Psychological Science OnlineFirst, published on January 20, 2015 as doi:10.1177/0956797614557867

Research Article

Psychological Science
1–11

DOI: 10.1177/0956797614557867
pss.sagepub.com

Psychological Language on Twitter Predicts County-Level Heart Disease Mortality

Johannes C. Eichstaedt[1], Hansen Andrew Schwartz[1,2], Margaret L. Kern[1,3], Gregory Park[1], Darwin R. Labarthe[4], Raina M. Merchant[5], Sneha Jha[2], Megha Agrawal[2], Lukasz A. Dziurzynski[1], Maarten Sap[1], Christopher Weeg[1], Emily E. Larson[1], Lyle H. Ungar[1,2], and Martin E. P. Seligman[1]

[1]Department of Psychology, University of Pennsylvania; [2]Department of Computer and Information Science, University of Pennsylvania; [3]Graduate School of Education, University of Melbourne; [4]School of Medicine, Northwestern University; and [5]Department of Emergency Medicine, University of Pennsylvania

Abstract

Hostility and chronic stress are known risk factors for heart disease, but they are costly to assess on a large scale. We used language expressed on Twitter to characterize community-level psychological correlates of age-adjusted mortality from atherosclerotic heart disease (AHD). Language patterns reflecting negative social relationships, disengagement, and negative emotions—especially anger—emerged as risk factors; positive emotions and psychological engagement emerged as protective factors. Most correlations remained significant after controlling for income and education. A cross-sectional regression model based only on Twitter language predicted AHD mortality significantly better than did a model that combined 10 common demographic, socioeconomic, and health risk factors, including smoking, diabetes, hypertension, and obesity. Capturing community psychological characteristics through social media is feasible, and these characteristics are strong markers of cardiovascular mortality at the community level.

Keywords

heart disease, risk factors, well-being, language, big data, emotions, social media, open data, open materials

Received 3/30/14; Revision accepted 10/10/14

Heart disease is the leading cause of death worldwide (World Health Organization, 2011). Identifying and addressing key risk factors, such as smoking, hypertension, obesity, and physical inactivity, have significantly reduced this risk (Ford & Capewell, 2011). Psychological characteristics, such as depression (Lett et al., 2004) and chronic stress (Menezes, Lavie, Milani, O'Keefe, & Lavie, 2011), have similarly been shown to increase risk through physiological effects (e.g., chronic sympathetic arousal) and deleterious health behaviors (e.g., drinking and smoking). Conversely, positive psychological characteristics, such as optimism (Boehm & Kubzansky, 2012) and social support (Tay, Tan, Diener, & Gonzalez, 2013), seem to decrease risk, most likely through similar pathways.

In its 2020 Strategic Impact Goal Statement, the American Heart Association suggested that to further reduce the risk for heart disease, "population-level strategies are essential to shift the entire distribution of risk" (Lloyd-Jones et al., 2010, p. 589). Like individuals, communities have characteristics, such as norms, social connectedness, perceived safety, and environmental stress, that

Corresponding Authors:
Johannes C. Eichstaedt, Department of Psychology, University of Pennsylvania, 3701 Market St., Ste. 220, Philadelphia, PA 19104
E-mail: johannes.penn@gmail.com, jeich@sas.upenn.edu

Hansen Andrew Schwartz, Department of Psychology, University of Pennsylvania, 3701 Market St., Ste. 219, Philadelphia, PA 19104
E-mail: andy.schwartz@gmail.com

contribute to health and disease (Cohen, Farley, & Mason, 2003). One challenge to addressing community-level psychological characteristics is the difficulty of assessment; traditional approaches that use phone surveys and household visits are costly and have limited spatial and temporal precision (Auchincloss, Gebreab, Mair, & Diez Roux, 2012; Chaix, Merlo, Evans, Leal, & Havard, 2009).

Rich information about the psychological states and behaviors of communities is now available in big social-media data, offering a flexible and significantly cheaper alternative for assessing community-level psychological characteristics. Social-media-based digital epidemiology can support faster response and deeper understanding of public-health threats than can traditional methods. For example, Google has used search queries to measure trends in influenza, providing earlier indication of disease spread than the Centers for Disease Control and Prevention (CDC; Ginsberg et al., 2009). Other studies have used Twitter to track Lyme disease, H1N1 influenza, depression, and other common ailments (Chew & Eysenbach, 2010; De Choudhury, Counts, & Horvitz, 2013; de Quincey & Kostkova, 2009; Paul & Dredze, 2011a, 2011b; Salathé, Freifeld, Mekaru, Tomasulo, & Brownstein, 2013; Seifter, Schwarzwalder, Geis, & Aucott, 2010; St Louis & Zorlu, 2012).

Methods for inferring psychological states through language analysis have a rich history (Pennebaker, Mehl, & Niederhoffer, 2003; Stone, Dunphy, Smith, & Ogilvie, 1966). Traditional approaches use dictionaries—predetermined lists of words—associated with different constructs (e.g., *sad, glum*, and *crying* are part of a negative-emotion dictionary; Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007). *Open-vocabulary* approaches identify predictive words statistically and are not based on traditional predetermined *dictionaries* (Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, et al., 2013), offering a complementary method of language analysis.

In this study, we analyzed social-media language to identify community-level psychological characteristics associated with mortality from atherosclerotic heart disease (AHD). Working with a data set of 10s of millions of Twitter messages (tweets), we used dictionary-based and open-vocabulary analyses to characterize the psychological language correlates of AHD mortality. We also gauged the amount of AHD-relevant information in Twitter language by building and evaluating predictive models of AHD mortality, and we compared the language models with traditional models that used demographic and socioeconomic risk factors.

Method

We collected tweets from across the United States, determined their counties of origin, and derived values for language variables (e.g., the relative frequencies with which people expressed anger or engagement) for each county. We correlated these county-level language measures with county-level age-adjusted AHD mortality rates obtained from the CDC. To gauge the amount of information relevant to heart disease contained in the Twitter language, we compared the performance of prediction models that used Twitter language with the performance of models that contained county-level (a) measures of socioeconomic status (SES; i.e., income and education), (b) demographics (percentages of Black, Hispanic, married, and female residents), and (c) health variables (incidence of diabetes, obesity, smoking, and hypertension). All procedures were approved by the University of Pennsylvania Institutional Review Board.

Data sources

We used data from 1,347 U.S. counties for which AHD mortality rates; county-level socioeconomic, demographic, and health variables; and at least 50,000 tweeted words were available. More than 88% of the U.S. population lives in the included counties (U.S. Census Bureau, 2010).[1]

Twitter data. Tweets are brief messages (no more than 140 characters) containing information about emotions, thoughts, behaviors, and other personally salient information. In 2009 and 2010, Twitter made a 10% random sample of tweets (the "Garden Hose") available for researchers through direct access to its servers. We obtained a sample of 826 million tweets collected between June 2009 and March 2010. Many Twitter users self-reported their locations in their user profiles, and we used this information to map tweets to counties (for details, see the Mapping Tweets to Counties section of the Supplemental Method in the Supplemental Material available online). This resulted in 148 million county-mapped tweets across 1,347 counties.

Heart disease data. Counties are the smallest socio-ecological level for which most CDC health variables and U.S. Census information are available. From the Centers for Disease Control and Prevention (2010b) we obtained county-level age-adjusted mortality rates for AHD, which is represented by code I25.1 in the International Classification of Disease, 10th edition (ICD 10; World Health Organization, 1992). This code has the highest overall mortality rate in the United States (prevalence = 51.5 deaths per 100,000 in 2010). We averaged AHD mortality rates across 2009 and 2010 to match the time period of the Twitter-language data set.

Demographic and health risk factors. We obtained county-level median income and the percentage of married residents from the American Community Survey (U.S. Census Bureau, 2009). We also obtained high school and college graduation rates from this survey, which we used to create an index of educational attainment. We obtained percentages of female, Black, and Hispanic residents from the U.S. Census Bureau (2010). From the Behavioral Risk Factor Surveillance System of the Centers for Disease Control and Prevention (2009, 2010a) we obtained prevalence of self-reported diabetes, obesity, smoking, and hypertension (common cardiovascular risk factors) for which county-level estimates had previously been derived (see Table S1 in the Supplemental Tables of the Supplemental Material for detailed source information).

Analytic procedure

Language variables from Twitter. We used an automatic process to extract the relative frequency of words and phrases (sequences of two to three words) for every county. For example, the relative frequency of the word *hate* ranged from 0.009% to 0.139% across counties (see the Tokenization section of the Supplemental Method).

We then derived two more types of language-use variables from counties' relative word-usage frequencies: variables based on (a) dictionaries and (b) topics. Dictionary-based variables were relative frequencies of psychologically related words from predetermined dictionaries (e.g., positive-emotion words accounted for 4.6% of all words in a county on average). Topic-based variables were the relative usage of 2,000 automatically created topics, which are clusters of semantically related words that can be thought of as latent factors (words can have loadings on multiple topics; see the Topic Extraction section of the Supplemental Method).

We used preestablished dictionaries for anger, anxiety, positive and negative emotions, positive and negative social relationships, and engagement and disengagement (Pennebaker et al., 2007; Schwartz, Eichstaedt, Kern, Dziurzynski, Lucas, et al., 2013). Topics had previously been automatically derived (Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, et al., 2013).

Because words can have multiple senses, act as multiple parts of speech, and be used in the context of irony or negation, it is important to gauge empirically how well such lists of words measure what is intended (Grimmer & Stewart, 2013). To that end, we had human raters evaluate the dictionaries to determine whether each accurately measured the psychological concept intended. For each of the eight dictionaries, two independent raters examined 200 tweets containing dictionary words and rated whether each dictionary word in the tweets expressed the associated dictionary concept. A third rater was brought in to break ties. Judges rated the dictionaries to have accuracy levels between 55% and 89% (see Table S2 in the Supplemental Tables).[2]

Statistical analysis. Dictionary and topic language variables were correlated with county AHD mortality rates using ordinary least squares linear regression. Each language variable was entered individually into the regression equation and then entered simultaneously with education and income as controls. We tested 2,000 topics, so we applied the Bonferroni correction to the significance threshold (i.e., for the correlation of 1 of 2,000 topics to be significant, its p value would have to be less than .05/2,000, or .000025).

Predictive models. A predictive model of county AHD mortality rates was created using all of the Twitter language variables. That is, we created a single model in which all of the word, phrase, dictionary, and topic frequencies were independent variables and the AHD mortality rate was the dependent variable. We used regularized linear regression (ridge regression) to fit the model (see the Predictive Models section of the Supplemental Method). We also created predictive models of county AHD mortality rates in which the predictors were different combinations of sets of variables: Twitter language, county demographics (percentages of Black, Hispanic, married, and female residents), and socioeconomic (income, education) and health (incidence of diabetes, obesity, smoking, and hypertension) variables.

We avoided distorted results (due to model overfitting—picking up patterns simply by chance) by using a 10-fold cross-validation process that compared model predictions with out-of-sample data. For this analysis, the counties were first randomly partitioned into 10 parts (*folds*). Then, a predictive model was created by fitting the independent variables to the dependent variable (AHD mortality) over 9 of the 10 folds of counties (the *training set*). We then evaluated how well the resulting model predicted the outcomes for the remaining fold (one 10th of the counties; the *hold-out set*). We evaluated the model by comparing its predicted rates with the actual CDC-reported mortality rates using a Pearson product-moment correlation. This procedure was repeated 10 times, allowing each fold to be the hold-out set. The results were averaged together to determine overall prediction performance across all counties for a given model.

To compare predictive performance between two models (e.g., a model based only on Twitter language versus a model based on income and education), we conducted paired t tests comparing the sizes of the standardized residuals of county-level predictions from the models.

Table 1. County-Level Correlations Between Atherosclerotic Heart Disease (AHD) Mortality and Twitter Language Measured by Dictionaries

Language variable	Correlation with AHD mortality
Risk factors	
Anger	.17 [.11, .22]**
Negative relationships	.16 [.11, .21]**
Negative emotions	.10 [.05, .16]**
Disengagement	.14 [.08, .19]**
Anxiety	.05 [.00, .11]†
Protective factors	
Positive relationships[a]	.02 [−.04, .07]
Positive emotions	−.11 [−.17, −.06]**
Engagement	−.16 [−.21, −.10]**

Note: The table presents Pearson *r*s, with 95% confidence intervals in square brackets (n = 1,347 counties). The anger and anxiety dictionaries come from the Linguistic Inquiry and Word Count software (Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007); the other dictionaries are our own (Schwartz, Eichstaedt, Kern, Dziurzynski, Lucas, et al., 2013). Positive correlations indicate that higher values for the language variables are associated with greater AHD mortality.
[a]This is the correlation without *love* included in the dictionary. See note 3 at the end of the article and the discussion for more information.
†$p < .10$. **$p < .001$.

Results

Dictionaries

Greater usage of anger, negative-relationship, negative-emotion, and disengagement words was significantly correlated with greater age-adjusted AHD mortality (*r*s = .10–.17; for specific results, including confidence intervals, see Table 1). After controlling for SES (income and education), all five negative language factors (including usage of anxiety words) were significant risk factors for AHD mortality (partial *r*s = .06, 95% confidence interval, or CI = [.00, .11], to .12, 95% CI = [.07, .17]). This suggests that Twitter language captures information not accounted for by SES. Greater use of positive-emotion and engagement words was associated with lower AHD mortality (r = −.11 and r = −.16, respectively). Use of engagement words remained significantly protective after controlling for SES (partial r = −.09, 95% CI = [−.14, −.04]), but use of positive-emotion words became only marginally significant (partial r = −.05, 95% CI = [−.00, −.11]). Usage of positive-relationships words[3] showed a nonsignificant association with AHD mortality (r = .02, 95% CI = [−.04, .07]; see Table 1).

Topics

We complemented the dictionaries with an open-vocabulary approach, using automatically created topics consisting of semantically coherent groups of words. For each county, we calculated the relative use of each topic, and we correlated topic use with AHD. Figure 1 shows topic composition and correlations for 18 topics whose use was significantly correlated with AHD mortality.[4] The risk factors we observed were themes of hostility and aggression (*shit, asshole, fucking;* *r*s = .18, 95% CI = [.12, .23], to .27, 95% CI = [.22, .32]), hate and interpersonal tension (*jealous, drama, hate;* *r*s = .16, 95% CI = [.11, .21], to .21, 95% CI = [.16, .26]), and boredom and fatigue (*bored, tired, bed;* *r*s = .18, 95% CI = [.12, .23], to .20, 95% CI = [.15, .25]). After controlling for SES, use of seven of the nine risk topics remained significantly correlated with AHD mortality at Bonferroni-corrected levels (partial *r*s = .12, 95% CI = [.07, .17], to .25, 95% CI = [.20, .30], $p < 7 \times 10^{-6}$).

Other topics were protective factors (Fig. 1, bottom panel). Use of topics related to positive experiences (*wonderful, friends, great;* *r*s = −.14, 95% CI = [−.19, −.08], to −.15, 95% CI = [−.21, −.10]) was associated with lower AHD mortality, a finding that mirrors the dictionary-based results. Also associated with lower AHD mortality was use of topics reflecting skilled occupations (*service, skills, conference;* *r*s = −.14, 95% CI = [−.20, −.09], to −.17, 95% CI = [−.22, −.12]) and topics reflecting optimism (*opportunities, goals, overcome;* *r*s = −.12, 95% CI = [−.18, −.07], to −.13, 95% CI = [−.18, −.07]), which has been found to be robustly associated with reduced cardiovascular disease risk at the individual level (Boehm & Kubzansky, 2012; Chida & Steptoe, 2008). After controlling for SES, the correlations between protective topics and AHD mortality remained significant at the traditional .05 level but were no longer significant at Bonferroni-corrected levels.

Prediction

In Figure 2, we compare the predictions of AHD mortality from regression models with different independent variables. Predictive performance was slightly but significantly better for a model combining Twitter and the 10 traditional demographic, SES, and health predictors than for a model that included only the 10 traditional predictors (Twitter plus 10 traditional factors: r = .42, 95% CI = [.38, .46]; 10 traditional factors only: r = .36, 95% CI = [.29, .43]), $t(1346) = -2.22$, p = .026. This suggests that Twitter has incremental predictive validity over and above traditional risk factors. A predictive model using only Twitter language (r = .42, 95% CI = [.38, .45]) performed slightly better than a model using the 10 traditional factors, $t(1346) = -1.97$, p = .049.

To explore these associations in greater detail, we compared the performance of prediction models containing stepwise combinations of Twitter and sets of demographic predictors (percentages of Black, Hispanic, married, and female residents), socioeconomic predictors (income and education), and health predictors (incidence

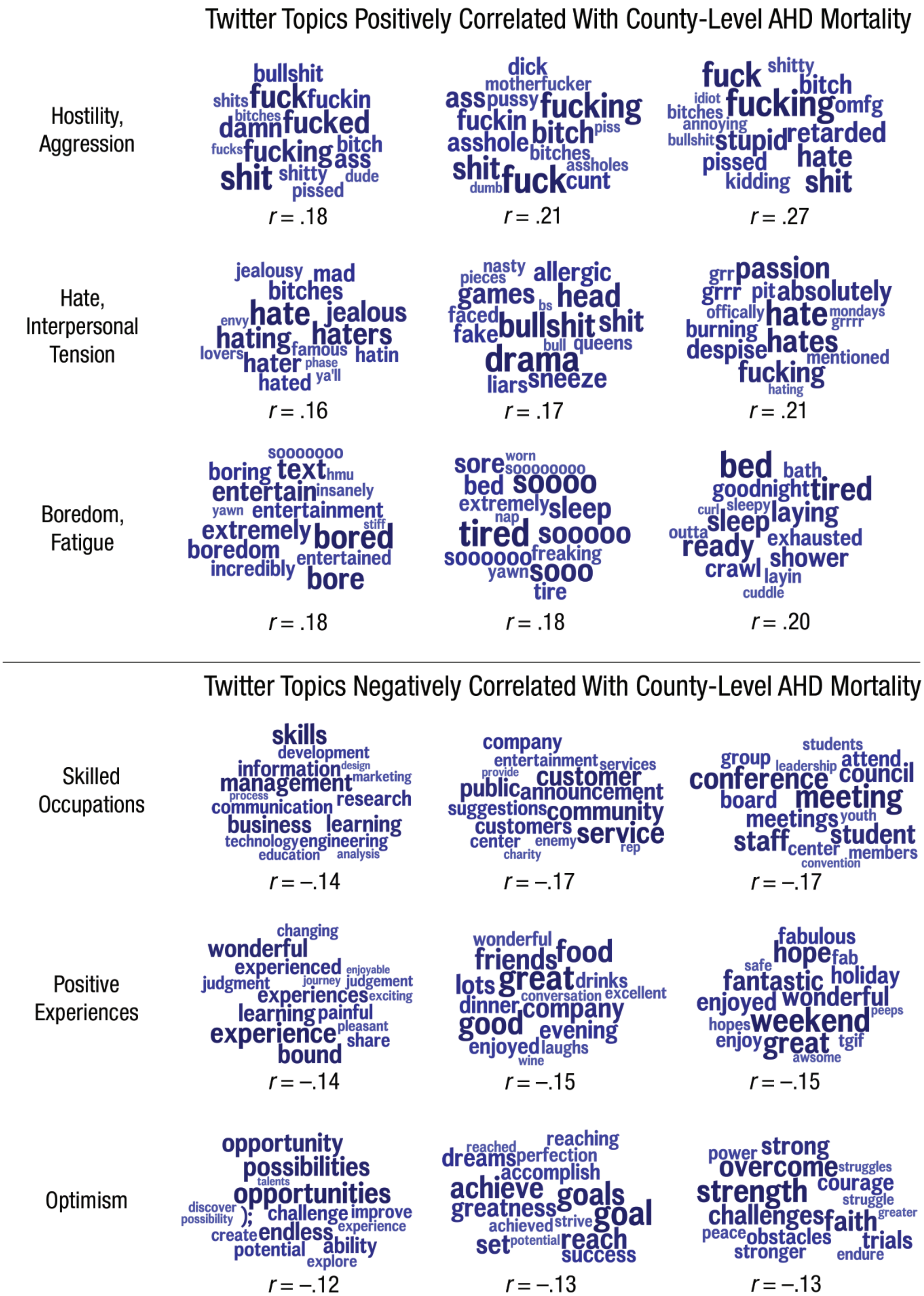

Fig. 1. Twitter topics most correlated with age-adjusted mortality from atherosclerotic heart disease (AHD; significant at a Bonferroni-corrected significance level of $p < 2.5 \times 10^{-5}$). The topics with positive correlations (top) and the topics with negative correlations (bottom) have each been grouped into sets, which are labeled at the left. The size of the word represents its prevalence relative to all words within a given topic (larger = more frequent; for details, see the Supplemental Method).

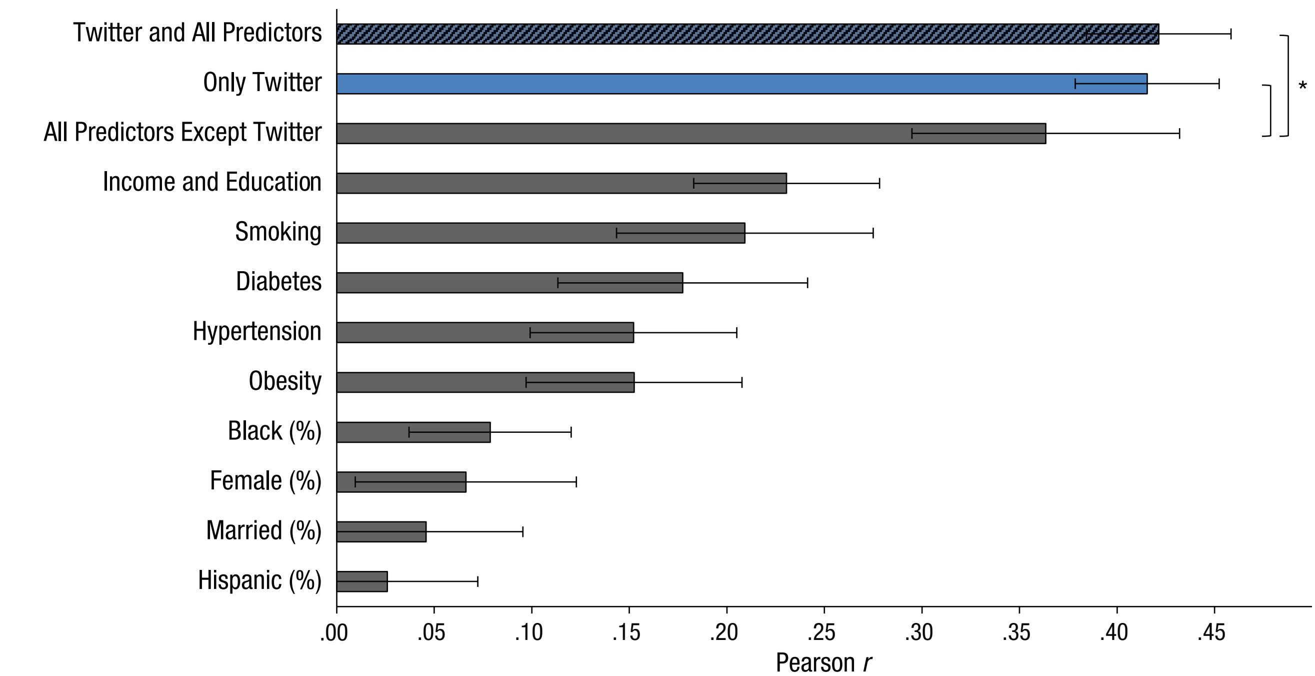

Fig. 2. Performance of models predicting age-adjusted mortality from atherosclerotic heart disease (AHD). For each model, the graph shows the correlation between predicted mortality and actual mortality reported by the Centers for Disease Control and Prevention. Predictions were based on Twitter language, socioeconomic status, health, and demographic variables singly and in combination. Higher values mean better prediction. The correlation values are averages obtained in a cross-validation process used to avoid distortion of accuracy due to chance (overfitting; for details, see the text). Error bars show 95% confidence intervals. Asterisks indicate significant differences between models (*$p < .05$).

of diabetes, obesity, smoking, and hypertension; see Table S4 in the Supplemental Tables). For all combinations of sets of traditional predictors, adding Twitter language significantly improved predictive performance, $t(1346) > 3.00$, $p < .001$. Adding traditional sets of predictors to Twitter language did not significantly improve predictive performance.

Taken together, these results suggest that the AHD-relevant variance in the 10 predictors overlaps with the AHD-relevant variance in the Twitter language features. Twitter language may therefore be a marker for these variables and in addition may have incremental predictive validity. Figure 3 shows CDC-reported AHD mortality averaged across 2009 and 2010 and Twitter-predicted mortality for the densely populated counties in the northeastern United States; a high degree of agreement is evident.

Discussion

Our study had three major findings. First, language expressed on Twitter revealed several community-level psychological characteristics that were significantly associated with heart-disease mortality risk. Second, use of negative-emotion (especially anger), disengagement, and negative-relationship language was associated with increased risk, whereas positive-emotion and engagement language was protective. Third, our predictive results suggest that the information contained in Twitter language fully accounts for—and adds to—the AHD-relevant information in 10 representatively assessed demographic, socioeconomic, and health variables. Taken together, our results suggest that language on Twitter can provide plausible indicators of community-level psychosocial health that may complement other methods of studying the impact of place on health used in epidemiology (cf. Auchincloss et al., 2012) and that these indicators are associated with risk for cardiovascular mortality.

Our findings point to a community-level psychological risk profile similar to risk profiles that have been observed at the individual level. County-level associations between AHD mortality and use of negative-emotion words (relative risk,[5] or RR, = 1.22), anger words (RR = 1.41), and anxiety words (RR = 1.11) were comparable to individual-level meta-analytic effect sizes for the association between AHD mortality and depressed mood (RR = 1.49; Rugulies, 2002), anger (RR = 1.22; Chida & Steptoe, 2009), and anxiety (RR = 1.48; Roest, Martens, de Jonge, & Denollet, 2010).

Although less is known at the individual level about the protective effects of positive psychological variables than about the risk associated with negative variables, our findings align with a growing body of research supporting the

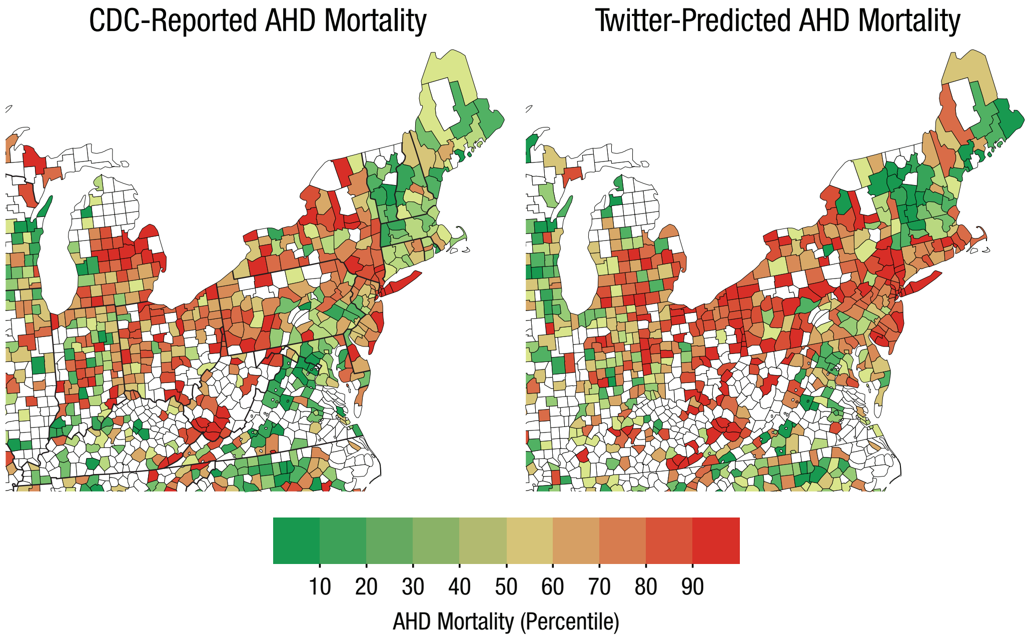

Fig. 3. Map of counties in the northeastern United States showing age-adjusted mortality from atherosclerotic heart disease (AHD) as reported by the Centers for Disease Control and Prevention (CDC; left) and as estimated through the Twitter-language-only prediction model (right). The out-of-sample predictions shown were obtained from the cross-validation process described in the text. Counties for which reliable CDC or Twitter language data were unavailable are shown in white.

cardiovascular health benefits of psychological well-being (Boehm & Kubzansky, in press). Engagement, which has long been considered an important component of successful aging (Rowe & Kahn, 1987), emerged as the strongest protective factor in our study. Use of positive-emotion words was also protective, which is in line with numerous findings that positive emotions convey protection from illness and disease (e.g., Howell, Kern, & Lyubomirsky, 2007; Pressman & Cohen, 2005). Fredrickson, Mancuso, Branigan, and Tugade (2000) have argued that positive emotions may undo the negative cardiovascular aftereffects of anxiety-induced cardiovascular reactivity. Optimism has been shown to have relatively robust association with reduced risk of cardiovascular events at the individual level (Boehm & Kubzansky, 2012; Chida & Steptoe, 2008). We did not have a predefined optimism dictionary, but our topic analyses seem to have identified this as a protective factor (as indicated by results for topics containing *opportunities, goals, overcome*; Fig. 1, bottom). This demonstrates the value of data-driven language analyses.

Overall, our topic findings were similar to and converged with our theory-based dictionary results (cross-correlations are given in Table S3 in the Supplemental Tables). Although theory-based analyses can be more easily tied to existing literature, topic analyses provide a richer portrait of specific behaviors and attitudes (e.g., cursing, frustration, being tired) that correspond to broad psychological characteristics (e.g., anger or stress) associated with an increased risk for AHD mortality. Data-driven analyses, such as our topic analyses, may help identify novel psychological, social, and behavioral correlates of disease.

When analyses use theory-based dictionaries, results can be driven by a few frequent but ambiguous words. For example, greater use of words in the original positive-relationships dictionary (Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, et al., 2013) was surprisingly associated with increased risk, as was the use of its most frequent word, *love*. *Love* accounted for more than a third of the total usage of the positive-relationships dictionary (5.3 million occurrences of *love* compared with 15.0 million occurrences of all words in the dictionary), which means that *love* drove the results for this dictionary. Reading through a random sample of tweets containing *love* revealed them to be mostly statements about loving things, not people.[6] Excluding *love* from the dictionary reduced the correlation between use of the words in the positive-relationships dictionary and AHD mortality (r = .08, 95% CI = [.03, .13]) to nonsignificance (r = .02, 95% CI = [−.04, .07]).

These results demonstrate the pitfalls of interpreting dictionary-based results at face value and underscore the importance of interpreting such results in light of the most frequent words contained in the dictionaries, which can drive the overall dictionary results in unexpected ways. For transparency, in Table S6 in the Supplemental

Tables, we have provided the correlations with AHD mortality for the 10 most frequently used words in each of the eight dictionaries. These findings also highlight the value of triangulating language analyses across different levels of analysis (words, topics, and dictionaries) for arriving at more robust interpretations.

Given that the typical Twitter user is younger (median age = 31 years; Fox, Zickurh, & Smith, 2009) than the typical person at risk for AHD, it is not obvious why Twitter language should track heart-disease mortality. The people tweeting are not the people dying. However, the tweets of younger adults may disclose characteristics of their community, reflecting a shared economic, physical, and psychological environment. At the individual level, psychological variables and heart-disease risk are connected through multiple pathways, including health behaviors, social relationships, situation selection, and physiological reactivity (Friedman & Kern, 2014). These pathways occur within a broader social context that directly and indirectly influences an individual's life experiences. Local communities create physical and social environments that influence the behaviors, stress experiences, and health of their residents (Diez Roux & Mair, 2010; Lochner, Kawachi, Brennan, & Buka, 2003). Epidemiological studies have found that the aggregated characteristics of communities, such as social cohesion and social capital, account for a significant portion of variation in health outcomes, independently of individual-level characteristics (Leyland, 2005; Riva, Gauvin, & Barnett, 2007), such that the combined psychological character of the community is more informative for predicting risk than are the self-reports of any one individual. The language of Twitter may be a window into the aggregated and powerful effects of the community context.

Our study has several limitations. Tweets constitute a biased sample in two ways. First, they may reflect social-desirability biases, because people manage their online identities (Rost, Barkhuus, Cramer, & Brown, 2013). Second, Twitter users are not representative of the general population. The Twitter population tends to be more urban and to have higher levels of education (Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2011). In 2009, the median age of Twitter users (Fox et al., 2009) was 5.8 years below the U.S. median age (U.S. Census Bureau, 2010). Nonetheless, our Twitter-based prediction model outperformed models based on classical risk factors in predicting AHD mortality; this suggests that, despite the biases, Twitter language captures as much unbiased AHD-relevant information about the general population as do traditional, representatively assessed predictors.

Another limitation is that our findings are cross-sectional; future research should address the stability of psychological characteristics of counties across time. Also, we relied on AHD mortality rates reported by the CDC, which draws on the underlying cause of death recorded on death certificates; however, the coding on death certificates may be inconsistent (Pierce & Denison, 2010). Finally, associations between language and mortality do not point to causality; analyses of language on social media may complement other epidemiological methods, but the limits of causal inferences from observational studies have been repeatedly noted (e.g., Diez Roux & Mair, 2010).

Traditional approaches for collecting psychosocial data from large representative samples, such as the Behavioral Risk Factor Surveillance System of the CDC and Gallup polls, tend to be expensive, are based on only thousands of people, and are often limited to a minimal, predefined list of psychological constructs. A Twitter-based system to track psychosocial variables is relatively inexpensive and can potentially generate estimates based on 10s of millions of people with much higher resolution in time and space. It is comparatively easy to create dictionaries automatically for different psychological or social constructs so that novel hypotheses can be tested. Our approach opens the door to a new generation of psychological informational epidemiology (Eysenbach, 2009; Labarthe, 2010) and could bring researchers closer to understanding the community-level psychological factors that are important for the cardiovascular health of communities and should become the focus of intervention.

Author Contributions

J. C. Eichstaedt led the project. J. C. Eichstaedt and H. A. Schwartz conceived of the study. H. A. Schwartz, J. C. Eichstaedt, G. Park, S. Jha, M. Agrawal, L. A. Dziurzynski, and M. Sap handled data acquisition and processing, development of the prediction models, and data analyses. J. C. Eichstaedt, M. L. Kern, H. A. Schwartz, and G. Park drafted the manuscript. D. R. Labarthe, R. M. Merchant, L. H. Ungar, and M. E. P. Seligman provided critical revisions. C. Weeg and E. E. Larson helped acquire, process, and analyze county-level information. All authors approved the final version of the manuscript for submission. L. H. Ungar and M. E. P. Seligman contributed equally to this article.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This work was supported by the Robert Wood Johnson Foundation's Pioneer Portfolio, through Exploring Concepts of Positive Health Grant 63597 (to M. E. P. Seligman), and by a grant from the Templeton Religion Trust.

Supplemental Material

Additional supporting information can be found at http://pss.sagepub.com/content/by/supplemental-data

Open Practices

All data and materials have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/rt6w2/. The complete Open Practices Disclosure for this article can be found at http://pss.sagepub.com/content/by/supplemental-data. This article has received badges for Open Data and Open Materials. More information about the Open Practices badges can be found at https://osf.io/tvyxz/wiki/view/ and http://pss.sagepub.com/content/25/1/3.full.

Notes

1. Analyses using the available heart disease, demographic, and socioeconomic information for the excluded counties revealed that, compared with the counties in the final sample, the excluded counties had smaller populations (median county population of 12,932 in 1,796 excluded counties vs. 78,265 in included counties), higher rates of AHD (Hedges's $g = 0.48$, 95% confidence interval, or CI = [0.38, 0.57]; $n = 597$ excluded counties with data available), lower income ($g = -0.42$, 95% CI = [−0.53, −0.32]; $n = 496$), and lower levels of education ($g = -0.61$, 95% CI = [−.72, −.51]; $n = 496$). The included and excluded counties did not differ in median age ($g = 0.003$, 95% CI = [−0.08, 0.08]; $n = 1{,}004$).
2. The anxiety and positive-relationships dictionaries were rated as having the lowest accuracies (55.0% and 55.5% respectively; see Table S2 in the Supplemental Tables), whereas the accuracy of the other dictionaries was markedly higher (average accuracy = 82.1%). Cross-correlations of dictionaries (see Table S3 in the Supplemental Tables) revealed that the frequency of use of the positive-relationships and anxiety dictionaries were unexpectedly positively correlated with the frequencies of use of all other dictionaries.
3. The word *love* was removed from the dictionary because it accounted for more than a third of the occurrences of words from this dictionary, and including it distorted the results (see Discussion, and note 6).
4. For ease of interpretation, we have grouped these topics into seemingly related sets and added labels to summarize our sense of the topics. These labels are open to interpretation, and we present for inspection the most prevalent words within the topics. County-level topic- and dictionary-frequency data can be downloaded from https://osf.io/rt6w2/files/.
5. To compare our findings with published effect sizes, we converted correlation coefficients to relative risk values following the method of Rosenthal and DiMatteo (2001).
6. In addition to having this word-sense ambiguity, mentions of *love* may signify a different kind of Twitter use in lower-SES areas. A factor analysis of the words in the positive-relationships dictionary revealed two factors with opposing correlations with SES. A general social factor (*friends, agree, loved*) correlated with higher SES ($r = .14$), and a partnership factor (*relationship, boyfriend, girlfriend*) correlated with lower SES ($r = -.43$), as well as higher AHD mortality ($r = .18$). Usage of the word *love* loaded much higher on this second factor than on the first one (see Table S5 in the Supplemental Tables). This finding may be an indication that in lower-SES areas, users share more about personal relationships on Twitter, which distorts the results obtained when using the original positive-relationships dictionary.

References

Auchincloss, A. H., Gebreab, S. Y., Mair, C., & Diez Roux, A. V. (2012). A review of spatial methods in epidemiology, 2000–2010. *Annual Review of Public Health*, *33*, 107–122. doi:10.1146/annurev-publhealth-031811-124655

Boehm, J. K., & Kubzansky, L. D. (2012). The heart's content: The association between positive psychological well-being and cardiovascular health. *Psychological Bulletin*, *138*, 655–691. doi:10.1037/a0027448

Boehm, J. K., & Kubzansky, L. D. (in press). Positive psychological well-being and cardiovascular disease. In W. Kop, L. Katzel, & S. Waldstein (Eds.), *Cardiovascular behavioral medicine*. New York, NY: Springer.

Centers for Disease Control and Prevention. (2009). *Behavioral Risk Factor Surveillance System annual survey data*. Retrieved from http://www.cdc.gov/brfss/annual_data/annual_2009.htm

Centers for Disease Control and Prevention. (2010a). *Behavioral Risk Factor Surveillance System annual survey data*. Retrieved from http://www.cdc.gov/brfss/annual_data/annual_2010.htm

Centers for Disease Control and Prevention. (2010b). *Underlying cause of death 1999-2010. CDC WONDER online database* [Data set]. Retrieved from http://wonder.cdc.gov/ucd-icd10.html

Chaix, B., Merlo, J., Evans, D., Leal, C., & Havard, S. (2009). Neighbourhoods in eco-epidemiologic research: Delimiting personal exposure areas. A response to Riva, Gauvin, Apparicio and Brodeur. *Social Science & Medicine*, *69*, 1306–1310. doi:10.1016/j.socscimed.2009.07.018

Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak. *PLoS ONE*, *5*(11), Article e14118. Retrieved from http://www.plosone.org/article/info:doi/10.1371/journal.pone.0014118

Chida, Y., & Steptoe, A. (2008). Positive psychological well-being and mortality: A quantitative review of prospective observational studies. *Psychosomatic Medicine*, *70*, 741–756. doi:10.1097/PSY.0b013e31818105ba

Chida, Y., & Steptoe, A. (2009). The association of anger and hostility with future coronary heart disease: A meta-analytic view of prospective evidence. *Journal of the American College of Cardiology*, *53*, 936–946. doi:10.1016/j.jacc.2008.11.044

Cohen, D. A., Farley, T. A., & Mason, K. (2003). Why is poverty unhealthy? Social and physical mediators. *Social Science & Medicine*, *57*, 1631–1641. doi:10.1016/S0277-9536(03)00015-7

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Social media as a measurement tool of depression in populations. In *WebSci '13: Proceedings of the 5th Annual ACM Web Science Conference* (pp. 47–56). doi:10.1145/2464464.2464480

de Quincey, E., & Kostkova, P. (2009). Early warning and outbreak detection using social networking websites:

The potential of Twitter. In P. Kostkova (Ed.), *Electronic Healthcare: Second International ICST Conference, eHealth 2009, Istanbul, Turkey*, September 23–25, 2009 (pp. 21–24). doi:10.1007/978-3-642-11745-9_4

Diez Roux, A. V., & Mair, C. (2010). Neighborhoods and health. *Annals of the New York Academy of Sciences*, *1186*, 125–145. doi:10.1111/j.1749-6632.2009.05333.x

Eysenbach, G. (2009). Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *Journal of Medical Internet Research*, *11*(1), Article e11. Retrieved from http://www.jmir.org/2009/1/e11/.

Ford, E. S., & Capewell, S. (2011). Proportion of the decline in cardiovascular mortality disease due to prevention versus treatments: Public health versus clinical care. *Annual Review of Public Health*, *32*, 5–22. doi:10.1146/annurev-publhealth-031210-101211

Fox, S., Zickurh, K., & Smith, A. (2009). *Twitter and status updating, fall 2009*. Retrieved from Pew Research Internet Project Web site: http://www.pewinternet.org/2009/10/21/twitter-and-status-updating-fall-2009

Fredrickson, B. L., Mancuso, R. A., Branigan, C., & Tugade, M. M. (2000). The undoing effects of positive emotions. *Motivation and Emotion*, *24*, 237–258. doi:10.1023/A:1010796329158

Friedman, H. S., & Kern, M. L. (2014). Personality, well-being, and health. *Annual Review of Psychology*, *65*, 719–742. doi:10.1146/annurev-psych-010213-115123

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*, 1012–1014. doi:10.1038/nature07634

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*, 267–297. doi:10.1093/pan/mps028

Howell, R., Kern, M. L., & Lyubomirsky, S. (2007). Health benefits: Meta-analytically determining the impact of well-being on objective health outcomes. *Health Psychology Review*, *1*, 83–136. doi:10.1080/17437190701492486

Labarthe, D. R. (2010). *Epidemiology and prevention of cardiovascular disease: A global challenge*. Sudbury, MA: Jones & Bartlett.

Lett, H. S., Blumenthal, J. A., Babyak, M. A., Sherwood, A., Strauman, T., Robins, C., & Newman, M. F. (2004). Depression as a risk factor for coronary artery disease: Evidence, mechanisms, and treatment. *Psychosomatic Medicine*, *66*, 305–315. doi:10.1097/01.psy.0000126207.43307.c0

Leyland, A. H. (2005). Socioeconomic gradients in the prevalence of cardiovascular disease in Scotland: The roles of composition and context. *Journal of Epidemiology & Community Health*, *59*, 799–803. doi:10.1136/jech.2005.034017

Lloyd-Jones, D. M., Hong, Y., Labarthe, D., Mozaffarian, D., Appel, L. J., Van Horn, L., . . . Rosamond, W. D. (2010). Defining and setting national goals for cardiovascular health promotion and disease reduction: The American Heart Association's strategic Impact Goal through 2020 and beyond. *Circulation*, *121*, 585–613. doi:10.1161/CIRCULATIONAHA.109.192703

Lochner, K. A., Kawachi, I., Brennan, R. T., & Buka, S. L. (2003). Social capital and neighborhood mortality rates in Chicago. *Social Science & Medicine*, *56*, 1797–1805. doi:10.1016/S0277-9536(02)00177-6

Menezes, A. R., Lavie, C. J., Milani, R. V., O'Keefe, J., & Lavie, T. J. (2011). Psychological risk factors and cardiovascular disease: Is it all in your head? *Postgraduate Medicine*, *123*, 165–176. doi:10.3810/pgm.2011.09.2472

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J. P., & Rosenquist, J. N. (2011). Understanding the demographics of Twitter users. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media* (pp. 554–557). Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2816/3234

Paul, M. J., & Dredze, M. (2011a). *A model for mining public health topics from Twitter*. Retrieved from the Johns Hopkins University Web site: http://www.cs.jhu.edu/~mpaul/files/2011.tech.twitter_health.pdf

Paul, M. J., & Dredze, M. (2011b). You are what you tweet: Analyzing Twitter for public health. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media* (pp. 265–272). Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2880/3264

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The development and psychometric properties of LIWC2007*. Austin, TX: LIWC.net.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*, 547–577. doi:10.1146/annurev.psych.54.101601.145041

Pierce, J. R., & Denison, A. V. (2010). Accuracy of death certificates and the implications for studying disease burdens. In V. R. Preedy & R. R. Watson (Eds.), *Handbook of disease burdens and quality of life measures* (pp. 329–344). New York, NY: Springer.

Pressman, S. D., & Cohen, S. (2005). Does positive affect influence health? *Psychological Bulletin*, *131*, 925–971. doi:10.1037/0033-2909.131.6.925

Riva, M., Gauvin, L., & Barnett, T. A. (2007). Toward the next generation of research into small area effects on health: A synthesis of multilevel investigations published since July 1998. *Journal of Epidemiology & Community Health*, *61*, 853–861. doi:10.1136/jech.2006.050740

Roest, A. M., Martens, E. J., de Jonge, P., & Denollet, J. (2010). Anxiety and risk of incident coronary heart disease: A meta-analysis. *Journal of the American College of Cardiology*, *56*, 38–46. doi:10.1016/j.jacc.2010.03.034

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, *52*, 59–82. doi:10.1146/annurev.psych.52.1.59

Rost, M., Barkhuus, L., Cramer, H., & Brown, B. (2013, February). *Representation and communication: Challenges in interpreting large social media datasets*. Paper presented at the 16th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW 2013), San Antonio, TX.

Rowe, J. W., & Kahn, R. L. (1987). Human aging: Usual and successful. *Science*, *237*, 143–149. doi:10.1126/science.3299702

Rugulies, R. (2002). Depression as a predictor for coronary heart disease: A review and meta-analysis. *American Journal of Preventive Medicine*, *23*, 51–61. doi:10.1016/S0749-3797(02)00439-7

Salathé, M., Freifeld, C. C., Mekaru, S. R., Tomasulo, A. F., & Brownstein, J. S. (2013). Influenza A (H7N9) and the importance of digital epidemiology. *New England Journal of Medicine*, *369*, 401–404. doi:10.1056/NEJMp1307752

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Lucas, R. E., Agrawal, M., . . . Ungar, L. H. (2013). Characterizing geographic variation in well-being using tweets. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media* (pp. 583–591). Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6138/6398.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., . . . Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, *8*(9), Article e73791. Retrieved from http://www.plosone.org/article/info:doi/10.1371/journal.pone.0073791

Seifter, A., Schwarzwalder, A., Geis, K., & Aucott, J. (2010). The utility of “Google Trends” for epidemiological research: Lyme disease as an example. *Geospatial Health*, *4*, 135–137.

St Louis, C., & Zorlu, G. (2012). Can Twitter predict disease outbreaks? *British Medical Journal*, *344*, Article e2353. Retrieved from http://www.bmj.com/content/344/bmj.e2353

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The general inquirer: A computer approach to content analysis*. Cambridge, MA: MIT Press.

Tay, L., Tan, K., Diener, E., & Gonzalez, E. (2013). Social relations, health behaviors, and health outcomes: A survey and synthesis. *Applied Psychology: Health and Well-Being*, *5*, 28–78.

U.S. Census Bureau. (2009). *Selected social characteristics in the United States: 2009. American Community Survey 1-year estimates*. Retrieved from http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_09_1YR_DP2&prodType=table

U.S. Census Bureau. (2010). *Profile of general population and housing characteristics: 2010 demographic profile data*. Retrieved from http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=DEC_10_DP_DPDP1&prodType=table

World Health Organization. (1992). Atherosclerotic heart disease. In *International statistical classification of diseases and related health problems* (10th ed.). Retrieved from http://apps.who.int/classifications/icd10/browse/2010/en#/I25.1

World Health Organization. (2011). *Global status report on noncommunicable diseases 2010. Description of the global burden of NCDs, their risk factors and determinants*. Retrieved from http://www.who.int/nmh/publications/ncd_report2010/en/

Online Supplement
Materials and Methods

Mapping Tweets to Counties

The method described in Schwartz et al. (2013a) was used to map language expressed on Twitter (tweets) to counties. This method relies on either the coordinates attached to a tweet (latitude, longitude) or the free-response "location" field for the Twitter user who posted the tweet to determine the tweet location. One percent of our tweets had coordinates. To map a pair of coordinates to a county, the point given by the coordinates was checked to see whether it was within the boundaries of a U.S. county. Other tweets were mapped to counties by the location text field. If the location field included city and state, we matched to the relevant county. For location fields with only city information, we could match counties if the name was unambiguous, defined as having a 90% likelihood of being one particular according to census population statistics (e.g., Chicago was unambiguously Chicago, Illinois, whereas Springfield could easily be Springfield in Pennsylvania, Virginia, or elsewhere). Large non-U.S. cities were also thrown out (e.g. London). This method favored fewer false positives (incorrect mappings) at the expense of mapping a more limited number of tweets. To access accuracy of this mapping process, human raters judged a sample of 100 tweets; 93% were true positives (correct mappings)**.** Approximately 16% of the tweets could be mapped to U.S. counties (about 148 million tweets).

Tokenization

Tokenization is the process of splitting sentences into words (also known as "tokens"). Typically, this involves identifying sequences of letters separated by spaces and disjoining punctuation where appropriate (e.g., "The C.D.C. reports heart disease rates aren't increasing." gets separated into "The", "C.D.C.", "reports", "heart", "disease", "rates", "aren't", "increasing", and "."). We used a tokenizer designed for social media that accurately captures emoticons such as ":)" (a smile) or "<3" (a heart) as words (Schwartz et al., 2013b). At the county-level, the frequencies of every unique word were summed, giving the word use for the county. From there, the dictionary and topic language features were derived.

Topic Extraction

Topics contain lists of semantically-related words. Unlike dictionaries, they are derived automatically, using a well-established algorithm from computer science, Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). LDA is a Bayesian mixture model, which groups words together that often appear together (e.g., one topic included *infection*, *ear*, *doctor*, *sinus*, *meds*, *antibiotics*, *poor*, *medicine*). Topics allow consideration of unanticipated categories of words. We used 2000 topics (available at wwbp.org/data.html) and derived their probability of usage per county from the relative word frequencies, as described in Schwartz et al. (2013b). The prevalence of a word in a topic is defined as the frequency for which that word appeared in the topic during fit of the LDA model. In other words, it is the estimated frequency for which the word is used as a representative of the topic.

Predictive Models

Our cross-sectional predictive models were fit via ridge regression (Hoerl & Kennard, 1970), which uses a standard machine learning approach of penalizing variable weights to avoid over-fitting due to variable multicollinearity. A10-fold cross-validation approach was used to fit and test models. Specifically, all 1,347 counties were divided randomly into ten nearly-equal sized groups ("folds"); nine folds were used as the "training set" in order to fit the model, and the final fold was used to test the model. The ridge regression method includes a penalization parameter (often called "alpha"), and we also used univariate feature selection, which includes a parameter automatically set by the algorithm by testing on a subset of the training data. Predictive accuracies (performance) were recorded as a Pearson *r* correlation between the predicted mortality rates and the Centers for Disease Control and Prevention (CDC) reported mortality rates (2010). The process was then repeated 10 times, such that a new fold became the test set each time, and the predictive accuracies were averaged across the 10 runs. Standard errors of the predictive accuracies were based on the accuracies across these 10 runs.

When using language features, we had many more independent variables (i.e., tens of thousands of language features) than we did units of analysis (counties). To avoid overfitting, we used univariate feature selection fed into Principal Component Analysis (PCA) (Hotelling, 1933; Martinsson, Rokhlin, Tygert, 2011*)* for each type of independent variable (i.e. running the word and phrase features separately from topics). In univariate feature selection for regression, we removed individual features that were not significantly correlated at a family-wise alpha of 60 with the mortality rates. PCA then reduced the number of dimensions to either 10% of its original size or half the number of counties – whichever was smaller. Both the significance level and dimensional reduction size were selected based on tests over the training sample. Such steps are common practice in the field of machine learning when dealing with large numbers of independent variables (Hastie, Tibshirani, Friedman, 2009). When creating a model based on non-language variables (i.e. the health and demographic values; at most 10 variables at a time), we entered the variables as independent variables into the linear ridge regression model without using univariate feature selection or dimensionality reduction, as these steps are unnecessary with simple conventional independent variables in a regression model.

References

American Community Survey (ACS). (2009). Selected social characteristics in the United States. Retrieved from http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_09_1YR_DP2&prodType=table

Behavioral Risk Factors Surveillance Survey. (2009-2010). Annual survey data. Centers for Disease Control and Prevention. Retrieved from www.cdc.gov/brfss/annual_data/annual_data.htm

Blei, D. M., Ng, A. Y., & Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Centers for Disease Control and Prevention (CDC). (2010). Underlying cause of death 1999-2010. *CDC WONDER Online Database*. Retrieved from http://wonder.cdc.gov/ucd-icd10.html

County Health Rankings and Roadmaps. (2010). Rankings data. Retrieved from http://www.countyhealthrankings.org/rankings/data

Diabetes Public Health Research. (2010). Diagnosed diabetes prevalence. Centers for Disease Control and Prevention. Retrieved from http://www.cdc.gov/diabetes/atlas/countydata/County_EXCELstatelistDM.html

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*, 2nd ed. Springer. http://dx.doi.org/10.1007/978-0-387-84858-7

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*, 55-67. http://dx.doi.org/10.1080/00401706.1970.10488634

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*, 417. http://dx.doi.org/10.1037/h0071325

Institute for Health Metrics and Evaluation (IHME). (2009). United States hypertension estimates by county 2001-2009. Global Health Data Exchange website. Retrieved from http://ghdx.healthmetricsandevaluation.org/record/united-states-hypertension-estimates-county-2001-2009

Martinsson, P. G., Rokhlin, V., & Tygert, M. (2011). A randomized algorithm for the decomposition of matrices. *Applied and Computational Harmonic Analysis, 30*, 47-68. http://dx.doi.org/10.1016/j.acha.2010.02.003

National Health Examination and Nutrition Survey. (2010). Centers for Disease Control and Prevention. Retrieved from http://www.cdc.gov/nchs/nhanes.htm

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Lucas, R. E., Agrawal, M., …, & Ungar, L. H. (2013a). Characterizing Geographic Variation in Well-Being using Tweets. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM)*. Boston, MA.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., …, & Ungar, L. H. (2013b). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE, 8*, e73791. http://dx.doi.org/10.1371/journal.pone.0073791

Table S1 *Variable Sources and Transformation*

Variable			Description of variable	Unit	Years covered	Source
Included variable	**Transformation**	**Categories**				
Atherosclerotic Heart Disease (AHD) mortality	averaged across years		International Classification of Disease (ICD) 10 code I25.1 recorded as underlying cause of death on death certificates, prepared for the county level and age-adjusted through the CDC (using year 2000 population estimates)	per 100,000 population	2009-2010	CDC Wonder, Underlying Cause of Death (CDC, 2010)
Income	log-transformed		Median household income	2010 inflation-adjusted US dollars	2008-2010	American Community Survey (ACS, 2010) 3-Year Estimates (Table DP03)
Educational Attainment Index	Independently standardized and then averaged	High school grad	Attainment of high school graduation or higher	% of population	2008-2010	ACS 3-Year Estimates (Table DP02) (ACS, 2010)
		College grad	Attainment of bachelor's degree or higher			
Diabetes			Adults (age 20+) diagnosed with diabetes	% of population	2008-2010	County-level estimates based on CDC's Behavioral Risk Factor Surveillance System (BRFSS) data (2009-2010), obtained through 2013 County Health Rankings (CHR; 2010) (see note).
Obesity			Body Mass Index >= 30, based on self-reported height and weight			
Smoking			Current adult smokers who have smoked >= 100 cigarettes in their lifetime		2005-2011	
Hypertension	averaged	male	Male adults (age 30+) who self-reported systolic BP of at least 140mm Hg and/or self-reported taking medication	% of population	2009	County-level estimates prepared through the Institute for Health Metrics and Evaluation (IHME; 2009) on the basis of CDC BRFSS data (see note).
		female	Female adults (age 30+) who self-reported systolic BP of at least 140mm Hg and/or self-reported taking medication			
% Black			Population of one race - Black or African American alone	% of population	2010	U.S. Census, Demographic Profile Data (Table DP01) (U.S. Census Bureau, 2010)
% Hispanic			Hispanic or Latino			
% Female			Female			
% Married	averaged	male	Male adults (age 15+) now married (not separated)	% of population	2008-2010	ACS 3-Year Estimates (Table DP02) (ACS, 2010)
		female	Female adults (age 15+) now married (not separated)			

Note on sources used for selected variables:

Diabetes and Obesity: County Health Rankings (CHR; 2010) used data from the National Center for Chronic Disease Prevention and Health Promotion's Division of Diabetes Translation (part of the CDC), which provides the Diabetes Public Health Resource (DPHR; 2010). DPHR used data from the CDC's Behavioral Risk Factor Surveillance System (BRFSS; 2009-2010), an ongoing national survey. DPHR developed county-level estimates from state-level BRFSS data using small area estimation techniques, including Bayesian multilevel modeling, multilevel logistic regression models, and a Markov Chain Monte Carlo simulation method.

Smoking: County-level estimates (based on BRFSS state-level data) were calculated for CHR by CDC staff.

Hypertension: The Institute for Health Metrics and Evaluation (IHME; 2009) used National Health Examination and Nutrition Survey data (1999-2008) to characterize the relationship between self-reported and physical measurements for various health factors. They used the resulting model to predict physical measurements for 2009 BRFSS participants (who supplied self-reported measures) and employed small area estimation techniques to estimate hypertension prevalence at the county-level.

Table S2

Dictionary Evaluation

	Dictionary	Top Ten Dictionary Words by Frequency	Two Rater Agreement	Accuracy
Risk Factors	Anger	shit f*** hate damn b*tch hell f***ing mad stupid b*tches	70.0%	60.0%
	Negative Relationships	hate alone jealous blame evil rude lonely independent hated ban	86.0%	75.5%
	Negative Emotion	sorry mad sad scared p*ssed crying horrible afraid terrible upset	87.0%	79.5%
	Disengagement	tired bored sleepy lazy blah meh exhausted yawn distracted boredom	91.0%	88.0%
	Anxiety	crazy pressure worry scared awkward scary fear doubt horrible afraid	81.5%	55.0%
Protective Factors	Positive Relationships	love home friends friend team social welcome together kind dear	75.0%	55.5%
	Positive Emotion	great happy cool awesome amazing glad excited super enjoy wonderful	93.0%	88.5%
	Engagement	learn interesting awake interested alive learning creative alert involved careful	74.5%	79.0%

Note. Each dictionary was evaluated by two independent raters. 200 random instances of tweets containing words from the dictionary in question were extracted, and the expert raters determined whether the word expressed the associated dictionary concept within the tweet. On average, the raters agreed 81.5% of the time, and a third rater was brought in to break ties. Accuracy refers to the percentage of tweets that expressed the associated dictionary concept, out of the 200 random instances sampled for every dictionary.

Table S3

Cross-Correlations between Dictionaries and Topics

	Anger	Negative Relation-ships	Negative Emotion	Disengage-ment	Anxiety	Positive Relation-ships†	Positive Emotion	Engage-ment
Anger	1	.76 [.73, .78]	.60 [.57, .64]	.72 [.69, .74]	.29 [.24, .34]	.18 [.26, .36]	-.33 [-.38, -.28]	-.30 [-.35, -.25]
Negative Relationships			.70 [.68, .73]	.67 [.64, .70]	.37 [.32, .41]	.42 [.50, .58]	-.04 [-.09, .01]	-.09 [-.14, -.04]
Negative Emotion				.55 [.51, .59]	.43 [.38, .47]	.45 [.50, .58]	.19 [.14, .24]	.04 [-.02, .09]
Disengagement					.29 [.24, .34]	.28 [.37, .46]	-.16 [-.21, -.11]	-.27 [-.32, -.22]
Anxiety						.38 [.29, .39]	.23 [.18, .28]	.16 [.11, .21]
Positive Relationships							.48 [.43, .52]	.23 [.18, .28]
Positive Emotion								.61 [.58, .64]

Topics	**Included Word**	Anger	Negative Relation-ships	Negative Emotion	Disengage-ment	Anxiety	Positive Relation-ships†	Positive Emotion	Engage-ment
Hostility, Aggression	bullsh*t	.94	.58	.43	.62	.19	-.03	-.45	-.40
	a**hole	.93	.62	.48	.61	.19	.00	-.41	-.39
	retarded	.81	.65	.56	.54	.21	.06	-.26	-.30
Hate, Inter-personal Tensions	hating	.88	.74	.54	.68	.23	.13	-.33	-.36
	drama	.87	.67	.53	.66	.26	.18	-.28	-.29
	passion	.67	.84	.66	.60	.33	.37	.02	-.08
Boredom, Fatigue	bored	.70	.60	.47	.87	.20	.16	-.26	-.35
	tired	.69	.70	.62	.87	.31	.32	-.04	-.21
	bed	.50	.61	.56	.69	.30	.41	.08	-.12
Skilled Occupations	management	-.42	-.32	-.23	-.41	.03	.29	.38	.69
	service	-.41	-.28	-.17	-.39	.08	.33	.51	.63
	conference	-.45	-.28	-.16	-.42	.11	.34	.56	.65
Positive Experiences	experience	-.30	-.12	-.01	-.26	.15	.42	.57	.76
	company	-.30	-.12	.11	-.21	.18	.54	.78	.67
	weekend	-.35	-.11	.09	-.22	.14	.55	.89	.62
Optimism, Resilience	opportunities	-.33	-.20	-.12	-.31	.10	.35	.41	.69
	achieve	-.21	-.07	.00	-.22	.17	.36	.39	.68
	strength	-.14	.06	.04	-.08	.29	.55	.48	.68

Note. Dictionary cross-correlations (Pearson *r*) are given, with 95% confidence intervals in brackets. To ease inspection, topic-dictionary correlations are color formatted, ranging from dark red (strongly negative) to dark green (strongly positive). Particularly strong correlations between topic clusters and dictionaries are emphasized with bolder boxes. Topics correspond to the topics shown in Figure 1, in the same order. The "included words" are dominant unique words in each cloud, which help identify the topic.
† The word "love" was removed from the dictionary, as it accounted for more than a third of all word occurrences in the dictionary, and distorted the results (see discussion).

Table S4

Performance of Regression Models Predicting AHD Mortality on the Basis of Different Sets of Predictors

Model	Demographic	SES	Health	Twitter	Accuracy of County-Level AHD Prediction	
1	X				.14 [.09, .19]	***
2	X			X	.42 [.38, .45]	
3		X			.23 [.18, .28]	***
4		X		X	.41 [.38, .45]	
5			X		.27 [.20, .34]	***
6			X	X	.42 [.38, .46]	
7	X	X			.32 [.27, .37]	***
8	X	X		X	.41 [.38, .45]	
9	X		X		.33 [.26, .40]	***
10	X		X	X	.42 [.38, .46]	
11		X	X		.29 [.23, .35]	***
12		X	X	X	.42 [.38, .46]	
13	X	X	X		.36 [.29, .43]	* *
14	X	X	X	X	.42 [.38, .46]	
15				X	.42 [.38, .45]	

Note. Performance of regression models predicting atherosclerotic heart disease (AHD) mortality from demographic variables (percentage of Blacks, Hispanics, married, and female residents), socioeconomic variables (income and education), health variables (incidence of diabetes, obesity, smoking, and hypertension), Twitter language, and all combinations of these sets of predictors. Accuracy refers to the Pearson *r* correlation between the set of predictors and CDC reported AHD. Brackets give 95% confidence intervals. The models are trained on one part of the data ("training set") and evaluated on another ("hold-out set"), to avoid distortion through chance. A model combining Twitter and all predictors (Model #14) significantly outpredicted the model with all predictors (Model 13), suggesting that Twitter has incremental predictive validity. Twitter language by itself significantly outpredicted a model with all SES, demographic and health predictors (Model 15 compared to Model 13). Predictive performance between two models was compared through paired t-tests, comparing the sizes of standardized residuals of county-level predictions from each model. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; † $p < 0.10$.

Table S5

Varimax-rotated Factor Structure of the County-level Frequencies of the 20 most Frequent Words in the Positive Relationship Dictionary

Words	Partnership factor	Social factor
love	**.65**	.39
home	.11	.35
friends	.47	**.53**
friend	.43	**.48**
team	-.07	.30
social	-.32	.13
welcome	-.09	.43
together	.40	.34
kind	-.23	.50
dear	.11	.41
agree	-.30	**.51**
loved	.03	**.51**
relationship	**.73**	.05
liked	.02	.12
loving	.18	.33
boyfriend	**.72**	.10
appreciate	.06	.27
girlfriend	**.66**	.06
helping	-.25	.38
united	-.27	.09

County-level correlations		
Socioeconomic Status (SES)†	-.43 [-.47, -.38]	.14 [.08, .19]
Atherosclerotic Heart Disease	.18 [.13, 23]	-.02 [-.07, .04]

Note. Examination of the eigenvalues and the Scree test revealed a clear two factor structure. Words are ordered in descending frequency of occurrence. Factor scores were imputed through regression (random factors, Thompson's method). Pearson correlations (r) are given with 95% confidence intervals in brackets. The 20 words shown account for 89.1% of all word occurrences of the positive relationship dictionary.

† SES index combining standardized high school and college graduation rates, and median income.

Table S6

Top Ten Dictionary Words by Frequency and Their Correlations with Atherosclerotic Heart Disease (AHD)

Anger Dictionary

Top Ten Words	Correlation with AHD Mortality (Pearson r with 95% CIs)	Correlation with AHD Mortality Controlled for Income and Education	Overall Frequency
shit	.12 [.06, .17]	.07 [.02, .13]	2,178,219
fuck	.20 [.15, .25]	.17 [.11, .22]	1,551,388
hate	.23 [.18, .28]	.19 [.13, .24]	1,307,810
damn	.03 [-.02, .09]	-.03 [-.08, .03]	1,252,834
bitch	.13 [.07, .18]	.06 [.01, .12]	864,810
hell	.01 [-.04, .07]	-.05 [-.11, .00]	781,102
fucking	.28 [.23, .33]	.29 [.24, .34]	651,694
mad	.13 [.08, .19]	.09 [.03, .14]	514,694
stupid	.11 [.06, .16]	.06 [.00, .11]	410,894
bitches	.13 [.08, .18]	.09 [.03, .14]	305,033

Negative Relationships Dictionary

Top Ten Words	Correlation with AHD Mortality (Pearson r with 95% CIs)	Correlation with AHD Mortality Controlled for Income and Education	Overall Frequency
hate	.23 [.18, .28]	.19 [.13, .24]	1,307,810
alone	.13 [.08, .18]	.09 [.03, .14]	292,621
jealous	.05 [-.01, .10]	.04 [-.02, .09]	177,374
blame	-.01 [-.07, .04]	-.01 [-.06, .04]	100,930
evil	-.07 [-.13, -.02]	-.07 [-.13, -.02]	94,161
rude	.04 [-.01, .10]	.02 [-.03, .08]	78,552
lonely	.05 [-.01, .10]	.01 [-.05, .06]	70,916
independent	-.04 [-.09, .01]	-.02 [-.08, .03]	39,313
hated	.10 [.05, .15]	.09 [.04, .14]	39,251
ban	-.05 [-.10, .00]	-.02 [-.07, .03]	36,417

Negative Emotions Dictionary

Top Ten Words	Correlation with AHD Mortality (Pearson r with 95% CIs)	Correlation with AHD Mortality Controlled for Income and Education	Overall Frequency
sorry	.04 [-.02, .09]	.04 [-.01, .09]	757,751
mad	.13 [.08, .19]	.09 [.03, .14]	514,694
sad	.00 [-.05, .06]	.00 [-.05, .05]	428,082
scared	.09 [.03, .14]	.03 [-.03, .08]	168,420
pissed	.19 [.14, .24]	.15 [.10, .20]	140,696
crying	.11 [.06, .17]	.09 [.04, .14]	123,994
horrible	.07 [.02, .12]	.08 [.02, .13]	113,522
afraid	.05 [-.01, .10]	.04 [-.02, .09]	104,582
terrible	.03 [-.03, .08]	.06 [.00, .11]	104,195
upset	.10 [.05, .15]	.08 [.02, .13]	93,648

Disengagement Dictionary

Top Ten Words	Correlation with AHD Mortality (Pearson r with 95% CIs)	Correlation with AHD Mortality Controlled for Income and Education	Overall Frequency
tired	.16 [.11, .21]	.10 [.05, .16]	580,979
bored	.18 [.13, .23]	.11 [.05, .16]	411,358
sleepy	-.01 [-.06, .04]	-.10 [-.16, -.05]	157,043
lazy	.04 [-.02, .09]	-.01 [-.06, .04]	138,761
blah	.07 [.02, .12]	.03 [-.02, .09]	110,085
meh	-.02 [-.07, .04]	-.04 [-.09, .01]	53,376
exhausted	.06 [.01, .12]	.09 [.03, .14]	49,955
yawn	-.03 [-.09, .02]	-.03 [-.08, .02]	21,398
distracted	-.06 [-.12, -.01]	-.04 [-.10, .01]	17,998
boredom	.04 [-.01, .10]	.04 [-.02, .09]	17,150

Anxiety Dictionary

Top Ten Words	Correlation with AHD Mortality (Pearson r with 95% CIs)	Correlation with AHD Mortality Controlled for Income and Education	Overall Frequency
crazy	.13 [.08, .18]	.09 [.04, .14]	696,947
pressure	.02 [-.03, .08]	.03 [-.02, .09]	193,805
worry	.05 [-.01, .10]	.02 [-.03, .08]	172,486
scared	.09 [.03, .14]	.03 [-.03, .08]	168,420
awkward	.09 [.04, .15]	.09 [.03, .14]	152,980
scary	-.02 [-.08, .03]	-.02 [-.07, .04]	121,521
fear	-.06 [-.12, -.01]	-.05 [-.10, .00]	120,542
doubt	.09 [.03, .14]	.09 [.03, .14]	115,207
horrible	.07 [.02, .12]	.08 [.02, .13]	113,522
afraid	.05 [-.01, .10]	.04 [-.02, .09]	104,582

Positive Relationships Dictionary

Top Ten Words	Correlation with AHD Mortality (Pearson r with 95% CIs)	Correlation with AHD Mortality Controlled for Income and Education	Overall Frequency
love	.13 [.08, .18]	.08 [.02, .13]	5,375,835
home	.11 [.05, .16]	.10 [.04, .15]	1,907,974
friends	.10 [.05, .15]	.09 [.04, .14]	1,005,756
friend	.05 [.00, .10]	.02 [-.03, .07]	721,639
team	-.07 [-.13, -.02]	-.05 [-.10, .01]	629,910
social	-.08 [-.14, -.03]	-.03 [-.09, .02]	448,731
welcome	-.04 [-.09, .01]	-.02 [-.07, .03]	421,685
together	.00 [-.05, .06]	-.02 [-.07, .04]	398,957
kind	-.09 [-.14, -.03]	-.04 [-.10, .01]	379,906
dear	.02 [-.03, .07]	.02 [-.03, .08]	289,738

Positive Emotion Dictionary

Top Ten Words	Correlation with AHD Mortality (Pearson r with 95% CIs)	Correlation with AHD Mortality Controlled for Income and Education	Overall Frequency
great	-.15 [-.21, -.10]	-.09 [-.15, -.04]	2,375,268
happy	.06 [.01, .12]	.06 [.01, .12]	1,830,533
cool	-.09 [-.14, -.04]	-.06 [-.12, -.01]	972,187
awesome	-.07 [-.12, -.01]	-.02 [-.08, .03]	971,447
amazing	.04 [-.01, .09]	.09 [.04, .15]	715,301
glad	-.07 [-.13, -.02]	-.09 [-.15, -.04]	499,789
excited	.00 [-.06, .05]	.04 [-.01, .09]	495,371
super	-.01 [-.06, .05]	.01 [-.04, .07]	473,677
enjoy	-.07 [-.12, -.01]	-.02 [-.07, .03]	381,689
wonderful	-.05 [-.10, .00]	-.04 [-.09, .02]	204,721

Engagement Dictionary

Top Ten Words	Correlation with AHD Mortality (Pearson r with 95% CIs)	Correlation with AHD Mortality Controlled for Income and Education	Overall Frequency
learn	-.08 [-.13, -.02]	-.05 [-.11, .00]	350,873
interesting	-.17 [-.22, -.12]	-.10 [-.15, -.04]	305,703
awake	.12 [.07, .17]	.11 [.05, .16]	158,400
interested	-.10 [-.15, -.05]	-.05 [-.10, .01]	137,553
alive	.07 [.01, .12]	.06 [.01, .11]	132,898
learning	-.11 [-.16, -.06]	-.07 [-.12, -.02]	118,337
creative	-.10 [-.16, -.05]	-.04 [-.10, .01]	89,367
alert	-.04 [-.09, .01]	-.02 [-.08, .03]	80,982
involved	-.09 [-.14, -.04]	-.05 [-.11, .00]	65,361
careful	-.07 [-.12, -.02]	-.09 [-.14, -.03]	63,719