Received: 14 January 2020 | Accepted: 12 October 2020

DOI: 10.1111/acps.13250

Acta Psychiatrica Scandinavica WILEY

ORIGINAL ARTICLE

Combining machine learning algorithms for prediction of antidepressant treatment response

Alexander Kautzky[1] | **Hans-Juergen Möller**[2] | **Markus Dold**[1] | **Lucie Bartova**[1] | **Florian Seemüller**[2,3] | **Gerd Laux**[4] | **Michael Riedel**[2,5] | **Wolfgang Gaebel**[6] | **Siegfried Kasper**[1]

[1]Department of Psychiatry and Psychotherapy, Medical University of Vienna, Vienna, Austria

[2]Department of Psychiatry and Psychotherapy, Ludwig-Maximilians-Q3 University Munich, Munich, Germany

[3]Department of Psychiatry and Psychotherapy, kbo-Lech-Mangfall-Klinik, Garmisch-Partenkirchen, Germany

[4]Department of Psychiatry and Psychotherapy, kbo-Inn-Salzach-Klinikum, Wasserburg, Germany

[5]Department of Psychiatry, Sächsisches Krankenhaus, Rodewisch, Germany

[6]Department of Psychiatry and Psychotherapy, Medical Faculty, Heinrich-Heine-University, Düsseldorf, Germany

Correspondence
Siegfried Kasper, Department of Psychiatry and Psychotherapy, Medical University of Vienna, Währinger Gürtel 18-20, A-1090 Vienna, Austria.
Email: gen-psychiatry@meduniwien.ac.at

Funding information
The study was performed within the framework of the German Research Network on Depression, which was funded by the German Federal Ministry for Education and Research BMBF (01GI0219). The BMBF had no further role in study design; in the collection, analysis, and interpretation of data and in the writing of the report. Dr. Kasper received grants/research support, consulting fees, and/or honoraria within the last three years from Angelini, AOP Orphan Pharmaceuticals AG, AstraZeneca, Eli Lilly, Janssen, KRKA-Pharma, Lundbeck, Neuraxpharm, Pfizer, Pierre Fabre, Schwabe, and Servier. Dr. Möller received consulting fees and /or honoraria from Otsuka, Schwabe, and Servier in the last three years.

Abstract

Objectives: Predictors for unfavorable treatment outcome in major depressive disorder (MDD) applicable for treatment selection are still lacking. The database of a longitudinal multicenter study on 1079 acutely depressed patients, performed by the German research network on depression (GRND), allows supervised and unsupervised learning to further elucidate the interplay of clinical and psycho-sociodemographic variables and their predictive impact on treatment outcome phenotypes.

Experimental Procedures: Treatment response was defined by a change of HAM-D 17-item baseline score ≥50% and remission by the established threshold of ≤7, respectively, after up to eight weeks of inpatient treatment. After hierarchical symptom clustering and stratification by treatment subtypes (serotonin reuptake inhibitors, tricyclic antidepressants, antipsychotic, and lithium augmentation), prediction models for different outcome phenotypes were computed with random forest in a cross-center validation design. In total, 88 predictors were implemented.

Results: Clustering revealed four distinct HAM-D subscores related to emotional, anxious, sleep, and appetite symptoms, respectively. After feature selection, classification models reached moderate to high accuracies up to 0.85. Highest accuracies were observed for the SSRI and TCA subgroups and for sleep and appetite symptoms, while anxious symptoms showed poor predictability.

Conclusion: Our results support a decisive role for machine learning in the management of antidepressant treatment. Treatment- and symptom-specific algorithms may increase accuracies by reducing heterogeneity. Especially, predictors related to duration of illness, baseline depression severity, anxiety and somatic symptoms, and

personality traits moderate treatment success. However, prospectives application of machine learning models will be necessary to prove their value for the clinic.

KEYWORDS
affective disorders, antidepressives, classification

1 | INTRODUCTION

Major depressive disorder (MDD) has steadily ranked up among the most burdensome diseases worldwide, reaching an estimated life time prevalence of about a fifth of the global population.[1] Despite decades of research, MDD is still a disease that is as common as challenging for clinicians. Considering high rates of non-response to standard antidepressive treatment of up to 50%, the outlook for patients at the initiation of treatment is alarmingly unsatisfactory.[2] While antidepressive treatment options are clearly effective,[3] the path to symptom remission is almost always time-consuming and often impeded by several unsuccessful trials. Even optimistic estimations assume resistance rates to continuous treatment with multiple trials of 15%.[4] Despite successful efforts to determine predictors of treatment response and resistance, even well-established markers such as baseline symptom severity or comorbid psychiatric disorders up to now did not impact treatment in the clinical setting.[5] While several guidelines highlight red flags such as side effects specific for a drug that are unfavorable for an individual patient or pharmacogenetic considerations,[6] treatment is characterized mostly by trial and error. While guidelines give some support for treatment optimization,[7,8] there is still no rationale for personalized treatment of MDD that provides symptom-oriented guidelines for the first antidepressant to prescribe or optimal augmentation.

With the rise and increased availability of large databases, multivariate models incorporating clinical and sociodemographic data were introduced to neuropsychiatric research only in recent years.[9] Concerning MDD, especially the American Sequenced Treatment Alternatives to Relieve Depression (STAR*D) database and European counterparts as the German research network on depression (GRND) or Group for Studies of Resistant Depression (GSRD), enabled progress in predicting treatment outcome on the individual patient level.[10-12] In the context of the GRND study, a logistic regression prediction model based on a set of predictors with univariate association with remission or response was presented earlier,[11] and similarly effective models based on other large naturalistic databases have been proposed.[10,13] Nevertheless, marginal progress was achieved on differential predictors for effectiveness of specific antidepressant categories.[14,15]

Significant outcomes

- Exploiting a large naturalistic database on treatment outcome in MDD, we detected data-driven symptom clusters with distinct patterns and predictors of response to antidepressant agents.
- Symptom severity, longer duration of illness reflected by number of episodes and hospitalizations and overall time living with depression, anxious, and somatic symptoms, high neuroticism, and low extraversion predicted the risk for disadvantageous treatment outcome over different classification models.
- Selections of clinical, sociodemographic, and personality variables enabled moderate to good classification accuracy up to 85% for treatment outcomes in a quasi-independent cross-center validation design. Specific predictor sets emerged for symptom clusters as well as treatment subtypes, and stratification generally increased model performance by reducing heterogeneity.

Limitations

- This is a naturalistic study, and patients differed considerably in treatment algorithms, number of episodes and previous treatments; thus, models for individual antidepressant agents could not be implemented.
- Despite a cross-trial validation design, in the absence of a fully independent test sample we cannot rule out overfitting and dependency of the machine learning results on the data context, nor prove generalizability.
- Because of low observation counts for treatment non-response and non-remission and divergent ratios in the different cross-center folds, oversampling of the minority class was applied, which may bring bias to our results.

1.1 | Aims of the study

Consequently, the aim of this study was to generate multivariate prediction models for treatment outcome specific for the common antidepressant drug entities serotonin reuptake inhibitors (SSRI), tricyclic antidepressants (TCA), antipsychotic (AP), and lithium augmentation. Furthermore, exploiting a combination of supervised learning for prediction of treatment outcome and unsupervised learning for definition of data-driven response phenotypes, subtypes of depressive symptoms were compared to the conventional HAM-D 17 item severity scores.

2 | METHODS

2.1 | Sample

All patents derive from the GRND study, a joint effort by twelve study centers across Germany (seven university hospitals, five district hospitals), funded by the German Federal Ministry of Education and Research (BMBF). Details on the study design and scope can be found in previous publications.[16] In short, the GRND was a large naturalistic study that aimed at longitudinal characterization of depressed patients and antidepressant treatment outcome in German psychiatric university and district hospitals. In total, 1079 patients with a depressive disorder diagnosed with the help of a Structured Clinical Interview for DSM-IV (SCID-I) were enrolled. 1014 showed longitudinal data availability for the inpatient treatment period and were eligible for further analysis. A description of the baseline characteristics of the total sample can be found in Ref.11 Missing data cannot be handled by machine learning techniques and imputation of clinical variables bears significant bias. Consequently, only patients with full data availability for all baseline variables were considered for this analysis ($n = 504$). Please also see Table 1 for an overview of baseline characteristics and refer to the supplements for details (Table S1, S2, Figure S1).

2.2 | Outcome phenotypes

The GRND registered a 17-item HAM-D score every other week until discharge from the hospital, enabling analysis of a broad spectrum of outcome phenotypes. Previous analyses of the GRND sample focused especially on early response and response and remission at discharge.[11]

For this analysis, response and remission after up to eight weeks of inpatient treatment were analyzed. Treatment response was defined by a HAM-D change equal to or greater than 50%. Remission was defined by reaching a HAM-D score of 7 or below.[17] Response and remission were compared to non-response or non-remission, defined by a failure to achieve a favorable treatment outcome after eight weeks of inpatient treatment.

The time points were chosen according to the estimated time of four weeks for adequacy for any antidepressant trial, indicating non-response to one trial at week four and a consecutive trial at week eight. Consequently, non-response after 8 weeks of continuous treatment reflects a state comparable to but less stringent than treatment resistance according to the European staging system.[18] Considering the naturalistic nature of the study, trials varied considerably between patients.

2.3 | Predictors

In addition to sex and age, 86 predictors grouped into five sets by modality, (i) baseline severity, (ii) clinical and sociodemographic variables, (iii) psychopathology and somatic symptomatology, (iv) psychiatric comorbidities, and (v) personality, were included in the analyses.

1. Depression severity was assessed by Montgomery-Asberg Rating Scale (MADRS,[19]) and the 17- and 21-item HAM-D (coded as numerical, total scores, and items; overall severity as binomial, severe vs moderate; 39 predictors).[20]
2. Sociodemographic and clinical predictors (relationship status, education, job qualification and current occupation, family history of psychiatric disorders, early life stress before the 6th and 15th year of age, number of previous hospitalizations, duration of episode, age of onset and duration of illness, presence or absence of suicidality, moderate vs severe depression, recurrent vs first episode) were assessed via the basic documentation (BADO), a systematic basic assessment of clinical and sociodemographic variables in psychiatry.[21] In this predictor group, also baseline scores of the Global Assessment of Functioning Scale (GAF) and Social and Occupational Functioning Scale (SOFAS) were implemented,[22,23] adding up to a total of 15 predictors.
3. Extensive psychopathology and somatic symptomatology were assessed with the scale of the Association for Methodology and Documentation in Psychiatry (AMDP; all numerical; 19 predictors based on AMDP subcategories).[24]
4. Psychiatric comorbidities were assessed with the SCID-I (presence or absence of eating, somatizing, anxiety and substance use disorder, PTSD, OCD, dysthymia), while axis II personality disorders (presence or absence) were determined with the SCID II (binomial; 8 predictors).

TABLE 1 Clinical characteristics and psychiatric comorbidities grouped by remission and response after up to 8 weeks of treatment. t Tests and Fisher tests were performed for continuous and categorical variables, respectively, and *p*-values are reported

Clinical characteristics	Response $n = 340$, 67.5%	Non-Response $n = 164$, 32.5%	Remission $n = 234$, 46.4%	Non-Remission $n = 270$, 53.6%	*p*-value Resp./Rem.
Age of onset					
Mean ± SD	39.09 ± 12.04	36.45 ± 12.54	39.36 ± 12.27	37.26 ± 12.18	n.s.
HAM-D 17 baseline					
Mean ± SD	22.97 ± 5.04	22.02 ± 4.80	22.48 ± 4.99	22.82 ± 4.97	n.s.
Recurrent depression					
Single	103 (30%)	35 (21%)	83 (35%)	55 (20%)	0.042/0.0002
Recurrent	237 (70%)	129 (79%)	151 (65%)	215 (80%)	
Duration MDD (in years)					
Mean ± SD	5.77 ± 8.76	7.59 ± 8.85	5.09 ± 8.57	7.47 ± 8.90	0.029/0.002
Duration episode					
<1 m	14 (4%)	16 (10%)	39 (17%)	28 (10%)	0.002/0.006
1–3 m	51 (15%)	37 (23%)	76 (32%)	72 (27%)	
3–6 m	111 (33%)	38 (23%)	61 (26%)	63 (23%)	
6 m–2 y	86 (25%)	58 (36%)	49 (21%)	87 (32%)	
>2 y	78 (23%)	15 (8%)	9 (4%)	20 (8%)	
Dysthymia					
Present	13 (4%)	17 (10%)	8 (3%)	22 (8%)	0.007/0.036
Absent	327 (96%)	147 (90%)	226 (97%)	248 (92%)	
Anxiety (GAD $n = 2$, PD $n = 32$, SP $n = 12$, AP $n = 20$, Specific Phobia $n = 6$)					
Present	28 (8%)	23 (14%)	16 (7%)	35 (13%)	0.057/0.026
Absent	312 (92%)	141 (86%)	218 (93%)	235 (87%)	
Personality disorder					
Present	44 (13%)	27 (16%)	27 (12%)	44 (16%)	n.s.
Absent	296 (87%)	137 (84%)	207 (88%)	226 (84%)	
Substance disorder					
Present	35 (10%)	14 (9%)	25 (11%)	24 (9%)	n.s.
Absent	305 (90%)	150 (91%)	209 (89%)	246 (91%)	
Suicidality					
Present	175 (51%)	85 (52%)	119 (51%)	141 (52%)	n.s.
Absent	165 (49%)	79 (48%)	115 (49%)	129 (48%)	
Sex					
Female	216 (62%)	110 (67%)	145 (62%)	176 (65%)	n.s.
Male	129 (38%)	54 (33%)	89 (38%)	94 (35%)	

MDD, Major Depressive Disorder; GAD, generalized anxiety disorder; PD, panic disorder; SD, social phobia; AP, agoraphobia.

5. Finally, personality traits extroversion, neuroticism, tolerance, conscientiousness, and openness as defined by the five-factor inventory (NEO-FFI) were included (numerical; 5 predictors).[25]

A complete list of the 88 predictors can be found in the supplements. Considering that the AMDP may not be familiar to most clinicians, the version used by the GRND study group can be found in the supplements.

2.4 | Unsupervised learning

Unsupervised learning allows to detect clusters or subgroups of data by a machine learning algorithm that has no prior knowledge of potential outcomes of interest. Hence, in this study the HAM-D items were used to define alternative outcome scores to the conventional total HAM-D score, solely based on observed patterns in the patients' data and unbiased from hypotheses of the analysts.

Thus, in order to improve the prediction performance, data-driven subtypes of response were computed from the 17 HAM-D items at baseline. Here, all patients with a fully documented baseline HAM-D were used for analysis ($n = 1079$). A hierarchical clustering solution was applied to detect co-occurring symptoms via the package "ClusOfVar" for the statistical software "R".[26] Hierarchical clustering is a distance-based algorithm suitable for categorical or ordinal variables with graphical determination of the number of clusters. The dendrogram branches are cut with maximum distance between horizontal lines, resulting in the most unsimilar clusters. In other words, the algorithm aims at defining groups that are as different from each other as possible. However, the optimal solution can also deviate from this rule and reflect considerations of the analysts based on the data type, structure, and context. Alternatively, an automated selection of the cluster number can be performed based on the Rand criterion.

However, using unsupervised learning in a specific database may produce results that cannot be generalized to other samples. In order to guarantee independence of the observed clusters from the data context of the GRND, a similar analysis was performed in the data pool of the European research consortium GSRD. The cross-sectional GSRD data pool comprises 1566 MDD patients suitable for clustering of HAM-D 17 items, deriving from two independent recruitment phases TRD-I and TRD-III.[27]

2.5 | Supervised learning

Contrary to the unsupervised learning algorithm described above, supervised learning targets an outcome of interest defined by the data analyst. Here, the aim was to build a model fit for differentiating treatment outcome phenotypes from each other, based on the 88 variables described above. Classification of treatment outcome was performed with "RandomForest" (RF) as implemented in the package "randomForest" for the statistical software "R".[28] RF is an ensemble decision tree algorithm that randomly picks data subsets and performs several splits based on one predictor until treatment outcome is classified for all observations. Usually, several thousand trees are computed with different random selections of predictors and subsamples. The final model is based on majority votes from all runs. Thereby, the number of randomly selected variables available at each split within a tree ("mtry") has to be set by the analyst, usually following the recommended rule of "mtry" $= \sqrt{\text{predictors}}$ In short, a large "mtry" leads to highly optimized models, always choosing the predictor which splits perfectly as an abundance of predictors are available. In contrast, low "mtry" restrains the model from using the best predictors all the time as only few random variables are available per split. This leads to generally weaker, but more diverse models that are potentially more practicable outside the training data context.

In other words, the RF algorithm tries to distinguish patients with unfavorable from patients with favorable treatment outcome by repeatedly applying subsets of the up to 88 predictors included in the models. As thousands of combinations of predictors are selected and compared by the algorithm, RF allows to assess the importance of each predictor in consideration of a wide variety of interaction effects, giving a more complete picture than conventional statistics that often rely on univariate or highly specific interaction effects. A graphical representation of the RandomForest classifier can be found in Figure S2.

Here, classification models were built for the outcome phenotypes of interest, (i) response and remission, (ii) response for the data-driven symptom clusters, and (iii) response stratified by treatment types SSRI, TCA, and AP and lithium augmentation.

The five predictor sets were first implemented separately. Next, modalities were combined to assess the benefit of additional predictors for model performance, and finally, feature selection was applied to choose the optimal subset of predictors. Thereby, a nested cross-center validation design was applied.[29] For a schematic depiction of the validation design, please refer to Figure 1. Each of the ten participating centers was treated as a fold, leading to ten models that were trained on nine centers and validated on the left-out, independent tenth center. Please see Table S3 for details on the centers. Variable selection was performed for each of these folds with the "varSelRF" package for "R," an algorithm for backwards variable elimination based on initial importance values of each predictors.[30,31] For each iteration, 3000 trees were grown with conventional settings for "mtry" and 0.01% of variables were dropped. The whole procedure was repeated fifty times with random starting seeds, and variables that were selected in more than 50% of runs were chosen for validation. Optimal "mtry" was determined in another tenfold cross-validation run within each training set with the "caret" package for "R".[32]

For estimation of predictor performance, for each model the optimal set of predictors for the whole sample was determined with "varSelRF" and the overlap with the predictor sets selected within each fold of the respective cross-center validation was computed to assess generalizability and stability of the most informative predictors. Model performance in dependency of the number of variables used was plotted with the "plot.varSelRF" function.

Data balancing was performed by artificially increasing the number of minor class observations (oversampling) whenever the minor class was registered in less than a third of observations. Oversampling was applied as provided by the Synthetic Minority Over-sampling Technique (SMOTE).[33] For low dimensional data with a favorable ratio of features to observations (n observations >n features), SMOTE was demonstrated to be more effective than other balancing techniques.[34] The SMOTE algorithm is basically a clustering approach that computes new

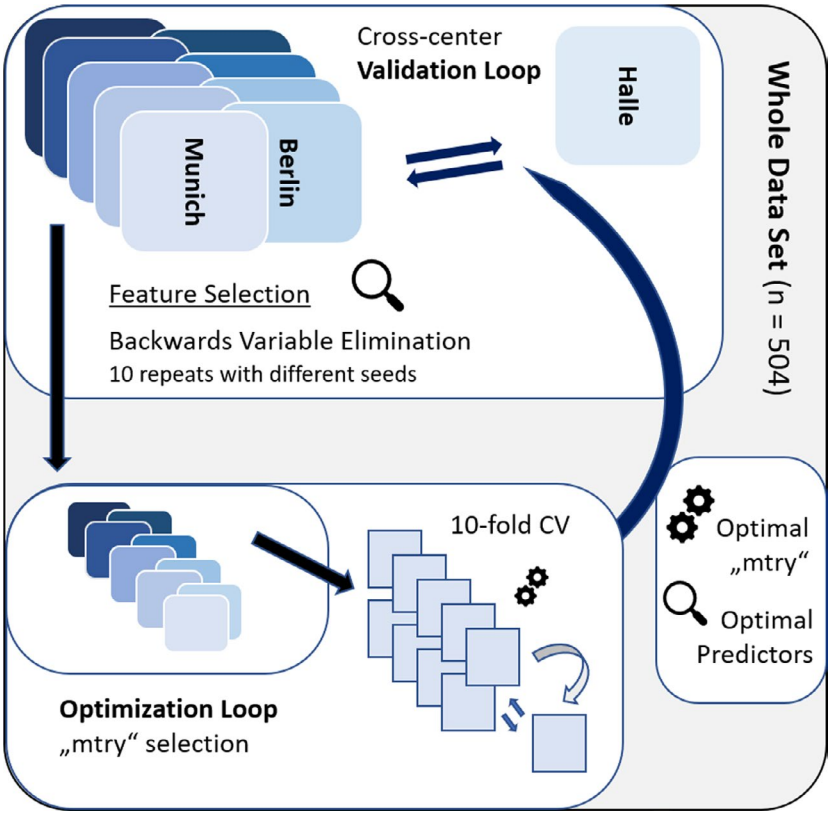

FIGURE 1 Nested cross-center validation design. The whole data set ($n = 504$) was split by recruiting centers, resulting in then folds of the outer loop. Within the inner loop, for each iteration of the outer loop the hyperparameter "mtry" was optimized in a 10-fold cross-validation. For variable selection within the outer loop, ten runs randomly seeded backwards variable elimination were performed and features selected in over 50% of the runs were chosen for "mtry" selection. Validation with optimized sets of predictors and "mtry" was performed in the left-out fold of the outer loop, represented by one independent center for each iteration [Colour figure can be viewed at wileyonlinelibrary.com]

observations for the less frequent class based on the nearest neighbors in the original sample. Applying standard settings, a new observation is created based on the five nearest neighbors in feature space. Thereby, a vector between the original sample and the nearest neighbors is computed and manipulated by multiplication with a random factor. The resulting rebalanced sample then shows even distribution of the outcome class. To prevent leakage of information from training samples to test samples trough balancing, SMOTE was applied to each fold of the cross-center validation design separately.

3 | RESULTS

3.1 | Unsupervised learning: clustering results

Hierarchical clustering of HAM-D 17 items revealed three to four easily distinguishable clusters in the GRND sample. Automated evaluation of the optimal number of clusters suggested four clusters with minimal advantage over a three-cluster solution. Similar results were found in the GSRD sample. Symptom clusters were similar, except for HAM-D item 17 that ended up in different clusters and was excluded. In the GSRD sample, the automated evaluation suggested two clusters with small advantage over a four-cluster solution (Figure S3). In synopsis with clinical considerations, the four-cluster solution was favored. In both samples, GRND and GSRD, similar clusters emerged that were named Cluster I "core emotional," Cluster II "anxious and somatic," Cluster III "sleep," and Cluster IV "appetite and weight." Cluster I was comprised of core emotional symptoms (HAM-D items 1 – 3 and 7 and 8; sadness, guilt, suicidality, and loss of interest in work and activities and psychomotor retardation), Cluster II contained anxiety-related symptoms (HAM-D items 9 – 11 and 13 and 14; psychomotor agitation, psychic and somatic anxiety, and general somatic symptoms and sexual symptoms), Cluster III represented sleep symptoms (HAM-D items 4–6; early, middle, and late insomnia), and Cluster IV appetite-related symptoms (HAM-D items 12 and 16; appetite and weight changes). A graphical presentation of clusters in both samples, GSRD and GRND, can be found in Figure 2.

Baseline values for each of the clusters correlated with baseline total HAM-D 17 score ($R = 0.43$–0.63, all $p < 0.001$) but not with baseline scores of the other clusters, except for a weak correlation of clusters I and III ($R = 0.12$, $p = 0.008$) as

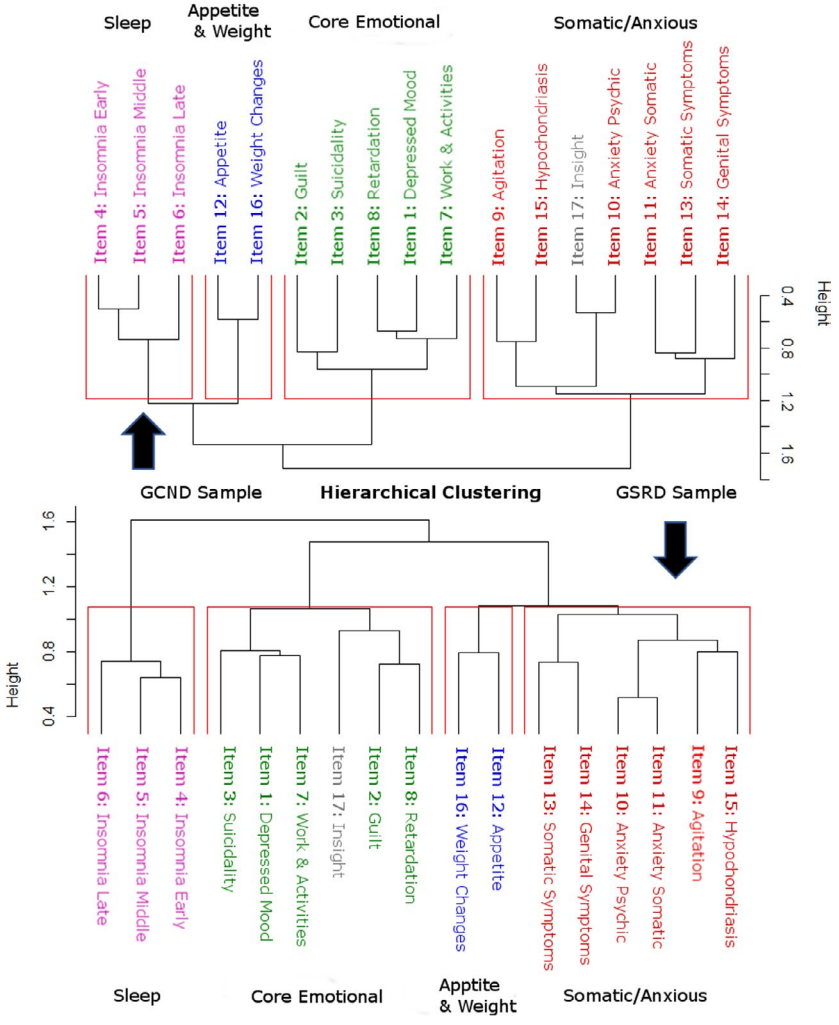

FIGURE 2 Symptom Clustering Results. Four clusters were chosen based on inspection of the hierarchical trees in two samples, the German Competence Network of Depression sample (GCND, $n = 504$) and the sample of the Group for the Studies of Resistant Depression (GSRD, $n = 1568$) as well as an automated evaluation based on the stability of partitions obtained from a hierarchy of the 17 HAM-D items in a bootstrap approach. Across both samples, similar cluster solutions were suggested, differing only by item 17 (insight). Based on their attributes, the clusters were named "Somatic & Anxious," "Core Emotional," "Sleep," and "Appetite and Weight" and are portrayed in different colors for easier interpretability [Colour figure can be viewed at wileyonlinelibrary.com]

well as clusters III and IV ($R = 0.22$, $p < 0.001$). Thus, severity of respective clusters differed within patients, but patients with high symptoms for any cluster were generally more severely affected. For correlation plots, please refer to the Figure S4.

Baseline total scores of the four clusters were added to the severity predictor set for classification analyses, resulting in a total of 88 predictors for the classification models. A plot of baseline cluster scores grouped by treatment outcome can be found in Figure S5.

3.2 | Supervised learning: prediction results

Response was reached by 55.2% of patients up to week four and 67.5% of patients up to week eight, while 46.4% of patients achieved symptom remission up to 8 weeks of treatment.

For prediction of all phenotypes, accuracy increased with the number of variables until a plateau was reached in most models at around 15 predictors. Dependency of accuracy on the number of predictors included is plotted in Figure S6.

Feature selection did not generally improve accuracies and using all predictors mostly did not compromise model performance. Performance of all predictor sets with and without feature section is listed in Table 2. For a summary of the most relevant predictors for each model and consistency of feature selection results through cross-center folds, please refer to Table S4.

3.2.1 | Response and remission

Remission after up to eight weeks of treatment could be predicted with maximal accuracy of 0.62, indicating a

TABLE 2 Accuracy of prediction models for all treatment outcome phenotypes and stratification groups. In the majority of models, using feature selection among all available predictors was most effective, with some models performing better using all predictors. Mostly, feature selection did improve accuracy by 5–10%. The optimal performing feature set for each model is highlighted in bold

Predictor set	Severity	BADO	AMDP	Comorb.	NEO-FFI	All	FS
Conventional outcome phenotypes						$n = 504$	
Remission	0.54	0.53	0.59	0.58	0.56	0.59	**0.62**
Response	0.67	0.64	0.64	0.53	0.56	0.68	**0.69**
HAM-D Clusters Cluster I – IV; $n = 393$, $n = 394$, $n = 389$, $n = 340$							
Cluster I emotional	0.66	0.59	0.61	0.55	0.60	**0.69**	0.67
Cluster II anxious	0.47	0.52	0.50	0.48	0.50	0.53	**0.56**
Cluster III sleep	0.77	0.62	0.7	0.63	0.61	**0.81**	0.79
Cluster IV appetite	0.79	0.68	0.73	0.70	0.69	**0.85**	0.84
Treatment type AP, Lithium, SSRI, TCA; $n = 204$, $n = 131$, $n = 121$, $n = 127$							
AP	0.60	0.59	0.66	0.47	0.53	0.62	**0.69**
Lithium	0.66	0.48	0.56	0.56	0.55	0.56	**0.69**
SSRI	0.78	0.78	0.75	0.64	0.62	**0.82**	**0.82**
TCA	0.77	0.72	0.67	0.52	0.69	0.79	**0.81**

HAM-D, Hamilton rating scale for depression; BADO, basic assessment scale of clinical and sociodemographic variables in psychiatry; AMDP, scale of the association for methodology and documentation in psychiatry; Comorb., comorbidities; FS, feature selection; AP, antipsychotics; SSRI, serotonin reuptake inhibitors; TCA, tricyclic antidepressants.

poor performance that was still better than chance level. Concerning the different predictor sets, the highest accuracy of 0.59 was reached with the AMDP set. Using all predictors resulted in a similar accuracy of 0.59, that was boosted modestly to 0.62 after feature selection.

For prediction of treatment response after up to eight weeks of treatment, an optimal accuracy of 0.69 was observed, indicating modest prediction performance. Thereby, a pattern of predictor set performance similar to prediction of remission was observed: the optimal model included all predictor sets and exploited feature selection.

Next, the most informative predictors were assessed for response and remission, respectively. For prediction of treatment response, age of disease onset, and overall duration, HAM-D 21 baseline score, the number of previous hospitalizations, baseline SOFAS score, HAM-D items 3 (suicidality) and 7 (work and activities), as well as MADRS item 5 (appetite), and the sleep and gastrointestinal subcategories of the AMDP were most informative.

For prediction of remission, recurrent episodes, the duration of the current episode, and of the illness, "core emotional" cluster baseline score, HAM-D item 19 (depersonalization and derealization) and MADRS item 5 (appetite), AMDP subcategories for cardiac, gastrointestinal and other somatic symptoms as well as delusion, NEO-FFI traits neuroticism, extraversion, and tolerance as well as education level were most predictive.

For "mtry" selection, ranges between 1 and 9 were tested. The optimal "mtry" settings varied between 2 and 7, notably differing from the generic rule that would suggest a less strict "mtry" at $\sqrt{90} \approx 9$.

3.2.2 | Symptom clusters

Treatment response was defined by a decline of 50% or more in cluster score within a timeframe of up to eight weeks of inpatient treatment, similar to the analyses of conventional treatment outcome phenotypes. Only patients with a relevant baseline score for as specific cluster were considered for the respective prediction model. Deduced from the maximal obtainable points within each cluster and established thresholds for severity of the total HAM-D 17 item score, a baseline score of 7 was required for Clusters I, 5 for Cluster II, 2 for Cluster III, and 1 for Cluster IV. Thus, 389, 394, 393, and 340 patients could be analyzed for Clusters I – IV, respectively. Response rates, defined by a 50% change of baseline symptoms, were similar to total HAM-D response rates (67.5%) for clusters I and III (62.7% and 67.4%), while cluster IV showed better response rates (79.4%) and cluster II considerably worse response (48.4%).

Again, optimal results were achieved using all sets of predictors and feature selection. Response for Cluster I could be predicted with an accuracy of 0.69. Response for Cluster II showed the lowest accuracy among all outcome phenotypes at 0.56. Response for Clusters III and

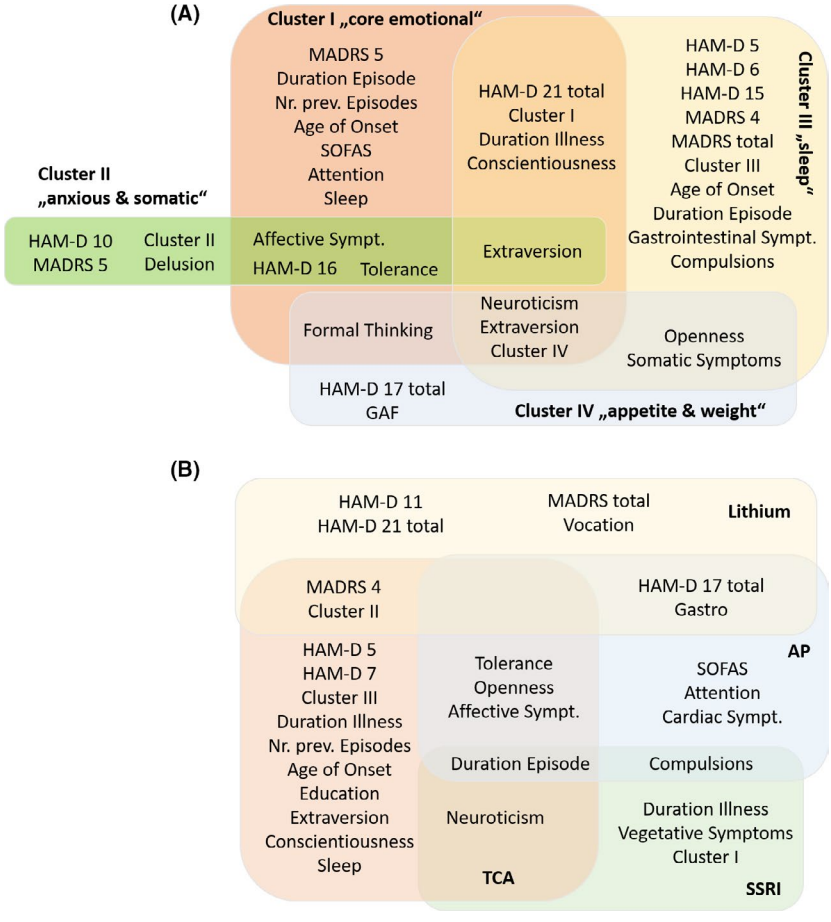

FIGURE 3 Schematic depiction of the most informative predictors for each model. Only predictors that were chosen by at least 50% of feature selection runs are shown. Models are depicted in different colors and grouped per cluster (section A) and per treatment type (section B), respectively [Colour figure can be viewed at wileyonlinelibrary.com]

IV showed high predictability with accuracies of 0.81 and 0.85, respectively. Here, optimal results were obtained using all available predictors.

The most important predictors for each cluster according to the feature selection algorithm are portrayed in Figure 3, section A.

3.2.3 | Treatment types

Patients were stratified by having received augmentation therapy with lithium ($n = 131$) or AP ($n = 208$). For those patients without augmentation therapy, further stratification by SSRI ($n = 121$) and TCA ($n = 127$) treatment was applied. There was some overlap between the lithium and neuroleptics and between the TCA and SSRI groups. Overall, stratification by treatment type enhanced the predictive power.

Prediction of response to SSRI treatment was accurate in 0.82 of observations. A comparable accuracy of 0.79 was computed for TCA. Here, again feature selection among all predictors yielded the optimal performance. For prediction of response to antipsychotic and lithium augmentation, accuracies of 0.69 were achieved. For augmentation prediction models, maximal performance was reached using all predictors.

The most important predictors for each treatment type according to the feature selection algorithm are portrayed in Figure 3, section B.

4 | DISCUSSION

Cross-center prediction models for predefined and data-driven treatment outcomes built within the multicenter database of the GRND reached accuracies from 0.56 to 0.85. For conventional treatment outcome phenotypes response and remission, moderate accuracies were achieved while stratification by treatment type and prediction of specific symptom clusters allowed higher accuracies.

These results are comparable to similar approaches in other large clinical databases, most notably the European GSRD and the American STAR*D sample. Accuracies around 0.7 were repeatedly reported for prediction of antidepressant treatment outcome,[5,10,12,15] as well as earlier decision tree-based findings in the GRND sample,[11] underlining the theorem that different learning algorithms often perform equally well as there is no gold-standard approach in machine learning.[9,35] Interestingly, contrary to the GSRD and this database, STAR*D was conducted in outpatients. The latter are known to differ from inpatients

in some clinical characteristics, most notably showing less symptom severity and suicidality.[36] This may be explaining already reported differences in clinical and sociodemographic characteristics between these samples; however, the fact that similarly effective models for prediction of treatment outcome could be computed suggests some generalizability of the results. Nevertheless, applying our models that were built from a sample with predominantly inpatients may be disrupted in a sample of outpatients.

Interestingly, in this analysis a better prediction accuracy could be achieved for response compared to remission (0.69 and 0.62, respectively). Considering that remission is the more extreme phenotype, requiring a stark decline in depressive symptoms, it may appear curious that the prediction model underperformed compared to the more broadly defined outcome of treatment response. This is also contrasting previous work by our group that showed somewhat better prediction performance for remission compared to other outcome phenotypes in a comparable sample.[10] While the differences in predicting remission and response may be exclusively related to the specifics of this particular data set, successful classification of remission also requires the model to distinguish responding from remitting patients, which may be more difficult than comparing response to non-response. In synopsis, differences of the two outcome phenotypes have been repeatedly reported for decades; thus, it can be expected that divergent results also emerge in data-driven analyses. Future research may need to include further variables, potentially addressing social support and negative cognitive styles, to further disentangle response and remission outcome phenotypes.

More importantly, however, this analysis highlights the advantages of addressing heterogeneity by sample stratification and application of data-driven response phenotypes instead of predefined total scores. While previous studies on conventional depression subtypes suggested little prognostic value for treatment outcome for different antidepressant classes,[37,38] recent reviews supported advantages of data-driven phenotypes.[39] Nevertheless, only few studies took advantage of combined unsupervised and supervised learning strategies for prediction of treatment outcome in MDD.[40,41] Still, a synthesis of machine learning applications into clinically relevant signatures of predictions for response of specific symptoms to specific therapeutics is lacking. The results of the GRND further close this gap by demonstrating that clinical features are implicated differently respective of the symptoms and drugs of interest.

Patterns for emotional and anxiety-related and sleep- and appetite-associated symptoms were observed. These clusters were detected almost identically in the GRND and GSRD samples and partly converge with earlier suggestions of data-driven HAM-D subscales. A previous study by Chekroud applied hierarchical clustering to the HAM-D and Quick Inventory of Depressive Symptomatology.[41] Three clusters emerged that were named "core emotional," "atypical," and "sleep" clusters. Similar to our results, the core emotional cluster consisted of symptoms related to mood, energy, concentration, interest, and self-worth. Interestingly, both emotional clusters resemble the traditional melancholic subtype of depression, indicating that data-driven subtypes can agree with clinical experience.[42] However, the "core emotional" cluster suggested by Chekroud also included suicidality, which sometimes is interpreted as atypical symptom and did not differ between atypical and other types of depression in other analyses.[43] Since anhedonia was demonstrated to act as risk factor for suicidality, a connection to core mood symptoms seems likely.[44,45] This is also in line with factorial analyses of the HAM-D.[46] Since the conventional concept of atypical depression also features hypersomnia and hyperphagia, both of which are not registered with the HAM-D, our results can neither support nor disagree with the relevance of this subtype of depression.

Both, the results by this analysis and by Chekroud match much earlier findings by several researchers that suggest core symptoms of depression, comprising the same symptoms of depressed mood, feelings of guilt, loss of interest in work and activities, and psychomotor retardation.[47-49] Contrary to the results of Chekroud and the previously reported core symptoms, in our analysis anxious symptoms were not connected to the core emotional symptoms but suggested to form a separate cluster together with somatic symptoms and agitation. Other investigations reported anxiety symptoms to be clustered together with sleep and/or weight loss. Appetite, weight loss, and insight were not included in the recent analysis by Chekroud, but an independent appetite cluster was suggested earlier by factorial analyses.[46] The clusters observed in the GRND and GSRD samples also support the recently established anxious subtype of depression as anxiety-related symptoms were connected to somatic symptoms and generally less favorable outcomes. While response rates for the other clusters were above or comparable to the total HAM-D response rate, cluster II response rates were considerably lower at 48.5%, reflecting worse treatment outcome reported for anxious depression.[50] On the other hand, the lower response rates could be related to treatment side effects that are most likely to manifest within cluster II and may thwart beneficial effects on other symptoms. This mechanism was suggested by recent single item analyses on SSRI response, but may be less relevant here since only patient with relevant baseline symptoms for the respective clusters were included.[51]

Concerning the most informative predictors across different models, results must be regarded preliminary. NEO-FFI traits of high neuroticism and low extraversion were associated with MDD with some consistency and balancing effects of these traits on AD treatment were reported before.[52] A recent study also reported genetic overlap between character

traits neuroticism, openness and conscientiousness and SSRI response.[53] Nevertheless, character traits failed to predict treatment outcome or performed considerably worse than clinical predictors in most investigations.[54] While NEO-FFI items were not selected for overall response, neuroticism, and extraversion as well as tolerance were selected for remission, indicating that character traits may be relevant for residual symptoms. There are suggestions in the literature that neuroticism may portray an alternative, broader, and less severe picture of MDD, which agrees with its role predicting remission.[55] Keeping in mind that NEO-FFI items by themselves performed almost at chance level, our results hint at interaction effects rather than a direct association with treatment outcome. Interestingly, NEO-FFI items were consistently chosen over personality disorder predictors by feature selection algorithms. However, this may be related to the dimensional measure of the NEO-FFI items that often outperform binomial predictors in the context of machine learning. Personality disorders had to be treated as a binomial predictor since specific disorders could not be accounted for owing to low comorbidity rates in the GCND sample.

In accordance with extensive previous work,[27,56-58] baseline severity and duration of the current episode as well as the overall illness were consistent predictors highlighted by most of the classification models, with unfavorable effects of early age of onset, longer duration of the current episode, and time lived with MDD. Also, the circumscribed clusters for sleep and appetite were predicted by these variables, indicating an overarching role for antidepressant treatment. Contrary to other reports, suicidality was not impactful for any model except prediction of overall treatment response.[10,59] While almost half of the patients in the GRND sample showed suicidality, a potential explanation is the rather low severity of suicidality compared to some other samples, indicated by an average HAM-D item 3 score of 1.65.

Interestingly, patients with favorable treatment outcome showed marginally higher baseline total and symptom cluster scores. This has previously been discussed and may be because of treatment response being defined by percentage change from a baseline score instead of an absolute threshold.[11] On the other hand, these effects were more pronounced for sleep and appetite symptom clusters which also showed higher response rates than total or emotional symptom scores. In contrast, baseline anxious symptom scores were higher among patients with unfavorable treatment outcome and showed considerably lower response rates. In synopsis, patients with more responsive symptoms such as sleep disturbances and loss of appetite may achieve a reduction in total score greater than 50% more easily.

Another surprising result was that psychiatric comorbidities hardly contributed to prediction performance. Personality disorders and a range of anxiety disorders, including panic disorder, social phobia, and generalized anxiety disorder, were previously demonstrated to hinder treatment response.[57,60-62] While anxiety disorders were not selected among the most informative features, HAM-D items related to anxiety as well as the baseline score for the anxious cluster were relevant for prediction models for response to TCA and lithium augmentation, with higher anxiety scores in the unfavorable treatment outcome groups. Only a small portion of patients showed personality (13.3%) or anxiety disorders (17.4%), and consequently, stratification by specific diagnosis was disregarded. Considering that most previous studies reported effects for some but not all anxiety disorders, currently no definite conclusion can be drawn.[57]

AMDP subcategories provide additional information compared to standard clinical interviews, further demonstrated by the consistent pick rates of these subcategories by feature selection algorithms. Overall, there are hardly data on the association of baseline AMDP items with antidepressant treatment outcome phenotypes. According to feature selection results, the sleep subcategory was generally more informative than respective HAM-D and MADRS items. Additionally, somatic and potentially side effect-related AMDP subcategories increased model performance for overall treatment outcome, especially gastrointestinal, cardiac, and other somatic symptoms. Interestingly, most psychopathology related subcategories as affective symptoms, attention, and formal thinking were only selected for cluster and treatment specific models, indicating a specialized role of these predictors. Generally, a higher symptom score was observed in the groups with unfavorable treatment outcome for all AMDP subscores. Interestingly, the AMDP may also provide some coverage of the so-called reverse vegetative symptoms, which are not assessed by the HAM-D and MADRS. Considering that reverse vegetative symptoms may occur in younger patients that are often treated as outpatients and the fact that we did not specifically address these symptoms, they may be underrepresented in our sample.

In synopsis, our results advocate looking at symptom clusters as potential predictors rather than total HAM-D scores. However, predictor performance across different clusters and treatment types must be interpreted with caution. Previous studies have demonstrated that certain antidepressants may be better suited to address specific symptom clusters and that treatment response to these drugs is predicted by distinctive features, but only study designs with precisely defined treatment arms allow for a clear attribution.[14,15,41] Since the GCND is a naturalistic sample, most patients received several antidepressant agents of various types. While some patients were drug-naïve or at least untreated for the current episode, others had already received antidepressants trials at study inclusion. Thus, only a basic stratification design by treatment type was possible.

Reflecting on the accuracy rates observed for the different prediction models, it seems that response to some clusters and

drugs can be predicted more easily than others. Especially, standard treatment with SSRI, but also TCA, showed good predictability with accuracies above 80%. Curiously, some previous reviews pointed out a decline of accuracy with increasing sample sizes to be common in machine learning analyses in neuropsychiatry.[35,63,64] Similarly, in the GCND sample models built on smaller groups showed better accuracies. This may be owed to decreased heterogeneity in the stratified samples. On the other hand, smaller models can be prone to overfitting despite optimal validation designs.[63] The same reservations hold true for cluster-based prediction models and the better predictability of sleep and appetite symptom scores compared to emotional, anxious, or total symptom scores.

Another relevant consideration for different prediction performance in the treatment groups may be baseline symptom severity that was significantly lower in the SSRI group (mean HAM-D score 21.4) compared to all three other treatment groups with average HAM-D scores of 22.7 for TCA and 23.3 and 23.7 for augmentation groups with antipsychotics and lithium, respectively. Similarly, the fraction of first episode depression was higher in the SSRI and TCA groups (both 29.1%) compared to the augmentation groups (21.2% and 12% for antipsychotics and lithium, respectively). The portion of patients with longer duration of the current episode (>6 month) was comparable between groups (62.67–68.75%). Comorbid anxiety disorders were most common in the TCA group (15.7%) and least common in the antipsychotic augmentation group (8.2%), while psychic anxiety (HAM-D item 10) did not substantially differ between groups and ranged from mean scores of 1.89 in the SSRI group to 2.13 in the lithium group.

Along these lines, the better predictability of the SSRI group may also be explained by higher importance of favorable predictor values such as lower symptom severity and shorter length of illness for model performance, as these values were over-represented in the SSRI group. Similarly, the fraction of treatment response was the higher in the SSRI and TCA groups (both 72%) compared to augmentation groups (61% and 57% for antipsychotics and lithium, respectively).

A decisive limitation is the lack of a completely independent sample for model validation. Nevertheless, the GCND database allows for quasi-independent cross-center validation since patients were recruited in ten different German university and communal hospitals. However, model performance may still be dependent on the exact definition of predictors as well as outcome phenotypes.

Another limitation to keep in mind is the fact that this was a naturalistic study, meaning that patients received a wide range of medication according to clinical judgment. The latter was at least partly based on the same variables used for prediction modeling, as for example patients with agitation may be more likely to receive sedating antidepressants or augmentation with antipsychotics. Thus, it is likely that predictors contributing to differential prediction dependent on treatment modality are biased by treatment selection itself and must be interpreted with caution.

Finally, the oversampling design may bear a risk of biased accuracies. Rebalancing of the data set is generally recommended for classification problems with very few observations with the minor outcome class. While the ratio was not extreme in this sample, it was demonstrated that data balancing can increase model performance when cross-validation folds differ in size and outcome ratios. Reflecting on previous investigations on SMOTE and other oversampling algorithms, the risk of boosted accuracies seems to be low as SMOTE hardly improved total accuracies but rather produced balanced sensitivity and specificity.[34]

Overall, our results further demonstrate that advanced statistics allow prediction of treatment outcome for MDD on a clinically relevant level.[5,9] Furthermore, treatment- and symptom-specific algorithms can be generated and bring along advantages for model precision. Unfavorable treatment outcome may increase with lifetime length of illness as shown by higher number of hospitalizations and longer duration of the current episode as well as overall illness. Similarly, more severe depression and especially anxiety-related and somatic symptoms may hinder successful treatment. Personality traits as neuroticism and extraversion also moderate treatment success. Even though these results agree with and expand on previous auspicious machine learning results in MDD, only prospective application of the established models will allow computer-aided diagnostic and predictive tools to prove their value for the clinic.

ACKNOWLEDGEMENTS

The network study was conducted in 12 psychiatric hospitals: Berlin Charite Campus Mitte (Andreas Heinz, Mazda Adli, Katja Wiethoff), Berlin Charité Campus Benjamin Franklin (Isabella Heuser, Gerd Bischof), Berlin Auguste Viktoria Klinik (Joachim Zeiler, Robert Fisher, Cornelia Fähser), Berlin St. Hedwig (Florian Standfest), Berlin St. Joseph (Dorothea Schloth), Düsseldorf (Wolfgang Gaebel, Joachim Cordes, Arian Mobascher), Gabersee (Gerd Laux, Sissi Artmann), Haar (Wolfram Bender, Nicole Theyson), Halle (Andreas Marneros, Dörthe Strube, Yvonne Reinelt, Peter Brieger), Heidelberg (Christoph Mundt, Klaus Kronmüller, Daniela Victor), München LMU (Hans-Jürgen Möller, Ulrich Hegerl, Roland Mergel, Michael Riedel, Florian Seemüller, Florian Wickelmaier, Markus Jäger, Thomas Baghai, Ingrid Borski, Constanze Schorr, Roland Bottlender), and München MPI (Florian Holsboer, Matthias Majer, Marcus Ising).

CONFLICT OF INTEREST

All other authors declare that they have no conflicts of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/acps.13250.

DATA AVAILABILITY STATEMENT

Data can be made available upon reasonable request.

ORCID

Alexander Kautzky https://orcid.org/0000-0001-9251-8285
Markus Dold https://orcid.org/0000-0001-8914-2192
Siegfried Kasper https://orcid.org/0000-0001-8278-191X

REFERENCES

1. Vos T, Allen C, Arora M, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016;388:1545-1602.
2. Akil H, Gordon J, Hen R, et al. Treatment resistant depression: A multi-scale, systems biology approach. Neurosci Biobehav Rev. 2018;84:272-288.
3. Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. Lancet. 2018;391:1357-1366.
4. Papakostas GI, Ionescu DF. Towards new mechanisms: an update on therapeutics for treatment-resistant major depressive disorder. Mol Psychiatry. 2015;20:1142-1150.
5. Cohen ZD, Derubeis RJ. Treatment selection in depression. Annu Rev Clin Psychol. 2018;14:209-236.
6. Serretti A. The present and future of precision medicine in psychiatry: focus on clinical psychopharmacology of antidepressants. Clin Psychopharmacol Neurosci. 2018;16:1-6.
7. Bauer M, Severus E, Köhler S, et al. World Federation of Societies of Biological Psychiatry (WFSBP) guidelines for biological treatment of unipolar depressive disorders. part 2: maintenance treatment of major depressive disorder-update 2015. World J Biol Psychiatry. 2015;16:76-95.
8. Bauer M, Pfennig A, Severus E, et al. World Federation of Societies of Biological Psychiatry (WFSBP) guidelines for biological treatment of unipolar depressive disorders, part 1: update 2013 on the acute and continuation treatment of unipolar depressive disorders. World J Biol Psychiatry. 2013;14:334-385.
9. Passos I, Mwangi B, Kapczinski F. Personalized Psychiatry: Big Data Analytics in Mental Health. Springer; 2018.
10. Kautzky A, Baldinger-Melich P, Kranz GS, et al. A new prediction model for evaluating treatment-resistant depression. J Clin Psychiatry. 2017;78:215-222.
11. Riedel M, Möller H-J, Obermeier M, et al. Clinical predictors of response and remission in inpatients with depressive syndromes. J Affect Disord. 2011;133:137-149.
12. Perlis RH. A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. Biol Psychiatry. 2013;74:7-14.
13. Kautzky A, Dold M, Bartova L, et al. Refining prediction in treatment-resistant depression: results of machine learning analyses in the TRD III sample. J Clin Psychiatry. 2017;79.
14. Iniesta R, Hodgson K, Stahl D, et al. Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables. Sci Rep. 2018;8:5530.
15. Chekroud AM, Zotti RJ, Shehzad Z, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry. 2016;3:243-250.
16. Seemüller F, Riedel M, Obermeier M, et al. Outcomes of 1014 naturalistically treated inpatients with major depressive episode. Eur Neuropsychopharmacol. 2010;20:346-355.
17. Riedel M, Möller H-J, Obermeier M, et al. Response and remission criteria in major depression–a validation of current practice. J Psychiatr Res. 2010;44:1063-1068.
18. Souery D, Amsterdam J, de Montigny C, et al. Treatment resistant depression: methodological overview and operational criteria. Eur Neuropsychopharmacol. 1999;9:83-91.
19. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry. 1979;134:382-389.
20. Hamilton M. Development of a rating scale for primary depressive illness. Br J Soc Clin Psychol. 1967;6:278-296.
21. Cording C, Gaebel W, Spengler A, et al. Die neue psychiatrische Basisdokumentation. Eine Empfehlung der DGPPN zur Qualitätssicherung im (teil-)stationären Bereich. Spektrum der Psychiatrie und. Nervenheilkunde. 1995;24:3-41.
22. Jones SH, Thornicroft G, Coffey M, Dunn G. A brief mental health outcome scale-reliability and validity of the Global Assessment of Functioning (GAF). Br J Psychiatry. 1995;166:654-659.
23. Morosini PL, Magliano L, Brambilla L, Ugolini S, Pioli R. Development, reliability and acceptability of a new version of the DSM-IV Social and Occupational Functioning Assessment Scale (SOFAS) to assess routine social functioning. Acta Psychiatr Scand. 2000;101:323-329.
24. Stieglitz RD, Haug A, Fahndrich E, Rosler M, Trabert W. Comprehensive Psychopathological Assessment Based on the Association for Methodology and Documentation in Psychiatry (AMDP) System: Development, Methodological Foundation, Application in Clinical Routine, and Research. Front Psychiatry. 2017;8:45.
25. Piedmont RL, McCrae RR, Costa PT. An assessment of the Edwards Personal Preference Schedule from the perspective of the five-factor model. J Pers Assess. 1992;58:67-78.
26. Chavent M, Kuentz-Simonet V, Liquet B, Saracco J. ClustOfVar: An R Package for the clustering of variables. J Stat Softw. 2012;50:1-16.
27. Kautzky A, Dold M, Bartova L, et al. Clinical factors predicting treatment resistant depression: affirmative results from the European multicenter study. Acta Psychiatr Scand. 2019;139:78-88.
28. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2:18-22.
29. Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. NeuroImage. 2017;145:166-179.
30. Diaz-Uriarte R. GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. BMC Bioinformatics. 2007;8.
31. Diaz-Uriarte R, Alvarez DE, Andres A. Gene selection and classification of microarray data using random forest. BMC Bioinform. 2006;7.
32. Kuhn M. Caret package. J Stat Softw. 2008;28.

33. Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTE: synthetic minority over-sampling technique. J Artificial Intelligence Res. 2002;16:321-357.
34. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics. 2013;14:106.
35. Gao S, Calhoun VD, Sui J. From classification to treatment outcome prediction. CNS Neurosci Ther; 2018.
36. Bartova L, Dold M, Kautzky A, et al. Results of the European Group for the Study of Resistant Depression (GSRD) - basis for further research and clinical practice. World J Biol Psychiatry. 2019;20:427-448.
37. Arnow BA, Blasey C, Williams LM, et al. Depression subtypes in predicting antidepressant response: a report from the iSPOT-D trial. Am J Psychiatry. 2015;172:743-750.
38. Uher R, Dernovsek MZ, Mors O, et al. Melancholic, atypical and anxious depression subtypes and outcome of treatment with escitalopram and nortriptyline. J Affect Disord. 2011;132:112-120.
39. van Loo HM, de Jonge P, Romeijn JW, Kessler RC, Schoevers RA. Data-driven subtypes of major depressive disorder: a systematic review. BMC Med. 2012;10:156.
40. Kautzky A, Baldinger P, Souery D, et al. The combined effect of genetic polymorphisms and clinical parameters on treatment outcome in treatment-resistant depression. Eur Neuropsychopharmacol. 2015;25:441-453.
41. Chekroud AM, Gueorguieva R, Krumholz HM, Trivedi MH, Krystal JH, McCarthy G. Reevaluating the efficacy and predictability of antidepressant treatments: a symptom clustering approach. JAMA Psychiatry. 2017;74:370-378.
42. Musil R, Seemüller F, Meyer S, et al. Subtypes of depression and their overlap in a naturalistic inpatient sample of major depressive disorder. Int J Methods Psychiatr Res. 2018;27:e1569.
43. Seemüller F, Riedel M, Wickelmaier F, et al. Atypical symptoms in hospitalised patients with major depressive episode: frequency, clinical characteristics, and internal validity. J Affect Disord. 2008;108:271-278.
44. Loas G, Lefebvre G, Rotsaert M, Englert Y. Relationships between anhedonia, suicidal ideation and suicide attempts in a large sample of physicians. PLoS One. 2018;13:e0193619.
45. Winer ES, Drapeau CW, Veilleux JC, Nadorff MR. The association between anhedonia, suicidal ideation, and suicide attempts in a large student sample. Arch Suicide Res. 2016;20:265-272.
46. Trivedi MH, Morris DW, Grannemann BD, Mahadi S. Symptom clusters as predictors of late response to antidepressant treatment. J Clin Psychiatry. 2005;66:1064-1070.
47. Bech P. Rating scales for affective disorders: their validity and consistency. Acta Psychiatr Scand Suppl. 1981;295:1-101.
48. Maier W, Philipp M, Heuser I, Schlegel S, Buller R, Wetzel H. Improving depression severity assessment–I. Reliability, internal validity and sensitivity to change of three observer depression scales. J Psychiatr Res. 1988;22:3-12.
49. McIntyre R, Kennedy S, Bagby RM, Bakish D. Assessing full remission. J Psychiatry Neurosci. 2002;27:235-239.
50. Gaspersz R, Nawijn L, Lamers F, Penninx B. Patients with anxious depression: overview of prevalence, pathophysiology and impact on course and treatment outcome. Curr Opin Psychiatry. 2018;31:17-25.
51. Hieronymus F, Lisinski A, Nilsson S, Eriksson E. Influence of baseline severity on the effects of SSRIs in depression: an item-based, patient-level post-hoc analysis. Lancet Psychiatry. 2019;6:745-752.
52. Wardenaar KJ, Conradi HJ, Bos EH, de Jonge P. Personality modulates the efficacy of treatment in patients with major depressive disorder. J Clin Psychiatry. 2014;75:e916-e923.
53. Amare AT, Schubert KO, Tekola-Ayele F, et al. Association of the polygenic scores for personality traits and response to selective serotonin reuptake inhibitors in patients with major depressive disorder. Front Psychiatry. 2018;9:65.
54. Blom MBJ, Spinhoven P, Hoffman T, et al. Severity and duration of depression, not personality factors, predict short term outcome in the treatment of major depression. J Affect Disord. 2007;104:119-126.
55. Kotov R, Gamez W, Schmidt F, Watson D. Linking, "big" personality traits to anxiety, depressive, and substance use disorders: a meta-analysis. Psychol Bull. 2010;136:768-821.
56. Schosser A, Serretti A, Souery D, et al. European Group for the Study of Resistant Depression (GSRD)–where have we gone so far: review of clinical and genetic findings. Eur Neuropsychopharmacol. 2012;22:453-468.
57. de Carlo V, Calati R, Serretti A. Socio-demographic and clinical predictors of non-response/non-remission in treatment resistant depressed patients: A systematic review. Psychiatry Res. 2016;240:421-430.
58. Iniesta R, Malki K, Maier W, et al. Combining clinical variables to optimize prediction of antidepressant treatment outcomes. J Psychiatr Res. 2016;78:94-102.
59. Souery D, Oswald P, Massat I, et al. Clinical factors associated with treatment resistance in major depressive disorder: results from a European multicenter study. J Clin Psychiatry. 2007;68:1062-1070.
60. Bock C, Bukh JD, Vinberg M, Gether U, Kessing LV. The influence of comorbid personality disorder and neuroticism on treatment outcome in first episode depression. Psychopathology. 2010;43:197-204.
61. Papakostas GI, Petersen TJ, Farabaugh AH, et al. Psychiatric comorbidity as a predictor of clinical response to nortriptyline in treatment-resistant major depressive disorder. J Clin Psychiatry. 2003;64:1357-1361.
62. Petersen T, Hughes M, Papakostas GI, et al. Treatment-resistant depression and Axis II comorbidity. Psychother Psychosom. 2002;71:269-274.
63. Pulini AA, Kerr WT, Loo SK, Lenartowicz A. Classification accuracy of neuroimaging biomarkers in attention-deficit/hyperactivity disorder: effects of sample size and circular analysis. Biol Psychiatry Cogn Neurosci Neuroimaging. 2018.
64. Schnack HG, Kahn RS. Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters. Front Psychiatry. 2016;7:50.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Kautzky A, Möller HJ, Dold M, et al. Combining machine learning algorithms for prediction of antidepressant treatment response. *Acta Psychiatr Scand*. 2021;143:36–49. https://doi.org/10.1111/acps.13250