From: stavroula.kousta@nature.com [mailto:stavroula.kousta@nature.com]
Sent: 08 March 2017 18:29
To:
Subject: Decision on Nature Human Behaviour submission NATHUMBEHAV-17021234

8th March 2017

Dear,

Thank you once again for your manuscript, entitled "Machine-learning identification of patients with suicidal ideation and suicidal behavior: Detecting alterations in the neural representations of death- and life-related concepts", and for your patience during the peer review process.

Your Article has now been evaluated by 3 referees. You will see from their comments copied below that, although they find your work of considerable potential interest, they have raised quite substantial concerns with respect to the generalizability and validation of your model, as well as the transparency of the methods and results. In light of these comments, we cannot accept the manuscript for publication, but would be interested in considering a revised version that addresses these serious concerns.

We hope you will find the referees' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions. Any revision must, at a minimum, thoroughly address the following:

*suitable validation of the model & demonstration of its generalizability;
*a principled, quantitative distinction of the population(s) to whom the model should apply (which is then used as a basis for selecting any subset of participants for your analyses);
*transparency in the presentation of methods and results (note that, although there is a word limit for the main text, there is no such limit for the supplementary material)

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

If revising your manuscript:

* Include a “Response to referees” document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

* If you have not done so already we suggest that you begin to revise your manuscript so that it conforms to our Article format instructions at http://www.nature.com/nathumbehav/info/final-submission. Refer also to any guidelines provided in this letter.

* Include completed version of the attached reporting checklists. They will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review. Completed checklists are essential for re-review of the paper.

If you wish to submit a suitably revised manuscript we would hope to receive it within 3 months. If you cannot send it within this time, please let us know. We will be happy to consider your revision so long as nothing similar has been accepted for publication at Nature Human Behaviour or published elsewhere. Should your manuscript be substantially delayed without notifying us in advance and your article is eventually published, the received date would be that of the revised, not the original, version.

Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Thank you for the opportunity to review your work.

Sincerely,

Stavroula Kousta

Stavroula Kousta, PhD
Chief Editor
Nature Human Behaviour

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
The manuscript by Just and colleagues – “Machine learning identification of patients with suicidal ideation and suicidal behavior: Detecting alterations in the neural representations of death – and life-related concepts” – employs a machine learning algorithm, applied to fMRI data, to identify from the brain data suicidal patients (vs. control subjects), and within the patient group, those patients who had previously actually attempted suicide. The algorithm did this by examing brain activation patterns evoked by death and life-related concepts. The results demonstrated remarkable sensitivity and specificity. Importantly, a major component of the neural signature that was used was associated with emotion.

I think this paper possesses a great deal of significance. It demonstrates how brain-derived data can be utilized to detect markers of a psychiatric disorder in single individuals, and as well, differentiate subtypes within the identified patient group. Also of importance, this paper shows how specific concepts can be differentially organized in the brains of patients vs. normal subjects.

There are, however, several problems with the manuscript that I think the authors need to address. The most important one concerns the absence of data figures. Two of the three figures that the authors provide in the paper show group data. However, in my view, the importance of the results concerns the ability to characterize individuals. Thus, I would like see individual brain data (in supplementary material). This would give the reader some feeling for the variability within the data and for inclination for how the machine learning algorithm is able to make the proper identification. In the results section the authors provide a number of statistical results, but they don't show the data points that went into the statistical analysis. It would be useful to actually see the distribution of data points for some of these results.
Figure 3 is also interesting, but it would be more so if the authors would indicate which data points corresponded to the attempters and which to the non-attempters. Do the residuals with respect to the regression line differentiate the two subgroups?

Finally, although machine learning is becoming more widespread in functional neuroimaging analysis, it is still not that common. Thus, in the supplementary material, I would suggest that the authors provide some information about the Gaussian Naïve Bayes method. In particular, what does this method use to discriminate the subjects' data?

Reviewer #2:
Remarks to the Author:

This manuscript proposed using fMRI to analyze (and predict) suicide ideality and suicidal behavior. It approaches a very difficult problem of identifying suicidal individuals. The accuracy is impressively high, and the stimulus (30 words related to suicide, negative concepts, and positive concepts) is appropriately novel to the intentions of identifying suicidality. The major limitations of this paper are the extensive feature mining, and the small sample size retained (34) of the original sample (79) for building the machine learning classifier.

Specifically, it is troubling that a GLM analyses was unable to identify any differences between groups for all 30 concepts or for subgroups of concepts (words). Because of this, the authors resorted to a concept similar yet statistically weaker to the GLM- averaging 4 TRs taken 4-8 seconds after the stimulus. This averaging TRs approach statistically has less power than a traditional GLM because it neglects the hemodynamic response and reduces degrees of freedom, yet yielded surprisingly high classification accuracy. However, this approach was only used for features within "regions" which were different between groups, and (more concerning) only within roughly 1/2 of the total subject pool (34/79 total subjects) for whom a difference appeared. This purposeful selection of regions, using a weaker variant of the GLM, in a limited subset of subjects, showed very strong classification accuracy, when a GLM (presumeably on the whole data set containing twice as many subjects) was unable to identify variability between these groups over the entire brain. Even if the methodology were completely unbiased (the machine learning part appears to be, yet the feature construction/5 region selection is ambiguous given that the same regions exist for all subjects), it shows that the generalizeability of this approach is likely limited even within the suicidal target sample.

This dataset appears well-collected with rich phenotyping and may yield many interesting and solid findings, yet it appears to be insufficient to answer this problem specifically. Identifying suicidality may be more likely when holding constant other (possibly causal) factors such as anxiety and childhood trauma. Approaches such as partial least squares may prove useful here, to adjust for the other measures (ASR, PHQ, etc...). The other measures (especially CTQ) are also very rich and insufficiently investigated in the literature.

Minor issues:
What is "sadness shame anger and pride" when only 3 types of words are presented, and none of these are anger or pride?

Control group: this study used subjects without a DSM diagnoses as a control group. However, subjects in the suicidal ideators group, subjects scored differently than controls in 7 assessments- only 1 of which was measuring suicidality. As the authors noted, the comparison then between healthy controls and individuals with suicidality does not necessarily suggest that suicidality itself is being selected for within the model- even within the suicide ideation group, subjects who may have acted on those ideations may have more depressive or anxiety-related symptoms.

Reviewer #3:
Remarks to the Author:
This study uses a Gaussian naïve Bayes classification to investigate the brain representations underlying thoughts related to death and life, and to develop a classifier the discriminates those with suicidal ideation from controls in N = 34 participants selected from larger sample. A second classifier discriminated those who made suicide attempts from those who did not in a smaller subsample (N = 17).

There are a number of strengths of the current paper, including the importance and novelty of the topic and the achievement of an impressively high classification rate. The idea of using activation of death and life related concepts also has strong face validity. The group's prior work on fMRI of concepts combined with machine learning has been groundbreaking and influential. It is highly appealing to extend this work in clinical directions and to particular populations. The use to examine suicidal ideation here is appropriate and exciting.

There are some limitations, and addressing these would help to strength in the scientific basis for the findings and increase the impact of the paper.

Part of my confusion about some details may stem from the fact that I did not seem to have access to the supplemental materials, though they were referenced in the manuscript.

First, the sample size is limited, particularly for a between person classifier that is subject to a number of potential confounding variables. Participants with suicidal ideation and controls were appropriately matched on several variables, but others are difficult to match on. Binary classifications are also more likely to pick up on extraneous variables rather than tracking suicidal ideation itself, though the correlation with individual differences in ideation is helpful in that regard.

The most helpful thing overall would be to see the model validated on a second sample, tested without changing or refitting the model. I note that application of the model to the initially excluded participants is very helpful in this regard, but it's not clear if the model is re-fit, as this application is described as a "leave-one-out" procedure and no cross-validation is required if no parameters are estimated.

A related concern is that the model seems to apply only to a subset of participants for whom the semantic concept classification worked well. While this is reasonable in many respects, it rules out the use (or perhaps development) of the model for use at a population level, and raises questions about whether those included are indeed representative of a broader population of suicidal individuals. It would be helpful to clearly and quantitatively indicate criteria for who the model should apply to and who it should not. Also, it's claimed that the model is applied to "those who demonstrated sustained attention and little head motion," but this is neither strictly accurate nor are attention and head movement assessed here. It seems premature to conclude that because the semantic classifier did not work for a given individual they must have had too much head movement or poor attention. There are many other potential causes, including individual differences and limitations in the original model. Thus, I found the statements about who is included and why to be somewhat misleading. I note that application of the model to the initially excluded participants is helpful in this regard, proviso understanding how this was actually done.

The second main issue relates to clarity and detail in the presentation of the model and results. There are a number of unusual steps included in the modeling process, and very little detail (in the main manuscript, at least) about how the relevant metrics are calculated. The result is a presentation of both model and results that is largely narrative-based and thus hard to understand concretely. It is not clear precisely how the discriminating concepts were defined, and how the four emotion vectors were applied. The presentation of results is sparse and tends to claim that the models worked well without showing any individual person-level and/or concept-level data that illustrates how well they worked and what the model is based on. This extends to the figures, which also display very little information and detail. This is a missed opportunity, and it also makes it less clear exactly how the concepts studied map onto brain activity, and how precisely the suicidal group is different. It is also not clear how strongly the concepts that are claimed to be related to suicidal ideation are actually related, or with the metric actually is. Are the concepts really significantly differentially involved? What is the distribution across concepts? How different is "death" from other concepts, and how different are the brain "representations" of "death" for those What did the maps for different concepts look like, and how similar are they? Do they have structured relationships that make sense according to establish models of semantics?

Third, the authors might be clear about what they mean about the emotion-related component of the model. It is very interesting to apply the patterns from Kassam et all, and on one hand this is innovative. However, it may be premature to conclude that any response related to these emotional patterns must be indexing a particular emotional state. This clearly seems to be assumed here. If, for example, I activate a "shame"-related pattern when I think of "death," it might not be that I am ashamed at all, but that the shame-related pattern is indexing a broader class of concepts and does not indexed the emotion per se. In addition, these emotional patterns have not been validated an independent data sets, nor examined for selectivity in additional independent data beyond the relatively limited initial sample. Can the emotion patterns be further validated in some way here -- i.e., do the regression weights of the emotion patterns on the concepts at least make sense overall? And how do we discern whether activation of an emotion-related pattern when thinking about a concept really means the *emotion* is activated, or whether the pattern is just not very specific? The authors may or may not agree with this, but some conceptual clarity is important.

It is also not clear why all 9 emotions from the Kassam maps were not characterized.

Fourth, two different class of fires are developed, though they might be indexing different levels of the same underlying construct, i.e. brain changes related to suicidal thoughts. To what degree are the two classifier a similar, and could a single classifier both discriminate those with suicidal ideation from those without and further show greater responses in those who actually attempt suicide? What does the suicidal attempt classifier even look like (it is not shown)? Does the original suicidal ideation classifier show a stronger response (distance from control norm) in attempters than non-attempters?

Additional thoughts

Here is a conceptual question: if many patients who died by suicide deny suicidal ideation, how can it be that brain measures accurately track those who report suicidal ideation, unless those brain measures are also relatively unrelated to actual suicide attempts? There may indeed be a good answer to this question but it is worth discussing.

The region described as anterior cingulate is not actually in the anterior cingulate.

If there are different sets of consistently engaged voxels when thinking about the concepts studied across those with suicidal ideation and those without, does that mean the classifier is based on the differential activation of these different regions? Or is it based on something else?

It would be helpful to have more details about how the Naive Bayes classifier deals with the continuous nature of fMRI signal. I assume that normal distributions are assumed and two distributions are estimated centered on the suicidal ideation and nonsuicidal edition groups. Perhaps this is in the missing Supp. Info.

Please clarify that leave one subject out classification was used. In addition, did the permutation test reveal any bias that could result from nine independence of subjects?

Copy of Final decision letter:

Subject: Decision on Nature Human Behaviour manuscript NATHUMBEHAV-17021234A

6th June 2017

Dear ,

Thank you once again for your revised manuscript, entitled "Machine-learning identification of suicidal ideation and behavior: Detecting alterations in the neural representations of key concepts," and for your patience during the re-review process.

Your manuscript has now been evaluated by Reviewers 2 and 3 from the original round of review. Their comments are included at the foot of this letter. Note that Reviewer 3 was asked to also provide additional commentary on the methods and the issues raised by Reviewer 2 in this round. This additional commentary is included after his/her independent review below. In the light of the feedback the reviewers provided, I regret that we cannot offer to publish your manuscript in Nature Human Behaviour.

As you can see, Reviewer 2 continues to raise objections about the data exclusions and the way the classifier was built. This is a significant concern, which could be overcome if out-of-sample validation were strong and convincing. However, Reviewer 3 has significant concerns about that, which cannot be addressed without additional, independent out-of-sample replication. If you do decide to perform independent replication in a new sample, we would be very pleased to consider a revision. Without these data, however, the reviewers' concerns with the present dataset & analyses are such as to preclude publication in Nature Human Behaviour.

I am sorry to be a bearer of discouraging news but hope that you will find the reviewers' comments helpful when preparing your paper for submission elsewhere.

Sincerely,

Stavroula Kousta

Stavroula Kousta, PhD
Chief Editor
Nature Human Behaviour

Reviewers' Comments:

Reviewer #2:
Remarks to the Author:
This is a revision of the paper "Machine-learning identification of suicidal ideation and behavior: Detecting alterations in the neural representations of key concepts"

The main original objection I had to the paper still has not been addressed. The abstract of the paper claims 91% accuracy in distinguishing suicidal individuals from healthy controls, but this accuracy is on 34 subjects (17 suicidal, 17 controls) subjects who were cherry-picked out of a pool of 79 total subjects because they showed discriminating activity. The authors have shown that their sub-model predicts ideators from attemptors, but this is a separate issue from intentional omission of 1/2 the total subject pool because it didn't fit the model.

"When these differences were statistically controlled for (using methods described by the classification accuracy slightly improved (from .91 to .94).
This process needs to be described in detail- controlling for multiple covariates when the groups differ on most of these covariates is not a simple process.

The sample sizes are not described in the abstract, and the details of the best accuracy (91%) needs to be specified that it is between healthy controls and patients.

Reviewer #3:
Remarks to the Author:
I have reviewed the manuscript, response letter, and supplementary information. Overall, the authors have responded with new analyses and figures, were responsive to issues raised in review. I believe this will be an important paper.

I have a few remaining comments, which are not intended to hold back publication, but to

The successful application of the model to the initially excluded participants is a very strong point, and strengthens the case for the importance of this paper.

The explanation for the use of simple averages across time vs. the GLM estimates is convincing.
The justification of the sample size is convincing.
The addition of analyses controlling for several psychological measures is very helpful.

I am a fan of the emotion classification work overall, and the application to this analysis. However, I suspect that what the field will find is that these patterns cannot be used to make strong inferences about emotional content. For instance, do you KNOW that there is no "emotion-free" cognitive task that engages any of your emotion patterns? As I mentioned before, these patterns have not been tested beyond the original Kassam publication by other research groups, on other datasets, to my knowledge. The authors might consider this in their discussion and make appropriate caveats.

The two full groups of participants (38 ideators and 41 controls) differed in age and gender. What is the classification accuracy controlling for these variables? And is it significant?

I would prefer to see the distributions of ideators vs. controls on responses in all of the Kassam emotion patterns, whether different across groups or not, even if a subset of these is selected as related to suicidal ideation. The argument about overfitting with 9 vs. 4 predictors does not make sense, if you are testing in out-of-sample participants. The request is simply to show data for ideators and controls (preferably individual person-level data) for 9 separate measures, one for each emotion pattern. There need not be any model fitting at all, necessarily – just emotion model application.

Sorry, but I have reviewed the SI and the manuscript, and I still don't have a very strong sense of what is driving the classifier at a brain level, and thus what we are learning about brain function from this. This is simply a request for more information. Figures 1 and 2 are very helpful in showing locations – but what concepts show the most differential activation in these regions, and in what direction? For example, you write: "The concepts that most strongly discriminated the groups were death, cruelty, trouble, carefree, good, and praise." Can you show, for example, line/bar plots across concepts in some of these regions, with error bars, for ideators and controls? One could imagine that this type of plot or similar would show the reader, e.g., in DMPFC the ideators show stronger responses than controls to death and trouble but not other concepts. There is a bit of this in the SI but it's hard to put the picture together. There is an opportunity here.

ADDITIONAL COMMENTS BY REVIEWER 2:

Upon a careful examination of the methods in this paper, my opinion is mixed about the strength of the findings. On the one hand, the authors DID answer criticisms about model overfitting and potential bias induced by feature selection, because they explicitly claimed to have tested the classification in a new, independent sample that was not part of the feature-selection process. This appears at first blush to be the case, but there are several irregularities in the details that leave me feeling not entirely confident that these results would replicate in a new sample with the exact same model.

1. They used a leave-one-out classification. This, while it sometimes has problems providing unbiased estimates, is a reasonable way of ensuring that the feature selection does not amount to a fishing expedition -- as long as feature selection is performed inside the cross-validation loop. But there are some potential issues here:
They write: "The voxel selection is based on only the training data for the model in each cross validation fold and is then applied to the test data." This is appropriate. One hopes that this was applied to all of the model-selection processes. However, upon careful review, it is not entirely clear that this is the case. They write: "To obtain a map of the clusters of stable voxels that characterized each group, a hit map was computed for the ideator group and the control group, such that only the voxels with a contribution of at least 4 (of 17) participants were considered." This implies that selection of clusters of voxels was based on the FULL sample (17 per group), not the training sample (ideally 16 per group, leaving one control and one ideator out; but here perhaps 16 and 17). This could constitute a model selection bias that could spuriously increase classification rates. In addition, the complex process described here would have to be automated and peformed within the cross-validation loop: "These voxels were spatially clustered, and only the clusters containing at least 5 voxels were included in the stability map of each group. Large clusters (with radius > 11mm) were subdivided into smaller clusters (either by finding within-cluster local maxima or by splitting the cluster along its longest axis)." The text below seems to suggest that this entire process was NOT, in fact, done within the cross-validation loop, as the stepwise procedure to obtain 5 clusters appears to have been done outside of cross-validation: "To determine which of these clusters would be useful in the group classification, the most discriminating clusters were identified using a stepwise procedure described below. Five such clusters were identified, as shown in Figure 2. These clusters were populated by stable voxels from one or both groups. During the group membership classification, the locations of the five discriminating clusters were recomputed to exclude the data of the left out participant (in order that the test participant's data be excluded from the training of the classifier) and compensating for the loss of these data by modifying the total number of initially selected voxels per participant to 1200 voxels." There are a number of complex and arbitrary steps here, which is an issue in terms of transparency and also over-optimism if these decisions were not automated within the cross-validation loop and NO changes were made after running an initial cross-validation analysis. e.g., if the authors tried this process with 7 clusters and without normalization to 1200 voxels, and obtained worse cross-validated results, and then re-ran the cross-validation with any such changes, this would produce over-optimistic accuracy.

2. They now report and additional, more conservative test of generalization: In additional quantitative assessment of the generalizability of the model applied a more conservative cross-validation technique. Instead of training the model on data from all but one participant, this

additional assessment left out the data of half of the participants (8 of 17) from each group for testing, and the model was trained on the remaining 9 participants' data. (There are a huge number of ways to leave out half of the participants from each group, so 1000 random selections of such partitionings were performed and the outcomes were averaged). The result was that the classification accuracy remained at a highly reliable level of .76, showing that a model based on a much smaller sample of the participants generalizes to the other half." This is helpful, but the accuracy is substantially lower. This MIGHT be related to the smaller training sample, but it might also be caused by bias due to model flexibility, re-fitting, and analysis choices made based on the WHOLE sample rather than the training subsets.

3. They show generalization to a new subsample of ideators independent from the training sample, which was not used in feature selection or model development. This is good, but there are two red flags to me in how this was done.

"Despite their exclusion from the main neurosemantic analysis, we show below that there remains valuable information in the fMRI data of the excluded suicidal ideator participants. The comparison of self-report data between the 34 participants included in the neurosemantic analyses and the remaining (excluded) participants is reported in Supplementary Information." and later, regarding this analysis: "no model re-fitting was performed." However, a limitation to this is that they used the SAME set of controls, not the 24 additional controls who were previously left out. This is unusual. Another limitation is that it is not exactly the same model, as their reply led me to believe. They write: "the number of voxels used in each location was increased from five to seven to counteract the higher noise level". These two limitations do, in fact, constiute two examples of the types of "model flexibility" that can be used to make results look better than they first appear, and are discussed in the "P-hacking" literature.

The other remaining concerns I outlined in my previous review still stand. For example, I think the argument that the emotion classifiers they apply really capture the essence of the emotion and can be used as markers for emotions with no caveats is naive. But this is an issue of interpretation. I'm also somewhat concerned, in light of the additional "model flexibility" mentioned above, about the fact that they declined to show the results from all of the emotion patterns they tested, claiming that it would "over-fit the data", which is not the case. These things make me a bit worried that some liberties are being taken to make the results better than they would be with a truly straightforward, unbiased replication of the findings in a new sample. Based on the authors' description, I think their model is interesting, and "paper tiger" is perhaps too strong a characterization. But I would feel more comfortable with an independent replication in a new sample without the "fudge factors" that seem to be present in this analysis.