nature human behaviour

MATTERS ARISING

https://doi.org/10.1038/s41562-021-01085-w

Check for updates

Towards increasing the clinical applicability of machine learning biomarkers in psychiatry

Juergen Dukart[1,2], Susanne Weis[1,2], Sarah Genon[1,2] and Simon B. Eickhoff[1,2] ✉

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

Due to a lack of objective biomarkers, psychiatric diagnoses still rely strongly on patient reporting and clinician judgement. The ensuing subjectivity negatively affects the definition and reliability of psychiatric diagnoses[1,2]. Recent research has suggested that a combination of advanced neuroimaging and machine learning may provide a solution to this predicament by establishing such objective biomarkers for psychiatric conditions, improving the diagnostic accuracy, prognosis and development of novel treatments[3].

These promises led to widespread interest in machine learning applications for mental health[4], including a recent paper that reports a biological marker for one of the most difficult yet momentous questions in psychiatry—the assessment of suicidal behaviour[5]. Just et al. compared a group of 17 participants with suicidal ideation with 17 healthy controls, reporting high discrimination accuracy using task-based functional magnetic resonance imaging signatures of life- and death-related concepts[3]. The authors further reported high discrimination between nine ideators who had attempted suicide versus eight ideators who had not. While being a laudable effort into a difficult topic, this study unfortunately illustrates some common conceptual and technical issues in the field that limit translation into clinical practice and raise unrealistic hopes when the results are communicated to the general public.

From a conceptual point of view, machine learning studies aimed at clinical applications need to carefully consider any decisions that might hamper the interpretation or generalizability of their results. Restrictiveness to an arbitrary setting may become detrimental for machine learning applications by providing overly optimistic results that are unlikely to generalize. As an example, Just et al. excluded more than half of the patients and healthy controls initially enrolled in the study from the main analysis due to missing desired functional magnetic resonance imaging effects (a rank accuracy of at least 0.6 based on all 30 concepts). This exclusion introduces a non-assessable bias to the interpretation of the results, in particular when considering that only six of the 30 concepts were selected for the final classification procedure. While Just et al. attempt to address this question by applying the trained classifier to the initially excluded 21 suicidal ideators, they explicitly omit the excluded 24 controls from this analysis, preventing any interpretation of the extent to which the classifier decision is dependent on this initial choice.

From a technical point of view, machine learning-based predictions based on neuroimaging data in small samples are intrinsically highly variable, as stable accuracy estimates and high generalizability are only achieved with several hundreds of participants[6,7]. The study by Just et al. falls into this category of studies with a small sample size. To estimate the impact of uncertainty on the results by Just et al., we adapted a simulation approach with the code and data kindly provided by the authors, randomly permuting (800 times) the labels across the groups using their default settings and computing the accuracies. These results showed that the 95% confidence interval for classification accuracy obtained using this dataset is about 20%, leaving large uncertainty with respect to any potential findings.

Special care is also required with respect to any subjective choices in feature and classifier settings or group selection. While ad-hoc selection of a specific setting is subjective, testing of different ones and outcome-based post-hoc justification of such leads to overfitting, thus limiting the generalizability of any classification. Such overfitting may occur when multiple models or parameter choices are tested with respect to their ability to predict the testing data and only those that perform best are reported. To illustrate this issue, we performed an additional analysis with the code and data kindly provided by Just et al. More specifically, in the code and the manuscript, we identified the following non-exhaustive number of prespecified settings: (1) removal of occipital cortex data; (2) subdivision of clusters larger than 11 mm; (3) selection of voxels with at least four contributing participants in each group; (4) selection of stable clusters containing at least five voxels; (5) selection of the 1,200 most stable features; and (6) manual copying and replacing of a cluster for one control participant. Importantly, according to the publication or code documentation, all of these parameters were chosen ad hoc and for none of these settings was a parameter search performed. We systematically evaluated the effect of each of these choices on the accuracy for differentiation between suicide ideators and controls in the original dataset provided by Just et al. As shown in Fig. 1, each of the six parameters represents an optimum choice for differentiation accuracy in this dataset, with any (even minor) change often resulting in substantially lower accuracy estimates. Similarly, data leakage may also contribute to optimistic results when information outside the training set is used to build a prediction model. More generally, whenever human interventions guide the development of machine learning models for the prediction of clinical conditions, a careful evaluation and reporting of any researcher's degrees of freedom is essential to avoid data leakage and overfitting. Subsequent sharing of data processing and analysis pipelines, as well as collected data, is a further key step to increase reproducibility and facilitate replication of potential findings.

For the field of clinical neuroscience to move towards being able to improve the diagnosis and prognosis of psychiatric conditions using objective machine learning-based biomarkers in mental health, it is crucial to deal with the above conceptual and technical pitfalls in the first place. Most importantly, realistic clinical populations need to be enrolled in such studies to increase the ecological validity and generalizability. Furthermore, strict technical and operational procedures need to be defined. At the stage

[1]Institute of Neuroscience and Medicine: Brain and Behaviour (INM-7), Jülich Research Centre, Jülich, Germany. [2]Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. ✉e-mail: S.Eickhoff@fz-juelich.de

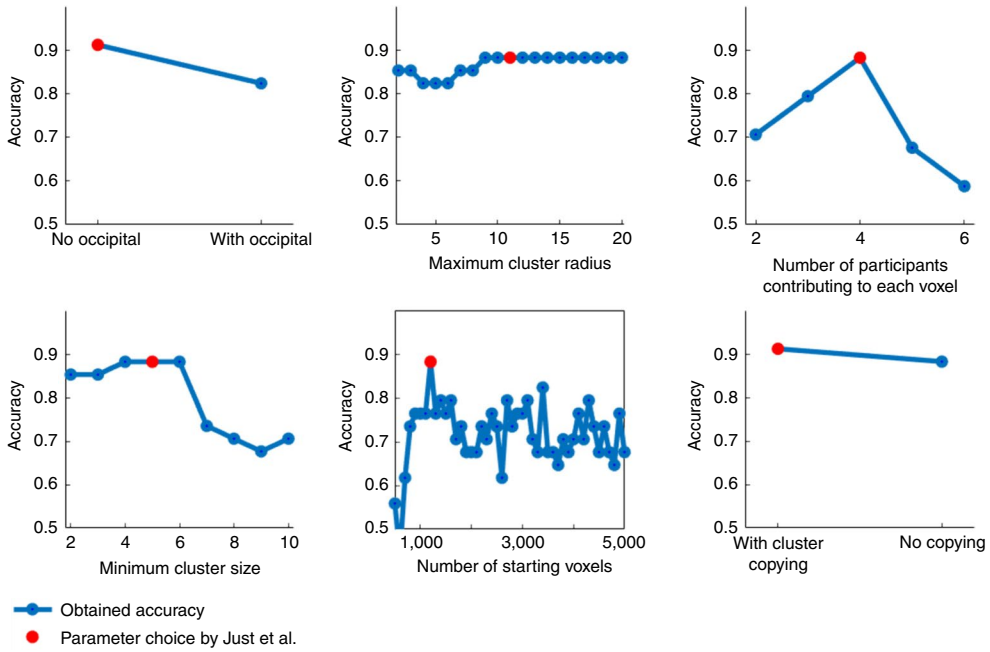

Fig. 1 | Effects of parameter choices on the accuracy for differentiation between suicide ideators and controls. Accuracies obtained for differentiation between suicide ideators and controls in the dataset used in Just et al. by systematically evaluating different parameter choices, while keeping all other parameters as default.

of identifying candidate machine learning biomarkers, strict nested cross-validation procedures, minimizing human interventions and subjective parameter choices, are essential to avoid data leakage or overfitting. Ideally, any results should be replicated in holdout data, and preferentially out-of-sample datasets, to evaluate their generalizability to other clinical populations. This comprises careful reporting and interpretation of any findings acknowledging specific limitations, to avoid inflation of public expectations or interpretation going beyond the actual scope of the specific study. Lastly, a prospective validation of the established machine learning biomarker with pre-registration of any settings and analyses is essential to establish its true generalizability. In particular, this last step, involving the collection of additional data in clinical populations that are typically difficult to recruit, requires prospective planning from researchers and clear commitment from grant and regulatory agencies to support such validation studies.

While these aims appear to be time and resource expensive, the benefit of the alternative (that is, investing resources into research with probably limited replicability, interpretability and generalizability) should be seriously questioned from a long-term perspective. From a research perspective, enticing but poorly replicable findings come with the danger of investing public resources on shaky scientific evidence and hence the risk of a public resource waste. Finally, from a societal and ethical perspective, there cannot be any doubt that treatment and medical intervention should build on strong and large evidence for the validity and efficacy of the proposed intervention in the clinical condition.

Methods

To perform the analyses reported here with respect to parameter choices and random group labelling, original code (http://www.ccbi.cmu.edu/Suicidal-ideation-NATHUMBEH2017/Just-NatHumBeh2017-code1.tgz) and data (http://www.ccbi.cmu.edu/Suicidal-ideation-NATHUMBEH2017/Just-NatHumBeh2017-data-and-code.html) used by Just et al. were downloaded from the respective public repositories. Random accuracy variation estimates were obtained by randomly permuting group labels between patients and controls 800 times and computing classification accuracies using the default settings from Just et al. In addition, to estimate the effect of parameter choices, each parameter was systematically changed (as displayed in Fig. 1) with recomputing of the classification accuracy while keeping all of the other parameters as default.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Received: 12 September 2019; Accepted: 1 March 2021;
Published online: 05 April 2021

References

1. Egger, H. L. et al. Test–retest reliability of the Preschool Age Psychiatric Assessment (PAPA). *J. Am. Acad. Child Adolesc. Psychiatry* **45**, 538–549 (2006).
2. Hyman, S. E. The diagnosis of mental disorders: the problem of reification. *Annu. Rev. Clin. Psychol.* **6**, 155–179 (2010).
3. Bzdok, D. & Meyer-Lindenberg, A. Machine learning for precision psychiatry: opportunities and challenges. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **3**, 223–230 (2018).
4. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
5. 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).
6. Cui, Z. & Gong, G. The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. *NeuroImage* **178**, 622–637 (2018).
7. Varoquaux, G. Cross-validation failure: small sample sizes lead to large error bars. *NeuroImage* **180**, 68–77 (2018).

Author contributions

J.D. and S.B.E. designed the work and wrote the manuscript. S.G. and S.W. contributed to interpretation and substantively revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-021-01085-w.

Correspondence and requests for materials should be addressed to S.B.E.

Peer review information *Nature Human Behaviour* thanks the anonymous reviewers for their contribution to the peer review of this work.

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

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2021

nature research

Corresponding author(s): Simon B. Eickhoff

Last updated by author(s): Jan 18, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The code and data used in this study were adopted from the original publication by Just et al. (2017). This original data are available at: http://www.ccbi.cmu.edu/Suicidal-ideation-NATHUM BEH2017/Just-NatHumBeh2017-data-and-code.html
Data analysis	The original data analysis code is available at: http://www.ccbi.cmu.edu/Suicidal-ideation-NATHUMBEH2017/Just-NatHumBeh2017-code1.tgz

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

ll data used in this study were derived from the publication by Just et al. and are kindly provided by the authors of the respective manuscript: http://www.ccbi.cmu.edu/Suicidal-ideation-NATHUMBEH2017/Just-NatHumBeh2017-data-and-code.html

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	34
Data exclusions	n.a
Replication	n.a
Randomization	n.a
Blinding	n.a

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Magnetic resonance imaging

Experimental design

Design type	Task-based cross-sectional fMRI study
Design specifications	event-related design
Behavioral performance measures	n.a

Acquisition

Imaging type(s)	functional
Field strength	3
Sequence & imaging parameters	20 slices, voxel size 3.125 x 3.125 x 5mm3, repetition time 1s
Area of acquisition	whole brain

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software	n.a

Normalization	n.a
Normalization template	n.a
Noise and artifact removal	n.a
Volume censoring	n.a

Statistical modeling & inference

Model type and settings	n.a
Effect(s) tested	n.a

Specify type of analysis: ☒ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016)	n.a
Correction	n.a

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Multivariate modeling and predictive analysis	Group labels (suicidal ideators vs controls) were used as labels for prediction using whole-brain derived contrast estimates as described in Just et al. (2017)