Refining Prediction in Treatment-Resistant Depression:
Results of Machine Learning Analyses in the TRD III Sample

Alexander Kautzky, MD[a]; Markus Dold, MD, PhD[a]; Lucie Bartova, MD[a]; Marie Spies, MD[a]; Thomas Vanicek, MD, PhD[a]; Daniel Souery, MD, PhD[b]; Stuart Montgomery, MD, Prof[c]; Julien Mendlewicz, MD, PhD, Prof[d]; Joseph Zohar, MD, Prof[e]; Chiara Fabbri, MD, PhD[f]; Alessandro Serretti, MD, PhD, Prof[f]; Rupert Lanzenberger, MD, Prof[a]; and Siegfried Kasper, MD, Prof[a,*]

ABSTRACT

Objective: The study objective was to generate a prediction model for treatment-resistant depression (TRD) using machine learning featuring a large set of 47 clinical and sociodemographic predictors of treatment outcome.

Method: 552 Patients diagnosed with major depressive disorder (MDD) according to *DSM-IV* criteria were enrolled between 2011 and 2016. TRD was defined as failure to reach response to antidepressant treatment, characterized by a Montgomery-Asberg Depression Rating Scale (MADRS) score below 22 after at least 2 antidepressant trials of adequate length and dosage were administered. RandomForest (RF) was used for predicting treatment outcome phenotypes in a 10-fold cross-validation.

Results: The full model with 47 predictors yielded an accuracy of 75.0%. When the number of predictors was reduced to 15, accuracies between 67.6% and 71.0% were attained for different test sets. The most informative predictors of treatment outcome were baseline MADRS score for the current episode; impairment of family, social, and work life; the timespan between first and last depressive episode; severity; suicidal risk; age; body mass index; and the number of lifetime depressive episodes as well as lifetime duration of hospitalization.

Conclusions: With the application of the machine learning algorithm RF, an efficient prediction model with an accuracy of 75.0% for forecasting treatment outcome could be generated, thus surpassing the predictive capabilities of clinical evaluation. We also supply a simplified algorithm of 15 easily collected clinical and sociodemographic predictors that can be obtained within approximately 10 minutes, which reached an accuracy of 70.6%. Thus, we are confident that our model will be validated within other samples to advance an accurate prediction model fit for clinical usage in TRD.

J Clin Psychiatry 2018;79(1):16m11385

To cite: 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.* 2018;79(1):16m11385.

To share: https://doi.org/10.4088/JCP.16m11385

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

[b]Université Libre de Bruxelles and Psy Pluriel Centre Européen de Psychologie Medicale, Brussels, Belgium

[c]Imperial College, University of London, London, United Kingdom

[d]School of Medicine, Free University of Brussels, Brussels, Belgium

[e]Psychiatric Division, Chaim Sheba Medical Center, Ramat Gan, Israel

[f]Department of Biomedical and Neuromotor Sciences, University of Bologna, Bologna, Italy

**Corresponding author:* Siegfried Kasper, MD, Department of Psychiatry and Psychotherapy, Medical University of Vienna, Währinger Gürtel 18-20, A-1090 Vienna, Austria (sci-biolpsy@meduniwien.ac.at).

Treatment-resistant depression (TRD) is characterized by a perseverance of depressive symptoms even after treatment trials of adequate dosage and duration have been applied.[1–4] As the broadest definition of TRD, nonresponse is described by the failure to achieve significant symptom relief after a single antidepressant treatment. Thereby, a reduction of at least 50% from baseline on a recognized rating scale for major depressive disorder (MDD) is commonly used to classify treatment response.[5–9] Most classifications require 2 failed treatment trials of adequate dosage and duration for TRD.

Major depressive disorder was ranked as the 11th most impairing disease overall as measured by disability-adjusted life-years in 2010.[10,11] As about a third of patients treated with antidepressant agents do not show sufficient symptom relief after a first antidepressant trial and 15% of patients remain afflicted even after multiple antidepressant trials are applied, TRD is in fact a substantial socioeconomic issue.[12–14] Consequently, for over a decade, our task force Group for the Study of Resistant Depression (GSRD), a multinational European research consortium, strove for potent strategy development for TRD, including characterization of risk markers and prediction tools to advance precision medicine.[15]

Previous single factor–based efforts for distinguishing risk markers in TRD have consistently highlighted the importance of clinical predictors for treatment outcome. However, none of these clinical markers was useful for detecting patients at high risk of resisting multiple antidepressant trials, leading to recommendations by recent reviews and think-tanks to focus on multivariate models.[16–20] The first study by our group adopting a multivariate approach suggested a combination of clinical and genetic factors as a signature characteristic of risk of nonresponse to therapy.[21] Furthermore, we achieved a classification based on a large set of clinical markers with over 70% accuracy for predicting TRD.[22] As studies featuring large sets of predictors for TRD are still limited, we investigated treatment outcome in TRD using a set of 47 clinical and sociodemographic predictors to confirm and substantiate our previous efforts.[21,22] All analyses were performed in a new patient collective of the GSRD data pool, independent from previous investigations.

METHOD

Sample Description

This analysis included 1,409 patients recruited from the participating centers in Italy (Bologna and Siena), Greece (Athens), Austria (Vienna), Switzerland (Geneva), Belgium (Brussels),

Clinical Points

- An advanced statistical prediction model with 47 clinical and sociodemographic predictors was established in a sample of 552 patients with major depressive disorder (MDD). Machine learning allows better prediction for treatment outcome in MDD than clinical evaluation, neuropsychiatric scales, or tools as EEG and fMRI, enabling accuracies around 70% with a set of 15 easily obtainable clinical and sociodemographic variables.
- Accounting for interactions between variables allowed prediction of treatment-resistant depression. Baseline MADRS score, quality of life, severity, and suicidal risk were the most informative predictors, followed by age, BMI, and the number of depressive episodes over lifetime. Sociodemographic variables such as profession, occupation, income type, and education also seem to be useful predictors.

Germany (Halle), France (Elancourt and Toulouse), and Israel (Tel Hashomer) between 2011 and 2016, labeled the TRD III sample, within the GSRD data pool. The study was approved by the ethical committees of all participating centers. The sample of the first 1,186 patients for whom data collection was completed (epidemiologic sample) has been described recently.[23] In short, participating subjects had to be at least 18 years old and be diagnosed with MDD according to *DSM-IV* criteria. Informed consent was obligatory for enrollment. A modified version of the MINI-International Neuropsychiatric Interview Version 5.0.0 (MINI),[24] as well as the Hamilton Depression Rating Scale (HDRS),[25] the Montgomery-Asberg Rating Scale for Depression (MADRS),[26] and the Young Mania Rating Scale (YMRS),[27] were applied for diagnosis and detailed assessment of MDD symptoms as well as psychiatric comorbidities. The primary diagnosis was required to be MDD, and patients showing only a secondary MDD diagnosis were excluded. Furthermore, substance abuse disorders except for nicotine abuse, as well as severe personality disorders, were regarded as exclusion criteria to rule out insufficient treatment response caused by these disorders.

Of the 1,409 patients, 917 showed the required outcome phenotypes of treatment response or TRD. Of these, 552 patients were eligible for this analysis based on full data availability for all 47 variables. Of these patients, 362 (235 female; mean ± SD age, 52.14 ± 14.07 years) were treatment-resistant, failing to respond to at least 2 antidepressant trials, and 190 reached an MADRS score below 22 and a score reduction of at least 50% and were thus classified as responders (112 female; mean age, 51.26 ± 14.74 years). Responders to antidepressant therapy and treatment-resistant patients showed comparable baseline MADRS scores of 34.54 and 36.23, respectively. No significant differences regarding sex, age, or baseline MADRS score were shown when excluded patients were compared to the 552 finally enrolled subjects or resistant patients were compared to responders. For a graphic depiction of recruitment and study cohorts, please see Supplementary eFigure 1.

Treatment Outcome Phenotypes

The 2 phenotypes analyzed for outcome were treatment response and TRD. To evaluate treatment outcome, MADRS scores were calculated for the time point of recruitment as well as retrospectively for the starting point of the first antidepressive treatment of the current major depressive episode (MDE), further referred to as the baseline MADRS score. A baseline MADRS score of at least 22 was required for inclusion, and phenotypes were determined by the changes in MADRS score. Treatment response was defined by reaching a MADRS score < 22 and a score decrease of at least 50% with treatment duration of at least 4 weeks at an adequate dose. TRD was defined by a failure to reach treatment response after at least 2 antidepressant trials of adequate duration and dosage. Patients who showed nonresponse after having received only 1 trial were excluded from this analysis, a necessary measure for maintaining a clear distinction between treatment response and TRD for classification.

Regarding mean baseline MADRS values, no significant differences could be observed between the outcome phenotypes treatment response (33.7), nonresponse (33.1), and TRD (35.7) in the full TRD III sample. Regarding patients included in the machine learning model, no differences could be observed between TRD (35.1) and treatment response (33.2) as well.

Predictors

Forty-seven predictors documented in the GSRD database were used for the prediction algorithm after exclusion of predictors with more than 20% missing values as well as redundant or unfeasible predictors. All 47 predictors are based on items of the MINI psychiatric interview. For a comprehensive list of all predictors, see Table 1.

The clinical symptoms fatigue, appetite change, sleep impairment, psychomotor agitation, social dysfunction, impaired decision making, feelings of guilt, and autoaggressive thoughts were recorded and included based on items of the MINI psychiatric interview (MINI A5A-G as well as 6A). Severity was defined by an abundance of symptoms at the worst stage of the current episode and coded as moderate or severe. Impairment in social, family, and work life was measured by the Sheehan Disability Scale (SDS).[28] In general, predictors derived from MINI items are based on a time range of 1 month before the rating, thereby covering the current depressive episode. Generalized anxiety disorder (GAD) symptoms are assessed for a time frame of 6 months, while panic disorder symptoms are assessed as lifetime. As some patients had already responded to therapy at the time of the interview, some scores, such as SDS, were evaluated retrospectively for the current episode.

Sociodemographic variables were characterized in a more sophisticated manner compared to our previous analysis; we distinguished occupation (eg, full time), profession (eg, employee, manager), and income type (eg, salary). Accommodation describes the social home environment, indicating if a patient is, for example, living at their parents' house or with a roommate.

Table 1. List of All 47 Predictors Featured in the Analysis Ordered by Groups

Group	Predictors
Sociodemographic (n = 10)	Age, sex, ethnicity, occupation, income type, profession, accommodation, education, relationship status, number of children
MDD history (n = 7)	Family history of MDD, family history of BD, number of relatives with MDD, number of relatives with BD, number of MDEs, timespan between first and last MDE, lifetime duration of hospitalization
Axis II comorbidity (n = 9)	GAD, social phobia, OCD, PTSD, panic disorder, agoraphobia, smoking, number of cigarettes, YMRS score
Axis III comorbidity (n = 3)	Diabetes, thyroid disorder, BMI
Clinical features (n = 13)	Severity, suicidality, suicidal risk, change of appetite, change of sleep, feelings of guilt, impaired decision making, fatigue, social dysfunction, psychomotor agitation, melancholia, autoaggressive thoughts, psychosis
Other predictors (n = 5)	Inpatient or outpatient, quality of social life, quality of work life, quality of family life, retrospective MADRS score

Abbreviations: BD = bipolar disorder, BMI = body mass index, GAD = generalized anxiety disorder, MADRS = Montgomery-Asberg Depression Rating Scale, MDD = major depressive disorder, MDE = major depressive episode, OCD = obsessive-compulsive disorder, PTSD = posttraumatic stress disorder, YMRS = Young Mania Rating Scale.

Statistical Analysis

We used the randomForest (RF) package for R software for our investigation (cran.r-project.org/), a machine learning algorithm suitable for classification of a dichotomous outcome variable as well as regression.[29] RF determines which predictor variables are most useful for distinguishing between TRD and treatment response by using a tree-based approach, calculating variable importance by residual sum of squares (RSS). Thereby, regression trees are computed, applying regression models to the outcome variable for each of the predictors. Consequently, splitting of the data is performed at several points for each independent variable. The difference between the predicted and actual values of the outcome variable is hence summed for all computations, and the resulting sum of squared error is used for determining the best variable for splitting. The binary outcome phenotypes TRD and treatment response were subsequently calculated based on the predicted MADRS score and the baseline MADRS score.

RF requires the input of a random starting value for decision tree growing; thus, multiple seeds were supplied, and the results of all these runs were averaged. The number of trees to grow was determined as 2,000 for each run, ensuring that every data row gets predicted a few times (ntree = 2,000 in the algorithm). We aligned our formula with the standard of using the square root of the number of predictors for sampling random candidate predictors for each split. While no power calculation has been established for RF yet, machine learning algorithms such as RF have been shown to produce reliable results even with a high ratio of predictors and observations, providing sufficient overall sample size and excluding all missing data.[30,31]

RF requires a learning and a test sample. For maximal efficiency and reliability, we computed our results with a 10-fold cross-validation. We performed a stepwise factor reduction to the training sample, repeating the prediction of treatment outcome with a gradually reduced number of predictors, starting with the least informative ones based on RSS. Finally, to test for more clinically practicable signatures of predictors, we tested smaller sets of predictors in the cross-validated sample as well as an additional test sample comprising patients formally excluded due to missing values for the whole set of predictors.

For comprehensive depiction of model performance, we also generated a receiver operating characteristic space in the form of a sensitivity vs (1 – specificity) plot for all models using R.

As RF does not provide information on how the importance of the predictors come into play, we also computed a generalized linear model using analysis of variance (ANOVA) to test for conventional single factor effects.

RESULTS

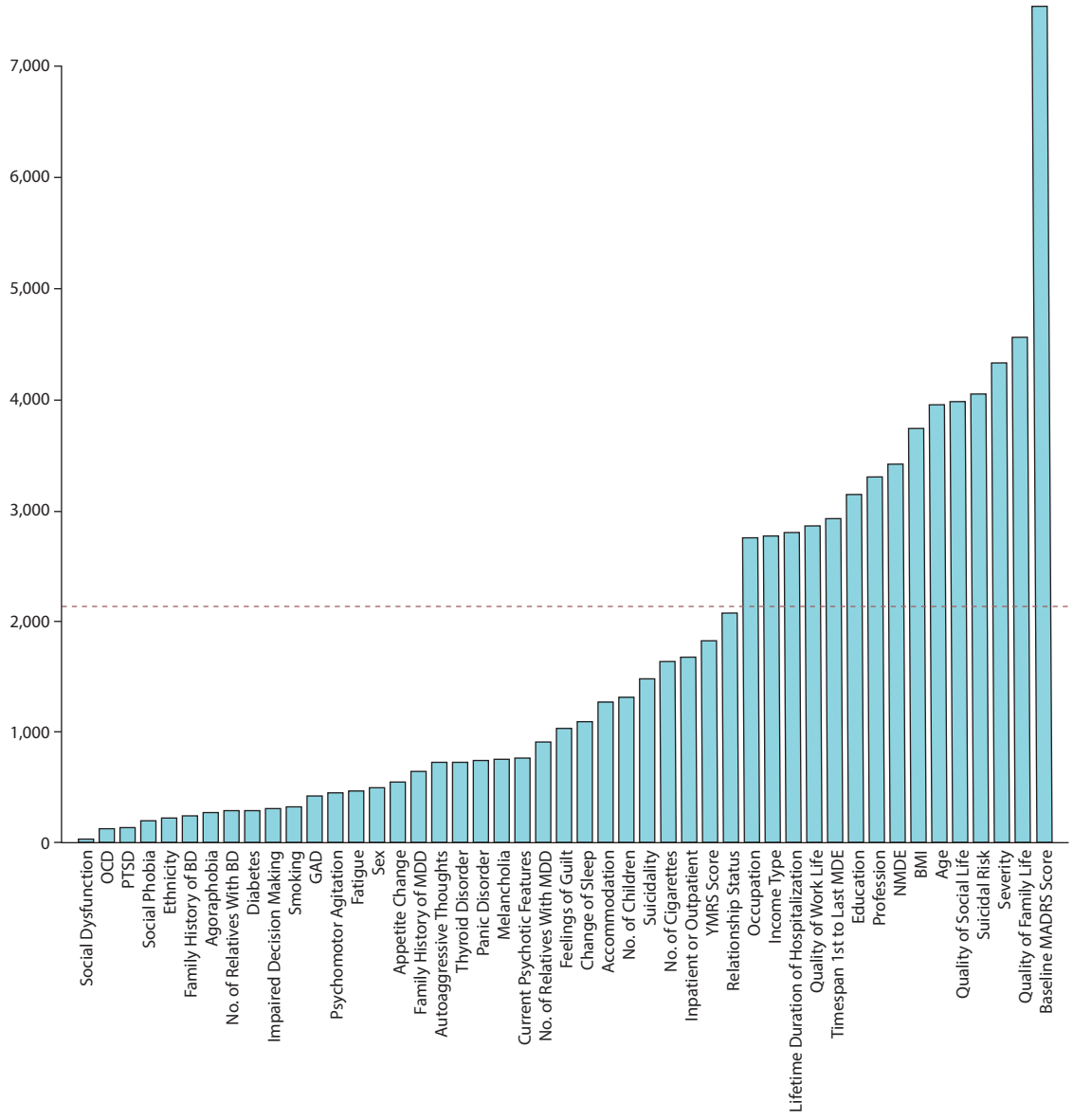
Importance values for evaluation of each of the predictors' contribution to the model highlighted baseline MADRS score, severity, suicidal risk, age, quality of family and social life, body mass index (BMI), lifetime number of major depressive episodes (NMDE), profession, occupation, income type, education, timespan between the first and last major depressive episode, quality of work life, and lifetime duration of hospitalization as the most influential predictors. As the error rate was increasing with reduction of predictors performed in a variable shrinkage analysis, we decided to keep all 47 predictors in our main model. Importance values measured by RSS values for all 47 predictors are portrayed in Figure 1.

Using all 47 predictors yielded the strongest results, with an accuracy of 75.0% for predicting TRD and treatment response. The sensitivity, representing the patients correctly predicted to have TRD, reached 82.2%, while the specificity was at 62.5% (Table 2). The positive predictive value (PPV) was at 79.6%, signifying the chance that a patient predicted to resist therapy will actually do so. The negative predictive value (NPV) reached 67.9%. Finally, our model yielded a false-positive rate (FPR), which describes the percentage of patients showing treatment response being misclassified as treatment-resistant, of 37.43%. In contrast, the false-negative rate, which describes the percentage of patients with TRD being misclassified as showing treatment response, was at 17.7%.

Regarding ANOVA, 12 factors showed association with TRD: severity, suicidality, poor quality of social and family life, comorbid panic disorder, high suicidal risk, family history of MDE, higher NMDE, higher YMRS score, appetite change, and feelings of guilt, as well as inpatient status (Table 3).

Relying on only the 15 most informative predictors resulted in the accuracy declining to 71.0%, while restricting the analysis to the 12 predictors showing association in

Figure 1. Average Importance Values for All 47 Predictors[a]



[a]Predictors are listed on the x-axis by increasing impact on prediction outcome, and mean residual sum of square (RSS) values are portrayed on the y-axis. Higher RSS signifies higher importance for predictive capability of the model. The red line indicates the cutoff of the 15 most valuable predictors, as featured in the simplified prediction model for enhanced clinical practicability.

Abbreviations: BD = bipolar disorder, BMI = body mass index, GAD = generalized anxiety disorder, MADRS = Montgomery-Asberg Depression Rating Scale, MDD = major depressive disorder, MDE = major depressive episode, NMDE = number of major depressive episodes, OCD = obsessive-compulsive disorder, PTSD = posttraumatic stress disorder, YMRS = Young Mania Rating Scale.

the ANOVA yielded an accuracy of 66.1%. While the PPV showed a robustness against predictor shrinkage (79.6% for all 47 predictors vs 80.1% for 15 predictors and 76.6% for 12 predictors), reduced efficiency of the simplified models was due to decline in NPV (67.9% vs 56.9% and 52.2%, respectively). A comprehensive depiction of model performance is provided by a receiver operating characteristic space diagram in Figure 2, plotting sensitivity against the FPR.

For the simplified models using 15 and 12 predictors, an additional validation of the prediction results was possible exploiting the set of patients excluded for the main analyses. One hundred nineteen patients and 88 patients that were excluded due to missing parameters for the full predictor set could be used as validation samples for the 15-predictor and 12-predictor analyses, respectively, as they showed full data availability for these reduced sets. Here, we could reach a comparable accuracy of 70.6% and 64.9% (Table 2 and Figure 2).

DISCUSSION

In this study, we strove for refinement and expansion of our previous endeavors by adopting the machine learning algorithm RF to a new sample that shows no overlap with previous investigations. We successfully established a

Table 2. Performance Measures for the Prediction Models[a]

Set of Predictors	Sensitivity	Specificity	FPR	PPV	NPV	Accuracy
All 47, 10-fold cross-validation	0.822	0.625	0.374	0.796	0.679	0.750
Top 15 RandomForest, 10-fold cross-validation	0.743	0.647	0.350	0.801	0.569	0.710
Top 15 RandomForest, validation set (n = 119)	0.803	0.603	0.396	0.819	0.603	0.706
Top 15 RandomForest, Elancourt (n = 60)	0.750	0.550	0.450	0.769	0.523	0.683
Top 15 RandomForest, Halle (n = 65)	0.725	0.600	0.400	0.743	0.576	0.676
12 ANOVA, 10-fold cross-validation	0.710	0.601	0.426	0.766	0.522	0.661
12 ANOVA, validation set (n = 88)	0.710	0.601	0.426	0.766	0.522	0.649

[a]Besides the main model featuring all predictors, we decided to implement clinically more accessible reduced feature sets. Thereby, constraint to the 15 most informative predictors from RandomForest caused a decline in accuracy to 0.710, while only fielding the predictors showing association in conventional analysis via analysis of variance (ANOVA) reached an accuracy of 0.661 in the cross-validated sample. For the reduced predictors sets, validation in a test sample of 119 (Top 15 RandomForest) and 88 (12 ANOVA) patients was performed. In addition, geographical data splitting as an alternative for cross-validation was performed with patients recruited in Elancourt and Halle as test samples. Besides accuracy as an overall measure of prediction efficacy, common evaluation parameters for binary classifiers are provided: sensitivity, specificity, false-positive rate (FPR), positive predictive value (PPV), and negative predictive value (NPV).

Table 3. ANOVA Results; Only Significant Predictors (n = 12) Are Shown and Ordered by *P* Value[a]

Predictor	TRD (n = 362) vs Response (n = 190)	*df*	Sum Square	Mean Square	*F* Value	*P* Value[b]
Severity	n: 258 vs 93	1	5.48	5.481	28.495	**< .0001**
NMDE	Mean: 4 vs 3	1	2.54	2.540	13.204	**.0002**
Quality of social life	Mean: 7.1 vs 6.1	1	2.70	2.705	14.062	**.0002**
Quality of family life	Mean: 6.8 vs 5.5	1	2.58	2.580	13.413	**.0003**
Suicidality	n: 208 vs 75	1	2.17	2.173	11.299	**.0008**
Inpatient status	n: 192 vs 78	2	1.54	0.771	4.010	***.0029***
Change in appetite	n: 284 vs 135	1	1.37	1.367	7.107	***.0079***
Panic disorder	n: 37 vs 12	1	1.27	1.266	6.582	***.0106***
Suicidal risk	Mean: 1.1 vs 0.6	1	1.22	1.223	6.356	***.0120***
Feelings of guilt	n: 316 vs 150	1	1.15	1.154	5.998	***.0146***
Family history of MDD	n: 174 vs 74	1	0.90	0.905	4.703	***.0186***
YMRS	Mean: 1.7 vs 0.7	1	1.03	1.0314	5.430	***.0202***

[a]Severity, number of depressive episodes, impairment of quality of social and family life, and suicidality showed the strongest associations, surviving correction for multiple comparison (corrected $P < .001$).
[b]Italics represent findings that did not withstand correction; boldface indicates significance after correction.
Abbreviations: ANOVA = analysis of variance, MDD = major depressive disorder, NMDE = number of depressive episodes over lifetime, TRD = treatment-resistant depression, YMRS = Young Mania Rating Scale.

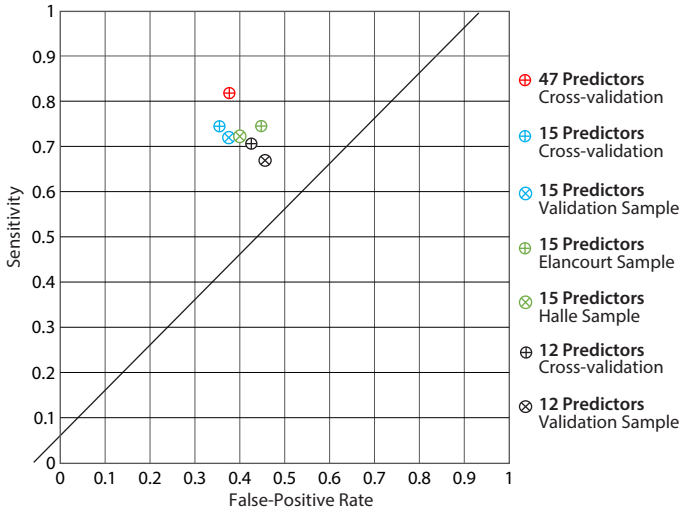
classification model that yielded an accuracy of 75.0% for forecasting treatment outcome using 47 predictors in 552 patients and that still showed an accuracy of 70.6% with a reduced set of 15 predictors. Thus, we outperformed our previous prediction model.

Machine learning techniques such as RF have shown promising results by considering interaction and main effects equally and producing reliable classification with high accuracy.[20] To our knowledge, besides our previous analyses in TRD and another study fielding a comparable technique in the Sequenced Treatment Alternatives to Relieve Depression sample, this is one of the first attempts to use new statistical approaches such as machine learning in TRD.[13] Predictors based on severity, education, relationship status, number of episodes, and suicidality showed significant influence on the prediction of outcome in these studies. Interestingly, posttraumatic stress disorder (PTSD), as well as ethnicity, were valued with high importance in some other models in TRD and MDD but failed to contribute much in this analysis.[32,33] This finding might be a result of the low frequency of comorbid PTSD in our sample (1.9%; percentages were not specified for the other studies), which also consists mainly of patients with Caucasian ethnicity.

Baseline MADRS score showed the strongest impact on prediction accuracy. This was expected, as baseline scores for determination of outcome evaluation have usually been shown to affect classification models in MDD and TRD.[32,34] Impairment of work, social, and family life as assessed by the SDS showed high importance as well. Considering that SDS score was shown to change in conformity with HDRS or MADRS scores in TRD and that subjectively high symptom burden was suggested as a risk factor for TRD independent of objective severity, this finding complements previous research.[35,36] Suicidal ideation, suicidal risk, and illness severity have consistently been associated with TRD,[16,37] as also observed in this study. Variables estimating the overall duration of illness, measured as the timespan between first and last MDE as well as the lifetime duration of hospitalization, ranked among the most informative predictors as well. On average, patients resistant to therapy showed a longer timespan between first and last MDE (11.69 vs 10.27 years). However, only NMDE showed an effect in the conventional ANOVA model out of these predictors, a replication of previous findings by our group and others.[38–40] Higher age has been associated with TRD by some studies[41–45] and was informative for the classification in this model. On the basis of the current data, predictors concerning time patterns of TRD seem to influence treatment outcome based on interactions. The association with clearly established predictors in TRD such as severity supports this proposition.

Concerning psychiatric comorbidities, panic disorder has repeatedly been shown to interfere with treatment response and was also linked to TRD in the conventional ANOVA model.[16,17] GAD, PTSD, social phobia, and agoraphobia failed to produce results in either model despite having previously been associated with TRD. However, with the exception of GAD and panic disorder, these comorbidities were rare in our sample.[15,46] This might be related to the focus on symptoms within the last month by the MINI, which does not account for lifetime symptoms except for panic disorder.

Figure 2. Receiver Operator Characteristics (ROC) for All Prediction Models[a]



[a]The sensitivity is portrayed on the y-axis, while the false-positive rate is scaled on the x-axis. The diagonal line indicates random guessing, meaning a prediction model with 50% accuracy. Optimal prediction quality would be achieved by an ROC in the upper left corner of the diagram. Here, the full model using all 47 predictors shows the strongest results, followed by the reduced 15-predictor model, while the analysis of variance–based 12-predictor model yielded the weakest predictive quality. For the 15- and 12-predictor models, additional validation in a previously untapped sample of 119 and 88 patients was possible, reaching comparable classification parameters as the cross-validation model. Finally, geographical data splitting with the patients deriving from Halle and Elancourt as test samples yielded slightly reduced accuracies for the 15-predictor model.

Regarding Axis III comorbidity, neither diabetes nor thyroid disorder showed strong contribution to the model, which conflicts with our last analysis.[22] The data on somatic disorders in TRD have been ambiguous, with few studies suggesting impact on TRD.[39,47–49] On the other hand, BMI, as a main parameter of obesity and associated diseases, yielded high importance estimation in this model.[50] On average, patients with TRD showed slightly higher BMI (26.29 vs 25.02). Previous research implicated no direct involvement of BMI in treatment response; however, linkage to risk for and severity of several comorbidities has been proposed.[51,52]

Concerning sociodemographic predictors, education, relationship status, profession, occupation, and income type were important contributors to our model. These predictors have all been investigated in the context of TRD before with ambiguous results.[44,53–55] Being in a relationship or married was suggested to increase one's chances of responding to treatment, while being divorced or widowed might constitute a higher risk of TRD. Lower income and being unemployed were often associated with MDD and have been implicated in retention of depressive states despite therapy.[56] On the other hand, suffering from an exhausting work schedule caused by high occupational level can also increase the risk for TRD.[57] However, in the single factor analysis, none of these predictors showed significant association, suggesting a complex interaction-based effect of these predictors.

Regarding predictors based on clinical symptoms, we addressed subsyndromal bipolarity by adding the YMRS score to our prediction model. Agitation and hidden bipolar features have often been discussed as major contributors to TRD.[58,59] However, despite single factor association of YMRS, the influence of this predictor as well as psychomotor agitation on the classification results was rather small.[46] The same was true for appetite change and feelings of guilt, which showed positive association with TRD but were not among the more informative variables for prediction.

Family history of unipolar or bipolar depression has previously been shown to increase the risk for MDD and has also been linked to treatment response in TRD.[17,60] In this model, their predictive importance was low; however, family history of MDD was positively associated with TRD in the ANOVA model.

It is important to note that we did not detect effects of sex on treatment outcome in TRD in these analyses, a finding that is in line with previous research in TRD.[15] Therefore, sex might not be involved in treatment response to antidepressants in TRD.

While some predictors showed clear preponderance in predictive information content, we have repeatedly concluded that models deteriorate with exclusion of predictors.[21,22] Here, as well as in our previous study in this field, using 47 predictors produced superior results to feature selection models excluding the lesser predictors. This contrasts the findings of other groups, which have suggested careful feature selection for maximizing effectiveness of prediction outcomes.[13,32] RF has been shown to exhibit a low vulnerability to overfitting when the recommended settings of trees to grow and variables chosen for each split are applied and sufficient training data are provided.[61,62] Therefore, we favor the consideration of all includable variables, adjusted for redundancy.

However, some limitations must be discussed in the context of this study. While machine learning algorithms such as RF can be expected to produce expedient results even when high numbers of predictors are fielded, the risk of false-positive results must be addressed. No independent data set was available for validation of our model. Therefore, our results cannot be generalized and are probably dependent on a narrow data context with regard to TRD staging methods, predictors classification, and sample selection by inclusion and exclusion criteria. Our use of a cross-validation design in a large data pool of 552 patients and our testing of the simplified prediction model also in 2 separate samples of patients not enrolled in the main analysis add weight to our results. As an alternative validation model, we performed a geographical split of the data to compare model performance between different centers. As some centers contributed only a small number of patients, we used patients from Elancourt and Halle (60 and 65, respectively) as a validation set and the rest of the patients for model generation with the 15 most informative predictors. Thereby, comparable accuracies of 67.67 and 68.34 could be achieved. However, we could not validate the prediction models

across TRD subdatabases due to differences in treatment outcome categorization and predictor recording. For TRD I and II, no baseline symptom score was registered, and treatment outcome was primarily defined by HDRS. The same holds true for several predictors that have been refined for TRD III on the basis of insight gained from our earlier results, for example, psychosocial predictors characterizing education and occupation or Sheehan and YMRS scores. Thus, we decided not to apply a validation using the whole GSRD data pool, as significant trade-off regarding quality of both predictor and outcome variables would have been inevitable. Hence, future studies by other groups will hopefully clarify if our findings are reproducible in other data sets.

Another important limitation is the dependence on retrospectively assessed variables. Especially the use of the baseline MADRS score for the calculation of treatment outcome phenotypes bears risks of patient as well as rater bias, as we did not assess these variables with outcome blinded. We did not find significant differences between baseline MADRS scores of the different outcome groups, and no significant contribution of baseline MADRS score was found in the conventional ANOVA model, indicating that considerable bias coming from retrospective assessment might be unlikely. Furthermore, this study did not have a tight treatment protocol beyond dosage and time thresholds as all patients were recruited in a retrospective cross-sectional approach. Consequently, the patients from our collective received a broad range of antidepressant as well as augmentation therapy. Most of them also received more than 1 antidepressant, and further stratification by therapeutic agent was discarded due to branching in subgroups too small for reliable interpretation.[23] In equal measure, only the simplified dichotomous outcome parameter treatment response and TRD was used, while nonresponse or further staging of TRD by number and type of antidepressant applied could not be implemented in our model. Developing prediction models for each antidepressant will be an important task for future studies as the predictive potential of certain variables for a specific antidepressant may better foster personalized medicine. On the other hand, our study quite likely depicts more real-life clinical settings, in which polypharmacy is common. Lastly, some predictors previously associated with TRD, such as duration of the current episode, were not implemented in our model as they would reduce the number of observation by more than 20% due to missing data.

In summary, we successfully elucidated a model based on clinical and sociodemographic variables with an accuracy of 75.0% for predicting treatment outcome. Thereby, the model is outperforming both clinical expert predictions and suggested classification tools based on neuropsychiatric scales or tools as electroencephalography and functional magnetic resonance imaging. To encourage testing of this model by other research groups and to increase practicability, we also implemented a prediction model using just 15 predictors with an accuracy of 71.0%. We show that variables that can easily be obtained in any clinical setting within a timeframe of approximately 10 minutes might be sufficient to markedly enhance assessment of treatment outcome. Our results encourage further studies utilizing multivariate approaches, as they emphasize that data mining and advanced statistics are auspicious in the quest for precision medicine in mental health.

Submitted: December 6, 2016; accepted October 11, 2017.

Published online: December 5, 2017.

Potential conflicts of interest: **Dr Dold** has received a travel grant from Janssen-Cilag. **Dr Spies** has received travel grants and/or reimbursement for congress participation from Janssen and AOP Orphan Pharmaceuticals, speaker honoraria from Janssen, and reimbursement for workshop participation from Eli Lilly. **Dr Souery** has received grant/research support from GlaxoSmithKline and Lundbeck and has served as a consultant or on advisory boards for AstraZeneca, Bristol-Myers Squibb, Eli Lilly, Janssen, and Lundbeck. **Dr Montgomery** has been a consultant or served on advisory boards for Forest, Johnson & Johnson, Lundbeck, Pierre Fabre, Pfizer, and Richter. **Dr Mendlewicz** is a member of the Board of the Lundbeck International Neuroscience Foundation and of the Advisory Board of Servier. **Dr Zohar** has received grant/research support from Lundbeck, Servier, and Pfizer; has served as a consultant or on advisory boards for Servier, Pfizer, Solvay, and Actelion; and has served on speakers bureaus for Lundbeck, GSK, Jazz, and Solvay. **Dr Serretti** is or has been consultant/speaker for Abbott, AstraZeneca, Clinical Data, Boehringer, Bristol-Myers Squibb, Eli Lilly, GlaxoSmithKline, Janssen, Lundbeck, Pfizer, Sanofi, and Servier. **Dr Lanzenberger** received travel grants and conference speaker honoraria from AstraZeneca, Lundbeck A/S, and Roche Austria GmbH. **Dr Kasper** received grant/research support, consulting fees, and/or honoraria within the last 3 years from Angelini, AOP Orphan Pharmaceuticals AG, AstraZeneca, Eli Lilly, Janssen, KRKA-Pharma, Lundbeck, Neuraxpharm, Pfizer, Pierre Fabre, Schwabe, and Servier. **Drs Kautzky**, **Bartova**, **Vanicek**, and **Fabbri** report no conflict of interest.

Funding/support: None.

Acknowledgments: The authors thank all persons involved in the GSRD project and the patients who made this analysis possible.

Supplementary material: See accompanying pages.

REFERENCES

1. Carvalho AF, Berk M, Hyphantis TN, et al. The integrative management of treatment-resistant depression: a comprehensive review and perspectives. *Psychother Psychosom*. 2014;83(2):70–88.
2. Rush AJ, Thase ME. Strategies and tactics in the treatment of chronic depression. *J Clin Psychiatry*. 1997;58(suppl 13):14–22.
3. Thase ME, Rush AJ. When at first you don't succeed: sequential strategies for antidepressant nonresponders. *J Clin Psychiatry*. 1997;58(suppl 13):23–29.
4. Nierenberg AA, Amsterdam JD. Treatment-resistant depression: definition and treatment approaches. *J Clin Psychiatry*. 1990;51(suppl):39–47, discussion 48–50.
5. Heimann H. Therapy-resistant depressions: symptoms and syndromes: contributions to symptomatology and syndromes. *Pharmakopsychiatr Neuropsychopharmakol*. 1974;7(3):139–144.
6. Berlim MT, Turecki G. Definition, assessment, and staging of treatment-resistant refractory major depression: a review of current concepts and methods. *Can J Psychiatry*. 2007;52(1):46–54.
7. Souery D, Papakostas GI, Trivedi MH. Treatment-resistant depression. *J Clin Psychiatry*. 2006;67(suppl 6):16–22.
8. Souery D, Amsterdam J, de Montigny C, et al. Treatment resistant depression: methodological overview and operational criteria. *Eur Neuropsychopharmacol*. 1999;9(1-2):83–91.
9. Papakostas GI, Ionescu DF. Towards new mechanisms: an update on therapeutics for treatment-resistant major depressive disorder. *Mol Psychiatry*. 2015;20(10):1142–1150.
10. Ferrari AJ, Charlson FJ, Norman RE, et al. Burden of depressive disorders by country, sex, age, and year: findings from the Global Burden of Disease Study 2010. *PLoS Med*. 2013;10(11):e1001547.
11. Whiteford HA, Degenhardt L, Rehm J, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet*. 2013;382(9904):1575–1586.
12. Whiteford HA, Harris MG, McKeon G, et al. Estimating remission from untreated major depression: a systematic review and meta-analysis. *Psychol Med*. 2013;43(8):1569–1585.
13. Perlis RH. Pharmacogenomic testing and personalized treatment of depression. *Clin Chem*. 2014;60(1):53–59.
14. Nierenberg AA, Katz J, Fava M. A critical overview of the pharmacologic management of

treatment-resistant depression. *Psychiatr Clin North Am*. 2007;30(1):13–29.
15. 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(7):453–468.
16. 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(7):1062–1070.
17. Balestri M, Calati R, Souery D, et al. Socio-demographic and clinical predictors of treatment resistant depression: a prospective European multicenter study. *J Affect Disord*. 2016;189:224–232.
18. Scarr E, Millan MJ, Bahn S, et al. Biomarkers for psychiatry: the journey from fantasy to fact, a report of the 2013 CINP Think Tank. *Int J Neuropsychopharmacol*. 2015;18(10):pyv042.
19. Roetker NS, Page CD, Yonker JA, et al. Assessment of genetic and nongenetic interactions for the prediction of depressive symptomatology: an analysis of the Wisconsin Longitudinal Study using machine learning algorithms. *Am J Public Health*. 2013;103(suppl 1):S136–S144.
20. Chen CC, Schwender H, Keith J, et al. Methods for identifying SNP interactions: a review on variations of Logic Regression, Random Forest and Bayesian logistic regression. *IEEE/ACM Trans Comput Biol Bioinformatics*. 2011;8(6):1580–1591.
21. 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(4):441–453.
22. Kautzky A, Baldinger-Melich P, Kranz GS, et al. A new prediction model for evaluating treatment-resistant depression. *J Clin Psychiatry*. 2017;78(2):215–222.
23. Dold M, Kautzky A, Bartova L, et al. Pharmacological treatment strategies in unipolar depression in European tertiary psychiatric treatment centers: a pharmacoepidemiological cross-sectional multicenter study. *Eur Neuropsychopharmacol*. 2016;26(12):1960–1971.
24. Sheehan DV, Lecrubier Y, Sheehan KH, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for *DSM-IV* and *ICD-10*. *J Clin Psychiatry*. 1998;59(suppl 20):22–33, quiz 34–57.
25. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry*. 1960;23(1):56–62.
26. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. *Br J Psychiatry*. 1979;134(4):382–389.
27. Young RC, Biggs JT, Ziegler VE, et al. A rating scale for mania: reliability, validity and sensitivity. *Br J Psychiatry*. 1978;133(5):429–435.
28. Sheehan DV, Harnett-Sheehan K, Raj BA. The measurement of disability. *Int Clin Psychopharmacol*. 1996;11(suppl 3):89–95.
29. Liaw A, Wiener M. Classification and regression by randomForest. *R News*. 2002;2(3):18–22.
30. Botta V, Louppe G, Geurts P, et al. Exploiting SNP correlations within random forest for genome-wide association studies. *PLoS One*. 2014;9(4):e93379.
31. Garcia-Magarinos M, Lopez-de-Ullibarri I, Cao R, et al. Evaluating the ability of tree-based methods and logistic regression for the detection of SNP-SNP interaction. *Ann Hum Genet*. 2009;73(pt 3):360–369.
32. Chekroud AM, Zotti RJ, Shehzad Z, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry*. 2016;3(3):243–250.
33. Perlis RH. A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. *Biol Psychiatry*. 2013;74(1):7–14.
34. Peeters FP, Ruhe HG, Wichers M, et al. The Dutch Measure for quantification of Treatment Resistance in Depression (DM-TRD): an extension of the Maudsley Staging Method. *J Affect Disord*. 2016;205:365–371.
35. De Berardis D, Marini S, Serroni N, et al. S-adenosyl-L-methionine augmentation in patients with stage II treatment-resistant major depressive disorder: an open label, fixed dose, single-blind study. *Scientific World Journal*;2013:204649.
36. Rane LJ, Fekadu A, Wooderson S, et al. Discrepancy between subjective and objective severity in treatment-resistant depression: prediction of treatment outcome. *J Psychiatr Res*. 2010;44(15):1082–1087.
37. Papakostas GI, Petersen T, Pava J, et al. Hopelessness and suicidal ideation in outpatients with treatment-resistant depression: prevalence and impact on treatment outcome. *J Nerv Ment Dis*. 2003;191(7):444–449.
38. Karp JF, Whyte EM, Lenze EJ, et al. Rescue pharmacotherapy with duloxetine for selective serotonin reuptake inhibitor nonresponders in late-life depression: outcome and tolerability. *J Clin Psychiatry*. 2008;69(3):457–463.
39. Sagud M, Mihaljevic-Peles A, Uzun S, et al. The lack of association between components of metabolic syndrome and treatment resistance in depression. *Psychopharmacology (Berl)*. 2013;230(1):15–21.
40. Maron E, Eller T, Vasar V, et al. Effects of bupropion augmentation in escitalopram-resistant patients with major depressive disorder: an open-label, naturalistic study. *J Clin Psychiatry*. 2009;70(7):1054–1056.
41. Bares M, Novak T, Kopecek M, et al. The early improvement of depressive symptoms as a potential predictor of response to antidepressants in depressive patients who failed to respond to previous antidepressant treatments: analysis of naturalistic data. *Eur Psychiatry*. 2012;27(7):522–527.
42. Bergman J, Miodownik C, Palatnik A, et al. Efficacy of bupropion XR in treatment-resistant elderly patients: a case series study. *Clin Neuropharmacol*. 2011;34(1):17–20.
43. Howland RH, Rush AJ, Wisniewski SR, et al. Concurrent anxiety and substance use disorders among outpatients with major depression: clinical features and effect on treatment outcome. *Drug Alcohol Depend*. 2009;99(1–3):248–260.
44. Nasso ED, Chiesa A, Serretti A, et al. Clinical and demographic predictors of improvement during duloxetine treatment in patients with major depression: an open-label study. *Clin Drug Investig*. 2011;31(6):385–405.
45. Petersen T, Hughes M, Papakostas GI, et al. Treatment-resistant depression and Axis II comorbidity. *Psychother Psychosom*. 2002;71(5):269–274.
46. 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(11):1357–1361.
47. Papakostas GI, Petersen T, Denninger J, et al. Somatic symptoms in treatment-resistant depression. *Psychiatry Res*. 2003;118(1):39–45.
48. Papakostas GI, Petersen T, Iosifescu DV, et al. Axis III disorders in treatment-resistant major depressive disorder. *Psychiatry Res*. 2003;118(2):183–188.
49. Degner D, Haust M, Meller J, et al. Association between autoimmune thyroiditis and depressive disorder in psychiatric outpatients. *Eur Arch Psychiatry Clin Neurosci*. 2015;265(1):67–72.
50. Yumuk V, Tsigos C, Fried M, et al. European guidelines for obesity management in adults. *Obes Facts*. 2015;8(6):402–424.
51. Mansoor B, Rengasamy M, Hilton R, et al. The bidirectional relationship between body mass index and treatment outcome in adolescents with treatment-resistant depression. *J Child Adolesc Psychopharmacol*. 2013;23(7):458–467.
52. Toups MS, Myers AK, Wisniewski SR, et al. Relationship between obesity and depression: characteristics and treatment outcomes with antidepressant medication. *Psychosom Med*. 2013;75(9):863–872.
53. Perlis RH, Alpert J, Nierenberg AA, et al. Clinical and sociodemographic predictors of response to augmentation, or dose increase among depressed outpatients resistant to fluoxetine 20 mg/day. *Acta Psychiatr Scand*. 2003;108(6):432–438.
54. Papakostas GI, Fava M, Thase ME. Treatment of SSRI-resistant depression: a meta-analysis comparing within- versus across-class switches. *Biol Psychiatry*. 2008;63(7):699–704.
55. Papakostas GI, Petersen T, Mischoulon D, et al. Functioning and interpersonal relationships as predictors of response in treatment-resistant depression. *Compr Psychiatry*. 2003;44(1):44–50.
56. Trivedi MH, Rush AJ, Wisniewski SR, et al. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. *Am J Psychiatry*. 2006;163(1):28–40.
57. Mandelli L, Serretti A, Souery D, et al. High occupational level is associated with poor response to treatment of depression. *Eur Neuropsychopharmacol*. 2016;26(8):1320–1326.
58. Iwanami T, Maeshima H, Baba H, et al. Psychomotor agitation in major depressive disorder is a predictive factor of mood-switching. *J Affect Disord*. 2015;170:185–189.
59. Rihmer Z, Dome P, Gonda X. Antidepressant response and subthreshold bipolarity in "unipolar" major depressive disorder: implications for practice and drug research. *J Clin Psychopharmacol*. 2013;33(4):449–452.
60. Serretti A, Chiesa A, Calati R, et al. Family history of major depression and residual symptoms in responder and non-responder depressed patients. *Compr Psychiatry*. 2014;55(1):51–55.
61. Gromski PS, Xu Y, Correa E, et al. A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data. *Anal Chim Acta*. 2014;829:1–8.
62. Liu X, Song M, Tao D, et al. Random forest construction with robust semisupervised node splitting. *IEEE Trans Image Process*. 2015;24(1):471–483.

Supplementary material follows this article.

THE JOURNAL OF CLINICAL PSYCHIATRY
THE OFFICIAL JOURNAL OF THE AMERICAN SOCIETY OF CLINICAL PSYCHOPHARMACOLOGY

Supplementary Material

Article Title: Refining Prediction in Treatment-Resistant Depression: Results of Machine Learning Analyses in the TRD III Sample

Author(s): Alexander Kautzky, MD; Markus Dold, MD, PhD; Lucie Bartova, MD; Marie Spies, MD; Thomas Vanicek, MD, PhD; Daniel Souery, MD, PhD; Stuart Montgomery, MD, Prof; Julien Mendlewicz, MD, PhD, Prof; Joseph Zohar, MD, Prof; Chiara Fabbri, MD, PhD; Alessandro Serretti, MD, PhD, Prof; Rupert Lanzenberger, MD, Prof; and Siegfried Kasper, MD, Prof

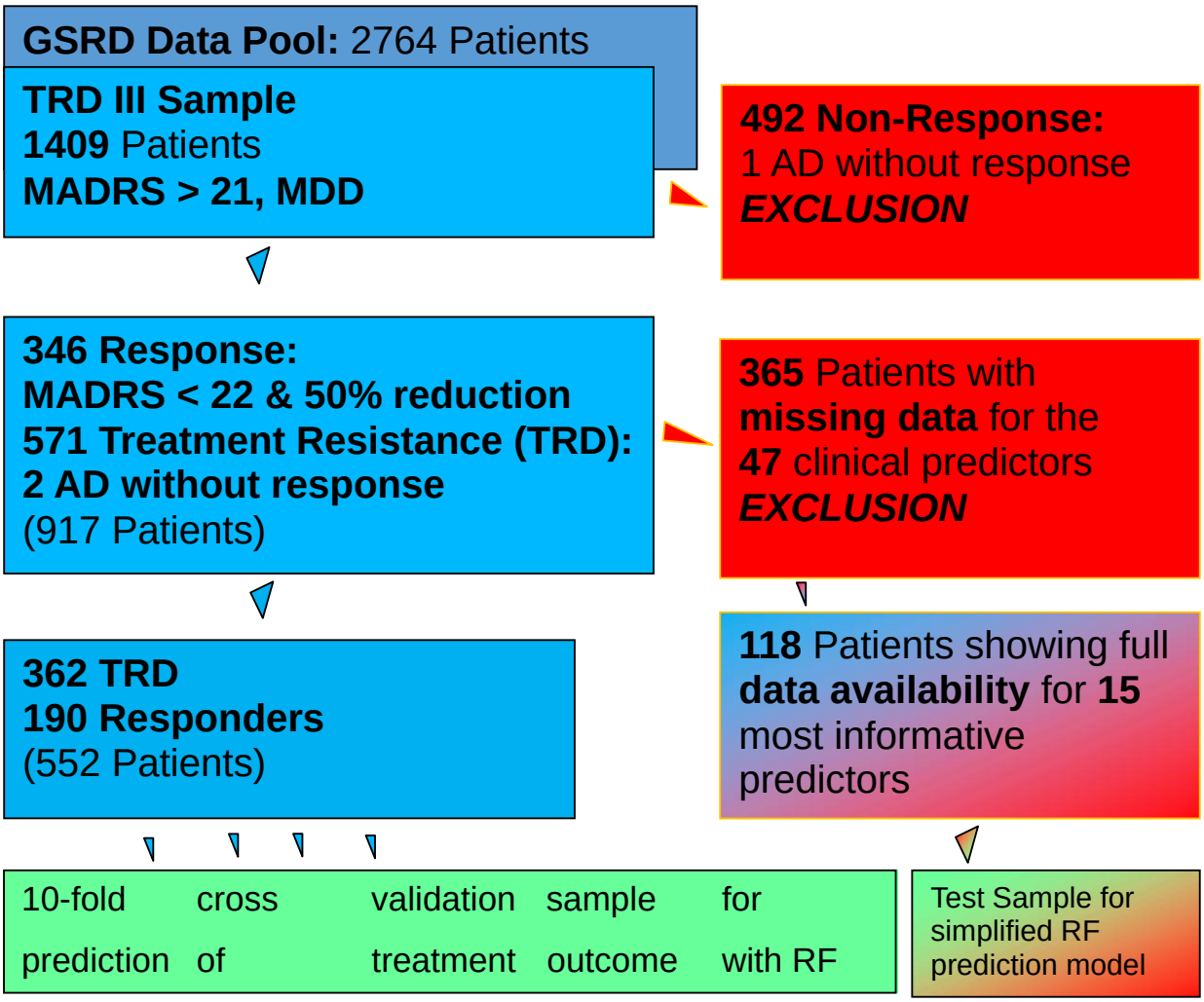
DOI Number: 10.4088/JCP.16m11385

List of Supplementary Material for the article

1. eFigure 1 Depiction of the Study Sample as Part of the Extensive Data Pool of the GSRD

Disclaimer

This Supplementary Material has been provided by the author(s) as an enhancement to the published article. It has been approved by peer review; however, it has undergone neither editing nor formatting by in-house editorial staff. The material is presented in the manner supplied by the author.



Supplementary eFigure1. Depiction of the study sample as part of the extensive data pool of the GSRD. For this analysis, the TRD III sample was used, comprising 1409 patients. Of those, 346 showed treatment response categorized by a MADRS reduction of at least 50% and final score below 22, while 571 patients did not achieve response and were determined as treatment resistant. Of those 917 patients, 552 showed full data availability for all 47 predictors and could hence be included in the prediction model. For model generation, 10-fold cross validation was used, cutting the sample into 10 subsamples and using one as a test set while the other nine are used as training set. This procedure is repeated for every subset so that every patient is predicted. For the simplified model with a reduced number of 15 predictors, 118 patients from the excluded sample could be used a validation sample for the prediction algorithm. Abbreviations: MADRS = Montgomery-**Å**sperg depression rating scale, GSRD = group for the studies of treatment resistant depression, RF = RandomForest, AD = antidepressive agent, TRD = treatment resistant depression.