Response to Rebuttal Letter

Original editor and reviewer comments highlighted in yellow, author responses in black, and new Johannes Eichstaedt responses in red.

I fully agree with the changes requested by the editor in the last revision, and, in the interest of efficiency, will focus my discussion on these necessary changes. The changes previously requested by the editor need to be addressed before this manuscript is suitable for publication.

Of particular concern is the misrepresentation in the title and abstract that this manuscript provides a "re-analysis" of the 2015 paper (which has machine learning prediction as its key contribution), and that no evidence could be found that Twitter predicts heart disease, given that the authors have now been able to reproduce the predictive model themselves (which they fail to adequately discuss in more than a footnote).

Editor's Comments

It is important that published studies are checked critically, and the manuscript has the potential to make an important contribution to the field, as also reflected by the comments of the reviewers. However, two main areas need to be revised before the manuscript can be considered for publication.

1) The relation to the original paper should be clarified. The manuscript comments on the data that was used, the analysis of this data and presents additional analyses. Eichstaedt et al. (2015) used data rather than acquiring the data themselves. Although the concerns on the reliability of this data are important, they do not specifically apply to the original article but to the field in general and should hence be discuss accordingly.

The manuscript in its current form does not adequately address this fundamental point about the framing of the article. The manuscript does not integrate with the existing literature and fails to acknowledge the larger context for methods and data used in the 2015 article, seemingly in an attempt to maximize the appearance of controversy. It is not suitable or publication in the current form.

The assessment of the analyses in Eichstaedt et al. (2015) is most relevant to the aim of the current manuscript. Some concerns that are raised are quantitatively tested, such as the inclusion of 'love', and this is insightful as pointed out by reviewer 3, while other concerns are not tested, e.g. inclusion of additional variables such as access to good-quality healthcare, socioeconomic status and Gini coefficient of income equality and the role of outliers. As the data and code is available these concerns should be quantitatively tested.

As the authors have now been able to reproduce the models and run our code (see Footnote 1), it is relatively easy to include additional variables in the analyses. As suggested by the editor, the authors are therefore very much in a position to empirically test the role additional variables may play, as is appropriate for an empirical article portending to be a "re-analysis" of previous work.

In its current form, the empirical results of the article do not support its rhetoric.

Most critically, the title "No Evidence That Twitter Language Reliably Predicts Heart Disease: A Reanalysis of Eichstaedt et al. (2015a)" is simply false, as the article does not report on a reanalysis of the core contribution of the 2015 article (the machine learning out-of-sample predictive results, to which the title of our article refers). It is not at all clear how they jump to the conclusion that there is "no evidence that twitter language reliably predicts heart disease" when they have themselves now been able to reproduce this result. The authors are now in a position to report on such a replication after getting our analysis code to work on their system. The title should be changed to accurately reflect the content of the article (see first editor comment).

Relatedly, the abstract is highly misleading, as it suggests that a re-analysis has been carried out where as their argument relies on a speculative alternative analysis (line 28: "Next, using the data files supplied by Eichstaedt et al., we reanalyze their numerical results, including testing their model with mortality from an alternative cause of death, namely suicide.") It should be changed to adequately represent the fact that the article does NOT present results of a reproduction of the original analysis. Further, the use of "model" is highly misleading, suggesting to the reader, inaccurately, that they used our specific statistical model for prediction rather than conducting a different statistical analysis simply on the relationship between specific language categories and suicide mortality

If this is not possible because the relevant data is not available, this should be clearly stated and speculations should be kept to a minimum. The additional analyses that are presented (relation with suicide) may help to put the original findings into context, but the robustness of this finding to confounding variables needs to be addressed.

The authors fully fail to address this point in any substantive way (see below for a more extensive discussion offered in the first review round). They instead appear to pursue a rhetorical strategy summarized by them as below:

"Concerning suicide, our point is not to claim that Twitter language can actually predict suicide. Rather, we show that replacing AHD mortality with suicide as the outcome variable in Eichstaedt et al.'s model produces a narrative that, we believe, appears to be at least as convincing as the original article. In fact we believe that in both cases the models are probably confounded by both unmeasured variables and noisy measurements of those that are

measured (e.g., the wide variability of county-level AHD mortality rates, and the considerably imprecision in some of the predictors such as smoking), with these confounds being similar for both outcomes. We are not sure how reviewer #4 arrived at his choice of additional variables to include in the regression to get the coefficients for suicide below the threshold of statistical significance (for example, we suspect that very few counties have a mean elevation sufficient to induce noticeable hypoxia, which reviewer #4 suggests might be driving increased levels of suicide), but we do not consider this especially relevant to our argument. We derived our figures for suicide using exactly the same correlational and OLS regression models as Eichstaedt et al., and we have not seen any evidence to explain why those results should be any more or less robust than the ones concerning AHD>'

This comment expresses the crux of their rhetorical strategy in its current form:

(a) Present an insufficiently developed analysis of (spurious) suicide correlations
(b) Suggest that a narrative could be told about these findings that resembles the heart disease findings
(c) Conclude that the heart disease findings are also spurious

This argument fails on a number of methodological grounds that are *critical* to empirical science. The heart disease findings were very carefully developed in light of the existing literature on the epidemiology of heart disease. The covariates included in the analyses were the strongest ones the existing literature pointed to. Additional covariates which have been missed by the original authors can be tested with the data and the code that has been released. The heart disease findings were developed to the best of our abilities in light of the existing literature. If the authors wanted to replicate our approach on suicide, then they should start with an adequate literature review on the epidemiology of suicide mortality, and they would quickly see that is uncorrelated with heart disease, and has different ecological correlates within the U.S. (namely rural and elevation) and is therefore not very comparable to heart disease (or rates of mental unhealth, see our preprint) at the ecological level.

Additional heart disease variables

The authors need not take our word for it, we encourage them to consider the heart disease literature and *empirically test* any additional variables that might be important. This is how science advances: through empirical tests, not through unsupported speculation.

Additional suicide variables

There exist highly relevant previous work on the county-level correlates of suicides (Kim et al., 2011; Brenner et al., 2011; Hirsch, 2006; Searles et al., 2014). It took us less than a couple hours to acquire the relevant variables and show that the correlations reported in the manuscript are fully accounted for (confounded by) variables well known to the suicide literature. If the

authors are not willing to engage with the empirical basis of their findings even to this level, their article is unsuitable for publication, as it purports itself to be an empirical analysis.

Please review:

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.

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.

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

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*.

The concern that additional lurking variables may drive the observed correlational findings is trivially true of all correlational research. As empirical scientists, we expect ourselves (and one another) to do what is in our power to identify and test such variables. In the case of heart disease, income and education are the strongest known correlates, and so we control the language findings for income and education in the original article. We encourage other scientists to explore additional variables, and offer well-developed empirical extensions (or refutations) of our work. I expect the same basic level of rigor of the authors.

The manuscript shows no such rigor or substantive engagement with the nature of the phenomena they have chosen to report analyses about.

Instead, the authors pursue a rhetorical strategy of "smearing the heart disease findings by means of analogy with their own shoddy analysis." This is not a means of carrying out empirical science, this is a means of maximizing controversy while lacking the empirical basis to support their claims.

So, before this article is suitable for publication, I ask that the authors fully address the editor's comments:

- Test and report additional confounding variables they think may drive the heart disease findings, which they are now in the position to do. Alter the manuscript accordingly to remove speculation unsupported by such analyses.

- Test additional confounding variables that likely drive the suicide findings, based on the existing county-level suicide literature (or our pre-print).

No speculation is appropriate where empirical tests are easily within the reach of the authors.

In its current form the article's empirical substance does not support the claims made by it ("re-analysis"), and is unsuitable for publication.

Response:
Concerning the testing of other variables

- For healthcare and parental SES (line ~100), we already noted in our original manuscript that these variables, which would have been desirable to include as predictors of AHD, were not present in Eichstaedt et al.'s dataset. We have added a few words to make it clear that these would be difficult to measure at a county level, but we do not believe that it is speculative to mention that the original authors' analyses did not include variables that are considered to be important predictors of their main outcome measure. It is they, not we, who have chosen to build a model out of county-level variables, the meaning of which is not always easy to establish.

It is the authors who have undertaken the project of critiquing the original analysis, and framed it as an empirical re-analysis. Given that the additional variables are well within your ability now to test as confounds, it is only appropriate that you do so, so this article is a contribution to empirical science and not just a speculative commentary.

2) The writing and structure of the manuscript should be revised. The wording should be more factual and the original study should be accurately discussed. Assumptions on the intentions of original authors that are not based on the article (e.g. Type A personality) and insinuations (violation of Psychological Science standards) should be removed. Misrepresentations of the original text (e.g. causal interpretation) should be corrected. The authors should also try to minimize redundancies in the manuscript.

Response:
We have made a number of changes at points where individual reviewers thought that our treatment of the original article was either inaccurately or infelicitously phrased. We have also spontaneously removed several points, not mentioned by the reviewers, where we felt that our language could have been misconstrued, or where we were factually wrong (e.g., the point on lines 161–165 of the original manuscript, which on a close re-reading of Eichstaedt et al.'s article we discovered they had in fact addressed).

In its current revision, the authors have only offered minimal, semantic changes, with the same misrepresentations stated in different form. This continues to include a claim about the article being based on Type A personality, and an overstatement about the claims made in the 2015 article about causality. The findings in the 2015 article are stated clearly in observational, correlational language ("markers") in the results section. I ask the authors properly reference the stated results.

We believe that it is appropriate to retain the discussion of **Type A personality** because, although Eichstaedt et al. did not mention this construct (a fact we now acknowledge explicitly), we feel that many readers of their article will have the idea that "angry / hostile / stressed people have more heart attacks" somewhere in their mind. Indeed, this association is the subject of the first sentence of Eichstaedt et al.'s abstract, although they only provided one reference to support it (Chida & Steptoe, 2009), a meta-analytic review article that notes that "In the studies fully controlled for possible behavioral covariates, there were no significant associations between anger and hostility and CHD in either disease or healthy studies" (p. 942). (Incidentally, the OR of 1.22 for anger cited without qualification by Eichstaedt et al. from Chida and Steptoe's article referred to *men*, not to all patients, for whom the OR was 1.19 [1.05, 1.35].) On a more general note, we feel that Eichstaedt et al.'s underlying theory is rather poorly articulated—for example, it really is not at all clear what "county-level psychological characteristics" might mean in theoretical terms, given (among other issues) the wide diversity in the nature of counties—so that any interpretation will require the reader to make some assumptions regarding the effects of negative affectivity, in its widest sense, on cardiovascular health.

Type A personality is not a theory endorsed in the 2015 article. It is an outdated theory that has received mixed empirical support. It's invocation by the authors as a framing of what is stated in the 2015 appears to be purposefully used as a straw man. There is strong empirical support for the link between hostility and cardiovascular hearth disease, established by a number of meta-analyses. As petty as it seems to have to comment on this, their discussion above of the Chida and Steptoe meta-analysis is misleading, it's conclusion in the abstract clearly states that: "The current review suggests that anger and hostility are associated with CHD outcomes both in healthy and CHD populations." (Chida & Steptoe, 2009).

I ask that the authors correctly report that the 2015 article merely claims that a connection between CHD and anger and hostility has been previously established, and avoids the straw man of invoking Type-A personality theory. I also ask that previous literature is correctly represented in its main findings.

Turning to the discussion of suggestions of **causality**, we would first note that we did not use this word (or any variant thereof) in our manuscript, except in a quote from another author in the discussion of TABP. From the context of Reviewer 4's comments, it seems that he is specifically objecting to what he sees as the implied claim of causal inference in our phrase "some geographically-localized psychological factor ... that *exerts* [emphasis added] a substantial influence on aspects of human life as different as vocabulary choice on social media and arterial plaque accumulation". This objection seems to be based on the fact that Eichstaedt et al. included a disclaimer about causality in their article. However, we feel that this disclaimer did not adequately compensate for some of Eichstaedt et al.'s use of language elsewhere in their article, such as "Local communities *create* [emphasis added] physical and social environments that *influence* [emphasis added] the behaviors, stress experiences, and health of their residents" (both italicized words here seem to us to imply causation at least as

strongly as our word “exerts”), and “Our approach ... 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” (the last part of which seeming to strongly imply that an intervention to change A, the psychological factor, is expected to lead to a change in B, cardiovascular health).

In its current form, the representation of the 2015 article clearly and apparently purposefully overstates the statement it makes about causality. At minimum, please acknowledge that the 2015 article explicitly disclaimed causality, and reports findings in unambiguous correlational language in the results section (“markers”). Specificaclly, the 2015 articles summarizes its results as following:

“"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)

Since the authors submitted their manuscript to PeerJ, the original authors have posted a preprint (http://doi.org/10.17605/OSF.IO/P75KU), responding to the critique in the present manuscript that was also posted as a preprint. As this preprint is highly relevant to the current manuscript, it should be cited to ensure the current manuscript is up-to-date when published at PeerJ. Although the authors do not need to provide a detailed response to all points raised in Eichstaedt et al. (2018), which I would advise against, they should revise their comments on the availability of data and code and address the role of confounding variables in the analysis of suicide.

Response:
We have referenced and cited Eichstaedt et al.’s new preprint, including an acknowledgement that we were able to download and install their modeling software.

These facts are not acknowledged in the body of the manuscript (only in a footnote), nor are results of running the code discussed.

Re. the question of confounding variables in the analysis of suicide, please see our previous replies.

Regarding the incompletely developed suicide analyses, please see my response above and original response from the first review below.

The comments above directly reference the comments of the editors after the previous review, and constitute the reasonable minimum of edits that need to be fully addressed before this manuscript should be published. Below are my original comments from the first review rounds, which in part expound on the points above in more detail (see suicide analysis), and raise additional points not yet sufficiently addressed.

Reviewer 4 (Johannes Eichstaedt)

Basic reporting

Note about this review

Fundamentally, their critique is not a reanalysis of our work, contrary to the title of the manuscript, and the claims made in the abstract. The authors' critique does not attempt a replication of our claim, implied in their title, that county-level Twitter language predicts county-level heart disease rates.

Instead, it contains a different analysis of language correlates of county-level suicide rates, which the authors claim should match the county-level correlates of heart disease mortality. In our response (Eichstaedt et al., 2018), we found suicide rates to be are uncorrelated with heart disease rates and their linguistic correlates to mostly 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 (see Eichstaedt et al., 2018).

These concerns have not been addressed.

The manuscript would be better framed as related work, or developed into an even-handed commentary on social media-based or epidemiological methods more generally. More details are given below.
C.f. the first editor's note. This concerns have been sufficiently addressed.

The most important limitations are flagged in the experimental designs section.

Concerns about basic reporting -- Summary

In terms of basic reporting, I would like to see the following shortcomings addressed before a

possible publication of the manuscript. (a) **Claims made in Eichstaedt et al., 2015 are misrepresented;** (b) the discussion of the sources of noise fails to acknowledge the principle claim of Eichstaedt et al., 2015, namely, that prediction models can use **Twitter language encodings to predict heart disease mortality out-of-sample**, that is, in an experimental set up in which sources of noise would work against (and not for) the ability to predict heart disease rates; (c) **concerns about different kinds of noise in county-level variables and Twitter data are speculative and one-sided**, lack empirical estimates, and do not properly acknowledge the large literature using these data sources.

These concerns have not been addressed.

(a) Misrepresentations of the original article

There are a number of misrepresentations in the manuscript, which are identified in detail in Appendix B of Eichstaedt et al., 2018. The most severe ones are flagged here.

Claims of independent psychological causation

The authors allege that:

"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." (p. 26, lines 599ff)

The 2015 article did do not claim independence of psychological markers from socioeconomic and demographic factors (and the principal claim was stated in our title, "Psychological Language on Twitter Predict County-level Heart Disease Mortality"). Regarding the authors' claim, the article stated at the end of the results section:

"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)

The article also specifically disclaimed making causal inferences, as it was purely cross-sectional:

"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)

Please clearly state the claims made by the 2015 article throughout the manuscript – psychological markers measured through Twitter mark, and in large part overlap in variance, with classical predictors, among them predominantly income and education (see Figure 2 in Eichstaedt et al., 2015).

Response:
Eichstaedt et al. used the terms "risk factor[s]" and "protective factor[s]" on multiple occasions in their article to describe aspects of Twitter language, whereas they used the term "marker" just twice. These two concepts have distinct meanings in epidemiology (cf. Kazdin AE, Kraemer HC, Kessler RC, Kupfer DJ, Offord DR. Contributions of risk-factor research to developmental psychopathology. Clinical Psychology Review. 1997 Jan 1;17(4):375-406); in particular, the term "marker" is preferred for explicitly non-causal relations (cf. Kazdin et al., p. 382).

However, that is perhaps a question of semantics. The point of our sentence ("The principal theoretical claim...") is that our understanding of the purported county-level psychological factor is that it (a) is reflected in whether people use positive or negative language on Twitter and (b) influences rates of AHD mortality. We believe that this is a reasonable summary of the paragraph on p. 166 that begins with "Given that the typical Twitter user is younger...", especially the sentence "Local communities create physical and social environments that influence the behaviors, stress experiences, and health of their residents".

It is not a reasonable summary. It is purposefully misleading. Please clearly state what the 2015 article states – particularly line 579 ("independently of other socioeconomic and demographic factors") is an outright falsehood that's clearly contradicted by a sentence in the results section ("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.") and the final sentence in the abstract ("Capturing community psychological characteristics through social media is feasible, and these characteristics are strong markers of cardiovascular mortality at the community level.")

We also note that Eichstaedt et al.'s article concludes with the claim that their findings "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" (p. 166). It seems to us that a claim that psychological factors "should become the focus of intervention" strongly implies that the authors believe that improving those factors will have a beneficial effect on cardiovascular health; indeed, such an intervention would appear to make little sense if it was thought that a different causal structure obtained.

This is the last sentence of the discussion section, pointing to possible future impact of this type of research. It is a misrepresentation to imply that the 2015 claims this as a result. The results of the 2015 article are clearly stated in abstract and results section.

Hence, we do not believe that our description of a factor that exerts an influence on cardiovascular health and Twitter language is an unreasonable characterization of Eichstaedt et al.'s claims.

It appears that the authors purposefully seek to overstate the places in which causality is referenced in the 2015 article, and fail to acknowledge the careful correlational language used in results section and abstract.

Type A personality

The authors suggest in their introduction (p. 4) that the claims made in the 2015 article hinge on the empirical adequacy of Type A personality theory, and then continue to discuss the mixed findings that Type A personality theory may have received. Our 2015 article did not mention Type A personality theory – it observed a set of correlations at the community level, and, in the discussion, compared them to what has been observed at the individual level. Specifically, it referenced individual-level associations with depressed mood, anger and anxiety:

"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)." (p. 164)

The authors should properly represent the claims made in the 2015 article (about hostility, depressed mood, anger and anxiety), and not suggest that the analysis is based on Type A personality theory, or require any specific personality theory to be correct.

Response:
It was certainly not our intention to suggest that the claims of the 2015 article hinged directly on the empirical adequacy of Type A personality theory or Type A behavior pattern (TABP). We have added some words to acknowledge that Eichstaedt et al. did not mention TABP.

Nevertheless, we consider the discussion of TABP to be relevant, as the popularly perceived association between hostility and cardiovascular disease at an individual level is likely to be salient in the minds of readers of Eichstaedt et al.'s article.

See comments above. TABP appears to be purposefully invoked as an implicit claim of the 2015 article based on the mixed evidence it has received—a rhetorical straw man. Please restate your summary of the background given in the 2015 article in terms used in the article, e.g., the link between hostility and cardiovascular disease.

Unavailability of data

The authors claim that "However, as far as we have been able to establish, Eichstaedt et al. did not provide any of the code needed to reproduce their analyses (…)" (e.g., page 11)

The 2015 author team released both the county-level (a) Twitter language and (b) outcome data in a way that allowed people with some effort to reproduce the 2015 findings (county-level topic, dictionary, and 1-to-3-gram frequencies, see https://osf.io/rt6w2/). An early version of the code base was released on the research group's homepage (wwbp.org) later in 2015. Since then, usability and documentation has been improved and it has been published and released open source in 2017 (Differential Language Analysis ToolKit, dlatk.wwbp.org; Schwartz et al, 2017). Additional step-by-step instructions to reproduce the original prediction accuracies can be found in Eichstaedt et al., 2018, Appendix A.

The authors should acknowledge that the code has be released publicly since the publication (in both 2015 and 2017), and that they made no attempt to contact the 2015 author team, who could have easily pointed them to these resources (as they already have to others).

Response:
We have removed the discussion about the availability of the code, as it was not of particular relevance to readers of *PeerJ*. We have also added a footnote to clarify the situation for people who may have read our preprint.

We experienced quite a few problems when trying to install the DLATK software, although in the end we got it to work (it seems to take about 18 hours of CPU time to run each model on our modest 64-bit CPU). We have been keeping extensive notes of the problems that we encountered and will gladly supply these to the authors once we have finished the current review process. (Indeed, we feel that there is the potential, in the not too distant future, for a fruitful discussion of the issues involved when results in psychology depend on large, complex, software-based models.)

We are delighted that you have managed to install the software.This is indeed research based on modern machine learning methods, which take CPU cycles. Note that that the key machine learning models are implemented by standard libraries (scikit-learn) widely used in research, and that it is not necessary to use our code base to reproduce the machine learning models.

(b) Misunderstanding about the role of noise/bias in out-of-sample Evaluation

The manuscript in its current form fails to acknowledge the fact that the principal claim that Twitter language encodings can be used to predict heart disease rates are determined out-of-sample, that is, in an experimental set up in which the prediction models are only evaluated on "test" counties not used during model fit ("training"). In other words, in our work, language patterns that generalize across both training and test counties contribute to correct predictions – and sources of noise in the data would likely degrade the ability of the models to find patterns that generalize across all counties.

The current discussion of sources of error, bias and noise in the manuscript (including the unrepresentativeness of Twitter users, Twitter bots, county conditions changing over time, potential unreliability of death certificates, etc.) is framed rhetorically to suggest that the possible sources of error make our finding that Twitter predicts AHD less reliable or robust. But these sources of noise no not disagree in substance with the results presented in the 2015 article, despite a rhetorical framing to the contrary used throughout the manuscript.

Instead, the 2015 article establishes empirically how much of the variance in heart disease can be predicted *despite* all these sources of error (which are mostly acknowledged in the original manuscript). In other words, the manuscript lists possible reasons why the out-of-sample prediction accuracies reported in the 2015 article are as relatively low as they are (out of sample accuracy r = 0.42) – accounting for 17% of the variance in heart disease--but offers no substantive critique of the methods to create these predictions, nor attempts a replication.

Please properly state the effect of the possible sources of noise/bias/error in light of the out-of-sample evaluation methods used in the 2015 article, and contextualize them in light of the relatively modest claims about how much of the variance in heart disease can be predicted used Twitter language. And please remove suggestions that the 2015 original analysis either claimed or claimed "implicitly" (line 161, p. 8) that the Twitter data sources are representative.

Response:
The use of out-of-sample validation is certainly likely to be helpful in preventing overfitting, but it does not tell us anything about the external validity of the model. For example, if (as we believe) a substantial part of the variance in per-county AHD mortality is explained about the practices of certifying physicians, it might be that the model is picking up on that (perhaps physicians in "angrier" counties are more inclined to believe in a causal link between TABP and AHD, for example).

We are not sure the authors understand our point. The out-of-sample validation empirically establishes an "upper bound" of how much various sources of noise speculated upon by the authors can indeed affect the predictions, as sources of noise work against, not for, the ability to predict the population mortality rates.

In other words, the following speculations on sources of noise are implicitly addressed by the out of sample evaluation -- the unrepresentativeness of Twitter users, Twitter bots and county conditions changing over time, and migration trends.

We would much appreciate it if the authors could acknowledge the use of out-of-sample evaluation in their manuscript.

The assumption that the CDC-reported mortality rates do not adequately represent the variance in heart disease rates is the only one that may limit external validity. In line with the first comment by the editor, as these mortality rates are widely used in research, their potential unreliability should be even-handedly discussed in the context of the existing literature.

(c) Speculative & one-sided critiques

The authors note that there are various sources of noise in the geo-located Twitter data and the county-level outcome data. In their role as a refutation of the claims made by the 2015 article, they are insufficiently contextualized considering the out-of-sample evaluation used in the analysis (see section (b) above).

As a commentary on social-media based and spatial / epidemiological methods more generally they lack (1) empirical estimates or estimates of the relative importance in their effect on these kinds of analysis, and more importantly, (2) an even-handed treatment of the large literature that uses these kinds of data, spanning computer science, public health and epidemiology among others. This includes critiques of big data and Twitter methods that have been published previously.

This has not been adequately addressed.

…Regarding Twitter Data

For example, as was noted in the 2015 article, users who tweet are not representatively selected, and some of the tweets (7%) are incorrectly mapped to counties. Further, some people may 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, etc. These critiques are valuable, but in their current form are insufficiently integrated with the rest of the literature. A large number of publications has used geo-tagged Twitter data for all sorts of applications including health predictions—the 2015 article is far from the only article using these methods.

e.g.

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.

Or, for critical reviews of these methods consider:

Pavalanathan, U., & Eisenstein, J. (2015). Confounds and consequences in geotagged Twitter data. arXiv preprint arXiv:1506.02275. (Published at EMNLP)

If a critical appraisal of social media-based methods is included, please provide a more balanced survey of the use and discussion of geo-tagged Twitter data in current research.

Response:
The fact that other authors have used data in a similar way does not in itself establish the validity or lack of bias of such methods, either in general or for any other specific case. While a more extensive discussion of the pros and cons of using data derived from social media would perhaps be interesting, we feel that it would be beyond the scope of our manuscript, in which the use of Twitter data is only one of several
.
I disagree, it is not beyond the scope of this manuscript to adequately situate your critique in light of previous work. Science builds across publications. See first editor comment.

…Regarding heart disease data

The authors doubt that the coding on death certificates for underlying cause of death is reliable (p. 13).

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 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, 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.

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.

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.

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.

Response:

Of these four references, two describe infectious diseases, which we believe are typically considerably easier to diagnose than a specific one out of the many forms of heart disease (AHD, code I25.1, is one of around 100 ICD-10 codes for "Ischemic heart disease"), and the other two (only one of which considers county-level data) consider all forms of heart disease together, whether as "heart disease" (Jemal) or "cardiovascular disease" (Hansen). None of them addresses the difficulties associated with specifically identifying AHD as the principal cause of death, or the wide range of prevalence of this diagnosis. We continue to believe that the enormous range of diagnoses of AHD that we identified is very likely to be an indication of lack of reliability in the measurement of Eichstaedt et al.'s principal outcome variable.

Please properly consider and acknowledge the use of these CDC-repported variables in other work, as requested by me and the editor.

Please acknowledge the wide use of these data sources in health research, and if a critique of the use of such data sets in epidemiological research is intended, please provide an even-handed accounting of the reliability of these data in previous research.

This has not been addressed.

Experimental design

Main Analyses

The manuscript claims to be a re-analysis of the 2015 Eichstaedt et al. paper in abstract and title. The 2015 paper had two major sections: (a) an evaluation of Twitter to cross-sectionally predict heart disease mortality, and (b) an exploration of the language correlates of heart disease mortality. This manuscript attempts a re-analysis of neither and thus its framing is inappropriate. (…)

However, I remain wholly unconvinced about (2), namely that the statistical analysis about suicide rates presented in the manuscript are meaningfully related to the heart disease correlations presented in the 2015 article.

Specifically, the authors connect the suicide analysis to the heart disease analysis through the following theoretical assumption: “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 6).

However plausible this may appear at the individual level, no empirical support is given for this critical assumption at the aggregate, county level, nor is the literature on the epidemiology of suicides properly considered.

In fact, the literature suggests suicides are a complex (and often baffling) 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; see Eichstaedt et al., 2018 for a full discussion). Implicit in these associations is also the increased availability of guns in rural communities, the preferred means of committing suicide.

In the original 2015 analysis, we tested the dictionary associations with heart disease rates using its strongest covariates as controls (income and education), finding that most of the negative language variables remained significantly associated.

In the response to this manuscript (Eichstaedt et al., 2018), we focused on an analysis of dictionary correlations for parsimony. When controlling the associations between Twitter dictionaries and suicide rates for its strongest covariates--elevation and rural population--the language correlations reported in this manuscript are no longer significant (see Eichstaedt et al., 2018, Table 1, column 2). In contrast, controlling the association between heart disease rates with Twitter dictionaries for elevation and rural population does not noticeably affect those coefficients – all significant associations remain significant (see Eichstaedt et al., 2018, Table 1, columns 3 and 4).

This suggests that the associations reported for county-level suicide rates are not robust, and likely in large part driven by county-level confounding variables.

To more directly test the authors’ hypothesis that more psychological variables ought to be better candidates for association with psychological Twitter language, we tested 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). Unlike suicides, mentally unhealthy days correlate with the psychological dictionaries in the same directions and roughly at the same

magnitudes as heart disease mortality (see Table 2 in Eichstaedt et al, 2018).

To summarize, a) the suicide language correlations are not robust and likely associated with confounds which are well known to the suicide literature, b) a clear CDC-reported measure of county psychological health shows the same correlations as does heart disease mortality.

Thus, in my view the empirical analysis offered in this manuscript is insufficiently developed to support the claims made in the title and abstract—that is, the suicide analyses are unrelated in important ways to the heart disease mortality and unsuitable as empirical support offered in a manuscript that is framed as a "re-analysis" of the 2015 heart disease article.

As its own contribution to the literature on the epidemiology of suicide, it would require further development and integration with the existing literature.

Response:

Our aim was not to formally investigate the epidemiology of suicide and how this might be affected by Twitter language. We agree that the apparent county-level associations between Twitter language and suicide rates are likely to be the product of unmeasured confounds. Our point is that such confounds, together with measurement unreliability (e.g., the large variations in AHD mortality, the censorship of Twitter language, and the problems associated with county-level aggregation) are also a plausible explanation of the results reported in Eichstaedt et al.'s (2015) article.

The above point has not been adequately addressed. The suicide findings are insufficiently developed, as outlined in my previous comments and at the beginning of this document, and as requested by the editor. Regarding lurking confounds in the heart disease findings, we have considered a wide range of county-level variables as possible confounds in light of the existing heart disease literature, and included the most predictive ones in the baseline model. You are in the position to reproduce and extend these analyses with any additional confounds and report your findings, so no innuendo or speculation is needed or appropriate.

The suicide correlations are spurious and disappear when the known county-level covariates are controlled for (see above and our response preprint).

The analyses in the manuscript continue to be insufficiently developed to merit its publication as an empirical contribution to the literature that has clear bearing on the heart disease findings.

Validity of the findings

As stated in the experimental design section, the claims made in title and abstract are not supported by the suicide-related analyses in the manuscript. The conclusions drawn by the authors are not supported by the experimental design.

I see a number of directions in which this manuscript could be developed.

(a) It could be reframed as an exploration of the language correlates of suicides. In that case, the analyses need to be properly developed to incorporate what is known about county-level suicide rates. In the current form, the analyses appear insufficiently developed.

(b) The manuscript could be developed into a full commentary of social-media based and/or epidemiological methods used in psychology. In that case, an even-handed review of the associated literatures would be required (in its current form it appears one-sided and largely speculative).

(c) The analyses could be replaced with an evaluation of the correlational profiles of a wider variety of more unambiguously psychological variables and considered in light of the associated literatures. To the extent such a manuscript is still framed as a commentary on the heart disease work, a clear empirical rationale ought to be developed how the analyses relate to the heart disease results.

(d) The manuscript could attempt a replication of the results presented by Eichstaedt et al., 2015 (see Eichstaedt et al., 2018, Appendix A for step-by-step instructions) and report on it.

(e) The manuscript could explore issues of spatial aggregation of psychological phenomena.

In the current form of the manuscript, none of these issues is sufficiently developed.

Response:
The purpose of our manuscript is to bring to the attention of the scientific community a number of questions that, we feel, are worth discussing as counterpoints to Eichstaedt et al.'s (2015) article. Several of the possible developments mentioned here by the reviewer might well be interesting to pursue subsequently, but we feel that it is important to start the discussion with the specifics of the case at hand, and let the debate continue from there.

It is OK to offer critiques, and even the intention to get a scientific debate started is defensible. Purposeful misrepresentations, however, do not serve such a debate.

To repeat the main concern, the claims made in title and abstract are not supported by the suicide-related analyses in the manuscript.

Comments for the Author

To reiterate, this critique is ill-framed as a re-analysis of the Eichstaedt et al., 2015 article, and it does not, in fact, attempt a replication of the key claim of that article (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 (a) uncorrelated with heart disease rates and (b) disappear when county elevation and rural populations are controlled for, suggesting that county-level suicide rates are not a straightforward measure of county-level psychological health. In addition, (c) 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.

Response:
See our previous comments; we are not claiming that Twitter language reliably predicts suicide. But the fact that adding a selection of predictors can eliminate the apparent relation between county-level positive affect and suicide makes us wonder what might happen if a different set of additional predictors were to be included in the regressions predicting AHD mortality.

Please respect the methods of empirical science and test any such confounds you can think of! You have been able to rerun our analyses, so you can now subject any such variable to the test. We have included the most predictive variables, and used them in the baseline model. We have found the suicide confounds by simply skimming the county-level suicide literature, and were quickly able to uncover the pattern of spurious correlations. We invite you to attempt to do the same with the heart disease work.

The manuscript contains a number of speculative doubts about data and methods that do not just apply to the Eichstaedt et al., 2015 paper, but to large literatures in computer science, public health and epidemiology, which are insufficiently taken into account in this manuscript.

Response:
Again, there are plenty of interesting ideas for other articles here, but they would take us way beyond the scope of our manuscript.

No, it is not beyond the scope of the current manuscript to properly acknowledge the large literature that shares data and methods (see first editor comment).

To repeat, most critically, the misrepresentation in the title and abstract need to be changed that this manuscript provides a “re-analysis” of the 2015 paper (which has machine learning prediction as its key contribution), and that no evidence could be found that Twitter predicts

heart disease, given that the authors have now been able to reproduce the predictive model themselves.