More Evidence that Twitter Language Predicts Heart Disease: A Response and Replication

Johannes C. Eichstaedt[1], H. Andrew Schwartz[2], Salvatore Giorgi[1], Margaret L. Kern[3], Gregory Park, Maarten Sap[4], Darwin R. Labarthe[5], Emily E. Larson[6], Martin E. P. Seligman[1], Lyle H. Ungar[1,7]

[1]Positive Psychology Center, University of Pennsylvania, USA

[2]Computer Science Department, Stony Brook University, USA

[3]Melbourne Graduate School of Education, University of Melbourne, Australia

[4]Paul G. Allen School of Computer Science & Engineering, University of Washington, USA

[5]Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, USA

[6]International Positive Education Network, London, UK

[7]Penn Medicine Center for Digital Health, University of Pennsylvania, USA

Acknowledgements

We thank our Research Assistant Meghana Nallajerla for her work on the manuscript and Michelle Schmitz for her help processing county-level elevation data.

Summary

A recent preprint by Brown and Coyne titled, "No Evidence That Twitter Language Reliably Predicts Heart Disease: A Reanalysis of Eichstaedt et al." asserts to re-analyze our 2015 article published in Psychological Science, “Twitter Language Predicts Heart Disease Mortality”, disputing its primary findings. While we welcome scrutiny of the study, Brown and Coyne’s paper does not in fact report on a reanalysis, but rather presents a new analysis relating Twitter language to suicide instead of heart disease mortality.

In our original article, we showed that Twitter language, fed into standard machine learning algorithms, was able to predict (i.e., estimate cross-sectionally) the out-of-sample heart disease rates of U.S. counties. Further, in a separate analysis, we found that the dictionaries and topics (i.e., sets of related words) which best predicted county atherosclerotic heart disease mortality rates included language related to education and income (e.g., “management,” “ideas,” “conference”) as well as negative social relationships (“hate”, “alone,” “jealous”), disengagement (“tired, “bored,” “sleepy”), negative emotions (“sorry,” “mad,” “sad”) as well as positive emotions (“great,” “happy,” “cool”) and psychological engagement (“learn,” “interesting,” “awake”).

Beyond conducting a new analysis (correlating Twitter language with suicide rates), Brown and Coyne also detail a number of methodological limitations of group-level and social media-based studies. We discussed most of these limitations in our original article, but welcome this opportunity to emphasize some of the key aspects and qualifiers of our findings, considering each of their critiques and how they relate to our findings. Of particular note, even though we discuss our findings in the context of what is known about the etiology of heart disease at the individual level, we reiterate here a point made in our original paper: that individual-level causal inferences cannot be made from the cross-sectional and group-level analyses we presented. Our findings are intended to provide a new epidemiological tool to take advantage of large amounts of public data, and to complement, not replace, definitive health data collected through other means.

We offer preliminary comments on the suicide language correlations: Previous studies have suggested that county-level suicides are relatively strongly associated with living in rural areas (Hirsch et al., 2006; Searles et al., 2014) and with county elevation (Kim et al., 2011; Brenner et al., 2011). When we control for these two confounds, we find the dictionary associations reported by Brown and Coyne are no longer significant. We conclude that their analysis is largely unrelated to our study and does not invalidate the findings of our original paper.

In addition, we offer a replication of our original findings across more years, with a larger Twitter data set. We find that (a) Twitter language still predicts county atherosclerotic heart disease mortality with the same accuracy, and (b) the specific dictionary correlations we reported are largely unchanged on the new data set. To facilitate the reproduction by other researchers of our original work, we also re-release the data and code with which to reproduce our original findings, making it more user-friendly. We will do the same for this replication upon publication.

More Evidence that Twitter Language Predicts Heart Disease: A Response and Replication

The 2015 Twitter and Heart Disease Study

In 2015, we published a study ("Twitter Language Predicts Heart Disease Mortality"). In it, we showed that when the language used by people in most U.S. counties on Twitter is fed into a standard machine learning algorithm, together with the mortality rates of a prevalent type of heart disease, the algorithm was able to output ("predict") the heart disease rates of the remaining counties. In other words, we showed that a machine learning algorithm could "learn" what kind of language was used on Twitter in communities with high heart disease rates and use that language to estimate the heart disease rates of counties for which it only knew the Twitter language profile, but not the actual heart disease rates. Not only was the algorithm able to predict beyond chance, but we showed that such models were slightly but significantly more accurate than the same type of model using ten leading variables together (e.g., demographics, income, education, smoking, hypertension).

Our analyses suggested that language shared on Twitter in a given county carries information about the health of the community. This was a group-level (or ecological) finding which is not be surprising, as many of the same variables that predict community heart disease--such as income-- have been predicted for Twitter users based on their Tweets (e.g., Flekova, Preoţiuc-Pietro, & Ungar, 2016). In other words, there is ample evidence to suggest that what is being said on Twitter changes with the characteristics of people who tweet. And very similar to the heart disease predictions work, other health factors (like obesity; Culotta, 2014) and psychological factors (like life satisfaction; Schwartz et al., 2013) of counties have been predicted from their aggregated Twitter language, suggesting that what is being shared on Twitter in a county changes with the characteristics of the people within that county.

Predictive Evaluation. In our study, the overall accuracy achieved by both the standard predictors and our Twitter model, as measured by correlation between algorithm-predicted and actual mortality rates, reached an effect size of $r = 0.42$. That is, the language variables we derived from Twitter accounted for 17% of the variance in heart disease mortality. Such a prediction is substantive, but not earth-shattering. In other words, we found evidence that there is some health-related signal that can be detected within all the noise of Twitter data.

Dictionary and Topic Correlates. We unpacked these predictions by asking what people talk about on Twitter that is correlated with greater county heart disease mortality. While typical survey-based research can only test known variables chosen *a priori* for the survey, the associated Twitter language can provide direction for new psychosocial insights. Correlated with greater mortality, we found topics related to hostility and aggression (*shit, asshole, fucking),* hate and interpersonal tension (*jealous, drama, hate*), and boredom and fatigue (*bored, tired, bed*). Correlated with less mortality, we found topics related to positive experiences (*wonderful, friends, great*), skilled occupations (*service, skills, conference*) and optimism (*opportunities,*

goals, overcome). All correlations were significant ($p < .01$; adjusted for multiple tests using a Bonferroni correction).

We obtained this result using both data-driven language analysis methods ("latent topics") as well as by using an approach that has been used for decades in psychology--counting the frequency of words in dictionaries (lists of words; for correlations between these methods, see Supplementary Table S2). Using dictionaries, we found similar results, suggesting that negative social relationships, disengagement and negative emotions on the one hand, and positive emotions and engagement on the other were associated with greater and less heart disease mortality, respectively (see Eichstaedt et al., 2015, Table 1). These approaches are not perfect, as we acknowledged in the original paper and discuss below, but provide meaningful findings.

The Critique

Conducting a new analysis with suicide as the outcome, Brown and Coyne (2018) argued that our findings are implausible and raise a variety of thoughtful concerns about epidemiological and social media-based methods. We summarize the main concerns as: a) both Twitter and the CDC-reported heart disease data contain various sample and reporting biases as well as inaccuracies and errors, and b) the specific language correlations we reported are not observed when county-level suicide rates are used as the outcome. We note that they did not perform analyses regarding whether or not data from Twitter can be used to predict heart disease (i.e., nothing akin to our "predictive evaluation"). We address the concerns they brought up regarding the dictionary and topic correlattions below, with more details in Appendix B.

Noise in Twitter Data

Brown and Coyne note that there are various sources of noise in Twitter data. As we noted in our article, (1) users who tweet are not representatively selected, and (2) some of the tweets (7%) are incorrectly mapped to counties. Further, people move from county to county, the way the "Garden Hose" Twitter sample is selected is non-random or otherwise imperfectly provided by Twitter, there are bots on Twitter, and so forth. We noted many of these concerns in the article and have continued to explore and to publish on limitations of social media data (e.g., Kern et al., 2016).

We agree with these concerns. This is partly why the results were surprising to us. As we note in the article:

> "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." (p. 166)

In other words, the noise works against our ability to predict mortality, since mortality data does not have the same selection biases and potential sources of errors. Specifically, we evaluated our accuracy "out of sample" – that is, we create the model on one set of data, and then

test the model (i.e., obtain the prediction accuracies) on a different part of the data ("cross-validation"). Greater noise in the data makes it harder to detect signal. Despite this noise, we were able predict heart disease from Twitter.

It bears repeating that our claim *how much* signal we were able to detect was modest (17% of the variance). A lot of factors impact how long a person lives and what they die from.

Noise in CDC Data

Brown and Coyne also repeated, as we point out in the original paper, that records of the underlying cause of death on death certificates have imperfections. As with nearly all health and psychological outcomes, no measure is perfect. While improvements in cause of death records are desirable, we note that use of such data is the standard for large-scale mortality studies in the U.S. (e.g., Pinner et al., 1996; Armstrong, Conn & Pinner, 1999; Jamal et al., 2005; Murray et al., 2006; Hansen et al., 2016) and we have no reason to believe that correcting the imperfections would change our key findings.

The Language Correlations for Suicide (Not a Replication)

Brown and Coyne used the aggregate Twitter data that we made available, but used suicide mortality rather than atherosclerotic heart disease (AHD) as the outcome. While this is an interesting analysis, it was not a reanalysis of our study on heart disease mortality. Additionally, out of the two types of analyses we conducted, *predictive evaluation* and *dictionary and topic correlates*, Brown and Coyne only present a result related to the later: dictionary and topic correlations of suicides (but not a predictive evaluation).

Brown and Coyne make a theoretical argument relying on the assumption that "we might expect county-level psychological factors that act directly on the health and welfare of members of the local community to be more closely reflected in the mortality statistics for suicide than those for a chronic disease such as AHD." (page 5). Across a smaller sample (N = 741) of counties for which 2009-2010 mortality from intentional self-harm was available, they observed that dictionaries they termed negative (Negative emotions, Anger, and Negative relationships) correlated *negatively* with suicides—unlike our heart disease findings--suggesting, for example, that communities expressing more anger on Twitter are those which commit *fewer* suicides—a surprising finding.

In previous work, we have observed approximately the same correlations, and we were able to closely reproduce the correlations reported by Brown and Coyne (left column in Table 1). We concur that they show a very different pattern than atherosclerotic heart disease mortality. Of note, across this sample of N = 741 counties, suicide rates are uncorrelated with heart disease mortality rates (r = -.06 [-.13, .02], p = .135). This is less surprising that it may first appear--others have shown that suicides are a complex mortality outcome that shows strong and robust links at the county level to (a) elevation (r = .51, as reported by Kim et al., 2011, and r = .50, as reported by Brenner et al., 2011) – perhaps because of the influence of hypoxia on serotonin metabolism (Bach et al., 2014), in addition to (b) living in a rural areas (e.g., see, Hirsch, 2006, for a review; Searles et al., 2014), attributed in part to social isolation and insufficient social

integration, a trend that has increased over time (Singh & Siahpush, 2002). This suggests Brown and Coyne's assumption, that suicide and heart disease should have the same county-level correlates, is not supported by literature on the epidemiology of suicide.

We next considered the question empirically ourselves: we looked for evidence of the same patterns suggested by the literature in our data. We found correlations between suicide mortality and the percentage of the population living in rural areas (r = .46 [.41, .52], p < .001) and county-level elevation (r = .45 [.39, .51], p < .001) that were (nominally) larger than any observed in the extensive list of all the socio, economic and demographic variables reported in and released with our 2015 paper (*countyoutcomes.csv* from https://osf.io/rt6w2/files/.).[1,2] In fact, as can be seen in Table 1, when we control for elevation and the percentage of the population living in rural areas, the dictionary associations reported by Coyne and Brown are no longer significant (and disengagement is again associated in the same direction as heart disease).

This suggests that elevation and the fraction of the population living in rural areas are two critical sources of ecological correlates underlying suicides: high suicide rates mark rural communities or those higher in elevation, which differ from lower-elevation and non-rural communities in complex ways (including gun-ownership; Searles et al., 2014). In contrast, applying the same statistical controls to the associations between Twitter dictionaries and heart disease rates does not substantively change them (Table 1, columns 3 and 4), suggesting that the heart disease rates are not affected by the same confounds. This suggests that suicide rates are unlike heart disease mortality (they are uncorrelated), and that county-level suicide rates cannot be used as a straight-forward estimate of the psychological health of communities.

Table 1

Correlations of 2009-2010 Twitter Language with Suicides and AHD, with and without Controlling for County Elevation and Population in Rural Area

Dictionary variable	Suicides		Atherosclerotic Heart Disease Mortality	
	Correlation (r)	Controlled for elevation and rural population (β)	Correlation (r)	Controlled for elevation and rural population (β)
Anger	-.17 [-.24, -.10] ***	-.04 [-.10, .02]	.21 [.14, .28] ***	.19 [.12, .27] ***
Negative relationships	-.09 [-.17, -.02] **	-.04 [-.10, .02]	.18 [.11, .25] ***	.15 [.08, .23] ***
Negative emotions	-.10 [-.17, -.03] **	-.03 [-.09, .02]	.13 [.06, .20] ***	.11 [.04, .19] **
Disengagement	.01 [-.06, .08]	.07 [.01, .13] *	.17 [.10, .24] ***	.15 [.07, .22] ***
Anxiety	-.05 [-.12, .03]	-.01 [-.07, .04]	-.01 [-.08, .06]	-.01 [-.08, .06]
Positive Relationships	.04 [-.03, .11]	.04 [-.02, .10]	.06 [-.02, .13]	.03 [-.04, .10]
Positive Emotions	.06 [-.01, .13]	.01 [-.04, .07]	-.16 [-.23, -.09] ***	-.15 [-.22, -.08] ***
Engagement	-.03 [-.10, .04]	-.02 [-.08, .03]	-.21 [-.27, -.14] ***	-.19 [-.26, -.11] ***

[1] This holds true despite the inclusion of only N = 741 more populous (and thus less rural) counties for which 2009/2010 suicide and other county data was available – suggesting that this relationship would be even stronger if more counties were included.

[2] A correlation of r = .46, for comparison, is also nominally higher than the performance of our best heart disease prediction model reported in the 2015 heart disease paper (r = .42, 95% CI = [.38, .46]).

Note: The table presents Pearson *r*s (correlation) and betas (standardized regression coefficients), with 95% confidence intervals in square brackets (across *n* = 741 counties for which AHD mortality, and percentage of population living in rural area, elevation, suicide and sufficient Twitter language data was available). 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). For simplicity, the word "love" has **not** been removed from the *Positive Relationships* dictionary, unlike Table 1 in Eichstaedt et al., 2015 (but as reported in the discussion). Suicides are age-adjusted rates exported from CDC Wonder as the Underlying Cause of Death on Death Certificates, following Brown & Coyne, 2018. ***p < .001, **p < .01, *p < .05

We more directly test Brown and Coyne's hypothesis that more psychological variables ought to be better candidates for association with psychological Twitter language by testing the most psychological variable we had released with the original 2015 paper: the number of mentally unhealthy days people reported on average in a county, based on the CDC's Behavioral Risk Factor Surveillance System (BRFSS), aggregated to the county level across 2005-2011 by CountyHealthRankings (2013). Table 2 shows its correlation with the dictionary-based language variables, with heart disease mortality and suicide correlations for comparison. Unlike suicides, mentally unhealthy days correlate with the psychological dictionaries in the same directions as heart disease mortality[3]. This preliminary analysis reaffirms the importance of considering ecological confounds. We have a manuscript in preparation that investigates county suicide predictions and their correlational profiles.

Table 2

Correlations of 2009-2010 Twitter Language with Heart Disease Mortality, Mentally Unhealthy Days, and Suicides

Dictionary variable	Correlation with AHD mortality	Correlation with Mentally Unhealthy Days	Correlation with Suicides
Anger	.21 [.14, .28] ***	.16 [.09, .23] ***	-.17 [-.24, -.10] ***
Negative relationships	.18 [.11, .25] ***	.20 [.13, .27] ***	-.09 [-.17, -.02] **
Negative emotions	.13 [.06, .20] ***	.10 [.03, .17] **	-.10 [-.17, -.03] **
Disengagement	.17 [.10, .24] ***	.30 [.23, .36] ***	.01 [-.06, .08]
Anxiety	-.01 [-.08, .06]	-.02 [-.09, .05]	-.05 [-.12, .03]
Positive Relationships	.06 [-.02, .13]	.17 [.10, .24] ***	.04 [-.03, .11]
Positive Emotions	-.16 [-.23, -.09] ***	-.12 [-.19, -.05] **	.06 [-.01, .13]
Engagement	-.21 [-.27, -.14] ***	-.22 [-.28, -.15] ***	-.03 [-.10, .04]

Note: The table presents Pearson *r*s, with 95% confidence intervals in square brackets (across *n* = 741 counties for which AHD mortality, mentally unhealthy days and suicide data was available). 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). For simplicity, the word "love" has **not** been removed from the *Positive Relationships* dictionary, unlike Table 1 in

[3]Associations remain significant (except for Anxiety) and in the same direction as reported in Table 2 when controlling for percentage of the population living in rural areas and elevation (all p's < .016).

Eichstaedt et al., 2015 (but as reported in the discussion). Mentally unhealthy days were estimated via phone survey to the CDC's Behavioral Risk Factor Surveillance System (2018), aggregated across 2005-2011 to the county-level, and released by CountyHealthRankings (2013). Suicides are age-adjusted rates exported from CDC Wonder as the Underlying Cause of Death on Death Certificates, following Brown & Coyne, 2018.
*** p < .001, **p < .01

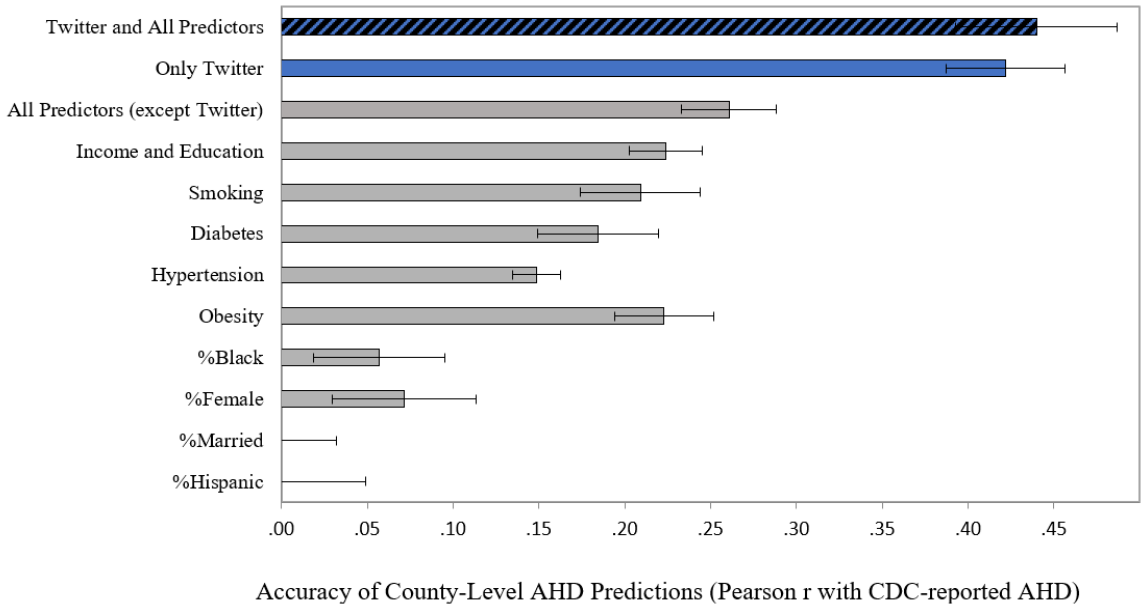
Replication of the Original Findings on New Data

The criticism also provides an opportunity to further test the original results. We reproduced the results on a much larger Twitter data set (951 rather than 148 million county-mapped tweets), spanning the years 2012/13 (in both Twitter and county-level demographic data; see supplementary table S1 for county data sources) across 1,536 rather than 1,347 counties. Figure 1a shows these new prediction results using Twitter and various demographic and health variables.[4] For simplicity, Twitter predictions are based only on 2,000 language topics (not additional dictionaries, word and phrase dictionaries which yield small improvements) used as predictors in a ridge regression model using cross-validation (see supplementary methods, Eichstaedt et al, 2015). Figure 1b shows the original results over 2009/2010 data, published in the 2015 article, for comparison.

In Eichstaedt et al., 2015, we reported the Twitter-only prediction model to reach an out-of-sample accuracy of r = .42 [.38, .45] across 2009-2010 data. Across 2012-2013 data, using only topics as language features in a ridge regression model without additional feature selection (but with updated data aggregation method[1]), we observe an equivalent accuracy (r = .42, [.39, .46]). In the 2009-2010 data, the prediction model combining all non-Twitter variables reaches a somewhat higher accuracy (r = .36, 95% CI = [.29, .43]) than an equivalent model does across 2012-2013 data (r = .26 [.23, .29]). As a result, the 2012-2013 Twitter model significantly out predicts this lower baseline by a larger margin than the 2009-2010 study ($t(1535) = -9.93$, one-tailed $p < .001$ across 2012-2013 vs. $t(1346) = -1.97$, $p = .049$ across 2009-2010, see Eichstaedt et al., 2015).

[4] These results draw on updated methods in which word use frequencies are not simply aggregated to the county-level, but first to the Twitter user-level, effectively assembling a sample of Twitter users per county, which is then averaged (Giorgi, Preotiuc-Pietro, & Schwartz, under review). With the publication of the corresponding manuscript introducing this method, we will release the improved person-weighted county-level language frequencies which again will allow for replication of the new results presented here.

a) Using 2012-2013 data (for both Twitter and traditional variables)



b) Using 2009-2010 data (as reported in Eichstaedt et al., 2015)

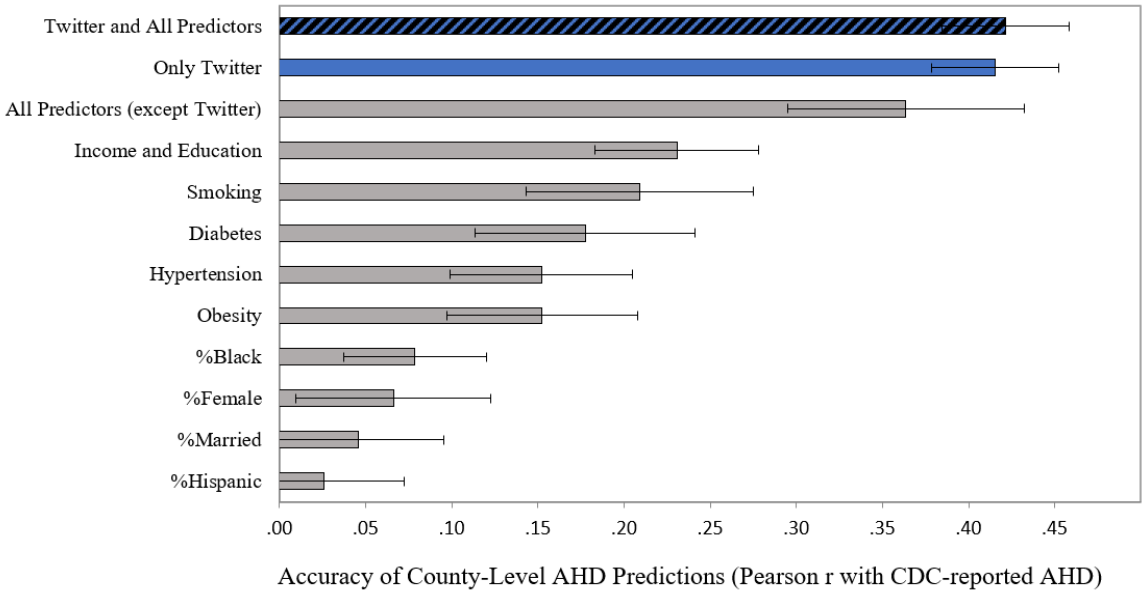


Figure 1. Performance of models predicting age-adjusted mortality from atherosclerotic heart disease (AHD) across (a) n = 1,536 and (b) n = 1,347 counties. For each model, the graph shows the correlation between predicted mortality and actual mortality reported by the Centers for Disease Control and 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. Error bars show 95% confidence intervals.

This thus adds further evidence for our original claim, that Twitter predicts (i.e., language patterns correlate with, but do not cause) county-level heart disease mortality. In Table 3, for completeness, we show correlations of the same dictionaries over this new data set, roughly matching those reported in the original article. As Supplementary Table S2 (adapted from Table S3 in Eichstaedt et al., 2015) shows, dictionaries and topics inter-correlate quite highly, so that we do not report the topic correlations here for brevity (these are easily reproducible from the 2012-2013 county frequency data we will be releasing).

Table 3

Correlations of Twitter Language Dictionaries and AHD Mortality from 2009-2010 (cf. Eichstaedt et al. 2015, Table 1) and 2012-2013

Dictionary variable	2009-2010 Twitter and AHD mortality data	2012-2013 Twitter and AHD mortality data
Anger	.17 [.11, .22] ***	.17 [.12, .22] ***
Negative relationships	.16 [.11, .21] ***	.24 [.19, .28] ***
Negative emotions	.10 [.05, .16] ***	.22 [.18, .27] ***
Disengagement	.14 [.08, .19] ***	.11 [.06, .16] ***
Anxiety	.05 [.00, .11] †	.11 [.06, .16] ***
Positive Relationships	.08 [.03, .13] **	.12 [.07, .17] ***
Positive Emotions	-.11 [-.17; -.06] ***	-.15 [-.20, -.10] ***
Engagement	-.16 [-.21, -.10] ***	-.21 [-.26, -.17] ***

Note: The table presents Pearson *r*s, with 95% confidence intervals in square brackets (across n = 1347 (2009-2010) and *n* = 1,536 (2012-2013) counties for which AHD mortality and sufficient Twitter data was available). 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). For simplicity, the word "love" has ***not*** been removed from the *Positive Relationships* dictionary across both time periods (unlike Table 1 in Eichstaedt et al., 2015).
*** p < .001, **p < .01, †p < .10

Release of Data and Code

We released both the county-level (a) Twitter language and (b) outcome data in a way that allowed people, in principle, to reproduce our findings (county-level topic, dictionary, and 1-to-3-gram frequencies, see https://osf.io/rt6w2/).

We also released an early version of our software on our homepage (wwbp.org), later in 2015. Since then, we have improved usability and documentation as well as released it open source in 2017 (Differential Language Analysis ToolKit, dlatk.wwbp.org; Schwartz et al, 2017). Along with this response, we are releasing step-by-step instructions to reproduce the prediction accuracies on the original data (2009-2010; see Appendix A), also re-shared on the Open Science Framework in a form suitable for direct database import (https://osf.io/7b9va/). Once everything

is installed on the user's system, it takes four DLATK commands (see Appendix A) to reproduce our original findings (reported in Eichstaedt et al., 2015) within the confidence intervals. We are committed to transparency and the values of open science.

Discussion

Brown and Coyne (2017) raise many concerns about our analysis, from the measurement error in the government-reported county-level variables to the many sources of noise in the Twitter data, many of which we acknowledged in our original paper. However, their critique does not attempt a replication of our claim that county-level Twitter language predicts county-level heart disease rates. Instead, it contains an exploratory analysis of language correlates of county-level suicide rates--which are uncorrelated with heart disease rates and disappear when county elevation and rural populations are controlled for (unlike heart disease associations), suggesting that county-level suicide rates are not a straightforward measure of county-level psychological health. A CDC-reported measure of poor mental health based on phone surveys, on the other hand, shows the same pattern of correlations with psychological Twitter Language as does heart disease mortality. This deserves further study, and we look forward to continued exploration about how social-media language relates to behavior and health outcomes.

In the spirit of replication, we have provided here a replication across different years of Twitter and county-level variable data, finding that Twitter language predicts heart disease at an equivalent accuracy to what we had originally reported (r = .42). We also observed largely similar dictionary correlations. We are also re-releasing the original data and the analysis code in a way that we hope will make reproductions of our results more accessible (see Appendix A).

One of the limitations we mentioned in the original paper, which we have since further explored and grown more concerned about (and which is not mentioned by Brown and Coyle) concerns the use of simple dictionaries to infer county-level psychological characteristics. We have since observed that while many standard dictionaries correlate as expected with traits at the individual level (for example, extraverted people use more words in positive emotion dictionaries), when applying them to county-level data, differences in language use across the U.S. may induce false positives in some dictionaries. In part, we had stumbled over this finding in the 2015 paper when realizing that "love," a highly frequent word with different word senses, correlated positively with heart disease rates. We have a manuscript in preparation to discuss these complexities (Jadika et al, in preparation). Our suggestion for future researchers is to instead use machine-learning based prediction models to estimate psychological factors, they seem to produce more robust estimates at the county-level than simple application of (unweighted) dictionaries.

The above is a limitation of applying dictionaries to county-level Twitter data in general. However, we have since been able to validate the dictionaries reported in this response and in the original 2015 paper against county-level well-being estimates from the Gallup-Healthways Well-Being Index, based on roughly 2 million respondents (G.H.W.B Index, 2017; Jadika et al, in preparation). All county-level dictionary frequencies reported here correlate significantly and in the expected directions with both county-level life satisfaction and happiness, except the

"Positive Relationships" dictionary, which correlates negatively with both well-being variables--similar to its negative correlation with county-level heart disease mortality. We have thus maintained confidence in all but the positive relationships dictionary to estimate psychological language use on Twitter.

Conclusion

Our original study, and replication on a new dataset presented here, show that machine learning models can detect county-level patterns in language use on Twitter which can predict county-level atherosclerotic heart disease mortality rates. While causal claims are not possible, we hypothesize that the characteristics of a community are reflected in what members of that community share on Twitter, and that Twitter may thus serve as a novel window into community-level health.

Materials and Methods

Main Replication

Twitter data. 2012-2013 Twitter data was a random sample of the 10%, aggregated to the user and then to the county level as described in Giorgi, Preotiuc-Pietro & Schwartz (under review). From the Twitter data, we extracted word frequencies as outlined in Eichstaedt et. al, 2015, and the same 2,000 topics (which can be found at wwbp.org/data.html). We will release these topic frequencies with the publication of the manuscript referenced above.

Economic, demographic and health variables. Supplementary Table 1 summarizes the sources of the 2012-2013 county-level data. We were not able to find county-level hypertension estimates after 2009, and thus used the same 2009 variable used in the Eichstaedt et al., 2015 analysis. Sufficient Twitter language data and county variables were available for N = 1,536 counties.

Prediction models. We built simple ridge regression models (a straightforward machine learning extension of linear regression models) using DLATK (Schwartz et al, 2017), picking ridge hyperparameters that are appropriate for the number of different predictors in the different models (2,000 Twitter topics: *alpha* = 10,000, Twitter and all predictors: *alpha* = 10,000, income and education model: *alpha* = 100 and all other predictors: *alpha* = 1).

Dictionary extraction and correlation. We extracted the same set of dictionaries described in Eichstaedt et al. (2015) from the 2012-2013 Twitter word frequencies and correlated them with the 2012-2013 heart disease data.

Exploratory Suicide Analysis

Twitter data. We used the dictionary frequencies originally released with Eichstaedt et al, 2015 (https://osf.io/rt6w2/) extracted from a 10% random sample of geo-tagged 2009-2010 Tweets.

Mentally unhealthy days. We used the county-level estimates of average number of self-reported mentally unhealthy days from the CDC's Behavioral Risk Factor Surveillance System , aggregated across 2005 to 2011, and provided by CountyHealthRankings (2013).

Suicide rates. We obtained estimates for suicide rates recorded as the underlying cause of death on death certificates from CDC Wonder (2018), using ICD-10 codes X60-X84, following Brown and Coyne (2017). Data from all three sources was available for N = 741 counties.

County elevation. We used the height of the surface above sea level at a county's centroid, determined by the CGIAR Consortium for Spatial Information.

Percentage of population living in rural area. We obtained this variable from the 2010 population census estimates, provided by CountyHealthRankings (2017).

References

Armstrong, G. L., Conn, L. A., & Pinner, R. W. (1999). Trends in infectious disease mortality in the United States during the 20th century. *Journal of the American Medical Association*, 281(1), 61-66.

Bach, H., Huang, Y. Y., Underwood, M. D., Dwork, A. J., Mann, J. J., & Arango, V. (2014). Elevated serotonin and 5-HIAA in the brainstem and lower serotonin turnover in the prefrontal cortex of suicides. Synapse, 68(3), 127-130.

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.

Berk, R. A. (1983). An introduction to sample selection bias in sociological data. *American Sociological Review*, 386-398.

Beyer, K. M., Schultz, A. F., & Rushton, G. (2008). Using ZIP codes as geocodes in cancer research. *Geocoding health data: The use of geographic codes in cancer prevention and control, research and practice*, 37-68.

Brenner, B., Cheng, D., Clark, S., & Camargo Jr, C. A. (2011). Positive association between altitude and suicide in 2584 US counties. High altitude medicine & biology, 12(1), 31-35.

Brown, N. J., & Coyne, J. (2018). No Evidence That Twitter Language Reliably Predicts Heart Disease: A Reanalysis of Eichstaedt et al (2015). Retrieved from https://psyarxiv.com/dursw (on 2/13/2018). DOI: 10.17605/OSF.IO/DURSW

Centers for Disease Control and Prevention. “CDC's Behavioral Risk Factor Surveillance System Website."

Centers for Disease Control and Prevention, & National Center for Health Statistics. (2015). Underlying cause of death 1999-2013 on CDC WONDER online database, released 2015. *Data are from the multiple cause of death files, 2013*.

Chida, Y., & Steptoe, A. (2009). Cortisol awakening response and psychosocial factors: a systematic review and meta-analysis. *Biological psychology*, *80*(3), 265-278.

Clark, A. M., DesMeules, M., Luo, W., Duncan, A. S., & Wielgosz, A. (2009). Socioeconomic status and cardiovascular disease: risks and implications for care. *Nature Reviews Cardiology*, *6*(11), 712.

County Health Rankings - 2013. (2013). Retrieved from http://www.countyhealthrankings.org/

Culotta, A. (2014, April). Estimating county health statistics with twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1335-1344). ACM.

Diez Roux, A. V., & Mair, C. (2010). Neighborhoods and health. *Annals of the New York Academy of Sciences*, *1186*(1), 125-145.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., ... & Weeg, C. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological science*, *26*(2), 159-169.

Flekova, L., Preoţiuc-Pietro, D., & Ungar, L. (2016). Exploring stylistic variation with age and income on twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Vol. 2, pp. 313-319).

Giorgi, S., Preotiuc-Pietro, D. and Schwartz, H. A., The Geo-User Lexical Bank and the Importance of Person-Level Aggregation (under review)

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. Psychological review, 114(2), 211.

Fontanella, C. A., Hiance-Steelesmith, D. L., Phillips, G. S., Bridge, J. A., Lester, N., Sweeney, H. A., & Campo, J. V. (2015). Widening rural-urban disparities in youth suicides, United States, 1996-2010. *JAMA pediatrics*, 169(5), 466-473.

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 Google Scholar

Friedman, M., & Rosenman, R. H. (1959). Association of specific overt behavior pattern with blood and cardiovascular findings: blood cholesterol level, blood clotting time, incidence of arcus senilis, and clinical coronary artery disease. *Journal of the American Medical Association*, *169*(12), 1286-1296.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. Psychological review, 114(2), 211.

Hansen, V., Oren, E., Dennis, L. K., & Brown, H. E. (2016). Infectious disease mortality trends in the United States, 1980-2014. Jama, 316(20), 2149-2151.

Hirsch, J. K. (2006). A review of the literature on rural suicide. *Crisis*, *27*(4), 189-199.

Hoyert, D. L., & Xu, J. (2012). Deaths: preliminary data for 2011. *Natl Vital Stat Rep*, *61*(6), 1-51.

Index, G. H. W. B. (2017). Gallup Healthways Well-Being Index.

Jaidka, K., Schwartz, A. H., Kern, M., Yaden, D. B., Giorgi, S., Ungar, L. H., Eichstaedt, J. C. (in preparation). The Pitfalls of Using Twitter to Measure the Well-being of U.S. Counties: A Comparison of the Leading Methods

Jemal, A., Ward, E., Hao, Y., & Thun, M. (2005). Trends in the leading causes of death in the United States, 1970-2002. Jama, 294(10), 1255-1259.

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological methods*, 21(4), 507.

Kim, N., Mickelson, J. B., Brenner, B. E., Haws, C. A., Yurgelun-Todd, D. A., & Renshaw, P. F. (2011). Altitude, gun ownership, rural areas, and suicide. American journal of psychiatry, 168(1), 49-54.

Kuper, H., Marmot, M., & Hemingway, H. (2002). Systematic review of prospective cohort studies of psychosocial factors in the etiology and prognosis of coronary heart disease. In *Seminars in vascular medicine*. Vol. 2, No. 03, pp. 267-314.

Lexhub tutorials. (n.d.). Retrieved March 10, 2018, from http://lexhub.org/tutorials.html

McAllum, C., St George, I., & White, G. (2005). Death certification and doctors' dilemmas: a qualitative study of GPs' perspectives. *Br J Gen Pract*, *55*(518), 677-683.

Mislove, A., Lehmann, S., Ahn, Y. Y., Onnela, J. P., & Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users. *ICWSM*, *11*(5th), 25.

Murray, C. J., Kulkarni, S. C., Michaud, C., Tomijima, N., Bulzacchelli, M. T., Iandiorio, T. J., & Ezzati, M. (2006). Eight Americas: investigating mortality disparities across races, counties, and race-counties in the United States. PLoS medicine, 3(9), e260.

Ormel, J., Rosmalen, J., & Farmer, A. (2004). Neuroticism: A non-informative marker of vulnerability to psychopathology. Social Psychiatry and Psychiatric Epidemiology, 39, 906–912. http://dx.doi.org/10.1007/s00127-004-0873-y

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

Pinner, R. W., Teutsch, S. M., Simonsen, L., Klug, L. A., Graber, J. M., Clarke, M. J., & Berkelman, R. L. (1996). Trends in infectious diseases mortality in the United States. Jama, 275(3), 189-193.

Robinson-Garcia, N., Costas, R., Isett, K., Melkers, J., & Hicks, D. (2017). The unbearable emptiness of tweeting—About journal articles. *PloS one*, *12*(8), e0183551.

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*(1), 38-46.

Rugulies, R. (2002). Depression as a predictor for coronary heart disease: a review and meta-analysis1. *American journal of preventive medicine*, *23*(1), 51-61.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Lucas, R. E., Agrawal, M., ... & Ungar, L. H. (2013a, July). Characterizing Geographic Variation in Well-Being Using Tweets. In *ICWSM* (pp. 583-591).

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*(9), e73791.

Schwartz, H. A., Giorgi, S., Sap, M., Crutchley, P., Ungar, L., & Eichstaedt, J. (2017). DLATK: Differential Language Analysis ToolKit. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 55-60).

Sedoc, J., Gallier, J., Foster, D., & Ungar, L. (2017). Semantic Word Clusters Using Signed Spectral Clustering. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 939-949).

Searles, V. B., Valley, M. A., Hedegaard, H., & Betz, M. E. (2014). Suicides in urban and rural counties in the United States, 2006–2008. *Crisis*.

Singh, G. K., & Siahpush, M. (2002). Increasing rural–urban gradients in US suicide mortality, 1970–1997. *American Journal of Public Health*, *92*(7), 1161-1167.

Appendix A

How to reproduce the Eichstaedt et al, 2015. results (2009-2010 Twitter data)

1. Install DLATK:
    a. Available via pip, GitHub, conda and docker
    b. Full instructions: http://dlatk.wwbp.org/install.html
2. Download the MySQL database from OSF
    a. https://osf.io/7b9va/
3. Unzip the file
4. Upload the database with the following command:
   mysql < ~/Downloads/twitter_heart_disease_2015.sql
5. Run the following DLATK queries:
    a. Twitter and Twitter + All Predictors
        - ./dlatkInterface.py -d twitter_heart_disease_2015 -t msgs -c cnty -f 'feat$cat_met_a30_2000_cp_w$msgs$cnty$16to16' 'feat$cat_dictionaries_w$msgs$cnty$16to16' 'feat$1to3gram$msgs$cnty$16to16' --combo_test_regression --model ridge10000 --folds 10 --outcome_table county_outcomes_2009to10 --outcomes ucd_I25_1_atheroHD\$0910_ageadj --where "userwordtotal >= 50000" --group_freq_thresh 0 --controls femalePOP165210D\$10 hispanicPOP405210D\$10 blackPOP255210D\$10 smokerCHR13\$0511 diabeticCHR13\$09 obeseCHR13\$09 hsbachgradHD03_ACS3\$10 hypertenaveIHME\$09 marriedaveHC03_AC3yr\$10 logincomeHC01_VC85ACS3yr\$10 --res_controls --all_controls_only
    b. All Predictors (except Twitter)
        - ./dlatkInterface.py -d twitter_heart_disease_2015 -t msgs -c cnty -f 'feat$cat_met_a30_2000_cp_w$msgs$cnty$16to16' 'feat$cat_dictionaries_w$msgs$cnty$16to16' 'feat$1to3gram$msgs$cnty$16to16' --combo_test_regression --model ridgefirstpasscv --folds 10 --outcome_table county_outcomes_2009to10 --outcomes ucd_I25_1_atheroHD\$0910_ageadj --where "userwordtotal >= 50000" --group_freq_thresh 0 --controls femalePOP165210D\$10 hispanicPOP405210D\$10 blackPOP255210D\$10 smokerCHR13\$0511 diabeticCHR13\$09 obeseCHR13\$09 hsbachgradHD03_ACS3\$10 hypertenaveIHME\$09 marriedaveHC03_AC3yr\$10 logincomeHC01_VC85ACS3yr\$10 --control_combo_size 10
    c. Income and Education
        - ./dlatkInterface.py -d twitter_heart_disease_2015 -t msgs -c cnty -f 'feat$cat_met_a30_2000_cp_w$msgs$cnty$16to16' 'feat$cat_dictionaries_w$msgs$cnty$16to16' 'feat$1to3gram$msgs$cnty$16to16' --combo_test_regression --model ridgefirstpasscv --folds 10 --outcome_table county_outcomes_2009to10 --outcomes ucd_I25_1_atheroHD\$0910_ageadj --where "userwordtotal >= 50000" --group_freq_thresh 0 --controls hsbachgradHD03_ACS3\$10 logincomeHC01_VC85ACS3yr\$10 --control_combo_size 2
    d. All single predictors
        - ./dlatkInterface.py -d twitter_heart_disease_2015 -t msgs -c cnty -f 'feat$cat_met_a30_2000_cp_w$msgs$cnty$16to16' 'feat$cat_dictionaries_w$msgs$cnty$16to16' 'feat$1to3gram$msgs$cnty$16to16' --combo_test_regression --model ridgefirstpasscv --folds 10 --outcome_table county_outcomes_2009to10 --outcomes ucd_I25_1_atheroHD\$0910_ageadj --where "userwordtotal >= 50000" --group_freq_thresh 0 --controls femalePOP165210D\$10 hispanicPOP405210D\$10 blackPOP255210D\$10 smokerCHR13\$0511 diabeticCHR13\$09 obeseCHR13\$09 hypertenaveIHME\$09 marriedaveHC03_AC3yr\$10 --control_combo_sizes 1

Appendix B: Detailed Responses

Most of Brown and Coyne's verbatim concerns are given as bullet points and in blue, our responses are in black. We've re-organized their concerns into sections for readability.

Noise in County Variables

- Quality of heart disease mortality rates
    - A definitive post-mortem diagnosis of AHD may require an autopsy, yet 5 the number of such procedures performed in the United States has halved in the past four decades (Hoyert, 2011).
    - .. is due, not to differences in the actual prevalence of AHD as the principal cause of death, but rather to the variation in the propensity of local physicians to certify the cause of death as AHD (cf. McAllum, St. George, & White, 2005).
    - It is also worth noting that, as reported by Eichstaedt et al. in their Supplemental Tables document (2015c), the "county-level" data for all of the variables that measure "county-level" health in 10 their study (obesity, hypertension, diabetes, and smoking) are in fact statistical estimates derived from state-level data using "Bayesian multilevel modeling, multilevel logistic regression models, and a Markov Chain Monte Carlo simulation method" (p. DS7). However, Eichstaedt et al. provided no estimates of the possible errors or biases that the use of such techniques might introduce.

We agree that, like outcomes used in nearly all public health studies, there is some degree of error in the heart disease mortality rates and other outcome variables, and we noted this in the original paper. The authors are correct in that we did not model the error in the original data, as is rarely done in these studies. The citation given by Brown and Coyne is based on the qualitative output of four teleconferenced focus groups across 16 General Practitioners in New Zealand, which concludes that "Improving death certification accuracy is a complex issue," providing no clear recommendation for an alternative. The source of mortality data we used (the mortality rates from the Centers for Disease Control and Prevention's Wide-ranging Online Data for Epidemiologic Research database, or CDC Wonder for short) are widely used in research. Our analysis and hundreds of others do in fact depend on the assumption the main source of variance within these officially-reported data to be what they profess to measure, in the same way that these outcomes and estimations are used throughout medical and public health research (e.g. Pinner et al., 1996; Armstrong, Conn & Pinner, 1999; Jamal et al., 2005; Murray et al., 2006; Hansen et al., 2016).

- Selection of counties
    - Thus, the selection of counties tends to include those with higher levels of the outcome variable, which has the potential to introduce selection bias (Berk, 1983).

Yes, this may contribute to error. The counties included contain over 88% of the population, suggesting that this more representative than most psychological studies, and provide two orders of magnitude more power than state-level analyses. Still we noted the selection bias in the article.

- County level Variables
    - First, the socioeconomic climate of an area can change substantially in less than a generation

We agree that that could happen to single counties, which contributes to noise in the data. The question is does is happen to all communities at such a rate that socioeconomic variables of counties are not be trusted? We think not.

The authors also note that 3.5% of the population moves each year, and suggests that this points to considerable mobility. But a considerable portion of those are the same people – and 96.5% are staying the same. This might point to two latent types of people – those who are mobile, and those who are more stable, and perhaps a community type argument would not be appropriate for the more mobile set. Of those who are mobile, they enter a community, and often either must assimilate to that culture, or quickly are dissatisfied and move on, helping to reinforce characteristics of that community. People are also averse to change. So while communities continually evolve and change, the extent to which health promoting or health demoting aspects also change is an open question for future research.

- Counties are not communities
    - Yet there seem to be several reasons to question such assumptions. First, the size and population of U.S. counties varies widely; their land areas range from 1.999 to 144,504.789 square miles, while their 2010 populations covered five orders of magnitude, between 82 and 9,818,605.Given such diversity in the scale and sociopolitical significance of counties, we find it difficult to conceive of a county-level factor, or set of factors, that might influence Twitter language and AHD prevalence with any degree of consistency across the United States.

They do indeed vary a lot in size and are imperfect units of analysis -- they are, however, better than all the alternatives that we know, such as U.S. states, of which there are 50. At the county level, other covariates of health (such as income) show consistent relationships across U.S. counties, and we do not find it difficult to conceive that predictors like income have psychological covariates. We used several precautions to ensure the generalizability of our findings, including corrections to significance thresholds to account for multiple-hypothesis testing, and cross-validating over held-out data. Others may replicate our findings using our data and software (see https://osf.io/rt6w2/, and Appendix A).

The authors give the example of two counties (Jackson County and Clay County in Indiana), noting that a tourist driving through would see few differences, and as such, it makes no sense that language would find psychological differences. Indeed, there are few demographic

differences between these counties, and so any differences could indeed be random noise. Note that we also claim to account for 17% of the variance in heart disease mortality (prediction accuracy of r = .42), so there will be quite a few specific counties for which predictions will be noisy.

- As Beyer, Schultz, and Rushton (2008, p. 40) put it, "The county often represents an area too large to use in determining true, local patterns of disease."

We agree that they are not ideal. We would appreciate the availability of data at lower levels of spatial aggregation, but unfortunately, county-level is the smallest level at which have found both Twitter and U.S. wide covariate data to be available. Importantly, we have found reliable evidence that variables derived from tweets within the county are able to predict atherosclerotic heart disease consistently. Future research should replicate these analyses at the zipcode or census tract level, if such data becomes available. Like in any type of research, we are limited to the data available.

Noise in Twitter Data

- Twitter sampling
  - The assumption that the users who provided enough information to allow their tweets to be associated with a county represented an unbiased sample of Twitter users in that county.
  - The implicit assumption that Twitter users represent a comparable fraction of the population of each county.

We have not made that assumption, and specifically disclaim representativeness it in the 2015 article:

"Our study has several limitations (...) 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)." (p. 166)

Still, the patterns of language use on Twitter provide enough information for a statistical model to learn to predict *representative* population-level mortality rates.

- that around 7% of tweets were incorrectly mapped to counties

As with most studies there is a degree of error in the data. This, as many other sources of noise, would only make things harder to predict, so work against our ability to predict heart disease, which we demonstrate nevertheless.

- Thus, it seems likely that a substantial proportion of the people who die from AHD each year in any given county may have lived in one or more other counties during the decades when AHD was developing, and thus been exposed to different forms, both favorable and unfavorable, of Eichstaedt et al.'s purported community-level psychological characteristics during that period.

Indeed, this is another source of noise. One could also argue that people likely move among similar counties, with similar socio-economic profiles. In addition, the same concerns would apply to the relationship between county income and heart disease – people may have lived in counties with different income levels–but nevertheless we and others observe consistent relationships between county-level income and heart disease.

We reiterate the fact that Twitter users in the same county as people dying of AHD (and some users from other counties incorrectly mapped) provide a reliable enough source of county-level information to predict rates of AHD mortality, at levels equivalent to typical risk factors.

- Twitter noise
  - Whether the omission of these words from their data set is due to a choice on the part of Eichstaedt and colleagues, or the consequence of a decision by Twitter to bowdlerize the "Garden Hose" dataset.

We have not touched the Twitter garden hose data. We grant that the purported irregularities may add a source of noise to the Twitter data.

- Some Twitter users have disproportionate effect
  - A corollary of this is that, despite the apparently large number of participants overall, a very small group of voluble Twitter users could have a substantial influence in smaller counties.
- There are bots on Twitter
  - On a related theme, Robinson- Garcia, Costas, Isett, Melkers, and Hicks (2017) warned that bots, or humans tweeting like bots, represent a considerable challenge to the interpretability of Twitter data;

We have recently updated our methods to limit the effect of any single Twitter account in the analysis, by weighing them equally within a county sample. This new method is applied in the replication on 2012/2013 data and explained in Schwartz et al. (under review). These results largely correspond to the original 2009/2010 results reported in Eichstaedt et al., 2015.

We have ourselves started to wonder how big the influence of bots could be in these analyses. We have identified some bots through looking for patterns of syntactically similar language and found mostly weather bots. overall, we were unable to find any strong indication that bots (or other "super-posters") drive the signal in the Twitter data.

Method Concerns

- **How similar are the comparative maps**
  - To this end, we wrote a program to extract the colors of each pixel across the two maps, convert these colors to the corresponding range of AHD mortality rates, and draw a new map that highlights the differences between these rates using a simple color scheme.

We feel that a correlation value (r = .42) summaries the overall accuracy well.

- **"Data was not made available"**
  - but we were not able to reproduce the ridge regression results that were described by Eichstaedt et al. under the heading "Predictive models" (p. 161) because neither the code nor the data were apparently made available.

This is false. Data was shared with the publication of the paper -- see https://osf.io/rt6w2/. The supplementary methods explain that we built a ridge regression model over the language frequencies we shared. We had also first released the code to run the analysis on the website of our research group (wwbp.org) later in 2015. We have also made a video tutorial on how to analyze the data from the OSF repository in R (see http://lexhub.org/tutorials.html).

We have since re-released the data on the OSF in more accessible form, ready for database import (osf.io/rt6w2/). The code base was published and released open source in 2017 with better documentation (dlatk.wwbp.org**,** Schwartz et al., 2017). The code base allows for the reproduction of the original findings within the confidence intervals using 4 DLATK interface calls (explained in Appendix A).

- **Potential outliers**
  - This aggregation into counties calls into question Eichstaedt et al.'s claim (p. 166) that an analysis of Twitter language can "generate estimates based on 10s of millions of people"; indeed, it could be that their results are being driven by just a few hundred outliers, particularly those living in smaller counties.

We have found no indication that that is the case. In fact, when we evaluate the accuracy of our Twitter prediction model weighing counties by the square root of population in the 2012/2013 reproduction, the predictive accuracy increases from r = .42 to a $r_{weighted}$ = .52 [0.48, 0.56]. We had chosen to reported the unweighted, and thus more conservative, predictive accuracies.

- **Dropping love from the dictionary**
    - Their justification for this was that "reading through a random sample of tweets containing love revealed them to be mostly statements about loving things, not people" (p. 165).
    - In fact, it turns out that hate dominated Eichstaedt et al.'s negative relationships dictionary (41.6% of all word occurrences) to an even greater degree than love did for the positive relationships dictionary (35.8%).

We largely agree with your point. Please also see the limitations paragraph in the main response addressing the larger point about dictionaries often being unreliable (specifically the "positive relationships" dictionary).

Dropping love from the dictionary was not ideal, and came out of a back and forth with a reviewer who requested further unpacking of the dictionary correlation -- which we did in Supplemental Table S5 and in footnotes 3 and 6. This is an example why we advocate for the transparent presentation of dictionary or topic-based results (i.e., showing the most prevalent words, which we did in Supplementary Table S6). While the correlations are significant and we have now replicated these on new data, seeing which words dominate helps to interpret dictionary-based results. For transparency, we reported the correlation with *love* included in the discussion within the main body of the manuscript, in the discussion.

    - Of course, it might be true of personal relationships, or indeed any other aspect of people's lives, that those who live in lower-SES areas—or, for that matter, those who are married, or smoke, or suffer from diabetes—tend to communicate more (or, indeed, less) about that topic on Twitter. But the factor analysis in Eichstaedt et al.'s Note 6 does not provide any direct evidence for their claim of a possible relation between residence in a lower SES area and a tendency to tweet about personal relationships.

See Supplementary Table S5, bottom, showing that there are two factors of word use within the Positive Relationship dictionary, one of which correlates both with higher heart disease mortality and lower socioeconomic status, the other with lower heart disease rates and not with socioeconomic status. Future research on these questions is needed, and we are pursuing it.

- **Prediction model**
    - "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." It is not clear exactly how this model was constructed or what weighting was given to the various components, even though the numbers of each category (words, phrases, dictionary entries, and topics) vary widely.

This was all learned automatically through a standard machine learning algorithm (ridge regression) discussed in the supplemental information and validated via a standard 10-fold cross-validation technique in which the data used to test the model is not used during model fit (*training*). See Appendix A for replication instructions, which fully specify the model.

- **Splitting the U.S. into regions**
  - We noted earlier that the diversity among counties made it difficult to imagine that the relation between Twitter language and AHD would be consistent across the entire United States.

As mentioned, the machine learning prediction model is able to predict on held-out counties out-of-sample. We have now replicated this across new years of Twitter and mortality data (2012-2013).

We agree that this raises a larger general point about how best to model spatial relationships in epidemiological research contexts. We have active lines of work trying to tackle the nature of these complexities, and the potential role of spatial regression techniques (not customarily used in psychology).

- **Topics are confusing**
  - Furthermore, some of the topics that were highlighted by Eichstaedt et al. in the word clouds in their Figure 1 contain words that directly contradict the topic label
  - Taken from Facebook: Thus, the extent to which these automatically extracted topics from Facebook really represent coherent psychological 8 or social themes that might appear in discussions on Twitter seems to be questionable, especially in view of the very different writing styles in use on these two networks.'

Single, less prevalent words do on occasion seem to have an antonymic relationship with other words in the topic. Topic modeling techniques incorporate the fact that antonyms are semantically related and thus can yield such subjective discrepancies. One of our authors (Ungar) has recently worked explicitly on addressing this problem in word clustering techniques (Sedoc et al., 2017). We have not integrated such techniques here but hope to and encourage others to in the future.

Topics are simply a way of clustering word use by co-occurrence, which is relatively invariant across social media platforms (for an excellent review of topic modelling, see Griffiths, Steyvers & Tenenbaum, 2007). We had used the topics in a number of previous papers and had been impressed with their nuance and specificity (e.g., Schwartz et al, 2013a). Using these Facebook topics on Twitter has worked well for predictive purposes evaluated on held-out-data (both in Eichstaedt et al., 2015 and also in Schwartz et al., 2013a). So in terms both of specificity and use as predictive features these topics have worked well on both Facebook on Twitter. But we agree that using LDA topics modelled over Facebook on Twitter is not ideal; specifically, LDA topics may miss some Twitter specific language use patterns (such as commenting and

retweeting). Future work should reproduce the findings with corpus-general or Twitter-modeled topics.

Notice that, unlike in dictionaries, with our choice of visualization it is plain which words drive topic frequencies -- what you see is what you get.

Suicides

- county-level psychological factors that act directly on the health and welfare of members of the local community to be more closely reflected in the mortality statistics for suicide than those for a chronic disease such as AHD.
- data for self-harm were only available for 741 counties; however, these represented 89.9% of the population of Eichstaedt et al.'s set of 1,347 counties.
- Apparently the "positive" versions of these factors, while acting via some unspecified mechanism to make the community as a whole less susceptible to developing hardening of the arteries, also simultaneously manage to make the same people more likely to commit suicide, and vice versa.

Please see main response document. County-level suicide rates diverge in their correlational profile over Twitter data not only from heart disease, but also from broader CDC-reported markers of poor mental health. The correlations reported by Brown & Coyne disappear once two major suicide confounds (elevation and population living in rural areas) are controlled for.

Causal Interpretation

"potential sources of distortion and bias in its assumptions about the nature of AHD"

- findings suggesting a link between Type A behavior pattern (TABP) and cardiac events and mortality in small samples (Friedman & Rosenman, 1959), an accumulation of evidence from more recent large-scale studies has consistently failed to show reliable evidence for such an association (Kuper, Marmot, & Hemingway, 2002).

It is that there is little evidence for TABP, but there is considerable evidence that specific aspects are indeed risky (e.g., hostility, aggression). Our findings point to these aspects (especially hostility) being predictive of risk, aligned with other findings in this area. It's a mis-reading of our arguments to say we are basing findings on the Type A Personality theory.

- At best, negative affectivity is likely to be no more than a non-informative risk marker (Ormel et al., 2004), not a risk factor for AHD.

We did not claim causality, which cannot, in principal, be inferred from the cross-sectional data analysis we have conducted here, as we noted in the original article:

"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." (p. 164)

"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)." (p. 166)

We specifically tried to communicate that the language correlates may be marking risk of AHD, not causing it. We did however note that some of the patterns of correlations in Twitter language were congruent with some accounts of the covariates of heart disease at the individual-level, which have stood up to meta-analyses:

"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)." (p. 165)

The typical approach in psychological research is to anchor findings to existing results, considering what replicates and what is different. We followed this approach. There is ample space for future research.

- In contrast to TABP, socioeconomic conditions have long been identified as playing a role in the development of AHD. For example, Clark, DesMeules, Luo, Duncan, and Wielgosz (2009) noted the importance role of access to good-quality healthcare and early-life factors such as parental socioeconomic status. However, neither of those variables appeared in Eichstaedt et al.'s (2015a) model.

We have included the strongest correlates of AHD for which data was available at the county level. Future research is always encouraged. Substantively, we presume both high-quality healthcare and parental SES to be moderately to highly correlated with current education and income levels of the counties, and thus included in the analyses by proxy.

Summary

- Summary & Discussion
  - The coding of AHD as the cause of death is subject to major variability;

We agree, this contributes to noise, see main response.

- the process that selects counties for inclusion is biased;

We agree, this contributes to noise, see main response.

- the model "predicts" suicide better than AHD mortality but with almost opposite results (in terms of the valence of language predicting positive or negative outcomes) to those found by Eichstaedt et al.;

We agree, see discussion in main response. Suicides are uncorrelated with heart disease, and the correlations the authors report disappear once elevation and population living in rural areas are controlled for. They are not a straight-forward measure of county-level psychological health. The broader CDC-reported county-level poor mental health variable correlates with psychological language categories in the same way as AHD.

- the Twitter-based dictionaries appear not to be a faithful summary of the words that were actually typed by users;

We disagree. We have not touched the Twitter data, see above and we have reported the leading correlated words for both topics and dictionaries. We have reproduced the findings over a different Twitter sample across different years, see main response.

- arbitrary choices were apparently made in some of the dictionary-based analyses;

In one case "love" was removed from one dictionary, which was explained in multiple places in the manuscript, and we reported the original correlations observed for the dictionary (without removing "love") in the discussion of the manuscript. See main response for further discussion about dictionaries.

- there are numerous problems associated with the use of counties as the unit of analysis;

We somewhat agree, but it is better than all alternatives we know, see above. We noted this in the original paper and were not claiming that this is an ideal final level of analysis.

- and the predictive power of the model, including the associated maps, appears to be questionable.

We disagree strongly. See main response for replication across different years, and the step-by-step replication guide in Appendix A.

- A more parsimonious explanation is that there is a very large amount of noise in the measures of the meaning of Twitter data used by Eichstaedt et al., and these authors' complex analysis techniques (involving, for example, several steps to deal with high multicollinearity)…

While the approaches that we used are somewhat more complex than basic multiple linear regression, they are standard machine learning techniques that are widely employed, fully understood, and can easily be reproduced. See Appendix A.

- …are merely modeling this noise to produce the illusion of a psychological mechanism that acts at the level of people's county of residence.

We strongly disagree, and do not follow what illusion we were trying to create. Predictions were done on held-out data (i.e., the model was trained on part of the data, and tested with another part, which is what yielded prediction accuracies), and can be replicated across multiple years (see main response). The non-Twitter county variables we use in the analysis are widely used in research – indeed, we purposely used data from the CDC, U.S. Census and similar official data sources to align with the use of data in other research (especially in sociological, epidemiological, and public health sectors; e.g. Pinner et al., 1996; Armstrong, Conn & Pinner, 1999; Jamal et al., 2005; Murray et al., 2006; Hansen et al., 2016).

- Jensen argued that "the extent of overlap between individuals' online and offline behavior and psychology has not been well established, but there is certainly reason to suspect that a gap exists between reported and actual behavior" (p. 2)

There is varied evidence on overlap between individuals' online & offline behavior and character, with some finding good overlap and others finding more variation. Most likely this varies by the user. Self-presentation and social-desirability biases all work against us, and yet we still found effects.

- "the principal claim"
    - The principal theoretical claim of Eichstaedt et al.'s (2015a) article appears to be that the best explanation for the associations that were observed between county-level Twitter language and AHD mortality is some geographically-localized psychological factor, shared by the inhabitants of an area, that exerts a substantial influence on aspects of human life as different as vocabulary choice on social media and arterial plaque accumulation, independently of other socioeconomic and demographic factors.

We do not claim independence of psychological markers from socioeconomic and demographic factors (and our principal claim was stated in our title, "Psychological Language on Twitter Predict County-level Heart Disease Mortality"). A secondary claim that relates to the above is given at the beginning of the discussion:

> "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." (p. 166)

We very much stress that what we measure are markers for (and *not* independent of) other ecological variables (such as income and education). As stated at the end of the results section in the paper:

> "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." (p. 164)

Conclusion (Detailed Reponses)

In conclusion, we appreciate the scrutiny of our work by Brown and Coyne (2018) and the opportunity to further discuss the work as well as release a user-friendly replication guide (see Appendix A). In their critique of our study, we were not able to identify any previously unacknowledged weaknesses of substantial import.

Supplementary Material

Table S1

Data Sources for 2012/2013 County-level Variables

Variable				Description of variable	Years covered	Source
Included variable	Transformation	Categories	MySQL variable name			
Atherosclerotic Heart Disease (AHD) mortality	averaged across years		ucd_i25_1_1213_ageAdj	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)	2012-2013	CDC Wonder, Underlying Cause of Death
Income	averaged across years, log transformed		log_med_income_2012to13_avg	Median household income	2012-2013	Small Area Income and Poverty Esitmates, obtained through County Health Rankings (CHR, 2014 and 2015)
High School			high_school_2012to13_chr16	Attainment of high school graduation or higher	2012-2013	EDFacts, obtained through County Health Rankings (CHR, 2016)
Bach Degree			bach_degree_2010to14_chr16	Attainment of bachelor's degree or higher	2010-2014	American Community Survey, obtained through County Health Rankings (CHR, 2015)
Diabetes	averaged across years		diabetes_2012to13_avg	Adults (age 20+) diagnosed with diabetes	2012-2013	CDC Diabetes Interactive Atlas. obtained through County Health Rankings (CHR, 2016 and 2017)
Obesity	averaged across years		obesity_2012to13_avg	Body Mass Index >= 30, based on selfreported height and weight	2012-2013	CDC Diabetes Interactive Atlas, obtained through County Health Rankings (CHR, 2016 and 2017)
Smoking			smoking_2014_chr14	Current adult smokers who have smoked >= 100 cigarettes in their lifetime	2014	County-level estimates based on CDC's Behavioral Risk Factor Surveillance System (BRFSS) data, obtained through County Health Rankings (CHR, 2016)
Hypertension	averaged	male female	hypertenaveIHME$09	Male adults (age 30+) who self-reported systolic BP of at least 140mm Hg and/or selfreported taking medication Female adults (age 30+) who self-reported systolic BP of at least 140mm Hg and/or selfreported taking medication	2009	County-level estimates prepared through the Institute for Health Metrics and Evaluation (IHME; 2009) on the basis of CDC BRFSS data (see note).
% Black	averaged across years		perc_black_2012to13_avg	Population of one race - Black or African American alone	2012-2013	U.S. Census, Demographic Profile Data, obtained through County Health Rankings (CHR, 2014 and 2015)
% Hispanic	averaged across years		perc_hispanic_2012to13_avg	Hispanic or Latino	2012-2013	U.S. Census, Demographic Profile Data, obtained through County Health Rankings (CHR, 2014 and 2015)
% Female	averaged across years		perc_female_2012to13_avg	Female	2012-2013	U.S. Census, Demographic Profile Data, obtained through County Health Rankings (CHR, 2014 and 2015)
% Married	averaged	male female	marriedaveHC03_AC3yr$10	Male adults (age 15+) now married (not separated) Female adults (age 15+) now married (not separated)		ACS 3-Year Estimates (Table DP02)

Table S2

Cross-Correlations between Dictionaries & Topics (taken from Eichstaedt et al., 2015, Table S3)

	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**								
Hostility, Aggression — bullsh*t	.94	.58	.43	.62	.19	-.03	-.45	-.40
Hostility, Aggression — a**hole	.93	.62	.48	.61	.19	.00	-.41	-.39
Hostility, Aggression — retarded	.81	.65	.56	.54	.21	.06	-.26	-.30
Hate, Inter-personal Tensions — hating	.88	.74	.54	.68	.23	.13	-.33	-.36
Hate, Inter-personal Tensions — drama	.87	.67	.53	.66	.26	.18	-.28	-.29
Hate, Inter-personal Tensions — passion	.67	.84	.66	.60	.33	.37	.02	-.08
Boredom, Fatigue — bored	.70	.60	.47	.87	.20	.16	-.26	-.35
Boredom, Fatigue — tired	.69	.70	.62	.87	.31	.32	-.04	-.21
Boredom, Fatigue — bed	.50	.61	.56	.69	.30	.41	.08	-.12
Skilled Occupations — management	-.42	-.32	-.23	-.41	.03	.29	.38	.69
Skilled Occupations — service	-.41	-.28	-.17	-.39	.08	.33	.51	.63
Skilled Occupations — conference	-.45	-.28	-.16	-.42	.11	.34	.56	.65
Positive Experiences — experience	-.30	-.12	-.01	-.26	.15	.42	.57	.76
Positive Experiences — company	-.30	-.12	.11	-.21	.18	.54	.78	.67
Positive Experiences — weekend	-.35	-.11	.09	-.22	.14	.55	.89	.62
Optimism, Resilience — opportunities	-.33	-.20	-.12	-.31	.10	.35	.41	.69
Optimism, Resilience — achieve	-.21	-.07	.00	-.22	.17	.36	.39	.68
Optimism, Resilience — 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 Eichstaedt et al, 2015, in the same order. The "included words" are dominant unique words in each cloud, which help identify the topic. † In these correlations, the word "love" was removed from the dictionary.