Characterizing Sociodemographic Error Disparities in Large-scale Language-based Health Predictions

Anonymous ACL submission

Abstract

Most NLP bias studies focus on individual- or document-level tasks, yet fields where bias has substantial consequences, like public health, operate at the community-level. We systematically examine sociodemographic error disparities in NLP models predicting *community-level* health outcomes across billions of community-mapped messages and evaluate four sociodemographic factor inclusion strategies. We introduce the *Bilateral Concentration Index* (BCI) to quantify non-monotonic disparities missed by traditional metrics, finding all baseline language-alone models had moderate disparities (average BCI=6.6%). However, while incorporating sociodemographics into modeling consistently improved *accuracy*, it often increased *disparities*, from negligible (concatenation: BCI=6.6%) to significantly (adaptation: BCI=8.2%), suggesting a cost-benefit trade-off. Largest disparities in error emerged over education and income (BCI= 2.7–16.4%), reducing accuracy for low-income (and sometimes high-income) communities, which could disadvantage them if used for policy decisions. These findings suggest the need to evaluate error disparities alongside accuracy to ensure fairness as models enter real-world applications.

1 Introduction

Regional disparities in health—reliable differences in outcomes by sociodemographic characteristics—are extensively studied in public health and social sciences to inform fair resource allocation (Beck et al., 2014; Lemstra et al., 2006; Shavers, 2007). Within NLP pipelines, biases leading to *error disparities* (varying model accuracy by sociodemographic attributes (Shah et al., 2020)) have typically been analyzed at the *document-* or *individual-level* (Salinas et al., 2023; Garimella et al., 2022; Rawat et al., 2024). However, for *community-level* tasks, such as predicting regional well-being, biases are less known. Understanding error disparities at the community-level is particularly critical for NLP models to inform public health policy.

Here, we systematically evaluate language-based predictive models across four community-level health tasks and three sociodemographic dimensions shown to have selection biases on Twitter (Giorgi et al., 2022a): percentage of foreign-born, percentage of educated, and median income of the population. We focus on sociodemographic ("human factor") inclusion techniques (Zamani et al., 2018) known to substantially improve model accuracy (Giorgi et al., 2023; Hovy, 2015), though their impact on bias or *error disparities* is unknown. 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, so we explore two different types: (1) *additive*, directly offsetting average outcome differences (e.g., heart disease rates for low versus high income), and (2) *adaptive*, adjusting language semantics to reflect sociodemographic context (e.g., different meaning of "club" for low- versus high-income).

We provide three **contributions**: **(1)** identifying community well-being tasks and sociodemographic factors most prone to model error disparities; **(2)** analyzing how *additive* and *adaptive* sociodemographic inclusion methods affect disparities and how this relates to their accuracy; and **(3)** proposing the Bilateral Concentration Index (BCI), an analog of the popular *Gini-coefficient* (Gini, 1912) from health disparity research, to quantify *error disparities*, capturing non-linear and non-monotonic sociodemographic-error relationships.

2 Related Work

The integration of sociodemographic factors into language-based predictive models, methods and challenges, has been investigated for at least a decade (Hovy, 2015; Lynn et al., 2017; Soni et al.,

Outc.\ Demog.	Forgn Born	HS Grad	Income
Heart Dis Mort.	4.1 %*	9.0 %**	9.2 %**
Life Satis.	9.9 %**	5.2 %**	5.8 %**
%FairPoor Hlth	4.8 %**	10.8 %**	7.5 %**
Suicide Mort.	4.4 %*	2.7 %	5.6 %*

Table 1: Error disparity (BCI) for the given sociodemographic factor (Demog.) and across language-based predictive models for the four community health tasks. Asterisk represents statistically significant difference from a random baseline (* $p < .05$, ** $p < .01$ from a permutation test).

2024). Sociodmeographic factors explored include, e.g., income, age, gender, and geographic location (Huang and Paul, 2019). Additionally, dialog systems are increasingly designed with human-like traits such as empathy and emotions (Rashkin et al., 2019; Omitaomu et al., 2022) or personas (Roller et al., 2021). Recent work has suggested that prompting generative LMs with personas reveals internal biases and simulates human roles in crowdsourcing tasks (Hu and Collier, 2024).

Work on *error disparity* (Shah et al., 2020) started approximately with the "Wall Street Journal effect," where POS taggers performed worse as user demographics diverged from WSJ training authors (Hovy and Søgaard, 2015); disparities in hate detection for Black authors due to annotators missing racial context (Sap et al., 2019); and lower accuracy in mental health prediction for Black versus matched White samples, even with Black-only training data (Rai et al., 2024). Though these studies did not address community-level tasks, they motivate exploring methods to account for sociodemographic differences in error, to calibrate models effectively for diverse populations.

3 Data Set

We use the open-source **County Tweet Lexical Bank** (CTLB) which contains 25,000 English-language lexical features across 2,041 US counties, derived from over 1.5 billion geolocated tweets (Giorgi et al., 2018). We focus on **three sociodemographic factors** that have had high predictive values 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. We consider **four county health tasks**: heart disease mortality (**HD**; $N = 1750$), life satisfaction (**LS**; $N = 1745$), percentage reporting 'fair'/'poor' health (**FP**; $N = 1703$), and suicide mortality (**SM**; $N = 1631$). These outcomes were chosen to be consistent with past community-level NLP tasks on selection bias (Giorgi et al., 2022a). See Appendix A for more details.

4 Methods

We describe the predictive models, factor inclusion techniques, and disparity metrics. With the focus being inclusion techniques and disparity metrics, we a use well-established technique for predictive modeling. Specifically, an ℓ_2 penalized (ridge) regression was used to estimate the outcomes (HD, LS, FP, SM) from county lexical and/or sociodemographic features. We recorded absolute errors for each county over 10-fold cross-validation with hyper-parameters set over a subset of training (§Appendix B).

Factor Inclusion Methods. We explored four factor inclusion techniques for integrating sociodemographic factors into language-based predictive models. Techniques spanned two overall strategies: (1) *additive* - direct inclusion accounting for baseline differences in outcomes depending on the sociodemographic factor (Preoţiuc-Pietro et al., 2015) and (2) *adaptive* - accounting for differences in the meaning of words or phrases depending on demographics. For example, the word "mean" might have one sense as "cruel," but among more educated populations could more often signify the mathematical average sense of the word (Lynn et al., 2017).

As additive techniques, we utilize: (1) **Factor Concatenation (FC)** – sociodemographic factors are concatenated with language features in a single feature vector; (2) **Residualized Controls (ResC)** – sociodemographic controls are first modeled independently and then the language-based model is fit to predict the residual from the control model (Zamani et al., 2018). By fitting to controls alone first, ResC ensures the they are not lost among the numerous language dimensions (Zamani and Schwartz, 2017).

As adaptive techniques, we utilize: (3) **Factor Adaptation (FA)** – linguistic features are composed with sociodemographic control variables allowing language features to have subtle difference in meaning depending on the author background (Lynn et al., 2017). We use the compositional function multiplying mean centered versions of the controls with the language features found

Demog Factor	Task	Disparity and Accuracy (Bilateral Concentration Index and Pearson r)											
		Lang (L)		L+C		ResC		FA		RFA		Cont (C)	
		BCI	r	BCI	r	BCI	r	BCI	r	BCI	r	BCI	r
Foreign Born	HD	4.1%	.749	4.2%	.750	**3.6%**	.747	5.8%	.764	5.5%	.763	4.4%	.351
	LS	<u>9.9%</u>	.450	<u>9.9%</u>	.451	9.6%	.447	9.4%	.502	9.1%	.491	11.1%	—-
	FP	4.8%	.764	4.8%	.764	4.4%	.754	5.9%	.773	5.8%	.770	4.9%	.078
	SM	4.4%	.635	4.7%	.633	7.0%	.671	8.2%**	.673	7.3%*	.670	1.9%	.354
High school Grad	HD	9.0%	.749	9.1%	.750	14.9%**	.730	12.2%*	.771	12.5%*	.765	13.7%	.526
	LS	5.2%	.450	5.0%	.456	3.3%	.505	3.5%	.541	3.6%	.518	4.1%	.306
	FP	10.8%	.764	11.0%	.769	<u>16.4%</u>**	.781	15.0%*	.808	15.1%*	.803	14.1%	.740
	SM	**2.7%**	.635	2.6%	.636	3.4%	.622	3.1%	.664	3.2%	.661	1.5%	—-
Income	HD	9.2%	.749	9.4%	.752	9.9%	.747	12.5%*	.780	<u>12.7%</u>*	.779	8.4%	.574
	LS	5.8%	.450	4.4%	.478	4.3%	.530	**3.7%**	.566	4.0%	.551	4.6%	.365
	FP	7.5%	.764	7.9%	.770	7.9%	.798	10.6%*	.813	10.4%	.811	7.0%	.649
	SM	5.6%	.635	5.8%	.637	7.6%	.636	8.5%	.655	8.1%	.647	5.3%	.304
Global Avg		**6.6%**	.649	**6.6%**	.654	7.7%**	.664	<u>8.2%</u>*	.692	8.1%*	.686	6.8%	.352

Table 2: **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*. Outcomes are heart disease (HD), life satisfaction (LS), fair/poor health (FP), and suicide mortality (SM). Factor inclusion methods beyond Language (L) and Sociodemographic Control (C) are Factor Concatenation (L + C), Residualized Controls (ResC), Factor Adaptation (FA), and Residualized Factor Adaptation (RFA). Dashes signify not significant results. **Bold** represents tests with the lowest disparity per sociodemographic factor. <u>Underline</u> 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).

beneficial in past work (Lynn et al., 2017); (2) **Residualized Factor Adaptation (RFA)** – combining FA and ResC, an FA model is fit to the residual of a control-only model offering the advantages of both (Zamani et al., 2018)[1].

Measuring Disparity. While past works in NLP-based predictive biases often compare error by sociodemographic groups, e.g., Hovy and Søgaard (2015); Zhao et al. (2017), community-level sociodemographic are often continuous (not group; e.g. percentages or averages). Social scientific works often utilize the Gini-coefficient (Gini, 1912) but it is limited to measures unidimensional disparities and require measuring disparities one variable (e.g. error) conditioned on another (e.g. median income of the community). We formulate an analog to Gini that captures the disparity in model performance with respect to a sociodemographic variable (sociodemographic factor), the **Bilateral Concentration Index (BCI)**.

BCI is adaptation of the concentration index based on the cumulative percent of total error for each county sorted by the sociodemographic variable (O'Donnell et al., 2007). To calculate BCI we take the integration of the difference between the concentration curve and a cumulative uniform distribution (a 45° diagonal – perfect equality):

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

where N is the total number of counties which are ordered sequentially from lowest to highest error. $f_i(x)$ represents the disparity at any point x between the interval $\frac{i}{N}$ to $\frac{i+1}{N}$– the cumulative error (e_i) compared to the expectation from the cumulative uniform (u_i) within the interval between counties:

$$f_i(x) = |(m_{e_i}x - e_{i+1}) - (m_{u_i}x - u_{i+1})| \quad (2)$$

$$m_{e_i} = \frac{e_{i+1} - e_i}{\frac{i}{N}}, m_{u_i} = \frac{u_{i+1} - u_i}{\frac{i}{N}} \quad (3)$$

Curves with large area under the cumulative uniform distribution indicate prediction error increases with the sociodemographic variable; curves above the diagonal indicate the opposite (Figure A2). Importantly, this approach treats observations continuously without binning, enabling a granular consideration of each observation's effect. The BCI metric is intuitive (maximum at 100%), but we also apply the Anderson-Darling test (AD) to assess significant disparities (see Appendix E).

[1] See Appendix D for mathematical notations

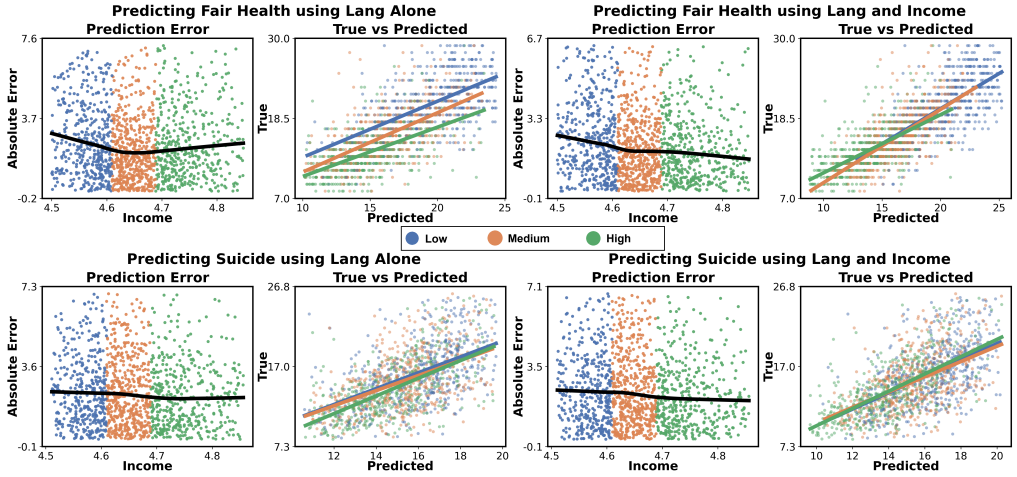

Figure 1: Scatter plots of outcomes with respect to income of a county. Prediction errors as a function of logged income in 1st and 3rd columns. Income is colored by tercile; LOESS curve in black. Predicted vs true outcome values in 2nd and 4th columns. Linear regression lines plotted for each income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

5 Results

We systematically evaluate error disparities of language-based predictive models across four tasks and three sociodemographic controls. We first establish overall disparities for language-alone models in Table 1, finding significant error disparities in every case except for suicide mortality with HS graduation. For example, substantial disparities were observed for Life Satisfaction predictions across foreign born percentage (BCI=9.9%). This means models were more accurate depending on the amount foreign born (less foreign born meant better accuracy in this case). Other large disparities included model predicting heart disease with income (BCI=9.2%) and Fair or Poor health with HS graduation (BCI=10.8%).

Table 2 shows results across the four types of inclusion techniques and controls alone (C). On average, all inclusion techniques improved accuracy over the language-alone results but often at the expense of an increase in error disparity. For example, factor adaptation (FA) while producing the best accuracies also had an average disparity BCI of 8.2%, an increase over the 6.6% observed from language alone. On the other hand, the simple concatenation approach (L+C) did not seem to increase disparities but it also did not substantially increase accuracy. Interestingly, control alone models did not have large error disparities, though this could be due to their low predictive performance overall, leaving less room for disparity.

To depict patterns disparities with respect to income, we visualize both error and prediction scatter plots for fair and poor health (high disparity) as well as suicide mortality (low disparity) in Figure 1. The slope of the LOESS (Cleveland, 1979) and the Bilateral Concentration Index are approximately proportional in magnitude. We observed non-linear patterns where simply being further from the mean in income meant worse performance, while for others, we observed models working better for those communities with higher income.

6 Conclusion

In a systematic evaluation of community-level health prediction tasks, we observed error disparities across three demographics and most tasks. We further analyzed the effect of sociodemographic factor inclusion methods on disparity in trade-off for accuracy improvements. We found that predicting outcomes such as heart disease and fair/poor health had much higher error for counties with lower education or income and accuracies for life satisfaction were lower for counties where the percentage of foreign born population was higher. While one might have expected factor inclusion methods to reduce error by better capturing differences in semantics by sociodemographic group, we found that, on average, such approaches, especially adaptive approaches, increased disparities. Overall results suggest that there are significant disparities in model performances at the county level for most sociodemographics and that the utility of introducing sociodemographic factors into such models depends the context, rather than having a universally positive or negative impact.

7 Limitations

To systematically study sociodemographic factor inclusion methods and their effects on bias (sociodemographic error disparities), we evaluated four methods across four outcomes. Despite this, this study is not exhaustive nor representative. For example, we evaluated a limited set of sociodemographic factors (foreign-born, education, and income). Several studies have shown race as a source of error disparities (Rai et al., 2024; Sap et al., 2019), which was not evaluated in the current study. Furthermore, the data set is limited in representation: we only consider communities in the US with sufficiently large number of Twitter users. Thus, our results may not extend to other regions or cultures. Finally, studies have shown error disparities at the document level (i.e., hate speech labels on social media posts; Sap et al., 2019), which was not evaluated in the current study. Though we think the factor inclusion approaches chosen are straightforward, and therefore provide a good basis for generalization, additional techniques could be tested as well.

8 Ethics

This study was reviewed and approved by the [redacted] Institutional Review Board. It is important to consider and discourage the potential negative applications of this work. Our approach can be utilized to uncover societal as well as individual error disparities, even within targeted recommendation systems. However, we recognize that, if misapplied, it could be leveraged to amplify algorithmic biases and exacerbate inequities. The results described could reinforce existing biases contributing to additional stigma towards a group. Additionally, "fairwashing" or blindly trusting models because they showed propensity for fairness in this study could lead to unaccounted for error disparity in new applications of these models. Our work is intended for researchers and practitioners of Social Science, and we don't condone the usage of such algorithms for malicious purposes.

References

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

Audrey N. Beck, Brian K. Finch, Shih-Fan Lin, Robert A. Hummer, and Ryan K. Masters. 2014. Racial disparities in self-rated health: Trends, explanatory factors, and the changing role of sociodemographics. Social Science & Medicine, 104:163–177.

William S. Cleveland. 1979. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association, 74(368):829–836.

Aparna Garimella, Rada Mihalcea, and Akhash Amarnath. 2022. 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, Online only. Association for Computational Linguistics.

C Gini. 1912. Variabilità e mutabilità (in italian). reprinted. Memorie di metodologia statistica.

Salvatore Giorgi, Veronica E Lynn, Keshav Gupta, Farhan Ahmed, Sandra Matz, Lyle H Ungar, and H Andrew Schwartz. 2022a. 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.

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. 2022b. Regional personality assessment through social media language. Journal of personality, 90(3):405–425.

Salvatore Giorgi, Daniel Preoţiuc-Pietro, Anneke Buffone, Daniel Rieman, Lyle Ungar, and H Andrew Schwartz. 2018. 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.

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

Dirk Hovy. 2015. 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.

Dirk Hovy and Anders Søgaard. 2015. 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.

Tiancheng Hu and Nigel Collier. 2024. 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, Bangkok, Thailand. Association for Computational Linguistics.

Xiaolei Huang and Michael Paul. 2019. 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.

Kokil Jaidka, Salvatore Giorgi, H Andrew Schwartz, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2020. 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.

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

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

Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2017. Human centered NLP with user-factor adaptation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1146–1155, Copenhagen, Denmark. Association for Computational Linguistics.

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

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

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and João Sedoc. 2022. Empathic conversations: A multi-level dataset of contextualized conversations. arXiv preprint arXiv:2205.12698.

Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H. Andrew Schwartz, and Lyle Ungar. 2015. The role of personality, age, and gender in tweeting about mental illness. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 21–30, Denver, Colorado. Association for Computational Linguistics.

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

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. 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.

Rajat Rawat, Hudson McBride, Rajarshi Ghosh, Dhiyaan Nirmal, Jong Moon, Dhruv Alamuri, Sean O'Brien, and Kevin Zhu. 2024. 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, Miami, Florida, USA. Association for Computational Linguistics.

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

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021. 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.

Abel Salinas, Parth Shah, Yuzhong Huang, Robert McCormack, and Fred Morstatter. 2023. 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, New York, NY, USA. Association for Computing Machinery.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. 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.

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

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. 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, Online. Association for Computational Linguistics.

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

Nikolai V Smirnov. 1939. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. Bull. Math. Univ. Moscou, 2(2):3–14.

Nikita Soni, H. Andrew Schwartz, João Sedoc, and Niranjan Balasubramanian. 2024. 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, Mexico City, Mexico. Association for Computational Linguistics.

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

Mohammadzaman Zamani and H. Andrew Schwartz. 2017. 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, Valencia, Spain. Association for Computational Linguistics.

Mohammadzaman Zamani, H. Andrew Schwartz, Veronica Lynn, Salvatore Giorgi, and Niranjan Balasubramanian. 2018. 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, Brussels, Belgium. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. 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, Copenhagen, Denmark. Association for Computational Linguistics.

Appendices

A Data Details

Community Language The County Tweet Lexical Bank (CTLB) is an open-source data set of US county-level language features. High-level details are described here and further details 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 US counties 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. 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 for each of the 6 million users. These user-level features were then averaged within each county to produce a set of US county lexical features (i.e., each county is represented by a vector of 25,000 unigram frequencies). 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 Five year estimates (2011-2015) for foreign born (percentage of a country's population that was born in another country), education (% of the population with a high school diploma), and income (median log annual household income) were obtained from United States Census Bureau's 2015 American Community Survey (ACS).

Community Outcomes We gathered age-adjusted mortality rates for heart disease and suicide from the Centers for Disease Control and Prevention (CDC), averaged over the years 2010-2015. Life satisfaction 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 came from the County Health Rankings, drawing on the Behavioral Risk Factor Surveillance System (BRFSS; 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?".

B Model Details

The same feature selection and modeling procedures were used across all four outcomes. In order to reduce the feature space, we performed a feature selection pipeline. First, we performed univariate feature selection, removing all features that were not significantly correlated at a family-wise error rate of 60. Next, we use principal component analysis (PCA) to further reduce the features. The dimension reduction size for PCA was chosen based on the size of the training fold.

All models were evaluated using 10-fold cross validation using a linear regression with ℓ_2 regularization (Ridge regression). The regression regularization parameter α was chosen via nested cross validation.

Feature extraction (unigrams) as well as predictive modeling were all done using the open-source Python package DLATK (Schwartz et al., 2017).

C Bilateral Concentration Index

Figure A1 is a visualization of a hypothetical concentration curve that crosses the line of equality. The light blue area represents the BCI.

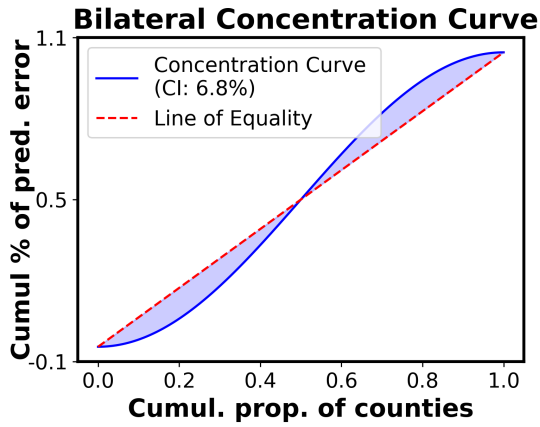

Figure A1: Zoomed in Bilateral Concentration Curve: BCI shown (where the red line is the cumulative uniform distribution, and the blue line is the predicted error of counties sorted by sociodemographic variable)

Figure A2 depicts another hypothetical concentration curve where n, or the number of counties, is ten and the cumulative error for each county crosses the line of equality between counties five and six. This example illustrates the distinction in behavior between the existing Concentration Index and the Bilateral Concentration Index as the BCI accounts summatively for all area difference between the line of equality and the cumulative error curve. The relevant variables used to solve the BCI using equations 1, 2, and 3 for this interval ($[\frac{i}{n}, \frac{i+1}{n}]$) are labeled.

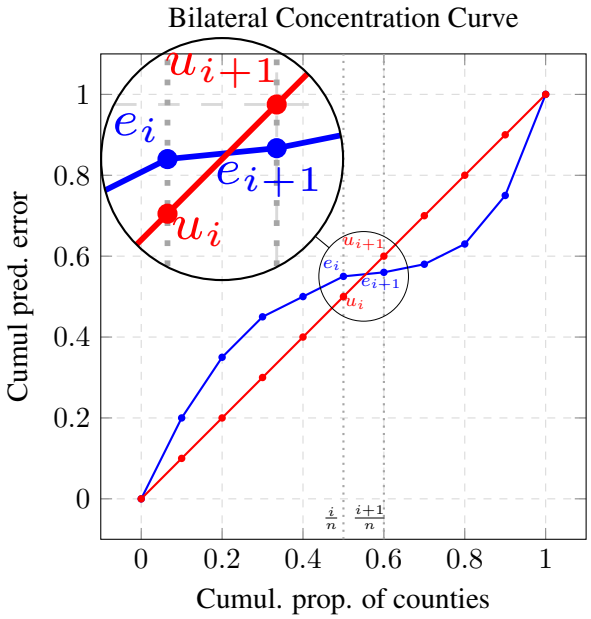

Figure A2: Bilateral Concentration Curve: the blue line is cumulative error curve and red line is the cumulative uniform line

D Factor Inclusion Methods

Residualized Controls can be represented mathematically as follows

$$\varepsilon = Y - \hat{Y}_C \quad (4)$$

$$\hat{\varepsilon}_L = \gamma \times X_L + \lambda \quad (5)$$

where ε is the residual 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.

In Factor Adaptation, the adapted language features (X_{A_i}) are combined as follows:

$$X_{A_i} = [X_L \cdot C_i], \forall i \in [1, |C|] \quad (6)$$

$$X_F = [X_L, X_{A_1}, ..., X_{A_{|C|}}] \quad (7)$$

Residualized Factor Adaptation can be represented as

$$\hat{\varepsilon}_L = \gamma \times \left[X_L, X_{A_1}, X_{A_2}, ..., X_{A_{|C|}}\right] + \lambda \quad (8)$$

E Additional Disparity Metrics

For comparison, we also ran existing disparity metrics Anderson-Darling, KS Tests, and Cross Entropy to evaluate disparity between cumulative prediction error and a cumulative uniform error. Cross

Demog Group	Task	Disparity Metrics: KS, CE, AD, BCI																							
		Lang (L)				Cont (C)				L+C				ResC				FA				RFA			
		KS	CE	AD	BCI	KS	CE	AD	BCI	KS	CE	AD	BCI	KS	CE	AD	BCI	KS	CE	AD	BCI	KS	CE	AD	BCI
Foreign Born	HD	.040	4.24	8157	4.1%	.037	4.25	9904	4.4%	.041	4.24	8523	4.2%	.038	4.24	6592	3.6%	.051	4.27	15881	5.8%	.049	4.27	14349	5.5%
	LS	.077	4.35	45767	9.9%	.090	4.39	58977	11.1%	.077	4.35	45446	9.9%	.075	4.34	42564	9.6%	.076	4.34	40610	9.4%	.073	4.33	38252	9.1%
	FP	.048	4.26	13016	4.8%	.051	4.28	16560	4.9%	.048	4.26	13060	4.8%	.045	4.26	11505	4.4%	.055	4.27	18419	5.9%	.055	4.27	17701	5.8%
	SM	.047	4.27	8726	4.4%	.017	4.21	2447	1.9%	.051	4.27	9902	4.7%	.065	4.29	20230	7.0%	**.076**	4.32	**27708**	**8.2**%	**.068**	4.30	**22017**	**7.3**%
High school Grad	HD	.074	4.34	39872	9.0%	.105	4.50	95668	13.7%	.074	4.34	40863	9.1%	**.108**	**4.57**	**111573**	**14.9**%	**.092**	**4.44**	**72498**	**12.2**%	**.095**	**4.45**	**76683**	**12.5**%
	LS	.043	4.27	16323	5.2%	.034	4.25	10566	4.1%	.041	4.26	15092	5.0%	.028	4.23	6731	3.3%	.029	4.23	7554	3.5%	.029	4.23	7325	3.6%
	FP	.100	4.45	61605	10.8%	.115	4.54	91734	14.1%	.100	4.45	62690	11.0%	**.131**	**4.66**	**124552**	**16.4**%	**.123**	4.58	**101910**	**15.0**%	**.124**	4.58	**103679**	**15.1**%
	SM	.027	4.24	4785	2.7%	.023	4.21	2132	1.5%	.027	4.24	4648	2.6%	.033	4.25	6596	3.4%	.030	4.23	4972	3.1%	.030	4.24	5056	3.2%
Income	HD	.076	4.35	42044	9.2%	.066	4.30	33119	8.4%	.077	4.35	44018	9.4%	.081	4.40	51295	9.9%	**.097**	**4.45**	**75203**	**12.5**%	**.099**	**4.46**	**76961**	**12.7**%
	LS	.051	4.29	18368	5.8%	.042	4.26	11029	4.6%	.037	4.25	9876	4.4%	.039	4.24	9711	4.3%	.037	4.25	7770	3.7%	.038	4.24	8508	4.0%
	FP	.073	4.35	32151	7.5%	.056	4.31	22060	7.0%	.073	4.35	33774	7.9%	.067	4.33	30154	7.9%	.083	4.39	**50514**	**10.6**%	.081	4.38	48925	10.4%
	SM	.052	4.25	12973	5.6%	.046	4.28	12254	5.3%	.052	4.25	13624	5.8%	.058	4.29	23406	7.6%	.068	4.32	28929	8.5%	.063	4.30	26258	8.1%
Global Avg		.059	4.305	25315	6.6%	.057	4.315	30537	6.8%	.058	4.301	25126	6.6%	.064	4.342	**37075**	**7.7**%	.068	4.341	**37664**	**8.2**%	.067	4.338	**37142**	**8.1**%

Table T1: **Disparities across county outcomes and different sociodemographic factor inclusion approaches:** Disparity is measured using the KS Test (KS), the Cross Entropy (CE), Anderson-Darling (AD), and the Bilateral Concentration Index (BCI) (as a percent) each comparing the cumulative error over counties, sorted by sociodemographic group, to a cumulative uniform distribution (Smirnov, 1939). Outcomes are heart disease (HD), life satisfaction (LS), fair/poor health (FP), and suicide mortality (SM). Factor inclusion methods beyond Language (L) and Demographic Control (C) are Factor Concatenation (L + C), Residualized Controls (ResC), Factor Adaptation (FA), and Residualized Factor Adaptation (RFA). **Bold** represents statistically significant difference from disparity with the same parameters using language alone (L). 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).

Entropy isnt as interpretable. KS test is much more interpretable, but fails to account for significant disparity in the tails of the county error distribution. The Anderson-Darling test is best equipped to account for the entirety of the distribution, but is also difficult to interpret. We use the BCI because it possesses the strengths of each of these methods. The results can be seen in Table T1.

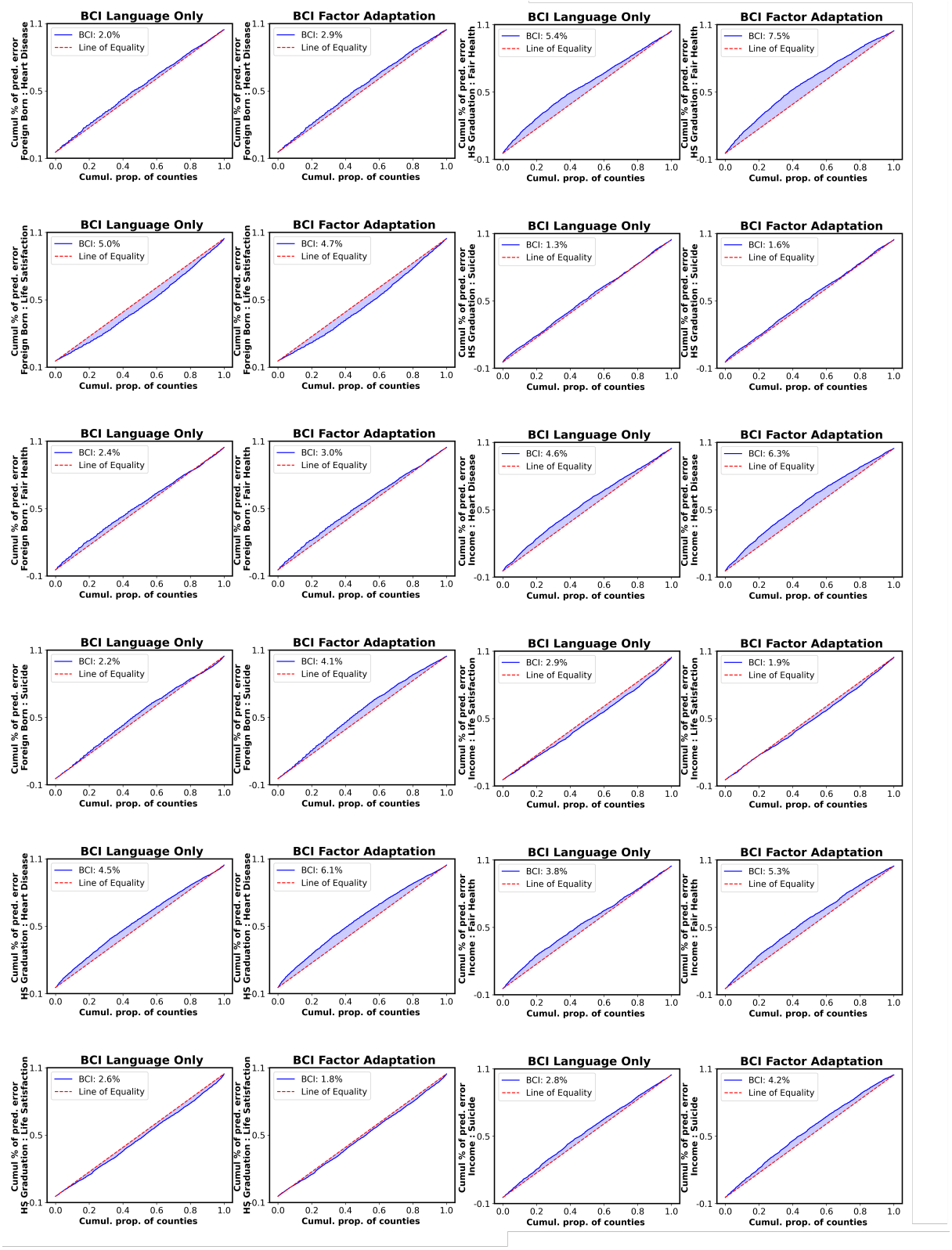

Figure A3: All Bilateral Concentration Curves: BCIs for all combinations of sociodemographic variable and outcome (where the red line is the cumulative uniform distribution, and the blue line is the predicted error of counties sorted by sociodemographic variable)