medRxiv preprint doi: https://doi.org/10.1101/2023.02.27.23286311; this version posted February 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

A Systematic Evaluation of Machine Learning-based Biomarkers for Major Depressive Disorder across Modalities

Nils R. Winter[1,2,✉], Julian Blanke[1], Ramona Leenings[1,3], Jan Ernsting[1,3,4], Lukas Fisch[1], Kelvin Sarink[1], Carlotta Barkhau[1], Katharina Thiel[1], Kira Flinkenflügel[1], Alexandra Winter[1], Janik Goltermann[1], Susanne Meinert[1,5], Katharina Dohm[1], Jonathan Repple[6,1], Marius Gruber[1,6], Elisabeth J. Leehr[1], Nils Opel[1,7,8,9], Dominik Grotegerd[1], Ronny Redlich[1,9,10], Robert Nitsch[2,5], Jochen Bauer[11], Walter Heindel[11], Joachim Groß[2,12], Till F. M. Andlauer[13], Andreas J. Forstner[14,15], Markus M. Nöthen[14], Marcella Rietschel[16], Stefan G. Hofmann[17,18], Julia-Katharina Pfarr[19,20], Lea Teutenberg[19,20], Paula Usemann[19,20], Florian Thomas-Odenthal[19,20], Adrian Wroblewski[19,20], Katharina Brosch[19,20], Frederike Stein[19,20], Andreas Jansen[19,20,21], Hamidreza Jamalabadi[19], Nina Alexander[19], Benjamin Straube[19], Igor Nenadić[19,20], Tilo Kircher[19,20], Udo Dannlowski[1,2*], and Tim Hahn[1,2*]

[1]University of Münster, Institute for Translational Psychiatry, Münster, Germany
[2]University of Münster, Otto Creutzfeldt Center for Cognitive and Behavioral Neuroscience, Münster, Germany
[3]University of Münster, Faculty of Mathematics and Computer Science, Münster, Germany
[4]Institute for Geoinformatics, University of Münster, Münster, Germany
[5]University of Münster, Institute for Translational Neuroscience, Münster, Germany
[6]Department of Psychiatry, Psychosomatic Medicine and Psychotherapy, University Hospital Frankfurt, Goethe University, Germany
[7]Department of Psychiatry and Psychotherapy, University Hospital Jena, Jena, Germany
[8]German Center for Mental Health (DZPG), Site Jena-Magdeburg-Halle, Germany
[9]Center for Intervention and Research on adaptive and maladaptive brain Circuits underlying mental health (C-I-R-C), Jena-Magdeburg-Halle, Germany
[10]Institute of Psychology, University of Halle, Halle, Germany
[11]University of Münster, Department of Clinical Radiology, Münster, Germany
[12]University of Münster, Institute for Biomagnetism and Biosignalanalysis, Münster, Germany
[13]Department of Neurology, Klinikum rechts der Isar, School of Medicine, Technical University of Munich, Munich, Germany
[14]Institute of Human Genetics, University of Bonn, School of Medicine and University Hospital Bonn, Bonn, Germany
[15]Institute of Neuroscience and Medicine (INM-1), Research Centre Jülich, Jülich, Germany
[16]Department of Genetic Epidemiology, Central Institute of Mental Health, Faculty of Medicine Mannheim, University of Heidelberg, Mannheim, Germany
[17]Department of Clinical Psychology, Philipps-University Marburg, Marburg, Germany
[18]Department of Psychological and Brain Sciences, Boston University, Boston, MA, USA
[19]Department of Psychiatry and Psychotherapy, Philipps-University Marburg, Marburg, Germany
[20]Center for Mind, Brain and Behavior (CMBB), Marburg, Germany
[21]Core Facility Brain Imaging, Faculty of Medicine, Philipps-University Marburg, Marburg, Germany

Background Biological psychiatry aims to understand mental disorders in terms of altered neurobiological pathways. However, for one of the most prevalent and disabling mental disorders, Major Depressive Disorder (MDD), patients only marginally differ from healthy individuals on the group-level. Whether Precision Psychiatry can solve this discrepancy and provide specific, reliable biomarkers remains unclear as current Machine Learning (ML) studies suffer from shortcomings pertaining to methods and data, which lead to substantial over- as well as underestimation of true model accuracy.

Methods Addressing these issues, we quantify classification accuracy on a single-subject level in N=1,801 patients with MDD and healthy controls employing an extensive multivariate approach across a comprehensive range of neuroimaging modalities in a well-curated cohort, including structural and functional Magnetic Resonance Imaging, Diffusion Tensor Imaging as well as a polygenic risk score for depression.

Findings Training and testing a total of 2.4 million ML models, we find accuracies for diagnostic classification between 48.1% and 62.0%. Multimodal data integration of all neuroimaging modalities does not improve model performance. Similarly, training ML models on individuals stratified based on age, sex, or remission status does not lead to better classification. Even under simulated conditions of perfect reliability, performance does not substantially improve. Importantly, model error analysis identifies symptom severity as one potential target for MDD subgroup identification.

Interpretation Although multivariate neuroimaging markers increase predictive power compared to univariate analyses, single-subject classification – even under conditions of extensive, best-practice Machine Learning optimization in a large, harmonized sample of patients diagnosed using state-of-the-art clinical assessments – does not reach clinically relevant performance. Based on this evidence, we sketch a course of action for Precision Psychiatry and future MDD biomarker research.

Funding The German Research Foundation, and the Interdisciplinary Centre for Clinical Research of the University of Münster.

Major Depression | Neuroimaging | Biomarker | Machine Learning
Correspondence: ***nils.r.winter@uni-muenster.de***

Introduction

Overcoming Cartesian mind-body dualism was the pivotal achievement of biological psychiatry in the 20th century, enabling the treatment of mental disorders as disorders of the brain.[1] Since the effectiveness of physical interventions such as neuropsychopharmacological treatments as well as the substantial heritability of many psychiatric disorders in principle support this dogma, hopes are high for biomarkers to inform diagnosis and treatment. However, identifying specific, reliable neurobiological deviations informative on the level of the individual patient has proven elu-

*These authors contributed equally to this work

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.02.27.23286311; this version posted February 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

sive even after decades of intense research, with the clinical reality of patients remaining largely unchanged.[2, 3] For Major Depressive Disorder (MDD) mounting evidence suggests that group-level, univariate neuroimaging or genetic markers only marginally differ between healthy controls and patients with MDD, with the distributions of patients and controls overlapping more than 85% even under optimal conditions.[4–6] Fuelled by the availability of large-scale datasets as well as substantial improvements regarding Machine Learning (ML) software and hardware, the field of Precision Psychiatry has gained increasing traction over the last decade. Precision Psychiatry aims to build models which allow for individual predictions, thereby moving from the investigation of univariate statistical group differences towards multivariate neurobiological patterns of individual patients. This focus on prediction and prognosis instead of group-level inference as well as the ability for a direct assessment of clinical utility renders Precision Psychiatry essential in all translational efforts.[7–11] While a consensus on best-practice guidelines for Precision Psychiatry and ML has been emerging[7, 10, 11], four broad issues in MDD biomarker research remain which may lead to substantial over- as well as underestimation of the true predictive performance: First, methodological shortcomings in predictive model validation (e.g. data leakage between training and test set) lead to an overestimation of predictive performance in many publications.[12] Strikingly, about a quarter of all published studies using predictive models in psychiatry do not provide any kind of model validation and thus do not provide any information regarding predictive performance in new patients.[13] In the same vein, small sample sizes for model evaluation, such as those most common in the literature today, often result in unreliable and eventually inflated estimates of predictive performance.[14] Second, many published studies rely on a single ML algorithm; often without optimizing model performance through hyperparameter tuning, thereby running the risk of greatly underestimating true predictive performance.[15] Third, current studies almost exclusively focus on a single data modality and studies integrating multiple modalities to increase predictive performance are rare.[7, 15] Fourth, clinical assessment of MDD diagnosis across studies is inconsistent and especially for larger studies often relies on self-report questionnaires rather than clinical interviews by a trained clinician, thus rendering diagnostic labels more heterogeneous and less reliable.[16, 17] Similarly, a lack of harmonization of study protocols, resulting in clinical heterogeneity of patient samples and recruitment modalities, quality control, and neuroimaging data acquisition in multi-site analyses has previously been used to explain small effect sizes and inconsistent results.[18] In summary, the existing literature on multivariate biomarker discovery in MDD does not allow for a conclusive evaluation of clinical utility of ML approaches. Here, we explicitly address these previous shortcomings to systematically evaluate ML-based multivariate biomarkers for MDD across neuroimaging modalities: We performed nested cross-validation to separate the model optimization step from the estimation of generalizability and ensured adequate test sets by using one of the largest single-study MDD cohorts for which multimodal data and in-depth diagnostic assessment is available (N=1,801 MDD patients and controls).[19, 20] Next, we did not rely on a single predictive algorithm, but capitalized on the advances in ML software[21, 22] and computational capabilities to combine multiple classifiers from complementary algorithmic categories including feature selection, dimensionality reduction, and extensive tuning of model hyperparameters, resulting in a total of 2.4 million machine learning models trained and evaluated in this study. Expanding previous work, we drew upon a comprehensive set of neuroimaging modalities including structural Magnetic Resonance Imaging (MRI), task-based and resting-state functional MRI (fMRI), Diffusion Tensor Imaging (DTI) as well as an MDD polygenic risk score and several environmental risk factors. This allowed us to directly compare predictive performance across modalities in the same sample and enabled us to quantify the potential benefit of multimodal data integration. In addition, clinical assessment of patients in our data was based on structured clinical interviews (SCID) which provided standardized DSM-based MDD diagnosis and therefore reduced the diagnostic uncertainty often hampering model performance in large-scale, multi-site data today. Likewise, methodological heterogeneity due to, e.g. differing exclusion criteria, recruitment modalities, clinical phenotyping, or MRI scanning protocols, could be alleviated in this well-curated, harmonized sample.[20] Finally, the low reliability of neuroimaging data and psychiatric diagnosis is being discussed as one of the major drivers for small effect sizes currently reported in the literature.[17, 23–26] To address this hypothesis, we systematically simulated classification performance in scenarios of optimal reliability and quantified expected improvements. Considering the substantial heterogeneity of patients with MDD, we finally conducted in-depth analyses of model errors to uncover characteristics of patients that contribute to misclassification, thereby shedding light on subgroups for which neuroimaging-based predictive models are successful or might fail.[27–29]

Methods

Study design and participants. The data used in this study are part of the Marburg-Münster Affective Disorders Cohort Study (MACS).[19, 20] Data were collected at two sites (Marburg and Münster, Germany) using identical study protocols and harmonized scanner settings.[20] The study was approved by the ethics committees of the medical faculties of the University of Marburg, Germany, and the University of Münster, Germany. Participants received financial compensation and gave written and informed consent. At the time of data analysis, a sample of N=2,036 healthy participants and patients with major depression were recruited as part of the MACS cohort (*eMethods 1-3*). Clinical diagnosis was assessed using the Structured Clinical Interview for DSM-IV, axis 1 disorders (SCID-I).[30] Patients were recruited from local in- and outpatient services and either fulfilled the DSM-IV criteria for an acute major depressive episode or had a

PREPRINT

medRxiv preprint doi: https://doi.org/10.1101/2023.02.27.23286311; this version posted February 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

lifetime history of a major depressive episode. Individuals with any history of neurological or medical conditions were excluded, resulting in a final sample of N=1,801. See *eMethods 1* for further information on exclusion criteria. Participants were recruited from September 11, 2014, to September 26, 2018. For every neuroimaging data modality, all participants for whom data of the specific modality were available and passed quality checks were used in subsequent analyses (see *eMethods 1 and 4-12*). This study followed Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines.[31]

Procedures and neuroimaging data modalities. The neuroimaging, genetic and behavioural data used in this study have been described previously.[5] Detailed information is available in *eMethods 4-12*. In short, voxel-based morphometry (VBM, CAT12 toolbox) and region-based surface, thickness and volume (FreeSurfer) were extracted from T_1-weighted structural MRI.[32, 33] Structural connectomes were derived from DTI as fractional anisotropy (FA) and mean diffusivity (MD).[34] Functional connectomes were derived from resting-state functional MRI (rsfMRI). Voxel-based local correlation (LCOR), the amplitude of low-frequency fluctuations (ALFF) as well as the fractional amplitude of low-frequency fluctuations (fALFF) were also computed from rsfMRI.[35] For both structural and functional connectomes commonly used graph network parameters such as betweenness centrality, degree centrality, or global efficiency were calculated.[36] Task-based fMRI was based on an established emotional face matching paradigm and a faces versus shapes contrast was used.[37, 38] In addition, we compared results to a commonly used polygenic risk score for depression (PRS, *eMethods 5*)[39, 40] as well as questionnaire data on adverse experiences during childhood (Childhood Trauma Questionnaire; CTQ) and current social support (F-SozU), since these variables are established risk or protective factors in the aetiology of major depression.[39, 41, 42] A medication load index was calculated expressing the current psychiatric medication. Current depressive symptoms were assessed using the Beck Depression Inventory (BDI) and Hamilton Depression Rating Scale (HAMD).[43, 44]

Choice of the primary measures. Accuracy of predicted diagnostic labels in all machine learning models was calculated using the widely used balanced classification accuracy (BACC), sensitivity, specificity, and area under the receiver operating characteristic curve (AUC), following STAR*D guidelines for reporting predictive accuracy. In addition, we calculated Matthew's correlation coefficient (MCC, Equation 1). For all metrics, mean and standard deviation across the 10 outer cross-validation splits were reported to assess the generalizability of the predictive models.

$$MCC = \frac{Cov(y, \hat{y})}{\sigma_y \cdot \sigma_{\hat{y}}} \quad \textbf{(1)}$$

Machine Learning analyses. A total of 2.4 million machine learning models to classify healthy participants and patients with MDD were trained, optimized and evaluated (see Figure 1, *eMethods 14*). A single ML pipeline consisted of a sequence of data transformation steps and a final classification algorithm. Data transformations included an imputation of missing data, a feature normalization, selection of a percentage of univariate features with the highest effect size, and a principal component analysis (PCA) to reduce the dimensionality of the brain data. Subsequently, a classification algorithm was trained to predict diagnosis, including support vector machines, random forests, logistic regression, k-nearest neighbour, Gaussian naive Bayes, and boosting classifiers. A nested cross-validation scheme with 10 inner validation and 10 outer test splits was used to optimize hyperparameters and assess final generalizability. These primary ML analyses were complemented by analyses for subgroups of acutely depressed patients (omitting remitted patients) or recurrently depressed patients (omitting single episode patients), males and females, as well as a homogeneous age group (age range 24 to 28). For more details see *eMethods 3*.

Modality integration. Brain modality integration was accomplished using two strategies. First, a PCA was performed for every data modality separately and the resulting components were then concatenated and used as input to the ML pipelines. Second, a voting ensemble strategy was used combining all diagnosis predictions from the unimodal models. Final predictions were calculated using a majority vote. All ML analyses were performed using PHOTONAI.[22] Scripts are available at https://github.com/wwu-mmll.

Simulation of perfect reliability. To quantify the effect of reliability on classification performance, we performed exploratory analyses using attenuation correction from classical test theory to simulate the true classification accuracy occurring if the reliability of the data was perfect.[45] We first computed MCC from the model predictions $\hat{y}$ and the actual diagnostic labels y (Equation 1).[46] This correlation was then corrected for an assumed reliability ρ using the attenuation formula (Equation 2).[47]

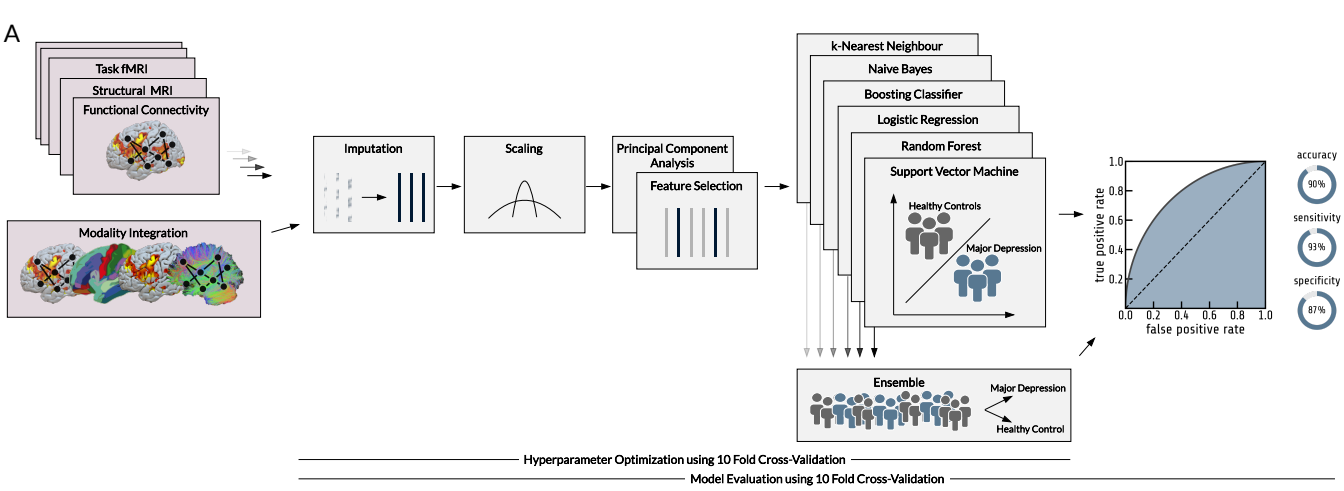
$$MCC_{corr} = \frac{MCC}{\sqrt{\rho}} \quad \textbf{(2)}$$

We conducted two separate attenuation correction analyses. First, we assume a reliability of $\rho_y = 0.28$ for an MDD diagnosis, which is based on the current literature on the interrater reliability of DSM-5 diagnoses.[17, 26] Second, we assumed reliabilities for neuroimaging data ranging from 0.1 to 1. The resulting corrected correlations were then converted back to BACC using prevalence ϕ and bias β with equations 15 and 21 in [46] (Equation 3, *eMethods 13*).

$$BACC = \frac{1}{2 \cdot \sqrt{\frac{\phi - \phi^2}{\beta - \beta^2}}} \cdot MCC + \frac{1}{2} \quad \textbf{(3)}$$

PREPRINT

medRxiv preprint doi: https://doi.org/10.1101/2023.02.27.23286311; this version posted February 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .



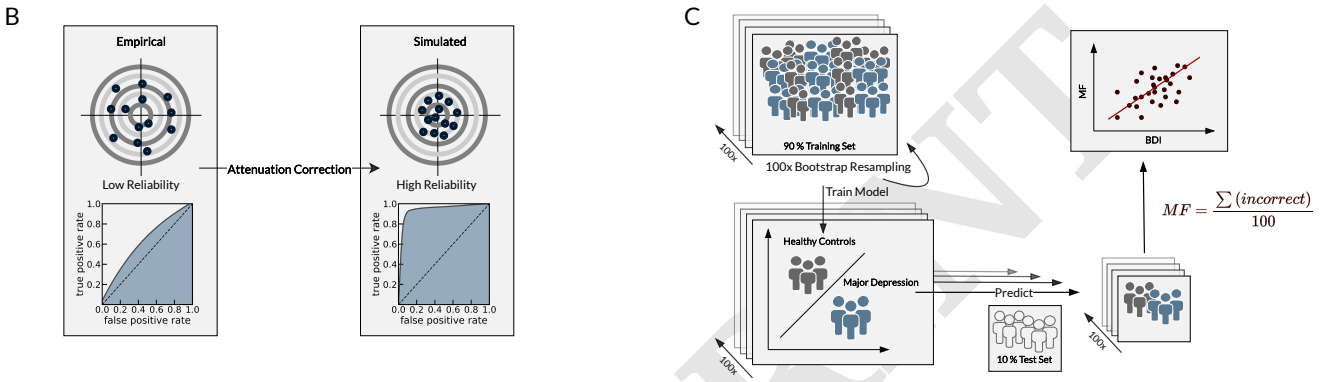


Fig. 1. Overview of all analyses. (A) illustrates steps of the Machine Learning pipeline. (B) illustrates reliability correction and its effect on classification accuracy. (C) illustrates model error analysis using misclassification frequency (MF) through repeated bootstrapping.

Analysis of systematic model error. Identifying subgroups of individuals for whom brain-based ML models routinely fail has been shown to improve the development of generalizable predictive models.[27] In order to quantify this tendency for misclassification in every individual, we performed 100 bootstrap resampling runs on the training set of the best performing neuroimaging modality. One ML pipeline for every bootstrap training set was then trained and diagnostic labels for the participants in the test set were collected, resulting in 100 predictions (healthy, depressed) for every participant. The sum over incorrect classifications then leads to the frequency of misclassification (MF).[27] Finally, MF was correlated with external measures describing depressive symptom severity and demographic or environmental characteristics using Spearman rank correlation.

Role of the funding source. The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

A total of 1,801 individuals (856 patients [47.5%] and 945 healthy controls [52.5%]) were included in the analyses (mean [SD] age, 36.1 [13.1] years; 555 female patients [64.8%] and 607 female healthy controls [64.2%], see Table 1 for details).

Multivariate classification accuracy. Across neuroimaging modalities and ML algorithms, BACC ranged between 48.1% and 61.5% (see eTable 1-2 detailed results and *eMethods* for neuroimaging feature descriptions). Results for the single best ML algorithm in each modality are shown in Figure 2. Highest BACC was found for resting-state connectivity, with mean [SD] BACC ranging between 51.5% [7.1%] and 61.5% [3.4%]. Structural MRI as well as task-based fMRI showed lower BACC compared to all resting-state fMRI modalities. Calculating graph network parameters from DTI or resting-state fMRI did not improve overall BACC compared to using the functional or structural connectome directly. To investigate the effect of remission status and chronicity of the MDD population, we performed additional analyses limited to, first, MDD patients with acute symptoms (N=599) thus excluding remitted patients and, second, MDD patients with recurrent episodes (N=297). Overall, ML pipelines on subgroups did not outperform the analysis containing all MDD patients ($BACC_{max}$=61.7%). Likewise, restricting analyses to male or female individuals or a more homogeneous age range of 24 to 28 did not change the overall results ($BACC_{max}$=61.6%, see eFigure 1-5 and eTables 5-19).

PREPRINT

medRxiv preprint doi: https://doi.org/10.1101/2023.02.27.23286311; this version posted February 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 1. Socio-demographics and clinical characteristics of all participants.

	Healthy	Major Depression	Difference
Sex			0.83
Male	338 (35.8%)	301 (35.2%)	
Female	607 (64.2%)	555 (64.8%)	
Age	34.40 (13.01)	36.76 (13.27)	0.001
HAMD	1.45 (2.18)	9.38 (7.17)	0.001
BDI	4.11 (4.27)	17.58 (11.02)	0.001
CTQ	32.59 (8.57)	45.06 (15.92)	0.001
Social Support	4.51 (0.54)	3.77 (0.87)	0.001
Medication Load Index		1.35 (1.48)	
Number of previous inpatient treatments		1.58 (2.08)	
Number of previous depressive episodes		3.99 (6.75)	
Total duration of previous inpatient treatments (in weeks)		11.95 (18.89)	
Total duration of all previous depressive episodes (in months)		45.36 (69.18)	
Comorbid diagnoses			
Any comorbid diagnosis		373 (43.6%)	
Anxiety disorder		269 (31.4%)	
Eating disorder		50 (5.8%)	
Dysthymic disorder		43 (5.0%)	
Substance use disorder		37 (0.8%)	
Somatic symptom disorder		27 (3.2%)	
Psychotic disorder		7 (0.8%)	

HAMD=Hamilton Rating Scale for Depression. BDI=Beck Depression Inventory. CTQ=Childhood Trauma Questionnaire. MRI=Magnetic Resonance Imaging. VBM=Voxel-Based Morphometry. *t or χ^2 tests. Lifetime comorbidities were derived from the structured clinical interview for DSM-IV (SCID). Multiple comorbidities were possible for any MDD patient.

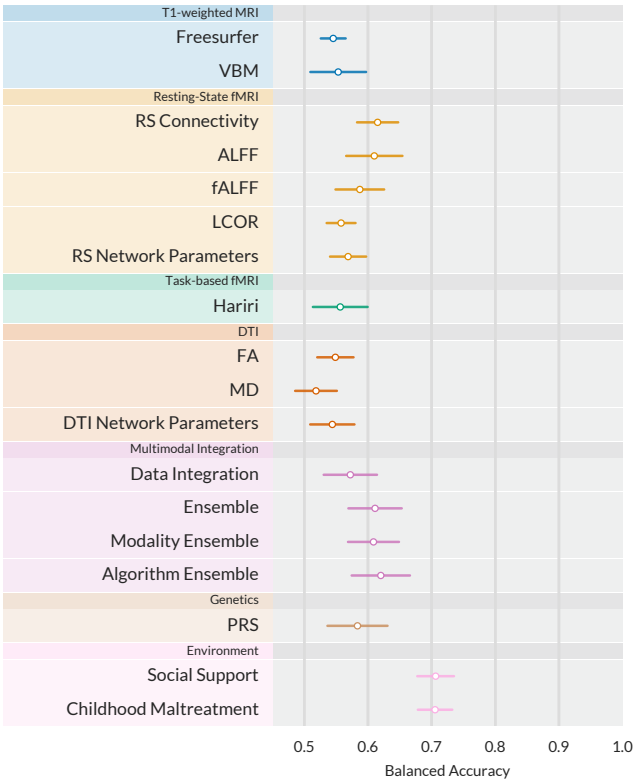


Fig. 2. Balanced accuracy for best machine learning pipeline in every modality. Error bars display +-1 standard deviation calculated across the 10 outer cross-validation folds. VBM=Voxel-based morphometry, ALFF=Amplitude of low-frequency fluctuations, fALFF=fractional ALFF, LCOR=Local correlation, FA=Fractional anisotropy, MD=Mean diffusivity, PRS=Polygenic risk score.

Multimodal Data Integration. Integration of neuroimaging modalities was evaluated using two alternative approaches. First, principal components from modality specific PCAs were concatenated and used as input to the previously described ML pipelines. This modality integration analysis achieved BACCs between 50.1% [4.0%] and 57.2% [4.4%] (eTable 3, Figure 2). Second, predicted labels from the unimodal models (across algorithm, across modalities, or across both) were combined into a majority-vote ensemble classifier. The voting ensemble classifier achieved a BACC of 61.1% [4.4%]. Both multimodal data integration methods did not improve the 61.5% accuracy reached in the best unimodal model. Combining predictions from all ALFF models achieved the highest BACC of 62.0% [4.8%].

Comparison with Genetic and Environmental Variables. We next compared the neuroimaging-based ML models to the predictive performance of univariate approaches using genetic and environmental variables. While the Howard et al. depression PRS[39] achieved similar results to neuroimaging (BACC = 58.4% [5.0%]), both self-reported childhood maltreatment and social support outperformed brain-based and PRS-based models, achieving a BACC of 70.5% [2.9%] and 70.6% [3.0%], respectively.

Effects of Reliability of Diagnosis and Neuroimaging Data. To investigate to what extent the reliability of neuroimaging data and diagnosis affect classification accuracy, we first converted BACC to MCC as a measure of the as-

medRxiv preprint doi: https://doi.org/10.1101/2023.02.27.23286311; this version posted February 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

sociation between the actual and predicted diagnostic label. This correlation coefficient could then be corrected using the attenuation correction formula, estimating classification performance given perfect reliability. Second, we corrected for the lower bound of the MDD diagnosis reliability of $\rho = 0.28$ as reported in the literature (Figure 3A). With this approach, BACC for the best machine learning algorithm on resting-state connectivity increased to 71.8% [6.4%]. BACC for the voting ensemble increased to 73.4% [7.4%]. Third, we assumed reliability coefficients of neuroimaging modalities between 0.1 and 1 (Figure 3B). For the best unimodal analysis (resting-state connectivity), BACC increases to 66.3% for an assumed reliability of 0.5. These reliability correction analyses suggest that improving reliability might only have a minor positive effect on classification accuracy.

Analysis of Systematic Model Errors. Recent work has shown that brain-based predictive models do not work equally well for all individuals, potentially displaying substantial bias.[27] Therefore, identifying patients for which ML models repeatedly fail will be informative for the development of clinical useful predictive models. The frequency with which each individual was incorrectly classified as either healthy or depressed was measured using the misclassification frequency (MF) based on the modality which achieved the highest performance in the unimodal analyses (rsfMRI connectivity). MF was significantly correlated with symptom severity in patients with depression (eTable 4). A higher score in current depressive symptom levels (BDI, HAMD) as well as a higher number of previous hospitalizations were associated with fewer misclassifications (BDI: n=621, r=-0.15, p<0.001; HAMD: n=628, r=-0.20, p<0.001, number of hospitalizations: n=622, r=-.10, p=0.01), showing that patients with more severe current depressive symptoms and a more unfavourable previous disease course were correctly classified as patients more often. Likewise, a higher score in global assessment of functioning (GAF) in patients and a lower GAF score in healthy controls was associated with more misclassifications (HC: n=690, r=-.10, p=0.007; MDD: n=620, r=.17, p<0.001). A higher medication load in patients was also associated with fewer misclassifications (n=631, r=-.21, p<0.001). Furthermore, a higher number of misclassifications was apparent in remitted patients compared to patients with acute depressive symptoms ($F_{1,628}$=7.24, p=0.007) and in patients without comorbidities compared to patients with comorbidities ($F_{1,628}$=7.79, p=0.005).

Discussion

Extending recent evidence in neuroimaging and other neurobiological research domains showing that univariate group-level differences between patients with MDD and healthy controls are small[5], we aimed to systematically evaluate ML approaches classifying patients and healthy controls on the basis of multivariate neuroimaging signatures. Importantly, we directly addressed the limitations of existing ML studies which have led to over- and underestimation of model performance, providing a much more accurate assessment of the potential of current predictive models in MDD diagnosis. In summary, training and testing a total of 2.4 million ML models on a large, harmonized sample, accuracy for predicting MDD diagnosis did not exceed 62%. Although slightly improving the 56-58% classification accuracy achieved using univariate neuroimaging and genetic markers[5], this systematic evaluation of multivariate methods revealed a disconcerting discrepancy to existing proof-of-concept studies, yielding considerably lower predictive accuracy than previously expected.[12] Our study provides four main improvements over existing ML studies: First, we reduced the common risk of producing systematically inflated predictive performance estimates due to data leakage and/or small test set size[14] by using nested cross-validation in a large sample of N=1,801 patients and controls, ensuring independent and sufficiently large test sets. Our results thus point towards small (test) sample sizes as a major driver in the current overestimation of neuroimaging-based predictability of MDD diagnosis.[12] Second, whereas previous studies mostly relied on single ML models, e.g., Support Vector Machines only, and did thus not systematically explore the space of possible ML pipelines, we employed an extensive multivariate approach providing substantially improved coverage of algorithmic search space. In addition, we also extensively sampled the hyperparameters for each algorithm. Despite these considerable efforts, classification accuracy still falls short of expectations. Note that we focused on classical ML algorithms and did not investigate more complex models e.g. based on Deep Learning.[48] Although deep learning (DL) has revolutionized ML applications, model performance will only improve if the data have nonlinear relationships exploitable at the available sample sizes, yet linear models have shown to perform on par with more complex DL approaches for structural and functional MRI up to sample sizes >10,000 subjects.[49] Future studies should, however, combine clinical samples with modality-specific, large-scale data of healthy controls (as available e.g. from the UK Biobank or ENIGMA) using for example self-supervised learning, transfer learning or semi-supervised learning approaches to increase sample size to tens of thousands to enable the exploitation of non-linear associations. Third, capitalizing on our multimodal dataset comprising structural and functional MRI as well as DTI, we tested if classification accuracy can be boosted by integrating data from neuroimaging modalities. However, even integrating all 11 modalities using different strategies did not increase performance. This suggests that modality specific models either learn so little that combining them is irrelevant or that model predictions are so highly correlated as to render their integration redundant. The latter seems plausible given that modalities with higher accuracy also show considerable correlation among each other (r_{max} = 0.47, eFigure 6). Fourth, we addressed two major shortcomings of large, multi-site ML studies, i.e. between-site variability due to data pooling and clinical heterogeneity due to unstandardized diagnostic procedures.[50] Our harmonized sample made it possible to run ML analyses on a large sample without the need of data pooling across multiple studies and acquisition processes, ef-

PREPRINT

medRxiv preprint doi: https://doi.org/10.1101/2023.02.27.23286311; this version posted February 27, 2023. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

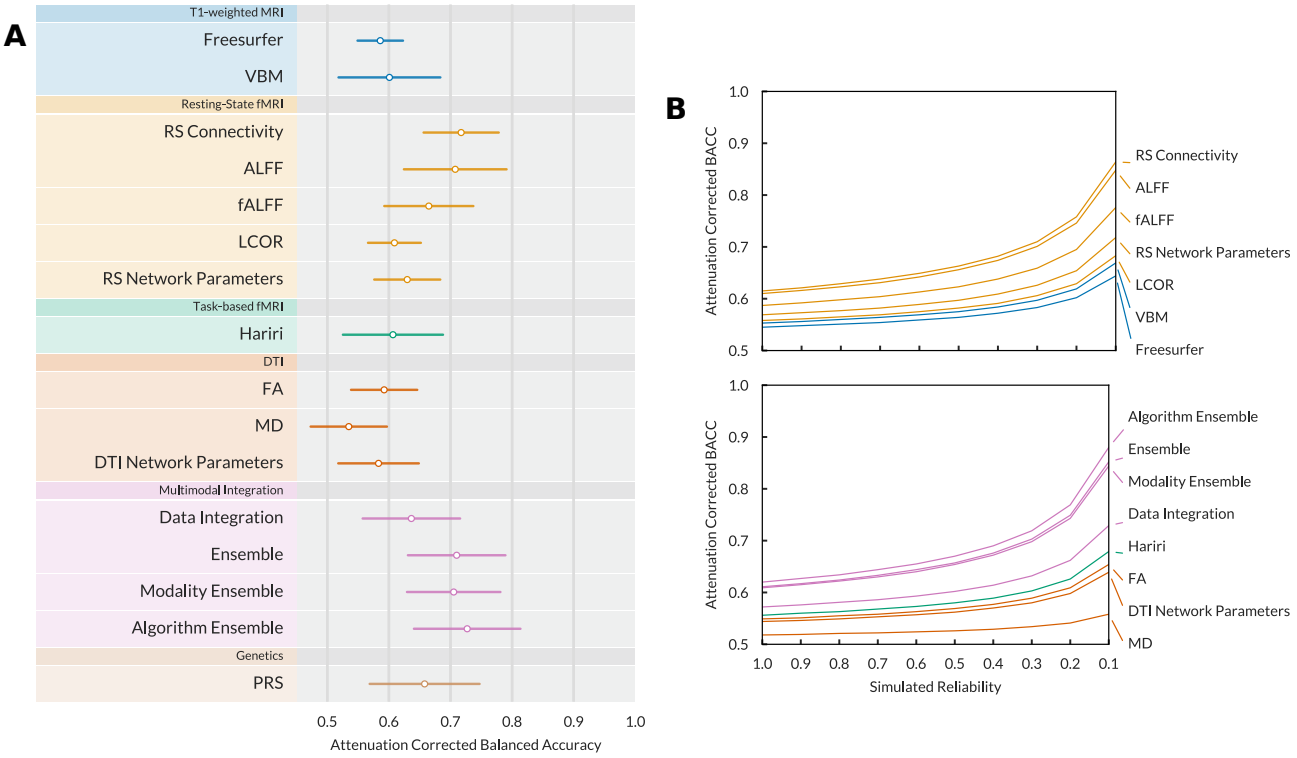


Fig. 3. (A) Balanced accuracy for best machine learning pipeline in every modality after performing an attenuation correction for the empirical reliability of the MDD diagnosis. Error bars display +-1 standard deviation calculated across the 10 outer cross-validation folds. (B) Balanced accuracy for best machine learning pipeline in every modality after performing an attenuation correction for simulated reliability of the neuroimaging data. A simulated reliability of 1 corresponds to the empirical results achieved in the unimodal analyses. Decreasing the simulated reliability results in a corrected BACC. VBM=Voxel-based morphometry, ALFF=Amplitude of low-frequency fluctuations, fALFF=fractional ALFF, LCOR=Local correlation, FA=Fractional anisotropy, MD=Mean diffusivity.

fectively minimizing methodological heterogeneity resulting from multiple scanning sites, neuroimaging preprocessing pipelines and population differences. In addition, we were able to reduce diagnostic uncertainty by relying upon structured clinical SCID interviews for MDD diagnostics. Thus, we provide evidence that low predictive performance cannot be explained by a lack of harmonization of studies or unstandardized diagnoses as previously suggested.[50]

Aiming to explain the apparent discrepancy between the popular belief in mainstream biological psychiatry that mental disorders are in fact brain disorders[2] and a lack of neurobiological manifestations of MDD informative on the level of the individual across the most commonly investigated modalities in research today, we will discuss a number of viewpoints concerning the reliability and validity of both the neuroimaging data and the conceptualization of MDD as well as the current research design.

Addressing the debate around reliability[17, 27, 51], we show that even under conditions of perfect reliability of diagnosis or neuroimaging data, clinically useful prediction on the level of the individual patient still remains elusive. Note that this approach can, by design, only simulate perfect reliability with regard to final model predictions and thus does not speak directly to the effect different data or pre-processing pipelines might have on model training.[51] Although improved reliability of neuroimaging data could potentially lead to more stable ML models, this seems unlikely given the complete lack of correlation between known reliability estimates of MRI data and our classification results.

Apart from concerns about reliability, we may also question the validity of neuroimaging data in terms of its ability to capture the neurobiological information necessary for explaining the MDD phenotype. If we assume current methods fall short in this regard, there are several research directions that could enhance our understanding of the disorder. These include higher spatial or temporal resolution, more advanced experimental paradigms or data preprocessing techniques, as well as longitudinal research designs that can model changes in an individual's neurobiology associated with current symptoms and episodes.[52, 53] Additionally, the complexity of the MDD phenotype might require a more comprehensive approach that incorporates interactions between neurobiology, the entire body, as well as the environment.[54] However, since there is no established formal theory of the neurobiology of depression, it is uncertain which neuroimaging methods will be best suited to capture clinically relevant information.[55]

On the other hand, if we assume that the information relevant for explaining behaviour and mental processes is present in current neuroimaging modalities, issues of biological validity of the MDD construct appear plausible. Since clinical heterogeneity in MDD has been extensively described[28], focusing on clinically relevant outcomes and longitudinal data, even across diagnoses, rather than MDD diagnosis itself might be better suited to yield high-accuracy predictions, e.g. associating neuroimaging markers with long-term

medRxiv preprint doi: https://doi.org/10.1101/2023.02.27.23286311; this version posted February 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

disease trajectories.[8, 56–58] Likewise, investigating symptoms rather than syndromes has been promoted lately, with network theory of psychopathology providing one conceptual framework possibly able to model symptom dynamics independent of psychiatric category.[59] Indeed, our results regarding correlations of misclassification frequency provide support for associations between symptom severity and neurobiological markers, suggesting that patients with higher levels of current symptoms, lower global functioning and more unfavourable disease courses in the past are easier to detect and correctly classify. Although providing a potential target for MDD subgroup identification beyond a more general MDD category, our complementary subgroup analyses focusing on acutely and recurrently depressed patients, respectively, did not increase predictive performance. This, however, might be due to the reduced sample sizes available during model training of depressive subgroups. Other research directions such as the Research Domain Criteria aim at identifying biologically motivated descriptions of mental disorders for which a direct link between neurobiology and cognitive processes is a necessary requirement.[60] However, the current results indicate that it might be difficult to find biological predictors for all patients currently covered under the umbrella of the MDD diagnosis.

In the same vein, a strictly reductionist case-control design in neuroimaging might be too simplistic to adequately model the complex relationship between brain and behaviour.[1, 61] Modelling complexity could be increased using e.g. normative modelling approaches that capture deviations of the individual patient, overcoming the necessity for a common biological cause across all MDD patients.[62] Similarly, identifying biotypes of mental disorders through clustering across DSM diagnoses might constitute a promising way forward.[40, 56, 63] Furthermore, given the complex, nonlinear dynamics of brain processes and symptom interactions, dynamical systems theory within computational psychiatry provides another conceptual framework that could be able to overcome simplistic reductionism and model the neurobiological complexity of MDD.[64] It also provides one way of moving towards quantitative theories of depression, e.g. network theory of psychopathology.[59] However, more research is needed to investigate whether these approaches are actually able to increase clinically relevant predictions on the level of the individual patient.

Conclusions

In summary, we show that although multivariate neuroimaging markers increase predictive performance compared to univariate analyses, classification on the level of the individual patient – even under optimal conditions – does not reach clinically relevant levels. How biological Precision Psychiatry can deliver more accurate individualized prediction to improve treatment and patient care remains a central open question at this point.

ACKNOWLEDGEMENTS

This work was funded by the German Research Foundation (DFG grants FOR2107 KI588/14-1, and KI588/14-2, and KI588/20-1, KI588/22-1 to Tilo Kircher, Marburg, Germany; STR1146/18-1 to Benjamin Straube, Marburg, Germany; HA7070/2-2, HA7070/3, and HA7070/4 to Tim Hahn, Münster, Germany; Dan3/012/17 to Udo Dannlowski) and MzH 3/020/20 from the Interdisciplinary Center for Clinical Research of the medical faculty of Münster to Tim Hahn. The project was further supported by the cluster project "The Adaptive Mind", funded by the Excellence Program of the Hessian Ministry of Higher Education, Science, Research and Art to Tilo Kircher and Benjamin Straube.

This work is part of the German multicenter consortium "Neurobiology of Affective Disorders. A translational perspective on brain structure and function", funded by the German Research Foundation (Deutsche Forschungsgemeinschaft DFG; Forschungsgruppe/Research Unit FOR2107).

Principal investigators (PIs) with respective areas of responsibility in the FOR2107 consortium are: Work Package WP1, FOR2107/MACS cohort and brainimaging: Tilo Kircher (speaker FOR2107; DFG grant numbers KI 588/14-1, KI 588/14-2), Udo Dannlowski (co-speaker FOR2107; DA 1151/5-1, DA 1151/5-2), Axel Krug (KR 3822/5-1, KR 3822/7-2), Igor Nenadic (NE 2254/1-2), Carsten Konrad (KO 4291/3-1). WP2, animal phenotyping: Markus Wöhr (WO 1732/4-1, WO 1732/4-2), Rainer Schwarting (SCHW 559/14-1, SCHW 559/14-2). WP3, miRNA: Gerhard Schratt (SCHR 1136/3-1, 1136/3-2). WP4, immunology, mitochondriae: Judith Alferink (AL 1145/5-2), Carsten Culmsee (CU 43/9-1, CU 43/9-2), Holger Garn (GA 545/5-1, GA 545/7-2). WP5, genetics: Marcella Rietschel (RI 908/11-1, RI 908/11-2), Markus Nöthen (NO 246/10-1, NO 246/10-2), Stephanie Witt (WI 3439/3-1, WI 3439/3-2). WP6, multi-method data analytics: Andreas Jansen (JA 1890/7-1, JA 1890/7-2), Tim Hahn (HA 7070/2-2), Bertram Müller-Myhsok (MU1315/8-2), Astrid Dempfle (DE 1614/3-1, DE 1614/3-2). CP1, biobank: Petra Pfefferle (PF 784/1-1, PF 784/1-2), Harald Renz (RE 737/20-1, 737/20-2). CP2, administration. Tilo Kircher (KI 588/15-1, KI 588/17-1), Udo Dannlowski (DA 1151/6-1), Carsten Konrad (KO 4291/4-1).

Data access and responsibility: All PIs take responsibility for the integrity of the respective study data and their components. All authors and coauthors had full access to all study data.

Acknowledgements and members by Work Package (WP): WP1: Henrike Bröhl, Katharina Brosch, Bruno Dietsche, Rozbeh Elahi, Jennifer Engelen, Sabine Fischer, Jessica Heinen, Svenja Klingel, Felicitas Meier, Tina Meller, Julia-Katharina Pfarr, Kai Ringwald, Torsten Sauder, Simon Schmitt, Frederike Stein, Annette Tittmar, Dilara Yüksel (Dept. of Psychiatry, Marburg University). Mechthild Wallnig, Rita Werner (Core-Facility Brainimaging, Marburg University). Carmen Schade-Brittinger, Maik Hahmann (Coordinating Centre for Clinical Trials, Marburg). Michael Putzke (Psychiatric Hospital, Friedberg). Rolf Speier, Lutz Lenhard (Psychiatric Hospital, Haina). Birgit Köhnlein (Psychiatric Practice, Marburg). Peter Wulf, Jürgen Kleebach, Achim Becker (Psychiatric Hospital Hephata, Schwalmstadt-Treysa). Ruth Bär (Care facility Bischoff, Neukirchen). Matthias Müller, Michael Franz, Siegfried Scharmann, Anja Haag, Kristina Spenner, Ulrich Ohlenschläger (Psychiatric Hospital Vitos, Marburg). Matthias Müller, Michael Franz, Bernd Kundermann (Psychiatric Hospital Vitos, Gießen). Christian Bürger, Katharina Dohm, Fanni Dzvonyar, Verena Enneking, Stella Fingas, Katharina Förster, Janik Goltermann, Dominik Grotegerd, Hannah Lemke, Susanne Meinert, Nils Opel, Ronny Redlich, Jonathan Repple, Katharina Thiel, Kordula Vorspohl, Bettina Walden, Lena Waltemate, Alexandra Winter, Dario Zaremba (Dept. of Psychiatry, University of Münster). Harald Kugel, Jochen Bauer, Walter Heindel, Birgit Vahrenkamp (Dept. of Clinical Radiology, University of Münster). Gereon Heuft, Gudrun Schneider (Dept. of Psychosomatics and Psychotherapy, University of Münster). Thomas Reker (LWL-Hospital Münster). Gisela Bartling (IPP Münster). Ulrike Buhlmann (Dept. of Clinical Psychology, University of Münster).

WP2: Marco Bartz, Miriam Becker, Christine Blöcher, Annuska Berz, Moria Braun, Ingmar Conell, Debora dalla Vecchia, Darius Dietrich, Ezgi Esen, Sophia Estel, Jens Hensen, Ruhkshona Kayumova, Theresa Kisko, Rebekka Obermeier, Anika Pützer, Nivethini Sangarapillai, Özge Sungur, Clara Raithel, Tobias Redecker, Vanessa Sandermann, Finnja Schramm, Linda Tempel, Natalie Vermehren, Jakob Vörckel, Stephan Weingarten, Maria Willadsen, Cüneyt Yildiz (Faculty of Psychology, Marburg University).

WP4: Jana Freff (Dept. of Psychiatry, University of Münster). Susanne Michels, Goutham Ganjam, Katharina Elsässer (Faculty of Pharmacy, Marburg University). Felix Ruben Picard, Nicole Löwer, Thomas Ruppersberg (Institute of Laboratory Medicine and Pathobiochemistry, Marburg University).

WP5: Helene Dukal, Christine Hohmeyer, Lennard Stütz, Viola Lahr, Fabian Streit, Josef Frank, Lea Sirignano (Dept. of Genetic Epidemiology, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University). Stefanie Heilmann-Heimbach, Stefan Herms, Per Hoffmann (Institute of Human Genetics, University of Bonn, School of Medicine & University Hospital Bonn). Andreas J. Forstner (Institute of Human Genetics, University of Bonn, School of Medicine & University Hospital Bonn).

WP6: Anastasia Benedyk, Miriam Bopp, Roman Keßler, Maximilian Lückel, Verena Schuster, Christoph Vogelbacher (Dept. of Psychiatry, Marburg University). Jens Sommer, Olaf Steinsträter (Core-Facility Brainimaging, Marburg University). Thomas W.D. Möbius (Institute of Medical Informatics and Statistics, Kiel University).

CP1: Julian Glandorf, Fabian Kormann, Arif Alkan, Fatana Wedi, Lea Henning, Alena Renker, Karina Schneider, Elisabeth Folwarczny, Dana Stenzel, Kai Wenk, Felix Picard, Alexandra Fischer, Sandra Blumenau, Beate Kleb, Doris Finholdt, Elisabeth Kinder, Tamara Wüst, Elvira Przypadlo, Corinna Brehm (Comprehensive Biomaterial Bank Marburg, Marburg University).

The FOR2107 cohort project (WP1) was approved by the Ethics Committees of the Medical Faculties, University of Marburg (AZ: 07/14) and University of Münster (AZ: 2014-422-b-S).

medRxiv preprint doi: https://doi.org/10.1101/2023.02.27.23286311; this version posted February 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Biosamples and corresponding data were sampled, processed and stored in the Marburg Biobank CBBMR.
Biomedical financial interests or potential conflicts of interest: Tilo Kircher received unrestricted educational grants from Servier, Janssen, Recordati, Aristo, Otsuka, neuraxpharm. Markus Wöhr is scientific advisor of Avisoft Bioacoustics.

Bibliography

1. Kenneth S. Kendler. Toward a Philosophical Structure for Psychiatry. *American Journal of Psychiatry*, 162(3):433–440, 2005. ISSN 0002-953X. doi: 10.1176/appi.ajp.162.3.433.
2. Thomas R. Insel and Bruce N. Cuthbert. Brain disorders? Precisely. *Science*, 348(6234): 499–500, 2015. ISSN 0036-8075. doi: 10.1126/science.aab2358.
3. Thomas Insel, Bruce Cuthbert, Marjorie Garvey, Robert Heinssen, Daniel S. Pine, Kevin Quinn, Charles Sanislow, and Philip Wang. Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. *American Journal of Psychiatry*, 167(7):748–751, 2010. ISSN 0002-953X. doi: 10.1176/appi.ajp.2010.09091379.
4. Jodie P. Gray, Veronika I. Müller, Simon B. Eickhoff, and Peter T. Fox. Multimodal Abnormalities of Brain Structure and Function in Major Depressive Disorder: A Meta-Analysis of Neuroimaging Studies. *American Journal of Psychiatry*, 177(5):422–434, 2020. ISSN 0002-953X. doi: 10.1176/appi.ajp.2019.19050560.
5. Nils R. Winter, Ramona Leenings, Jan Ernsting, Kelvin Sarink, Lukas Fisch, Daniel Emden, Julian Blanke, Janik Goltermann, Nils Opel, Carlotta Barkhau, Susanne Meinert, Katharina Dohm, Jonathan Repple, Marco Mauritz, Marius Gruber, Elisabeth J. Leehr, Dominik Grotegerd, Ronny Redlich, Andreas Jansen, Igor Nenadic, Markus M. Nöthen, Andreas Forstner, Marcella Rietschel, Joachim Groß, Jochen Bauer, Walter Heindel, Till Andlauer, Simon B. Eickhoff, Tilo Kircher, Udo Dannlowski, and Tim Hahn. Quantifying Deviations of Brain Structure and Function in Major Depressive Disorder Across Neuroimaging Modalities. *JAMA Psychiatry*, 79(9):879–888, 2022. ISSN 2168-622X. doi: 10.1001/jamapsychiatry.2022.1780.
6. Masashi Ikeda, Takeo Saito, Tetsufumi Kanazawa, and Nakao Iwata. Polygenic risk score as clinical utility in psychiatry: a clinical viewpoint. *Journal of Human Genetics*, 66(1):53–60, 2021. ISSN 1434-5161. doi: 10.1038/s10038-020-0814-y.
7. Elvisha Dhamala, B.T. Thomas Yeo, and Avram J. Holmes. Methodological Considerations for Brain-Based Predictive Modelling in Psychiatry. *Biological Psychiatry*, 2022. ISSN 0006-3223. doi: 10.1016/j.biopsych.2022.09.024.
8. Martin P. Paulus. Pragmatism Instead of Mechanism: A Call for Impactful Biological Psychiatry. *JAMA Psychiatry*, 72(7):631–632, 2015. ISSN 2168-622X. doi: 10.1001/jamapsychiatry.2015.0497.
9. Danilo Bzdok, Gael Varoquaux, and Ewout W. Steyerberg. Prediction, Not Association, Paves the Road to Precision Medicine. *JAMA Psychiatry*, 78(2):127–128, 2021. ISSN 2168-622X. doi: 10.1001/jamapsychiatry.2020.2549.
10. Tim Hahn, A A Nierenberg, and S Whitfield-Gabrieli. Predictive analytics in mental health: applications, guidelines, challenges and perspectives. *Molecular Psychiatry*, 22(1):37–43, 2017. ISSN 1359-4184. doi: 10.1038/mp.2016.201.
11. Nils R. Winter, Micah Cearns, Scott R. Clark, Ramona Leenings, Udo Dannlowski, Bernhard T. Baune, and Tim Hahn. From multivariate methods to an AI ecosystem. *Molecular Psychiatry*, pages 1–5, 2021. ISSN 1359-4184. doi: 10.1038/s41380-021-01116-y.
12. Joseph Kambeitz, Carlos Cabral, Matthew D. Sacchet, Ian H. Gotlib, Roland Zahn, Mauricio H. Serpa, Martin Walter, Peter Falkai, and Nikolaos Koutsouleris. Detecting Neuroimaging Biomarkers for Depression: A Meta-analysis of Multivariate Pattern Recognition Studies. *Biological Psychiatry*, 82(5):330–338, 2017. ISSN 0006-3223. doi: 10.1016/j.biopsych.2016.10.028.
13. Alan J. Meehan, Stephanie J. Lewis, Seena Fazel, Paolo Fusar-Poli, Ewout W. Steyerberg, Daniel Stahl, and Andrea Danese. Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. *Molecular Psychiatry*, pages 1–9, 2022. ISSN 1359-4184. doi: 10.1038/s41380-022-01528-4.
14. Claas Flint, Micah Cearns, Nils Opel, Ronny Redlich, David M. A. Mehler, Daniel Emden, Nils R. Winter, Ramona Leenings, Simon B. Eickhoff, Tilo Kircher, Axel Krug, Igor Nenadic, Volker Arolt, Scott Clark, Bernhard T. Baune, Xiaoyi Jiang, Udo Dannlowski, and Tim Hahn. Systematic misestimation of machine learning performance in neuroimaging studies of depression. *Neuropsychopharmacology*, 46(8):1510–1517, 2021. ISSN 0893-133X. doi: 10.1038/s41386-021-01020-7.
15. Mohammad R. Arbabshirani, Sergey Plis, Jing Sui, and Vince D. Calhoun. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, 145(Pt B):137–165, 2017. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2016.02.079.
16. Aleks Stolicyn, Mathew A. Harris, Xueyi Shen, Miruna C. Barbu, Mark J. Adams, Emma L. Hawkins, Laura de Nooij, Hon Wah Yeung, Alison D. Murray, Stephen M. Lawrie, J. Douglas Steele, Andrew M. McIntosh, and Heather C. Whalley. Automated classification of depression from structural brain measures across two independent community-based cohorts. *Human Brain Mapping*, 41(14):3922–3937, 2020. ISSN 1065-9471. doi: 10.1002/hbm.25095.
17. Eiko I. Fried, Jessica K. Flake, and Donald J. Robinaugh. Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology*, 1 (6):358–368, 2022. doi: 10.1038/s44159-022-00050-2.
18. L Schmaal, D J Veltman, T G M van Erp, P G Sämann, T Frodl, N Jahanshad, E Loehrer, H Tiemeier, A Hofman, W J Niessen, M W Vernooij, M A Ikram, K Wittfeld, H J Grabe, A Block, K Hegenscheid, H Völzke, D Hoehn, M Czisch, J Lagopoulos, S N Hatton, I B Hickie, R Goya-Maldonado, B Krämer, O Gruber, B Couvy-Duchesne, M E Rentería, L T Strike, N T Mills, G I de Zubicaray, K L McMahon, S E Medland, N G Martin, N A Gillespie, M J Wright, G B Hall, G M MacQueen, E M Frey, A Carballedo, L S van Velzen, M J van Tol, N J van der Wee, I M Veer, H Walter, K Schnell, E Schramm, C Normann, D Schoepf, C Konrad, B Zurowski, T Nickson, A M McIntosh, M Papmeyer, H C Whalley, J E Sussmann, B R Godlewska, P J Cowen, F H Fischer, M Rose, B W J H Penninx, P M Thompson, and D P Hibar. Subcortical brain alterations in major depressive disorder: findings from the ENIGMA Major Depressive Disorder working group. *Molecular Psychiatry*, 21(6):806–812, 2016. ISSN 1359-4184. doi: 10.1038/mp.2015.69.
19. Tilo Kircher, Markus Wöhr, Igor Nenadic, Rainer Schwarting, Gerhard Schratt, Judith Alferink, Carsten Culmsee, Holger Garn, Tim Hahn, Bertram Müller-Myhsok, Astrid Dempfle, Maik Hahmann, Andreas Jansen, Petra Pfefferle, Harald Renz, Marcella Rietschel, Stephanie H. Witt, Markus Nöthen, Axel Krug, and Udo Dannlowski. Neurobiology of the major psychoses: a translational perspective on brain structure and function—the FOR2107 consortium. *European Archives of Psychiatry and Clinical Neuroscience*, 269(8): 949–962, 2019. ISSN 0940-1334. doi: 10.1007/s00406-018-0943-x.
20. Christoph Vogelbacher, Thomas W.D. Möbius, Jens Sommer, Verena Schuster, Udo Dannlowski, Tilo Kircher, Astrid Dempfle, Andreas Jansen, and Miriam H.A. Bopp. The Marburg-Münster Affective Disorders Cohort Study (MACS): A quality assurance protocol for MR neuroimaging data. *NeuroImage*, 172:450–460, 2018. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2018.01.079.
21. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(null):2825–2830, 2011. ISSN 1532-4435.
22. Ramona Leenings, Nils Ralf Winter, Lucas Plagwitz, Vincent Holstein, Jan Ernsting, Kelvin Sarink, Lukas Fisch, Jakob Steenweg, Leon Kleine-Vennekate, Julian Gebker, Daniel Emden, Dominik Grotegerd, Nils Opel, Benjamin Risse, Xiaoyi Jiang, Udo Dannlowski, and Tim Hahn. PHOTONAI—A Python API for rapid machine learning model development. *PLOS ONE*, 16(7):e0254062, 2021. doi: 10.1371/journal.pone.0254062.
23. Scott Marek, Brenden Tervo-Clemmens, Finnegan J. Calabro, David F. Montez, Benjamin P. Kay, Alexander S. Hatoum, Meghan Rose Donohue, William Foran, Ryland L. Miller, Timothy J. Hendrickson, Stephen M. Malone, Sridhar Kandala, Eric Feczko, Oscar Miranda-Dominguez, Alice M. Graham, Eric A. Earl, Anders J. Perrone, Michaela Cordova, Olivia Doyle, Lucille A. Moore, Gregory M. Conan, Johnny Uriarte, Kathy Snider, Benjamin J. Lynch, James C. Wilgenbusch, Thomas Pengo, Angela Tam, Jianzhong Chen, Dillan J. Newbold, Annie Zheng, Nicole A. Seider, Andrew N. Van, Athanasia Metoki, Roselyne J. Chauvin, Timothy O. Laumann, Deanna J. Greene, Steven E. Petersen, Hugh Garavan, Wesley K. Thompson, Thomas E. Nichols, B. T. Thomas Yeo, Deanna M. Barch, Beatriz Luna, Damien A. Fair, and Nico U. F. Dosenbach. Reproducible brain-wide association studies require thousands of individuals. *Nature*, pages 1–7, 2022. ISSN 0028-0836. doi: 10.1038/s41586-022-04492-9.
24. Aki Nikolaidis, Andrew A. Chen, Xiaoning He, Russell Shinohara, Joshua Vogelstein, Michael Milham, and Haochang Shou. Suboptimal phenotypic reliability impedes reproducible human neuroscience. *bioRxiv*, page 2022.07.22.501193, 2022. doi: 10.1101/2022.07.22.501193.
25. Maxwell L. Elliott, Annchen R. Knodt, David Ireland, Meriwether L. Morris, Richie Poulton, Sandhya Ramrakha, Maria L. Sison, Terrie E. Moffitt, Avshalom Caspi, and Ahmad R. Hariri. What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychological Science*, 31(7):792–806, 2020. ISSN 0956-7976. doi: 10.1177/0956797620916786.
26. Darrel A. Regier, William E. Narrow, Diana E. Clarke, Helena C. Kraemer, S. Janet Kuramoto, Emily A. Kuhl, and David J. Kupfer. DSM-5 Field Trials in the United States and Canada, Part II: Test-Retest Reliability of Selected Categorical Diagnoses. *American Journal of Psychiatry*, 170(1):59–70, 2013. ISSN 0002-953X. doi: 10.1176/appi.ajp.2012.12070999.
27. Abigail S. Greene, Xilin Shen, Stephanie Noble, Corey Horien, C. Alice Hahn, Jagriti Arora, Fuyuze Tokoglu, Marisa N. Spann, Carmen I. Carrión, Daniel S. Barron, Gerard Sanacora, Vinod H. Srihari, Scott W. Woods, Dustin Scheinost, and R. Todd Constable. Brain–phenotype models fail for individuals who defy sample stereotypes. *Nature*, pages 1–10, 2022. ISSN 0028-0836. doi: 10.1038/s41586-022-05118-w.
28. Eiko I. Fried and Randolph M. Nesse. Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. *Journal of Affective Disorders*, 172:96–102, 2015. ISSN 0165-0327. doi: 10.1016/j.jad.2014.10.010.
29. Eric Feczko, Oscar Miranda-Dominguez, Mollie Marr, Alice M. Graham, Joel T. Nigg, and Damien A. Fair. The Heterogeneity Problem: Approaches to Identify Psychiatric Subtypes. *Trends in Cognitive Sciences*, 23(7):584–601, 2019. ISSN 1364-6613. doi: 10.1016/j.tics.2019.03.009.
30. H-U Wittchen, U. Wunderlich, S. Gruschwitz, and M. Zaudig. SKID I. Strukturiertes Klinisches Interview für DSM-IV. Achse I: Psychische Störungen. Interviewheft und Beurteilungsheft. Eine deutschsprachige, erweiterte Bearb. d. amerikanischen Originalversion des SKID I. 1997.
31. Erik von Elm, Douglas G Altman, Matthias Egger, Stuart J Pocock, Peter C Gøtzsche, Jan P Vandenbroucke, and STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. *PLoS Medicine*, 4(10):e296, 2007. ISSN 1549-1277. doi: 10.1371/journal.pmed.0040296.
32. Bruce Fischl. FreeSurfer. *NeuroImage*, 62(2):774–781, 2012. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2012.01.021.
33. Christian Gaser and Florian Kurth. Computational Anatomy Toolbox CAT12.
34. Siemon C. de Lange and Martijn P. van den Heuvel. Structural and functional connectivity reconstruction with CATO - A Connectivity Analysis TOolbox. *bioRxiv*, page 2021.05.31.446012, 2021. doi: 10.1101/2021.05.31.446012.
35. Susan Whitfield-Gabrieli and Alfonso Nieto-Castanon. Conn: A Functional Connectivity Toolbox for Correlated and Anticorrelated Brain Networks. *Brain Connectivity*, 2(3):125–141, 2012. ISSN 2158-0014. doi: 10.1089/brain.2012.0073.
36. Farzad V. Farahani, Waldemar Karwowski, and Nichole R. Lighthall. Application of Graph Theory for Identifying Connectivity Patterns in Human Brain Networks: A Systematic Review. *Frontiers in Neuroscience*, 13:585, 2019. ISSN 1662-4548. doi: 10.3389/fnins.2019.00585.
37. Ahmad R. Hariri, Alessandro Tessitore, Venkata S. Mattay, Francesco Fera, and Daniel R. Weinberger. The Amygdala Response to Emotional Stimuli: A Comparison of Faces and Scenes. *NeuroImage*, 17(1):317–323, 2002. ISSN 1053-8119. doi: 10.1006/nimg.2002.1179.

PREPRINT

medRxiv preprint doi: https://doi.org/10.1101/2023.02.27.23286311; this version posted February 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

38. Udo Dannlowski, Anja Stuhrmann, Victoria Beutelmann, Peter Zwanzger, Thomas Lenzen, Dominik Grotegerd, Katharina Domschke, Christa Hohoff, Patricia Ohrmann, Jochen Bauer, Christian Lindner, Christian Postert, Carsten Konrad, Volker Arolt, Walter Heindel, Thomas Suslow, and Harald Kugel. Limbic Scars: Long-Term Consequences of Childhood Maltreatment Revealed by Functional and Structural Magnetic Resonance Imaging. *Biological Psychiatry*, 71(4):286–293, 2012. ISSN 0006-3223. doi: 10.1016/j.biopsych.2011.10.021.
39. David M Howard, Mark J Adams, Toni-Kim Clarke, Jonathan D Hafferty, Jude Gibson, Masoud Shirali, Jonathan R I Coleman, Saskia P Hagenaars, Joey Ward, Eleanor M Wigmore, Clara Alloza, Xueyi Shen, Miruna C Barbu, Eileen Y Xu, Heather C Whalley, Riccardo E Marioni, David J Porteous, Gail Davies, Ian J Deary, Gibran Hemani, Klaus Berger, Henning Teismann, Rajesh Rawal, Volker Arolt, Bernhard T Baune, Udo Dannlowski, Katharina Domschke, Chao Tian, David A Hinds, 23andMe Research Team, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, Maciej Trzaskowski, Enda M Byrne, Stephan Ripke, Daniel J Smith, Patrick F Sullivan, Naomi R Wray, Gerome Breen, Cathryn M Lewis, and Andrew M McIntosh. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nature Neuroscience*, 22(3):343–352, 2019. ISSN 1097-6256. doi: 10.1038/s41593-018-0326-7.
40. Helena Pelin, Marcus Ising, Frederike Stein, Susanne Meinert, Tina Meller, Katharina Brosch, Nils R. Winter, Axel Krug, Ramona Leenings, Hannah Lemke, Igor Nenadić, Stefanie Heilmann-Heimbach, Andreas J. Forstner, Markus M. Nöthen, Nils Opel, Jonathan Repple, Julia Pfarr, Kai Ringwald, Simon Schmitt, Katharina Thiel, Lena Waltemate, Alexandra Winter, Fabian Streit, Stephanie Witt, Marcella Rietschel, Udo Dannlowski, Tilo Kircher, Tim Hahn, Bertram Müller-Myhsok, and Till F. M. Andlauer. Identification of transdiagnostic psychiatric disorder subtypes using unsupervised learning. *Neuropsychopharmacology*, 46 (11):1895–1905, 2021. ISSN 0893-133X. doi: 10.1038/s41386-021-01051-0.
41. D P Bernstein, L Fink, L Handelsman, J Foote, M Lovejoy, K Wenzel, E Sapareto, and J Ruggiero. Initial reliability and validity of a new retrospective measure of child abuse and neglect. *American Journal of Psychiatry*, 151(8):1132–1136, 1994. ISSN 0002-953X. doi: 10.1176/ajp.151.8.1132.
42. Thomas Fydrich, Gert Sommer, Stefan Tydecks, and Elmar Brähler. Fragebogen zur sozialen Unterstützung (F-SozU): Normierung der Kurzform (K-14). *Zeitschrift für Medizinische Psychologie*, 18:43, 2009.
43. Max Hamilton. A Rating Scale for Depression. *Journal of Neurology, Neurosurgery & Psychiatry*, 23(1):56, 1960. ISSN 0022-3050. doi: 10.1136/jnnp.23.1.56.
44. Aaron T. Beck, Robert A. Steer, and Margery G. Carbin. Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, 8 (1):77–100, 1988. ISSN 0272-7358. doi: 10.1016/0272-7358(88)90050-5.
45. Christine E. DeMars. Classical Test Theory and Item Response Theory. In *The Wiley Handbook of Psychometric Testing*, pages 49–73. 04 2018. ISBN 9781118489772. doi: 10.1002/9781118489772.ch2. Wiley Online Books.
46. Davide Chicco, Niklas Tötsch, and Giuseppe Jurman. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, 14(1):13, 2021. ISSN 1756-0381. doi: 10.1186/s13040-021-00244-z.
47. Tenko Raykov and George A Marcoulides. *Introduction to psychometric theory*. Routledge, New York, NY u.a, 2011. ISBN 0415878225.
48. Nicola K. Dinsdale, Emma Bluemke, Vaanathi Sundaresan, Mark Jenkinson, Stephen M. Smith, and Ana I.L. Namburete. Challenges for machine learning in clinical translation of big data imaging studies. *Neuron*, 2022. ISSN 0896-6273. doi: 10.1016/j.neuron.2022.09.012. doi: 10.1016/j.neuron.2022.09.012.
49. Marc-Andre Schulz, B. T. Thomas Yeo, Joshua T. Vogelstein, Janaina Mourao-Miranada, Jakob N. Kather, Konrad Kording, Blake Richards, and Danilo Bzdok. Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nature Communications*, 11(1):4238, 2020. doi: 10.1038/s41467-020-18037-z.
50. L Schmaal, D J Veltman, T G M van Erp, P G Sämann, T Frodl, N Jahanshad, E Loehrer, M W Vernooij, W J Niessen, M A Ikram, K Wittfeld, H J Grabe, A Block, K Hegenscheid, D Hoehn, M Czisch, J Lagopoulos, S N Hatton, I B Hickie, R Goya-Maldonado, B Krämer, O Gruber, B Couvy-Duchesne, M E Rentería, L T Strike, M J Wright, G I de Zubicaray, K L McMahon, S E Medland, N A Gillespie, G B Hall, L S van Velzen, M-J van Tol, N J van der Wee, I M Veer, H Walter, E Schramm, C Normann, D Schoepf, C Konrad, B Zurowski, A M McIntosh, H C Whalley, J E Sussmann, B R Godlewska, F H Fischer, B W J H Penninx, P M Thompson, and D P Hibar. Response to Dr Fried & Dr Kievit, and Dr Malhi et al. *Molecular Psychiatry*, 21(6):726–728, 2016. ISSN 1359-4184. doi: 10.1038/mp.2016.9.
51. Martin Gell, Simon B Eickhoff, Amir Omidvarnia, Vincent Küppers, Kaustubh R Patil, Theodore D Satterthwaite, Veronika I Müller, and Robert Langner. The Burden of Reliability: How Measurement Noise Limits Brain-Behaviour Predictions. *biorXiv*, 2023. doi: 10.1101/2023.02.09.527898.
52. Giulia Cattarinussi, Giuseppe Delvecchio, Eleonora Maggioni, Cinzia Bressi, and Paolo Brambilla. Ultra-high field imaging in Major Depressive Disorder: a review of structural and functional studies. *Journal of Affective Disorders*, 290:65–73, 2021. ISSN 0165-0327. doi: 10.1016/j.jad.2021.04.056.
53. Peter J. Uhlhaas, Peter Liddle, David E.J. Linden, Anna C. Nobre, Krish D. Singh, and Joachim Gross. Magnetoencephalography as a Tool in Psychiatric Research: Current Status and Perspective. *Biological Psychiatry*, 2(3):235–244, 2017. ISSN 2451-9022. doi: 10.1016/j.bpsc.2017.01.005.
54. K S Kendler. Levels of explanation in psychiatric and substance use disorders: implications for the development of an etiologically based nosology. *Molecular Psychiatry*, 17(1):11–21, 2012. ISSN 1359-4184. doi: 10.1038/mp.2011.70.
55. Eiko I. Fried. Lack of Theory Building and Testing Impedes Progress in The Factor and Network Literature. *Psychological Inquiry*, 31(4):271–288, 2021. ISSN 1047-840X. doi: 10.1080/1047840x.2020.1853461.
56. Frederike Stein, Elena Buckenmayer, Katharina Brosch, Tina Meller, Simon Schmitt, Kai Gustav Ringwald, Julia Katharina Pfarr, Olaf Steinsträter, Verena Enneking, Dominik Grotegerd, Walter Heindel, Susanne Meinert, Elisabeth J Leehr, Hannah Lemke, Katharina Thiel, Lena Waltemate, Alexandra Winter, Tim Hahn, Udo Dannlowski, Andreas Jansen, Igor Nenadić, Axel Krug, and Tilo Kircher. Dimensions of Formal Thought Disorder and Their Relation to Gray- and White Matter Brain Structure in Affective and Psychotic Disorders. *Schizophrenia Bulletin*, 48(4):902–911, 2022. ISSN 0586-7614. doi: 10.1093/schbul/sbac002.
57. Jonathan Repple, Marius Gruber, Marco Mauritz, Siemon C. de Lange, Nils Ralf Winter, Nils Opel, Janik Goltermann, Susanne Meinert, Dominik Grotegerd, Elisabeth J. Leehr, Verena Enneking, Tiana Borgers, Melissa Klug, Hannah Lemke, Lena Waltemate, Katharina Thiel, Alexandra Winter, Fabian Breuer, Pascal Grumbach, Hannes Hofmann, Frederike Stein, Katharina Brosch, Kai G. Ringwald, Julia Pfarr, Florian Thomas-Odenthal, Tina Meller, Andreas Jansen, Igor Nenadic, Ronny Redlich, Jochen Bauer, Tilo Kircher, Tim Hahn, Martijn van den Heuvel, and Udo Dannlowski. Shared and Specific Patterns of Structural Brain Connectivity Across Affective and Psychotic Disorders. *Biological Psychiatry*, 93(2):178–186, 01 2023. ISSN 0006-3223. doi: 10.1016/j.biopsych.2022.05.031.
58. Hannah Lemke, Lina Romankiewicz, Katharina Förster, Susanne Meinert, Lena Waltemate, Stella M. Fingas, Dominik Grotegerd, Ronny Redlich, Katharina Dohm, Elisabeth J. Leehr, Katharina Thiel, Verena Enneking, Katharina Brosch, Tina Meller, Kai Ringwald, Simon Schmitt, Frederike Stein, Olaf Steinsträter, Jochen Bauer, Walter Heindel, Andreas Jansen, Axel Krug, Igor Nenadic, Tilo Kircher, and Udo Dannlowski. Association of disease course and brain structural alterations in major depressive disorder. *Depression and Anxiety*, 39 (5):441–451, 2022. ISSN 1091-4269. doi: 10.1002/da.23260.
59. Denny Borsboom. A network theory of mental disorders. *World Psychiatry*, 16(1):5–13, 2017. ISSN 1723-8617. doi: 10.1002/wps.20375.
60. Bruce N Cuthbert and Thomas R Insel. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Medicine*, 11(1):126–126, 2013. doi: 10.1186/1741-7015-11-126.
61. Denny Borsboom, Angélique O. J. Cramer, and Annemarie Kalis. Brain disorders? Not really: Why network structures block reductionism in psychopathology research. *Behavioral and Brain Sciences*, 42:e2, 2018. ISSN 0140-525X. doi: 10.1017/s0140525x17002266.
62. Saige Rutherford, Seyed Mostafa Kia, Thomas Wolfers, Charlotte Fraza, Mariam Zabihi, Richard Dinga, Pierre Berthet, Amanda Worker, Serena Verdi, Henricus G. Ruhe, Christian F. Beckmann, and Andre F. Marquand. The normative modeling framework for computational psychiatry. *Nature Protocols*, 17(7):1711–1734, 2022. ISSN 1754-2189. doi: 10.1038/s41596-022-00696-5.
63. Frederike Stein, Tina Meller, Katharina Brosch, Simon Schmitt, Kai Ringwald, Julia Katharina Pfarr, Susanne Meinert, Katharina Thiel, Hannah Lemke, Lena Waltemate, Dominik Grotegerd, Nils Opel, Andreas Jansen, Igor Nenadić, Udo Dannlowski, Axel Krug, and Tilo Kircher. Psychopathological Syndromes Across Affective and Psychotic Disorders Correlate With Gray Matter Volumes. *Schizophrenia Bulletin*, 47(6):1740–1750, 2021. ISSN 0586-7614. doi: 10.1093/schbul/sbab037.
64. Quentin J. M. Huys, Michael Browning, Martin P. Paulus, and Michael J. Frank. Advances in the computational understanding of mental illness. *Neuropsychopharmacology*, 46(1): 3–19, 2021. ISSN 0893-133X. doi: 10.1038/s41386-020-0746-4.

PREPRINT