Article

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

Nils R. Winter, MSc[1,2]; Julian Blanke, BSc[1]; Ramona Leenings, PhD[1,3]; Jan Ernsting, MSc[1,3,4]; Lukas Fisch, MSc[1]; Kelvin Sarink, MSc[1]; Carlotta Barkhau, MSc[1]; Daniel Emden, MSc[1]; Katharina Thiel, MSc[1]; Kira Flinkenflügel, MSc[1]; Alexandra Winter, MSc[1]; Janik Goltermann, PhD[1]; Susanne Meinert, PhD[1,5]; Katharina Dohm, PhD[1]; Jonathan Repple, MD[6,1]; Marius Gruber, MSc[1,6]; Elisabeth J. Leehr, PhD[1]; Nils Opel, MD[1,7,8,9]; Dominik Grotegerd, PhD[1]; Ronny Redlich, PhD[1,9,10]; Robert Nitsch, MD, PhD[2,5]; Jochen Bauer, PhD[11]; Walter Heindel, MD[11]; Joachim Groß, PhD[2,12]; Benjamin Risse, PhD[2,3,13], Till F. M. Andlauer, PhD[14]; Andreas J. Forstner, MD[15,16]; Markus M. Nöthen, MD[15]; Marcella Rietschel, MD[17]; Stefan G. Hofmann, PhD[18]; Julia-Katharina Pfarr, MSc[19,20]; Lea Teutenberg, MSc[19,20]; Paula Usemann, MSc[19,20]; Florian Thomas-Odenthal, MSc[19,20]; Adrian Wroblewski, PhD[19,20]; Katharina Brosch, PhD[19,20]; Frederike Stein, PhD[19,20]; Andreas Jansen, PhD[19,20,21]; Hamidreza Jamalabadi, PhD[19;] Nina Alexander, PhD[19,20]; Benjamin Straube, PhD[19,20]; Igor Nenadić, MD[19,20]; Tilo Kircher, MD[19,20]; Udo Dannlowski, MD, PhD[1,2*]; Tim Hahn, PhD[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]University of Münster, Institute for Geoinformatics, Münster, Germany
[14]Department of Neurology, Klinikum rechts der Isar, School of Medicine, Technical University of Munich, Munich, Germany
[15]Institute of Human Genetics, University of Bonn, School of Medicine and University Hospital Bonn, Bonn, Germany
[16]Institute of Neuroscience and Medicine (INM-1), Research Centre Jülich, Jülich, Germany
[17]Department of Genetic Epidemiology, Central Institute of Mental Health, Faculty of Medicine Mannheim, University of Heidelberg, Mannheim, Germany
[18]Department of Clinical Psychology, Philipps-University Marburg, Marburg, Germany
19Department 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

[*]Equally contributing last authors.

Corresponding author:

Nils R. Winter, nils.r.winter@uni-muenster.de, +49 251 83 51847

Albert-Schweitzer-Campus 1, 48149 Münster, Germany

Word count: 3231

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 1,801 patients with depression and healthy controls, even 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 to univariate analyses, no depression biomarker could be identified that classified individuals with clinically relevant performance.

Abstract

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: Whether precision psychiatry can solve this discrepancy and provide 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.

Design: Data is part of the Marburg-Münster Affective Disorders Cohort Study (MACS), a case-control clinical neuroimaging study.

Setting: Patients and controls were recruited from primary care and the general population in Münster and Marburg, Germany, from September 11, 2014, to September 26, 2018.

Participants: Patients with acute and lifetime MDD as well as healthy controls in the age range of 18 to 65 years. The Münster Neuroimaging Cohort (MNC) was used as an independent partial replication sample. Data were analyzed from April 2022 to June 2023.

Main Outcome and Measure: We quantify diagnostic classification accuracy on an individual level employing an extensive ML-based multivariate approach across a comprehensive range of neuroimaging modalities, including structural and functional Magnetic Resonance Imaging (MRI), Diffusion Tensor Imaging as well as a polygenic risk score for depression.

Results: A total of 1,801 individuals (856 patients [47.5%]) were included in the main analyses (mean [SD] age, 36.1 [13.1] years; 555 female patients [64.8%]). The MNC replication sample included 1,198 individuals (362 patients [30.1%]). Training and testing a total of 4 million ML models, we find accuracies for diagnostic classification between 48.1% and 62.0%. 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 MRI within MNC. Even under simulated conditions of perfect reliability, performance does 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 to univariate neuroimaging markers, no informative individual-level MDD biomarker – even under extensive ML optimization in a large sample of diagnosed patients – could be identified.

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 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]

Fuelled by the availability of large-scale datasets 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], 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, 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 multi-site analyses has previously been used to explain small effect sizes and inconsistent results.[13]

In summary, the existing literature on multivariate biomarker discovery in MDD does not allow for a conclusive evaluation of the clinical utility of ML approaches. Therefore, this study aims to establish an upper bound on 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 multivariate 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 is available.[14,15] Next, we did not rely on a single predictive algorithm but capitalized on the advances in ML software[16] 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 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 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 interviews (SCID), which provided standardized DSM-based MDD diagnosis and reduced the diagnostic uncertainty often hampering model performance in large-scale, multi-site data today.[12,17] 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.[15] All analyses were replicated within the two study sites. Additionally, an independent sample was used to replicate ML analyses for structural T_1-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 two sites (Marburg and Münster, Germany) using identical study protocols and harmonized scanner settings.[15] A sample of N=2,036 healthy participants and patients with major depression were recruited as part of the MACS cohort from September 11, 2014, to September 26, 2018 (*eMethods 1-3*). MDD diagnosis was assessed using the Structured Clinical Interview for DSM-IV, axis 1 disorders (SCID-I). Individuals with any history of neurological or medical conditions were excluded, resulting in a final sample of N=1,801 (see *eMethods 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 (see *eMethods 1* and *4-12*). Similar inclusion and exclusion criteria were used in the MNC replication sample (*eMethods 16*). This study follows TRIPOD reporting guidelines. The FOR2107 cohort project 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). Participants received financial compensation and gave written and informed consent.

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.[23,24] Structural connectomes were derived from DTI as fractional anisotropy (FA) and mean diffusivity

(MD). Functional connectomes, voxel-based local correlation (LCOR), the amplitude of low-frequency fluctuations (ALFF) as well as the fractional amplitude of low-frequency fluctuations (fALFF) were derived from resting-state functional MRI (rsfMRI). 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 to a polygenic risk score for depression (PRS) 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 etiology of major depression.[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 (HAMD). Structural T_1-weighted MRI data (FreeSurfer and VBM) was used in the MNC replication analysis.

Main outcomes

Accuracy of predicted diagnostic labels in all machine learning models was calculated using balanced classification accuracy (BACC). In addition, we calculated Matthew's correlation coefficient (MCC, Equation 1). For all metrics, mean and standard deviation across the 10 outer CV splits were reported to assess the generalizability of the predictive models.

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

Machine Learning analyses

A total of 4 million ML models to classify healthy participants and patients with MDD were trained, optimized, and evaluated (*eMethods 14, eTable 40*). 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. 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 CV 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 (omitting remitted patients) or recurrently depressed patients (omitting single episode patients), patients with or without comorbidities, patients currently receiving medication and those not currently on medication, males and females, as well as a homogeneous age group (age range 24 to 28) and replicated within the two study sites (*eMethods 3)*. 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] Scripts are available at https://github.com/wwu-mmll/A-Systematic-Evaluation-of-Machine-Learning-based-Biomarkers-for-Major-Depressive-Disorder.

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.[30] We

first computed MCC from the model predictions $\hat{y}$ and the actual diagnostic labels y (Equation 1).[31] This correlation was then corrected for an assumed reliability ρ using the attenuation formula (Equation 2).[32]

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

We conducted two separate attenuation correction analyses. First, we assumed the reliability $\rho_y = 0.28$ of an MDD diagnosis 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 in [31] (Equation 3, *eMethods 13*).

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

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 (see eMethods 15). In short, one 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.

Results

A total of 1,801 individuals (856 patients [47.5%]) 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). The MNC sample included 1,198 individuals (362 patients [30.1%], mean [SD] age, 35.3 [12.6] years; 209 female patients [57.7%] and 473 female healthy controls [56.6%], see e*Table 26*).

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*. The 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.

ML pipelines on subgroups of only acutely depressed (N=599), only remitted patients (N=297), only patients with (N=373) or without (N=482) comorbidities, or either patients that were (N=535) or were not currently medicated (N=321) showed similar results compared to the analysis containing all MDD patients ($BACC_{max}$=64.8%). 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 5-31*). Assessing the relationship between the training sample size and model performance, additional analyses suggest that models may quickly reach a performance plateau (eMethods 17 and eFigure 9).

Multimodal Data Integration

Integration of neuroimaging modalities using principal components from modality-specific PCAs achieved BACCs between 50.1% [4.0%] and 57.2% [4.4%] (*eTable 3*, *Figure 2*). Combining predicted labels from the unimodal models (across algorithms, across modalities, or both) into a majority-vote 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%].

Replicability Analyses

The main findings were replicated in the Marburg and Münster samples independently. Highest BACC was 59.2% [5.1%] in the Marburg and 60.0% [9.0%] in the Münster sample (see *eTable 20-25*). In the independent MNC sample, the highest BACC for regional cortical and subcortical surface, thickness, and volume was 54.0% [5.1%] while BACC for VBM analysis reached 53.4% [4.4%].

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[27] 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

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 of $\rho = 0.28$ as reported in the literature (*Figure 3a*). BACC for the best machine learning algorithm on resting-state connectivity increased to 71.8% [6.4%] after correction. 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 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

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. Similar effects were found for lower global assessment of functioning (GAF), higher medication load and the presence of comorbidities (*eTable 4*).

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 and 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 56-58% 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 upon 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[22,33], 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 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.[33] 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.[34,35]

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 most plausible. Rather than solely focusing on the diagnosis of MDD, looking at longitudinal data and clinically relevant outcomes across diagnoses may result in more accurate predictions, such as linking neuroimaging markers with long-term disease trajectories.[36,37] 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 and more unfavourable 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 behaviour.[1] Normative modelling capture

deviations of the individual patient, thereby overcoming the necessity for a common biological cause across all MDD patients.[38] Similarly, identifying MDD biotypes through clustering across DSM diagnoses might constitute a promising way forward.[39,40] 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 four main improvements over existing ML studies: First, a reduced risk of inflated predictive performance estimates by using nested cross-validation with sufficiently large test sets. Second, a systematic optimization of many possible ML pipelines and algorithms. Third, an integration of data from 11 neuroimaging modalities and preprocessing methods. Fourth, using 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 upon structured clinical 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 two study sites, along with the independent replication results for T1-weighted sMRI in the MNC sample, adds significant robustness to our findings and further supports our conclusions. Our additional sample size analyses indicate that model performance plateaus relatively quickly, suggesting that larger samples may not significantly enhance performance. However, our subgroup analyses, particularly those focusing on patients with or without comorbidities or treatment, result 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 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.

Conclusion

Based on this evidence from state-of-the-art methods and one of the most comprehensive datasets, 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 MDD patients.

Acknowledgments

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), MzH 3/020/20 from the Interdisciplinary Center for Clinical Research of the medical faculty of Münster to Tim Hahn, and by the German Science Foundation CRC 1451/A7 to Robert Nitsch. 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. Stefan G. Hofmann is supported through the Alexander von Humboldt Professorship and the LOEWE Spitzenprofessur.

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).

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.

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: Tim Hahn and Nils Ralf Winter had full access to all the data and take responsibility for the integrity of the data and accuracy of the data analysis

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).

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.

Figures

Figure 1. Overview of all analyses.

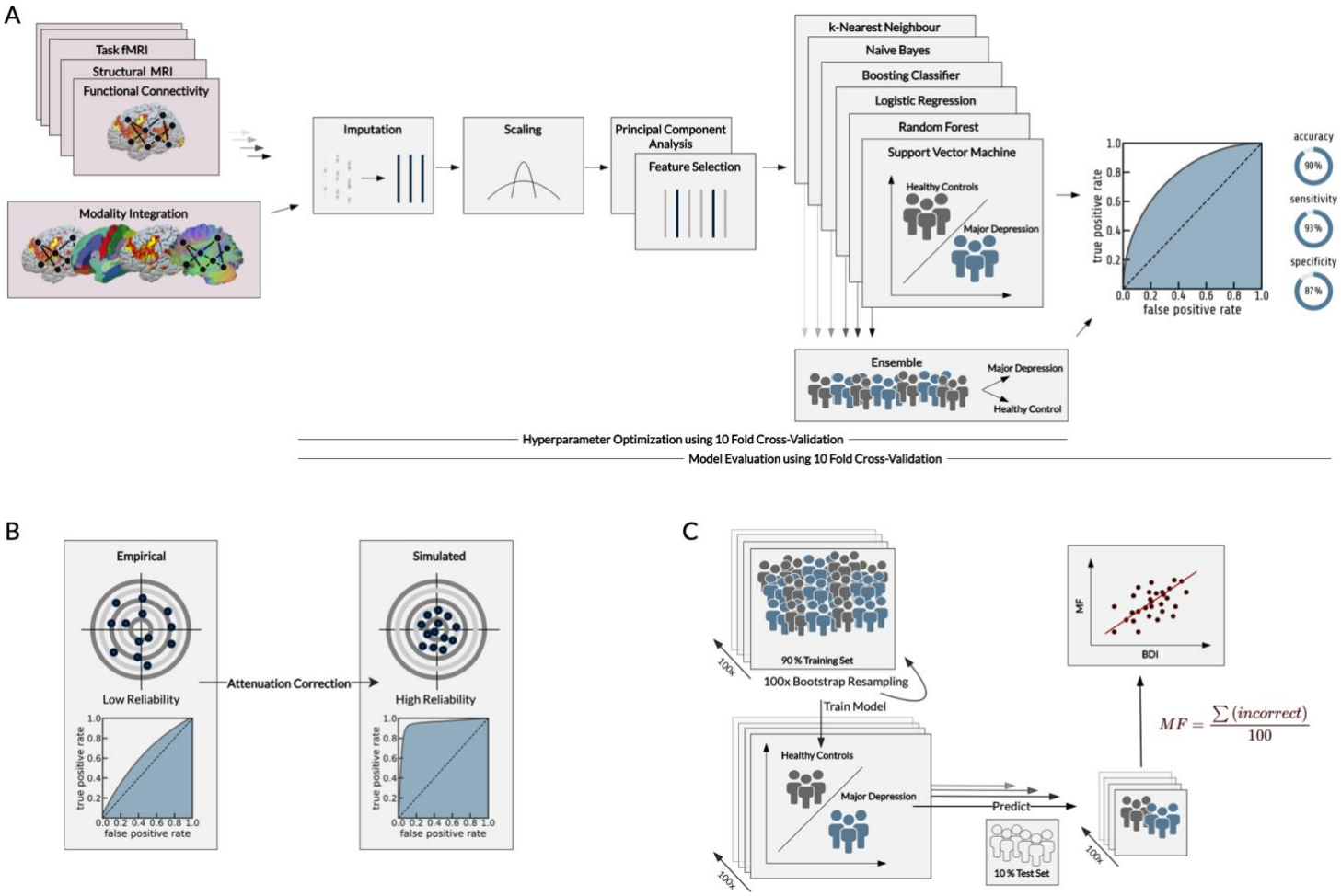

Figure 1. Overview of all analyses. (A) illustrates steps of the Machine Learning pipeline. (B) illustrates reliability correction and its effect on classification accuracy. (C) illustrates model error analysis using misclassification frequency (MF) through repeated bootstrapping.

Figure 2. Balanced accuracy for best machine learning pipelines.

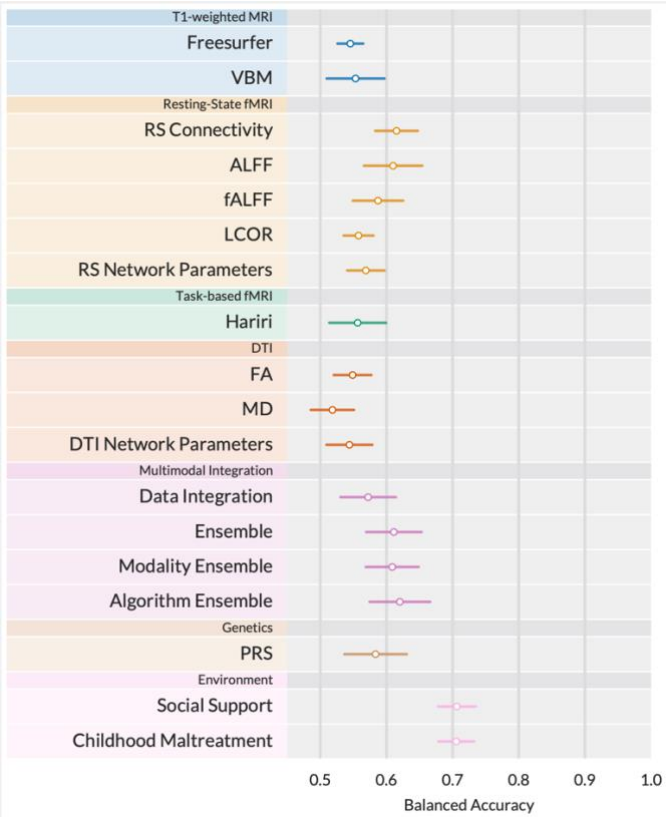

Figure 2. Balanced accuracy for best machine learning pipeline in every modality. Error bars display +-1 standard deviation calculated across the 10 outer cross-validation folds. VBM=Voxel-based morphometry, ALFF=Amplitude of low-frequency fluctuations, fALFF=fractional ALFF, LCOR=Local correlation,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

Figure 3. Balanced accuracy after attenuation correction.

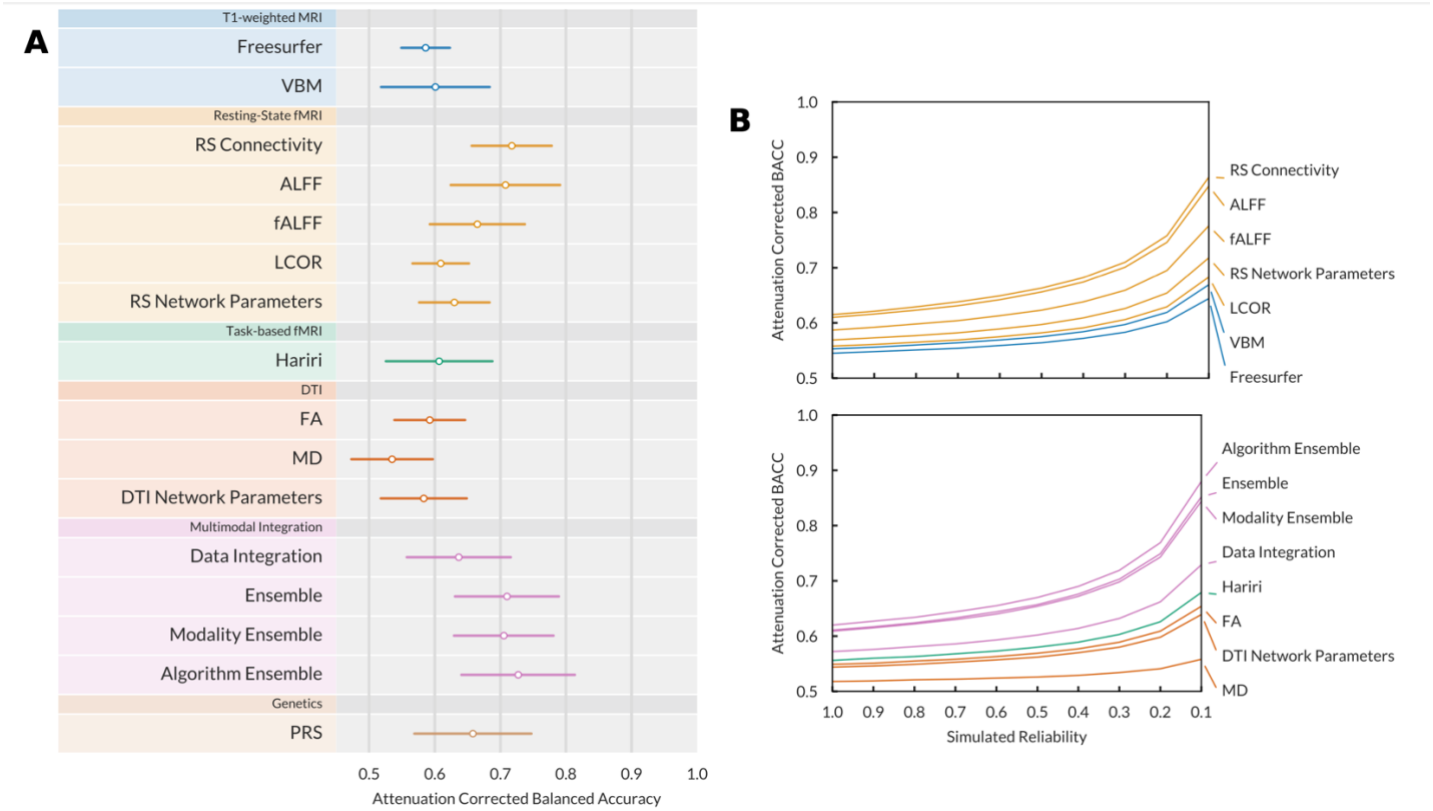

Figure 3. (a) Balanced accuracy for best machine learning pipeline in every modality after performing an attenuation correction for the empirical reliability of the MDD diagnosis. Error bars display +-1 standard deviation calculated across the 10 outer cross-validation folds. (b) Balanced accuracy for 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.

Tables

Table 1. Socio-demographic 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. For continuous variables, mean (SD) is reported.

References

1. Kendler KS. Toward a Philosophical Structure for Psychiatry. *Am J Psychiat*. 2005;162(3):433-440. doi:10.1176/appi.ajp.162.3.433
2. Insel TR, Cuthbert BN. Brain disorders? Precisely. *Science*. 2015;348(6234):499-500. doi:10.1126/science.aab2358
3. Insel T, Cuthbert B, Garvey M, et al. Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. Am J Psychiat. 2010;167(7):748-751. doi:10.1176/appi.ajp.2010.09091379
4. 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 Psychiat. 2020;177(5):422-434. doi:10.1176/appi.ajp.2019.19050560
5. 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
6. Dhamala E, Yeo BTT, Holmes AJ. Methodological Considerations for Brain-Based Predictive Modelling in Psychiatry. *Biol Psychiat*. Published online 2022. doi:10.1016/j.biopsych.2022.09.024
7. Bzdok D, Varoquaux G, Steyerberg EW. Prediction, Not Association, Paves the Road to Precision Medicine. *Jama Psychiat*. 2021;78(2):127-128. doi:10.1001/jamapsychiatry.2020.2549
8. Hahn T, Nierenberg AA, Whitfield-Gabrieli S. Predictive analytics in mental health: applications, guidelines, challenges and perspectives. *Mol Psychiatr*. 2017;22(1):37-43. doi:10.1038/mp.2016.201
9. Kambeitz J, Cabral C, Sacchet MD, et al. Detecting Neuroimaging Biomarkers for Depression: A Meta-analysis of Multivariate Pattern Recognition Studies. Biol Psychiat. 2017;82(5):330-338. doi:10.1016/j.biopsych.2016.10.028
10. Meehan AJ, Lewis SJ, Fazel S, et al. Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. Mol Psychiatr. Published online 2022:1-9. doi:10.1038/s41380-022-01528-4
11. Flint C, Cearns M, Opel N, et al. Systematic misestimation of machine learning performance in neuroimaging studies of depression. *Neuropsychopharmacol*. 2021;46(8):1510-1517. doi:10.1038/s41386-021-01020-7
12. Stolicyn 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
13. Schmaal L, Veltman DJ, Erp TGM van, et al. Subcortical brain alterations in major depressive disorder: findings from the ENIGMA Major Depressive Disorder working group. Mol Psychiatr. 2016;21(6):806-812. doi:10.1038/mp.2015.69
14. 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 Psy Clin N*. 2019;269(8):949-962. doi:10.1007/s00406-018-0943-x
15. 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
16. 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
17. Schmaal L, Veltman DJ, Erp TGM van, et al. Response to Dr Fried & Dr Kievit, and Dr Malhi et al. Mol Psychiatr. 2016;21(6):726-728. doi:10.1038/mp.2016.9
18. Marek S, Tervo-Clemmens B, Calabro FJ, et al. Reproducible brain-wide association studies require thousands of individuals. *Nature*. Published online 2022:1-7. doi:10.1038/s41586-022-04492-9
19. Nikolaidis A, Chen AA, He X, et al. Suboptimal phenotypic reliability impedes reproducible human neuroscience. *Biorxiv*. Published online 2022:2022.07.22.501193. doi:10.1101/2022.07.22.501193
20. 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 Psychiat. 2013;170(1):59-70. doi:10.1176/appi.ajp.2012.12070999
21. Fried EI, Flake JK, Robinaugh DJ. Revisiting the theoretical and methodological foundations of depression measurement. *Nat Rev Psychology*. 2022;1(6):358-368. doi:10.1038/s44159-022-00050-2
22. Greene AS, Shen X, Noble S, et al. Brain–phenotype models fail for individuals who defy sample stereotypes. Nature. Published online 2022:1-10. doi:10.1038/s41586-022-05118-w
23. Gaser C, Kurth F. *Computational Anatomy Toolbox CAT12*. 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-switz*. 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. 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 Psychiat. 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). *Zeitschrift für Medizinische Psychologie*. 2009;18:43.
30. DeMars CE. Classical Test Theory and Item Response Theory. In: *The Wiley Handbook of Psychometric Testing*. ; 2018:49-73. doi:10.1002/9781118489772.ch2
31. 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
32. Raykov T, Marcoulides GA. *Introduction to Psychometric Theory*. Routledge; 2011.
33. Gell M, Eickhoff SB, Omidvarnia A, et al. The Burden of Reliability: How Measurement Noise Limits Brain-Behaviour Predictions. *biorXiv*. Published online 2023. doi:10.1101/2023.02.09.527898
34. 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 Disorders*. 2021;290:65-73. doi:10.1016/j.jad.2021.04.056
35. Uhlhaas PJ, Liddle P, Linden DEJ, Nobre AC, Singh KD, Gross J. Magnetoencephalography as a Tool in Psychiatric Research: Current Status and Perspective. Biol Psychiat. 2017;2(3):235-244. doi:10.1016/j.bpsc.2017.01.005
36. 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 Heal*. 2021;8(1):e24333. doi:10.2196/24333
37. Zaremba D, Dohm K, Redlich R, et al. Association of Brain Cortical Changes With Relapse in Patients With Major Depressive Disorder. Jama Psychiat. 2018;75(5):484. doi:10.1001/jamapsychiatry.2018.0123
38. Rutherford S, Kia SM, Wolfers T, et al. The normative modeling framework for computational psychiatry. *Nat Protoc*. Published online 2022:1-24. doi:10.1038/s41596-022-00696-5
39. 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. *Schizophrenia Bull*. 2022;48(4):902-911. doi:10.1093/schbul/sbac002
40. Pelin H, Ising M, Stein F, et al. Identification of transdiagnostic psychiatric disorder subtypes using unsupervised learning. Neuropsychopharmacol. 2021;46(11):1895-1905. doi:10.1038/s41386-021-01051-0