

A Systematic Evaluation of Machine Learning–Based Biomarkers for Major Depressive Disorder

Nils R. Winter, MSc; Julian Blanke, BSc; Ramona Leenings, PhD; Jan Ernsting, MSc; Lukas Fisch, MSc; Kelvin Sarink, MSc; Carlotta Barkhau, MSc; Daniel Emden, MSc; Katharina Thiel, MSc; Kira Flinkenflügel, MSc; Alexandra Winter, MSc; Janik Goltermann, PhD; Susanne Meinert, PhD; Katharina Dohm, PhD; Jonathan Repple, MD; Marius Gruber, MSc; Elisabeth J. Leehr, PhD; Nils Opel, MD; Dominik Grotegerd, PhD; Ronny Redlich, PhD; Robert Nitsch, MD, PhD; Jochen Bauer, PhD; Walter Heindel, MD; Joachim Gross, PhD; Benjamin Risse, PhD; Till F. M. Andlauer, PhD; Andreas J. Forstner, MD; Markus M. Nöthen, MD; Marcella Rietschel, MD; Stefan G. Hofmann, PhD; Julia-Katharina Pfarr, MSc; Lea Teutenberg, MSc; Paula Usemann, MSc; Florian Thomas-Odenthal, MSc; Adrian Wroblewski, PhD; Katharina Brosch, PhD; Frederike Stein, PhD; Andreas Jansen, PhD; Hamidreza Jamalabadi, PhD; Nina Alexander, PhD; Benjamin Straube, PhD; Igor Nenadić, MD; Tilo Kircher, MD; Udo Dannlowski, MD, PhD; Tim Hahn, PhD

IMPORTANCE 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), no informative biomarkers have been identified.

OBJECTIVE To evaluate whether machine learning (ML) can identify a multivariate biomarker for MDD.

DESIGN, SETTING, AND PARTICIPANTS This study used data from the Marburg-Münster Affective Disorders Cohort Study, a case-control clinical neuroimaging study. Patients with acute or lifetime MDD and healthy controls aged 18 to 65 years were recruited from primary care and the general population in Münster and Marburg, Germany, from September 11, 2014, to September 26, 2018. The Münster Neuroimaging Cohort (MNC) was used as an independent partial replication sample. Data were analyzed from April 2022 to June 2023.

EXPOSURE Patients with MDD and healthy controls.

MAIN OUTCOME AND MEASURE Diagnostic classification accuracy was quantified on an individual level using an extensive ML-based multivariate approach across a comprehensive range of neuroimaging modalities, including structural and functional magnetic resonance imaging and diffusion tensor imaging as well as a polygenic risk score for depression.

RESULTS Of 1801 included participants, 1162 (64.5%) were female, and the mean (SD) age was 36.1 (13.1) years. There were a total of 856 patients with MDD (47.5%) and 945 healthy controls (52.5%). The MNC replication sample included 1198 individuals (362 with MDD [30.1%] and 836 healthy controls [69.9%]). Training and testing a total of 4 million ML models, mean (SD) accuracies for diagnostic classification ranged between 48.1% (3.6%) and 62.0% (4.8%). Integrating neuroimaging modalities and stratifying individuals based on age, sex, treatment, or remission status does not enhance model performance. Findings were replicated within study sites and also observed in structural magnetic resonance imaging within MNC. Under simulated conditions of perfect reliability, performance did not significantly improve. Analyzing model errors suggests that symptom severity could be a potential focus for identifying MDD subgroups.

CONCLUSION AND RELEVANCE Despite the improved predictive capability of multivariate compared with univariate neuroimaging markers, no informative individual-level MDD biomarker—even under extensive ML optimization in a large sample of diagnosed patients—could be identified.

JAMA Psychiatry. doi:10.1001/jamapsychiatry.2023.5083
Published online January 10, 2024.

 Multimedia

 Supplemental content

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Nils R. Winter, MSc, Institute for Translational Psychiatry, University of Münster, Albert-Schweitzer-Campus 1, 48149 Münster, Germany (nils.r.winter@uni-muenster.de).

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 elusive 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.^{4,5}

Fueled by the availability of large-scale data sets and substantial improvements regarding machine learning (ML) software and hardware, precision psychiatry has gained increasing traction over the last decade. Precision psychiatry aims to build models that allow for individual predictions, thereby moving from the investigation of univariate statistical group differences toward multivariate neurobiological patterns of individual patients.^{6,7}

While a consensus on best-practice guidelines for precision psychiatry and ML has been emerging,^{6,8} 4 broad issues in MDD biomarker research remain, which may lead to substantial overestimation as well as underestimation of the true predictive performance. First, methodological shortcomings in predictive model validation (eg, data leakage between training and test set, lack of validation) lead to an overestimation of predictive performance in many publications.^{9,10} 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.¹¹ 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. Third, current studies almost exclusively focus on a single data modality, and studies integrating multiple modalities to increase predictive performance are rare.⁶ 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.¹² 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 multisite analyses has previously been used to explain small effect sizes and inconsistent results.¹³

In summary, the current body of literature does not provide enough evidence to definitively assess the effectiveness of ML approaches in identifying a multivariate biomarker for MDD. Therefore, this study aims to systematically evaluate the classification accuracy achievable by neuroimaging-based biomarkers attainable in the present state of the field. We explicitly address previous shortcomings to evaluate ML-based mul-

Key Points

Question Can multivariate machine learning approaches identify the neural signature of major depressive disorder in individual patients?

Findings In this case-control neuroimaging study that included 1801 patients with depression and healthy controls, the best machine learning algorithm only achieved a diagnostic classification accuracy of 62% across major neuroimaging modalities.

Meaning Although multivariate neuroimaging markers increase predictive power compared with univariate analyses, no depression biomarker could be uncovered that is able to identify individual patients.

tivariate biomarkers for MDD systematically; 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 are available.^{14,15} Next, we did not rely on a single predictive algorithm but capitalized on the advances in ML software¹⁶ to combine multiple classifiers from complementary algorithmic categories, including feature selection, dimensionality reduction, and extensive tuning of model hyperparameters, resulting in a total of 4 million ML models trained and evaluated in this study. Expanding previous work, we drew on a comprehensive set of neuroimaging modalities, including structural magnetic resonance imaging (MRI), task-based and resting-state functional MRI (fMRI), and diffusion tensor imaging as well as an MDD polygenic risk score (PRS) and several environmental risk factors. This allowed us to compare predictive performance across modalities in the same sample directly and enabled us to quantify the potential benefit of multimodal data integration. In addition, the clinical assessment of patients in our data was based on Structured Clinical Interview for *DSM-IV* (SCID), which provided standardized *DSM*-based MDD diagnosis and reduced the diagnostic uncertainty often hampering model performance in large-scale, multisite data today.^{12,17} Likewise, methodological heterogeneity, eg, due to differing exclusion criteria, recruitment modalities, clinical phenotyping, or MRI scanning protocols, could be alleviated in this well-curated, harmonized sample.¹⁵ All analyses were replicated within the 2 study sites. Additionally, an independent sample was used to replicate ML analyses for structural T1-weighted MRI modalities.

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.^{18–21} 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.²²

Methods

Study Design and Participants

The data used in the main analyses are part of the Marburg–Münster Affective Disorders Cohort Study (MACS).^{14,15} Data were collected at 2 sites (Marburg and Münster, Germany) using identical study protocols and harmonized scanner settings.¹⁵ A sample of 2036 healthy participants and patients with MDD were recruited as part of the MACS cohort from September 11, 2014, to September 26, 2018 (eMethods 1–3 in [Supplement 1](#)). MDD diagnosis was assessed using the SCID, axis 1 disorders. Individuals with any history of neurological or medical conditions were excluded, resulting in a final sample of 1801 (eMethods 1 in [Supplement 1](#)). 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 (eMethods 1 and 4 to 12 in [Supplement 1](#)). Similar inclusion and exclusion criteria were used in the MNC replication sample (eMethods 16 in [Supplement 1](#)). This study follows the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guideline. The FOR2107 cohort project was approved by the Ethics Committees of the Medical Faculties, University of Marburg and University of Münster. Participants received financial compensation and gave written informed consent.

Procedures and Neuroimaging Data Modalities

The neuroimaging, genetic, and behavioral data used in this study have been described previously.⁵ Detailed information is available in eMethods 4–12 in [Supplement 1](#). In short, voxel-based morphometry (VBM; CAT12 toolbox version r1450) and region-based surface, thickness, and volume (FreeSurfer version 5.3) were extracted from T1-weighted structural MRI.^{23,24} Structural connectomes were derived from diffusion tensor imaging as fractional anisotropy and mean diffusivity. Functional connectomes, voxel-based local correlation, the amplitude of low-frequency fluctuations (ALFF), and the fractional ALFF were derived from resting-state fMRI. For both structural and functional connectomes, commonly used graph network parameters, such as betweenness centrality, degree centrality, or global efficiency, were calculated.²⁵ Task-based fMRI was based on an established emotional face-matching paradigm.²⁶ In addition, we compared results with a PRS for depression as well as questionnaire data on adverse experiences during childhood (Childhood Trauma Questionnaire) and current social support (F-SozU), since these variables are established risk or protective factors in the etiology of MDD.^{27–29} 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 (HAM-D). Structural T1-weighted MRI data (FreeSurfer and VBM) were used in the MNC replication analysis.

Main Outcomes

Accuracy of predicted diagnostic labels in all ML models was calculated using balanced classification accuracy (BACC). In addition, we calculated Matthew's correlation coefficient (MCC). For all metrics, means and SDs across the 10 outer cross-validation splits were reported to assess the generalizability of the predictive models.

ML Analyses

A total of 4 million ML models to classify healthy participants and patients with MDD were trained, optimized, and evaluated (eMethods 14 and eTable 40 in [Supplement 1](#)). A single ML pipeline included imputation of missing data, feature normalization, and—optionally—feature selection or principal component analysis (PCA) to reduce dimensionality of the brain data ([Figure 1](#)). Subsequently, a classification algorithm was trained to predict diagnosis, including support vector machines, random forests, logistic regression, k-nearest neighbor, Gaussian-naïve 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 patients with acute depression (omitting remitted depression) or recurrent depression (omitting single-episode depression), patients with or without comorbidities, patients currently receiving and not currently receiving medication, male and female patients, as well as a homogeneous age group (age range, 24 to 28 years) and replicated within the 2 study sites (eMethods 3 in [Supplement 1](#)). Brain modality integration was accomplished either using a combination of PCA components from every data modality or a voting ensemble strategy combining all diagnosis predictions from the unimodal models. All ML analyses were performed using PHOTONAI.^{16,30} Scripts are available online.³¹

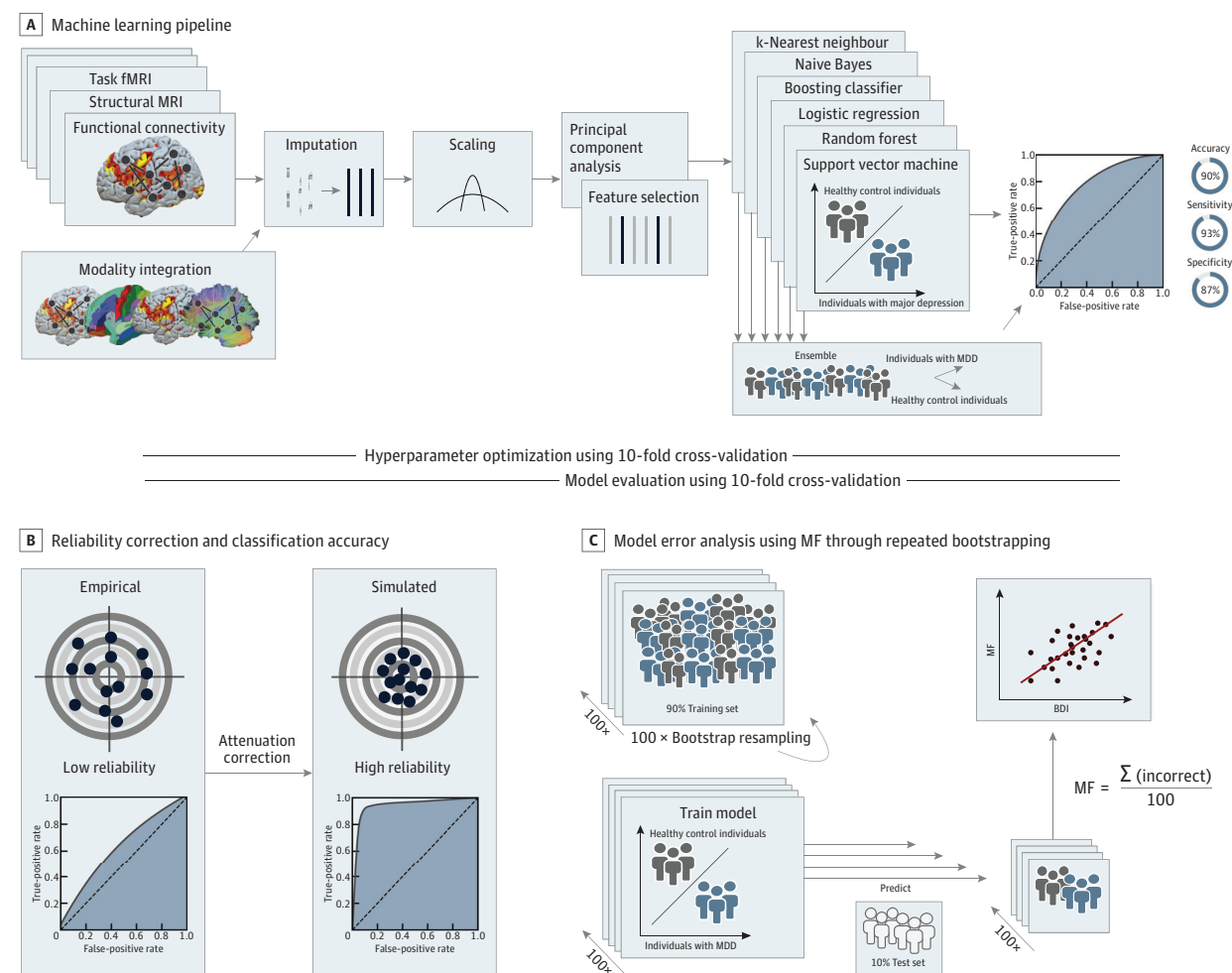
Simulation of Perfect Reliability

To quantify the effect of reliability on classification performance, we performed exploratory analyses using attenuation correction from classical test theory to estimate the true classification accuracy occurring if the reliability of the data was perfect.³² We first computed MCC from the model predictions ($\hat{y}$) and the actual diagnostic labels (y) using the equation $MCC = \text{Covariance}(y, \hat{y}) / (\sigma_y \times \sigma_{\hat{y}})$.³³ This correlation was then corrected for an assumed reliability (ρ) using the attenuation formula $MCC \text{ Correlation} = MCC / \sqrt{\rho}$.³⁴

We conducted 2 separate attenuation correction analyses. First, we assumed the reliability of an MDD diagnosis ($\rho_y = 0.28$) based on the current literature on the interrater reliability of *DSM-5* diagnoses.^{20,21} 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 from Chicco et al³³ (eMethods 13 in [Supplement 1](#)) using the following equation:

$$BACC = \frac{1}{2 \times \sqrt{\frac{\phi - \phi^2}{\beta - \beta^2}}} \times MCC + 0.5$$

Figure 1. Overview of All Analyses



Overview of all analyses. A, Steps of the machine learning pipeline. B, Reliability correction and its effect on classification accuracy. C, Model error analysis using misclassification frequency (MF) through repeated bootstrapping. BDI indicates

Beck Depression Inventory; fMRI, functional magnetic resonance imaging; MDD, major depressive disorder; MRI, magnetic resonance imaging.

Analysis of Systematic Model Error

In each individual, we quantified the tendency for misclassification based on 100 bootstrap resampling runs on the training set of the best-performing neuroimaging modality (eMethods 15 in Supplement 1). In short, 1 ML pipeline for every bootstrap training set was trained and diagnostic labels for the participants in the test set were collected, resulting in 100 predictions for every participant. The sum of incorrect classifications then leads to the frequency of misclassification (MF).²² Finally, MF was correlated with external measures describing depressive symptom severity and demographic or environmental characteristics using Spearman rank correlation.

Statistical Analysis

P values for group comparisons were based on *t* or χ^2 tests, as appropriate. Significance was set at *P* < .05, and all *P* values were 2-tailed. Statistical tests were conducted using the Python-based statistical software statsmodels version 0.14.

Results

Of 1801 included participants, 1162 (64.5%) were female, and the mean (SD) age was 36.1 (13.1) years (Table). There were a total of 856 patients with MDD (47.5%) and 945 healthy controls (52.5%). The MNC sample included 1198 individuals (362 patients with MDD [30.1%] and 836 healthy controls [69.9%]; mean [SD] age, 35.3 [12.6] years; 682 female participants [56.9%]; eTable 38 in Supplement 1).

Multivariate Classification Accuracy

Across neuroimaging modalities and ML algorithms, mean (SD) BACC ranged between 48.1% (3.6%) and 62.0% (4.8%) (eTables 1, 2, and 40 and eFigure 8 in Supplement 1). Results for the single best ML algorithm in each modality are shown in Figure 2. The highest BACC was found for resting-state connectivity, with a mean (SD) BACC ranging between 51.5% (7.1%)

Table. Sociodemographic and Clinical Characteristics of All Participants

	No. (%)		P value ^a
Characteristic	Healthy controls	Patients with MDD	
Sex			
Female	607 (64.2)	555 (64.8)	.83
Male	338 (35.8)	301 (35.2)	
Age, mean (SD), y	34.40 (13.01)	36.76 (13.27)	<.001
HAMD, mean (SD)	1.45 (2.18)	9.38 (7.17)	<.001
BDI, mean (SD)	4.11 (4.27)	17.58 (11.02)	<.001
CTQ, mean (SD)	32.59 (8.57)	45.06 (15.92)	<.001
Social support, mean (SD)	4.51 (0.54)	3.77 (0.87)	<.001
Medication load index, mean (SD)	NA	1.35 (1.48)	NA
Previous inpatient treatments, mean (SD)	NA	1.58 (2.08)	NA
Previous depressive episodes, mean (SD)	NA	3.99 (6.75)	NA
Total duration of previous inpatient treatments, mean (SD), wk	NA	11.95 (18.89)	NA
Total duration of all previous depressive episodes, mean (SD), mo	NA	45.36 (69.18)	NA
Comorbid diagnoses ^b			
Any comorbid diagnosis	NA	373 (43.6)	NA
Anxiety disorder	NA	269 (31.4)	NA
Eating disorder	NA	50 (5.8)	NA
Dysthymic disorder	NA	43 (5.0)	NA
Substance use disorder	NA	37 (0.8)	NA
Somatic symptom disorder	NA	27 (3.2)	NA
Psychotic disorder	NA	7 (0.8)	NA

Abbreviations: BDI, Beck Depression Inventory; CTQ, Childhood Trauma Questionnaire; HAMD, Hamilton Rating Scale for Depression; MDD, major depressive disorder; MRI, magnetic resonance imaging; NA, not applicable; VBM, voxel-based morphometry.

^a Calculated using t or χ^2 tests, as appropriate.

^b Lifetime comorbidities were derived from the Structured Clinical Interview for *DSM-IV*. Multiple comorbidities were possible for any patient with MDD.

and 61.5% (3.4%). Structural MRI as well as task-based fMRI showed lower BACC compared with all resting-state fMRI modalities.

ML pipelines on subgroups of only patients with acute depression (n = 599), only patients with remitted depression (n = 297), only patients with comorbidities (n = 373) or without comorbidities (n = 482), or either patients that were currently receiving medication (n = 535) or not currently receiving medication (n = 321) showed similar results compared with the analysis containing all patients with MDD (max BACC = 64.8%). Likewise, restricting analyses to male or female individuals or a more homogeneous age range of 24 to 28 years did not change the overall results (max BACC = 61.6%; eFigures 1 to 5 and eTables 5 to 31 in [Supplement 1](#)). Additional sample size simulation analyses did not reveal any discernible association between the training sample size and model performance (eMethods 17 and eFigure 9 in [Supplement 1](#)). Nonlinear dimensionality reduction methods did not outperform linear methods (eMethods 18 and eFigure 10 in [Supplement 1](#)).

Multimodal Data Integration

Integration of neuroimaging modalities using principal components from modality-specific PCAs achieved mean (SD) BACCs between 50.1% (4.0%) and 57.2% (4.4%) (Figure 2; eTable 3 in [Supplement 1](#)). Combining predicted labels from the unimodal models (across algorithms, across modalities, or both) into a majority-vote ensemble classifier achieved a mean (SD) 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 mean (SD) BACC of 62.0% (4.8%).

Replicability Analyses

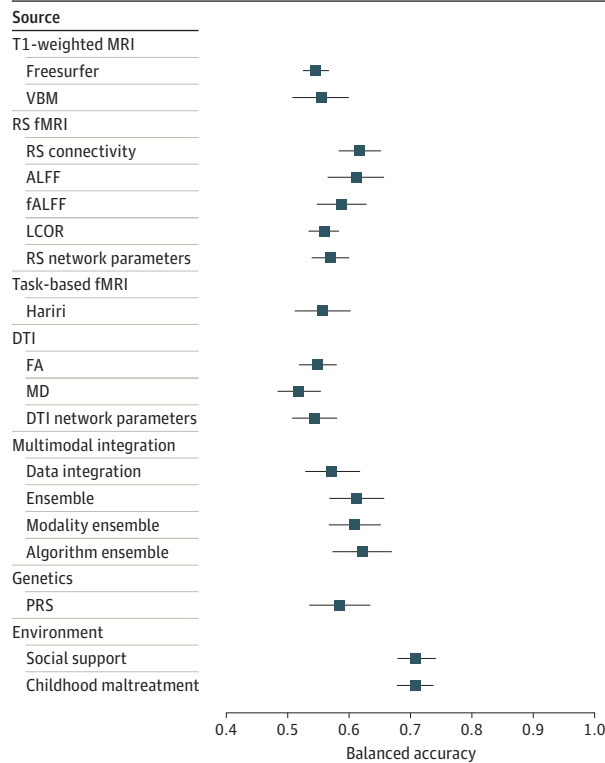
The main findings were replicated in the Marburg and Münster samples independently. The highest mean (SD) BACC was 59.2% (5.1%) in the Marburg sample and 60.0% (9.0%) in the Münster sample (eTables 32 to 37 and eFigures 6 and 7 in [Supplement 1](#)). In the independent MNC sample, the highest mean (SD) BACC for regional cortical and subcortical surface, thickness, and volume was 54.0% (5.1%), while the mean (SD) BACC for VBM analysis reached 53.4% (4.4%) (eTable 39 in in [Supplement 1](#)).

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 depression PRS used by Howard et al²⁷ achieved similar results to neuroimaging (mean [SD] BACC, 58.4% [5.0%]), both self-reported childhood maltreatment and social support outperformed brain-based and PRS-based models, achieving mean (SD) BACCs of 70.5% (2.9%) and 70.6% (3.0%), respectively.

Effects of Reliability of Diagnosis and Neuroimaging Data

The MCC correlation coefficient between actual and predicted diagnosis was corrected using the attenuation correction formula, estimating classification performance given perfect reliability. We first corrected for the lower bound of the MDD diagnosis reliability ($\rho_y = 0.28$) as reported in the litera-

Figure 2. Balanced Accuracy for the Best Machine Learning Pipelines

Balanced accuracy for the best machine learning pipeline in every modality. Error bars display 1 SD calculated across the 10 outer cross-validation folds. ALFF, amplitude of low-frequency fluctuations; DTI, diffusion tensor imaging; FA, fractional anisotropy; fALFF, fractional amplitude of low-frequency fluctuations; fMRI, functional magnetic resonance imaging; LCOR, local correlation; MD, mean diffusivity; MRI, magnetic resonance imaging; PRS, polygenic risk score; RS, resting state; VBM, voxel-based morphometry.

ture (Figure 3A). The mean (SD) BACC for the best ML algorithm on resting-state connectivity increased to 71.8% (6.4%) after correction. The mean (SD) BACC for the voting ensemble increased to 73.4% (7.4%). Next, we assumed reliability coefficients of neuroimaging modalities between 0.1 and 1 (Figure 3B). For the best unimodal analysis (resting-state connectivity), BACC increased 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

The frequency with which each individual was incorrectly classified as either healthy or having depression was measured using the MF based on the modality that achieved the highest performance in the unimodal analyses (resting-state fMRI connectivity). MF was significantly correlated with symptom severity in patients with depression (eTable 4 in Supplement 1). A higher score in current depressive symptom levels (BDI and HAMD scores) as well as a higher number of previous hospitalizations were associated with fewer misclassifications (BDI: 621 patients; $r = -0.15$; $P < .001$; HAMD: 628 patients; $r = -0.20$; $P < .001$; number of hospitalizations: 622 patients; $r = -0.10$; $P = .01$), showing that patients with more severe current de-

pressive symptoms and a more unfavorable previous disease course were correctly classified as patients with MDD more often. Similar effects were found for lower global assessment of functioning, higher medication load, and the presence of comorbidities (eTable 4 in Supplement 1).

Discussion

Extending recent evidence showing that univariate group-level differences between patients with MDD and healthy controls are small,⁵ we systematically evaluated ML approaches classifying patients with MDD and healthy controls based on multivariate neuroimaging signatures. In summary, training and testing a total of 4 million ML models on a large, harmonized sample, the accuracy for predicting MDD diagnosis did not exceed 62%. Although slightly improving the 56% to 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.⁹

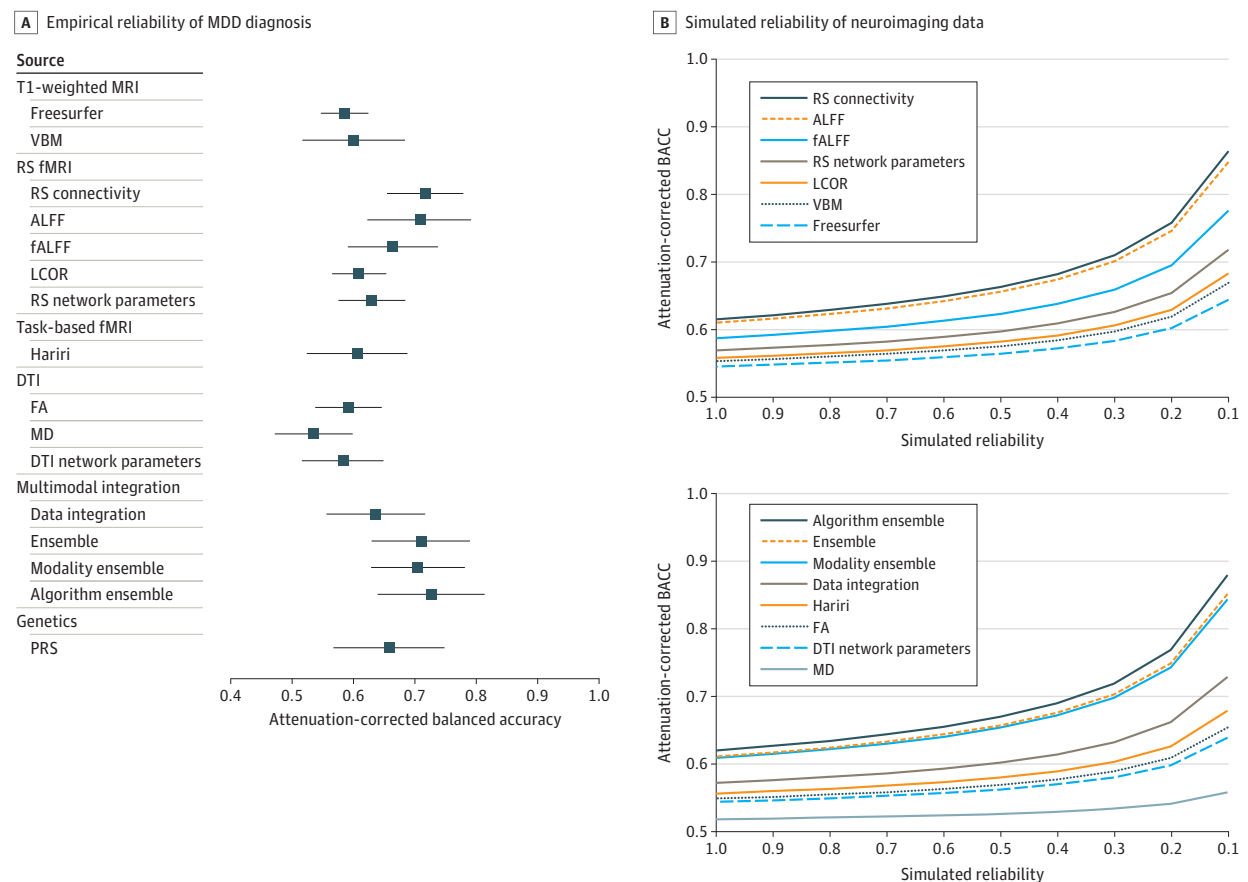
Considering that biological psychiatry is built on the premise that mental disorders have a neural basis, it is essential for the field to explain the lack of neurobiological manifestations of MDD informative on the level of the individual across the most commonly investigated modalities today. We will discuss several viewpoints concerning the reliability and validity of both the neuroimaging data and the MDD conceptualization.

Addressing the debate around reliability,^{22,35} we show that even under conditions of perfect reliability of diagnosis or neuroimaging data, reliable predictions on the level of the individual patient still remain elusive. Note that this approach can only simulate perfect reliability with regard to final model predictions and thus does not speak directly to the effect different data or preprocessing pipelines might have on model training.³⁵ 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, several research directions 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.^{36,37}

On the other hand, if we assume that the information relevant for explaining behavior and mental processes is present in current neuroimaging modalities, issues of biological validity of the MDD construct appear most plausible. Rather than solely focusing on the diagnosis of MDD, looking at lon-

Figure 3. Balanced Accuracy After Attenuation Correction



A, Balanced accuracy for the best machine learning pipeline in every modality after performing an attenuation correction for the empirical reliability of the major depressive disorder (MDD) diagnosis. Error bars display 1 SD calculated across the 10 outer cross-validation folds. B, Balanced accuracy for the 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 balanced classification accuracy (BACC). ALFF, amplitude of low-frequency fluctuations; DTI, diffusion tensor imaging; FA, fractional anisotropy; fALFF, fractional amplitude of low-frequency fluctuations; fMRI, functional magnetic resonance imaging; LCOR, local correlation; MD, mean diffusivity; MRI, magnetic resonance imaging; PRS, polygenic risk score; RS, resting state; VBM, voxel-based morphometry.

gitudinal data and clinically relevant outcomes across diagnoses may result in more accurate predictions, such as linking neuroimaging markers with long-term disease trajectories.^{38,39} Indeed, our results regarding correlations of MF provide support for associations between symptom severity and neurobiological markers, suggesting that patients with higher levels of current symptoms and more unfavorable disease courses in the past are easier to detect and correctly classify.

In the same vein, case-control designs might be too simplistic to adequately model the complex relationship between brain and behavior.¹ Normative modeling captures deviations of the individual patient, thereby overcoming the necessity for a common biological cause across all patients with MDD.⁴⁰ Similarly, identifying MDD biotypes through clustering across *DSM* diagnoses might constitute a promising way forward.^{41,42} However, more research is needed to investigate whether these approaches are actually able to increase clinically informative predictions for individual patients.

Strengths and Limitations

Our study provides 4 main improvements over existing ML studies. First, there is a reduced risk of inflated predictive performance estimates by using nested cross-validation with sufficiently large test sets. Second, we systematically optimized many possible ML pipelines and algorithms. Third, we integrated data from 11 neuroimaging modalities and preprocessing methods. Fourth, we used a large sample collected in a single study without the need for data pooling across multiple studies and acquisition processes, effectively 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 on SCID interviews. Thus, we provide evidence that low predictive performance cannot be explained by a lack of harmonization of studies or unstandardized diagnoses, as previously suggested.¹⁷ The replication conducted within the 2 study sites along with the independent replication results for

T1-weighted structural MRI in the MNC sample adds significant robustness to our findings and further supports our conclusions.

This study has limitations. Our additional analyses on sample size did not demonstrate any discernible association between sample size and model performance. This suggests that increasing the sample size may not necessarily lead to a significant improvement in performance. Alternatively, it is possible that the overall sample size is still insufficient to reliably detect a consistent increase in model performance. In the latter scenario, this might imply that larger sample sizes are necessary to establish conclusive evidence regarding the potential of neuroimaging data as a biomarker for MDD. In the same vein, our subgroup analyses, particularly those focusing on patients with or without comorbidities or treatment, resulted in a substantial reduction in sample size, which could potentially impede the robust identification of an ML-based

biomarker. In future studies, recruiting larger cohorts, especially for these more homogeneous samples, could be advantageous. In addition, it's important to note that our study primarily focused on classical ML algorithms rather than deep learning methods, which represent another avenue for potential exploration in future research.

Conclusions

In this study, no informative individual-level MDD biomarker—even under extensive ML optimization in a large sample of diagnosed patients—could be identified. It is imperative for researchers, journals, and funding agencies to reflect on the next steps in advancing biological psychiatry. These steps should prioritize delivering more accurate individualized predictions to enhance the treatment and care of patients with MDD.

ARTICLE INFORMATION

Accepted for Publication: November 5, 2023.

Published Online: January 10, 2024.

doi:10.1001/jamapsychiatry.2023.5083

Author Affiliations: Institute for Translational Psychiatry, University of Münster, Münster, Germany (N. R. Winter, Blanke, Leenings, Ernsting, Fisch, Sarink, Barkhau, Emden, Thiel, Flinkenflügel, A. Winter, Goltermann, Meinert, Dohm, Repple, Gruber, Leehr, Opel, Grotegerd, Redlich, Dannlowski, Hahn); Otto Creutzfeldt Center for Cognitive and Behavioral Neuroscience, University of Münster, Münster, Germany (N. R. Winter, Nitsch, Gross, Risse, Dannlowski, Hahn); Faculty of Mathematics and Computer Science, University of Münster, Münster, Germany (Leenings, Ernsting, Risse); Institute for Geoinformatics, University of Münster, Münster, Germany (Ernsting, Risse); Institute for Translational Neuroscience, University of Münster, Münster, Germany (Meinert, Nitsch); Department of Psychiatry, Psychosomatic Medicine and Psychotherapy, University Hospital Frankfurt, Goethe University, Frankfurt am Main, Germany (Repple, Gruber); Department of Psychiatry and Psychotherapy, University Hospital Jena, Jena, Germany (Opel); Center for Intervention and Research on Adaptive and Maladaptive Brain Circuits Underlying Mental Health, Jena, Germany (Opel, Redlich); German Center for Mental Health (DZPG), Jena, Germany (Opel); Department of Psychology, University of Halle, Halle, Germany (Redlich); German Center for Mental Health (DZPG), Halle, Germany (Redlich); Clinic for Radiology, University of Münster, University Hospital Münster, Münster, Germany (Bauer, Heindel); Institute for Biomagnetism and Biosignalanalysis, University of Münster, Münster, Germany (Gross); Department of Neurology, Klinikum rechts der Isar, School of Medicine, Technical University of Munich, Munich, Germany (Andlauer); Institute of Human Genetics, University of Bonn, School of Medicine and University Hospital Bonn, Bonn, Germany (Forstner, Nöthen); Institute of Neuroscience and Medicine (INM-1), Research Centre Jülich, Jülich, Germany (Forstner); Department of Genetic Epidemiology, Central Institute of Mental Health, Faculty of Medicine Mannheim, University of Heidelberg, Mannheim, Germany (Rietschel); Department of Clinical

Psychology, Philipps-University Marburg, Marburg, Germany (Hofmann); Department of Psychiatry and Psychotherapy, Philipps-University Marburg, Marburg, Germany (Pfarr, Teutenberg, Usemann, Thomas-Odenthal, Wroblewski, Brosch, Stein, Jansen, Jamalabadi, Alexander, Straube, Nenadić, Kircher); Center for Mind, Brain and Behavior (CMBB), Marburg, Germany (Pfarr, Teutenberg, Usemann, Thomas-Odenthal, Wroblewski, Brosch, Stein, Jansen, Alexander, Straube, Nenadić, Kircher); Core Facility Brain Imaging, Faculty of Medicine, Philipps-University Marburg, Marburg, Germany (Jansen).

Author Contributions: Mr Winter and Dr Hahn had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Drs Dannlowski and Hahn contributed equally as co-senior authors.

Concept and design: N. Winter, Blanke, Jansen, Jamalabadi, Nenadić, Dannlowski, Hahn.

Acquisition, analysis, or interpretation of data: N. Winter, Blanke, Leenings, Ernsting, Fisch, Sarink, Barkhau, Emden, Thiel, Flinkenflügel, A. Winter, Goltermann, Meinert, Dohm, Repple, Gruber, Leehr, Opel, Grotegerd, Redlich, Nitsch, Bauer, Heindel, Gross, Risse, Andlauer, Forstner, Nöthen, Rietschel, Hofmann, Pfarr, Teutenberg, Usemann, Thomas-Odenthal, Wroblewski, Brosch, Stein, Jamalabadi, Alexander, Straube, Nenadić, Kircher, Dannlowski, Hahn.

Drafting of the manuscript: N. Winter, Dannlowski, Hahn.

Critical review of the manuscript for important intellectual content: All authors.

Statistical analysis: N. Winter, Blanke, Leenings, Ernsting, Barkhau, Dohm, Repple, Jamalabadi, Dannlowski, Hahn.

Obtained funding: Redlich, Risse, Nöthen, Rietschel, Jansen, Jamalabadi, Straube, Nenadić, Kircher, Dannlowski, Hahn.

Administrative, technical, or material support: Blanke, Leenings, Ernsting, Fisch, Sarink, A. Winter, Goltermann, Repple, Opel, Grotegerd, Bauer, Heindel, Andlauer, Forstner, Nöthen, Rietschel, Pfarr, Teutenberg, Thomas-Odenthal, Brosch, Stein, Jansen, Jamalabadi, Nenadić, Kircher, Dannlowski.

Supervision: Leehr, Heindel, Risse, Rietschel, Hofmann, Kircher, Dannlowski, Hahn.

Conflict of Interest Disclosures: Dr Andlauer reported employment by Boehringer Ingelheim outside the submitted work. Dr Nöthen reported grants from German Research Foundation during the conduct of the study; personal fees from HMG Systems Engineering GmbH, Deutsches Ärzteblatt, and EVERIS Belgique outside the submitted work; employment by Life & Brain GmbH; and holding shares in Life & Brain GmbH. Dr Hofmann reported professorship from Alexander von Humboldt Foundation and personal fees from the National Institute of Mental Health during the conduct of the study. Dr Nenadić reported grants from the German Research Foundation during the conduct of the study. Dr Kircher received unrestricted educational grants from Servier, Janssen, Recordati, Aristo, Otsuka, and Neuraxpharm. Dr Dannlowski reported grants from the German Research Foundation and Interdisciplinary Centre for Clinical Research Münster during the conduct of the study. No other disclosures were reported.

Funding/Support: This work was funded by the German Research Foundation (grants FOR2107 KI588/14-1, KI588/14-2, KI588/20-1, and KI588/22-1 to Dr Kircher; grant STR1146/18-1 to Dr Straube; grants HA7070/2-2, HA7070/3, and HA7070/4 to Dr Hahn; grants FOR2107 DA1151/5-1, DA1151/5-2, DA1151/9-1, DA1151/10-1, DA1151/11-1, and SFB-TRR58 projects C09 and Z02 to Dr Dannlowski; grants NE2254/1-2, NE2254/2-1, NE2254/3-1, and NE2254/4-1 to Dr Nenadić; grants NO246/10-1 and NO246/10-2 to Dr Nöthen), the Interdisciplinary Centre for Clinical Research of the medical faculty of Münster (grant MzH 3/020/20 to Dr Hahn; grants Dan3/012/17 and Dan3/022/22 to Dr Dannlowski), the German Science Foundation (grant CRC 1451/A7 to Dr Nitsch), and the European Commission (REFORM/SC2020/029 to Dr Nöthen).

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Sharing Statement: See [Supplement 2](#).

Additional Contributions: Acknowledgments and members by Work Package (WP): WPI: 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, and Dilara Yüksel (Department of Psychiatry, Marburg University, Marburg, Germany); Mechthild Wallnig and Rita Werner (Core-Facility Brain Imaging, Marburg University); Carmen Schade-Brittinger and Maik Hahmann (Coordinating Centre for Clinical Trials, Marburg, Germany); Michael Putzke (Psychiatric Hospital, Friedberg, Germany); Rolf Speier and Lutz Lenhard (Psychiatric Hospital Vitos, Marburg, Germany); Birgit Köhnlein (Psychiatric Practice, Marburg, Germany); Peter Wulf, Jürgen Kleebach, and Achim Becker (Psychiatric Hospital Hephata, Schwalmstadt-Treysa, Germany); Ruth Bär (Care Facility Bischoff, Neukirchen, Germany); Matthias Müller, Michael Franz, Siegfried Scharmann, Anja Haag, Kristina Spenner, and Ulrich Ohlenschläger (Psychiatric Hospital Vitos, Marburg, Germany); Matthias Müller, Michael Franz, and Bernd Kundermann (Psychiatric Hospital Vitos, Gießen, Germany); 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, and Dario Zaremba (Department of Psychiatry, University of Münster, Münster, Germany); Harald Kugel, Jochen Bauer, Walter Heindel, and Birgit Vahrenkamp (Department of Clinical Radiology, University of Münster); Gereon Heuft and Gudrun Schneider (Department of Psychosomatics and Psychotherapy, University of Münster); Thomas Reker (LWL-Hospital Münster, Münster, Germany); Gisela Bartling (Institut für Psychologische Psychotherapieausbildung Münster, Münster, Germany); and Ulrike Buhlmann (Department 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 Raitel, Tobias Redecker, Vanessa Sandermann, Finnja Schramm, Linda Tempel, Natalie Vermehren, Jakob Vörckel, Stephan Weingarten, Maria Willadsen, and Cüneyt Yildiz (Faculty of Psychology, Marburg University); WP4: Jana Freff (Department of Psychiatry, University of Münster); Susanne Michels, Goutham Ganjam, and Katharina Elsässer (Faculty of Pharmacy, Marburg University); Felix Ruben Picard, Nicole Löwer, and Thomas Ruppertsberg (Institute of Laboratory Medicine and Pathobiochemistry, Marburg University); WP5: Helene Dukal, Christine Hohmeyer, Lennard Stütz, Viola Lahr, Fabian Streit, Josef Frank, and Lea Sirignano (Department of Genetic Epidemiology, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany); Stefanie Heilmann-Heimbach, Stefan Herms, and Per Hoffmann (Institute of Human Genetics, University of Bonn, School of Medicine & University Hospital Bonn, Bonn, Germany); 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, and Christoph Vogelbacher (Department of Psychiatry, Marburg University); Jens Sommer and Olaf

Steinsträter (Core-Facility Brainimaging, Marburg University); Thomas W. D. Möbius (Institute of Medical Informatics and Statistics, Kiel University, Kiel, Germany); *Central Project 1*: Julian Glandorf, Fabian Kormann, Arif Alkan, Fatana Wedi, Lea Henning, Alena Renker, Karina Schneider, Elisabeth Folwaczny, Dana Stenzel, Kai Wenk, Felix Picard, Alexandra Fischer, Sandra Blumenau, Beate Kleb, Doris Finholdt, Elisabeth Kinder, Tamara Wüst, Elvira Przypadlo, and Corinna Brehm (Comprehensive Biomaterial Bank Marburg, Marburg University).

Additional Information: Principal investigators with areas of responsibility in the FOR2107 consortium include: *Work Package 1 (WP1)*—Marburg-Münster Affective Disorders Cohort Study cohort and brain imaging: Tilo Kircher (speaker FOR2107), Udo Dannlowski (cospeaker FOR2107), Axel Krug, Igor Nenadic, and Carsten Konrad; *WP2*—animal phenotyping: Markus Wöhr and Rainer Schwarting; *WP3*—micro RNA: Gerhard Schratz; *WP4*—immunology and mitochondria: Judith Alferink, Carsten Culmsee, and Holger Garn; *WP5*—genetics: Marcella Rietschel, Markus Nöthen, and Stephanie Witt; *WP6*—multimethod data analytics: Andreas Jansen, Tim Hahn, Bertram Müller-Myhsok, and Astrid Dempfle; *Central Project 1*—biobank: Petra Pfefferle and Harald Renz; *CP2*—administration: Tilo Kircher, Udo Dannlowski, and Carsten Konrad. Biosamples and corresponding data were sampled, processed, and stored in the Comprehensive Biomaterial Bank Marburg.

REFERENCES

- Kendler KS. Toward a philosophical structure for psychiatry. *Am J Psychiatry*. 2005;162(3):433-440. doi:10.1176/appi.ajp.162.3.433
- Insel TR, Cuthbert BN. Medicine. Brain disorders? Precisely. *Science*. 2015;348(6234):499-500. doi:10.1126/science.aab2358
- Insel T, Cuthbert B, Garvey M, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *Am J Psychiatry*. 2010;167(7):748-751. doi:10.1176/appi.ajp.2010.09091379
- Gray JP, Müller VI, Eickhoff SB, Fox PT. Multimodal abnormalities of brain structure and function in major depressive disorder: a meta-analysis of neuroimaging studies. *Am J Psychiatry*. 2020;177(5):422-434. doi:10.1176/appi.ajp.2019.19050560
- Winter NR, Leenings R, Ernsting J, et al. Quantifying deviations of brain structure and function in major depressive disorder across neuroimaging modalities. *JAMA Psychiatry*. 2022;79(9):879-888. doi:10.1001/jamapsychiatry.2022.1780
- Dhamala E, Yeo BTT, Holmes AJ. One size does not fit all: methodological considerations for brain-based predictive modeling in psychiatry. *Biol Psychiatry*. 2023;93(8):717-728. doi:10.1016/j.biopsych.2022.09.024
- Bzdok D, Varoquaux G, Steyerberg EW. Prediction, not association, paves the road to precision medicine. *JAMA Psychiatry*. 2021;78(2):127-128. doi:10.1001/jamapsychiatry.2020.2549
- Hahn T, Nierenberg AA, Whitfield-Gabrieli S. Predictive analytics in mental health: applications, guidelines, challenges and perspectives. *Mol Psychiatry*. 2017;22(1):37-43. doi:10.1038/mp.2016.201
- Kambeitz J, Cabral C, Sacchet MD, et al. Detecting neuroimaging biomarkers for depression: a meta-analysis of multivariate pattern recognition studies. *Biol Psychiatry*. 2017;82(5):330-338. doi:10.1016/j.biopsych.2016.10.028
- Meehan AJ, Lewis SJ, Fazel S, et al. Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. *Mol Psychiatry*. 2022;27(6):2700-2708. doi:10.1038/s41380-022-01528-4
- Flint C, Cearns M, Opel N, et al. Systematic misestimation of machine learning performance in neuroimaging studies of depression. *Neuropsychopharmacology*. 2021;46(8):1510-1517. doi:10.1038/s41386-021-01020-7
- Stolcyn A, Harris MA, Shen X, et al. Automated classification of depression from structural brain measures across two independent community-based cohorts. *Hum Brain Mapp*. 2020;41(14):3922-3937. doi:10.1002/hbm.25095
- Schmaal L, Veltman DJ, van Erp TGM, et al. Subcortical brain alterations in major depressive disorder: findings from the ENIGMA Major Depressive Disorder working group. *Mol Psychiatry*. 2016;21(6):806-812. doi:10.1038/mp.2015.69
- Kircher T, Wöhr M, Nenadic I, et al. Neurobiology of the major psychoses: a translational perspective on brain structure and function-the FOR2107 consortium. *Eur Arch Psychiatry Clin Neurosci*. 2019;269(8):949-962. doi:10.1007/s00406-018-0943-x
- Vogelbacher C, Möbius TWD, Sommer J, et al. The Marburg-Münster Affective Disorders Cohort Study (MACS): A quality assurance protocol for MR neuroimaging data. *Neuroimage*. 2018;172:450-460. doi:10.1016/j.neuroimage.2018.01.079
- Leenings R, Winter NR, Plagwitz L, et al. PHOTONAI—a Python API for rapid machine learning model development. *PLoS One*. 2021;16(7):e0254062. doi:10.1371/journal.pone.0254062
- Schmaal L, Veltman DJ, van Erp TGM, et al. Response to Dr Fried & Dr Kievit, and Dr Malhi et al. *Mol Psychiatry*. 2016;21(6):726-728. doi:10.1038/mp.2016.9
- Marek S, Tervo-Clemmens B, Calabro FJ, et al. Reproducible brain-wide association studies require thousands of individuals. *Nature*. 2022;603(7902):654-660. doi:10.1038/s41586-022-04492-9
- Nikolaidis A, Chen AA, He X, et al. Suboptimal phenotypic reliability impedes reproducible human neuroscience. *bioRxiv*. Preprint posted online July 23, 2022. doi:10.1101/2022.07.22.501193
- Regier DA, Narrow WE, Clarke DE, et al. DSM-5 field trials in the United States and Canada, part II: test-retest reliability of selected categorical diagnoses. *Am J Psychiatry*. 2013;170(1):59-70. doi:10.1176/appi.ajp.2012.12070999
- Fried EI, Flake JK, Robinaugh DJ. Revisiting the theoretical and methodological foundations of depression measurement. *Nat Rev Psychol*. 2022;1(6):358-368. doi:10.1038/s44159-022-00050-2
- Greene AS, Shen X, Noble S, et al. Brain-phenotype models fail for individuals who defy sample stereotypes. *Nature*. 2022;609(7925):109-118. doi:10.1038/s41586-022-05118-w

23. Gaser C, Kurth F. CAT: a computational anatomy toolbox for SPM. Accessed December 3, 2023. <http://www.neuro.uni-jena.de/cat/>
24. Fischl B. FreeSurfer. *Neuroimage*. 2012;62(2):774-781. doi:10.1016/j.neuroimage.2012.01.021
25. Farahani FV, Karwowski W, Lighthall NR. Application of graph theory for identifying connectivity patterns in human brain networks: a systematic review. *Front Neurosci*. 2019;13:585. doi:10.3389/fnins.2019.00585
26. Hariri AR, Tessitore A, Mattay VS, Fera F, Weinberger DR. The amygdala response to emotional stimuli: a comparison of faces and scenes. *Neuroimage*. 2002;17(1):317-323. doi:10.1006/nimg.2002.1179
27. Howard DM, Adams MJ, Clarke TK, et al. 23andMe Research Team; Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat Neurosci*. 2019;22(3):343-352. doi:10.1038/s41593-018-0326-7
28. Bernstein DP, Fink L, Handelsman L, et al. Initial reliability and validity of a new retrospective measure of child abuse and neglect. *Am J Psychiatry*. 1994;151(8):1132-1136. doi:10.1176/ajp.151.8.1132
29. Fydrich T, Sommer G, Tydecks S, Brähler E. Fragebogen zur sozialen Unterstützung (F-SozU): normierung der Kurzform (K-14). *Z Med Psychol*. 2009;18:43.
30. PHOTONAI. Homepage. Accessed December 3, 2023. <https://photon-ai.com/>
31. Winter N. A-Systematic-Evaluation-of-Machine-Learning-based-Biomarkers-for-Major-Depressive-Disorder. Accessed December 3, 2023. <https://github.com/www-mml/A-Systematic-Evaluation-of-Machine-Learning-based-Biomarkers-for-Major-Depressive-Disorder>
32. DeMars CE. Classical test theory and item response theory. In: Irwing P, Booth T, Hughes DJ, eds. *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*. Wiley-Blackwell; 2018:49-73. . doi:10.1002/9781118489772.ch2
33. Chicco D, Tötsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Min*. 2021;14(1):13. doi:10.1186/s13040-021-00244-z
34. Raykov T, Marcoulides GA. *Introduction to Psychometric Theory*. Routledge; 2011. doi:10.4324/9780203841624
35. Gell M, Eickhoff SB, Omidvarnia A, et al. The burden of reliability: how measurement noise limits brain-behaviour predictions. *bioRxiv*. Preprint posted online February 10, 2023. doi:10.1101/2023.02.09.527898
36. Cattarinussi G, Delvecchio G, Maggioni E, Bressi C, Brambilla P. Ultra-high field imaging in major depressive disorder: a review of structural and functional studies. *J Affect Disord*. 2021;290:65-73. doi:10.1016/j.jad.2021.04.056
37. Uhlhaas PJ, Liddle P, Linden DEJ, Nobre AC, Singh KD, Gross J. Magnetoencephalography as a tool in psychiatric research: current status and perspective. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2017;2(3):235-244. doi:10.1016/j.bpsc.2017.01.005
38. Goltermann J, Emden D, Leehr EJ, et al. Smartphone-based self-reports of depressive symptoms using the Remote Monitoring Application in Psychiatry (ReMAP): interformat validation study. *JMIR Ment Health*. 2021;8(1):e24333. doi:10.2196/24333
39. Zaremba D, Dohm K, Redlich R, et al. Association of brain cortical changes with relapse in patients with major depressive disorder. *JAMA Psychiatry*. 2018;75(5):484-492. doi:10.1001/jamapsychiatry.2018.0123
40. Rutherford S, Kia SM, Wolfers T, et al. The normative modeling framework for computational psychiatry. *Nat Protoc*. 2022;17(7):1711-1734. doi:10.1038/s41596-022-00696-5
41. Stein F, Buckenmayer E, Brosch K, et al. Dimensions of formal thought disorder and their relation to gray- and white matter brain structure in affective and psychotic disorders. *Schizophr Bull*. 2022;48(4):902-911. doi:10.1093/schbul/sbac002
42. Pelin H, Ising M, Stein F, et al. Identification of transdiagnostic psychiatric disorder subtypes using unsupervised learning. *Neuropsychopharmacology*. 2021;46(11):1895-1905. doi:10.1038/s41386-021-01051-0