Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on AI and Social Good

Quantifying Error Disparities in Population Health Models

Aaron Marker[1] , **Salvatore Giorgi**[1] , **Adithya V Ganesan**[1,2] , **Vasudha Varadarajan**[3]
Ojas Deshpande[2] , **Laura Brandt**[4] , **Gabriel Odom**[5] , **H. Andrew Schwartz**[1]

[1]Vanderbilt University
[2]Stony Brook University
[3]Carnegie Mellon University
[4]City College of New York
[5]OdomScience, LLC
{aaron.marker, hansen.schwartz}@vanderbilt.edu

Abstract

Many high-stakes social applications of AI, such as public health surveillance and policy planning, operate at the community- rather than individual-level. However, existing AI fairness metrics are primarily designed for individual-level predictions and discrete demographic categories, limiting their applicability to community-level settings. In this work, we first introduce the *Bilateral Concentration Index* (BCI) to quantify nonmonotonic error disparities missed by the category-based metrics used at individual or data-levels. Then we conduct a large-scale audit of sociodemographic error disparities in both lexical- and transformer-based models of county-level health outcomes, over a dataset cover billions of community-mapped messages. All tasks exhibited significant disparities, with BCI ranging from 2.1% for life satisfaction to 17.0% for fair or poor health. We further evaluate four approaches for incorporating sociodemographic information, as potential bias mitigation strategies, finding that while demographic inclusion consistently improved predictive accuracy, it frequently amplified error disparities. The largest disparities were associated with education and income (BCI = 2.7–16.4%), often reducing accuracy for low-income, and in some cases, high-income communities. These findings highlight a critical accuracy–fairness trade-off in community-level models for public health tasks, demonstrating how seemingly beneficial modeling choices can lead to increased disparities which could disadvantage communities if model predictions are used to make policy decisions in geographic areas where this public health data is not available.

1 Introduction

Population health surveillance traditionally relies on surveys and administrative data, which are costly, temporally lagged, and often lack fine-grained spatial resolution. Recent work shows that community-aggregated social media language can predict health outcomes at the U.S. county level [Mangalik *et al.*, 2024]. However, as with public health disparity analyses that guide fair resource allocation [Beck *et al.*, 2014; Lemstra *et al.*, 2006; Shavers, 2007], before such systems can inform public health policy, we must understand their *error disparities*—systematic differences in model performance across sociodemographic groups [Shah *et al.*, 2020].

Error disparities have been extensively studied at the *document-* or *person-level* [Salinas *et al.*, 2023; Garimella *et al.*, 2022; Rawat *et al.*, 2024], but remain poorly understood for *community-level* prediction tasks such as regional health and well-being. To address this gap, we systematically evaluate language-based predictive models across four community-level health outcomes, examining error disparities along three sociodemographic dimensions with known selection biases on Twitter [Giorgi *et al.*, 2022a]: percentage of foreign-born, percentage of educated, and median household income.

We focus on sociodemographic ("human factor") inclusion techniques [Zamani *et al.*, 2018] – methods that incorporate demographic factors to substantially improve model accuracy [Giorgi *et al.*, 2023; Hovy, 2015] – though their impact on bias or *error disparities* is unexplored. While sociodemographic inclusion could theoretically increase bias, past studies indicate it can also reduce it [Shah *et al.*, 2020]. We hypothesize this could depend on the inclusion strategy. We therefore evaluate two classes of methods: *additive* approaches, which account for baseline outcome differences across demographic groups [Zamani and Schwartz, 2017], and *adaptive* approaches, which allow language features to vary with sociodemographic context [Lynn *et al.*, 2017; Zamani *et al.*, 2018].

Our **key contributions** are three-fold: **(1) Systematic bias assessment:** We analyze which community health tasks and sociodemographic factors are most prone to error disparities in NLP models utilizing both traditional ngram and modern transformer LM approaches; **(2) Evaluation of mitigation methods:** We evaluate four sociodemographic factor inclusion strategies, spanning *additive* and *adaptive* approaches, and examine the relationship between accuracy and error disparities; and **(3) Novel metric:** We propose the Bilateral Con-

centration Index (BCI), inspired by the *Gini coefficient*, to quantify error disparities for continuous-valued community outcomes, capturing non-linear and non-monotonic sociodemographic–error relationships.

Our findings reveal a trade-off between accuracy and fairness in community-level prediction, with implications for the responsible deployment of NLP systems in policy-relevant settings.

Interdisciplinary Domain Expertise. This work was carried out in close collaboration with (i) a clinical psychologist with expertise in end-point bias and (ii) a clinical and public health biostatistician. The psychologist contributed to problem formulation and framing, leveraging their expertise in ethics and fairness to contextualize potential community-level harms, guide the analysis, and lead development of the paper's ethics section. This input shaped how we interpret model error disparities in terms of societal impact and policy relevance. The biostatistician contributed to the statistical methodology, including guidance on regression and correlation analyses, validation procedures, and population-level inference, and refined the methodological language to ensure statistical rigor and clarity. Overall, these collaborations informed both the technical design and the ethical framing of the study.

2 Related Work

Prior work has explored incorporating sociodemographic information into language-based models for over a decade [Hovy, 2015; Soni *et al.*, 2024], primarily to improve predictive performance or language understanding [Lynn *et al.*, 2017; Huang and Paul, 2019]. More recent studies have examined conditioning models on human traits such as empathy, emotion, or personas in dialog and generative settings [Rashkin *et al.*, 2019; Roller *et al.*, 2021; Hu and Collier, 2024]. However, the impact of such sociodemographic inclusion on *error disparities* has not been systematically studied.

Separately, a substantial body of work has documented demographic disparities in NLP model errors. Early evidence includes the "Wall Street Journal effect," where tagging errors increased for text authored by individuals demographically distant from the training data [Hovy and Søgaard, 2015]. More recent studies show persistent performance gaps in health-related prediction tasks [Rai *et al.*, 2024] and in social media content moderation, where language from Black individuals is more likely to be mislabeled as hate speech [Sap *et al.*, 2019]. Although these studies focus on individual- or document-level tasks, they motivate the need to examine and mitigate disparities in community-level modeling.

In population health settings, evaluating bias remains challenging, particularly for regression-based community outcomes. Existing fairness metrics largely target binary classification or discrete groups [Pessach and Shmueli, 2022], and a comparatively small fraction of prior work addresses regression settings [Mehrabi *et al.*, 2021]. Widely used measures such as statistical parity and bounded group loss [Agarwal *et al.*, 2019] do not naturally extend to continuous outcomes or continuous sociodemographic attributes, while alternatives such as Individual Fairness [Berk *et al.*, 2017] are difficult to operationalize for model evaluation. These limitations motivate the need for metrics tailored to continuous, community-level disparity analysis.

3 Data Set

We draw on county-level datasets during two distinct time periods that pair large-scale language use with sociodemographics and health outcomes. Below, we describe the language features, control variables, and outcome measures used in our analyses.

Community Language We use the **County Tweet Lexical Bank** (CTLB), which is an open-source dataset of U.S. county-level language features. We describe the details relevant to our experiments here, while other specifics can be found in [Giorgi *et al.*, 2018]. This dataset is derived from a 10% sample of Twitter from 2009–2015. From this sample, Twitter users were mapped to U.S. counties – administrative regions within each state – via self-reported location (via a free-text field in their profile) and latitude/longitude coordinates associated with their tweets. To be included in the dataset, county-mapped Twitter users needed at least 30 tweets across the 10% sample, and counties were included if at least 100 such users were mapped to the county (we acknowledge as a limitation of this method that the 30 tweet a year minimum may under represent individuals who do not use online social media). A total of 6 million users across 2,041 counties met this threshold, for a final dataset of 1.5 billion tweets. From these tweets, lexical features (25,000 of the most frequent unigrams) were extracted and embeddings (RoBERTa-Large 1024 dimensional embeddings) were taken for each of the 6 million users. These user-level features were then averaged within each county to produce a set of U.S. county features (i.e., each county is represented by a vector of 25,000 unigram frequencies or a embedding vector of 1024 dimensions). This dataset has been validated across several studies and shown to predict community health [Matero *et al.*, 2023; Abebe *et al.*, 2020], well-being [Jaidka *et al.*, 2020], and psychology [Giorgi *et al.*, 2022b].

Community Controls We focus on **three sociodemographic factors** that have had high predictive value in past work [Giorgi *et al.*, 2022a]: percentage of foreign-born residents, percentage of the county's population with a high school diploma, and the log of the county's median income. Five-year estimates (2011–2015) for percent of foreign-born population (percentage of a county's population that was born in another country), percent of high school graduate population (percentage of the population with a high school diploma), and median income of population (median log annual household income of a population) were obtained from the United States Census Bureau's 2015 American Community Survey (ACS).

Community Outcomes We consider prediction of **four county health outcomes**: heart disease mortality, life satisfaction, fair or poor health, and suicide mortality. We gathered age-adjusted mortality rates per 100,000 people for heart disease (**HD**; $N = 1750$, mean = 189.67, SD = 45.58) and suicide (**SM**; $N = 1631$, mean = 15.55, SD = 4.67) from the

Centers for Disease Control and Prevention (CDC), averaged over the years 2010–2015. Life satisfaction (**LS**; $N = 1745$, mean = 3.39, SD = 0.03) was assessed using individual responses to the question: "In general, how satisfied are you with your life?" on a scale from 1 (very dissatisfied) to 5 (very satisfied), with scores averaged across 2009 and 2010 [Lawless and Lucas, 2011].

Lastly, data on poor or fair health (**FP**; $N = 1703$, mean = 17.15, SD = 5.71) came from the County Health Rankings, drawing on the Behavioral Risk Factor Surveillance System [Remington *et al.*, 2015]. This age-adjusted metric reflects the percentage of adults who rated their health as "fair" or "poor" in response to the question: "In general, would you say that your health is Excellent/Very good/Good/Fair/Poor?"

Replication To examine how generalizable the findings are to different time periods and approaches, we also rerun the experiment using a more recent set of 425 million tweets from 2019–2021 [Mangalik *et al.*, 2024] across the same 2628 counties we call the County Tweet Lexical Bank 2 (CTLB2). Each of the sociodemographic factors and outcomes was sourced from the same databases as the first dataset but taken from 2020 to fit the recency of the dataset. Not all of the same data was collected for the more recent period, so this study does not analyze foreign born percentage or life satisfaction.

4 Metrics for Error Disparity

We study community-level disparity where sociodemographic attributes are continuous rather than categorical, requiring metrics that capture distributional error differences without binning.

Measuring Disparity Prior work on predictive bias in NLP often compares errors across discrete demographic groups (e.g., [Hovy and Søgaard, 2015; Zhao *et al.*, 2017]); however, community-level sociodemographic attributes are typically continuous (e.g., percentages or averages). In social science, the Gini coefficient and Lorenz curve are commonly used to measure inequality, but they require ordering by the outcome variable itself and therefore cannot assess disparities conditioned on an external attribute. Concentration Curves [O'Donnell *et al.*, 2007] generalize the Lorenz curve by allowing observations to be ranked by an auxiliary variable (e.g., median income), enabling the measurement of error disparities conditioned on continuous sociodemographic factors (Figure 1).

The concentration index does not however account for the non monotonic bias present in a concentration curve like the curve in Figure 1. To address this limitation, we introduce the **Bilateral Concentration Index (BCI)**, a modification of the Concentration Index that captures disparity magnitude irrespective of directionality, enabling detection of nonlinear relationships between prediction error and sociodemographic variables. We clarify strengths of this metric in section 6.1. We define the BCI below.

$$f_i(x) = |(m_{e_i}x - e_{i+1}) - (m_{u_i}x - u_{i+1})| \quad (1)$$

$$m_{e_i} = \frac{e_{i+1} - e_i}{\frac{1}{N}}, m_{u_i} = \frac{u_{i+1} - u_i}{\frac{1}{N}} \quad (2)$$

where N is the number of ranked observations (counties ordered from lowest to highest error), and $f_i(x)$ represents the absolute deviation between cumulative error (e_i) and uniform expectation (u_i) within each interval (variable $f_i(x)$ and m separated out for readability). The absolute value ensures that disparities in opposing directions contribute additively.

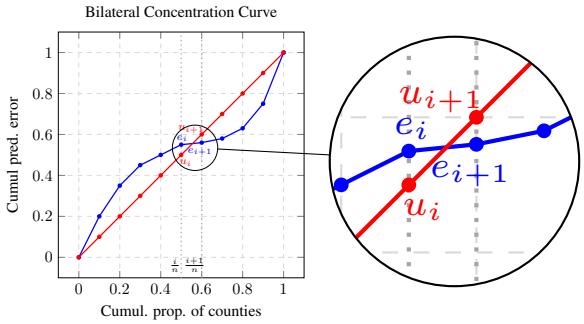

Figure 1: Zoomed in Bilateral Concentration Curve: The blue line is cumulative error curve and red line is the cumulative uniform line of equality. This example illustrates the scenario where there is low error in the middle such that the cumulative error line crosses the line of equality on an interval

All disparity metrics are reported in Table 2 for Experiment 1 (language-only). Across experiments, metrics generally tracked similar disparity patterns, but differed in interpretability. The Anderson–Darling (AD) test yields unbounded values that are difficult to compare across settings, while the Concentration Index (CI) may understate disparity when opposing deviations cancel. In contrast, BCI remains bounded, direction-agnostic, and comparable across conditions.

Due to its bilateral and continuous formulation, BCI captures nonlinear relationships between prediction error and sociodemographic rank that are missed by KS and CI. For KS, this distinction appears empirically in Figure 4: disparity among counties with respect to income is low (BCI = 7.5%, KS = 0.072). Adaptation increases disparity primarily at the high-income end, reflected by a larger increase in BCI (10.6%) than KS (0.083). For CI, this is illustrated in Figure 3, where the concentration curve for predicting county-level suicide rates as a function of percent foreign-born crosses the line of equality multiple times, corresponding to a U-shaped relationship in the loess plot where the minimum error is in the center as opposed to either extreme. Figure 1 further illustrates how BCI accounts for opposing regions that cancel under CI.

BCI treats observations continuously without binning, enabling fine-grained measurement of disparity across the distribution. However, because CI encodes disparity directionality through its sign, it provides complementary information. Accordingly, BCI is best interpreted in conjunction with CI or the concentration curve to provide a complete view of community-level disparity.

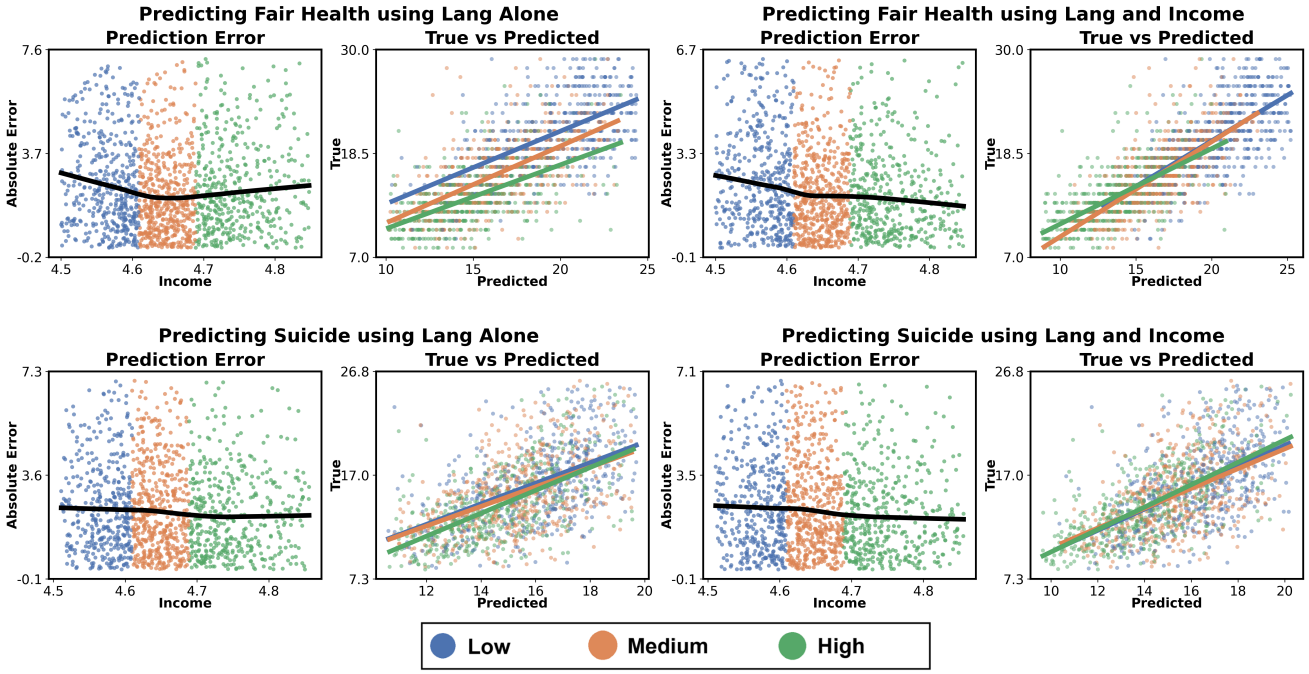

Figure 2: Scatter plots of outcomes with respect to income of a county. Prediction errors as a function of logged income in 1st and 3rd columns. Income is colored by tercile; loess curve in black. Predicted vs true outcome values in 2nd and 4th columns. Linear regression lines plotted for each income tercile use the same color mapping. loess plot is included to show nonmonotonic error disparities.

5 Methods

We evaluate how sociodemographic information can be incorporated into language-based prediction using established factor inclusion techniques. Below, we describe the inclusion strategies and modeling pipeline used consistently across all experiments.

Factor Inclusion Methods. We examine four techniques for integrating sociodemographic factors into language-based predictive models, spanning two strategies: *additive* methods, which account for baseline outcome differences across sociodemographic variables, and *adaptive* methods, which allow linguistic meaning to vary by demographic context [Lynn *et al.*, 2017]. For example, the word "mean" might have one sense as "cruel", but among more educated populations it could more often signify the mathematical average sense of the word [Lynn *et al.*, 2017].

As additive approaches, we use: (1) **Factor Concatenation (Factor Concat)**, where sociodemographic factors are concatenated with language features into a single feature vector; and (2) **Residualized Controls (Res Control)**, where sociodemographic factors are first modeled independently and the language model predicts the residual from this control-only model [Zamani *et al.*, 2018]: Residualized Controls can be represented mathematically as follows

$$\varepsilon = Y - \hat{Y}_C \tag{3}$$

$$\hat{\varepsilon}_L = \gamma \times X_L + \lambda \tag{4}$$

where ε is the residual, γ and λ are the coefficients and bias respectively for the linear prediction model, and $\hat{Y}_C$ represents the predictions of the controls model for the outcome variable, Y. The residual is minimized by a subsequent model that uses the language features, X_L. This approach ensures that sociodemographic effects are preserved rather than overwhelmed by high-dimensional language features [Zamani and Schwartz, 2017].

As adaptive techniques, we utilize: (3) **Factor Adaptation (Fact Adapt)** – this approach augments linguistic features by transforming them with control variables (C). This allows the model to use language features both independently, and in the context of controls [Lynn *et al.*, 2017]. In Factor Adaptation, the adapted language features (X_{A_i}) are combined as follows:

$$X_{A_i} = [X_L \cdot C_i], \forall i \in [1, |C|] \tag{5}$$

$$X_F = [X_L, X_{A_1}, ..., X_{A_{|C|}}] \tag{6}$$

where controls are mean-centered prior to composition [Lynn *et al.*, 2017]. (4) **Residualized Factor Adaptation (Res Fact Adapt)** – combining **Fact Adapt** and **Res Control**, a **Fact Adapt** model is fit to the residual of a control-only model offering the advantages of both [Zamani *et al.*, 2018]. Residualized Factor Adaptation can be represented as

$$\hat{\varepsilon}_L = \gamma \times \left[X_L, X_{A_1}, X_{A_2}, ..., X_{A_{|C|}}\right] + \lambda \tag{7}$$

Baselines. We compare all factor inclusion methods against a language-only baseline (Language Alone) to isolate the effects of sociodemographic integration. Used alone, sociodemographic factors were generally weaker predictors than language, with foreign-born population performing substantially worse and education and income marginally worse.

Model Accuracy Across Sociodemographic Factors, Outcomes, Approaches, and Feature Types											
Sociodemog.	Outcome	Baseline		Additive Approaches				Adaptive Approaches			
		Language Alone		Factor Concat		Res Control		Fact Adapt		Res Fact Adapt	
Factor		ngram	embed	ngram	embed	ngram	embed	ngram	embed	ngram	embed
% Foreign Born Population	Heart Disease	.749	.770	.750	.772	.747	.762	.764	.775	.763	.776
	Life Satisfact.	.450	.496	.451	.502	.447	.492	.502*	.525	.491*	.525
	Poor Health	.764	.766	.764	.767	.754	.760	.773	.772	.770	.772
	Suicide Mort.	.635	.667	.633	.683	.671*	.677	.673*	.684	.670*	.682
% High School Grad. Pop.	Heart Disease	.749	.770	.750	.781	.730	.766	.771	.780	.765	.777
	Life Satisfact.	.450	.496	.456	.531	.505*	.527	.541**	.576**	.518**	.562**
	Poor Health	.764	.766	.769	.799**	.781	.788	.808**	.803*	.803**	.801**
	Suicide Mort.	.635	.667	.636	.669	.622	.665	.664	.675	.661	.677
Median Household Income	Heart Disease	.749	.770	.752	.788	.747	.775	.780**	.786	.779*	.784
	Life Satisfact	.450	.496	.478	.559**	.530**	.554**	.566**	.566**	.551**	.562**
	Poor Health	.764	.766	.770	.802**	.798*	.788	.813**	.808**	.811**	.807**
	Suicide Mort.	.635	.667	.637	.672	.636	.668	.655	.671	.647	.668
Global Avg		.649	.675	.654	.694*	.664*	.685	.692*	.702*	.686*	.699*

Table 1: for both ngram and embedding approaches across county outcomes and different sociodemographic factor inclusion approaches: Accuracies measured using *Pearson r* Asterisks represent statistically significant difference from disparity with the same parameters using language alone (L) (*: $p < .05$, **: $p < .01$). Significance for global average calculated using harmonic mean of p values for all tests conducted for that factor inclusion method, which controls the family wise error rate[Wilson, 2019].

Implementation Details. All experiments use a consistent feature selection and modeling pipeline across outcomes and factor inclusion methods. To emphasize inclusion strategies and disparity metrics, we employ a standard linear modeling pipeline commonly used in county-level language analysis [Eichstaedt *et al.*, 2015; Giorgi *et al.*, 2018], implemented using DLATK [Schwartz *et al.*, 2017].

All experiments use the same modeling pipeline to isolate the effects of factor inclusion methods. Following prior county-level language prediction work [Eichstaedt *et al.*, 2015; Giorgi *et al.*, 2018], we use DLATK [Schwartz *et al.*, 2017] to construct language-based predictors. Starting from 25,000 unigram features, we apply FWER-controlled univariate feature selection (threshold = 0.60) followed by PCA dimensionality reduction. For each factor inclusion method, the resulting language features are combined with the relevant sociodemographic factor and used to train one ℓ_2-regularized linear regression (ridge) model to predict each outcome (HD, LS, FP, SM). Models are evaluated using 10-fold cross validation, with the regularization parameter α selected via nested cross validation. Experiments are conducted across all combinations of outcomes (4), sociodemographic factors (3), and inclusion methods (4).

The functions used are publicly released for reproducibility[1]. Additionally, we provide a web application[2] for computing BCI, accuracy metrics, significance tests, and related disparity measures from user-provided data.

6 Results

We first examine the characteristics of the BCI metric against other measures of bias, then, using the BCI, we systematically evaluate error disparities of language-based predictive models across four outcomes and three sociodemographic controls.

[1]https://github.com/AaronMarker/AiFairnessPipeline

[2]https://humanlanguage.org/BCI

6.1 BCI in Comparison with Alternative Metrics

An effective disparity metric should be globally sensitive to shifts anywhere in the distribution, interpretable, and capable of capturing nonlinear relationships between error and sociodemographic rank. We empirically compare common disparity metrics listed below, in Table 2.

The **Kolmogorov–Smirnov (KS)** statistic measures the maximum deviation (as a score between 0 and 1) between a concentration curve and the line of equality, making it intuitive but insensitive to global distributional differences, especially near the distribution's edges (see Figure 4).

The **Anderson–Darling (AD)** test addresses this limitation by weighting distribution tails more heavily, but produces unbounded values that are difficult to interpret and compare across settings.

Disparities Using Language Alone to predict Median Household Income				
Outcome	Anders-Darling	KS	CI	BCI
Heart Disease	13823	.047	-1.5%	4.8%
Poor Health	62676	.088	1.5%	9.2%
Suicide Mort.	8026	.035	0.4%	3.5%

Table 2: Disparity according to four different metrics using embeddings alone to predict median household income for Twitter data from 2019-2021. For all metrics, further from zero indicates more disparity.

The **Concentration Index (CI)** provides a bounded and interpretable measure by integrating the signed area between the concentration curve and the line of equality.

However, the CI assumes monotonic relationships between error and sociodemographic rank. As shown in Figure 1, non-monotonic disparities—where error is high in the middle of the distribution but low at both extremes—can result in cancellation of opposing areas, yielding a CI of zero despite substantial disparity.

On the other hand, the **BCI metric** is defined as the total ab-

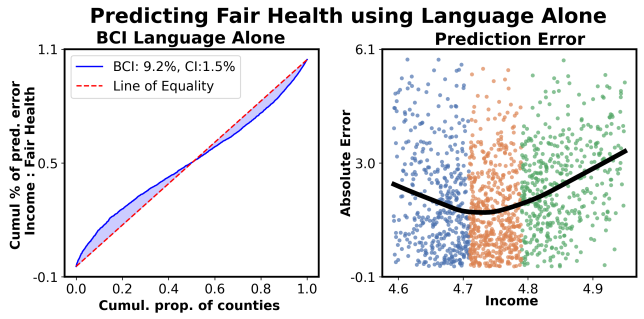

Figure 3: CDF of prediction error for fair/poor health from language, ordered by income (left), and error versus log income (right). When error varies nonlinearly with a demographic variable, the standard concentration index (CI) underestimates disparity (1.5%), whereas the bilateral concentration index (BCI) better captures it (9.2%).

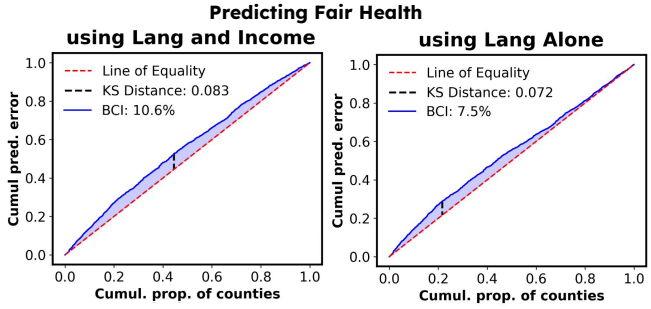

Figure 4: Comparison of BCI and KS metrics: The graphs show that similar KS values may reflect either concentrated or diffuse disparities, whereas BCI better captures overall disparity magnitude.

solute area between the concentration curve and the line of equality, aggregating disparities in either direction from the line of equality (See Figure 3):

$$BCI = 2 \sum_{i=0}^{N-1} \left(\int_{\frac{i}{N}}^{\frac{i+1}{N}} f_i(x)dx \right) \tag{8}$$

6.2 Evaluation of Factor Inclusion Methods

BCI Reference Points. To contextualize disparity magnitudes, we compare observed BCIs to two references: a random permutation of errors (noise-driven lower bound) and a sorted error ordering (high-inequality upper benchmark). Across models, lower-bound BCIs ranged from 0.01–0.03 and upper-bound BCIs from 0.44–0.46.

Outcome	Foreign Born		High School Grad		Median Income	
	ngram	embed	ngram	embed	ngram	embed
Heart Disease	4.1%*	6.5%**	9.0%**	10.2%**	9.2%**	11.1%**
Life Satisfact.	9.9%**	7.6%**	5.2%**	3.9%*	5.8%**	3.9%*
Poor Health	4.8%**	4.8%**	10.8%**	10.4%**	7.5%**	8.7%**
Suicide Mort.	4.4%*	7.9%**	2.7%	4.4%*	5.6%*	8.1%**

Table 3: Error disparity (BCI) for the given sociodemographic factor (Demog.) and across language-based predictive models (language alone) for the four community health outcomes on the CTLB dataset. Ngram and embedding-based model results are shown side by side. Asterisk represents statistically significant difference from a random baseline (* $p < .05$, ** $p < .01$).

Disparity in Language Predictions. Table 3 shows significant error disparities for language-only models across all outcome–demographic pairs except suicide mortality with HS graduation. For example, life satisfaction predictions exhibit substantial disparity across foreign-born population share (BCI = 9.9% ngram, 7.6% embedding). Disparity magnitudes vary across outcomes for the same sociodemographic variable, indicating that prediction error is not solely determined by the demographic factor itself.

N-gram vs. Embedding Models. Embedding models generally exhibit higher error disparity than n-gram models, except for life satisfaction, while consistently achieving higher predictive accuracy (Table 1). Thus, disparity differences should be interpreted jointly with accuracy gains.

Factor Inclusion Methods. We next evaluate whether incorporating sociodemographic factors mitigates disparity. We find that all factor inclusion methods improve accuracy relative to language-only models, but often increase error disparity (Tables 1 and 4). For example, factor adaptation achieves the highest accuracy (Pearson $r = 0.702$ embedding) but increases average disparity relative to language-only baselines (BCIs 8.3% and 7.3% for embedding respectively).

Figure 2 shows that lower overall prediction error does not necessarily imply lower disparity. While factor inclusion often reduces error, the fitted error curves show a larger slope over varying income, corresponding to higher BCI values.

Replication with Recent Data. Using county-level RoBERTa embeddings from 2019–2021 tweets, we observe lower overall accuracy and disparity, but replicate the central finding: factor inclusion improves accuracy, but often increases disparity (Table 5).

7 Conclusion

We conducted a systematic evaluation of error disparities in community-level health prediction tasks, revealing substantial performance differences across sociodemographic variables and outcomes. To quantify these effects, we introduced the Bilateral Concentration Index (BCI), which captures asymmetric error distributions and highlights disparities not evident from aggregate accuracy alone.

Across both n-gram and transformer-based language models, we observed consistent disparities, with higher errors for counties with lower income or educational attainment on multiple health outcomes, and reduced accuracy for life satisfaction in counties with higher foreign-born populations. We further evaluated factor inclusion methods as potential mitigators and found a robust trade-off between accuracy and disparity. While incorporating sociodemographic factors generally improved predictive accuracy, methods that were most effective in this regard—particularly adaptive approaches—often increased error disparities. These patterns replicated across modeling approaches and time periods.

Overall, our findings suggest that sociodemographic disparities in model performance are pervasive at the county level, and that incorporating demographic factors does not yield universally beneficial effects. Instead, the impact of such factors depends on the outcome, demographic vari-

(CTLB) Model Disparity Across Sociodemographic Factors, Outcomes, Approaches, and Feature Types											
Sociodemog. Factor	Outcome	Baseline Language Alone		Additive Approaches Factor Concat		Additive Approaches Res Control		Adaptive Approaches Fact Adapt		Adaptive Approaches Res Fact Adapt	
		ngram	embed	ngram	embed	ngram	embed	ngram	embed	ngram	embed
% Foreign Born Population	Heart Disease	4.1%	6.5%	4.2%	6.6%	**3.6%**	5.2%	5.8%	7.2%	5.5%	7.0%
	Life Satisfact.	9.9%	7.6%	9.9%	7.3%	9.6%	7.8%	9.4%	7.9%	9.1%	7.9%
	Poor Health	4.8%	4.8%	4.8%	4.9%	4.4%	4.4%	5.9%	5.6%	5.8%	5.7%
	Suicide Mort.	4.4%	7.9%	4.7%	8.9%	7.0%	8.1%	8.2%**	9.3%	7.3%*	8.9%
% High School Grad. Pop	Heart Disease	9.0%	10.2%	9.1%	11.4%	14.9%**	13.3%**	12.2%*	13.0%*	12.5%*	12.2%
	Life Satisfact.	5.2%	3.9%	5.0%	2.6%	3.3%	2.3%	3.5%	**2.1%**	3.6%	2.2%
	Poor Health	10.8%	10.4%	11.0%	12.8%	16.4%*	17.0%**	15.0%*	15.0%*	15.1%*	15.3%**
	Suicide Mort.	2.7%	4.4%	2.6%	4.2%	3.4%	4.6%	3.1%	4.6%	3.2%	4.5%
Median Household Income	Heart Disease	9.2%	11.1%	9.4%	12.3%	9.9%	11.1%	12.5%*	12.9%	12.7%*	12.8%
	Life Satisfact	5.8%	3.9%	4.4%	**3.0%**	4.3%	3.3%	3.7%	3.3%	4.0%	3.3%
	Poor Health	7.5%	8.7%	7.9%	10.4%	7.9%	8.7%	10.6%*	10.3%	10.4%	10.2%
	Suicide Mort.	5.6%	8.1%	5.8%	8.5%	7.6%	8.6%	8.5%	8.8%	8.1%	9.0%
Global Avg		**6.6%**	7.3%	**6.6%**	7.7%	7.7%**	7.9%*	8.2%*	8.3%	8.1%*	8.3%*

Table 4: across county outcomes and different sociodemographic factor inclusion approaches: Disparity is measured for both ngram and embedding approaches using the *Bilateral Concentration Index* (BCI) (as a percent), each comparing the cumulative error over counties, sorted by sociodemographic factor, to a cumulative uniform distribution. **Bold** represents tests with the lowest disparity per sociodemographic factor. Underline represents tests with the highest disparity per sociodemographic factor. Asterisks represent statistically significant difference from disparity with the same parameters using language alone (L) (*: $p < .05$, **: $p < .01$). Significance for global average calculated using harmonic mean of p values for all tests conducted for that factor inclusion method, which controls the family wise error rate [Wilson, 2019].

CTLB2 Embedding Model Disparity and Accuracy (Bilateral Concentration Index and Pearson r)													
Sociodemog. Factor	Outcome	Language Alone		Factor Concat		Res Control		Fact Adapt		Res Fact Adapt		Control Alone	
		BCI ↓	r ↑	BCI ↓	r ↑	BCI ↓	r ↑	BCI ↓	r ↑	BCI ↓	r ↑	BCI ↓	r ↑
% High School Graduate Population	Heart Disease	**4.0%**	.285	7.7%**	.534	12.6%**	.537	12.2%**	.538	12.6%**	.531	12.2%	.518
	Poor Health	3.8%	.652	**1.3%**	.850	6.2%	.852	5.7%	.870	5.6%	.863	5.2%	.792
	Suicide Mort.	4.4%	.240	**1.6%**	.365	2.3%	.389	3.2%	.388	2.8%	.382	1.4%	.313
% Median Household Income	Heart Disease	**4.8%**	.293	9.6%**	.531	12.7%**	.542	12.9%**	.546	12.7%**	.545	13.0%	.541
	Poor Health	9.2%	.649	8.6%	.721	**2.4%**	.714	6.0%	.733	4.9%	.732	1.6%	.463
	Suicide Mort.	**3.6%**	.251	5.0%	.338	8.0%**	.368	8.2%**	.374	7.6%*	.387	7.4%	.258
Global Avg:	Lexical (CTLB1)	**6.6%**	.649	**6.6%**	.654	7.7%**	.664	8.2%*	.692	8.1%*	.686	6.8%	.352
	Embeddings (CTLB 1)	**7.3%**	.675	7.7%	.694	7.9%*	.685	8.3%	.702	8.3%*	.699	6.8%	.352
	Embeddings (CTLB 2)	**5.0%**	.395	5.6%*	.557	7.4%**	.567	8.0%*	.575	7.7%*	.574	6.8%	.481

Table 5: Disparities and Accuracies across county outcomes and different sociodemographic factor inclusion approaches: Disparity is measured using the *Bilateral Concentration Index* (BCI) (as a percent), each comparing the cumulative error over counties, sorted by sociodemographic factor, to a cumulative uniform distribution. Accuracies measured using *Pearson r*. **Bold** represents tests with the lowest disparity per sociodemographic factor. Underline represents tests with the highest disparity per sociodemographic factor. Asterisks represent statistically significant difference from disparity with the same parameters using language alone (L) (*: $p < .05$, **: $p < .01$). Significance for global average calculated using harmonic mean of p values for all tests conducted for that factor inclusion method, which controls the family wise error rate[Wilson, 2019].

able, and modeling approach, underscoring the need to evaluate fairness and accuracy jointly rather than in isolation. These findings are essential to consider when modeling public health via social media to make real world predictions for geographic areas with less available data.

Ethical Statement

This study was reviewed and approved by an academic institutional review board (found to be exempt, non-human subjects data). No individual-level data has been reported or released in the current study. We report only aggregate data at the community level, so there are no privacy concerns to individuals. It is important to consider and discourage the potential negative applications of this work. While our approach can uncover error disparities in systems such as targeted recommendations, we recognize it could also be misapplied to amplify algorithmic biases and exacerbate inequities. The results described could reinforce existing biases contributing to additional stigma towards a group. Our work is intended for researchers and practitioners of Social Science, and we don't condone the usage of such algorithms for malicious purposes.

Acknowledgements

This research was, in part, funded by the National Institutes of Health (NIH) Agreement NO. 1OT2OD032581-01. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH.

References

[Abebe *et al.*, 2020] Rediet Abebe, Salvatore Giorgi, Anna Tedijanto, Anneke Buffone, and H Andrew Andrew Schwartz. Quantifying community characteristics of maternal mortality using social media. In *Proceedings of The Web Conference 2020*, pages 2976–2983, 2020.

[Agarwal *et al.*, 2019] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International conference on machine learning*, pages 120–129. PMLR, 2019.

[Beck *et al.*, 2014] Audrey N. Beck, Brian K. Finch, Shih-Fan Lin, Robert A. Hummer, and Ryan K. Masters. Racial disparities in self-rated health: Trends, explanatory factors, and the changing role of socio-demographics. *Social Science & Medicine*, 104, 2014.

[Berk *et al.*, 2017] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

[Eichstaedt *et al.*, 2015] Johannes C Eichstaedt, Hansen Andrew Schwartz, Margaret L Kern, Gregory Park, Darwin R Labarthe, Raina M Merchant, Sneha Jha, Megha Agrawal, Lukasz A Dziurzynski, Maarten Sap, et al. Psychological language on twitter predicts county-level heart disease mortality. *Psychological science*, 26(2):159–169, 2015.

[Garimella *et al.*, 2022] Aparna Garimella, Rada Mihalcea, and Akhash Amarnath. Demographic-aware language model fine-tuning as a bias mitigation technique. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 311–319, November 2022.

[Giorgi *et al.*, 2018] Salvatore Giorgi, Daniel Preoţiuc-Pietro, Anneke Buffone, Daniel Rieman, Lyle Ungar, and H Andrew Schwartz. The remarkable benefit of user-level aggregation for lexical-based population-level predictions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1167–1172, 2018.

[Giorgi *et al.*, 2022a] Salvatore Giorgi, Veronica E Lynn, Keshav Gupta, Farhan Ahmed, Sandra Matz, Lyle H Ungar, and H Andrew Schwartz. Correcting sociodemographic selection biases for population prediction from social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 228–240, 2022.

[Giorgi *et al.*, 2022b] Salvatore Giorgi, Khoa Le Nguyen, Johannes C Eichstaedt, Margaret L Kern, David B Yaden, Michal Kosinski, Martin EP Seligman, Lyle H Ungar, H Andrew Schwartz, and Gregory Park. Regional personality assessment through social media language. *Journal of personality*, 90(3):405–425, 2022.

[Giorgi *et al.*, 2023] Salvatore Giorgi, David B Yaden, Johannes C Eichstaedt, Lyle H Ungar, H Andrew Schwartz, Amy Kwarteng, and Brenda Curtis. Predicting us county opioid poisoning mortality from multi-modal social media and psychological self-report data. *Scientific reports*, 13(1):9027, 2023.

[Hovy and Søgaard, 2015] Dirk Hovy and Anders Søgaard. Tagging performance correlates with author age. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)*, pages 483–488, 2015.

[Hovy, 2015] Dirk Hovy. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, 2015.

[Hu and Collier, 2024] Tiancheng Hu and Nigel Collier. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, August 2024.

[Huang and Paul, 2019] Xiaolei Huang and Michael Paul. Neural user factor adaptation for text classification: Learning to generalize across author demographics. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)*, pages 136–146, 2019.

[Jaidka *et al.*, 2020] Kokil Jaidka, Salvatore Giorgi, H Andrew Schwartz, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. Estimating geographic subjective well-being from twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the national academy of sciences*, 117(19):10165–10171, 2020.

[Lawless and Lucas, 2011] Nicole M Lawless and Richard E Lucas. Predictors of regional well-being: A county level analysis. *Social Indicators Research*, 101(3):341–357, 2011.

[Lemstra *et al.*, 2006] Mark Lemstra, Cory Neudorf, and John Opondo. Health disparity by neighbourhood income. *Canadian Journal of Public Health*, 97:435–439, 2006.

[Lynn *et al.*, 2017] Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz. Human centered NLP with user-factor adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1155, September 2017.

[Mangalik *et al.*, 2024] Siddharth Mangalik, Johannes C Eichstaedt, Salvatore Giorgi, Jihu Mun, Farhan Ahmed, Gilvir Gill, Adithya V. Ganesan, Shashanka Subrahmanya, Nikita Soni, Sean AP Clouston, et al. Robust language-based mental health assessments in time and space through social media. *NPJ Digital Medicine*, 7(1):109, 2024.

[Matero *et al.*, 2023] Matthew Matero, Salvatore Giorgi, Brenda Curtis, Lyle H Ungar, and H Andrew Schwartz. Opioid death projections with ai-based forecasts using social media language. *NPJ Digital Medicine*, 6(1):35, 2023.

[Mehrabi *et al.*, 2021] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[O'Donnell *et al.*, 2007] Owen Andrew O'Donnell, Eddy K.A. Van Doorslaer, Adam Wagstaff, and Magnus Lindelow. *Analyzing Health Equity Using Household Survey Data: A Guide to Techniques and Their Implementation*, volume 1. November 2007.

[Pessach and Shmueli, 2022] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[Rai *et al.*, 2024] Sunny Rai, Elizabeth C Stade, Salvatore Giorgi, Ashley Francisco, Lyle H Ungar, Brenda Curtis, and Sharath C Guntuku. Key language markers of depression on social media depend on race. *Proceedings of the National Academy of Sciences*, 121(14):e2319837121, 2024.

[Rashkin *et al.*, 2019] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, 2019.

[Rawat *et al.*, 2024] Rajat Rawat, Hudson McBride, Rajarshi Ghosh, Dhiyaan Nirmal, Jong Moon, Dhruv Alamuri, Sean O'Brien, and Kevin Zhu. DiversityMedQA: A benchmark for assessing demographic biases in medical diagnosis using large language models. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 334–348, November 2024.

[Remington *et al.*, 2015] Patrick L Remington, Bridget B Catlin, and Keith P Gennuso. The county health rankings: rationale and methods. *Population health metrics*, 13(1):11, 2015.

[Roller *et al.*, 2021] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, 2021.

[Salinas *et al.*, 2023] Abel Salinas, Parth Shah, Yuzhong Huang, Robert McCormack, and Fred Morstatter. The unequal opportunities of large language models: Examining demographic biases in job recommendations by chatgpt and llama. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, 2023.

[Sap *et al.*, 2019] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678, 2019.

[Schwartz *et al.*, 2017] H Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Lyle Ungar, and Johannes Eichstaedt. Dlatk: Differential language analysis toolkit. In *Proceedings of the 2017 conference on empirical methods in natural language processing: System demonstrations*, pages 55–60, 2017.

[Shah *et al.*, 2020] Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, July 2020.

[Shavers, 2007] Vickie L. Shavers. Measurement of socioeconomic status in health disparities research. *Journal of the National Medical Association*, 99(9):1013–1023, Sep 2007.

[Soni *et al.*, 2024] Nikita Soni, H. Andrew Schwartz, João Sedoc, and Niranjan Balasubramanian. Large human language models: A need and the challenges. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8631–8646, June 2024.

[Wilson, 2019] Daniel J. Wilson. The harmonic mean value for combining dependent tests. *Proceedings of the National Academy of Sciences*, 116(4):1195–1200, 2019.

[Zamani and Schwartz, 2017] Mohammadzaman Zamani and H. Andrew Schwartz. Using Twitter language to predict the real estate market. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 28–33, April 2017.

[Zamani *et al.*, 2018] Mohammadzaman Zamani, H. Andrew Schwartz, Veronica Lynn, Salvatore Giorgi, and Niranjan Balasubramanian. Residualized factor adaptation for community social media prediction tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3560–3569, October-November 2018.

[Zhao *et al.*, 2017] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, September 2017.