SSM - Population Health 13 (2021) 100750

Contents lists available at ScienceDirect

SSM - Population Health

journal homepage: http://www.elsevier.com/locate/ssmph

Evaluating associations between area-level Twitter-expressed negative racial sentiment, hate crimes, and residents' racial prejudice in the United States

Thu T. Nguyen[a,*], Dina Huang[b], Eli K. Michaels[c], M. Maria Glymour[d], Amani M. Allen[c,e], Quynh C. Nguyen[b]

[a] Department of Family and Community Medicine, University of California San Francisco, San Francisco, CA, 94110, USA
[b] Department of Epidemiology & Biostatistics, University of Maryland School of Public Health, College Park, MD, 20742, USA
[c] Division of Epidemiology, University of California, Berkeley, CA, 94704, USA
[d] Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, CA, 94158, USA
[e] Divisions of Community Health Sciences, University of California, Berkeley, CA, 94704, USA

ARTICLE INFO

Keywords:
Racial bias
Racial discrimination
Big data
Social media
Measures

ABSTRACT

Background: The objective of the current study is to investigate whether an area-level measure of racial sentiment derived from Twitter data is associated with state-level hate crimes and existing measures of racial prejudice at the individual-level.

Methods: We collected 30,977,757 tweets from June 2015–July 2018 containing at least one keyword pertaining to specific groups (Asians, Arabs, Blacks, Latinos, Whites). We characterized sentiment of each tweet (negative vs all other) and averaged at the state-level. These racial sentiment measures were merged with other measures based on: hate crime data from the FBI Uniform Crime Reporting Program; implicit and explicit racial bias indicators from Project Implicit; and racial attitudes questions from General Social Survey (GSS).

Results: Living in a state with 10% higher negative sentiment in tweets referencing Blacks was associated with 0.57 times the odds of endorsing a GSS question that Black-White disparities in jobs, income, and housing were due to discrimination (95% CI: 0.40, 0.83); 1.64 times the odds of endorsing the belief that disparities were due to lack to will (95% CI: 0.95, 2.84); higher explicit racial bias (β: 0.11; 95% CI: 0.04, 0.18); and higher implicit racial bias (β: 0.09; 95% CI: 0.04, 0.14). Twitter-expressed racial sentiment was not statistically-significantly associated with incidence of state-level hate crimes against Blacks (IRR: 0.99; 95% CI: 0.52, 1.90), but this analysis was likely underpowered due to rarity of reported hate crimes.

Conclusion: Leveraging timely data sources for measuring area-level racial sentiment can provide new opportunities for investigating the impact of racial bias on society and health.

Introduction

In the U.S., racial disparities persist for a variety of health outcomes (Alhusen, Bower, Epstein, & Sharps, 2016; Pool, Ning, Lloyd-Jones, & Allen, 2017; Sternthal, Slopen, & Williams, 2011). Racism creates and perpetuates these disparities via both personal interactions and more systemic forms of inequality (Paradies et al., 2015; Pascoe & Smart Richman, 2009; Phelan & Link, 2015). However, the measurement of racism remains a challenge to evaluating the its full impact on health.

Racism exists at multiple levels, including individual, interpersonal, and institutional, and represents both negative normative beliefs (stereotypes) and attitudes (prejudice) towards minoritized groups and differential treatment resulting in inequitable access to resources and opportunities (discrimination) (Williams, Lawrence, & Davis, 2019) and psychological strain due to increased stress burden. Self-reported racial attitudes and beliefs, generally assessed via survey, are subject to a number of limitations including social desirability bias and self-censorship (An, 2015; Stocké, 2007), risking invalid exposure assessment (Nuru-Jeter et al., 2018). Experiences with racial discrimination are also commonly measured at the individual level by self-report

* Corresponding author. 995 Potrero Ave., Bldg 20, San Francisco, CA, 94110, USA.
E-mail addresses: thu.nguyen@ucsf.edu (T.T. Nguyen), dinahuang26@gmail.com (D. Huang), maria.glymour@ucsf.edu (M.M. Glymour), amaniallen@berkeley.edu (A.M. Allen), qtnguyen@umd.edu (Q.C. Nguyen).

https://doi.org/10.1016/j.ssmph.2021.100750
Received 30 August 2020; Received in revised form 27 December 2020; Accepted 1 February 2021
Available online 6 February 2021

(Krieger, Smith, Naishadham, Hartman, & Barbeau, 2005; Williams, Yan, Jackson, & Anderson, 1997). Self-reports of racial discrimination can be influenced by coping (e.g., denial), trait- or state-based aspects of personality (e.g., stigma consciousness, race-based rejection sensitivity), and aspects of racial identity (e.g., internalized racism) (Nuru-Jeter et al., 2018).

Millions of tweets are sent daily by users across the globe, and 90% of Twitter users make their profile public (Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2011). Perceived anonymity associated with online spaces may decrease self-censorship of socially unacceptable views and increased willingness to express attitudes that are less likely to be reported in survey interviews due to social desirability response bias (Chae et al., 2018; Suler, 2004). Leveraging data from social media may be one way to circumvent some of the limitations of traditional self-report measures and help capture attitudes about sensitive topics such as racial prejudice and bias (Chae et al., 2015). In addition to providing a proxy measure for the typical racial attitude in a place, Twitter expressions may capture a racial climate that has influences on health above and beyond individual level attitudes.

An ecosocial approach to the study of racism views racism as operating across multiple levels over the life course and reflecting systemic prejudice, which has emergent properties of its own despite individual level experiences and institutional racial discrimination. Cultural racism is defined as infusion of the ideology of inferiority in the values, language, imagery, symbols, and unstated assumptions of the larger society (Williams et al., 2019). Cultural racism is displayed through media, stereotyping, and norms within society and its institutions. In this way, cultural racism is systemic and produces an environment where institutional and individual-level discrimination can thrive (Williams et al., 2019). Geronimus described this as the "surround" (Geronimus et al., 2016) and describes how it can deplete physical and psychological well-being independent of individual-level discrimination experiences.

To capture the "surround", or racial climate of an area, our research team developed an area-level measure of racial sentiment from Twitter data analyzed with machine learning models algorithms. Research examining area-level racial sentiment is in its infancy. Building on prior research, our measure offers new, cost-efficient data sources for characterizing area-level racial sentiment (Nguyen et al., 2019). The measure demonstrated associations with adverse birth outcomes (T. Nguyen et al., 2020; Nguyen et al., 2018) and cardiovascular outcomes (Huang, Huang, Adams, Nguyen, & Nguyen, 2020).

One potential mechanism linking area-level sentiment to poorer health outcomes is that it creates an environment that may encourage racism or the tolerance of racism. In this paper, we investigate whether individuals living in an area with higher negative racial sentiment harbor more racial prejudice. Hate crimes represent an extreme form of intentional, explicit discrimination. Existing national sources of racial prejudice come from the General Social Survey (GSS) and Project Implicit. The objective of the current study is to investigate whether area-level racial sentiment, using the Twitter measure we previously developed, is associated with 1) race-related hate crimes, and 2) with existing measures of prejudice based on individual level reports from more traditional data sources.

Methods

Measures

Twitter data

A random 1% sample of publicly available tweets were collected from June 2015 to July 2018 using Twitter's Streaming Application Programming Interface (API). We restricted our analyses to English language tweets from the United States with latitude and longitude coordinates or other place attributes that permitted identification of the state or county location where the tweet was associated. All tweets included in the analysis used one or more of 518 race-related keywords. These keywords were compiled from racial and ethnic categories used by prior studies examining race-related online conversations (Bartlett, Reffin, Rumball, & Williamson, 2014; Pew Research Center, 2016) and an online database of racial slurs ("The Racial Slur Database," 2018). Tweets were classified into five main racial/ethnic categories: Asians, Arabs, Blacks, Latinos, and Whites according to the keywords used. Details of the data collection process including the full keyword list have been previously published (Nguyen et al., 2019).

We performed a sentiment analysis on the Twitter data set. This procedure has been previously described (Nguyen et al., 2020). Briefly, we utilized Support Vector Machines (SVM), a supervised machine learning model to label the tweets. We obtained training data from manually labeled Sentiment140 (n=498) (Sentiment140, 2011), Kaggle (n=7,086) (Kaggle in Class, 2011), Sanders (n=5,113) (Sanders Analytics, 2011) and 6,481 tweets labeled by our research group. Sentiment140, Sanders, and Kaggle datasets are all publicly available training datasets specifically labeled for sentiment analysis. We compared negative tweets (assigned a value of 1) to all other tweets (assigned value of 0). We used 5-fold cross validation to assess model performance and reached a high level of accuracy for the negative classification of tweets (91%) and a high F1 score (84%). We then labeled all the collected tweets using SVM model by assigning a dichotomized sentiment value (1 versus 0) to each tweet. State and year specific sentiment variables were created by averaging the dichotomous sentiment value of tweets referencing various racial/ethnic groups. State-level sentiment scores are a continuous measure of the proportion of tweets that are negative and scaled so that a one-unit increase represents a 10% increase in proportion of tweets that have a negative sentiment. State and year specific racial sentiment data are then merged with state and year specific data on hate crimes, racial attitudes from the General Social Survey, and explicit and implicit bias from Project Implicit.

Hate crimes

We used 2015–2018 hate crime data from the FBI Uniform Crime Reporting (UCR) Program Data. All hate crime data are made publicly available due to the enactment of Hate Crime Statistics Act of 1990 (FBI, 2018). We examined any hate crimes as well as hate crimes targeting specific groups (Asians, Blacks, Latinos, Arabs, and Whites) at the state level.

General Social Survey

We used racial prejudice questions from the 2016 and 2018 General Social Survey (GSS). The GSS has tracked trends in attitudes, behaviors, and attributes from American society since 1972 and is a repeated cross-sectional survey of a nationally representative sample of non-institutionalized adults 18 years of age and older (NORC, 2019). We used the following items: 1) "Do Blacks tend to be unintelligent or tend to be intelligent?" 2) "Do blacks tend to be hard working or lazy?" Response options were selected from a 7-level Likert scale. Identical questions were asked in reference to Whites. The GSS also asked respondents, "On the average (Negroes/Blacks/African-Americans) have worse jobs, income, and housing than white people. Do you think these differences are …" A. "Mainly due to discrimination?"; B. "Because most (Negroes/Blacks/African-Americans) have less in-born ability to learn?"; C. "Because most (Negroes/Blacks/African-Americans) don't have the chance for education that it takes to rise out of poverty?"; D. "Because most (Negroes/Blacks/African-Americans) just don't have the motivation or will power to pull themselves up out of poverty?" All response options were yes or no and modeled as individual items.

Project Implicit

Project Implicit represents the largest repository of data on explicit and implicit racial bias with over 3 million tests performed since 2002 (Leitner, Hehman, Ayduk, & Mendoza-Denton, 2016). Respondents are self-selected; anyone can volunteer to take the test online. We used

Project Implicit data from all race (Black-White) tests completed between January 1, 2015 to December 31, 2018. The sample was restricted to respondents residing in the United States with available data on state and both implicit and explicit racial bias measures. We excluded respondents who made errors on >30% of trials or had reaction times <300 ms on >10% of trials in order to exclude responses with low accuracy or high response latencies (Greenwald, Nosek, & Banaji, 2003; Hehman, Calanchini, Flake, & Leitner, 2019).

Implicit racial bias is assessed using the Implicit Association Test (IAT), which measures speed of keyboard associations between images of Black and White faces and positive and negative words. In this way, the IAT measures automatic, or unconscious, racial biases of respondents. Scores range from approximately −2 to approximately 2. Negative scores indicate an anti-White/pro-Black bias, zero indicates no bias, and positive scores indicate a pro-White/anti-Black bias.

Explicit racial bias is assessed by asking respondents to rank their feelings of warmth/coldness toward Black people and White people on an 11-point Likert scale ranging from 0 "extremely cold" to 10 "extremely warm." Hence, the explicit racial bias measure captures attitudes that respondents are willing to self-report. Following previous work, we calculated the difference between the White score and the Black score to create a difference measure ranging from −10 to 10, with negative values representing greater feelings of warmth toward Black individuals, positive values representing greater feelings of warmth toward White individuals, and 0 representing a neutral score (Hehman et al., 2019). Explicit and implicit racial bias scores were standardized so that a one-unit increase represents a one standard deviation increase in the scores.

Covariates

State-level characteristics

All models adjusted for state-level characteristics including percent non-Latino Black, percent Latino, population density (per square mile), southern state indicator (yes/no), and economic disadvantage (standardized factor score (DeVellis, 1991) summarizing the following variables: percent unemployed; percent with some college as the highest level of education, percent with high school diploma as the highest level of education, percent children in poverty, percent single parent households, and median household income), to account for potential confounding by state-level demographic and economic characteristics. Use of this factor score has been previously published.[24] State-level covariates were derived from 2013-2017 5-year estimates from the American Community Survey (U.S. Census Bureau, 2020).

Individual-level covariates-GSS

Individual-level covariates for GSS respondents were included in models examining the association between the Twitter-derived racial sentiment scores and racial attitude questions on the GSS. These included age (years), sex (male, female), race (Non-Latino White, Non-Latino Black, Other), education (years), and family income (continuous). To reduce skewness, family income was square root transformed following previous research (Morey, Gee, Muennig, & Hatzenbuehler, 2018).

Individual-level covariates-Project Implicit

Individual-level covariates for Project Implicit respondents were included in models examining the association between the racial sentiment scores and implicit and explicit racial bias. These included age (years), sex (male, female), race (Non-Latino White, Non-Latino Black, Other), and education (less than high school, high school graduate, and bachelor's degree or higher).

Statistical analyses

Hate crimes

Negative binomial regression models were fitted to estimate incidence rate ratios (IRRs) for the relationship between sentiment scores and hate crimes. State and year specific total population were used as the offset variable in the models. Models controlled for year and the state-level characteristics described above. Standard errors were calculated specifying clustering at the state level. Complete cases analyses were performed. The IRRs can be interpreted as the relative increase in the hate crime rate associated with 10% increase in negative racial sentiment.

GSS

Linear and logistic regression models were fitted for continuous and dichotomous GSS items, respectively. A factor analysis of GSS questions was conducted to examine how well the individual items "hang" together (DeVellis, 1991). The eigenvalues for the first factor was 1.16, and all other factors had eigenvalues less than 1 (Online Supplementary Materials Table S1), indicating no factor explained more of the variance than single observed variable. As a result, we examined the associations between the Twitter-derived sentiment scores and individual GSS items. Models used the study's sampling weight to appropriately account for the GSS sampling design, and standard errors were calculated specifying clustering at the state level. Models controlled for year and individual-level covariates of GSS participants and state-level characteristics in a complete case analysis.

Project Implicit

Linear regression models were fitted for continuous explicit and implicit variables in a complete case analysis. Models controlled for year of the test as well as state-level characteristics and individual-level Project Implicit respondent characteristics.

Sensitivity analyses

The main analyses investigated the association between state-level racial sentiment and individual-level racial attitudes and racial bias. In our data set, all the tweets (N=30,977,757) have location information to identify the state associated with the tweet. For a very small subset of tweets (4.26%, N=1,320,647), geolocation information is available that would permit the identification of the county. In order to use as much of the data as possible, we choose state as the area of aggregation for our main analyses. Sensitivity analyses were performed to examine associations with county-level racial sentiment. To conduct these analyses, we merged county and year specific racial sentiment data with county and year specific data on hate crimes, racial attitudes from the General Social Survey, and explicit and implicit bias from Project Implicit.

Results

Tweets referencing Blacks (45.44%), Whites (44.88%) and Arabs (39.14%) had the highest proportion negative sentiment. Tweets referencing Latinos and Asians had the lowest proportion of negative sentiment (11.66%, and 6.90%), respectively. Proportion of tweets referencing racial/ethnic minorities expressing negative sentiment by state are graphed and presented in the online supplementary Figure 1 and Table S2. Mississippi (46.6%) and Louisiana (46.2%) were the states with the highest proportion of tweets referencing racial/ethnic minorities that were negative and Utah (33.0%) and Hawaii (32.0%) were the states with the lowest proportion of negative race-related tweets. Temporal and spatial resolution for Twitter, General Social Survey, Project Implicit, and hate crimes for years 2015–2018 are presented in online supplementary Table S3.

Hate crimes

There were 111,085 total hate crimes reported in the FBI UCR data program in 2015–2018. Of these, 10,643 hate crimes against racial and ethnic minorities. Hate crimes against Blacks represent the largest minority group targeted with 7,520 hate crimes reported during this period (Table 1). Because of the rarity of hate crimes, confidence intervals for the association of the Twitter-derived bias measure and hate crimes were very wide and in all cases included the null. A 10% increase in the proportion of negative tweets referencing racial or ethnic minorities was associated with higher incidence rate of hate crimes against minorities (IRR: 1.38 (95% CI: 0.66, 2.85) in the state (Table 2). Greater negative sentiment of tweets referencing Whites and negative sentiment referencing Latinos were associated with elevated anti-White hate crimes (IRR: 1.56, 95% CI: 0.68, 3.58) and anti-Latino hate crimes (IRR: 1.50; 95% CI: 0.46, 4.84). Negative sentiment referencing other subgroups (Black, Asians, and Arabs) were also not significantly associated with hate crimes against the respective subgroup (Table 2). Negative sentiment tweets referencing racial minorities as a group was not related to *any* hate crimes (IRR: 1.03; 95% CI: 0.65, 1.65).

GSS

Demographic characteristics of the GSS analytic sample are presented in Table 3. The mean age of GSS participants was 48 years; 64% of the participants were White, and 55% of the study sample were women (Table 3). GSS respondents tended to rate Whites higher than Blacks in terms of work ethic and intelligence. The mean score for the work ethic question for Whites was 4.39 (out of 7) compared to 3.98 for Blacks (β: −0.41, $p < 0.001$). The mean score for intelligence for Whites was 4.60 (out of 7) compared to 4.36 for Blacks (Table 3) (β: −0.24, $p < 0.001$). Approximately 45% and 52% of GSS respondents believed Black-White disparities in jobs, income and housing were due to discrimination and lack of education, respectively. These responses are consistent with identifying systemic causes for persistent disparities. Forty-one percent of respondents also believed disparities were due to lack of will, which is more consistent with identifying individuals as the cause of their circumstances. Only 9% endorsed the belief that differences were due to in-born ability to learn (Table 3).

GSS respondents living in states with a 10% higher negative sentiment for tweets referencing Blacks were less likely to believe that Blacks were hard working (β: −0.22; 95% CI: −0.50, 0.05) compared to GSS respondents living in states with lower negative sentiment for tweets referencing Blacks. GSS respondents living in states with a 10% higher negative sentiment for tweets referencing Blacks also had 0.57 times the odds of endorsing that Black-White disparities in jobs, income, and housing were due to discrimination (95% CI: 0.40, 0.83), and 0.81 times the odds of endorsing that disparities were due to lack of education (95% CI: 0.42, 1.55). In contrast, they had 1.64 times the odds of endorsing the belief that disparities were due to lack to will (95% CI: 0.95, 2.84) (Table 4). Higher negative sentiment for tweets referencing Whites was also associated with lower scores for the belief that Whites were hard working (β=−0.27; 95% CI: −0.46, −0.08 (Table 4).

Table 1
Sentiment scores and hate crimes, 2015–2018.

	N	%
Negative sentiment scores for tweets referencing		
Minorities	29,063,011	40.64%
Blacks	18,562,433	45.44%
Whites	1,914,726	44.88%
Arabs	33,682	39.14%
Latino	1,848,756	11.66%
Asians	2,207,120	6.90%
Hate crime categories		
Total hate crimes	111,085	100%
Anti-Minority	10,643	9.58%
Anti-Black	7,520	6.77%
Anti-White	2,902	2.61%
Anti-Arab	1,036	0.93%
Anti-Latino	1,579	1.42%
Anti-Asian	511	0.46%

Table 2
Association between state-level sentiment and hate crimes occurring in that state (N=200).

	Incidence Rate Ratio	95% CI	
Negative minority sentiment and any hate crime	1.03	0.65	1.65
Negative minority sentiment and hate crimes against minorities	1.38	0.66	2.85
Negative Black sentiment and hate crimes against Blacks	0.99	0.52	1.90
Negative White sentiment and hate crimes against Whites	1.56	0.68	3.58
Negative Arab sentiment and hate crimes against Arabs	1.00	0.88	1.13
Negative Hispanic sentiment and hate crimes against Hispanics	1.50	0.46	4.84
Negative Asian sentiment and hate crimes against Asians	0.28	0.04	1.77

Adjusted negative binomial regression models were run for each outcome separately. Models controlled for year and state-level % non-Latino Black, % Latino, southern state indicator, population density, and economic disadvantage (standardized factor score summarizing the following four variables: percent unemployed; percent with some college, percent with high school diploma, percent children in poverty, percent single parent households, and percent median household income).

Table 3
Demographic characteristics of the GSS analytic sample (N=2,644).

Characteristic	N	%
Age (Mean, SD)	48.46	17.72
Female	1,460	55.22
Non-Hispanic Black	452	17.1
Non-Hispanic White	1,687	63.8
Other	275	10.4
Education (years) (Mean, SD)	13.59	2.84
Racial Attitudes GSS questions		
Hard working (Whites) (Mean, SD)	4.39	1.10
Hard working (Blacks) (Mean, SD)	3.98	1.16
Intelligent (Whites) (Mean, SD)	4.60	1.15
Intelligent (Blacks) (Mean, SD)	4.36	1.05
Black-White Disparities in jobs, income, housing due to:		
Discrimination (Mean, SD)	0.46	0.50
Lack chance for education (Mean, SD)	0.51	0.50
In-born ability (Mean, SD)	0.09	0.28
Lack of will (Mean, SD)	0.42	0.49

Responses for questions related to hard work and intelligence ranged from 1 to 7 with higher scores indicating belief the group is more hardworking or intelligent. Questions related to Black White disparities had 0 (yes) or 1 (no) as response options.

Project Implicit

Demographic characteristics of the Project Implicit analytic sample are presented in Table 5. Study respondents had a mean age of 28 years; 63% were female, and 64% were non-Hispanic White. Project Implicit respondents living in states with a 10% higher negative sentiment for tweets referencing Blacks had higher scores (indicating greater anti-Black bias) on both the standardized explicit racial bias (β: 0.11; 95% CI: 0.04, 0.18) and the standardized implicit racial bias (β: 0.09; 95% CI: 0.04, 0.14) measures (Table 6).

Sensitivity analyses to examine associations between county-level

Table 4
Association of state-level negative sentiment in tweets referencing Blacks with individual-level responses to GSS racial attitude questions for residents in that state (N=2,644).

	Estimate (β or OR)	95% CI	
Response about Black People			
Linear regression			
Hard working	−0.22	−0.50	0.05
Intelligent	−0.08	−0.38	0.22
Logistic Regression			
Discrimination	0.57	0.40	0.83
Lack chance for Education	0.81	0.42	1.55
In-born ability	0.87	0.42	1.80
Lack of will	1.64	0.95	2.84
Response about White People			
Hard working	−0.27	−0.46	−0.08
Intelligent	0.06	−0.36	0.48

Adjusted models were estimated for each outcome separately. Models specified clustering at the state level and controlled for year and state-level % non-Latino Black, % Latino, southern state indicator, population density, and economic disadvantage (standardized factor score summarizing the following variables: percent unemployed; percent with some college, percent with high school diploma, percent children in poverty, percent single parent households, and median household income) as well as individual-level respondent age, sex, race/ethnicity, education, and family income. Models used the GSS's sampling weight to appropriately account for the study's sampling design.

Table 5
Demographic characteristics of the Project Implicit analytic sample (N=867,950).

Characteristic	N	%
Age (Mean, SD)	28.05	12.97
Female	547,584	63.09
Race/Ethnicity		
Non-Hispanic Black	90,016	10.37
Non-Hispanic White	558,070	64.30
Other	219,864	25.33
Education		
<High School	129,752	14.95
High School	376,671	43.40
College degree or greater	361,527	41.65
Explicit racial bias (Mean, SD)	−0.16	1.95
Implicit racial bias (Mean, SD)	0.30	0.43

The Project Implicit explicit racial bias measure had a range of −10 to 10. The implicit racial bias measure had a range of −1.9 to 1.8 out of a possible −2 to 2.

Table 6
Association between state-level negative sentiment for tweets referencing Blacks and individual-level Project Implicit bias measures for residents in that state, 2015–2018 (N=867,950).

	β Estimate	95% CI	
Explicit bias	0.11	0.04	0.18
Implicit bias	0.09	0.04	0.14

Explicit and implicit measures standardized. Adjusted models were estimated for each outcome separately. Models specified clustering at the state level and controlled for year and state-level % non-Latino Black, % Latino, southern state indicator, population density, and economic disadvantage (standardized factor score summarizing the following variables: percent unemployed; percent with some college, percent with high school diploma, percent children in poverty, percent single parent households, and median household income) as well as individual-level respondent age, sex, race/ethnicity, and education.

racial sentiment and individual-level racial attitudes and implicit and explicit racial biases showed attenuation at the county level, but the pattern and direction of the estimates are similar across the two levels of aggregation (online supplemental Tables S4-S5). As seen with state-level racial sentiment, county-level racial sentiment was not consistently associated with county-level hate crimes (Online supplementary Tables S6).

Discussion

Individuals living in states with higher levels of negative racial sentiment assessed from Twitter expressed greater racial bias in the GSS and Project Implicit. These results indicate that racial climates that are less welcoming to minorities are associated with higher racial bias at the individual level. Area-level negative racial sentiment was positively associated with hate crimes targeting racial and ethnic minorities, but these estimates were not statistically significant, possibly due to lower statistical power given the relative rarity of documented hate crimes. Consistent with the concept of cultural racism, commonplace and continuous negative expressions regarding racial and ethnic minorities may create an environment where individual and interpersonal forms of biases and discrimination can flourish (Williams et al., 2019).

Racial bias at both the community (Chae et al., 2015; Morey et al., 2018) and individual-level (Alhusen et al., 2016; Pascoe & Smart Richman, 2009) have been associated with adverse health outcomes. While few studies have examined racial bias at multiple social levels, nascent evidence suggests the two constructs may operate independently on health outcomes. For example, in a multilevel survival analysis, Lee and colleagues found community-level racial bias was associated with mortality independent of individual-level racial bias (Lee, Muennig, Kawachi, & Hatzenbuehler, 2015). Emerging work has also shown that area-level racial bias is associated with worse health outcomes for both racial/ethnic minorities as well as Whites (Leitner et al., 2016; T.; Nguyen et al., 2020; Nguyen et al., 2018). Taken together, the growing body of evidence indicates the potential influence of area-level racial bias on the health of communities. The current work demonstrates the correspondence between different measures of racial bias, confirming that state-level negative racial sentiment as measured with Twitter is predictive of individual-level expressions of racial prejudice in the GSS and in the IAT assessments. Although further research is needed, these findings have potential policy implications. A strong body of evidence documents racial bias within institutions and systems including in the employment, educational and housing sectors (Phelan & Link, 2015; Williams, Lawrence, & Davis, 2019). Hence, areas with more negative racial sentiment may have greater prejudicial racial attitudes that absolve responsibility at the institutional and structural level which may impede policies and other actions aimed at reducing inequities.

Study findings should be interpreted in light of several limitations. Our main analyses examined state-level racial sentiment due to limited data at lower levels of aggregation. Future work can explore heterogeneity of racial sentiment within states. Sensitivity analyses at the county-level revealed attenuation of the estimates observed at the state-level. However, only 4% of the Twitter data had latitude and longitude coordinates that permitted the identification of the county associated with the tweet. Twitter users who enable location information may be different from users who do not enable location information. Twitter data represent what people are willing to express in the online public sphere. Twitter users are not representative of the U.S. population; younger and higher socioeconomic populations are over-represented on Twitter (Pew Research Center, 2018). Hate crime data come from the FBI UCR. Participation in the FBI UCR program is mandated for federal law enforcement agencies (LEAs), but is voluntary for local, state, and tribal LEAs (FBI, 2018). Hate crimes reported in the UCR are an underestimate (Pezzella, Fetzer, & Keller, 2019). For bias incidents to be reflected in the UCR Hate Crime reporting program, several steps need to occur. Victims need to report the hate crime. Police need to record the incident as a potential hate crime and to determine and verify the bias motivation. Finally, police agencies must classify the incident and report it to the UCR Hate Crime Unit (Nolan & Akiyama, 1999). The underreporting of hate crimes may be correlated with racism such that areas with greater racial bias may be less likely to report hate crimes, which would bias estimates toward the null. Hate crimes represent an extreme form of discrimination, which is different from the other measures capturing racial attitudes and beliefs. Project Implicit respondents are self-selected; hence the results are not necessarily generalizable to the U. S. population. However, consistency of results from implicit and explicit bias scores and GSS questions (which are nationally representative) is encouraging. Triangulation of evidence from multiple sources, each with its own strengths and limitations, is more compelling than any individual measure alone (Lawlor, Tilling, & Davey Smith, 2016).

Interpersonal and structural racial bias are leading explanations for the continuing racial inequities across an array of negative health outcomes but research to confirm the role of racism has been hampered by challenges in both measuring racial bias and evaluating its impact. This study demonstrates that Twitter-derived measures of racial sentiment are associated with more traditional individual-level measures of racial bias and discrimination. The detection of associations in the hypothesized direction adds credence to the use of social media measures for the timely and cost-efficient measurement area-level racial climate and may improve understanding of the multilevel ways in which racism impacts health.

Ethics statement

This study was determined exempt by the University of California, San Francisco Institutional Review Board (18–24255).

Author statement

Thu Nguyen: Funding acquisition, Conceptualization, Methodology, Data Curation, Formal Analysis, Writing-Original Draft Preparation, Review, and Editing; **Dina Huang**: Formal Analysis, Review, and Editing; **Eli Michaels**: Data Curation, Review, and Editing; **M. Maria Glymour**: Conceptualization, Review, and Editing; **Amani Allen**: Conceptualization, Review, and Editing; **Quynh Nguyen**: Conceptualization, Review, and Editing.

Declaration of competing interest

Authors declare no conflicts of interest.

Acknowledgements

Funding**:** Research reported in this publication was supported by the National Institute on Minority Health and Health Disparities, the National Library of Medicine, and the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award Number R00MD012615 (Nguyen, T., PI), R01LM012849 (Nguyen, Q., PI), F31HL151284 (Michaels, E., PI). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ssmph.2021.100750.

References

Alhusen, J. L., Bower, K., Epstein, E., & Sharps, P. (2016). Racial discrimination and adverse birth outcomes: An integrative review. *Journal of Midwifery & Women's Health, 61*(6), 707–720. https://doi.org/10.1111/jmwh.12490

An, B. P. (2015). The role of social desirability bias and racial/ethnic composition on the relation between education and attitude toward immigration restrictionism. *The Social Science Journal, 52*(4), 459–467. https://doi.org/10.1016/j.soscij.2014.09.005

Bartlett, J., Reffin, J., Rumball, N., & Williamson, S. (2014). *Anti-social media*. Demos (pp. 1–51).

Chae, D. H., Clouston, S., Hatzenbuehler, M. L., Kramer, M. R., Cooper, H. L., Wilson, S. M., … Link, B. G. (2015). Association between an internet-based measure of area racism and black mortality. *PLoS One, 10*(4), Article e0122963.

Chae, D. H., Clouston, S., Martz, C. D., Hatzenbuehler, M. L., Cooper, H. L. F., Turpin, R., … Kramer, M. R. (2018). Area racism and birth outcomes among Blacks in the United States. *Social Science & Medicine, 199*, 49–55.

DeVellis, R. (1991). *Factor analytic strategies scale development: Theory and applications* (Vol. 26, pp. 91–109). Newbury Park, CA: SAGE.

FBI. (2018). Uniform crime reporting (UCR) program's hate crime frequently asked questions (FAQ) retrieved from. https://ucr.fbi.gov/hate-crime-faqs.

Geronimus, A. T., James, S. A., Destin, M., Graham, L. F., Hatzenbuehler, M. L., Murphy, M. C., … Thompson, J. P. (2016). Jedi public health: Co-creating an identity-safe culture to promote health equity. *SSM - Population Health, 2*, 105–116. https://doi.org/10.1016/j.ssmph.2016.02.008

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*(2), 197.

Hehman, E., Calanchini, J., Flake, J. K., & Leitner, J. B. (2019). Establishing construct validity evidence for regional measures of explicit and implicit racial bias. *Journal of Experimental Psychology: General, 148*(6), 1022–1040.

Huang, D., Huang, Y., Adams, N., Nguyen, T. T., & Nguyen, Q. C. (2020). Twitter-characterized sentiment towards racial/ethnic minorities and cardiovascular disease (CVD) outcomes. *Journal of Racial and Ethnic Health Disparities*. https://doi.org/10.1007/s40615-020-00712-y

Kaggle in Class. (2011). Sentiment classification. https://inclass.kaggle.com/c/si650winter11.

Krieger, N., Smith, K., Naishadham, D., Hartman, C., & Barbeau, E. M. (2005). Experiences of discrimination: Validity and reliability of a self-report measure for population health research on racism and health. *Social Science & Medicine, 61*(7), 1576–1596. https://doi.org/10.1016/j.socscimed.2005.03.006

Lawlor, D. A., Tilling, K., & Davey Smith, G. (2016). Triangulation in aetiological epidemiology. *International Journal of Epidemiology, 45*(6), 1866–1886. https://doi.org/10.1093/ije/dyw314

Lee, Y., Muennig, P., Kawachi, I., & Hatzenbuehler, M. L. (2015). Effects of racial prejudice on the health of communities: A multilevel survival analysis. *American Journal of Public Health, 105*(11), 2349–2355. https://doi.org/10.2105/AJPH.2015.302776

Leitner, J. B., Hehman, E., Ayduk, O., & Mendoza-Denton, R. (2016). Blacks' death rate due to circulatory diseases is positively related to whites' explicit racial bias: A nationwide investigation using Project implicit. *Psychological Science, 27*(10), 1299–1311. https://doi.org/10.1177/0956797616658450

Mislove, A., Lehmann, S., Ahn, Y., Onnela, J. P., & Rosenquist, J. N. (2011). Understanding the demographics of twitter users. In *Paper presented at the proceedings of the fifth international AAAI conference on weblogs and social media* (Barcelona, Spain).

Morey, B. N., Gee, G. C., Muennig, P., & Hatzenbuehler, M. L. (2018). Community-level prejudice and mortality among immigrant groups. *Social Science & Medicine, 199*, 56–66. https://doi.org/10.1016/j.socscimed.2017.04.020

Nguyen, T., Adams, N., Huang, D., Glymour, M. M., Allen, A. M., & Nguyen, Q. C. (2020). The association between state-level racial attitudes assessed from twitter data and adverse birth outcomes: Observational study. *JMIR Public Health and Surveillance, 6*(3), Article e17103. https://doi.org/10.2196/17103

Nguyen, T., Criss, S., Allen, A. M., Glymour, M. M., Phan, L., Trevino, R., et al. (2019). Pride, love, and twitter rants: Combining machine learning and qualitative techniques to understand what our tweets reveal about race in the US. *International Journal of Environmental Research and Public Health, 16*(10), 1766.

Nguyen, T., Meng, H.-W., Sandeep, S., McCullough, M., Yu, W., Lau, Y., … Nguyen, Q. C. (2018). Twitter-derived measures of sentiment towards minorities (2015–2016) and associations with low birth weight and preterm birth in the United States. *Computers in Human Behavior, 89*, 308–315. https://doi.org/10.1016/j.chb.2018.08.010

Nolan, J. J., & Akiyama, Y. (1999). An analysis of factors that affect law enforcement participation in hate crime reporting. *Journal of Contemporary Criminal Justice, 15*(1), 111–127. https://doi.org/10.1177/1043986299015001008

NORC. (2019). The general social survey. https://gss.norc.org/.

Nuru-Jeter, A. M., Michaels, E. K., Thomas, M. D., Reeves, A. N., Thorpe, R. J., Jr., & LaVeist, T. A. (2018). Relative roles of race versus socioeconomic position in studies of health inequalities: A matter of interpretation. *Annual Review of Public Health, 39*, 169–188. https://doi.org/10.1146/annurev-publhealth-040617-014230

Paradies, Y., Ben, J., Denson, N., Elias, A., Priest, N., Pieterse, A., … Gee, G. (2015). Racism as a determinant of health: A systematic review and meta-analysis. *PLoS One, 10*(9), Article e0138511.

Pascoe, E. A., & Smart Richman, L. (2009). Perceived discrimination and health: A meta-analytic review. *Psychological Bulletin, 135*(4), 531–554. https://doi.org/10.1037/a0016059

Pew Research Center. (2016). Social media conversations about race. http://www.pewinternet.org/2016/08/15/social-media-conversations-about-race/.

September 18, 2018 Pew Research Center. (2018). Social media use in 2018 http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/.

Pezzella, F. S., Fetzer, M. D., & Keller, T. (2019). The dark figure of hate crime underreporting. *American Behavioral Scientist.* , Article 0002764218823844. https://doi.org/10.1177/0002764218823844

Phelan, J. C., & Link, B. G. (2015). Is racism a fundamental cause of inequalities in health? *Annual Review of Sociology, 41*(1), 311–330. https://doi.org/10.1146/annurev-soc-073014-112305

Pool, L. R., Ning, H., Lloyd-Jones, D. M., & Allen, N. B. (2017). Trends in racial/ethnic disparities in cardiovascular health among US adults from 1999-2012. *Journal of the American Heart Association, 6*(9).

Sanders Analytics. (2011). Twitter sentiment corpus. http://www.sananalytics.com/lab/twitter-sentiment/.

Sentiment140. (2011). For academics retrieved from. http://help.sentiment140.com/for-students.

Sternthal, M. J., Slopen, N., & Williams, D. R. (2011). Racial disparities in health. *Du Bois Review: Social Science Research on Race, 8*(1), 95–113. https://doi.org/10.1017/S1742058X11000087

Stocké, V. (2007). Determinants and consequences of survey respondents' social desirability beliefs about racial attitudes. *Methodology, 3*(3), 125–138. https://doi.org/10.1027/1614-2241.3.3.125

Suler, J. (2004). The online disinhibition effect. *CyberPsychology and Behavior, 7*, 321–326.

The Racial Slur Database. (2018). http://www.rsdb.org/.

U.S. Census Bureau. (2020). American community survey data retrieved from. https://www.census.gov/programs-surveys/acs/data.html.

Williams, D. R., Lawrence, J. A., & Davis, B. A. (2019). Racism and health: Evidence and needed research. *Annual Review of Public Health, 40*, 105–125. https://doi.org/10.1146/annurev-publhealth-040218-043750

Williams, D. R., Yan, Y., Jackson, J. S., & Anderson, N. B. (1997). Racial differences in physical and mental health: Socio-economic status, stress and discrimination. *Journal of Health Psychology, 2*(3), 335–351. https://doi.org/10.1177/135910539700200305