nature human behaviour

Matters arising

https://doi.org/10.1038/s41562-023-01560-6

Overfitting to 'predict' suicidal ideation

Received: 1 September 2020

Accepted: 16 February 2023

Published online: 6 April 2023

Check for updates

Timothy Verstynen [1] ✉ & Konrad Paul Kording [2]

ARISING FROM: M. A. Just et al. *Nature Human Behaviour* https://doi.org/10.1038/s41562-017-0234-y (2017).

Unlike many areas of medicine, the fields of psychiatry and clinical psychology suffer from a critical lack of ability to directly measure the internal processes that are the root of most psychiatric disorders[1]. Instead, these fields rely on indirect assessments, via verbal report or behavioural analyses, that can often be unreliable indicators of internal thoughts and experiences. Over the past few years, machine learning methods applied to functional neuroimaging data have presented a promising avenue for the field of computational psychiatry to potentially measure preverbal internal processes[2], offering hope for the development of neural biomarkers of psychiatric diseases.

In one such study, Just and colleagues[3] reported promising findings on a potential neural biomarker for suicidal ideation. The authors reported a 91% classification accuracy for predicting a participant's group membership (suicidal ideating individuals, $n = 17$; non-ideating control; $n = 17$), using leave-one-out cross-validation (LOOCV) with a classifier trained on functional magnetic resonance imaging responses to a list of words. Such a robust ability to identify individuals who are probably suicidally ideating on the basis of preverbal neural processes could revolutionize psychiatric approaches to suicide.

However, the procedures described in the original paper suggest several problems. First, the use of LOOCV can inflate the estimated classification accuracy, as well as overall type-I error[4]. Second, and most importantly, the feature selection appears to have relied on the same data that are used in the final model evaluation. In the supplementary information section of their paper titled 'Identifying the most discriminable concepts and locations', the authors state that they used a forward stepwise selection procedure to identify the best combination of concepts (words) and locations (sphere of voxels from anatomically defined regions of interest) that maximized their model accuracy in predicting whether a participant was in the suicidal ideation or control group. According to the text, feature selection happened along two different dimensions: words and regions. For words, the authors only used data from 6 out of 30 words in their final model. The authors present no a priori reason for why this subset of words would be better at discriminating between groups. Therefore, we assume that this subset was determined solely by the described forward stepwise search process. We cannot be entirely sure because the authors did not share all their code. However, as we shall point out below, we have reasons to believe that we understand their approach reasonably well. For regions, the authors used multiple selection procedures for determining which clusters of voxels to include in their model, resulting in 5 out of approximately 25 (on average 25 based on group analysis) regions being included in the final model. First, the authors evaluate voxels on the basis of a stability score of responses across trials. No information is provided for how this stability is quantified. For each fold of the cross-validation procedure, the hold-out test subject was not included in the voxel stability analysis, although it is not clear why because voxel stability is already an independent measure from the classifier performance. Second, the authors selected the best subset of stable voxel clusters on each fold, separately for each group. On average there were 11 stable regions for the suicidal ideation group and 14 regions for the control group. The final analysis only included two from the suicidal ideation group and three from the control group, again presumably identified using the stepwise selection procedure. This is problematic, however, because group assignment is already influencing features included in the final classifier analysis. It is in the group subset that the forward stepwise search appears to have been applied.

Given the sample size and structure of the classification problem, we can see no way that a consistent set of features (one set of words and regions to serve as a biomarker across all subjects) can be identified without using data from all participants. This reflects what we are calling 'feature hacking'[5], a form of circular inference[6] that contaminates the validity of out-of-sample validation tests. Feature hacking is the process of inflating model performance in cross-validation tests by selecting the best subset of the features that maximizes performance of the hold-out test set that is used as a benchmark for how well the classifier generalizes to predicting unseen data.

Using code and data provided by the authors, notably missing the exact code for implementing the feature selection steps, we conducted a re-analysis of the feature selection process. We started by simply attempting to replicate the deterministic feature selection method as described in the original manuscript, using a logistic regression classifier on group membership and a forward stepwise feature selection in three stages. First, we selected the subset of words that best distinguished the two groups using average response data from all stable regions (with stability determined excluding the data from the out-of-sample participant). Second, we selected the set of stable regions from the suicidal ideation group that best distinguished group membership using all 30 words. Finally, we selected the stable regions from the control group using all 30 words. Feature selection on regions

[1] Departments of Psychology, Carnegie Mellon Neuroscience Institute, and Biomedical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA.
[2] Departments of Bioengineering and Neuroscience, University of Pennsylvania, Philadelphia, PA, USA. ✉e-mail: timothyv@andrew.cmu.edu

Table 1 | Change in LOOCV accuracy when different feature selection approaches are applied

Configuration	LOOCV accuracy
Preselected features (words and regions) from ref. [3]	91%
Features from forward stepwise search using logistic regression	32%
All words, preselected regions	59%
Preselected words, all regions	65%
All words, all regions	41%

was run separately for the two groups because this follows the logic of the original analysis. It is worth noting that this method still suffers from circular inference because all data are being used in the feature selection process. As highlighted above, the sample size is too small to enable a completely unbiased feature selection process.

Our analysis could not replicate the original feature set. We identified only one word, 'vitality', one region from the patient group (left angular gyrus) and one region from the control group (left anterior cingulum). Only the last region overlapped with the original set of features. Importantly, using these words and regions, the LOOCV classifier method from the original paper falls to 32% (Table 1). Thus, using a standard forward stepwise selection procedure we were unable to either replicate the features or model accuracy reported in the original paper.

We next set out to see how much the feature selection process used by the authors affected the classifier performance. To start we re-ran the original classifier reported by ref. [3] but removed feature selection along the two dimensions, words and regions, separately. These results are reported in Table 1. Removing feature selection on words, but including the same set of selected regions as used in the original paper, reduced classifier accuracy by 32%. Removing feature selection on the set of stable regions, while keeping the original six words used in the original paper, dropped classifier accuracy by 26%. Thus the classifier accuracy reported by ref. [3] is highly sensitive to the unique set of words and regions used.

The only feature selection method used in the original paper that does not suffer from circularity is the original selection of the stable voxel clusters (regions). Here, stability was determined by excluding the held-out test subject for each run of the LOOCV classifier. Thus, the only truly unbiased model that can be run on the data is one in which all words and all stable regions, for both groups, are used. This model returns a classification accuracy of 41%, well below chance and a full 50% below the accuracy reported by ref. [3].

Using information from data in a validation set to determine the structure of a model leads to inflated estimates of performance. This can happen either by selecting the observations (for example, only including the subset of participants that maximize validation set performance) or features (for example, applying arbitrary transformations of variables based on validation set performance) based on information from what should be a protected part of the sample. Our re-analysis shows that the classification results reported by ref. [3] are probably inflated due to the presence of information leakage somewhere in the feature selection process. Our analysis clearly shows that the most conservative approach (using all words and all stable regions) yields classification performance that does not outperform chance and no combination of features returns a classification accuracy near what was previously reported. Without a more detailed description of the methods and independent evaluation of the feature selection process itself, we are forced to conclude that the reported ability to discriminate the suicidally ideating from non-ideating controls is not supported by the available code and data provided from in the original report.

Data availability

The data used in these analyses can be found at https://doi.org/10.1184/R1/22086995.v1 and http://www.ccbi.cmu.edu/Suicidal-ideation-NATHUMBEH2017/Just-NatHumBeh2017-data-and-code.html.

Code availability

The code used in these analyses can be found at https://doi.org/10.1184/R1/22086995.v1 and the original code shared by ref. [3] can be found at http://www.ccbi.cmu.edu/Suicidal-ideation-NATHUMBEH2017/Just-NatHumBeh2017-data-and-code.html.

References

1. Fuchs, T. Subjectivity and intersubjectivity in psychiatric diagnosis. *Psychopathology* **43**, 268–274 (2010).
2. Bzdok, D. & Meyer-Lindenberg, A. Machine learning for precision psychiatry: opportunities and challenges. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **3**, 223–230 (2018).
3. Just, M. A. et al. Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nat. Hum. Behav.* **1**, 911–919 (2017).
4. Flint, C. et al. Systematic misestimation of machine learning performance in neuroimaging studies of depression. *Neuropsychopharmacology* **46**, 1510–1517 (2021).
5. Cawley, G. C. & Talbot, N. L. C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **11**, 2079–2107 (2010).
6. Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F. & Baker, C. I. Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* **12**, 535–540 (2009).

Author contributions

T.V. and K.P.K. contributed to writing and conceptualization. T.V. wrote the analysis.

Competing interests

T.V. works at the same institution as the senior and first author of ref. [3]. The authors have no other competing interests.

Additional information

Correspondence and requests for materials should be addressed to Timothy Verstynen.

Reprints and permissions information is available at www.nature.com/reprints.