Interrogating Racism in the Medical Literature Using Word Embeddings

Lauren D. Liao[2], Sajia Darwish[3], Caroline Figueroa[4], Erin Manalo-Pedro[5], Swetha Pola[5], Maithili Jha[6], Fernando De Maio[7,8], Claudia von Vacano[1], Chris J. Kennedy[9,10], and Pratik S. Sachdeva[1]

[1] D-Lab, University of California, Berkeley, Berkeley, USA
[2] Division of Research, Kaiser Permanente, Oakland, USA
[3] Department of Biostatistics, Harvard University, Cambridge, USA
[4] Delft University of Technology, Delft, Netherlands
[5] Independent Scholar
[6] Rimtec Corporation, Addison, USA
[7] Department of Sociology, DePaul University, Chicago, USA
[8] Health Equity Research, American Medical Association, Chicago, USA
[9] Center for Precision Psychiatry, Massachusetts General Hospital, Boston, USA
[10] Department of Psychiatry, Harvard Medical School, Boston, USA

Abstract

The medical literature has an important role to play in establishing anti-racist practice that may alleviate racial health inequities. Recent critical discourse analyses have demonstrated that medical literature often fails to explicitly name racism or discuss it through a structural lens, instead employing euphemistic language that obscures structural determinants of health inequities. Here, we build upon this work by using Word2Vec word embeddings to interrogate a corpus of 871 published articles containing the word "racism" sourced from top medical journals between 1999 and 2020. Our findings reveal distinct patterns in medical discourse around racism. First, hierarchical clustering of discrimination-, power-, and wealth-related words demonstrated clear separation between racism-related concepts and structural determinants, with racism showing minimal similarity to wealth-related words while clustering more closely with other forms of discrimination. Second, we found that qualifying language denoting uncertainty (e.g., "maybe", "possibly") showed higher similarity to racism-related words than more confident language, suggesting qualifying language serves as a hedge against direct assertions about racist processes. Finally, we conducted a network analysis revealing how concepts cluster within medical discourse, with bridging words between health inequities and racism clusters predominantly reflecting interpersonal rather than structural framings of racism. Specifically, words associated with health inequities, such as "stress" and "homelessness" connected to racism primarily through person-level gatekeepers such as "interpersonal," "prejudice," and "overt," while structural concepts remained notably absent from common pathways. Overall, understanding these linguistic patterns is crucial as the medical community works to build anti-racist norms, while simultaneously relying on medical text to train artificial intelligence systems deployed in clinical settings.

Keywords: racism, medicine, digital humanities, medical humanities, embeddings, natural language processing

Lauren D. Liao, Sajia Darwish, Caroline Figueroa, Erin Manalo-Pedro, Swetha Pola, Maithili Jha, Fernando De Maio, Claudia von Vacano, Chris J. Kennedy, and Pratik S. Sachdeva. "Interrogating Racism in the Medical Literature Using Word Embeddings." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 673–690. https://doi.org/10.63744/voHr9u5XsC0n.

1 Introduction

Racism has deep historical roots within the field of medicine [10; 39]. While racism can manifest at the interpersonal level through the conscious or unconscious biases of individual care providers [26; 29], *structural racism* – the broad array of mechanisms and systems by which societies perpetuate racial discrimination [4] – has a deeper and more insidious impact on health outcomes [18]. Structural racism systematically shapes medical practices, establishing self-perpetuating cycles of health inequity that endure even when individual clinicians and practitioners espouse egalitarian values [20]. Structural racism is particularly evident in the medical literature, where the framing of race-related health differences often disconnects observed inequities from their underlying structural causes, either by avoiding the word "racism" entirely or by naming racism without linking it to its role as a driver of health inequities [7; 8; 30].

The manner in which the medical literature discusses racism – which Figueroa et al. term *racism narratives* [17] – shapes policy decisions and clinical beliefs across medicine [23]. For example, the failure of practitioners to explicitly name racism as a causal factor in health inequities research diminishes awareness of structural determinants among readers, leading to interventions that focus on individual behaviors rather than addressing systemic causes [30]. Even when racism is explicitly acknowledged, if practitioners discuss it solely through the lens of interpersonal bias rather than as a structural phenomenon, there is correspondingly less pressure to enact the institutional changes necessary to address root causes of health inequity [15; 17; 22].

This issue has become particularly pressing as medical texts authored by doctors and clinicians are increasingly being used to train large language models (LLMs) and artificial intelligence models in healthcare applications [2; 43; 51]. Structural racism is evident in these medical texts, such as racial bias in the electronic health record [13; 25; 46]. Recent computational studies reveal pervasive racial biases embedded within these systems, demonstrating that LLMs trained on medical corpora exhibit biases in clinical decision-making [41; 53; 54; 55]. As these biased language models are deployed in clinical decision support systems and diagnostic tools, they risk perpetuating and scaling discriminatory practices across healthcare institutions, making it crucial to understand how medical practitioners conceptualize and discuss racism in their written discourse.

Despite these pressing concerns, recent discourse analyses have empirically demonstrated that racism is often not explicitly named in medical literature [11; 23; 30]. Furthermore, there remains comparatively limited work examining racism in medicine from a structural perspective [30]. Notably, Hardeman et al. conducted a review revealing that public health literature avoids naming structural racism as a fundamental cause of health inequities, and thus fails to adequately address them in practice [23]. Krieger et al. constructed a corpus of medical articles mentioning the word racism, finding that they represent a small fraction of the larger medical literature on health inequities [30]. Using this corpus, Figueroa et al. developed a typology of *racism narratives* present in the medical literature, with broad categories spanning "dismissal", "person-level", "societal", and "actionable" [17]. Here, we build upon the foundation of critical discourse analysis established by these works by utilizing computational methods to provide a distant reading of racism in the medical literature at scale.

Natural language processing methods, such as word embeddings, can serve as powerful tools for textual interrogation. Word embedding analyses have revealed gender and ethnic stereotypes in both general language corpora [19; 33] and medical texts, including ICU notes and clinical documentation [13; 25]. In this work, we apply word embeddings to the corpus developed by Krieger et al. [30] to interrogate how racism is discussed and conceptualized within the medical literature. By utilizing a variety of secondary analyses, including hierarchical clustering and network analysis, we identify patterns in how racism is linguistically framed within medical discourse. Our findings provide evidence that medical literature conceptualizes racism through individual-level, rather than structural frameworks, which may have implications for how the field understands and

addresses structural determinants of health inequities.

2 Methods

All code used to carry out the analyses and generate the figures in this paper is publicly available on GitHub.[1]

2.1 Text collection and preprocessing

We conducted our analysis on the corpus of medical articles containing the word "racism" created by Krieger et al. [30]. This corpus consisted of articles obtained from four leading medical research journals – The British Medical Journal (BMJ), The Journal of the American Medical Association (JAMA), The New England Journal of Medicine (NEJM), and The Lancet, published from 1999-2020. The final corpus consists of 871 articles, with 391 articles from BMJ, 128 from JAMA, 91 from NEJM, and 260 from The Lancet. We used raw text extracted from these articles with optical character recognition (OCR).

We applied a robust text cleaning pipeline to each article in the corpus. The pipeline consisted of preprocessing steps commonly performed in natural language processing, as well as additional steps to clean text obtained from OCR. We used a custom preprocessing pipeline rather than a pretrained tokenizer to control the level of granularity of tokens. The pipeline is as follows: (1) Identify and correct valid words separated into two lines connected by a dash (e.g., "`hea- lth`" $\rightarrow$ "`health`"); (2) Remove digital object identifiers, URLs, digits, common punctuation, line separations, and blank spaces between valid words; (3) Remove stop words (e.g., "`the`", "`and`") using a custom curated list; (4) Convert text to lowercase; (5) Replace words in British English with their American English equivalents; [28] (6) Manually correct common misspellings; and (7) Add spaces between words where OCR removed spaces (e.g., "`noblankspaces`" $\rightarrow$ "`no blank spaces`").

After applying the text cleaning pipeline to each article in the corpus, we tokenized the dataset by breaking down each word or word-like unit (e.g., a contraction) into distinct "tokens" using the `nltk` (Natural Language Toolkit) package [5], which streamlines downstream analyses. Importantly, our tokenizing included lemmatization, where words were converted to their root form (e.g., "`roots`"$\rightarrow$ "`root`" and "`rooted`"$\rightarrow$"`root`"). For the rest of this paper, we refer to tokens as "words."

We subsequently expanded the vocabulary by creating bigrams and trigrams via the package `gensim` [44]. The inclusion of bigrams and trigrams allows commonly used pairs and trios of words to be used as single tokens (e.g., `health_disparity` or `social_determinants_health`). We omitted common English connector words when forming bigrams and trigrams (e.g., *also*, *then*, etc.). We only included bigrams and trigrams when they occurred with a minimum frequency phrase count of 10. Bigrams and trigrams were further chosen via a scoring function that assesses the likelihood of co-occurrences of words.

The above preprocessing pipeline ensured a more robust analysis by removing extraneous words from the text, leaving behind a cleaner corpus dataset for subsequent word embedding calculations. After applying the preprocessing pipeline to the corpus, we obtained a vocabulary with 9,242 unique words (including bigrams and trigrams).

2.2 Word embeddings

Overview. Word embeddings are numerical representations of words, constructed based on their usage in large corpora. In practice, the embedding is a vector of numbers that represents a word

[1] `https://github.com/dlab-projects/interrogating_racism_medical_literature`

in a given corpus. Word embeddings are useful because they capture the meanings, semantic relationships, and syntactic properties of the words. While the raw numerical representation of the word is not interpretable alone, the embedding can be used in downstream quantitative tasks to elucidate the underlying structure of the text. For example, word embeddings have been used to perform semantic tasks quantitatively, such as finding synonyms and testing analogies [36].

We employed Word2Vec to construct word embeddings (the continuous bag-of-words variant) [35]. Word2Vec is a model that generates word embeddings by either predicting a target word from its surrounding context words or predicting context words from a target word. A Word2Vec model is generally specified by two main hyperparameters: (i) a context window W that specifies the number of words surrounding a target word used in training, and (ii) a vector size V specifying the dimensionality of the word embedding.

Word embeddings enable analysis of pairwise relationships between words with a similarity measure. The similarity measure quantifies the degree to which two words share conceptual or functional similarities based on how they are used in the corpus. We chose to use the cosine similarity, which quantifies similarity as the angle between the tokens' word vectors. The cosine similarity ranges from -1 to 1, where values closer to 1 represent higher similarity. Although two words with high similarity may not be interchangeable, they are suggested to be used in similar contexts.

Training Procedure. We used the package `gensim` to carry out word embeddings analyses [44]. Due to randomness in the optimization algorithm used by `gensim`, each Word2Vec fit generates different embeddings (and therefore, word similarities). This variation can result in differing interpretations, especially in small corpora. To create a more robust word similarity matrix, we fit word embeddings for 100 random seeds. For each fit, we calculated word similarities between every pairwise combination of words, resulting in a *word similarity matrix* with dimensionality 9242×9242. We then averaged the word similarity matrices across the 100 fits to produce a final word similarity matrix for the corpus.

To select the vector and window size hyperparameters, we evaluated the consistency of word similarities for a set of pre-specified seed words across fits. Specifically, for the pre-specified seed words, we identified which words appeared at least 75% of the time in the list of the top-10 most similar words. We chose the hyperparameter configuration which most consistently maintained the top-10 list. We tested all pairwise combinations of window sizes $W \in \{5, 10\}$ and vector sizes $V \in \{32, 64, 128\}$. In the process of evaluating the consistency of top similar words, we found that a window size $W = 5$ and vector size $V = 128$ produced the most consistent and stable results for this corpus. Other hyperparameters were left at their default values.

2.3 Word Similarity Matrix

Seed Words. We examined word similarities from a subset of the full 9242×9242 word similarity matrix. We used 10 seed words related to (1) discrimination (*homophobia*, *sexism*, *racism*, *discrimination*, *bias*), (2) wealth (*wealth*, *poverty*, *economic*), and (3) power (*power*, *structural*). To choose the seeds words, we began with anchor words for each category of interest: *racism*, *wealth*, and *power*. We chose the final seed words in an iterative fashion, by examining the most similar words to the anchor words that fell within the pre-defined categories. We began with a list of 21 candidate words which we narrowed down to the 10 final seed words with the aim of balancing the number of seeds words in each cluster, the total number of seed words, and the magnitude of word similarities.

Dendrogram. To further elucidate hierarchical structure within the 10×10 similarity matrix, we computed a dendrogram. Dendrograms provide evidence of hierarchical structure in data by sequentially clustering pairs or groups of words. Dendrograms, at their finest levels, pair words with high similarity. These pairings are treated as a unit to form subsequent groupings. The broadest

groupings describe more global structure of the word similarity matrix. We conducted hierarchical agglomerative clustering with Ward's linkage using *scipy* [49] to construct a dendrogram for the seed word similarity matrix.

2.4 Qualifying Language

Scientists, clinicians, and editors often use *qualifying language* – words such as *likely*, *probably*, *possibly*, etc. – to indicate degree of confidence in claims asserted in publications [27]. Such language can also be used to soften or hedge statements [32]. Given the hesitancy of medical practitioners to name racism, we may find that qualifying language is used to distance authors from strong claims about racism or to minimize its perceived importance. Thus, we used word similarities to quantify the degree to which qualifying language was used in the context of words denoting race or racism.

2.5 Network analysis

Network Construction. To visualize and interpret the 9242×9242 similarity matrix, we built a network representation by treating the similarity matrix as an adjacency matrix, following past work which uses word embeddings as conceptual space frameworks [34]. In this construction, each word acts as a node, and edges connect word pairs with weights defined as 1 minus their similarity, so that more similar words are positioned closer together. Because every word pair has a similarity score, this network would be extremely dense, making analysis and visualization difficult. To address this, we sparsified the network by removing edges.

We sought to sparsify the network in a manner that reduced the number of edges while preserving rare or informative words. For example, we could sparsify the network by setting a minimum word similarity for inclusion, which would be akin to keeping the top k percent of word similarities. This approach, however, would not be desirable because rare but informative words may have lower similarity scores and would risk being excluded. We instead use an approach similar to the *local degree* method of sparsification in social networks, which has been shown to preserve the global structure of networks [21]. For each word, we chose to only include edges to its 30 most similar words (we refer to this as the "top-30 list"). This approach ensures every word is included in the network, while retaining only the most meaningful edges. To validate that only connections with high similarity were included, we examined the distribution of least similar pairings for each node in the network. The IQR for this distribution was [0.958, 0.987], suggesting that few to no edges in the network connect dissimilar words. We treated the resulting network as undirected. Thus, in scenarios where one word was in the top-30 list of another, but not vice versa, we included an undirected edge connecting the two words. Sparsifying by keeping only the 30 most similar words for each node reduced the number of edges by 45.3% compared to only keeping the top 1% of word similarities in the network.

Visualization. To visualize the network, we extracted a sub-network from the 9242 nodes using 61 hand-selected terms motivated by the construct of racism developed by [17], encompassing discrimination-related terms at the individual, societal, and actionable level. We supplemented this list with words related to contributors to health disparities or health outcomes. We visualized the 61-node network with Pyvis, which uses force-directed layout algorithms to identify meaningful clusters based on edge lengths [42].

Identifying Gatekeepers. We sought to identify "gatekeepers", or words that mediated conceptual pathways in the network. Gatekeepers can be thought of as having the property of "betweenness" (as in betweenness centrality), a common property found in small-word networks characterized by high clustering and short average path lengths [34; 52]. Like brokers in social network

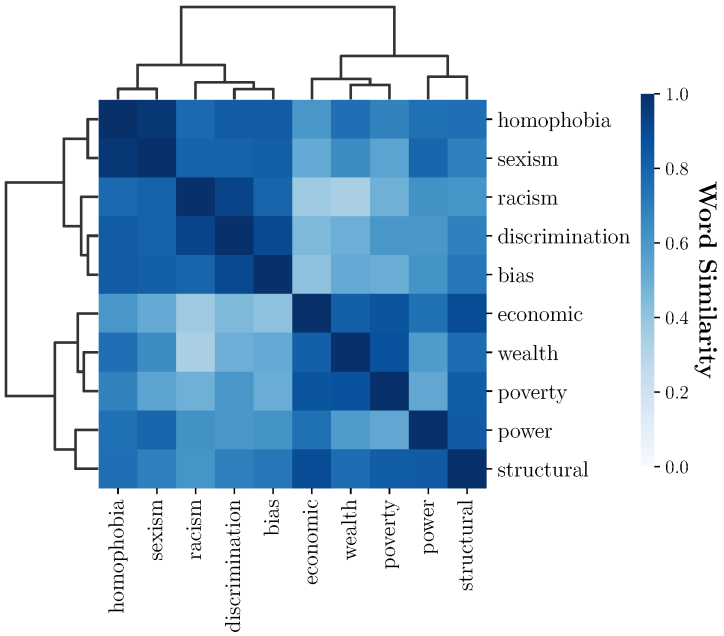

Figure 1: Word similarities reveal hierarchical relationships among thematic groups of tokens. The heat map depicts word similarities for a sub-matrix of 10 hand-picked words, chosen from the broader 9242×9242 matrix. Word similarity (colorbar: bluer/darker indicates higher similarity) is measured by the cosine similarity. Words are clustered hierarchically using a dendrogram, indicated by the tree structure.

analysis [31], they occupy strategic positions that potentially shape how medical practitioners understand connections between social determinants and health outcomes.

We focused on gatekeepers between the word *racism* and words related to health inequities. To identify the gatekeepers, we determined the most common words existing on the shortest paths between *racism* and the health inequity words. For example, in this scheme, the shortest path from a word related to health inequities – *stress* – and *racism* is traced through:

$$\textit{stress} \rightarrow \textit{psychological} \rightarrow \textit{stigma} \rightarrow \textit{interpersonal} \rightarrow \textit{discrimination} \rightarrow \textit{racism}$$

Here, *interpersonal* serves as a potential gatekeeper (which we validate by examining many paths). We emphasize that these paths reflect correspondences among word usage in medical texts, and do not reflect causality.

We calculated shortest paths between words w_i, w_j on the network, where edge lengths (conditional on an edge existing between two words) were defined as

$$|e_{ij}| = 1 - \text{cosine}(w_i, w_j)$$

where $\text{cosine}(w_i, w_j)$ is the cosine word similarity between words w_i and w_j. Thus, pairs of words with high similarity had shorter edges connecting them (and words not in each other's top-30 list had no edge connecting them, due to the sparsification procedure discussed above). Past work leveraging word embeddings networks as conceptual space frameworks generally utilize the shortest path as a meaningful indicator of conceptual relatedness [3; 34].

Labeling Health Inequity Words. To identify a subset of words to use for the shortest path analysis, we used OpenAI's GPT-4 [1] (the most powerful LLM available to use at the time of analysis) to extract words from the corpus that are directly related to health inequities (see Appendix A for the prompt). We used few-shot learning, providing the model with example input-output to guide selection of relevant words. To account for the size of GPT-4's context window,

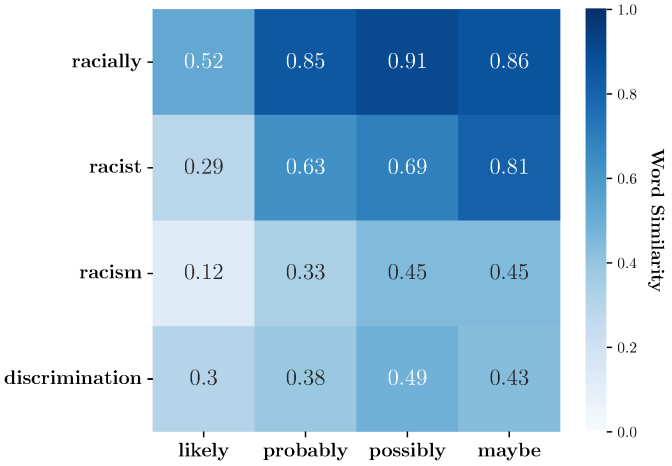

Figure 2: Qualifying language reflecting uncertainty more frequently aligns with racism-related words. The cell color denotes the strength of the word similarity (colorbar: bluer/darker indicates higher similarity) between racism-related words (rows) and qualifying language (columns). Qualifying language is sorted in decreasing order of perceived probability. Exact word similarities are provided in each cell.

we prompted 1000 words at a time in batches. From the full corpus of 9,242 words, we identified 110 words related to health inequities.

3 Results

We applied word embeddings to a corpus of 871 medical articles containing the word *racism* from four leading medical journals (BMJ, JAMA, NEJM, and The Lancet) published between 1999-2020 [30]. After implementing a comprehensive text preprocessing pipeline (Section 2.1), we obtained a vocabulary of 9,242 unique tokens (which we refer to as "words"). We computed word embeddings by averaging over multiple model fits, improving stability and robustness of the representations (Section 2.2) [6]. These embeddings enabled us to quantify semantic relationships between words and conduct downstream analyses examining how racism is conceptualized and discussed within medical literature through hierarchical clustering (Section 2.3), examination of qualifying language (Section 2.4), and a network analysis, using shortest path calculations between racism and words related to health inequities (Section 2.5).

3.1 Word similarities reveal hierarchical relationships among thematic groups of tokens

We calculated word similarities between all pairs of words in the corpus, resulting in a 9242×9242 word similarity matrix. Examining all pairwise combinations of similarities would be infeasible. Thus, we analyzed word similarities for a set of hand-picked words using a heat map (Fig. 1). The raw frequencies for these words are summarized in Appendix Table 1. The subset of words we chose embodied three groups, corresponding to discrimination, wealth, and power (Section 2.3). We additionally calculated a dendrogram using the similarity matrix.

The dendrogram (Fig. 1: black lines) consists of two large clusters: one containing discrimination-related words and the other containing the power-related and wealth-related words. The second cluster further breaks down into the power-related and wealth-related groupings, demonstrating that the word similarities and resulting dendrogram are able to reveal the semantic structure of these words. The heat map indicates a clear distinction between the two

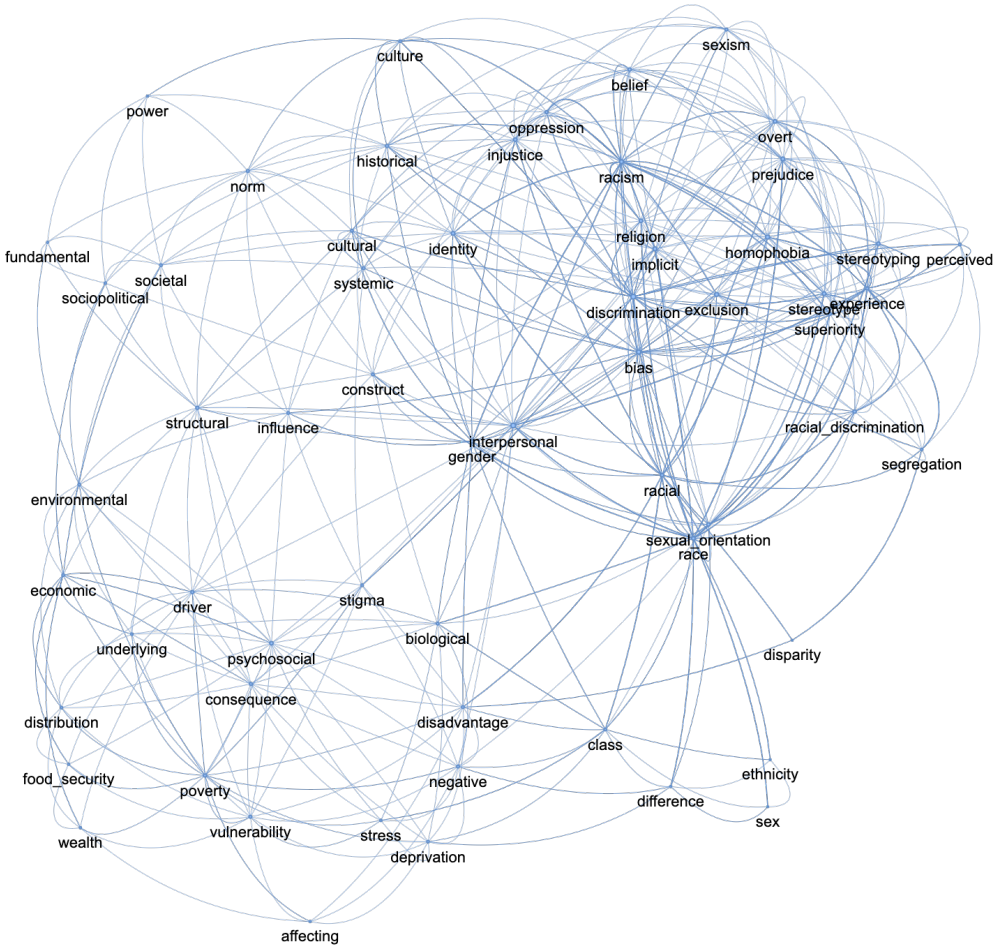

Figure 3: Word similarity network. The visualized network depicts the structure of a sub-network for 61 hand-selected words. A force-directed layout algorithm positions nodes based on edge lengths, causing similar words to cluster more tightly. Communities of words with greater collective similarity will cluster more tightly.

largest clusters. The low similarity between discrimination-related words and power- or wealth-related words suggests they are not often used in similar contexts. Notably, the word similarities between *racism* and the wealth-related words are among the lowest in the similarity matrix, with similarities nearing 0, while *racism* shows slightly higher similarities with power-related words.

The dendrogram further visualized hierarchical structure within these two clusters. For example, among the discrimination-related words, *homophobia* and *sexism* form their own cluster with high similarity, while *racism* couples more tightly with *discrimination* and *bias*. We may expect this result, given that the corpus was constructed with articles containing the word *racism*, but it raises questions about how *discrimination* and *bias* might relate to specific forms of discrimination in a broader medical corpus. Within the power- and wealth-related cluster, *power* and *structural* are grouped together but show distinct patterns in their relationships with other words, such as *poverty*, suggesting context-dependent interpretations. Together, these results demonstrate how the word similarities, via a dendrogram, can facilitate a hierarchical composition of thematic groups of words, shedding light on how tightly coupled notions of racism are with discrimination-, wealth-, and power-related words.

3.2 Qualifying language reflecting uncertainty more frequently aligns with racism-related words

Next, we used word embeddings to quantify the relationship between words associated with racism and qualifying language (Section 2.4). We chose four words related to race and discrimination (*racially*, *racist*, *racism*, *discrimination*) and calculated their corresponding similarities with four words used to qualify claims: *likely*, *probably*, *possibly*, and *maybe* (listed in decreasing order of perceived probability) [9; 50]. We plotted the word similarities among the 16 pairs in a heat map (Fig. 2).

Among the words analyzed, *racially* showed the highest similarity to qualifying language, while *racism* exhibited the lowest similarity across all qualifiers. As the perceived uncertainty of the qualifier increases (*likely* to *maybe*), the similarity with all four racism-related words also generally increases. For example, the similarity between *likely* and *racism* is 0.12, whereas the similarity between *maybe* and *racism* is 0.45. The other three words exhibited similar patterns, with word similarities increasing from 0.52 to 0.86 (*racially*), 0.29 to 0.81 (*racist*), and 0.3 to 0.43 (*discrimination*). These patterns suggest a reluctance to make confident claims about racism. We note these results do not clarify whether that uncertainty is unwarranted, i.e., whether the conclusions of a given paper are strong enough to warrant direct claims on racism, but the authors hedge nonetheless. However, the substantially higher word similarities between qualifying language and *racially* compared to *racism* and *racist* reflect how medical discourse generally employs more neutral, descriptive language as a hedge against making direct assertions about racist processes or discriminatory practices, warranting further examination.

3.3 Gatekeeper words mediate pathways between racism and health inequities

The heat map and dendrogram in Figure 1 facilitated examination of only a small vocabulary of words in the corpus. To better understand the relationship of a broader range of words, we turned to a network analysis (Section 2.5). First, we visualized a subset of the full vocabulary (Fig. 3). We provide raw frequencies for these words in Appendix Table 2. We found several distinct clusters, including discrimination-related words (top right) and health inequity outcomes (bottom left). The discrimination cluster exhibited tighter connectivity, with words such as *implicit*, *homophobia*, *exclusion*, and *superiority* showing high similarity to one another. In contrast, health inequity outcomes, including *poverty*, *psychological*, *vulnerability*, and *food_security*, formed a more dispersed cluster with lower internal connectivity. The upper left region contains words related to expectations and structures, such as *cultural*, *norm*, *power*, and *structural*, which demonstrate relatively low similarity to the discrimination cluster. Notably, several words occupied intermediate positions between the larger clusters, serving as connecting bridges. For instance, *stigma*, *interpersonal*, *biological*, and *construct* all lie along pathways linking discrimination-related words with health outcome words. The positioning of broader words such as *norm*, *power*, and *societal* suggests connections to structural and systemic concepts, though these remain distant (lower in similarity) from the primary clusters.

The visualization in Figure 3 shows that there are specific words that may connect separate clusters such as health outcomes and words related to racism. Following established work in distributional semantics, we assume that semantic proximity in word embeddings – reflecting co-occurrence patterns in text – indicates conceptual associations in medical discourse, such that shortest paths through the network reveal how medical practitioners connect these ideas in their writing [48]. For example, the shortest path between *racism* and *poverty*, a key determinant of health inequities, passes through the word *interpersonal*, suggesting a particular framing of racism (akin to person-level racism from Figueroa et al.'s construct [17]). If we view the network of word similarities as a conceptual space framework [34], then these connecting words – which we call

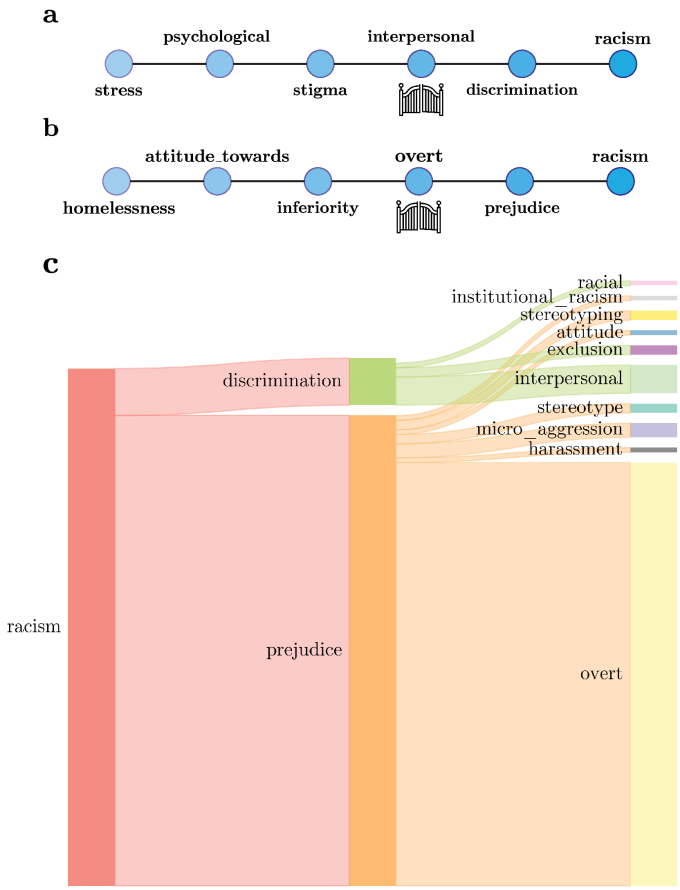

Figure 4: Gatekeeper words mediate pathways between racism and health inequities. a, b. Example shortest pathways between two candidate targets – *stress* and *homelessness* – and *racism*. The words *interpersonal* and *overt* are gatekeepers. **c.** Sankey plot showing the most common pathways taken up to 3 levels away from *racism*. The width of each flow denotes the fraction of paths passing through that word: the majority of paths pass through *overt*, followed by *interpersonal*.

gatekeepers – control how different domains are linked in the semantic mappings of medical text. We emphasize that a gatekeeper is not the only way to conceptually relate two nodes in a network, but highlights the most prominent semantic path between two concepts in the network, as suggested by the word embeddings. We hypothesized that a few gatekeepers dominate the pathways between racism and words related to health inequities and outcomes.

We identified gatekeepers between *racism* and 110 words related to health inequities by calculating shortest paths from *racism* to these words (Section 2.5). We then identified the gatekeepers as the words most consistently appearing on the pathways. For example, in Figure 4a, the shortest path to the word *stress* passes through *discrimination* and then *interpersonal*, both of which could be potential gatekeepers. Meanwhile, in Figure 4b, the shortest path to *homelessness* passes through *prejudice* and *overt*.

We identified clear patterns in how the targets connect to racism through the network. The mean pathway distance from the health inequity words to *racism* was 0.292 (ranging from 0.074 to 0.416). The five closest words to racism were *discrimination*, *sexism*, *homophobia*, *microaggression*, and *xenophobia*, while the five most distant were *tuberculosis*, *malaria*, *social_determinant*, *obesity*, and *poverty*. We visualized the flow of shortest pathways from racism to all health inequity

words using a Sankey diagram (Fig. 4c), with the width of each flow proportional to the number of pathways passing through each gatekeeper. No health inequity words lacked a pathway to racism. Among the pathways, we found that 81.8% of paths reached their targets through the subsequence *overt*→*prejudice*→*racism*, while 9.1% connected primarily through *interpersonal* (Fig. 4c), providing evidence for our hypothesis that a few gatekeepers (e.g., *overt* and *interpersonal*) dominate conceptual pathways between *racism* and health inequities.

4 Discussion

We leveraged word embeddings to interrogate how racism is conceptualized and discussed in medical literature published in prestigious journals. Our analysis identified distinct patterns in medical discourse around racism, including the separation of racism-related concepts from structural determinants, such as wealth and power, the use of qualifying language (words such as *likely*, *probably*, *possibly*) to hedge discussions of racism, and the dominance of person-level gatekeepers connecting health inequities to racism. These findings provide quantitative evidence that the medical literature discusses racism primarily through individual-level mechanisms rather than structural determinants, and tends to use qualifying language exhibiting uncertainty when discussing it [24].

Our analyses focused solely on articles that use the word *racism*. However, in accordance with our hypothesis that medical researchers and clinicians may refrain from naming racism in the medical literature, studying the medical literature on health inequities at large is of paramount importance. Our results serve as an upper bound for engagement with structural racism in the medical literature, because articles that avoid using *racism* are unlikely to engage with it conceptually on a deep level. Indeed, authors may intentionally add qualifying language to words like *racism* in order to discuss it without being required by editors to remove it entirely. Thus, there is a selection bias in the corpus we studied, and future work should conduct similar analyses on medical texts not using the word *racism*. Such analyses could leverage large-scale medical corpora including PubMed [40] and PubMed Central's Open Access Subset [38] to examine patterns of health inequity discourse in the broader medical literature where racism may not be explicitly named. These datasets provide rich opportunities to interrogate discussion (or lack thereof) of racism across time, medical fields, and publication types.

We found gatekeepers (Fig. 4) that predominantly reflect person-level conceptualizations of racism rather than structural ones. The presence of words like *overt*, *interpersonal*, and *prejudice* as gatekeepers aligns with Figueroa et al.'s person-level racism framework [17], suggesting that the portion of the medical literature we study conceptualizes racism primarily through individual attitudes and behaviors. Notably absent from common pathways are words indicative of structural racism such as *institutional*, *systemic*, or *structural*, reinforcing the pattern of individual-focused rather than systems-focused discourse around racism in medical texts.

We focused on medical literature from leading journals that, while internationally circulated, predominantly reflect medical discourse and research conducted within the United States and other Western contexts. The conceptualization and discussion of racism in medical literature may vary significantly across different cultural contexts. For instance, the framing of racism in medical discourse may differ in countries with distinct colonial histories, different racial and ethnic compositions, or alternative healthcare systems. Future research should study how racism narratives manifest in medical literature from diverse global contexts, including journals published in non-English languages and medical systems with different historical relationships to race and ethnicity.

While our analysis focused primarily on racism, our findings briefly touched on related concepts such as sexism and homophobia. Future research could extend these methods to examine how different forms of discrimination are conceptualized and discussed in medical literature, as each carries unique historical trajectories and social contexts within the United States. For instance, the medicalization of homosexuality and its subsequent depathologization [16] represents a funda-

mentally different historical relationship between medicine and LGBTQ+ communities compared to the legacy of medical racism rooted in slavery and segregation. Similarly, discussions of sexism in medicine may be shaped by the field's gender demographics and the particular ways that gender bias manifests in clinical care and research. Comparative analyses across these different forms of discrimination could illuminate how medical discourse varies in its willingness to name and address different types of bias.

Our analysis relied on word embeddings, which capture semantic relationships between individual words but are limited in their ability to represent higher-order linguistic structures and contextual interactions between concepts. We opted for word embeddings to set a foundation for future work relying on more advanced techniques. For example, sentence embeddings offer a promising avenue for deepening our understanding of how racism is discussed in medical literature by preserving the broader semantic context in which words appear [12; 45]. Future research could employ sentence embeddings to cluster sentences or paragraphs containing the word *racism* to identify distinct narrative frameworks in the spirit of Figueroa et al.'s typology of racism narratives. Additionally, large language models present a promising direction in mixed-methods approaches, and could serve to support qualitative analyses by facilitating thematic coding and iterative analysis of racism narratives [47].

As the medical community continues to reckon with its historical and contemporary role in perpetuating racial health inequities and establish anti-racist norms, studies of this nature will become increasingly important to ensure accountability [14]. Clinical journal language is powerful in that it establishes dominant narratives and therefore the norms in the medical community. Furthermore, since clinical journals are used in continuing medical education, their language impacts how medical students are trained and practice medicine [37]. Even more pressing, clinical journals – and other related texts written by medical practitioners – are increasingly used to train large language models for medical AI systems [2; 43; 51]. Their language will directly shape the biases and perspectives embedded in these AI tools, influencing future clinical decision-making and patient care. The size and scope of these systems will necessitate continuing work and innovation in computational humanities to critically examine and audit their usage and impact.

Acknowledgements

We acknowledge funding from the American Medical Association which supported this work in its early stages. The ideas in this article are those of the authors and do not necessarily represent policy of the American Medical Association.

References

[1] Achiam, Josh, Adler, Steven, Agarwal, Sandhini, Ahmad, Lama, Akkaya, Ilge, Aleman, Florencia Leoni, Almeida, Diogo, Altenschmidt, Janko, Altman, Sam, Anadkat, Shyamal, et al. "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774* (2023).

[2] Alsentzer, Emily, Murphy, John R, Boag, Willie, Weng, Wei-Hung, Jin, Di, Naumann, Tristan, and McDermott, Matthew. "Publicly available clinical BERT embeddings". In: *arXiv preprint arXiv:1904.03323* (2019).

[3] Amancio, Diego R, Machicao, Jeaneth, and Quispe, Laura VC. "Leveraging word embeddings to enhance co-occurrence networks: A statistical analysis". In: *PloS one 20*, no. 7 (2025), e0327421.

[4] Bailey, Zinzi D, Krieger, Nancy, Agénor, Madina, Graves, Jasmine, Linos, Natalia, and Bassett, Mary T. "Structural racism and health inequities in the USA: evidence and interventions". In: *The lancet 389*, no. 10077 (2017), pp. 1453–1463.

[5] Bird, Steven and Loper, Edward. “NLTK: The Natural Language Toolkit”. In: *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 214–217. URL: https://aclanthology.org/P04-3031/.

[6] Bloem, Jelke, Fokkens, Antske, and Herbelot, Aurélie. “Evaluating the consistency of word embeddings from small data”. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. 2019, pp. 132–141.

[7] Boyd, Rhea, Krieger, Nancy, Maio, Fernando De, and Maybank, Aletha. “The World’s Leading Medical Journals Don’t Write About Racism. That’s a Problem”. In: *TIME* (Apr. 2021). Accessed 2025-07-18. URL: https://time.com/5956643/medical-journals-health-racism/.

[8] Boyd, Rhea W, Lindo, Edwin G, Weeks, Lachelle D, and McLemore, Monica R. “On racism: a new standard for publishing on racial health inequities”. In: *Health Affairs Forefront* (2020).

[9] Budescu, David V and Wallsten, Thomas S. “Consistency in interpretation of probabilistic phrases”. In: *Organizational behavior and human decision processes 36*, no. 3 (1985), pp. 391–405.

[10] Byrd, W Michael and Clayton, Linda A. “Race, medicine, and health care in the United States: a historical survey”. In: *Journal of the National Medical Association 93*, no. 3 Suppl (2001), 11S.

[11] Castle, Billie, Wendel, Monica, Kerr, Jelani, Brooms, Derrick, and Rollins, Aaron. “Public health’s approach to systemic racism: a systematic literature review”. In: *Journal of Racial and Ethnic Health Disparities 6*, no. 1 (2019), pp. 27–36.

[12] Cer, Daniel, Yang, Yinfei, Kong, Sheng-yi, Hua, Nan, Limtiaco, Nicole, John, Rhomni St, Constant, Noah, Guajardo-Cespedes, Mario, Yuan, Steve, Tar, Chris, et al. “Universal sentence encoder”. In: *arXiv preprint arXiv:1803.11175* (2018).

[13] Cobert, Julien, Mills, Hunter, Lee, Albert, Gologorskaya, Oksana, Espejo, Edie, Jeon, Sun Young, Boscardin, W John, Heintz, Timothy A, Kennedy, Christopher J, Ashana, Deepshikha C, et al. “Measuring implicit bias in ICU notes using word-embedding neural network models”. In: *Chest 165*, no. 6 (2024), pp. 1481–1490.

[14] Crear-Perry, Joia, Maybank, Aletha, Keeys, Mia, Mitchell, Nia, and Godbolt, Dawn. “Moving towards anti-racist praxis in medicine”. In: *The Lancet 396*, no. 10249 (2020), pp. 451–453.

[15] Dean, Lorraine T and Thorpe Jr, Roland J. “What structural racism is (or is not) and how to measure it: clarity for public health and medical researchers”. In: *American Journal of Epidemiology 191*, no. 9 (2022), pp. 1521–1526.

[16] Drescher, Jack. “Out of DSM: Depathologizing homosexuality”. In: *Behavioral sciences 5*, no. 4 (2015), pp. 565–575.

[17] Figueroa, Caroline A, Manalo-Pedro, Erin, Pola, Swetha, Darwish, Sajia, Sachdeva, Pratik, Guerrero, Christian, Vacano, Claudia von, Jha, Maithili, De Maio, Fernando, and Kennedy, Chris J. “The stories about racism and health: the development of a framework for racism narratives in medical literature using a computational grounded theory approach”. In: *International journal for equity in health 22*, no. 1 (2023), p. 265.

[18] Ford, Chandra L and Airhihenbuwa, Collins O. “The public health critical race methodology: praxis for antiracism research”. In: *Social science & medicine 71*, no. 8 (2010), pp. 1390–1398.

[19] Garg, Nikhil, Schiebinger, Londa, Jurafsky, Dan, and Zou, James. “Word embeddings quantify 100 years of gender and ethnic stereotypes”. In: *Proceedings of the National Academy of Sciences 115*, no. 16 (2018), E3635–E3644.

[20] Gee, Gilbert C and Ford, Chandra L. “Structural racism and health inequities: Old issues, New Directions1”. In: *Du Bois review: social science research on race 8*, no. 1 (2011), pp. 115–132.

[21] Hamann, Michael, Lindner, Gerd, Meyerhenke, Henning, Staudt, Christian L, and Wagner, Dorothea. “Structure-preserving sparsification methods for social networks”. In: *Social Network Analysis and Mining 6*, no. 1 (2016), p. 22.

[22] Hamed, Sarah, Bradby, Hannah, Ahlberg, Beth Maina, and Thapar-Björkert, Suruchi. “Racism in healthcare: a scoping review”. In: *BMC public health 22*, no. 1 (2022), p. 988.

[23] Hardeman, Rachel R, Murphy, Katy A, Karbeah, J’Mag, and Kozhimannil, Katy Backes. “Naming institutionalized racism in the public health literature: a systematic literature review”. In: *Public Health Reports 133*, no. 3 (2018), pp. 240–249.

[24] Harper, Shaun R. “Race without racism: How higher education researchers minimize racist institutional norms”. In: *The Review of Higher Education 36*, no. 1 (2012), pp. 9–29.

[25] Harrigian, Keith, Zirikly, Ayah, Chee, Brant, Ahmad, Alya, Links, Anne, Saha, Somnath, Beach, Mary Catherine, and Dredze, Mark. “Characterization of stigmatizing language in medical records”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2023, pp. 312–329.

[26] Hoffman, Kelly M, Trawalter, Sophie, Axt, Jordan R, and Oliver, M Norman. “Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites”. In: *Proceedings of the National Academy of Sciences 113*, no. 16 (2016), pp. 4296–4301.

[27] Hyland, Ken. “The Author in the Text: Hedging Scientific Writing.” In: *Hong Kong papers in linguistics and language teaching 18* (1995), pp. 33–42.

[28] hyperreality. “American-British-English-Translator”. `https://github.com/hyperreality/American-British-English-Translator`. CLI for American English and British English translation. Accessed: 2025-07-18. 2025.

[29] Jones, Camara Phyllis. “Levels of racism: a theoretic framework and a gardener’s tale”. In: *American journal of public health 90*, no. 8 (2000), p. 1212.

[30] Krieger, Nancy, Boyd, Rhea W, De Maio, Fernando, and Maybank, Aletha. “Medicine’s privileged gatekeepers: producing harmful ignorance about racism and health”. In: *Health Affairs Forefront* (2021).

[31] Kwon, Seok-Woo, Rondi, Emanuela, Levin, Daniel Z, De Massis, Alfredo, and Brass, Daniel J. “Network brokerage: An integrative review and future research agenda”. In: *Journal of Management 46*, no. 6 (2020), pp. 1092–1120.

[32] Lewin, Beverly A. “Hedging: an exploratory study of authors’ and readers’ identification of ‘toning down’in scientific texts”. In: *Journal of English for Academic Purposes 4*, no. 2 (2005), pp. 163–178.

[33] Lewis, Molly and Lupyan, Gary. “Gender stereotypes are reflected in the distributional structure of 25 languages”. In: *Nature human behaviour 4*, no. 10 (2020), pp. 1021–1028.

[34] Liu, Zhu, Liu, Ying, Luo, KangYang, Kong, Cunliang, and Sun, Maosong. “Exploring the Small World of Word Embeddings: A Comparative Study on Conceptual Spaces from LLMs of Different Scales”. In: *arXiv e-prints* (2025), arXiv–2502.

[35] Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781* (2013).

[36] Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S, and Dean, Jeff. “Distributed representations of words and phrases and their compositionality”. In: *Advances in neural information processing systems 26* (2013).

[37] Milota, Megan M, Van Thiel, Ghislaine JMW, and Van Delden, Johannes JM. “Narrative medicine as a medical education tool: a systematic review”. In: *Medical teacher 41*, no. 7 (2019), pp. 802–810.

[38] National Library of Medicine. “PMC Open Access Subset”. Bethesda, MD, 2003. URL: `https://pmc.ncbi.nlm.nih.gov/tools/openftlist/`.

[39] Nelson, Alan. “Unequal treatment: confronting racial and ethnic disparities in health care”. In: *Journal of the national medical association 94*, no. 8 (2002), p. 666.

[40] Noroozizadeh, Shahriar, Kumar, Sayantan, Chen, George H, and Weiss, Jeremy C. “PMOA-TTS: Introducing the PubMed Open Access Textual Times Series Corpus”. In: *arXiv preprint arXiv:2505.20323* (2025).

[41] Omar, Mahmud, Sorin, Vera, Agbareia, Reem, Apakama, Donald U, Soroush, Ali, Sakhuja, Ankit, Freeman, Robert, Horowitz, Carol R, Richardson, Lynne D, Nadkarni, Girish N, et al. “Evaluating and addressing demographic disparities in medical large language models: a systematic review”. In: *International Journal for Equity in Health 24*, no. 1 (2025), p. 57.

[42] Perrone, Giancarlo, Unpingco, Jose, and Lu, Haw-minn. “Network visualizations with Pyvis and VisJS”. In: *arXiv preprint arXiv:2006.04951* (2020).

[43] Rasmy, Laila, Xiang, Yang, Xie, Ziqian, Tao, Cui, and Zhi, Degui. “Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction”. In: *NPJ digital medicine 4*, no. 1 (2021), p. 86.

[44] Řehůřek, Radim, Sojka, Petr, et al. “Gensim—statistical semantics in python”. In: *Retrieved from genism. org* (2011).

[45] Reimers, Nils and Gurevych, Iryna. “Sentence-bert: Sentence embeddings using siamese bert-networks”. In: *arXiv preprint arXiv:1908.10084* (2019).

[46] Sun, Michael, Oliwa, Tomasz, Peek, Monica E, and Tung, Elizabeth L. “Negative Patient Descriptors: Documenting Racial Bias In The Electronic Health Record: Study examines racial bias in the patient descriptors used in the electronic health record.” In: *Health Affairs 41*, no. 2 (2022), pp. 203–211.

[47] Tai, Robert H, Bentley, Lillian R, Xia, Xin, Sitt, Jason M, Fankhauser, Sarah C, Chicas-Mosier, Ana M, and Monteith, Barnas G. “An examination of the use of large language models to aid analysis of textual data”. In: *International Journal of Qualitative Methods 23* (2024), p. 16094069241231168.

[48] Turney, Peter D and Pantel, Patrick. “From frequency to meaning: Vector space models of semantics”. In: *Journal of artificial intelligence research 37* (2010), pp. 141–188.

[49] Virtanen, Pauli et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods 17* (2020), pp. 261–272. DOI: `10.1038/s41592-019-0686-2`.

[50] Wallsten, Thomas S, Budescu, David V, Rapoport, Amnon, Zwick, Rami, and Forsyth, Barbara. “Measuring the vague meanings of probability terms.” In: *Journal of Experimental Psychology: General 115*, no. 4 (1986), p. 348.

[51] Wang, Guangyu, Yang, Guoxing, Du, Zongxin, Fan, Longjun, and Li, Xiaohu. “ClinicalGPT: large language models finetuned with diverse medical data and comprehensive evaluation”. In: *arXiv preprint arXiv:2306.09968* (2023).

[52] Watts, Duncan J and Strogatz, Steven H. “Collective dynamics of ‘small-world’networks”. In: *nature 393*, no. 6684 (1998), pp. 440–442.

[53] Yang, Yifan, Liu, Xiaoyu, Jin, Qiao, Huang, Furong, and Lu, Zhiyong. “Unmasking and quantifying racial bias of large language models in medical report generation”. In: *Communications Medicine 4*, no. 1 (2024), p. 176.

[54] Zack, Travis, Lehman, Eric, Suzgun, Mirac, Rodriguez, Jorge A, Celi, Leo Anthony, Gichoya, Judy, Jurafsky, Dan, Szolovits, Peter, Bates, David W, Abdulnour, Raja-Elie E, et al. “Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study”. In: *The Lancet Digital Health 6*, no. 1 (2024), e12–e22.

[55] Zhang, Haoran, Lu, Amy X, Abdalla, Mohamed, McDermott, Matthew, and Ghassemi, Marzyeh. “Hurtful words: quantifying biases in clinical contextual word embeddings”. In: *proceedings of the ACM Conference on Health, Inference, and Learning*. 2020, pp. 110–120.

A System Prompt for Identifying Words Related to Health Inequities

Round-robin, Head-to-Head Deliberation System Prompt

```
You are a tool for labeling health related terms.

You will be given a list of health related terms with each term
separated by a comma.

Please return the words from the given list that is related or
contributing to health disparities.

Return "poverty", "homophobia", "stigma", "stress", and "racism"
because they are directly related and contributing to health
disparities.

Do not return "africa", "male", "health", "patient", "black",
"cocaine", "invasion", "horrible", and "cancer" because they are
not directly contributing to health disparities.

If none of the terms are health related terms, return "NONE".
```

B Frequencies of Seed Words

Word	Count	Word	Count
homophobia	33	economic	685
sexism	62	wealth	123
racism	1550	poverty	465
discrimination	734	power	453
bias	391	structural	323

Table 1: Frequencies of seed words for Figure 1.

Word	Count	Word	Count	Word	Count
stereotype	107	sexual_orientation	37	environmental	313
exclusion	133	systemic	123	economic	685
prejudice	179	gender	562	vulnerability	110
superiority	13	interpersonal	71	food_security	46
experience	1005	injustice	157	distribution	161
overt	62	historical	211	underlying	138
homophobia	33	culture	767	poverty	465
bias	391	cultural	662	societal	126
racial_discrimination	145	construct	51	norm	163
religion	89	identity	258	sociopolitical	20
oppression	78	stress	268	fundamental	141
implicit	77	stigma	130	structural	323
racism	1550	deprivation	116	influence	272
discrimination	734	negative	232	difference	882
segregation	95	disadvantage	186	sex	483
perceived	177	biological	188	class	321
sexism	62	driver	116	ethnicity	351
stereotyping	74	psychosocial	86	disparity	911
belief	293	wealth	123	power	453
race	1025	consequence	318		
racial	893	affecting	84		

Table 2: Frequencies of seed words used to generate network in Figure 3.