, Emily L Neely, and Chuan-Fen Liu. 2017. Methods for Measuring Racial Differences in Hospitals Outcomes Attributable to Disparities in Use of High-Quality Hospital Care. , 826–848 pages.

[28] Dawn Hershman, Russell McBride, Judith S Jacobson, Lois Lamerato, Kevin Roberts, Victor R Grann, and Alfred I Neugut. 2005. Racial disparities in treatment and survival among women with early-stage breast cancer. *J. Clin. Oncol.* 23, 27 (Sept. 2005), 6639–6646.

[29] Elizabeth A Howell, Jennifer Zeitlin, Paul Hebert, Amy Balbierz, and Natalia Egorova. 2013. Paradoxical Trends and Racial Differences in Obstetric Quality and Neonatal and Maternal Mortality. , 1201–1208 pages.

[30] Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Sci Data* 3 (May 2016), 160035.

[31] Rafi Kabarriti, N Patrik Brodin, Maxim I Maron, Chandan Guha, Shalom Kalnicki, Madhur K Garg, and Andrew D Racine. 2020. Association of Race and Ethnicity With Comorbidities and Survival Among Patients With COVID-19 at an Urban Medical Center in New York. *JAMA Netw Open* 3, 9 (Sept. 2020), e2019795.

[32] Riikka Koulu. 2020. Proceduralizing control and discretion: Human oversight in artificial intelligence policy. *Maastrich. J. Eur. Comp. Law* 27, 6 (Dec. 2020), 720–735.

[33] Paulyne Lee, Maxine Le Saux, Rebecca Siegel, Monika Goyal, Chen Chen, Yan Ma, and Andrew C Meltzer. 2019. Racial and ethnic disparities in the management of acute pain in US emergency departments: Meta-analysis and systematic review. *Am. J. Emerg. Med.* 37, 9 (Sept. 2019), 1770–1777.

[34] Li-Wei Lehman, Mohammed Saeed, William Long, Joon Lee, and Roger Mark. 2012. Risk stratification of ICU patients using topic models inferred from unstructured progress notes. *AMIA Annu. Symp. Proc.* 2012 (Nov. 2012), 505–511.

[35] J C Lester, J L Jia, L Zhang, G A Okoye, and E Linos. 2020. Absence of images of skin of colour in publications of COVID-19 skin manifestations. *Br. J. Dermatol.* 183, 3 (Sept. 2020), 593–595.

[36] Yen-Fu Luo and Anna Rumshisky. 2016. Interpretable Topic Features for Post-ICU Mortality Prediction. *AMIA Annu. Symp. Proc.* 2016 (2016), 827–836.

[37] Courtney Lyder. 2009. Closing the skin assessment disparity gap between patients with light and darkly pigmented skin. *J. Wound Ostomy Continence Nurs.* 36, 3 (May 2009), 285.

[38] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning Adversarially Fair and Transferable Representations. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 3384–3393.

[39] Emmanuel Martinez and Lauren Kirchner. 2021. The Secret Bias Hidden in Mortgage-Approval Algorithms.

[40] Alan Nelson. 2002. Unequal treatment: confronting racial and ethnic disparities in health care. *J. Natl. Med. Assoc.* 94, 8 (Aug. 2002), 666–668.

[41] Yue-Harn Ng, V Shane Pankratz, Yuridia Leyva, C Graham Ford, John R Pleis, Kellee Kendall, Emilee Croswell, Mary Amanda Dew, Ron Shapiro, Galen E Switzer, Mark L Unruh, and Larissa Myaskovsky. 2020. Does Racial Disparity in Kidney Transplant Waitlisting Persist After Accounting for Social Determinants of Health? , 1445–1455 pages.

[42] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (Oct. 2019), 447–453.

[43] Neesha Oozageer Gunowa, Joanne Brooke, Marie Hutchinson, and Debra Jackson. 2020. Embedding skin tone diversity into undergraduate nurse education: Through the lens of pressure injury. *J. Clin. Nurs.* 29, 21-22 (Nov. 2020), 4358–4367.

[44] Neesha Oozageer Gunowa, Marie Hutchinson, Joanne Brooke, and Debra Jackson. 2018. Pressure injuries in people with darker skin tones: A literature review. *J. Clin. Nurs.* 27, 17-18 (Sept. 2018), 3266–3275.

[45] Jenny Park, Somnath Saha, Brant Chee, Janiece Taylor, and Mary Catherine Beach. 2021. Physician Use of Stigmatizing Language in Patient Medical Records. *JAMA Netw Open* 4, 7 (July 2021), e2117052.

[46] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and Others. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[47] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 33–44.

[48] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med* 4, 1 (May 2021), 86.

[49] Margaret E Roberts, Brandon M Stewart, and Dustin Tingley. 2019. Stm: An R package for structural topic models. *Journal of Statistical Software* 91 (2019), 1–40.

[50] Paul Robinette, Ayanna Howard, and Alan R Wagner. 2017. Conceptualizing Overtrust in Robots: Why Do People Trust a Robot That Previously Failed? In *Autonomy and Artificial Intelligence: A Threat or Savior?*, W F Lawless, Ranjeev Mittu, Donald Sofge, and Stephen Russell (Eds.). Springer International Publishing, Cham, 129–155.

[51] A Rumshisky, M Ghassemi, T Naumann, P Szolovits, V M Castro, T H McCoy, and R H Perlis. 2016. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Transl. Psychiatry* 6, 10 (Oct. 2016), e921.

[52] Jeffrey H Silber. 2009. Hospital Teaching Intensity, Patient Race, and Surgical Outcomes. , 113 pages.

[53] Jonathan Skinner, Amitabh Chandra, Douglas Staiger, Julie Lee, and Mark Mc-Clellan. 2005. Mortality after acute myocardial infarction in hospitals that disproportionately treat black patients. *Circulation* 112, 17 (Oct. 2005), 2634–2641.

[54] Michael Sun, Tomasz Oliwa, Monica E Peek, and Elizabeth L Tung. 2022. Negative patient descriptors: Documenting racial bias in the electronic health record. *Health Aff.* (Jan. 2022), 101377hlthaff202101423.

[55] Eric J Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 25, 1 (Jan. 2019), 44–56.

[56] Megan M Tschudy and Tina L Cheng. 2016. The "Black Box" of Racial Disparities in Asthma. *JAMA Pediatr.* 170, 7 (July 2016), 644–645.

[57] Carl van Walraven, Peter C Austin, Alison Jennings, Hude Quan, and Alan J Forster. 2009. A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. *Med. Care* 47, 6 (June 2009), 626–633.

[58] Jennifer Vanderminden and Jennifer J Esala. 2019. Beyond Symptoms: Race and Gender Predict Anxiety Disorder Diagnosis. *Soc. Ment. Health* 9, 1 (March 2019), 111–125.

[59] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, Pilar N Ossorio, Sonoo Thadaney-Israni, and Anna Goldenberg. 2019. Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* 25, 9 (Sept. 2019), 1337–1340.

[60] David R Williams and Michelle Sternthal. 2010. Understanding racial-ethnic disparities in health: sociological contributions. *J. Health Soc. Behav.* 51 Suppl (2010), S15–27.

[61] Monnica T Williams, Diana A Beckmann-Mendez, and Eric Turkheimer. 2013. Cultural Barriers to African American Participation in Anxiety Disorders Research. *J. Natl. Med. Assoc.* 105, 1 (March 2013), 33–41.

[62] Jiancheng Ye, Liang Yao, Jiahong Shen, Rethavathi Janarthanam, and Yuan Luo. 2020. Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. *BMC Med. Inform. Decis. Mak.* 20, Suppl 11 (Dec. 2020), 295.

[63] Jiaming Zeng, Michael F Gensheimer, Daniel L Rubin, Susan Athey, and Ross D Shachter. 2022. Uncovering interpretable potential confounders in electronic medical records. *Nat. Commun.* 13, 1 (Feb. 2022), 1014.

[64] Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning* (Toronto, Ontario, Canada) *(CHIL '20)*. Association for Computing Machinery, New York, NY, USA, 110–120.

A   DESCRIPTIVE STATISTICS FOR PATIENTS

Table S1: Descriptive statistics of the patient cohort in each dataset, including demographics, insurance provider, unit type, and comorbidities. We report the number of patients in each category with the percentage of total patients in parentheses.

		Dataset	
		MIMIC	Columbia
Total		28,032	29,807
Race	Black	2,833 (10%)	5,883 (20%)
	White	25,199 (90%)	23,924 (80%)
Gender	Male	15,594 (56%)	16,696 (56%)
	Female	12,438 (44%)	13,110 (44%)
Insurance	Public	15,195 (54%)	19,206 (64%)
	Private	12,562 (45%)	10,158 (34%)
	Self-Pay	275 (1%)	443 (1%)
Age	<1	5,495 (20%)	3,728 (13%)
	1-17	0 (0%)	35 (0%)
	18-24	623 (2%)	545 (2%)
	25-34	902 (3%)	1,286 (4%)
	35-44	1,725 (6%)	1,553 (5%)
	45-54	3,176 (11%)	2,998 (10%)
	55-64	4,332 (15%)	5,280 (18%)
	65-74	4,497 (16%)	6,489 (22%)
	75+	7,282 (26%)	7,893 (26%)
Unit	MICU	8,763 (31%)	5,770 (19%)
	NICU	5,495 (20%)	3,767 (13%)
	SICU	4,247 (15%)	3,583 (12%)
	CCU	3,805 (14%)	5,175 (17%)
	TSICU	3,334 (12%)	0 (0%)
	CSRU	4,761 (17%)	7,145 (24%)
	NUICU	0 (0%)	4,367 (15%)
Number of Comorbidities	0	7,023 (25%)	4,135 (14%)
	1	3,614 (13%)	2,533 (8%)
	2	4,682 (17%)	3,459 (12%)
	3	4,513 (16%)	4,031 (14%)
	4	3,470 (12%)	4,236 (14%)
	5	2,303 (8%)	3,652 (12%)
	6	1,266 (5%)	2,738 (9%)
	7+	1,161 (4%)	5,023 (17%)

B   WORDS REMOVED FROM THE COLUMBIA NOTES

Table S2: A list of PHI removed from the Columbia notes to redact race. These terms often served as strong proxies for race, and were thus removed.

Area Codes	Places / Hospitals
347	brooklyn
646	harlem
201	interfaith
845	olmstead
908	downstate
430	5gn
185	cosgrove
718	zaire
	africa
	jamaica
	regional
	samaritan
	valley

C   BIGRAM REPRESENTATION

In addition to a unigram BoW representation of the clinical notes, we also tested models that used a unigram + bigram BoW representation. The results of these models on the MIMIC dataset are provided in Table S3. As these models did not meaningfully improve performance, we restricted our focus in the main paper to unigram only models.

Table S3: Classification accuracy of bigram representations. We found that adding bigrams to the BoW representation did not improve accuracy. As before, we evaluated the classifier on ten random train-test splits, and report the mean and standard deviation of test set AUCs across splits.

BoW Representation	Model	AUC
Unigram	Logistic Regression	0.78 (0.004)
	xgBoost	0.81 (0.003)
	Ensemble	0.83 (0.003)
Unigram + Bigram	Logistic Regression	0.78 (0.005)
	xgBoost	0.82 (0.004)
	Ensemble	0.83 (0.003)

D STM DETAILS

Here, we provide some more details about the structural topic model (STM) discussed in section 4. Each document in the STM is all the notes written for a single patient. The text was tokenized, and words were converted to lowercase and stemmed. Next, stop words, numbers, and punctuation were removed. Lastly, words that appear only in one document were dropped, as these words provide very little new information. Many such words are misspellings (e.g., "abswer" and "absolutl"). The final vocabulary contains 145,362 words. The patient's primary diagnosis, unit, and self-reported race were included as covariates to evaluate variation in topic prevalence.

To determine the optimal number of topics k, each topic in an STM was both computationally and manually evaluated using two metrics: topic cohesiveness and topic exclusivity. We also manually labeled each topic. Out of STMs with $k \in \{20, 30, 40, 50, 100, 20\}$, the STM with $k = 200$ was selected as it is highly exclusive and coherent. The top 30 words by probability and frequency for topics with significantly different topic prevalence by race are listed below. "Highest Prob" indicates words that had the highest probability in each topic, while the FREX metric identifies words both by their overall frequency and exclusivity to that topic [49].

- **Topic 6 (Hypertension) Top Words:**
  - Highest Prob: hypertens, sbp, hydralazin, htn, lopressor, benign, med, given, dose, start, increas, antihypertens, continu, systol, metoprolol, labetalol, gtt, prn, receiv, labetolol, plan, lisinopril, control, tid, pressur, emerg, order, effect, improv, remain
  - FREX: hypertens, antihypertens, hydralazin, hydral, antihypertens, labetalol, hydralizin, crisi, lisinopril, labetolol, htn, amlodipin, norvasc, metoprolol, benign, enalapril, nifedipin, lopressor, sbp, clonidin, atenolol, systol, emerg, urgenc, med, tid, hctz, labetelol, sbps, hydrochlorothiazid

- **Topic 22 (Breastfeeding) Top Words:**
  - Highest Prob: infant, feed, parent, monitor, note, resp, stabl, will, start, sat, distress, date, continu, warmer, respiratori, grunt, updat, well, mild, cont, remain, stool, mom, retract, void, plan, temp, lib, resolv, clear
  - FREX: grunt, ttn, lib, breastfeed, idm, polycos, adlib, dstick, pathway, bme, warmer, breastfe, dstix, lga, latch, mec, meconium, parent, heplock, babe, retain, d-stick, hypoglycemia, sim, bfing, near-term, dad, hat, t-shirt, gaggi

- **Topic 28 (Speech Translation) Top Words:**
  - Highest Prob: speak, interpret, english, russian, languag, translat, understand, improv, given, need, deaf, stenosi, plan, assess, start, elev, barrier, htn, chang, admit, aortic, per, portugues, name, abl, sat, arriv, communic, sever, italian
  - FREX: russian, translat, english, deaf, portugues, languag, interpret, italian, creol, greek, portuges, russian-speak, interpretor, bifasicular, steal, spanish, speak, earlob, nauseaepigastr, etohtobacco, voltag, newact, non-english, scurvi, armr, polish, gestur, -negat, cpod, fin

- **Topic 30 (Family Members) Top Words:**
  - Highest Prob: name, mother, social, will, call, state, sister, work, visit, also, need, contact, brother, friend, father, inform, today, first, time, locat, phone, live, home, get, pts, ask, come, worker, involv, want
  - FREX: sister, brother, worker, friend, father, mother, telephonefax, aunt, phone, said, told, social, dss, live, situat, girlfriend, cousin, school, health, custodi, work, messag, info, know, young, court, inquir, identifi, contactnam, inform

- **Topic 37 (Atrial Fibrillation) Top Words:**
  - Highest Prob: afib, atrial, fibril, rate, lopressor, coumadin, dilt, diltiazem, -fib, control, dose, inr, continu, remain, given, plan, rvr, start, monitor, fib, digoxin, increas, respons, gtt, receiv, metoprolol, assess, stabl, rapid, order
  - FREX: dilt, fibril, diltiazem, afib, -fib, atrial, rvr, digoxin, coumadin, fib, cardizem, dig, diltazem, anticoagul, inr, diltizem, rapid, metoprolol, raf, lopressor, wrvr, afibrvr, paf, rate, diltiazam, paroxysm, warfarin, arrhythmia, esmolol, pvi

- **Topic 48 (Blood Thinner) Top Words:**
  - Highest Prob: cath, ccu, note, groin, stent, lab, puls, sat, resp, cont, sheath, clear, post, toler, stabl, lopressor, dcd, follow, deni, good, progress, site, given, cardiac, nurs, soft, start, monitor, complet, foley
  - FREX: integrelin, integrellin, imi, integrilin, sheath, cpain, aggrastat, integrillin, ptca, bppp, bleedinghematoma, ptcastent, soundsclear, painfre, cks, vsdataev, artven, cpks, palpab, rca, cvri, condom, aggrestat, cpsob, o-msaox, foleycd, cathstent, aivr, reperfus, treponin

- **Topic 63 (Delivery) Top Words:**
  - Highest Prob: will, mom, infant, score, note, bilater, nas, min, normal, name, babi, negat, well, stabl, intact, activ, born, murmur, gram, mean, temp, nicu, admiss, blood, room, sound, rate, mass, regular, clear
  - FREX: nas, ros-, obdoctor, pedifirst, obfirst, fh-otherwis, heent-normocephal, sh-otherwis, exam-gen, cv-stabl, atraumat, opium, exam-inf, score, abstin, id-wil, resp-stabl, dto, imp-inf, pedi-undecid, sh-parent, dss, nrrubella, nas-, lenght, pedi-dr, pedidoctor, atruamat, obhospit, oblast

- **Topic 83 (DNR/DNI) Top Words:**
  - Highest Prob: son, comfort, care, morphin, dnrdni, dnr, made, pts, status, cmo, discuss, code, decis, arriv, want, famili, wish, measur, time, note, team, meet, will, given, make, nurs, appear, place, poor, start
  - FREX: cmo, dnr, dnrdni, wish, hospic, son, pronounc, nephew, priest, dni, autopsi, morgu, expir, decis, death, proxi, measur, hcp, palliat, cathol, dnidnr, funer, prognosi, mortem, morphin, neic, agreement, heroic, die, grave

- **Topic 90 (Transfer) Top Words:**
  - Highest Prob: hospit, transfer, osh, manag, arriv, name, admit, ward, show, found, initi, note, admiss, floor, locat, icu, old, upon, report, start, present, evalu, receiv, place, male, unit, prior, stabl, went, develop

– FREX: hospit, osh, transfer, manag, ward, arriv, found, floor, locat, upon, admit, admiss, icu, show, old, initi, hosp, evalu, went, report, outsid, male, taken, name, avail, develop, unit, year, prior, present

- **Topic 94 (Mood Disorder) Top Words:**
  – Highest Prob: agit, haldol, confus, bed, restraint, restless, pull, safeti, continu, attempt, time, given, frequent, wrist, get, requir, calm, increas, monitor, reorient, prn, ativan, alarm, dose, sitter, place, soft, effect, sedat, period

  – FREX: haldol, agit, restraint, restless, safeti, zyprexa, rail, combat, confus, aggit, pull, wrist, alarm, restrain, sitter, siderail, calm, reorient, thrash, redirect, qtc, yell, mitt, kick, disori, precedex, climb, posey, seroquel, re-orient

- **Topic 102 (Patient Condition) Top Words:**
  – Highest Prob: clear, note, resp, neuro, urin, sat, soft, abd, ectopi, floor, diet, void, deni, lung, skin, nurs, yellow, intact, alert, orient, well, stabl, plan, nsr, foley, good, afebril, sob, mae, social

  – FREX: aaox, pleasant, axox, diet, commod, bedpan, aao, guvoid, anzemet, dns, univers, aox, periph, cracker, gingeral, admissionprogress, spontcommand, cooper, maex, cvpulm, bp-s, cos, non-tend, flds, viii, toast, supper, pivx, rac, appetit

- **Topic 104 (Dialysis) Top Words:**
  – Highest Prob: dialysi, renal, line, today, remov, anur, cath, note, receiv, plan, hemodialysi, fistula, place, fluid, stool, micu, cont, start, sent, pain, dialyz, will, day, quinton, cathet, tomorrow, blood, arm, follow, prior

  – FREX: dialyz, quinton, dialysi, plasmapheresi, thrill, anur, permacath, tank, hemodialysi, dwell, bruit, explod, heaveswretch, pheresi, abusedepend, bruitthril, witj, dializ, allergist, cloud, quentin, ultrafiltr, abn, dust, tessalon, thrillbruit, dialys, methylnaltrexon, plasmaphoresi, -rd

- **Topic 109 (HBP Medication) Top Words:**
  – Highest Prob: sbp, gtt, nitro, ntg, niprid, wean, titrat, keep, hydralazin, lopressor, start, goal, increas, maintain, labetolol, control, given, mcgkgmin, nicardipin, clear, nsr, sat, dose, labetalol, esmolol, good, per, med, remain, monitor

  – FREX: niprid, nitro, ntg, snp, nitroprussid, labetolol, labatelol, labatolol, nicardapin, esmolol, norvasc, labetelol, nicardipin, disect, nitroglycerin, labetalol, antihtn, carden, zestril, accupril, hydralzin, labetol, hydralazin, lowgrad, nirpid, -mgmin, cpsobdyspnea, lobetolol, hydralyzin, esmelol

- **Topic 110 (Respiratory Distress Syndrome) Top Words:**
  – Highest Prob: cpap, resp, cont, note, mild, remain, retract, care, monitor, continu, fio, clear, wean, infant, increas, respiratori, stabl, support, will, sat, need, toler, spell, shift, npo, nasal, place, breath, status, soft

– FREX: rds, surfact, survanta, nco, surf, pectus, prong, cbg, hmd, ccmin, ncpap, subcost, hyalin, retract, pecti, cpap, scr, cmho, ampgent, premat, cannula, cbgs, gfring, intercost, icsc, dlyte, wga, -ccmin, ampig, cannual

- **Topic 123 (Premature Babies) Top Words:**
  – Highest Prob: infant, feed, continu, cont, remain, note, fio, nasal, care, will, cpap, name, soft, lasix, plan, resp, flow, crib, retract, clear, hernia, support, monitor, today, cannula, increas, mild, need, well, activ

  – FREX: cld, prune, rop, bob, diuril, prednisilon, nco, bounci, prilosec, aldacton, hmf, prednisolon, hypospadia, hypospadius, hfnc, zantac, gday, bpd, tobradex, stuffi, calciferol, diruil, -bs, scr, tfr, nutrient, mildmod, mwf, diurilkcl, inguin

- **Topic 125 (Insulin) Top Words:**
  – Highest Prob: gtt, insulin, titrat, drip, sugar, blood, start, endo, remain, given, per, glucos, dose, increas, continu, uhr, lasix, nph, unit, obes, hour, protocol, control, restart, monitor, unitshr, scale, current, infus, elev

  – FREX: insulin, nph, sugar, drip, uhr, gtt, titrat, obes, reg, endo, unitshr, -uhr, glucos, endocrin, fingerstick, fsbs, unithr, accord, protocol, insul, adjust, slide, scale, regular, restart, endo-insulin, -unitshr, endoinsulin, glu, qhr

- **Topic 140 (Drug Abuse) Top Words:**
  – Highest Prob: methadon, abus, use, cocain, hep, drug, narcan, heroin, ivdu, receiv, last, substanc, polysubst, ago, tox, show, plan, urin, recent, narcot, also, ivda, hepat, hiv, name, status, micu, monitor, opiat, team

  – FREX: cocain, heroin, ivda, ivdu, methadon, substanc, opiat, polysubst, addict, marijuana, narcan, meth, psa, fundus, lochia, abus, methodon, crystal, overus, hep, intraven, tox, user, vpc, timelin, vpb, drug, requirng, rhabdomyolysi, hcv

- **Topic 154 (Rash) Top Words:**
  – Highest Prob: rash, red, appli, area, powder, nystatin, cellul, order, note, bodi, cream, back, yeast, miconazol, antibiot, thigh, lotion, infect, groin, name, extrem, buttock, rais, last, sarna, itch, abx, benadryl, sent, lower

  – FREX: rash, cellul, miconazol, nystatin, powder, sarna, lotion, itch, yeast, itchi, red, perineum, erythemat, trunk, cream, topic, dermatolog, bodi, appli, erythema, benadryl, thigh, derm, fungal, ointment, peri, torso, antifung, area, dermat

- **Topic 156 (Asthma) Top Words:**
  – Highest Prob: asthma, throat, swell, prednison, stridor, sore, steroid, benadryl, monitor, edema, present, difficulti, note, tongu, receiv, reaction, solumedrol, allergi, upper, sat, day, breath, without, home, laryng, swallow, admit, start, ent, dose

  – FREX: myasthenia, angioedema, heliox, racem, asthma, nif, stridor, laryng, anaphylact, goiter, throat, mestinon, sore, benadryl, allerg, adr, epipen, prednison, stidor, hoars, uvula, solumedrol, cepacol, pyridostigmin, ade, thyroid, thyroidectomi, hive, anaphylaxi, nif-

- **Topic 178 (Renal Diseases) Top Words:**
  – Highest Prob: renal, chronic, diseas, failur, kidney, end, stage, esrd, plan, today, crf, assess, respons, baselin, lab, monitor, bun, dialysi, urin, cad, follow, fistula, creat, team, cri, creatinin, remov, ckd, will, buncr

  – FREX: esrd, crf, kidney, ckd, end, mwf, nephropathi, nephrectomi, stage, hyperpotassemia, amyloidosi, diseas, nosocomi, abluterolatrov, tropck, leukeran, renal, rcc, pmhcopd, chronic, mrar, cadaver, kayaxel, monwedfri, hyperkalemia, creatin, urticari, glomerulonephr, polycyst, jvp

- **Topic 181 (Anxiety) Top Words:**
  – Highest Prob: anxieti, ativan, anxious, husband, support, time, feel, emot, need, state, sleep, get, want, reassur, refus, given, ask, much, provid, also, night, take, abl, med, day, request, effect, encourag, calm, home

  – FREX: anxious, anxieti, husband, reassur, ativan, emot, xanax, frustrat, refus, relax, cope, feel, ineffect, want, fear, calm, explain, attack, panic, help, explan, die, angri, nervous, slept, someth, talk, much, express, upset

- **Topic 189 (Skin) Top Words:**
  – Highest Prob: skin, area, appli, integr, care, impair, turn, coccyx, cream, continu, barrier, frequent, chang, reposit, breakdown, tear, redden, heel, intact, plan, buttock, excori, note, open, dri, multipl, hour, posit, alo, back

  – FREX: integr, excori, alo, vesta, barrier, cream, fold, coccyx, heel, criticaid, waffl, redden, impair, area, breakdown, cleans, skin, perian, appli, gluteal, tear, buttock, moistur, scab, blanch, boot, reposit, blanchabl, mepilex, antifung

- **Topic 190 (Diabetes) Top Words:**
  – Highest Prob: diabet, insulin, type, blood, sugar, scale, mellitus, unit, slide, glucos, humalog, hyperglycemia, plan, dose, continu, monitor, per, given, glargin, lantus, receiv, cover, coverag, regular, check, assess, fsbs, start, dka, hour

  – FREX: diabet, mellitus, humalog, dka, hyperglycemia, glargin, lantus, type, metformin, ketoacidosi, iddm, glyburid, sugar, hyperglycem, gap, fsbs, slide, dmii, anion, hypoglycemia, hypoglycem, humolog, fingerstick, keton, insulin, glipizid, hiss, bedtim, niddm, scale