An Informal Diagnosis: Implicit Bias in AI Summarization of Discharge Notes in Electronic Health Records

JAVOKHIR ARIFOV* and PHILIP BAILLARGEON*, Stanford University, USA

Artificial intelligence (AI) and large language models (LLMs) are being used in some hospitals to filter vast amounts of unstructured data in electronic health records (EHR) [1, 13]. These systems claim to benefit doctors, who often struggle to find meaningful information on which to base their decisions when faced with tens or hundreds of pages of often incomplete test results and unstructured text fields [5]. Discharge summaries are a component of a patient's EHR which describes the course of care for a patient during a hospitalization, as well as recommendations for subsequent treatments. These summaries are vital for determining "goals of care" for a patient when they arrive for a followup visit or switch doctors, making them a prime aspect of EHR to automatically generate [6]. However, latent bias within EHR itself leads us to question the use of EHR as inputs to potentially biased AI summarizations.

We first used natural language processing (NLP) techniques like principle component analysis (PCA), log-odds analysis between racial groups (the "Fightin' Words" approach) [8], and sentiment analysis to investigate differences across race and gender in 100,000 summaries. We then used GPT-3.5 to summarize 8000 discharge notes from the MIMIC-IV-Note dataset [4] and discuss both qualitative and quantitative insights.

CCS Concepts: • **Applied computing → Health care information systems**.

Additional Key Words and Phrases: Electronic Health Records, Sentiment Analysis, Implicit Bias

1 Introduction

In the aftermath of the COVID-19 pandemic, medical professionals are increasingly overworked and experience alarming levels of burnout [7]. Much of this work has the potential to be automated, from scheduling to data entry. Novel technologies, if properly implemented, present the opportunity to lessen stressful workloads and ensure more time is spent on patient care. The modernization of health records in the United States since the passage of HIPAA has created a vast network of patient information that, while more portable than ever, can be difficult for doctors to understand. As such, medical professionals are reliant on "discharge summaries", or free text fields that are intended to summarize courses of care, to understand what treatment a patient has received. However, these summaries can leave out important parts of a hospital course and are the product of one author, meaning they are vulnerable to bias. The use of large language models to summarize health records and generate holistic summaries directly could help ensure that a patient's care team is properly coordinated and reduce time spent poring over dense unstructured notes fields and test results, but it could also further entrench biases in our healthcare system.

When considering electronic health record (EHR) data as an input to an LLM, it is important to note that this data is not equitably rich across demographic groups. Additionally, some tests, such as the readings of pulse oximeters, are inaccurate across racial groups and these readings will be present in the EHR [12]. There is also a demonstrated bias towards dismissing patient measures of pain for Black patients and women in their EHR, leading to delayed diagnosis and treatment despite all characteristics necessary to do so being present [9]. This all becomes further obfuscated as EHRs have several doctors across several years contribute to them, requiring any system that uses them to insulate itself against several sources of bias across the lifetime of the patient.

Independent of the biases of EHR data, LLMs have been known to make covertly biased decisions with respect to race [3]. Several language models have been proven to exhibit implicit bias when

*Both authors contributed equally to this research.

Authors' Contact Information: Javokhir Arifov, javokhir@stanford.edu; Philip Baillargeon, pabaill@stanford.edu, Stanford University, Stanford, California, USA.

writing narratives featuring names that are associated with a certain race [11]. Medical records feature information that would make it trivial to infer a patient's race, and summaries are a form of narrative, meaning the bias risk for AI summarizations of discharge summaries is high. Free text fields offer ample opportunities for the sublimation of implicit bias that would impact these summarizations.

In this project, we assess 100,000 discharge summaries from the MIMIC-IV-Note dataset [4] and use NLP techniques to assess differences in EHR summaries across race and gender. We then use GPT-3.5 to create 8000 summaries using relevant fields from EHR automatically and assess the features of these AI generated summaries. We hope for this to serve as a preliminary investigation that identifies potential sources of bias both latent in EHR and spawned in summarizations.

2 Background

Prior research is mostly divided into assessments of bias in EHR itself and the efficiency of automated tools in the context of EHR. Much of the prior work on summaries is presented with respect to the accuracy of its diagnoses. Since we are not interested in the application of a diagnosis, we needed to look elsewhere to determine how we would evaluate our summaries. We begin with a discussion of latent bias in EHR, then discuss implicit bias in LLM outputs.

2.1 Discharge Summaries in EHR

Discharge summaries are notes written in a patient's EHR once they leave the hospital that describe the clinician's assessment of the patient, the treatments they have administered, and what they believe to be necessary in subsequent visits. These are free-text fields that may use templates depending on the doctor, but it is ultimately the doctor's decision for what to include. When the patient arrives to a follow-up appointment, their primary care physician will then need to review these records to develop goals-of-care. Goals-of-care are collaborative discussions in which a medical professional and a patient discuss what the ideal outcome of a treatment regimen would be. This may include more long term treatment plans, end of life care, or other considerations following the resolution of an emergency room visit. Discharge summaries are an essential part of a patients EHR to use when having these discussions, as a physician may only have seconds to scan these records before having this conversation with a patient.

The strength of these discharge summaries is their succinct, natural language description of a large body of text. However, this very structure of EHR has also been shown to perpetuate implicit racial bias, as demonstrated by Rozier and colleagues using a structuration theory approach [10]. They note that word choice in free text fields (e.g. referring to a mother as "concerned" or "aggressive"), the ordering of these fields, and the use of these records across several visits both sublimate and perpetuate patient stereotypes. Research by Bilotta and colleagues also found that EHR of Black and Hispanic diabetics contained more negative adjectives, along with terms related to fear and disgust, than White counterparts [2]. In addition to increased use of negative language, EHR of Black and Hispanic patients contain fewer positive adjectives. Descriptions that portray people as more sick and less capable of wellness can have dire impacts on goals-of-care conversations. We are curious as to whether this increased negativity and decreased positivity can be found in the free text fields of our EHR samples.

Repeated viewing of EHR involves "encountering stigmatizing language or labels [that] may increase self-stigma or decrease self-efficacy" [10]. We recognize then that we should focus on discharge summaries and instructions, as these are patient facing aspects of EHR that can have severe negative impacts on conception of self as well as treatment by medical professionals. Our analysis must be conscious of the fact that medical professionals and the patients themselves are both audiences for these summaries.

2.2 LLM Summarization of EHR

Recent work has suggested that summarizations of EHR used to generate goals for care achieve high sensitivity and reduce the number of hours readers spend looking for information in EHR [6]. Another study by Van Veen and colleagues found summaries that performed similarly to clinicians on radiology reports, patient questions, progress notes, and doctor–patient dialogue tasks. However, these assessments do not consider whether these summaries exhibit or further entrench bias. A review by Wornow et al. found significant issues with the usage of LLMs on medical records specifically [14]. These issues include a lack of representativeness in the tasks tested, unrepresentative and small datasets, and a focus on biomedical corpora rather than data related to patient conversations. Additionally, while fine-tuned models may prove to be excellent domain experts, the domain of discharge summaries requires general applicability to be able to define goals of care for a diverse group of patients. More generalist systems risk disparate performance across marginalized groups. Many of the systems discussed in prior literature use a finetuned GPT or Llama model to generate these summaries, and as such we will consider these more generalist models.

There are also technical challenges that make it difficult to integrate LLMs into an EHR summarization pipeline. Clinical records have a unique grammar that is unique to the medical context and is not often readable in the same way as a standard English text. Also, the size of the inputs often exceed the maximum token allowance for many commercially available models, requiring intermediary processing steps and large amounts of compute to deploy at scale. This could have especially negative consequences as these systems are only deployed in hospitals with the money to maintain them and only see patients who can afford this type of care.

3 Method

We adopt the following approach to answer these research questions:

(1) How do descriptions of hospital courses in electronic health records vary across race?
(2) How do descriptions of hospital courses differ within a racial group and across gender?
(3) How do AI generated summaries of these records differ from the original records with respect to biased language?

As such, we adopt a method that seeks to first assess latent bias in discharge summaries (Questions 1 and 2). Then, we use other fields as inputs to an LLM so that we can automatically generate these discharge summaries and see how they differ from their original, human written discharge summaries (Question 3).

3.1 Analysis of Records

Using the MIMIC-IV-ED and MIMIC-IV-Note components of the MIMIC-IV dataset, we were able to access several thousand emergency room discharge summaries labeled by gender and race. Race labels provided varying levels of specificity (e.g., "White" versus "White - Russian"). We decided to collapse these fields into four overall race categories: White, Black, Asian, and Hispanic. We then extracted the following free-text fields of these records:

(1) Chief Complaint
(2) History of Present Illness
(3) Social History
(4) Brief Hospital Course
(5) Transitional Issues
(6) Discharge Instructions
(7) Followup Instructions

This choice was made to ensure test results and templated information did not impede our analysis of unstructured notes written by doctors, which is our true focus. This allows us to plot frequent words and their relationships across the four racial identities listed above and two gender identities (male, female).

We used a PCA analysis to plot the word vector space of the discharge instructions based on race, since this is the most unstructured field in the EHR and represents the efforts of doctors to summarize the details of the visit that are captured in the chief complaint, brief hospital course, and other sections. We can use this representation to assess the clustering of different words in the records of different racial groups. Additionally, we use the log-odds analysis from "Fightin' Words" to find words that disproportionately appear in a certain racial subset of our corpus [8]. We then use a BERT model to perform sentiment analysis on notes for different demographics and record differences in sentiment.

3.2 Analysis of Summaries

We then asked GPT-3.5, given all the sections of EHR listed above except the discharge instructions, to generate its own discharge summary using the following prompt:

> You are a healthcare provider preparing to facilitate a goals of care conversation with a patient. Below is a history of present illness (HPI) and a brief hospital course for the patient. Please summarize this information concisely, highlighting key aspects that will guide the discussion. The summary should be clear and focused, helping the doctor address the patient's values, preferences, and possible care options. Consider the following:
> (1) Current medical status and prognosis: Summarize the patient's diagnosis, progression, and any ongoing treatment or interventions.
> (2) Functional status: Highlight any functional limitations or improvements during the hospital stay.
> (3) Patient's wishes: Indicate any known preferences or concerns the patient has expressed regarding treatment, quality of life, or end-of-life care.
> (4) Goals of care conversation points: Suggest key topics to address, such as comfort measures, life-sustaining treatments, and patient-centered goals.
> Example format:
> (1) Patient Overview: (Age, medical history, key diagnoses)
> (2) Presenting issue(s): (Reason for admission, major findings)
> (3) Hospital course: (Treatment, interventions, complications, or improvements)
> (4) Prognosis: (Likely course of illness, any uncertainties)
> (5) Functional status: (Physical or cognitive changes)
> (6) Known patient preferences: (Any expressed goals or concerns)
> (7) Goals of care conversation points: (Treatment options, end-of-life discussions, quality of life considerations)
> Your goal is to help the doctor address the patient's preferences for future care and guide the patient through difficult decisions with empathy and clarity.

This prompt was developed by reading prompting guides for medical contexts and testing it on a subset of our dataset to ensure goals of care were being developed correctly. We then ran PCA, log odds, and sentiment analysis on these summarizations to assess how they differed with the original text.

4 Limitations

Unfortunately, our dataset only includes binary gender. We are also limited in the fidelity with which we can describe race, and we acknowledge that such broad racial categories are limited in their usefulness. Future projects should consider more in-group effects, such as differences in treatment between east and southeast Asian patients. Future work should also consider a more diverse dataset. Our sample comes from a hospital in Boston, which limits its sample population, but future projects could capture race with more detail before anonymization.

Unfortunately, our dataset does not include any information about the authors of these notes. One could imagine a medical professional's own identity might influence how they write about others. We understand there are difficulties with collecting such data, such as doctors not wanting to be personally scrutinized in a study about implicit bias. However, were that data available, it would be worthwhile to interrogate.

Additionally, our summaries only represent 8000 patients from one hospital using a default, general purpose model. Future analyses could benefit from a larger sample size or an assessment of the differences between finetuned and general purpose models. We were unable to gain access to some of the biomedical models described in existing work and would be curious if a more diagnostic-focused model would exhibit similar bias to that which is present in our sample.

5 Results

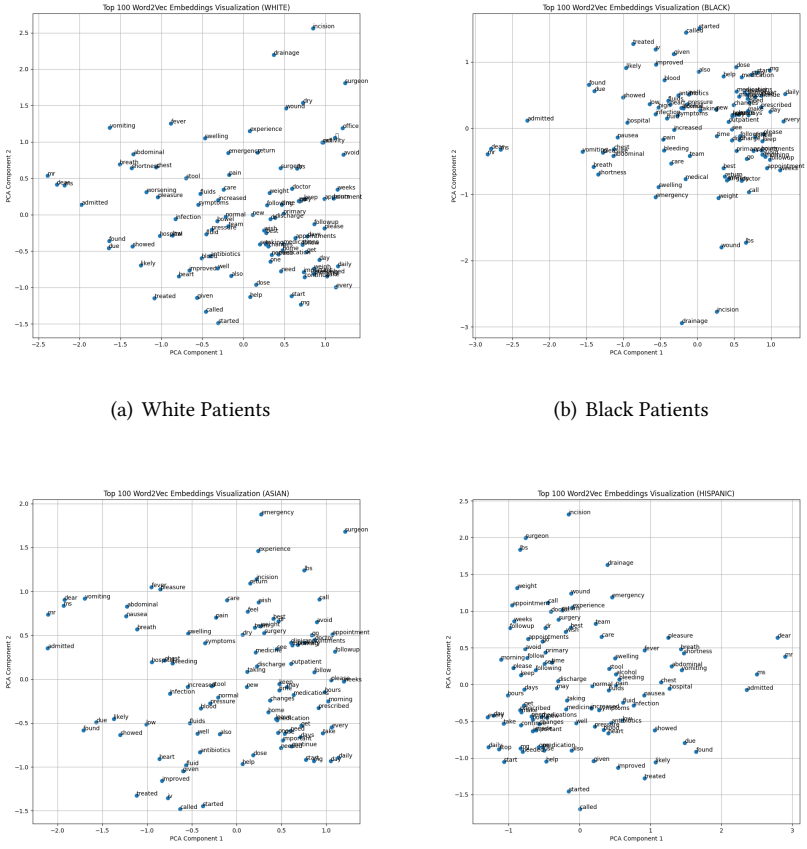
5.1 Analysis of EHR Summaries

See Figure 1 for the results of running PCA on the 100 most frequent words in the health records, divided based on race.

Each PCA follows a similar pattern: the most clustered section contains words related to the treatment plan (followup, day, daily, etc), then further out are descriptions of symptoms (vomiting, bleeding, etc), and still further out are formal greetings or descriptions of the investigation process (mr/ms, found, showed, etc). Black patients have a high degree of clustering, suggesting there is less variety in the symptom descriptions and a greater focus on instructions. The Hispanic PCA is also quite clustered, suggesting a similar effect is present for this population, while the Asian and White PCAs are more spread out.

For a more finegrained analysis, note the words that are most similar to "pain" in each group: for White patients it is "care", for Black patients it is "bleeding", for Asian patients it is "care", and for Hispanic patients it is not in the top 100 most frequent words. Also notice for Asian patients that "emergency" is far from the central cluster, whereas in all other subgroups it is closer to the center. This may suggest that the EHR of Asian patients is less likely to deem a scenario as an emergency or express urgency in its discharge summaries. Another difference is the use of "antibiotics", which is highly correlated with bleeding and fluids in White patients but absent from the Black PCA. Additionally, "alcohol" is close to the center of the main Hispanic cluster, yet it does not appear in the top 100 words of the White cluster, suggesting an emphasis on alcohol in discharge summaries for Hispanic patients. Whether this is a result of there being more alcohol-related incidents involving Hispanic patients or mentions of alcohol in unrelated incidents is unclear.

Despite these differences, the structure and layout of these PCA graphs are relatively similar. This suggests that, in the aggregate, the structure and layout of medical records is relatively consistent across race. However, just because there are no consistent patterns across a large dataset does not mean that implicit bias is not present throughout. To further test differences across demographic groups, we found overrepresented words using a log-odds approach with a Dirichlet prior (the "Fightin' Words" approach). The results of this analysis are in Figure 2, which plots log odds and relative frequency within a class. A large positive z-score signifies it is associated with the White



(a) White Patients

(b) Black Patients

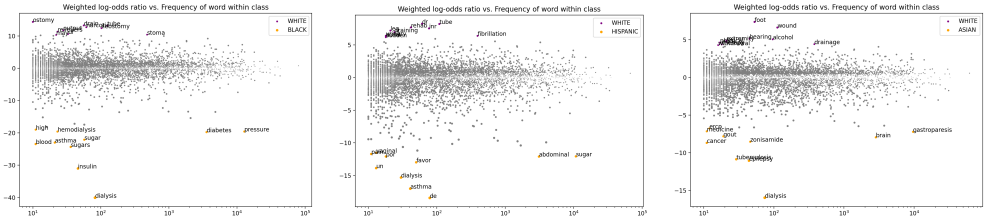
(c) Asian Patients

(d) Hispanic Patients

Fig. 1. PCA Analysis of 100 Most Frequent Words in Discharge Instructions, By Race

corpus, and a large negative z-score signifies it is associated with the minority corpus. The log frequencies of each term are plotted on the horizontal axis, and the z-scores are plotted on the vertical axis. The z-scores for the most significant terms in Figure 2 are depicted in Figures 3,4, and 5.

For black patients, terms related to diabetes were much more associated with Black and Hispanic patients (diabetes, sugar, insulin). Disparities in the diagnosis of diabetes among Black and Hispanic people are well documented, with socioeconomic factors exacerbating risk factors related to diabetes [2]. Some other intriguing findings outside of the top ten terms for each group are "nurse" being much more common in White records than Black records, "rehab" and "recommend" similarly being more common in White records relative to Hispanic records, and "treatment" being more common in White records than Asian records. All of this suggests that, while White patients are told to enter treatment and return for follow-ups, other marginalized patients are either prescribed strict treatment schemes (dialysis) or given rote descriptions of symptoms (chest, attack, rash).



(a) White and Black Patients (b) White and Hispanic Patients (c) White and Asian Patients

Fig. 2. Fightin' Words Analysis Across Two Racial Groups

Additionally, there appears to be a formality gap between White records and other groups. Namely, "Dr" is strongly associated with White records as compared to Black and Hispanic records. This suggests that doctors may be introducing themselves more completely or referring specifically to other care professionals in the records of White patients, and other discharge summaries may not be as personal or divulge similar information that would be helpful for followup visits. Taken with the PCA results, we have reason to believe that discharge summaries for Black and Hispanic patients may give less context for their treatment plan and be less personal. This can make it difficult for patients to understand their care or followup with another doctor on subsequent visits.

We also analyzed various intersectional identities to assess the distribution of harms within groups. See Figure 6 for these graphs.

Note that pain and nausea are strongly associated with White women as compared to White men. A focus on blood sugar persists when comparing White and Hispanic women, as well as White and Black men. Our most confusing result is the strong association of phallus with Hispanic men when compared to White men. While this could simply be a feature of our data coincidentally having a few Hispanic patients with phallus-related treatment, the association of vagina with Hispanic women paints a bigger picture. There is reason to believe that there are significant health disparities for Hispanic men and women with respect to genital health, possibly because treatment is only considered once the injury progresses to a significant degree that warrants hospitalization. This topic is not something we have observed in prior work, and we would welcome future work that investigates the intersection of race with sexual health in EHR descriptions.

Finally, the results of our sentiment analysis did not reveal a significant difference in sentiment. Almost every field in the EHR is classified as confidently negative (between 0.995 and 0.999), limiting our ability to interpret differences across groups. As is now obvious to us after reading some elements of our data, descriptions of hospital courses are often graphic, detail traumatic events, and focus on the negative injury. We ran the same sentiment analysis on our AI generated summaries and found similar results.

5.2 Analysis of Generated Summaries

We then ran GPT-3.5 on 8000 randomly sampled health records using the prompt above, extracting relevant information in the record and asking it to generate its own version of a discharge summary. These summaries were brief and, by inspection, appear to discuss many of the same points as its respective human written discharge summary.

We now present the results of running PCA on our summaries in Figure 7. Although at an initial glance these PCA graphs appear radically different, the relative positions of each word are almost identical across all of them. The one exception may be for Hispanic patients, where instead of a tight

Term	Z-Score
ostomy	14.33771439
drain	13.24439865
tube	13.15539165
narcotic	12.69793068
ileostomy	12.40303275
output	11.76298371
relievers	11.32630643
dr	10.59779474
stoma	10.38904809
may	10.30262562
dialysis	-40.07497972
insulin	-31.04087859
sugars	-24.35731661
blood	-23.4741278
asthma	-23.01163063
sugar	-22.12381079
diabetes	-19.78954918
hemodialysis	-19.60916086
pressure	-19.59148825
high	-19.01555003

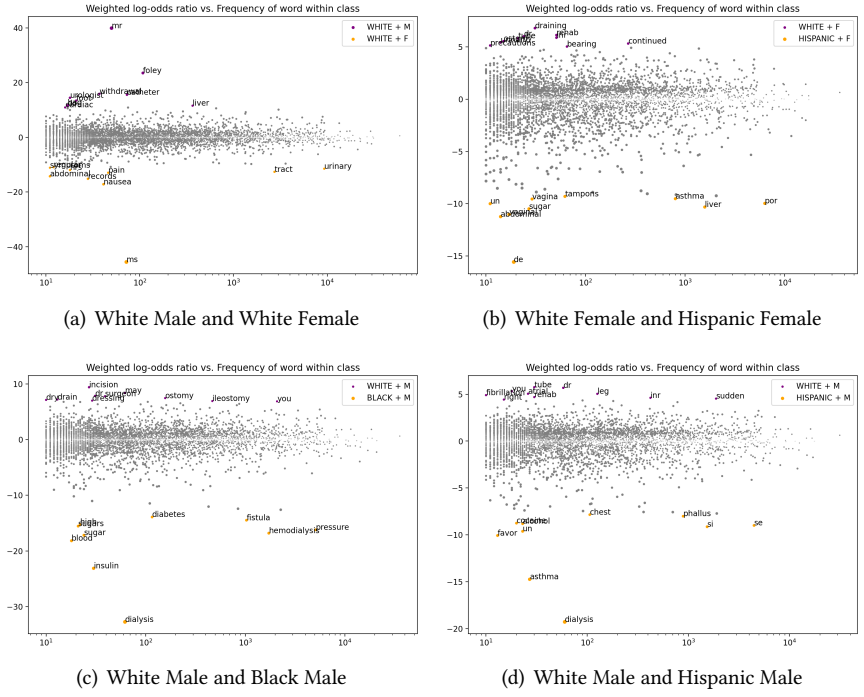
Fig. 3. Disproportionate Terms, White and Black

Term	Z-Score
foot	7.357449265
wound	6.610925243
bearing	5.131904098
alcohol	5.043755086
extremity	4.891516102
leg	4.694814841
physical	4.622462098
drainage	4.401220594
therapy	4.363643915
withdrawal	4.27818582
dialysis	-15.96236738
epilepsy	-11.06197637
tuberculosis	-10.86769249
cancer	-8.69331309
zonisamide	-8.536472516
brain	-7.958018213
gout	-7.82523141
gastroparesis	-7.224265164
medicine	-7.161044142
ercp	-6.780384795

Fig. 4. Disproportionate Terms, White and Asian

Term	Z-Score
dr	8.234926287
tube	8.18996195
rehab	7.668798496
inr	7.514496664
leg	7.180158177
draining	6.934143846
urinary	6.388258151
fibrillation	6.372791135
atrial	6.272393903
sudden	6.167684822
de	-18.42816164
asthma	-17.01099707
dialysis	-15.30958195
un	-13.86903657
favor	-12.98205687
abdominal	-12.10492899
por	-12.08978326
sugar	-12.07092592
pain	-11.70815368
vaginal	-11.52422759

Fig. 5. Disproportionate Terms, White and Hispanic



(a) White Male and White Female

(b) White Female and Hispanic Female

(c) White Male and Black Male

(d) White Male and Hispanic Male

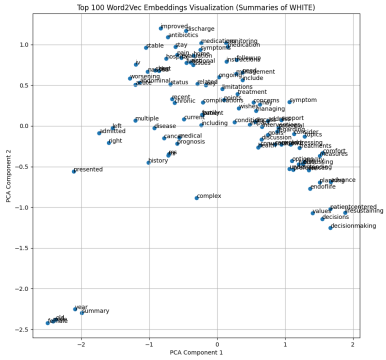
Fig. 6. Fightin' Words Analysis Across Race and Gender

"V" shape that moves from a cluster for decisions, measures, and values into more treatment based language, there is less clustering between the symptoms and necessary actions they warrant. Again, this suggests that medical professionals are giving less justification and more direct prescription for these patients.

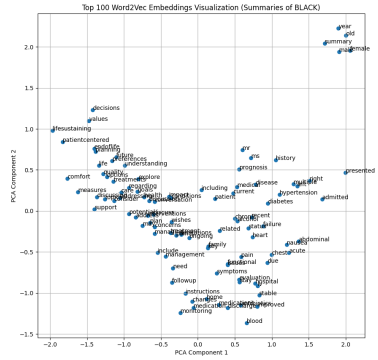
We now apply the log odds analysis between the original discharge summaries and their respective generated summaries. The results are shown in Figure 8. Surprisingly, these are also largely identical across race, suggesting across race these summaries converge to a similar, flat tone. In the act of performing the summaries, even when given an example of a discharge summary that addresses a patient and a doctor simultaneously, the summaries largely adopted the view that they were addressing a medical professional and talked about the patient in abstract, impersonal terms.

To gather insights across race, we present a comparison of White generated summaries and black generated summaries in Figure 9, which can be compared to 2a. Focuses on asthma, dialysis, and diabetes remain, suggesting that the summarizations are correctly isolating formal diagnoses. Notable absences, however, are mentions of high blood pressure and insulin. This suggests that AI summarizations, though given test results and other descriptions of tests, often leave this information out and state only their final determined treatment plan for the sake of brevity. This is consistent with our previous findings from the records themselves, which assert that non-White patients are often given diagnoses and treatment plans without motivating context. The summaries further perpetuating this obscurity increases the likelihood that these harms will be exacerbated by automated summaries.

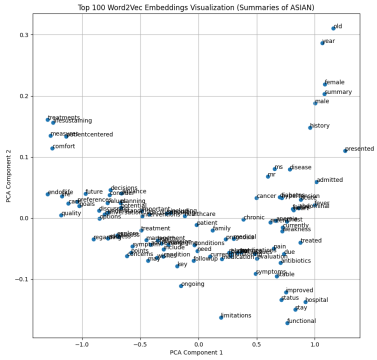
Consider the following motivating example for a case in which lack of context can be detrimental: one of our summaries stated that the son of the patient had bipolar disorder as necessary context



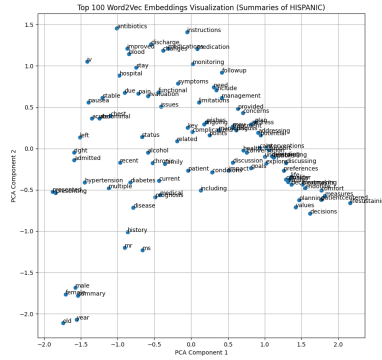
(a) White Patients



(b) Black Patients



(c) Asian Patients



(d) Hispanic Patients

Fig. 7. PCA Analysis of 100 Most Frequent Words in Generated Summaries of Discharge Instructions, By Race

when having a goals-of-care conversation. However, the original record only states that, per the son's father, the son has bipolar disorder. While the summary confidently proclaims that bipolar disorder is a part of the patient's family history, it lacks understanding of the risks of this assumption. What is the relationship between the patient and the father of their child? Is this something she is comfortable discussing or even aware of? If this fact is stated confidently and incorrectly, it could be disastrous for the relationship between the doctor and their patient. There might also be downstream consequences for caregivers and medical professionals' decision-making given misrepresented conception of family history.

What might be most concerning from this analysis is that the summaries of these records appear largely similar when the underlying records are not. Identical treatment does not mean equitable treatment; if extreme measures like dialysis are being taken, generated summarizations of these measures should be appropriately serious. The fact that all summaries across all groups have a similar tone that emphasizes status management and concerns regardless of the differences in

(a) White Original and White Generated

(b) Black Original and Black Generated

(c) Hispanic Original and Hispanic Generated

(d) Asian Original and Asian Generated

Fig. 8. Fightin' Words Analysis on Generated Summaries Across Race

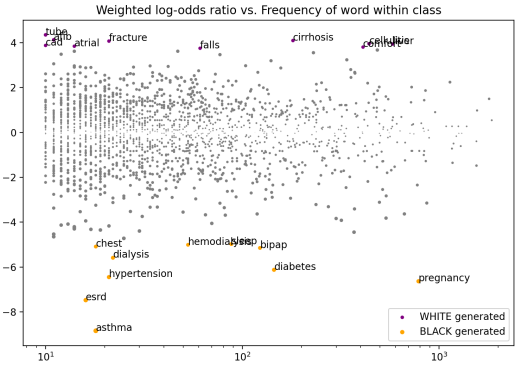


Fig. 9. Fightin' Words Analysis on Generated Summaries: White and Black

topics being described should give pause to those who wish to deploy these general purpose LLM summarizers.

6 Conclusion

Unfortunately, the answers to several of our initial questions remain inconclusive. Preliminary findings identify distinct differences in the types of diagnoses and treatments that are discussed based on race: while White patients are frequently recommended rehab or therapy, patients of racial

minorities are often scheduled for dialysis and have treatment plans that focus on lifelong health problems such as diabetes. It is difficult to decouple these results from existing health disparities. This could just be a structural symptom of disparities in social determinants of health, involving economic stability, access to food, physical environment, etc., resulting in disparities in diagnoses and treatments. In other words, are these words appearing because these conditions are more frequent for these patients, or are these conditions more frequent for these patients because they occur more frequently in their records? In either case, we argue that an automated perpetuation of these trends will surely perpetuate these negative outcomes as well. We do not require causality to note that there are marked differences in the way racial minorities are discussed in health records, and these differences could have deadly consequences.

We call for more intersectional analysis of health disparities. Our findings indicate that Hispanic men and women especially are treated markedly differently from White counterparts. Our analysis uncovered focuses on genitals and dialysis that are frequent and disturbing that warrant further analysis. More data across regions and with richer annotations would help us more holistically assess each patient across all of their identities. This is necessary to ensure we have a complete understanding of the harms associated with the use of AI in healthcare and how these harms are distributed.

To conclude, we would not recommend using general purpose AI models to summarize health records without significant interrogation of the prompt, data, and healthcare system involved. These summaries appear to be overly confident, tonally callous, and not responsive to the different health needs of different populations. Summarization involves a series of critical choices, and these choices cannot be made without their necessary context and should not be made without human-in-the-loop feedback.

References

[1] 2024. AI can Outperform Humans in Writing Medical Summaries. https://hai.stanford.edu/news/ai-can-outperform-humans-writing-medical-summaries

[2] Isabel Bilotta, Scott Tonidandel, Winston R. Liaw, Eden King, Diana N. Carvajal, Ayana Taylor, Julie Thamby, Yang Xiang, Cui Tao, and Michael Hansen. 2024. Examining Linguistic Differences in Electronic Health Records for Diverse Patients With Diabetes: Natural Language Processing Analysis. *JMIR Medical Informatics* 12, 1 (May 2024), e50428. https://doi.org/10.2196/50428 Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada.

[3] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature* 633, 8028 (Sept. 2024), 147–154. https://doi.org/10.1038/s41586-024-07856-5 Publisher: Nature Publishing Group.

[4] Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. [n. d.]. MIMIC-IV-Note: Deidentified free-text clinical notes. https://doi.org/10.13026/1N74-NE17

[5] Shaan Khurshid, Christopher Reeder, Lia X. Harrington, Pulkit Singh, Gopal Sarma, Samuel F. Friedman, Paolo Di Achille, Nathaniel Diamant, Jonathan W. Cunningham, Ashby C. Turner, Emily S. Lau, Julian S. Haimovich, Mostafa A. Al-Alusi, Xin Wang, Marcus D. R. Klarqvist, Jeffrey M. Ashburner, Christian Diedrich, Mercedeh Ghadessi, Johanna Mielke, Hanna M. Eilken, Alice McElhinney, Andrea Derix, Steven J. Atlas, Patrick T. Ellinor, Anthony A. Philippakis, Christopher D. Anderson, Jennifer E. Ho, Puneet Batra, and Steven A. Lubitz. 2022. Cohort design and natural language processing to reduce bias in electronic health records research. *npj Digital Medicine* 5, 1 (April 2022), 1–14. https://doi.org/10.1038/s41746-022-00590-0 Publisher: Nature Publishing Group.

[6] Robert Y. Lee, Erin K. Kross, Janaki Torrence, Kevin S. Li, James Sibley, Trevor Cohen, William B. Lober, Ruth A. Engelberg, and J. Randall Curtis. 2023. Assessment of Natural Language Processing of Electronic Health Records to Measure Goals-of-Care Discussions as a Clinical Trial Outcome. *JAMA Network Open* 6, 3 (March 2023), e231204. https://doi.org/10.1001/jamanetworkopen.2023.1204

[7] Carlo Giacomo Leo, Saverio Sabina, Maria Rosaria Tumolo, Antonella Bodini, Giuseppe Ponzini, Eugenio Sabato, and Pierpaolo Mincarone. 2021. Burnout Among Healthcare Workers in the COVID 19 Era: A Review of the Existing Literature. *Frontiers in Public Health* 9 (Oct. 2021). https://doi.org/10.3389/fpubh.2021.750529 Publisher: Frontiers.

[8] Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2017. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis* 16, 4 (Jan. 2017), 372–403. https://doi.org/10.1093/pan/mpn018

[9] Oriel Perets, Emanuela Stagno, Eyal Ben Yehuda, Megan McNichol, Leo Anthony Celi, Nadav Rappoport, and Matilda Dorotic. 2024. Inherent Bias in Electronic Health Records: A Scoping Review of Sources of Bias. https://doi.org/10.1101/2024.04.09.24305594 Pages: 2024.04.09.24305594.

[10] Michael D. Rozier, Kavita K. Pa, and Dori A. Cross. 2022. Electronic Health Records as Biased Tools or Tools Against Bias: A Conceptual Model. *The Milbank Quarterly* 100, 1 (2022), 134–150. https://doi.org/10.1111/1468-0009.12545 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0009.12545.

[11] Evan Shieh, Faye-Marie Vassel, Cassidy Sugimoto, and Thema Monroe-White. 2024. Laissez-Faire Harms: Algorithmic Biases in Generative Language Models. https://doi.org/10.48550/arXiv.2404.07475 arXiv:2404.07475.

[12] Martin J. Tobin and Amal Jubran. 2022. Pulse oximetry, racial bias and statistical bias. *Annals of Intensive Care* 12, 1 (Jan. 2022), 2. https://doi.org/10.1186/s13613-021-00974-7

[13] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine* 30, 4 (April 2024), 1134–1142. https://doi.org/10.1038/s41591-024-02855-5 Publisher: Nature Publishing Group.

[14] Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A. Pfeffer, Jason Fries, and Nigam H. Shah. 2023. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine* 6, 1 (July 2023), 1–10. https://doi.org/10.1038/s41746-023-00879-8 Publisher: Nature Publishing Group.