Multimodal data integration advances longitudinal prediction of the naturalistic course of depression and reveals a multimodal signature of remission during 2-year follow-up

Habets, P.C.; Thomas, R.M.; Milaneschi, Y.; Jansen, R.; Pool, R.; Peyrot, W.J.; ... ; Vinkers, C.H.

Biological Psychiatry

Archival Report

Multimodal Data Integration Advances Longitudinal Prediction of the Naturalistic Course of Depression and Reveals a Multimodal Signature of Remission During 2-Year Follow-up

Philippe C. Habets, Rajat M. Thomas, Yuri Milaneschi, Rick Jansen, Rene Pool, Wouter J. Peyrot, Brenda W.J.H. Penninx, Onno C. Meijer, Guido A. van Wingen, and Christiaan H. Vinkers

ABSTRACT

BACKGROUND: The ability to predict the disease course of individuals with major depressive disorder (MDD) is essential for optimal treatment planning. Here, we used a data-driven machine learning approach to assess the predictive value of different sets of biological data (whole-blood proteomics, lipid metabolomics, transcriptomics, genetics), both separately and added to clinical baseline variables, for the longitudinal prediction of 2-year remission status in MDD at the individual-subject level.

METHODS: Prediction models were trained and cross-validated in a sample of 643 patients with current MDD (2-year remission n = 325) and subsequently tested for performance in 161 individuals with MDD (2-year remission n = 82).

RESULTS: Proteomics data showed the best unimodal data predictions (area under the receiver operating characteristic curve = 0.68). Adding proteomic to clinical data at baseline significantly improved 2-year MDD remission predictions (area under the receiver operating characteristic curve = 0.63 vs. 0.78, p = .013), while the addition of other omics data to clinical data did not yield significantly improved model performance. Feature importance and enrichment analysis revealed that proteomic analytes were involved in inflammatory response and lipid metabolism, with fibrinogen levels showing the highest variable importance, followed by symptom severity. Machine learning models outperformed psychiatrists' ability to predict 2-year remission status (balanced accuracy = 71% vs. 55%).

CONCLUSIONS: This study showed the added predictive value of combining proteomic data, but not other omics data, with clinical data for the prediction of 2-year remission status in MDD. Our results reveal a novel multimodal signature of 2-year MDD remission status that shows clinical potential for individual MDD disease course predictions from baseline measurements.

https://doi.org/10.1016/j.biopsych.2023.05.024

Major depressive disorder (MDD) is a heterogeneous disorder in which both treatment response and prognosis vastly differ among individuals. Around 20% to 25% of patients with MDD are at risk for chronic depression, independent of initial treatment type (1). The ability to predict the disease course of individuals with MDD early on is essential for optimal treatment planning because this could allow for early treatment intensification for patients with a low long-term chance of remission and the potential bypassing of initial first-choice treatments.

Previous studies have yielded insights into clinical, psychological, and biological markers for chronicity in depression. Chronicity in depression has been related to longer symptom duration, increased symptom severity, and earlier age of onset (1,2); higher levels of neuroticism and lower levels of extraversion and conscientiousness (3); and various inflammatory markers (4), low levels of vitamin D (5), metabolic syndrome (6), and lower cortisol awakening response (7). However, statistically significant differences on a group level will not always be useful for the prediction of disease course at the individual level, due to either low effect sizes or redundancy with respect to other more predictive variables. While multiple studies have shown that biological data can be used to make accurate diagnostic predictions of MDD cases and healthy control participants (8–10), individual prediction of disease course in depression has proven to be a difficult task, with a recent systematic review and meta-analysis showing an average accuracy of only 60% for predicting remission or resistance after treatment in adequate-quality studies (i.e., studies that used a follow-up time span of 8–24 weeks) (11).

One challenge of predicting MDD outcomes is that the etiology and phenotype of MDD differ widely between individuals, and large interindividual variation may exist with regard to

SEE COMMENTARY ON PAGE 908

ISSN: 0006-3223

relevant predictors (12–15). Especially when only a limited set of predictor variables are included for prediction modeling, the chances of accurately capturing complex multimodal system dynamics (i.e., the biopsychosocial model of depression) with those variables decrease further. With the availability of novel machine learning methods that can learn complex, high-dimensional nonlinear patterns in data, a solution to this problem may be to incorporate multiple high-dimensional data sources, each containing putative predictive factors.

While several studies have tried to predict MDD course from a range of different data modalities (e.g., clinical variables, metabolomics, imaging data, epigenetics) (16–20), combining multiple data modalities for predicting MDD chronicity has been relatively uncommon. In one recent analysis of the NESDA (Netherlands Study of Depression and Anxiety) cohort (21) that integrated clinical, psychological, and biomarker data, predictions of 2-year chronicity (defined as 2-year remission status) in MDD reached a balanced accuracy of 62% using a penalized linear model (22). Adding limited biological data yielded no improvement in prediction accuracy over the combination of clinical and psychological data (22). Another NESDA analysis using a similar model with epigenetic data showed an area under the curve of 0.571 for predicting the same 2-year remission status outcome (20). Remission after 6 years was predicted more accurately, but the reported area under the receiver operating characteristic curve (AUROC) (0.724) was based on 10-fold cross-validation results rather than on outheld test set results, possibly leading to over-optimistic performance metrics (23,24). Interestingly, neither adding genome-wide single nucleotide polymorphism data nor adding 27 clinical, demographic, and lifestyle variables improved predictions (20). This is an important finding because no other studies have integrated features from multiple high-dimensional biological data (i.e., multiomic data) and clinical data to improve predictions of MDD disease course. This contrasts with other fields of medicine, where multimodal data integration has led to significant advancements in the field of precision medicine (25), most notably in the field of precision oncology (26,27).

To further investigate the potential of multimodal data in the field of precision psychiatry, the current study explored the potential of integrating multiomic, clinical, psychological, and demographic data. To this end, we used high-dimensional multimodal data collected from 804 NESDA subjects with MDD (21). In a subset of 643 individuals (80% of the total sample), using combinations of lipid metabolomic, proteomic, transcriptomic, genetic, demographic, psychological, and clinical data measured at baseline (i.e., from the moment of MDD diagnosis), we used cross-validated machine learning models to predict MDD remission after 2 years of follow-up. To allow for nonlinear pattern detection, and to assess the potential benefit of nonlinear models over linear models in multimodal pattern detection, we used several linear and nonlinear machine learning algorithms (elastic net, support vector machine, random forest, XGBoost, artificial neural network). The validity of the models' predictions was then tested in a separate outheld test group of 161 individuals (20% of the total sample).

To embed our machine learning models' performance metrics in the context of clinical expertise (i.e., how good or bad predictive performances are from a clinicians' point of view), we also had 4 clinical psychiatrists predict 2-year remission status in a subset of 200 individuals on the basis of extensive clinical information.

METHODS AND MATERIALS

Participants

In the current study, we included data that were collected as part of a larger, multicenter longitudinal study (NESDA, N = 2981) (see Supplemental Methods in Supplement 1) (21). We included a subsample from the NESDA cohort consisting of 804 subjects and used the following inclusion criteria [identical to our previous study (22)]: 1) presence of a DSM-IV MDD or dysthymia diagnosis (or both) during the past 6 months at baseline, established using the structured Composite International Diagnostic Interview (version 2.1) (28); 2) confirmation of depressive symptoms in the month before baseline either by the Composite International Diagnostic Interview or by the Life Chart Interview (29); and 3) availability of 2-year follow-up data on DSM-IV diagnosis and depressive symptoms measures with the Composite International Diagnostic Interview. The ethical review board of the Vrije Universiteit University Medical Center and subsequently the review boards of each participating center approved the NESDA research protocol (reference No. 2003/183). After providing complete verbal and written information about the study, informed consent was obtained from all participants at the start of the baseline assessment.

We defined 2 outcome groups: remission or no remission 2 years after follow-up. We based the outcome on the presence or absence of a current unipolar depression diagnosis (6-month recency of an MDD diagnosis or dysthymic disorder) at 2-year follow-up according to DSM-IV criteria. The label "remitted" was given to individuals who were in stable remission for at least 6 months, and the label "nonremitted" was given to participants who, at the 2-year time point, were diagnosed with depression and had experienced active symptoms during the past 6 months. This approach was aimed at improving the reliability of 2-year remission status labels by reducing misclassification from recent diagnoses or nearby relapses. Sample characteristics and statistics for both outcome groups of all 804 included subjects are shown in Table 1. Additionally, we provide statistics on comorbid anxiety disorders at baseline and at the 2-year time point in Table S6 in Supplement 2.

Clinical Variables

We included a set of 10 relevant clinical, psychological, and demographic predictor variables (which we will refer to hereafter as "clinical variables"), including age, sex, years of education, depressive symptom severity (Inventory of Depressive Symptoms—Self-Report questionnaire) (30) (both as total score and as severity category ranging from 1 to 5), and 5 personality dimensions (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness) measured with the NEO Five-Factor Inventory (31). Additional information on variable inclusion is provided in Supplemental Methods in Supplement 1.

Table 1. Sample Characteristics

	Presence of Unipolar Depression at Follow-up				
	No	Yes	Statistics	*p* Value	*p* Value (Bonferroni Corrected)
Sample Size	407 (51%)	397 (49%)	–	–	–
Age, Years	41.07 (12.55)	42.89 (11.83)	$F = 4.49$	.03	.28
Sex, Male	133 (33%)	145 (37%)	$\chi^2 = 1.15$	.28	>.99
Education, Years	11.60 (3.17)	11.51 (3.37)	$F = 0.14$	.71	>.99
Body Mass Index	26.06 (5.43)	26.10 (5.54)	$F = 0.0086$	.93	>.99
Recruitment Type, Primary Care/Specialized Care/General Population	162/209/36	143/229/25	$\chi^2 = 3.96$	.14	>.99
Diagnosis at Baseline, DD/Dysth/MDD	75/16/316	122/18/257	$\chi^2 = 17.28$	<.0002	<.002
Antidepressant Use at Baseline	166 (41%)	189 (48%)	$\chi^2 = 3.52$	.06	.49
Antidepressant Use at 2-Year Follow-up	127 (31%)	175 (44%)	$\chi^2 = 13.66$	.0002	<.002
Psychopharmaca Use (Any Type) Past 3 Years at Baseline	173 (43%)	194 (49%)	$\chi^2 = 3.03$	.08	.66

Data are given as *n* (%) or mean (SD). The table shows characteristics of the total sample separately by the presence or absence of a unipolar depression diagnosis (MDD or dysthymia) 2 years after baseline measurement. DD indicates both MDD and dysthymia diagnoses.

DD, double depression; Dysth, dysthymia; MDD, major depressive disorder.

Proteomic Variables

Proteomic data was collected and available at baseline for only 611 of the total 804 subjects. For these 611 individuals, a panel of 243 analytes involved in endocrinological, immunological, metabolic, and neurotrophic pathways were measured in serum at baseline using a multiplex enzyme-linked immunosorbent assay. A full list of the 243 analytes and their inclusion in predictive modeling with missing percentages per variable can be found in Table S1 in Supplement 2. Supplemental Methods in Supplement 1 provides additional details on data collection, missingness, imputation, and processing.

Lipid and Metabolite Variables

A lipid-focused metabolomics platform was used to measure 231 lipids, metabolites, and metabolite ratios in plasma at baseline for 790 of the 804 included subjects (14 individuals had no metabolomic data available and were excluded from metabolomic-informed predictions). From now on, we refer to this data as "lipidomic data." Additional lipidomic data processing details are described in Supplemental Methods in Supplement 1.

Transcriptomic Variables

Transcriptome-wide expression levels were measured in whole blood for 669 of the 804 included individuals. For each subject, 44,241 microarray probes targeting 23,588 genes were available for analysis. We used a data-driven feature reduction and processing pipeline to select a final set of 87 genes for machine learning modeling (see full description in Supplemental Methods in Supplement 1). We also analyzed whether a proteomic panel–focused approach to selecting genes in the transcriptomic data would yield improved results. We did this by matching genes in the transcriptomic data based on being included in Kyoto Encyclopedia of Genes and Genomes pathway categories (32) that have been found to be enriched in the coverage of the proteomic panel. See Supplemental Methods in Supplement 1 for details.

Genotype Data

Using the LDpred package in R (33), a total of 29 polygenic risk scores (PRSs) were calculated for 701 of the 804 included subjects with available genotype data. Additional details about DNA extraction and PRS calculation can be found in Supplemental Methods in Supplement 1. Table S3 in Supplement 2 lists all 29 phenotypes for which PRSs were calculated (e.g., MDD, anxiety, neuroticism).

Analysis

All analyses were performed using the programming languages R (version 4.0.3) and Python (version 3.8.5). All R and Python codes are made publicly available on GitHub at https://github.com/pchabets/chronicity-prediction-depression.

Machine Learning Analysis

Full details about data preprocessing, imputation methods, hyperparameter tuning, training, validation, test procedures, and model evaluation are described in Supplemental Methods in Supplement 1. Table S4 in Supplement 2 also lists preprocessing, imputation, and hyperparameter settings for each data type and model. In short, first, XGBoost models (34) were trained using each data modality separately to predict 2-year remission. Second, to investigate possible prediction augmentation effects of combining clinical and high-dimensional biological data, separate XGBoost models were trained using the combination of clinical data added to each of the separate omics data sources. Third, another XGBoost model was trained using the combination of all data modalities together. We also investigated the nature of the multimodal predictive signature by running different linear and nonlinear algorithms including elastic net, support vector machine, random forest, and a feed-forward densely connected artificial neural network using 1) only clinical features (i.e., severity scores, psychological and demographic variables); 2) only proteomics data; and 3) the combination of both data modalities. For each model, prediction augmentation by adding proteomics data to the clinical data was evaluated using the

AUROC. Additional details about model evaluation and comparison are provided in Supplemental Methods in Supplement 1.

All performance metrics reported are from validating the trained models on outheld test data.

Feature Importance Analysis

Feature importance analysis was based on computing Shapley values for every feature included in the best-performing XGBoost model using the Shapley additive explanations (SHAP) implementation for XGBoost (35,36). Protein-protein interaction and enrichment analysis was performed using the metascape platform (37). Additional details about SHAP and enrichment analysis are provided in Supplemental Methods in Supplement 1.

Human Predictions

Four human raters (trained and board-certified psychiatrists) independently predicted 2-year remission status for 200 subjects with MDD using clinical baseline data. Each rater was given 2 sets of samples for prediction. In the first sample, raters had to predict the 2-year remission status of subjects based on the same 10 clinical baseline predictor variables used by the machine learning models. In the second sample, raters also had access to baseline data on 1) dysthymia diagnosis, 2) MDD history, 3) anxiety diagnosis (lifetime), 4) 1-month recency of anxiety disorder symptoms, 5) alcohol diagnosis status (lifetime), 6) recency of alcohol abuse or dependency, and 7) total disease history (totaling 17 baseline predictor variables). In addition, we trained another XGBoost model on the same set of additional clinical data to allow for a more direct comparison of predictive performance between human raters and trained models. The human rating was set up single blinded, meaning that none of the human raters were given information on any of the model's performances before finishing their predictions. Additional details about the human prediction process, inter-rater agreement analysis, and preprocessing of the data for the XGBoost model are described in Supplemental Methods in Supplement 1.

RESULTS

Proteomic Data Are Most Informative for Predicting 2-Year Remission Status

First, we tested how well 2-year remission in MDD could be predicted for each data modality separately. For each of the available data modalities, the train, validation, and test sets used for the classification models approximated balanced distributions of the 2 outcome classes (Table S4 in Supplement 2). For unimodal data predictions, the model using proteomic data showed the highest performance (AUROC = 0.67, balanced accuracy = 0.68), followed by the models informed by clinical data (AUROC = 0.63, balanced accuracy = 0.62) and genetic data (AUROC = 0.61, balanced accuracy = 0.60) (Figure 1; Table S4 in Supplement 2). All models reached accuracy levels significantly above chance level. For the model informed by PRSs, accuracy only reached a significant level when using a cutoff on the ROC curve that made the model significantly biased toward false-negative classifications (McNemar's test, $p = 1.86 \times 10^{-6}$) (Table S4 in Supplement 2). The XGBoost model using all 63 clinical variables that were included in the previous prediction study by Dinga *et al.* (22) did not outperform the XGBoost model using the selected 10 clinical variables (AUROC = 0.61 vs. AUROC = 0.63, $p = .71$), indicating that no superior nonlinear information was detected in the discarded 53 clinical variables. Using a proteomic panel–focused approach to feature selection in the transcriptomic data resulted in the model's accuracy level failing to perform statistically significantly above chance level (AUROC = 0.54, $p = .90$) (see Figure S3 in Supplement 1).

Combining Clinical and Proteomic Data Augments Prediction Performance

Models informed by both clinical and omics data outperformed models informed by unimodal omics data in every case, most robustly for combining proteomic and clinical data (Figure 2). All combinations of clinical and omics data resulted in higher predictive performance than the model informed by clinical data only, except for combining clinical and transcriptomic data (Figure 2). Although a clear trend in augmented predictions by combining omics with clinical data was observed for all omics data (Figure 2), only the augmented performance of adding proteomic to clinical data reached statistical significance (AUROC = 0.78 vs. AUROC = 0.63, $p = .013$).

To further investigate the augmented prediction of 2-year remission status when adding proteomic to clinical data, we used several linear and nonlinear machine learning models informed by clinical, proteomic, and the combination of both types of data (Table S4 in Supplement 2). Informing machine learning models by only proteomic data resulted in low predictive performance for linear models compared with nonlinear models (Figure S2 in Supplement 1). Augmented predictive performance by adding proteomic data to clinical data was not found for any linear model, but it was observed for all nonlinear models, was most pronounced for XGBoost, and reached statistical significance only for the XGBoost model ($p = .013$) (Figure S2 in Supplement 1).

Additionally, we assessed to what extent the predictive performance of the XGBoost model could be attributed to (captured interactions with) possible confounding factors (i.e., antidepressant use at baseline, other psychopharmaceuticals used at baseline, years of education, center of patient recruitment, body mass index, age, and sex). Using a method recently proposed method by Dinga *et al.* (38) (see Supplemental Methods in Supplement 1) that provides post hoc controlling for confounders on the level of machine learning predictions using component analysis of deviance explained (D^2) in outcome labels, we found that the multimodal XGBoost predictions not only explained deviance in the 2-year remission outcomes independent of the possible confounders (independent $D^2 = 0.13$) but also explained a substantially larger amount of deviance than all confounders (independent $D^2 = 0.038$) or the interaction of confounder information and predictions (shared $D^2 = 0.028$) (Figure S4 in Supplement 1).

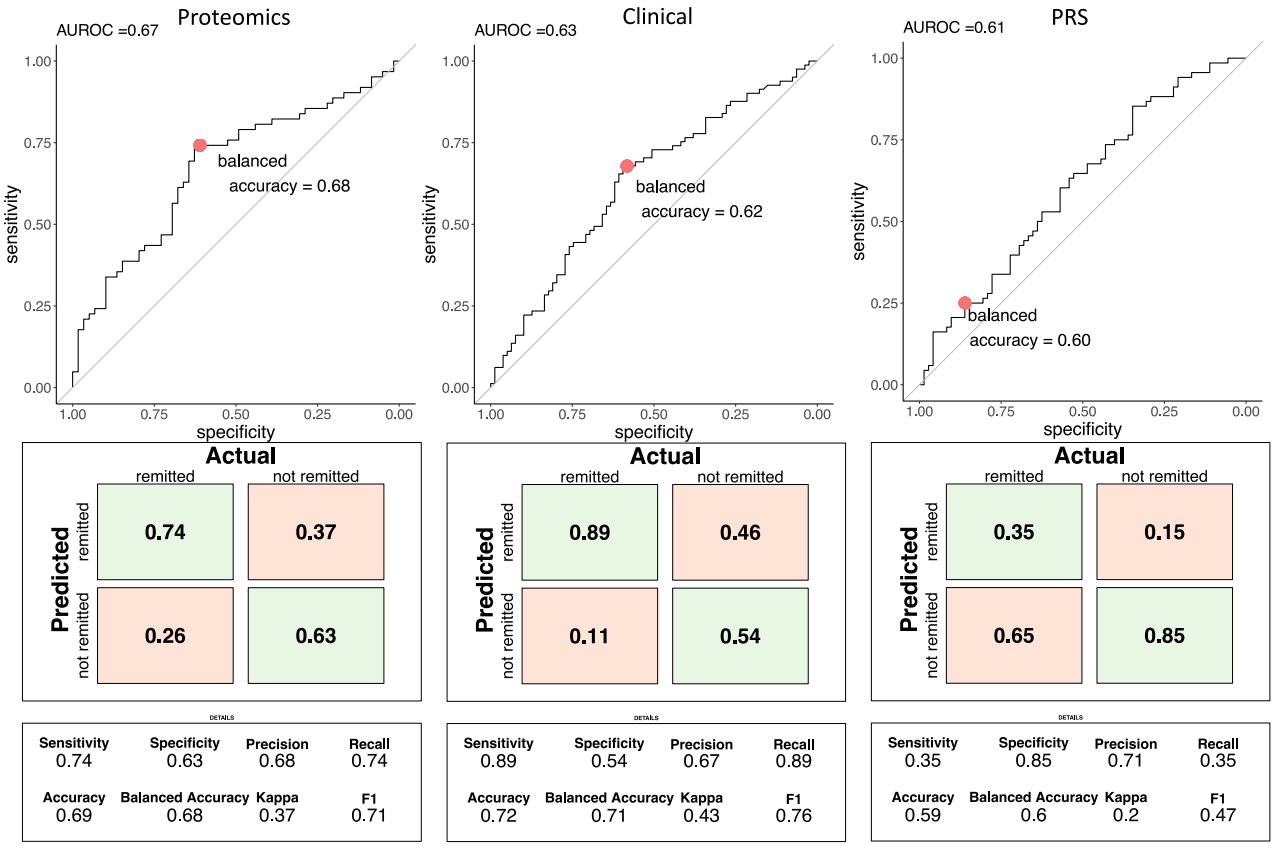


Figure 1. Predictive performance of XGBoost models informed by proteomic data (left), clinical data (middle), or polygenic risk score (PRS) (genetic) data (right). Receiver operating characteristic curves are plotted separately, with the reported area under the receiver operating characteristic curve (AUROC) and the maximum balanced accuracy shown for the optimal class probability cutoff. For each model, a confusion matrix is shown with additional performance metrics.

Variable Importance Analysis Shows Predictive Pattern Enrichment of Analytes Involved in Inflammatory Response and Lipid Metabolism

SHAP analysis was performed on the best-performing unimodal and multimodal informed models (i.e., XGBoost informed by proteomic data and XGBoost informed by clinical and proteomic data). For both the proteomics-only model and the model informed by both clinical and proteomic data, blood fibrinogen levels showed the highest mean absolute SHAP values (Figure 3; Table S5 in Supplement 2). Symptom severity at baseline was the most predictive clinical feature for 2-year remission status in MDD (Figure 3; Table S5 in Supplement 2). For the proteomics-only model, 109 analytes had an average absolute SHAP value > 0 (i.e., were informative for predictions). For the combined data model, 42 features were informative for predicting 2-year remission in the MDD model, including 38 proteomic analytes. Age, years of education, and sex were not attributed any SHAP values in the multimodal XGBoost model (Table S5 in Supplement 2).

Proteomic analytes that were informative in the combined data model and in the proteomics-only model were analyzed separately for protein-protein interactions and pathway enrichments. Network analysis of protein-protein interactions revealed densely connected subnetworks associated with inflammatory response and lipid metabolism for both the unimodal and multimodal XGBoost models, with enrichment of Reactome, gene ontology, and WikiPathway terms related to interleukin-10 signaling, chemokine signaling pathway, cholesterol esterification, and reverse cholesterol transport (Figure 4).

Human Prediction of 2-Year Remission Status From Clinical Data

Four clinical psychiatrists independently predicted 2-year remission status retrospectively from baseline data for 200 subjects with balanced subjects' outcome distribution (2-year nonremitted $n = 100$, remitted $n = 100$). Using the 10 clinical features that the clinical XGBoost model was informed by, human raters had an average accuracy of 0.51 (minimum = 0.35, maximum = 0.63) (Figure 5). When additional relevant clinical baseline data were available to the human raters, the raters' average prediction accuracy increased to 0.55 (minimum = 0.33, maximum = 0.65) (Figure 5). Inter-rater reliability between the 4 raters was low (Fleiss' kappa = 0.32, $p = 7.26 \times 10^{-7}$). Both XGBoost models (trained on the limited and extended clinical data) outperformed the human predictions

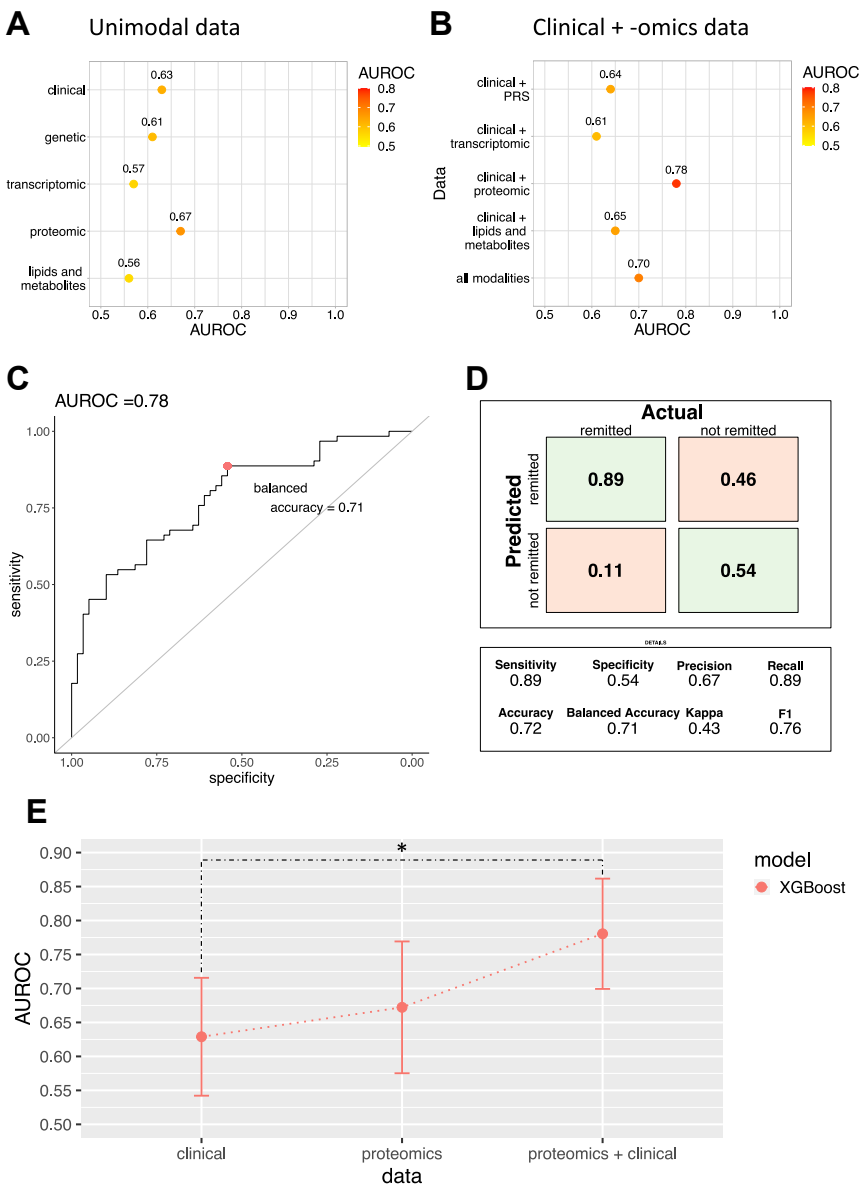


Figure 2. **(A)** Predictive performance of XGBoost models trained on several single data modalities. The y-axis shows the data modality used for training and testing the model. The x-axis shows the area under the receiver operating characteristic curve (AUROC) score of each model tested on the outheld test set. **(B)** The same as **(A)**, but results for multimodal data are shown. **(C, D)** ROC curve plotted with AUROC for the XGBoost model informed by both clinical and proteomic data. The maximum balanced accuracy is shown on the ROC curve, showing the optimal probability cutoff for highest classification accuracy. On the right, the confusion matrix with additional performance metrics is shown for this model. **(E)** Performance metrics (AUROC) of the XGBoost models informed by only clinical data, only proteomic data, or the combination of both are shown, with the 95% CI of the AUROC indicated by the whiskers. A significant difference in AUROC values between models is indicated by an asterisk. PRS, polygenic risk score.

(Figure 5). The XGBoost model trained on the extended set of clinical information performed slightly better than the model trained on the limited set of 10 clinical variables but not significantly better (AUROC = 0.65 vs. AUROC = 0.63, p = .75). Additionally, it was still significantly outperformed by the model trained on both clinical and proteomic data (AUROC = 0.78 vs. AUROC = 0.65, p = .03).

DISCUSSION

In this study, we showed that longitudinal prediction of 2-year MDD remission status substantially benefited from integrating multimodal data compared with relying exclusively on unimodal data. More specifically, model predictions improved significantly when combining proteomic and clinical data (Figure 2). Our model that was informed only by clinical data showed identical performance to the previously reported performance of a linear model using multiple data modalities, including those 10 clinical variables, in the same dataset (22). The performance of our model predictions increased significantly when proteomic data were added, but only for nonlinear models, suggesting superior multimodal predictive pattern detection by nonlinear models over linear models (Figure S2 in Supplement 1). Subsequent SHAP analysis revealed a multimodal predictive signature consisting of baseline symptom

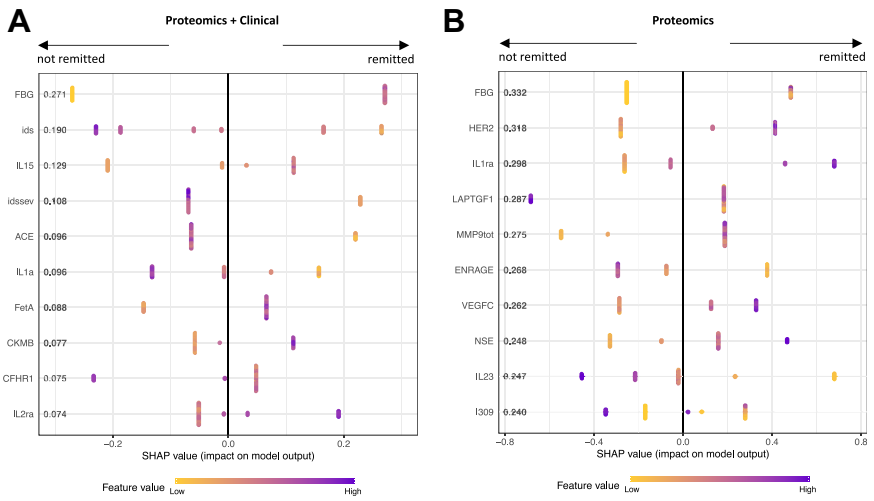


Figure 3. **(A)** Shapley additive explanations (SHAP) analysis results visualized for the XGBoost model informed by both clinical and proteomic data. On the y-axis, the 10 most important features (ranked by absolute average SHAP value, i.e., global SHAP value) are shown with their respective global SHAP value. Each dot in the graph indicates a single prediction (i.e., 1 subject). The position of the dots on the x-axis shows the impact that the feature value had for that individual prediction, with a negative SHAP value meaning that the model's decision was pushed toward a "not remitted" classification and a positive SHAP value meaning that the feature's value pushed the model's decision toward predicting the "remitted" class. Colors indicate the relative feature value measured in a subject (relative to the mean of all subjects), with yellow indicating a relatively low value and purple indicating a relatively high value. Note that most dots are stacked on each other because a range of feature values for several individuals can result in similar SHAP values (i.e., different feature values can influence the model's decision by the same magnitude and direction). **(B)** Similar to **(A)**, but here SHAP results of the XGBoost model informed by proteomic data only are shown. Proteomic abbreviations shown on the y-axis are listed by full name in Table S1 in Supplement 2. IDS (Inventory of Depressive Symptoms) is a continuous measure of symptom severity. IDSSEV (Inventory of Depressive Symptoms Severity) is a categorical measure of symptom severity.

severity, personality traits, and peripheral blood biomarkers related to the immune system and lipid metabolism.

Interpretation of Predictive Features

We have included a full, in-depth discussion of the feature importance and enrichment analysis results and their limitations in Supplemental Discussion in Supplement 1. A short summary follows below.

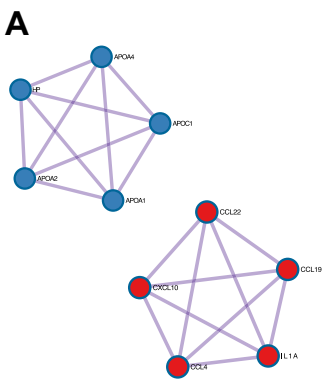
The findings of the SHAP analysis support previous research that suggests that symptom severity is predictive of MDD chronicity (22). While a set of 3 inflammatory markers (C-reactive protein, interleukin 6, tumor necrosis factor α) did not improve predictions in the same previous study (22), the current study's proteomic analysis showed a predictive inflammatory component as part of a multimodal signature that was predictive of 2-year remission status in depression, indicating the need for higher proteomic resolution. Due to the multivariate nature of the predictive model, no valid conclusion can be drawn about whether high or low fibrinogen is a risk factor for 2-year nonremission (39). Although the model informed by all data modalities performed better than the clinical-only model (AUROC = 0.70 vs. 0.63), integrating all omics data with clinical data did not yield improved predictions over the combined proteomics and clinical model, possibly indicating redundancy of the other omics data in our NESDA sample considering 2-year MDD remission predictions (Figure 2A, B).

While combining several PRSs resulted in prediction accuracy similar to the model informed by clinical data only, our results suggest that future PRS improvements or genetic dimensionality reduction techniques may be needed to further improve the multimodal prediction of complex traits. In constructing our PRSs, we based them on all genetic variants found in primary genome-wide association studies (Table S3 in Supplement 2 lists all studies), such as the one conducted for depression (40). This approach was taken because restricting our analysis to significantly associated variants would considerably diminish predictive power, as supported by recent studies (41,42). The model that significantly outperformed the one based on clinical data was the model that was informed by both clinical and proteomic data, with only the XGBoost model having a large enough effect to have adequate power within the test sample, while proteomic data were most informative for unimodal data predictions. Furthermore, our study suggests that while proteomic data had a larger feature space than clinical and PRS data, their superiority in predicting 2-year MDD remission status cannot be explained by dimensionality alone, and it may be the most informative omics data modality when combined with clinical data.

Prediction and Models' Performance

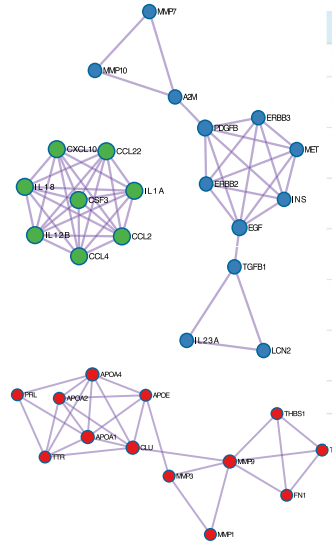
Applying machine learning models entails predicting at the individual-subject level (n = 1 prediction), which may ultimately pave the way for individual clinical application, i.e., enable personalized psychiatry (43). For personalized psychiatry, accuracy and other prediction performance metrics are arguably more valuable than traditional statistical measures because they 1) indicate how well a model works on the individual-subject level and 2) are the result of the model being put to practice in separate new individuals not previously seen by the model.

One might regard a balanced accuracy of 71% (i.e., our best-performing model) as too low for use in clinical practice. However, consistent with previous findings (44), we showed that the next best thing for patients—namely interpretive predictions by clinicians—performed substantially worse. Moreover, we showed that clinicians' predictions showed high inter-rater variability resulting in low inter-rater reliability (Fleiss' kappa = 0.32) (45). One of the clinicians was part of the research design, and so to mitigate incentive bias, we used a blinded rater design, meaning that the raters had no information on the models' performance until after they had made their predictions. When excluding the involved clinician from the



Combined Proteomics + Clinical Model

Color	MCODE	GO	Description	Log10(P)
	MCODE_1	R-HSA-6783783	Interleukin-10 signaling	-5.0
	MCODE_1	GO:0009615	Response to virus	-3.8
	MCODE_1	WP3929	Chemokine signaling pathway	-3.8
	MCODE_2	GO:0010873	Positive regulation of cholesterol esterification	-6.9
	MCODE_2	GO:0010872	Regulation of cholesterol esterification	-6.9
	MCODE_2	GO:0033700	Phospholipid efflux	-6.4



Proteomics-Only Model

Color	MCODE	GO	Description	Log10(P)
	MCODE_1	GO:0043691	Reverse cholesterol transport	-6.1
	MCODE_1	R-HSA-975634	Retinoid metabolism and transport	-5.6
	MCODE_1	R-HSA-9709957	Sensory perception	-5.6
	MCODE_2	GO:0032885	Regulation of polysaccharide biosynthetic process	-5.6
	MCODE_2	GO:0032881	Regulation of polysaccharide metabolic process	-5.6
	MCODE_2	R-HSA-6811558	PI5P, PP2A and IER3 regulate PI3K/AKT signaling	-4.8
	MCODE_3	R-HSA-6783783	Interleukin-10 signaling	-8.2
	MCODE_3	GO:0071216	Cellular response to biotic stimulus	-5.2
	MCODE_3	GO:0071222	Cellular response to lipopolysaccharide	-5.2

Figure 4. **(A)** Densely connected protein-protein subnetworks and subsequent enrichment analysis results for the informative proteomic analytes in the combined clinical and proteomic data model. **(B)** Similar to **(A)**, but results of the informative analytes of the proteomics-only model are shown. GO, gene ontology database; WP, WikiPathways database; R-HSA, Reactome pathway database.

analysis (rater 1 in Figure 5), the mean accuracy of the remaining 3 human raters was 0.57 when using 10 clinical variables and 0.53 when having access to the complementary clinical data. One can argue that the retrospective data shown to the clinicians in our study did not approximate a live clinical impression. However, previous studies have shown that this added source of information for predictions results in even worse predictions by clinicians (46,47). This does not mean that live clinical impressions are uninformative for future predictions per se. In the light of multimodal prediction, live clinical impressions may yet prove to hold complementary information for augmented predictions. Future studies will have to clarify for what type of outcome predictions (e.g., therapy response, remission), and in combination with what type of data, clinical impressions add a complementary layer of predictive information.

A much-needed next step for personalized psychiatry is to implement and test machine learning–based clinical decision making in clinical trials. Clinical implementation of machine learning models has shown promise in preliminary studies (48,49). Importantly, recent clinical trials successfully showed the superiority of machine learning–based clinical decision making compared with conventional clinical decision making for medical fields outside of psychiatry (50,51). To facilitate the future prospects of personalized, machine learning–aided

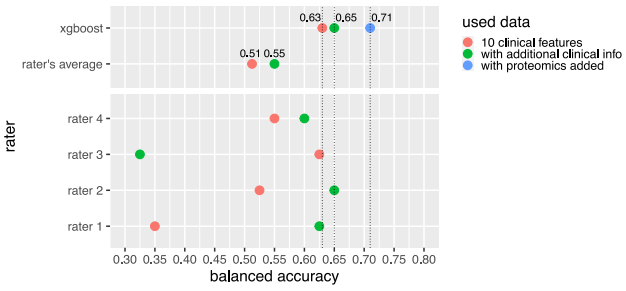


Figure 5. Results of clinical psychiatrists' predictions of 2-year major depressive disorder chronicity compared with the XGBoost model's performance. The x-axis shows the balanced accuracy of predictions. The dots represent a psychiatrist's or model's predictive performance, with the color indicating on what basis information predictions were made.

decision making in psychiatry, we view our study as an important clue regarding the type of data that can be informative for prediction models put into practice specifically with the aim of predicting an individual's naturalistic course of MDD from baseline data. Such predictions may ultimately benefit patient outcomes by providing clinically actionable information. For example, in cases predicted to show nonremission after 2 years, intensifying therapy early on may improve disease course.

Strengths and Limitations

We were able to train, (cross-)validate, and test our models on subject data that were collected as part of the longitudinal NESDA study. Likewise, human prediction evaluation was solely based on data of subjects included in the NESDA database. Unfortunately, without any validation of our model on data external to the NESDA dataset, robust assessment of the generalizability of our model's performance is currently lacking. However, given that we carefully prevented any data leakage from final test set to the train set in all our imputation, preprocessing, and feature selection procedures (i.e., prevented double dipping) (23,52,53); used balanced train and test sets (54); used separate repeated 10-fold cross-validation procedures in the train sample independent of the final test set (23,24); used a sample size for prediction analysis of several hundred (23); and tested final model performance on an outheld test sample (23,53,55), we believe that, at the least, the multimodal features that were found to be most predictive represent robust findings. Therefore, we argue that the importance of this work lies primarily in the observations that 1) there are blood-based variables that can be individually predictive for the naturalistic 2-year course of MDD; 2) individual predictions become more accurate when using multimodal data; and 3) high-dimensional predictive signatures may not be detected using conventional linear machine learning models. Secondarily, we found that the best-performing multimodal predictive signature comprised baseline symptom severity, personality traits, and blood biomarkers related to immune system and lipid metabolism. These findings can aid variable inclusion decisions in future MDD remission prediction studies. However, due to the lack of precise disease trajectory data before and after the 2-year interval and the noncausal nature of SHAP values, the predictive signature's causal interpretation remains unclear. Additionally, the NESDA sample does not offer information on biomarker status over time, thereby limiting the evaluation of the signature's stability for predicting remission at other time intervals and extrapolative value beyond the 2-year interval.

Age, sex, body mass index, years of education, center of patient recruitment, antidepressant use at baseline, and other psychopharmaceuticals used at baseline were not significantly different between the 2 outcome groups (Table 1). Furthermore, our post hoc confounder analysis showed that the performance of the multimodal predictions could not be explained by any of these confounding factors (Figure S4 in Supplement 1) and was independent from any possible interaction with these confounding factors. These results suggest some biological validity of the multimodal signature independent of any of the included confounders.

We chose 2-year remission status instead of a longer period of enduring depression because 1) persistent depressive disorder is defined as depression lasting for at least 2 years in DSM-5 ($F_{34.1}$ = 300.4) (56), and 2) sufficient sample size was available only for the 2-year time period (although still suboptimal for the model informed by all modalities). Despite using stable 6-month outcomes, our predicted outcome label hinges on a somewhat arbitrary cutoff concerning both time span and diagnostic criteria, which may conflate chronic phenotypes and delayed remission cases under a single designation. Likewise, the remitted label does not represent lifetime remission because individuals who show remission after 2 years can eventually relapse into depressive episodes. This may account for the moderate prediction performance found with unimodal omics data and suboptimal results using multimodal data. To what extent the described predictive signature generalizes to longer remission rates remains to be seen and may not hold given that an approximated 30% to 85% of remissions in depression are followed up by a relapse at some point (57). Indeed, analyzing the NESDA subsample of our data with available outcome data at a 2-year, 4-year, and 6-year interval (n = 503), we found that 4-year consistency of remission status was reported for 67% of the individuals, and 6-year consistency of remission status was reported for 49% of the individuals. Notwithstanding this, accurate prediction of remission status at a 2-year interval—even if inconsistent with remission status at other points in time—provides clinically relevant information (although only for the 2-year period).

Similarly, comorbid anxiety disorders may obscure true remission rates. A previous study showed that recovery rate is lower when including both depressive and anxiety disorder symptoms (58). Looking at our own sample, we found that the 2-year remitted group showed a significantly lower prevalence of anxiety disorders at the 2-year time point (27% vs. 53%, $p < 2 \times 10^{-4}$), while no such difference was found at baseline (Table S6 in Supplement 2). Beck Anxiety Inventory scores (59) were also significantly lower for the remitted group at the 2-year time point (14.6 vs. 20.41, $p < 2 \times 10^{-4}$). Interestingly, at baseline, neither group showed any significant difference (Table S6 in Supplement 2). Furthermore, the model including anxiety scores and disorder information as clinical data did not improve but instead slightly worsened predictions (AUROC = 0.63 vs. AUROC = 0.61) (see Results). This suggests that anxiety disorder–related information at baseline added no information beyond the 10 clinical variables for 2-year depression remission predictions despite the nonremission group showing higher average Beck Anxiety Inventory scores at baseline (20.09 vs. 17.58, $p = 9 \times 10^{-4}$) (Table S6 in Supplement 2). It is questionable to what extent the predictive signature in our study can be expected to be depression specific. The possibility of overlapping diagnoses and a predictive pattern that may indicate combined remission rather than depression-specific remission could contribute to the suboptimal prediction performance.

Conclusions

To our knowledge, this is the first study to show that the combination of multimodal biological and clinical data significantly improves the accuracy of individual longitudinal

predictions of remission status in MDD in a relatively large sample (N = 804). Moreover, this study shows that what is predictive of remission of MDD within 2 years is a combined signature of symptom severity, personality traits, and immune- and lipid metabolism–related proteins at baseline. We argue that future studies that investigate the potential of clinical application of MDD course prediction models are much needed and should consider including both clinical and proteomic data focused on immune and lipid metabolism markers in their data.

ACKNOWLEDGMENTS AND DISCLOSURES

This work was supported by the Geestkracht program of the Netherlands Organization for Health Research and Development (Grant No. 10-000-1002 [to BWJHP]) and is also supported by participating universities and mental health care organizations (Vrije Universiteit University Medical Center, GGZ inGeest, Arkin, Leiden University Medical Center, GGZ Rivierduinen, University Medical Center Groningen, Lentis, GGZ Friesland, GGZ Drenthe, Institute for Quality of Health Care, Netherlands Institute for Health Services Research, and Netherlands Institute of Mental Health and Addiction). The collaboration project is cofunded by the public-private partnerships Allowance made available by Health ~ Holland (Topsector Life Sciences & Health) to stimulate public-private partnerships (to CHV).

We thank Animal Genetics for their partnership and funding contribution. We thank Dr. E. Bosdriesz and Dr. F. Bennis for their valuable input on the machine learning analysis and Dr. J. Tijdink and Dr. J. Luykx for their aid in complementing human predictions.

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

ARTICLE INFORMATION

From the Department of Anatomy & Neurosciences, Amsterdam University Medical Center, Vrije Universiteit, Amsterdam, the Netherlands (PCH, RMT, YM, RJ, GAvW, CHV); Department of Psychiatry, Amsterdam Neuroscience, Amsterdam University Medical Center, Vrije Universiteit, Amsterdam, the Netherlands (PCH, YM, RJ, WJP, BWJHP, CHV); Department of Internal Medicine, section Endocrinology, Leiden University Medical Center, Leiden, the Netherlands (PCH, OCM); Department of Biological Psychology, Vrije Universiteit Amsterdam, Neuroscience Campus Amsterdam, Amsterdam, the Netherlands (RP); and Department of Complex Traits Genetics, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, Vrije Universiteit, Amsterdam, the Netherlands (WJP).

Address correspondence to Philippe C. Habets, M.D., Ph.D., at p.c.habets@amsterdamumc.nl.

Received Oct 4, 2022; revised May 11, 2023; accepted May 30, 2023.

Supplementary material cited in this article is available online at https://doi.org/10.1016/j.biopsych.2023.05.024.

REFERENCES

1. Penninx BWJH, Nolen WA, Lamers F, Zitman FG, Smit JH, Spinhoven P, *et al.* (2011): Two-year course of depressive and anxiety disorders: Results from the Netherlands Study of Depression and Anxiety (NESDA). J Affect Disord 133:76–85.
2. Pettit JW, Lewinsohn PM, Roberts RE, Seeley JR, Monteith L (2009): The long-term course of depression: Development of an empirical index and identification of early adult outcomes. Psychol Med 39:403–412.
3. Wiersma JE, van Oppen P, van Schaik DJF, van der Does AJW, Beekman ATF, Penninx BWJH (2011): Psychological characteristics of chronic depression: A longitudinal cohort study. J Clin Psychiatry 72:288–294.
4. Lamers F, Vogelzangs N, Merikangas KR, de Jonge P, Beekman ATF, Penninx BWJH (2013): Evidence for a differential role of HPA-axis function, inflammation and metabolic syndrome in melancholic versus atypical depression. Mol Psychiatry 18:692–699.
5. Milaneschi Y, Hoogendijk W, Lips P, Heijboer AC, Schoevers R, van Hemert AM, *et al.* (2014): The association between low vitamin D and depressive disorders. Mol Psychiatry 19:444–451.
6. Vogelzangs N, Beekman ATF, Boelhouwer IG, Bandinelli S, Milaneschi Y, Ferrucci L, Penninx BWJH (2011): Metabolic depression: A chronic depressive subtype? Findings from the InCHIANTI study of older persons. J Clin Psychiatry 72:598–604.
7. Vreeburg SA, Hoogendijk WJG, DeRijk RH, van Dyck R, Smit JH, Zitman FG, Penninx BWJH (2013): Salivary cortisol levels and the 2-year course of depressive and anxiety disorders. Psychoneuroendocrinology 38:1494–1502.
8. Qi B, Fiori LM, Turecki G, Trakadis YJ (2020): Machine learning analysis of blood microRNA data in major depression: A case-control study for biomarker discovery. Int J Neuropsychopharmacol 23:505–510.
9. Han SYS, Tomasik J, Rustogi N, Lago SG, Barton-Owen G, Eljasz P, *et al.* (2020): Diagnostic prediction model development using data from dried blood spot proteomics and a digital mental health assessment to identify major depressive disorder among individuals presenting with low mood. Brain Behav Immun 90:184–195.
10. Bhak Y, Jeong HO, Cho YS, Jeon S, Cho J, Gim JA, *et al.* (2019): Depression and suicide risk prediction models using blood-derived multi-omics data. Transl Psychiatry 9:262.
11. Sajjadian M, Lam RW, Milev R, Rotzinger S, Frey BN, Soares CN, *et al.* (2021): Machine learning in the prediction of depression treatment outcomes: A systematic review and meta-analysis. Psychol Med 51:2742–2751.
12. Milaneschi Y, Lamers F, Peyrot WJ, Abdellaoui A, Willemsen G, Hottenga JJ, *et al.* (2016): Polygenic dissection of major depression clinical heterogeneity. Mol Psychiatry 21:516–522.
13. Park SC, Kim JM, Jun TY, Lee MS, Kim JB, Yim HW, Park YC (2017): How many different symptom combinations fulfil the diagnostic criteria for major depressive disorder? Results from the CRESCEND study. Nord J Psychiatry 71:217–222.
14. Milaneschi Y, Lamers F, Penninx BWJH (2021): Dissecting depression biological and clinical heterogeneity-The importance of symptom assessment resolution. JAMA Psychiatry 78:341–341.
15. Jermy BS, Glanville KP, Coleman JRI, Lewis CM, Vassos E (2021): Exploring the genetic heterogeneity in major depression across diagnostic criteria. Mol Psychiatry 26:7337–7345.
16. Mocking RJT, Naviaux JC, Li K, Wang L, Monk JM, Bright AT, *et al.* (2021): Metabolic features of recurrent major depressive disorder in remission, and the risk of future recurrence. Transl Psychiatry 11:37.
17. Rubenstein LV, Rayburn NR, Keeler EB, Ford DE, Rost KM, Sherbourne CD (2007): Predicting outcomes of primary care patients with major depression: Development of a depression prognosis index. Psychiatr Serv 58:1049–1056.
18. Klein NS, Holtman GA, Bockting CLH, Heymans MW, Burger H (2018): Development and validation of a clinical prediction tool to estimate the individual risk of depressive relapse or recurrence in individuals with recurrent depression. J Psychiatr Res 104:1–7.
19. Koutsouleris N, Kambeitz-Ilankovic L, Ruhrmann S, Rosen M, Ruef A, Dwyer DB, *et al.* (2018): Prediction models of functional outcomes for individuals in the clinical high-risk state for psychosis or with recent-onset depression: A multimodal, multisite machine learning analysis. JAMA Psychiatry 75:1156–1172.
20. Clark SL, Hattab MW, Chan RF, Shabalin AA, Han LKM, Zhao M, *et al.* (2020): A methylation study of long-term depression risk. Mol Psychiatry 25:1334–1343.
21. Penninx BWJH, Beekman ATF, Smit JH, Zitman FG, Nolen WA, Spinhoven P, *et al.* (2008): The Netherlands Study of Depression and Anxiety (NESDA): Rationale, objectives and methods. Int J Methods Psychiatr Res 17:121–140.
22. Dinga R, Marquand AF, Veltman DJ, Beekman ATF, Schoevers RA, van Hemert AM, *et al.* (2018): Predicting the naturalistic course of depression from a wide range of clinical, psychological, and biological data: A machine learning approach. Transl Psychiatry 8:241.
23. Poldrack RA, Huckins G, Varoquaux G (2020): Establishment of best practices for evidence for prediction: A Review. JAMA Psychiatry 77:534–540.

24. Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, *et al.* (2012): Reporting and methods in clinical prediction research: A systematic review. PLoS Med 9:1–12.
25. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ (2022): Multimodal biomedical AI. Nat Med 28:1773–1784.
26. Boehm KM, Aherne EA, Ellenson L, Nikolovski I, Alghamdi M, Vázquez-García I, *et al.* (2022): Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. Nat Cancer 3:723–733.
27. Boehm KM, Khosravi P, Vanguri R, Gao J, Shah SP (2022): Harnessing multimodal data integration to advance precision oncology. Nat Rev Cancer 22:114–126.
28. Robins LN, Wing J, Wittchen HU, Helzer JE, Babor TF, Burke J, *et al.* (1988): The Composite International Diagnostic Interview. An epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. Arch Gen Psychiatry 45:1069–1077.
29. Lyketsos CG, Nestadt G, Cwi J, Heithoff K, *et al.* (1994): The Life Chart Interview: A standardized method to describe the course of psychopathology. Int J Methods Psychiatr Res 4:143–155.
30. Rush AJ, Giles DE, Schlesser MA, Fulton CL, Weissenburger J, Burns C (1986): The inventory for depressive symptomatology (IDS): Preliminary findings. Psychiatry Res 18:65–87.
31. Costa PT, McCrae RR (1995): Domains and facets: Hierarchical personality assessment using the revised NEO personality inventory. J Pers Assess 64:21–50.
32. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M (2023): KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res 51:D587–D592.
33. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, *et al.* (2015): Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am J Hum Genet 97:576–592.
34. Chen T, Guestrin C (2016): Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. New York, NY: Association for Computing Machinery, 785–794.
35. Lundberg SM, Lee S-I (2017): A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30:4765–4774.
36. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, *et al.* (2020): From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2:56–67.
37. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, *et al.* (2019): Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun 10:1523.
38. Dinga R, Schmaal L, Penninx BWJH, Veltman DJ, Marquand AF (2020): Controlling for effects of confounding variables on machine learning predictions. biorXiv https://doi.org/10.1101/2020.08.17.255034.
39. Winter NR, Goltermann J, Dannlowski U, Hahn T (2021): Interpreting weights of multimodal machine learning models-Problems and pitfalls. Neuropsychopharmacology 46:1861–1862.
40. Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, *et al.* (2018): Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat Genet 50:668–681.
41. Howard DM, Adams MJ, Clarke TK, Hafferty JD, Gibson J, Shirali M, *et al.* (2019): Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat Neurosci 22:343–352.
42. Trubetskoy V, Pardiñas AF, Qi T, Panagiotaropoulou G, Awasthi S, Bigdeli TB, *et al.* (2022): Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature 604:502–508.
43. Koppe G, Meyer-Lindenberg A, Durstewitz D (2021): Deep learning for small and big data in psychiatry. Neuropsychopharmacology 46:176–190.
44. Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C (2000): Clinical versus mechanical prediction: A meta-analysis. Psychol Assess 12:19–30.
45. McHugh ML (2012): Interrater reliability: The kappa statistic. Biochem Med 22:276–282.
46. Bell I, Mellor D (2009): Clinical judgements: Research and practice. Aust Psychol 44:112–121.
47. Smith M, Francq B, McConnachie A, Wetherall K, Pelosi A, Morrison J (2020): Clinical judgement, case complexity and symptom scores as predictors of outcome in depression: An exploratory analysis. BMC Psychiatry 20:125.
48. Wu CS, Yang AC, Chang SS, Chang CM, Liu YH, Liao SC, Tsai HJ (2021): Validation of machine learning-based individualized treatment for depressive disorder using target trial emulation. J Pers Med 11:1316.
49. Rajpurkar P, Yang J, Dass N, Vale V, Keller AS, Irvin J, *et al.* (2020): Evaluation of a machine learning model based on pretreatment symptoms and electroencephalographic features to predict outcomes of antidepressant treatment in adults with depression: A Prespecified Secondary Analysis of a Randomized Clinical Trial [published correction appears in JAMA Netw Open 2020; 3:e2016001]. JAMA Netw Open 3:e206653.
50. Adams R, Henry KE, Sridharan A, Soleimani H, Zhan A, Rawat N, *et al.* (2022): Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med 28:1455–1460.
51. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R (2017): Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: A randomised clinical trial. BMJ Open Respir Res 4:e000234.
52. Shim M, Lee SH, Hwang HJ (2021): Inflated prediction accuracy of neuropsychiatric biomarkers caused by data leakage in feature selection. Sci Rep 11:7980.
53. Whelan R, Garavan H (2014): When optimism hurts: Inflated predictions in psychiatric neuroimaging. Biol Psychiatry 75:746–748.
54. Vandewiele G, Dehaene I, Kovács G, Sterckx L, Janssens O, Ongenae F, *et al.* (2021): Overly optimistic prediction results on imbalanced data: A case study of flaws and benefits when applying over-sampling. Artif Intell Med 111:101987.
55. Yeung AWK, More S, Wu J, Eickhoff SB (2022): Reporting details of neuroimaging studies on individual traits prediction: A literature survey. Neuroimage 256:119275.
56. American Psychiatric Association, DSM-5 Task Force (2013): Diagnostic and Statistical Manual of Mental Disorders, 5th ed. Washington, DC: American Psychiatric Publishing.
57. Touya M, Lawrence DF, Kangethe A, Chrones L, Evangelatos T, Polson M (2022): Incremental burden of relapse in patients with major depressive disorder: A real-world, retrospective cohort study using claims data. BMC Psychiatry 22:152.
58. Verduijn J, Verhoeven JE, Milaneschi Y, Schoevers RA, van Hemert AM, Beekman ATF, Penninx BWJH (2017): Reconsidering the prognosis of major depressive disorder across diagnostic boundaries: Full recovery is the exception rather than the rule. BMC Med 15:215.
59. Beck AT, Epstein N, Brown G, Steer RA (1988): An inventory for measuring clinical anxiety: Psychometric properties. J Consult Clin Psychol 56:893–897.