Biological Psychiatry

Reply to: Sample Size, Model Robustness, and Classification Accuracy in Diagnostic Multivariate Neuroimaging Analyses

To the Editor:

In our recent publication, we report a meta-analysis of the diagnostic performance of neuroimaging-based classification models for the differentiation of patients with a depressive disorder from healthy control subjects (1). In summary, our results indicate that across studies patients can be identified with an estimated accuracy of 77%. Moderator analysis provided some evidence for effects of moderating factors, including neuroimaging modality. However, despite theoretical arguments and similar findings in comparable analyses in schizophrenia (2,3), we did not find evidence for an effect of sample size on classification accuracy.

In their comment on Kambeitz *et al.* (1), Neuhaus and Popescu (4) discuss the potential role of sample size on classification accuracy in neuroimaging-based diagnostic models for psychiatric disorders such as schizophrenia, attention-deficit/hyperactivity disorder, and depression. They report significant correlations between sample size and diagnostic accuracy for schizophrenia and attention-deficit/hyperactivity disorder and a nonsignificant correlation (of comparable magnitude) for depression. Thus, they conclude that small sample sizes might lead to overly optimistic estimates of the classification accuracy of such models.

We thank Neuhaus and Popescu (4) for their important and valuable comment about our recent work. In general, we agree with their view that small sample sizes are associated with an increased risk of positive bias. To assess this possibility, we did test the influence of sample size within our bivariate meta-analytic model. Consistent with Neuhaus and Popescu's report (4), we did not find a significant effect, but we discussed the small sample sizes of the included studies as a potential limitation. We found that other factors did have a significant effect on classification accuracy such as neuroimaging modality and age in our analysis of depression (1) and neuroimaging modality, age, medication, and clinical symptoms in our previous work on schizophrenia (5). Thus, while an effect of sample size cannot be ruled out [e.g., as Neuhaus and Popescu (4) suggest, because of a smaller number of available studies], other factors might be equally important when assessing the robustness of neuroimaging-based classification models. While our results overall are consistent with those of Neuhaus and Popescu (4), we would like to use this opportunity highlight several additional methodological aspects that are relevant in this context.

Given the recent advent of machine learning methods and their application in psychiatry, we expect that there will be a growing number of publications using these methodologies and, as a result, an increasing need for quantitative and qualitative synthesis of the available evidence (e.g., in the form of meta-analyses). Because studies of diagnostic tests typically use different outcome measures (sensitivity, specificity, accuracy, etc.), conducting a meta-analysis poses the challenge of deriving a common metric from each study. If possible, we recommend that researchers avoid conducting meta-analyses that are based on simple diagnostic accuracies or related measures such as the diagnostic odds ratio (6). Instead, models such as bivariate meta-analytic models allow researchers to preserve the two-dimensional nature of the data (sensitivity and specificity), which is important given that sensitivity and specificity are often negatively correlated within studies, and that ignoring this dependency will lead to bias (7). Moreover, averaging sensitivities with specificities to derive accuracies might make it difficult to assess the potential of the diagnostic models to be used in clinical practice (8). Bivariate meta-analytic models allow the separate estimation of between-study heterogeneity of sensitivity and specificity and the assessment of the correlation between sensitivity and specificity. Another important aspect when analyzing diagnostic models concerns the distribution of the outcome metric. Accuracy measures are subject to an upper bound (100%); thus, it might not be reasonable to assume a normal distribution or a linear relation as is required for the application of correlational analysis. The nonlinearity of the association between accuracy and sample size is nicely demonstrated in the recent work of Schnack and Kahn (2). There are several options to circumvent this problem, such as logit transformation in the context of the bivariate meta-analytic model (8) or the natural logarithm in the context of the hierarchical summary receiver operating characteristic model (7). Finally, given the importance of analyzing sensitivity and specificity in parallel in a bivariate meta-analytic model and the presence of further moderator effects (e.g., neuroimaging modality), it is advisable that researchers investigate moderator effects by using meta-regression analysis instead of isolated correlational analysis. This allows the investigation of moderator effects separately on sensitivity and specificity (5,8) and the interaction of multiple moderators.

Lastly, the role of sample size can be directly investigated on the level of the original study by systematically undersampling data and testing the effect on classification performance. This approach would allow researchers to derive the highly needed recommendations for minimum sample sizes that provide reliable performance estimates. Most importantly, it is expected that the role of sample size is problem-dependent and will depend on a multitude of factors such as the disorder under investigation [see Neuhaus and Popescu (4)], the prediction target (e.g., diagnosis, treatment response, or functional outcome), the classification algorithm used, and the neuroimaging modality or the sample heterogeneity (e.g., different research sites or different magnetic resonance imaging scanners).

In summary, we would like to emphasize the importance of statistical methodology when assessing the performance of diagnostic classification models. In addition, we are in full agreement with the arguments put forward by Neuhaus and Popescu (4) that small sample sizes represent a potential risk for the assessment of diagnostic models. The key to overcoming this potential hazard is the most rigorous validation of the generated models in independent samples and the strict separation of train- and test-samples (e.g., as is done in

https://doi.org/10.1016/j.biopsych.2018.01.023
ISSN: 0006-3223

cross-validation). Most importantly, following reporting standards for studies of diagnostic tests (9) will allow efficient synthesis of studies and will facilitate the efficient development of machine learning in psychiatry.

Joseph Kambeitz
Carlos Cabral
Matthew D. Sacchet
Ian H. Gotlib
Roland Zahn
Mauricio H. Serpa
Martin Walter
Peter Falkai
Nikolaos Koutsouleris

Acknowledgments and Disclosures

This work was supported by National Institute of Mental Health Grant No. MH101495 (to IG) and Friedrich-Baur Stiftung, the Förderung Forschung und Lehre (881/856), and the German Research Foundation Grant No. KA 4413/1-1 (to JK).

The authors report no biomedical financial interests or potential conflicts of interest.

Article Information

From the Department of Psychiatry (JK, CC, PF, NK), Ludwig-Maximilians University Munich, Munich, Germany; Department of Psychiatry and Behavioral Sciences (MDS) and the Neurosciences Program and Department of Psychology (IHG), Stanford University, Stanford, California; Institute of Psychiatry (RZ), King's College London, London, United Kingdom; Laboratory of Psychiatric Neuroimaging Institute (MHS), Department of Psychiatry (MHS), and the Center for Interdisciplinary Research on Applied Neurosciences (MHS), University of São Paulo, São Paulo, Brazil; and the Clinical Affective Neuroimaging Laboratory (MW), Department of Behavioural Neurology, Leibniz Institute for Neurobiology, Magdeburg, and Department of Psychiatry and Psychotherapy (MW), Eberhard Karls University, Tuebingen, Germany.

Address correspondence to Joseph Kambeitz, M.D., Department of Psychiatry, Ludwig-Maximilians University Munich, Nußbaumstrasse 7, 80336 Munich, Germany; E-mail: joseph.kambeitz@med.uni-muenchen.de.

See also associated correspondence: https://doi.org/10.1016/j.biopsych.2017.09.032.

Received Jan 30, 2018; accepted Jan 31, 2018.

References

1. Kambeitz J, Cabral C, Sacchet MD, Gotlib IH, Zahn R, Serpa MH, *et al.* (2017): Detecting neuroimaging biomarkers for depression: A meta-analysis of multivariate pattern recognition studies. Biol Psychiatry 82:330–338.
2. Schnack HG, Kahn RS (2016): Detecting neuroimaging biomarkers for psychiatric disorders: Sample size matters. Front Psychiatry 7:50.
3. Neuhaus AH, Popescu FC (2018): Impact of sample size and matching on single-subject classification of schizophrenia: A meta-analysis. Schizophr Res 192:479–480.
4. Neuhaus AH, Popescu FC (2018): Sample size, model robustness, and classification accuracy in diagnostic multivariate neuroimaging analyses. Biol Psychiatry 84:e81–e82.
5. Kambeitz J, Kambeitz-Ilankovic L, Leucht S, Wood S, Davatzikos C, Malchow B, *et al.* (2015): Detecting neuroimaging biomarkers for schizophrenia: A meta-analysis of multivariate pattern recognition studies. Neuropsychopharmacology 40:1742–1751.
6. Lee J, Kim KW, Choi SH, Huh J, Park SH (2015): Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: A practical review for clinical researchers-part II. Statistical methods of meta-analysis. Korean J Radiol 16:1188–1196.
7. Rutter CM, Gatsonis CA (2001): A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 20:2865–2884.
8. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH (2005): Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 58:982–990.
9. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, *et al.* (2015): STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. BMJ 351:h5527.