American College of Neuropsychopharmacology

www.nature.com/npp

CORRESPONDENCE

Reply to Winter et al: Interpreting weights of multimodal machine learning models—problems and pitfalls

Neuropsychopharmacology (2021) 46:1863; https://doi.org/10.1038/s41386-021-01082-7

In their correspondence, Winter et al. [1] raised concerns with the application of machine learning to examine associations between brain variables and childhood maltreatment in [2]. The primary concern was that the association between maltreatment and brain variables may have been obscured because the reported model contained non-brain covariates. Specifically, the results may have fallen victim to the Rashomon effect – the possibility that there are numerous combinations of brain variables that yield comparable findings to the reported model due to the inclusion of clinically-relevant covariates. This concern is important given the possible instability of machine learning results [3]. We addressed this concern in two ways. First, the brain regions were selected by aggregating over a set of 500 models, which is consistent with Breiman's recommendation for combating the Rashomon effect [3]. Second, we evaluated 250,000 competing models, constructed from permuting features from the entire feature set. As indicated in supplemental materials, the reported model (AUC = .90) outperformed all competitor models (AUC_{Mean} = .74), which suggested the specific features in the reported model were likely associated with maltreatment.

Winter et al. correctly noted that the inclusion of covariates in a model alters the association between brain regions and maltreatment. Accounting for such variables, however, is imperative as machine learning methods often identify patterns among the variables of interest that serve as proxies for these covariates. For example, if an association between sex and the outcome variable exists but sex is not in the model, the elastic net may include brain regions that diverge across the sexes. This would lead to the incorrect conclusion that these regions were associated with the outcome variable. A proposed remedy to this proxy concern is to regress out covariates from each brain region and then use the residualized regions in the analysis.

We used this residualized approach to determine if the selected brain regions reported in [2] were associated with maltreatment. Using the residualized brain regions reported in [2], ridge regression using 5-fold cross-validation obtained an $AUC_{residualized} = 0.69$. This result was comparable to the model reported in [2] that included non-residualized brain regions and covariates, AUC = .71. We repeated the permutation analysis described above using residualized brain regions. The residualized model obtained an $AUC_{residualized} = 0.81$, which was superior to all 250,000 competing residualized models ($AUC_{Mean} = 0.67$). The comparable performance between a residualized brain-only model and the model that included brain regions and covariates further supports the association between maltreatment and the regions identified in [2]. We recommend future researchers use this residualized method to evaluate brain-only models while accounting for confounding variables.

Winter et al. also raised concerns on the misinterpretation of multivariate weights from machine learning models. We agree and, indeed, our manuscript interpreted results based solely on the inclusion/exclusion of features. The bar graph showing feature weights was included to list the selected features, show their relative weightings, and communicate the directionality of the associations. We too caution others to avoid interpreting weights as indicative of association strength.

Matthew Price[1]✉, Nicholas Allgaier[2] and Hugh Garavan[2]
[1]*Center for Research on Emotion Stress and Technology, Department of Psychological Science, University of Vermont, Burlington, VT, USA.* [2]*Department of Psychiatry, University of Vermont, Burlington, VT, USA.* ✉*email: Matthew.Price@uvm.edu*

REFERENCES

1. Winter NR, Goltermann J, Dannlowski U, Hahn T. Interpreting weights of multimodal machine learning models—problems and pitfalls. Neuropsychopharmacology. 2021:1–2. https://www.nature.com/articles/s41386-021-01030-5#citeas.
2. Price M, Albaugh M, Hahn S, Juliano AC, Fani N, Brier ZMF, et al. Examination of the association between exposure to childhood maltreatment and brain structure in young adults: a machine learning analysis. Neuropsychopharmacology. 2021:1–7. https://www.nature.com/articles/s41386-021-01030-5#citeas.
3. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci. 2001;16:199–231.

FUNDING & DISCLOSURES

The authors report no biomedical financial interests or potential conflicts of interest. This work received support from the following sources: NIMH K08MH107661-A1 and the European Union-funded FP6 Integrated Project IMAGEN (Reinforcement-related behavior in normal brain function and psychopathology) (LSHM-CT- 2007-037286).

AUTHOR CONTRIBUTIONS

MP: conceptualized the response, conducted follow-up analyses, and drafted the majority of the manuscript. NA: Provided substantial guidance on the analysis and reviewed drafts of the manuscript. HG: Reviewed drafts of the manuscript and provided input on the narrative direction.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to M.P.

Received: 14 June 2021 Accepted: 18 June 2021
Published online: 13 July 2021