

Predicting patients who will drop out of out-patient psychotherapy using machine learning algorithms

Björn Bennemann, Brian Schwartz, Julia Giesemann and Wolfgang Lutz

Background

About 30% of patients drop out of cognitive-behavioural therapy (CBT), which has implications for psychiatric and psychological treatment. Findings concerning drop out remain heterogeneous.

Aims

This paper aims to compare different machine-learning algorithms using nested cross-validation, evaluate their benefit in naturalistic settings, and identify the best model as well as the most important variables.

Method

The data-set consisted of 2543 out-patients treated with CBT. Assessment took place before session one. Twenty-one algorithms and ensembles were compared. Two parameters (Brier score, area under the curve (AUC)) were used for evaluation.

Results

The best model was an ensemble that used Random Forest and nearest-neighbour modelling. During the training process, it was significantly better than generalised linear modelling (GLM) (Brier score: $d = -2.93$, 95% CI $(-3.95, -1.90)$); AUC: $d = 0.59$, 95% CI $(0.11$ to $1.06)$). In the holdout sample, the ensemble was able to correctly identify 63.4% of cases of patients, whereas the GLM only identified 46.2% correctly. The most important predictors

were lower education, lower scores on the Personality Style and Disorder Inventory (PSSI) compulsive scale, younger age, higher scores on the PSSI negativistic and PSSI antisocial scale as well as on the Brief Symptom Inventory (BSI) additional scale (mean of the four additional items) and BSI overall scale.

Conclusions

Machine learning improves drop-out predictions. However, not all algorithms are suited to naturalistic data-sets and binary events. Tree-based and boosted algorithms including a variable selection process seem well-suited, whereas more advanced algorithms such as neural networks do not.

Keywords

drop out; machine learning; algorithms; ensembles; variable selection.

Copyright and usage

© The Author(s), 2022. Published by Cambridge University Press on behalf of the Royal College of Psychiatrists. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

Cognitive-behavioural therapy (CBT) is an effective treatment for mental health problems.¹ However, approximately one in five patients drop out of treatment,² leading to many problems including a lack of adequate treatment.^{3,4} Because of these negative consequences, identifying patients at a high risk of dropping out could lead to the development of clinical support tools that minimise the risk of drop out in individual patients.^{5,6} However, findings from past studies examining CBT treatments have been heterogeneous with only younger age and lower education level being consistently associated with drop out.^{7–9} Most studies used small samples and heterogeneous methods. An increase in statistical precision and large data-sets are therefore necessary to reliably identify patients at risk of dropping out of therapy.

Methodological developments

Over recent years, machine-learning approaches in particular have had a large impact on prediction modelling and on the most recent debate about the implementation of personalised or precision medicine concepts in mental health.^{10,11} Machine learning has been applied in various prediction contexts,^{12–15} taking advantage of the ability to capture non-linear relationships.¹⁶

Nevertheless, machine learning does not always have an advantage over more traditional methods,¹⁷ indicating that personalised medical care faces serious challenges that cannot be addressed through algorithmic complexity alone.¹⁸ It remains unclear which machine-learning methods are most suited to data from an out-patient CBT setting and whether previous findings can be generalised to this context.^{15,18} Further, to our knowledge, there is no study that has investigated the use of machine-learning algorithms for the

prediction of a binary event in a naturalistic setting. For this reason, we pursued two aims in this study.

- Various machine-learning algorithms will be systematically compared with regard to their personalised drop-out predictions and under routine care out-patient CBT conditions.
- Findings from these comparisons will be used to generate a clinically useful drop-out prediction model that can be used in clinical practice before the first session has occurred.

Method

Patients and treatment

The analyses were based on a sample comprising 2543 patients treated at the University of Trier out-patient CBT clinic in Southwest Germany between 2007 and 2021. Patients were included when they had completed a battery of questionnaires at intake, had begun therapy after the diagnostic phase (i.e. completed at least three sessions) and completed (i.e. consensual termination) or dropped out of treatment (see Supplementary Materials 1 available at <https://doi.org/10.1192/bjp.2022.17> for a flow chart of selected patients).

Written informed consent was obtained from all patients. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. All procedures involving human patients were approved by the ethics committee of the University of Trier.

All patient data collected from 2007 to 2017 were used for the model-generating process (training sample) and the remaining

data were used for testing purposes (holdout sample). Therapy took place once a week (range 3–113 sessions). When patients dropped out, the number of sessions was significantly lower than when they completed therapy (mean for those who dropped out 17.2 sessions; mean for those with completion 43.4 sessions; $t(2541) = 33.46$; $P < 0.001$; Cohen's $d = 1.49$). This held true for the training sample (mean for those who dropped out 17.5 sessions; mean for those with completion 44.3 sessions; $t(2041) = 31.36$; $P < 0.001$; Cohen's $d = 1.51$) and for the holdout sample (mean for those who dropped out 6.4 sessions; mean for those with completion 39.0 sessions; $t(498) = 12.51$; $P < 0.001$; Cohen's $d = 1.21$).

Diagnoses were based on the German version of the Structured Clinical Interview for Axis I DSM-IV Disorders—Patient Edition¹⁹ and the International Diagnostic Checklist for Personality Disorders.²⁰ Interviews were conducted by intensively trained independent clinicians before actual therapy began. All sessions were videotaped for establishing diagnoses and supervision; interviews and diagnoses were discussed in expert consensus teams that included four senior clinicians. Final diagnoses were determined by consensual agreement of at least 75% of the team members. For an overview of patient characteristics and differences between the groups see Supplementary materials 2.

The mean scores on the short-form of the Outcome Questionnaire²¹ and the Brief Symptom Inventory (BSI;²² German translation of Derogatis²³) were 1.90 (s.d. = 0.56) and 1.30 (s.d. = 0.71), respectively, indicating a moderate-to-severe general level of distress.

Therapists

Patients were treated by 220 therapists (173 female, 40 male, 7 unknown) who participated in a 3-year (full-time) or 5-year (part-time) postgraduate training programme with a CBT focus. Each therapist had at least 1 year of clinical training before beginning to treat patients. On average, therapists treated 11.6 patients each (s.d. = 6.2, range 1–26). Each therapist received 1 h of group or individual supervision on a monthly basis. The session videos were used for supervision and research. Supervisors were senior clinicians with at least 5 years of clinical experience after completing training. In treatment, therapists scored a mean of 3.81 (s.d. = 0.89) on the 'overall adherence item' of the Inventory of Therapeutic Interventions and Skills that has a range from 0 to 6.²⁴ For this reason, adherence can be described as adequate.

Measures

Drop out

Drop out was assessed via clinical judgement at the end of treatment. When the patient and therapist agreed on a consensual end of therapy, the treatment was considered regularly completed. In contrast, when the patient stopped coming to therapy, despite the therapist's appraisal that more sessions were necessary the form of termination was considered as a drop out. Examples of this operationalisation of drop out include the patient stopped coming to sessions and was unable to be reached by phone or email or the patient told the therapist that they would no longer be coming to therapy anymore, despite the therapist's advice to continue therapy.

Intake variables

A total of 77 variables measured at intake (i.e. before the first session) were included in the analyses. Table 1 shows all 77 variables as well as the mean differences. All variables were assessed via questionnaires.

Selection of machine-learning algorithms

In order to get an accurate picture of common algorithms used in sociological/ scientific/ medical contexts, we decided to use and compare only those algorithms that have already found application

in the relevant literature. For this purpose, we particularly focused on the Stratified Medicine Approaches for Treatment Selection Mental Health Prediction Tournament at the 2019 Treatment Selection Idea Lab conference, in which 13 different research groups developed different prediction models using the same data-set (for a further review, see Cohen et al³²). Using the information from this tournament, as well as an examination of the literature provided by the tournament organiser, we selected a total of 21 algorithms for closer examination (see Table 2). As we aimed to compare different algorithms regardless of them being linear or non-linear, we decided to include linear algorithms alongside the machine-learning algorithms, as suggested by Brownlee.¹⁶

Data analytic strategy

Data preparation

All analyses were conducted using the free software environment R version 4.1.1.³³ No variables that had more than 10% missing values were included in the analyses. Therefore, we had to exclude a total of five variables (total scores of the Patient Health Questionnaire (PHQ-9), Affective Style Questionnaire (ASQ), Generalised Anxiety Disorder Assessment (GAD-7), Ten Item Personality Measure (TIPI) and Work and Social Adjustment Scale (WSAS)). No patient was excluded from the analyses because of too many missing values. Variables with less than 10% missing values were imputed using a trained Random Forest in the R package missForest v1.4.³⁴ The imputations for the training and holdout samples were conducted separate from the cross-validation framework before the actual analyses.

For model training, we used the R-packages caret v6.0-90³⁵ and caretEnsemble v2.0.1.³⁶ These packages tune the hyperparameters to their optimal settings depending on which one is being used. To ensure a fair comparison of the algorithms, we did not change the packages' default settings. Table 2 shows the algorithms used and the different tuning parameters that were tested (for a further review see Kuhn³⁷). Identification of the best model was always based on the receiver operating characteristic curve. The model with the largest area under the curve (AUC) was considered the best model. All models predicted drop out as a binary event (drop out versus non-drop out).

Ranking and correlation of algorithms

First, we ranked all individual algorithms based on the two parameters (i.e. Brier (for a description see below), AUC) and compared the correlations of the predictions of all algorithms during the model-building process using the corresponding function in the Caret package. For this purpose, we conducted a nested cross validation with 20 outer and 10 inner loops according to Brownlee's³⁸ recommendations. All continuous variables were centered separately for each outer cross-validation loop of the training and test sets. Subtrahend was always the mean value from the training data of the respective variable to avoid data leakage and to ensure appropriate data preparation for the algorithms.¹⁶ Drop out was dichotomised with 1 (drop out) and 0 (regular termination). Subsequently, each of the 21 machine-learning algorithms generated a drop-out prediction model based on each outer and inner cross-validation training set that was then evaluated in the respective outer or inner cross-validation test set to minimise overfitting³⁹ and the influence of sample characteristics. For the inner cross validation, we also applied a sampling method (synthetic minority oversampling technique; SMOTE)⁴⁰ to address the problem of class imbalance.^{41,42} SMOTE is a hybrid method combining up and down sampling. It artificially generates new examples of the minority class using the nearest neighbours of these cases.

Table 1 Predictors used for model generation. Predictors were routinely collected at intake^a

Variables	Training sample, patients who drop-out versus regularly completed		Holdout sample, patients who drop-out versus regularly completed		Training sample versus holdout sample	
	t-test/ χ^2	P	t-test/ χ^2	P	t-test/ χ^2	P
Male gender	-0.67	0.41	1.89	0.17	1.62	0.20
High education ^b	-29.83	<0.001	-12.16	<0.001	2.11	0.15
Middle education ^b	0.08	0.78	0.16	0.69	-0.62	0.43
Sick leave	6.03	<0.05	0.01	0.93	-2.82	0.09
Children	0.31	0.58	-0.00	0.95	-5.40	<0.05
Marital status	-8.54	<0.01	-1.96	0.16	-3.06	0.08
Medication intake	-1.35	0.25	-0.00	1.00	-2.15	0.14
Age	-3.75	<0.001	-1.70	0.09	-0.46	0.65
Outcome Questionnaire (OQ) – Total score	3.75	<0.001	2.12	<0.05	1.02	0.31
OQ – Symptom distress	3.14	<0.01	2.02	<0.05	1.49	0.14
OQ – Social role functioning	1.31	0.19	0.69	0.49	0.72	0.47
OQ – Interpersonal relationship	5.75	<0.001	2.54	<0.05	-0.70	0.48
Questionnaire for the Evaluation of Psychotherapy (FEP2) ²⁵ – Total score	3.20	<0.01	2.38	<0.05	0.08	0.93
FEP2 – Well-being	1.80	0.07	1.93	0.05	1.40	0.16
FEP2 – Discomfort	3.54	<0.001	2.87	<0.01	0.99	0.32
FEP2 – Incongruence	3.60	<0.001	2.66	<0.01	0.03	0.97
FEP2 – Interpersonal	1.92	0.06	0.89	0.37	-1.66	0.10
Emotionality Inventory (EMI) ²⁶ – Total score	2.71	<0.01	1.45	0.15	0.65	0.52
EMI – Anxiety	1.75	0.08	2.39	<0.05	-0.85	0.40
EMI – Depression	3.42	<0.001	1.63	0.10	1.71	0.09
EMI – Inhibition	0.81	0.42	0.38	0.71	-0.64	0.52
EMI – Security	3.09	<0.01	1.58	0.12	-0.16	0.88
EMI – Well-being	2.71	<0.01	0.23	0.41	0.68	0.50
Brief Symptom Inventory (BSI) – Total score ^c	5.92	<0.001	2.96	<0.01	0.13	0.90
BSI – Somatic problem	4.12	<0.001	2.61	<0.01	-0.04	0.97
BSI – Obsessive compulsive	2.86	<0.01	1.22	0.22	0.28	0.78
BSI – Uncertainty	4.32	<0.001	1.94	0.05	-0.57	0.57
BSI – Depression	5.09	<0.001	2.17	<0.05	0.76	0.45
BSI – Anxiety	3.72	<0.001	2.67	<0.01	-0.31	0.76
BSI – Hostility	5.25	<0.001	2.79	<0.01	-0.92	0.36
BSI – Phobia	4.32	<0.001	2.42	<0.05	0.65	0.51
BSI – Paranoid	5.87	<0.001	3.05	<0.01	-1.58	0.11
BSI – Psychoticism	5.16	<0.001	1.94	0.05	1.36	0.17
BSI – Additional	6.81	<0.001	3.07	<0.01	1.58	0.11
Interpersonal Problems (IIP32) ²⁷ – Total score	1.86	0.06	0.97	0.33	-0.39	0.70
IIP – Autocratic/ dominant	4.82	<0.001	2.60	<0.01	-1.86	0.06
IIP – Confrontational	3.19	<0.01	2.06	<0.05	-0.65	0.51
IIP – Unapproachable	2.86	<0.01	2.01	<0.05	1.01	0.31
IIP – Introverted	1.58	0.11	0.81	0.42	-0.33	0.74
IIP – Submissive	-2.06	<0.05	-1.87	0.06	0.19	0.85
IIP – Exploitable	-2.95	<0.01	-1.91	0.05	-0.24	0.81
IIP – Caring	1.40	0.16	1.44	0.15	0.11	0.92
IIP – Expressive	1.16	0.25	-0.54	0.59	-0.28	0.78
Incongruence Questionnaire (INK23) ²⁸ – Total score	3.47	<0.001	1.43	0.15	-1.08	0.28
INK – Approach	2.74	<0.01	1.42	0.16	0.03	0.97
INK – Avoidance	3.78	<0.001	1.27	0.21	-2.26	<0.05
Dysfunctional Attitudes Scale – short-form (DASK) ²⁹ – Total score	2.42	<0.05	0.92	0.36	-0.38	0.70
DASK – Recognition	0.36	0.72	1.05	0.30	-1.37	0.17
DASK – Performance	2.91	<0.01	0.72	0.47	-0.08	0.94
Inventory of Stressful Events (ILE) ³⁰ – Score for number of events	3.19	<0.01	1.61	0.11	-2.31	<0.05
ILE – Score for stress	3.18	<0.01	1.41	0.16	-2.64	<0.01
ILE – Number in patient's life	3.54	<0.001	2.24	<0.05	-1.20	0.23
ILE – Number of events in close relationships	-1.55	0.12	-1.04	0.30	-0.22	0.83
ILE – Number of events in distant relationships	-3.86	<0.001	-2.11	<0.05	5.24	<0.001
General Perceived Self-Efficacy Scale ^d	-0.26	0.79	-0.80	0.43	0.90	0.37
Personality Style and Disorder Inventory – short-form (PSSIK) ³¹ – Antisocial	4.49	<0.001	3.16	<0.01	-0.08	0.93
PSSIK – Paranoid	6.07	<0.001	2.47	<0.05	-1.63	0.10
PSSIK – Schizoid	3.23	<0.01	0.74	0.46	-0.17	0.87
PSSIK – Avoidant	0.34	0.74	-1.13	0.26	-0.67	0.50
PSSIK – Compulsive	-4.66	<0.001	-1.52	0.13	-0.80	0.42
PSSIK – Schizotypal	1.42	0.16	-0.05	0.96	-1.71	0.09
PSSIK – Rhapsodic	-0.63	0.53	1.15	0.25	0.28	0.78
PSSIK – Narcissistic	1.44	0.15	0.94	0.35	0.79	0.43
PSSIK – Negativistic	6.06	<0.001	2.85	<0.01	-2.30	<0.05
PSSIK – Dependent	4.23	<0.001	1.93	0.05	-1.15	0.25
PSSIK – Borderline	4.51	<0.001	2.19	<0.05	0.33	0.74

(Continued)

Table 1 (Continued)

Variables	Training sample, patients who drop-out versus regularly completed		Holdout sample, patients who drop-out versus regularly completed		Training sample versus holdout sample	
	<i>t</i> -test/ χ^2	<i>P</i>	<i>t</i> -test/ χ^2	<i>P</i>	<i>t</i> -test/ χ^2	<i>P</i>
PSSIK – Histrionic	3.43	<0.001	2.28	<0.05	-1.42	0.15
PSSIK – Depressive	4.23	<0.001	1.63	0.10	-0.11	0.91
PSSIK – Altruistic	1.88	0.06	1.83	0.07	-0.04	0.97
Patient-rated well-being ^d	-2.91	<0.01	-1.70	0.09	-0.90	0.37
Current emotional and psychological functioning ^d	-3.05	<0.01	-2.69	<0.01	1.92	0.06
Therapy Expectations – Importance of psychotherapy ^d	-1.70	0.09	0.64	0.53	0.03	0.97
Therapy Expectations – Difficulties attending psychotherapy ^d	-1.93	0.05	1.58	0.11	-1.75	0.08
Therapy Expectations – Confidence in the helpfulness of psychotherapy ^d	-2.82	<0.01	-3.22	<0.01	-0.64	0.52
Therapy Expectations – Amount of previous psychotherapy ^d	-0.09	0.93	0.67	0.50	2.49	<0.05
Therapy Expectations – Chronicity of the problem ^d	1.37	0.17	2.10	<0.05	-0.24	0.81
Therapy Expectations – Estimated future coping ^d	-2.22	<0.05	-1.47	0.14	-0.47	0.64

a. Negative values indicate a negative correlation with the drop-out variable or a higher value/ratio in the holdout sample. For dichotomous variables (first seven variables) a χ^2 -test was used, for continuous variables, a *t*-test was used.
b. High education, university entrance qualification; middle school, middle school graduation.
c. The total score of the BSI additional scale is the mean of the four additional items of the BSI.
d. This item was used as a single question and has no scale.

Furthermore, the majority class examples are also undersampled, leading to a more balanced data-set. Then, a performance ranking based on the Brier score and the AUC was generated as well as the correlation matrix for all algorithms.

Brier score

The Brier score ranges from 0 (best prediction) to 1 (worst prediction) by measuring probabilistic predictions.⁴³ Thus, it takes the certainty of the prediction into account. In effect, it is the mean squared error of the forecast:

$$\frac{1}{N} \sum_{t=1}^N (f_t - o_t)^2$$

Hereby, *N* is the total number of observations, *f* is the probability of the event (i.e. drop out) and *o* is the actual outcome (i.e. 0 or 1) of the event at instance *t*.

AUC

The AUC uses the sensitivity and specificity of a prediction and ranges from 0 (worst prediction) to 1 (best prediction). Based on signal detection theory,⁴⁴ the AUC takes the base rate of the dependent variable into account.

Ensembles

We used the ranking of single algorithms and the correlation matrix to generate ensembles. Ensembles show better performance and greater robustness in certain contexts by reweighting the results of different algorithms, which can produce better overall results.⁴⁵ We decided to use five types of ensembles. The two and three best algorithms, the two and three least correlating algorithms and the best algorithm with the respective least correlating algorithm. The idea of merging algorithms with low correlations is that they probably assess different aspects of the data-set.⁴⁶ Therefore, it is possible that an ensemble of such algorithms improves the prediction significantly, even though one algorithm makes poor predictions on its own. These ensembles were merged either via a generalised linear modelling (GLM) algorithm or via the best algorithm across both parameters (i.e. Brier score and AUC) according to our ranking using the stacking method. Again we used Caret with its default settings to create an ensemble with the best parameters. In total, we generated ten ensembles (five types of ensemble \times two ways of merging).

Comparing ensembles and single algorithms

Next, we compared all ten ensembles and the five best single algorithms. Again, we used a nested cross validation as described above. However, this time we used a ten-fold inner cross validation with three repetitions. Repeating the cross validation leads to a more precise result,⁴⁷ so we conducted this procedure for a more adequate comparison.

Extending the procedure

In order to gain a more comprehensive picture, we repeated the entire procedure twice. For the first repetition, we only used the significant predictors (i.e. initial impairment, male gender, lower education status, more histrionic and less compulsive personality style and negative treatment expectations) from Zimmermann et al.⁸ Thus, we evaluated the changes in the prediction when using these relevant predictors only. For the second repetition, we performed variable selection using an elastic net regularisation with the Caret package for the training set after each split. As we examined a large set of variables, we evaluated whether some models improve, when preceded by variable selection. This was done 20 times for each training set of the outer loops inside the cross-validation framework, preventing data leakage from the test set. For this elastic net selection after each split, we did not use Caret to choose the optimal setting, but set alpha to 0.1 for the first analysis and then altered alpha in increments of 0.1 until 1 was reached. An alpha of 0 is equal to a ridge regression, whereas an alpha of 1 equals a least absolute shrinkage and selection operator regression. We also defined lambda's range analogue to the alpha parameter. Lambda defines the magnitude of the regression penalty. This resulted in 100 different possible combinations of these two parameters (ten values for alpha \times ten values for lambda) to identify the best fitting model. Identification of the best model was always based on the AUC. The model with the highest value was considered the best model. At the end of the second repetition, we only included the predictors that had predictive power in the best model.

Conducting this entire procedure three times (with all variables, with only seven predictors, and with variables that had predictive power in the preceding elastic net analysis) led to a total of 30 ensembles (10 ensembles \times 3 procedures) and 15 single algorithms (5 single algorithms \times 3 procedures). Each ensemble and single algorithm generated a model via a nested cross validation with 20 outer loops and 10 inner loops with three repetitions. The model

Table 2 Classification of all machine learning algorithms¹⁶ that were used in this study^a

Category and algorithm	Tuning parameters
Regression ^b	
Generalised Linear Model (GLM)	–
GLM with stepwise feature selection using AIC (GLMAIC)	–
Bayesian	
Bayesian GLM (BAYESGLM)	–
Naïve Bayes	usekernel(y/n); laplace correction = 0; bandwidth adjustment = 1
Decision Tree	
C4.5-like Trees (C4.5)	confidence threshold (0.01; 0.255; 0.5); minimum instances per leaf (1; 2; 3)
Conditional Inference Trees (CTREE)	max tree depth(1; 2; 3); mincriterion(0.01; 0.5; 0.99)
Artificial neural networks	
Feed-Forward Neural Network with single hidden layer (NNET)	number of hidden units(1; 3; 5); decay(0; 0.1; 0.0001)
Averaged feed-forward Neural Network with single hidden layer over different seeds (AVNN)	size(1; 3; 5); decay(0; 0.1; 0.0001); bagging = FALSE
Monotone Multi-Layer Perceptron Neural Network (MONMLP)	number of hidden units(1; 3; 5); number of models = 1
Dimensionality reduction	
Linear Discriminant Analysis (LDA)	–
Regularisation	
Elastic net	alpha(0.1; 0.55; 1); lambda(0.0001; 0.001; 0.01)
Instance-based	
K-fold-nearest-neighbors (kNN)	Number of nearest neighbors(5; 7; 9)
Support Vector Machines (SVM)	cost(0.25; 0.5; 1);
Ensembles	
Stochastic Gradient Boosting (GBM)	max tree depth(1; 2; 3); #boosting iterations(50; 100; 150); n.minobsinnode = 10; shrinkage = 0.1
Boosted Logistic Regression (LOGIT)	niter(11; 21; 31)
Extreme Gradient Boosting (XGB)	shrinkage(0.3; 0.4); max tree depth(1; 2; 3); colsample_bytree(0.6; 0.8); subsample(0.5; 0.75; 1); number of boosting iteration(50; 100; 150); minimum loss reduction = 0; min_child_weight = 1;
Random Forest	number of randomly selected predictors (2; 4; 7)
Bagged Multivariate Adaptive Regression Splines (MARS)	number of terms(2; 8; 14); degree = 1
Bagged Classification and Regression Tree (CART)	–
Boosted Classification Trees (ADA)	max tree depth(1; 2; 3); number of trees(50; 100; 150); learning rate = 0.1
Boosted Generalised Linear Model (GLMBOOST)	number of boosting iterations(50; 100; 150)
y, yes; n, no.	
a. The numbers in the square brackets indicate the different tuning parameters tested using the R package Caret.	
b. Categories are shown followed by the respective algorithms.	

with the best mean prediction scores across all cross validations and across both parameters was chosen. Generating 20 models via the outer cross validations resulted in one distribution consisting of 20 Brier scores and one distribution consisting of 20 AUC scores for each algorithm/ensemble. In order to quantitatively compare the differences and distributions as well as the robustness against sampling artefacts, *t*-tests between the best and worst model as well as between the best model and a single GLM were conducted for each parameter.

For a final test we used the best ensemble/algorithm and let it generate a model with the whole training sample via a ten-fold cross validation with three repetitions. This model was then tested in the still unused and independent holdout sample to assess the generalisability of the model and to prevent overfitting.

Last, the holdout sample's confusion matrix was examined in order to assess the improvement of the prediction. Therefore, each individual that had a higher risk than the mean of the training sample to drop out of therapy (i.e. 30.6%) was considered a predicted 'dropout case'. Finally, the Caret package was used to determine the most important variables.

Results

After the first step, the algorithm with the best predictions when using all variables was Stochastic Gradient Boosting. When only using predictors that showed predictive power in a preceding elastic net analysis, Random Forest was the best algorithm.

Boosted Classification Trees (ADA) made the best predictions when only the seven significant predictors were used (see Supplementary Materials 3 for an overview of all algorithms). Especially boosting and tree-based approaches seemed to make the best predictions. Further, algorithms from different classes seemed to correlate the least with each other (see Supplementary Materials 4 for the low correlating algorithms).

Next, by using the rankings (Supplementary Materials 3) and correlations (Supplementary Materials 4), we generated the ensembles as described above for the final analyses. Comparing the different algorithms and ensembles, the best model across both parameters was generated by an ensemble with the best machine-learning algorithm and its least correlating algorithm (i.e. Random Forest and K-Fold-Nearest-Neighbors (kNN)) that was merged via a GLM and had a preceding elastic net variable selection (Brier score 0.1983, AUC = 0.6581). Table 3 provides an overview of all algorithms and ensembles.

The distributions of each algorithm/ensemble revealed that the best ones hardly differed from each other (see Fig. 1). Nevertheless, some models seemed to make significantly worse predictions. For the Brier score, the pattern was very similar.

As a result of these distributions, we were able to compare model accuracy/robustness via *t*-tests. A paired one-sided *t*-test revealed a highly significant effect between the overall best and overall worst models concerning the AUC score (AUC_{best} = 0.6581; AUC_{worst} = 0.5465; *t*(19) = 8.30, *P* < 0.001, Cohen's *d* = 1.86, 95% CI (0.11, 2.58)). Comparing the overall best model

Table 3 Mean scores of the models generated by all 45 algorithms and ensembles^a

Algorithm/ensemble	Stacking method	Variables used	Brier score	AUC	Training AUC
Best with lowest correlation	GLM	Selected with elastic net	0.1983	0.6581	0.6617
Two best	GLM	Selected with elastic net	0.1983	0.6577	0.6674
Three best	GLM	Selected with elastic net	0.1985	0.6535	0.6673
Two best	GBM	All	0.1989	0.6550	0.6515
Three best	GLM	All	0.1994	0.6513	0.6549
Best with lowest correlation	GLM	All	0.1992	0.6497	0.6492
Two Best	GLM	All	0.1995	0.6518	0.6530
Three best	GBM	All	0.1998	0.6523	0.6557
GBM	–	Selected with elastic net	0.2022	0.6661	0.6608
Two best	GLM	Manually selected	0.1995	0.6493	0.6464
Random Forest	–	Selected with elastic net	0.2041	0.6605	0.6602
Best with lowest correlation	GLM	Manually selected	0.1997	0.6488	0.6430
Best with lowest correlation	GBM	All	0.2004	0.6506	0.6483
Three best	GLM	Manually selected	0.1998	0.6468	0.6461
Three least correlating	GLM	All	0.2010	0.6435	0.6494
Three least correlating	GBM	All	0.2006	0.6412	0.6485
Three best	ADA	Manually selected	0.2011	0.6435	0.6488
ADA	–	All	0.2071	0.6525	0.6485
GBM	–	All	0.2055	0.6457	0.6497
Best with lowest correlation	ADA	Manually selected	0.2017	0.6403	0.6424
XGB	–	Selected with elastic net	0.2069	0.6480	0.6584
Two best	ADA	Manually selected	0.2014	0.6349	0.6482
Random Forest	–	All	0.2058	0.6392	0.6428
ADA	–	Selected with elastic net	0.2099	0.6475	0.6591
XGB	–	All	0.2075	0.6448	0.6451
Two least correlating	GLM	All	0.2049	0.6197	0.6129
Three least correlating	GLM	Manually selected	0.2053	0.6193	0.6095
Two least correlating	GLM	Manually selected	0.2056	0.6143	0.6060
GBM	–	Manually selected	0.2208	0.6525	0.6369
GLMBOOST	–	Selected with elastic net	0.2309	0.6408	0.6516
Three least correlating	ADA	Manually selected	0.2059	0.6066	0.6150
Two least correlating	ADA	Manually selected	0.2064	0.6087	0.6092
Two least correlating	GBM	All	0.2060	0.6010	0.6121
ADA	–	Manually selected	0.2240	0.6349	0.6440
GLMBOOST	–	Manually selected	0.2342	0.6364	0.6379
GLMBOOST	–	All	0.2306	0.6349	0.6487
Two least correlating	GLM	Selected with elastic net	0.2064	0.5971	0.5872
LDA	–	Manually selected	0.2347	0.6364	0.6377
Three least correlating	GLM	Selected with elastic net	0.2074	0.5986	0.6058
Three best	Random Forest	Selected with elastic net	0.2180	0.6085	0.6143
GLMAIC	–	Manually Selected	0.2342	0.6342	0.6392
Two best	Random Forest	Selected with elastic net	0.2376	0.5893	0.5902
Three least correlating	Random Forest	Selected with elastic net	0.2490	0.5661	0.5607
Best with lowest correlation	Random Forest	Selected with elastic net	0.2586	0.5864	0.5838
Two least correlating	Random Forest	Selected with elastic net	0.2859	0.5465	0.5489

AUC, area under the curve; GLM, Generalised linear model; ADA, Boosted Classification Trees; XGB, Extreme Gradient Boosting; GBM, Stochastic Gradient Boosting; GLMBOOST, boosted generalised linear model; LDA, Linear Discriminant Analysis; GLMAIC, GLM with stepwise feature selection using Akaike information criterion.
a. All ensembles and algorithms are ranked.

with the models of a GLM using all variables, the effect was still significant ($AUC_{best} = 0.6581$; $AUC_{GLM} = 0.6253$; $t(19) = 2.63$, $P < 0.01$, Cohen's $d = 0.59$, 95% CI (0.11, 1.06)). For the Brier score, the effects were also significant when comparing the best with the worst model ($Brier_{best} = 0.1983$; $Brier_{worst} = 0.2859$; $t(19) = -13.03$, $P < 0.001$, Cohen's $d = -2.91$ 95% CI (-3.92, -1.89)) and when comparing the best model with the GLM model using all variables ($Brier_{best} = 0.1983$; $Brier_{GLM} = 0.2384$; $t(19) = -13.11$, $P < 0.001$, Cohen's $d = -2.93$ 95% CI (-3.95, -1.90)). All boxplots are shown in Supplementary Materials 5.

Before the first session occurred the best model was able to identify 63.4% of all holdout cases of patients dropping out correctly (the confusion matrix is shown in Supplementary Material 6) having an AUC of 0.6694 and a Brier score of 0.1988. Thus, it achieved a substantial improvement over the model generated by a GLM using all variables (46.2%).

The main predictors of drop out that made a substantial contribution (i.e. relative importance >90%) to the model were lower education level, younger age, lower scores on the compulsive scale

of the Personality Style and Disorder Inventory (PSSI), higher scores on the negativistic and antisocial scale of the PSSI and higher scores on the additional scale of the BSI as well as a higher total score (see Supplementary materials 7 for an overview of all variables; see Liaw and Wiener⁴⁸ for a description of how variable importance is calculated in a Random Forest model). The BSI additional scale is the mean of the four additional items not included in any of the dimension scores ('poor appetite', 'trouble falling asleep', 'thoughts of death and dying', 'feeling of guilt').

Discussion

Main findings

The aim was to evaluate the use of different machine-learning algorithms in a naturalistic routine care setting by generating a predictive model to identify patients who are at risk of dropping out. Two different indices were used to gain a more comprehensive picture of the results. We selected 21 algorithms for our study and used nested

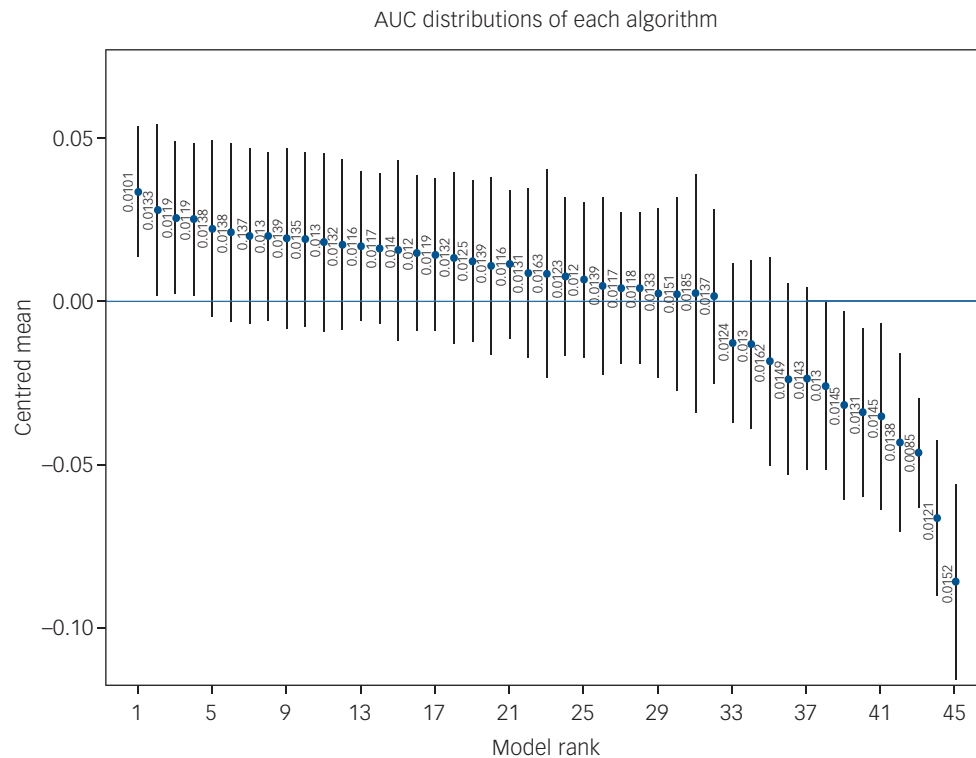


Fig. 1 Distribution of the 20 outer cross-validation models generated by each algorithm and ensemble ranked from best to worst using the area under the curve.

Each value was grand-mean centred; the horizontal line represents the total average of all models. The numbers on the graphs are the standard deviations.

cross validation to compare them. We used the best algorithms and least correlating algorithms to generate ensembles that were also compared. The best model was an ensemble of the best algorithm with its least correlating algorithm (i.e. Random Forest, kNN) that used only predictive variables and was merged via a GLM. Differences between the best ensemble and a single GLM as well as the worst algorithm were highly significant, independent of the examined parameters. When comparing the distributions of the best model and the GLM, a large effect size of up to $d = -2.93$ was found, indicating the superiority of the best model independent of the training sample used.

The best model was able to correctly identify 63.4% of all cases of patients dropping out in an independent holdout sample. Although this does not seem very precise at first, it must be acknowledged that this prediction was made before the first session of routine CBT and that a single GLM correctly identified only 46.2%. Therefore, this model is of high clinical value and is able to identify before the first session has occurred patients who tend to drop out of therapy. The mostly identical values of the AUC and the Brier score in the holdout sample compared with the test set in the modelling process indicate good stability and generalisability of the model.

Interpretation of our findings relating to identification of patients at risk for treatment discontinuation

The most important variables used in the final prediction model also appear to differ significantly between individuals who dropped out and those where there was consensual termination. Nevertheless, this is not true for all variables, suggesting that the model uses more than just the different mean values for prediction. Based on the relevant variables in the model, therapists should take time to build a complementary relationship with the patient and invest

time in explaining how therapy can concretely help them. Particularly high levels of interpersonal variables that make it difficult to establish a functional therapeutic relationship (for example negativistic or antisocial personality style) appear to increase the risk of drop out. It is important for clinicians to pay attention to the complementarity of the relationship in order to establish a good alliance. This is especially crucial in the first session, as this is where the first impression is made. Here model predictions can be used to better prepare for potential interpersonal difficulties.

With regard to the BSI additional scale, it seems reasonable to first treat symptoms such as sleep problems, poor appetite, suicidal thoughts and feelings of guilt. Although general symptom burden or functionality do not play a particularly important role, symptoms that are very obvious to the patient (such as sleep problems, distressing suicidal thoughts) appear to be important indicators. It is obvious that patients hope for a quick improvement in these symptoms resulting from therapy, which has an important signal effect for therapists to focus on the treatment of these symptoms, especially at the beginning of therapy.

Interestingly, lower education and younger age also seem to increase the probability of individuals dropping out. Other studies have also identified these variables,⁸ so these should be considered in therapy, even if they are invariant. Future studies should explore the underlying mechanisms of these on drop-out probability to better understand the effects and to improve future models. Nevertheless, using the information from our model, clinicians could generate a more precise case concept for the individual patient before the first session to help patients gain confidence in therapy, facilitate the establishment of a functional therapeutic relationship, and thus reduce the risk of patients dropping out. Therefore, the best model from our analyses could improve and further support measurement-based care with regard to drop-out prediction and prevention.

Interpretation of our findings relating to the use of machine learning

Further, the results indicate that machine-learning algorithms/ensembles can have a true predictive advantage in naturalistic settings. However, this does not apply to all algorithms. Some produced significantly worse predictions, indicating that not all machine-learning algorithms/ensembles are suited to naturalistic settings. Our results revealed that ensembles consisting of low-correlating algorithms did not perform well except when a powerful algorithm that delivers good predictions on its own is included. The idea that low-correlating algorithms assess different aspects of the data-set and thus should perform better than ensembles with more similar algorithms did not hold true. Mayer⁴⁶ states that an optimal ensemble of low correlating algorithms consists of those that perform similarly on their own. This could explain the worse prediction quality in our data, as this was not the case in our analyses.

Furthermore, the assumption ‘the more the better’ also did not hold true. Ensembles that used more algorithms did not automatically perform better. This finding is in line with previous argumentation that selecting algorithms to create an ensemble does not follow easy rules like ‘the more the better’, but is a research topic of its own.⁴⁹ In addition, when using a large set of variables, a variable selection procedure should be part of model generation, either by using an algorithm that includes a selection procedure or a preceding variable selection. Our findings indicate that algorithms that had to handle many variables and did not include a variable selection procedure performed worse (for example linear discriminant analysis). This finding is in line with the existing literature, stating that, in clinical settings, not every variable has predictive power for a certain outcome⁵⁰ and can thus weaken the power of the model.

Interestingly, tree-based and boosted algorithms seem to perform better compared with more advanced algorithms such as neural networks. This finding appeared consistently, independent of the examined parameters. Therefore, for this kind of naturalistic binary data, boosted linear algorithms and tree-based approaches such as random forest seem very well suited.

Limitations

Although this study has many strengths, several limitations must be mentioned. One reason for the poor performance of neural networks could be the data quality. Albeit naturalistic assessments include crucial predictive information, they are nowhere near perfect and always have measurement errors. Although this topic is not new,⁵¹ these errors prevent the algorithms from assessing the relevant relationships. These suggestions are in line with the existing literature.^{50,52} A solution to this problem could be the usage of ecological momentary assessment data, which provides more accurate descriptions of within-person processes at a higher resolution. For future studies, it is of great interest whether complex algorithms such as neural networks are more suitable for such data and are thus able to improve predictions of drop out. Electrophysiological variables and neural imaging variables could also improve predictions,⁵³ but such assessments are expensive and time-consuming and therefore unlikely to be used in routine care. In addition, the amount of data could have played an important role. Complex algorithms that are able to assess high-order interactions need a lot of data.¹⁶ Thus, the size of our data-set limits the evaluation of these algorithms. Future studies should try to generate even larger data-sets in order to evaluate the possible benefit of advanced algorithms.

Furthermore, it is possible that certain predictive variables were not collected. For example, we only collected whether patients were taking medication or not, regardless of what they were taking or for

how long. Although this variable did not play a role in our model, it cannot be ruled out that more precise information could improve the model. The same applies to the variables that we had to exclude because of too many missing values (i.e. PHQ-9, ASQ, GAD-7, TIPI, WSAS). These variables contain important clinical information that could be important for prediction.

Moreover, we used only a small number of possible machine-learning algorithms. Although we used many models that have already been applied in psychological studies to create a representative picture, it cannot be ruled out that an even more suitable approach for this kind of data exists. Also, as mentioned above, the use of ensembles requires a profound understanding of this topic. For our own ensembles, we used the stacking method. However, there are other options to create ensembles such as bagging or boosting.



Although the model is well protected against overfitting by the use of repeated and nested cross validation as well as a separate holdout sample, the possibility of overfitting cannot be completely ruled out. Furthermore, our holdout sample is quite similar to the training sample, which limits the generalisability of the results. Nevertheless, it must be noted that there are differences between the samples, especially with regard to the diagnosis, which is why a certain degree of generalisability can be assumed. Nevertheless, holdout samples from other institutes should be used in future studies to more robustly test generalisability.

In addition, although this model helps identifying patients who are at risk of dropping out of therapy, it does not reveal the reasons for this increased risk. No causal conclusions can be drawn from this model, which is a limitation of our model and of machine learning in general. Nevertheless, the identified predictors provide first clues as to which risk factors may be relevant to drop out. Moreover, the identification of patients at risk for treatment discontinuation is the first step to reducing the number of patients who drop out.

Implications

To our knowledge, this study is the first to use such a large naturalistic data-set to evaluate different machine-learning algorithms and ensembles to identify a useful drop-out prediction model. The current study compared several machine-learning methods in order to evaluate the benefit of machine learning in naturalistic contexts and to generate a model that has high clinical value for identifying drop-out risk at an individual level. The model identified over 60% of patients’ type of therapy termination correctly. This study’s findings highlight that it is possible to identify, before the first session has occurred, patients at risk of dropping out and that machine learning algorithms provide an important contribution to model generation. Tree-based and boosted algorithms that include a variable selection procedure (for example elastic net) seem especially suited to building prediction models for psychotherapy drop out.

Future research should further explore treatment data to improve prediction models and use them to develop strategies to reduce the risk of drop out. By implementing these models into clinical support systems, the number of individuals who drop out could be reduced, resulting in more effective therapy outcomes and less burden on patients and society.

Björn Bennemann , Department of Clinical Psychology and Psychotherapy, University of Trier, Germany; **Brian Schwartz**, Department of Clinical Psychology and Psychotherapy, University of Trier, Germany; **Julia Giesemann** , Department of Clinical Psychology and Psychotherapy, University of Trier, Germany; **Wolfgang Lutz**, Department of Clinical Psychology and Psychotherapy, University of Trier, Germany

Correspondence: Björn Bennemann. Email: bennemann@uni-trier.de

First received 1 Sep 2021, final revision 12 Jan 2022, accepted 17 Jan 2022

Supplementary material

To view supplementary material for this article, please visit <https://doi.org/10.1192/bjp.2022.17>.

Data availability

The data that support the findings of this study are available from the corresponding author, B.B., upon reasonable request.

Acknowledgements

We are thankful to Dr Kaitlyn Poster for proofreading the manuscript.

Author contributions

W.L. provided the data-set and advised on the structure and design of the study. B.B. developed the idea and performed the analyses. B.S. J.G. and W.L. advised on statistical issues. All authors were involved in interpreting data, drafting the work or critically revising it for important intellectual content. All authors provided final approval of the submitted version of the manuscript.

Funding

This work was partly supported by the German Research Foundation (DFG) under Grant Nr. LU 660/10-1.

Declarations of interest

None.

References

- von Brachel R, Hirschfeld G, Berner A, Willutzki U, Teismann T, Cwik JC, et al. Long-term effectiveness of cognitive behavioral therapy in routine outpatient care: a 5-to 20-year follow-up study. *Psychother Psychosom* 2019; **88**: 225–35.
- Swift JK, Greenberg RP, Tompkins KA, Parkin SR. Treatment refusal and premature termination in psychotherapy, pharmacotherapy, and their combination: a meta-analysis of head-to-head comparisons. *Psychotherapy (Chic)* 2017; **54**: 47–57.
- Wells JE, Browne MO, Aguilar-Gaxiola S, Al-Hamzawi A, Alonso J, Angermeyer MC, et al. Drop out from out-patient mental healthcare in the World Health Organization's World Mental Health Survey initiative. *Br J Psychiatry* 2013; **202**: 42–9.
- Rossi A, Amadeo F, Bisoffi G, Ruggeri M, Thornicroft G, Tansella M. Dropping out of care: inappropriate terminations of contact with community-based psychiatric services. *Br J Psychiatry* 2002; **181**: 331–8.
- Lutz W, Rubel JA, Schiefele A-K, Zimmermann D, Böhnke JR, Wittmann WW. Feedback and therapist effects in the context of treatment outcome and treatment length. *Psychother Res* 2015; **25**: 647–60.
- Kessler RC. The potential of predictive analytics to provide clinical decision support in depression treatment planning. *Curr Opin Psychiatry* 2018; **31**: 32–9.
- Swift JK, Greenberg RP. Premature discontinuation in adult psychotherapy: a meta-analysis. *J Consult Clin Psychol* 2012; **80**: 547–59.
- Zimmermann D, Rubel JA, Page AC, Lutz W. Therapist effects on and predictors of non-consensual dropout in psychotherapy. *Clin Psychol Psychother* 2017; **24**: 312–21.
- Tehrani E, Krussel J, Borg L, Munk-Jørgensen P. Dropping out of psychiatric treatment: a prospective study of a first-admission cohort. *Acta Psychiatr Scand* 1996; **94**: 266–71.
- Delgado J, Lutz W. A development pathway towards precision mental health care. *JAMA Psychiatry* 2020; **77**: 889–90.
- Fabbri C, Kasper S, Kautzky A, Bartova L, Dold M, Zohar J, et al. Genome-wide association study of treatment-resistance in depression and meta-analysis of three independent samples. *Br J Psychiatry* 2019; **214**: 36–41.
- Legge SE, Dennison CA, Pardiñas AF, Rees E, Lynham AJ, Hopkins L, et al. Clinical indicators of treatment-resistant psychosis. *Br J Psychiatry* 2020; **216**: 259–66.
- Lutz W, Deisenhofer A-K, Rubel JA, Bennemann B, Giesemann J, Poster K, et al. Prospective evaluation of a clinical decision support system in psychological therapy. *J Consult Clin Psychol* [Epub ahead of print] 24 Jun 2021. Available from: <https://doi.org/10.1037/ccp0000642>.
- Maj M, Stein DJ, Parker G, Zimmermann M, Fava GA, Hert M, et al. The clinical characterization of the adult patient with depression aimed at personalization of management. *World Psychiatry* 2020; **19**: 269–93.
- Kessler RC, Bossarte RM, Luedtke A, Zaslavsky AM, Zubizarreta JR. Machine learning methods for developing precision treatment rules with observational data. *Behav Res Ther* 2019; **120**: 103412.
- Brownlee J. *Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch [e-book]*. v1.14. Machine Learning Mastery, 2019 (<https://machinelearningmastery.com/master-machine-learning-algorithms> [cited 22 Sept 2020]).
- Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019; **110**: 12–22.
- Wilkinson J, Arnold KF, Murray EJ, van Smeden M, Carr K, Sippy R, et al. Time to reality check the promises of machine learning-powered precision medicine. *Lancet Digital Health*; 2020; **2**: 677–680.
- Wittchen H-U, Wunderlich U, Gruschwitz S, Zaudig M. *SKID I. Strukturiertes Klinisches Interview für DSM-IV. Achse I: Psychische Störungen. Interviewte und Beurteilungsheft. Eine deutschsprachige, erweiterte Bearb. d. amerikanischen Originalversion des SKID, I [SCID I. Structured Clinical Interview for DSM-IV. Axis 1: Mental disorders. Interview booklet and assessment booklet. A German-language, expanded adaptation of the American original version of the SKID I]*. Hogrefe, 1999.
- Bronisch T, Hiller W, Mombour W, Zaudig M. *International Diagnostic Checklists for personality disorders according to ICD-10 and DSM-IV—IDCL-P*. Hogrefe and Huber Publishers, 1996.
- Ellsworth JR, Lambert MJ, Johnson J. A comparison of the Outcome Questionnaire-45 and Outcome Questionnaire-30 in classification and prediction of treatment outcome. *Clin Psychol Psychother* 2006; **13**: 380–91.
- Franke GH. *Brief Symptom Inventory (BSI) von LR Derogatis (Kurzform der SCL-90-R) [Brief Symptom Inventory (BSI) by LR Derogatis (short form of SCL-90-R)]*. Beltz Test, 2000.
- Derogatis LR. *SCL-90-R: Symptom Checklist-90-R Administration, Scoring, and Procedures Manual*. NCS Pearson, 1975.
- Boyle K, Deisenhofer A-K, Rubel JA, Bennemann B, Weinmann-Lutz B, Lutz W. Assessing treatment integrity in personalized CBT: the inventory of therapeutic interventions and skills. *Cogn Behav Ther* 2020; **49**: 210–27.
- Lutz W, Schürch E, Stulz N, Böhnke JR, Schöttke H, Rogner J, et al. Entwicklung und psychometrische Kennwerte des Fragebogens zur Evaluation von Psychotherapieerläufen (FEP) [Development and psychometric parameters of the questionnaire for evaluating the course of psychotherapy (FEP)]. *Diagnostica* 2009; **55**: 106–16.
- Lutz W, Tholen S, Schürch E, Berking M. Die Entwicklung, Validierung und Reliabilität von Kurzformen gängiger psychometrischer Instrumente zur Evaluation des therapeutischen Fortschritts in Psychotherapie und Psychiatrie [The development, validation and reliability of short forms of common psychometric instruments for the evaluation of therapeutic progress in psychotherapy and psychiatry]. *Diagnostica* 2006; **52**: 11–25.
- Horowitz LM, Strauß B, Kordy H. *Inventar zur Erfassung interpersonaler Probleme - Deutsche Version [Inventory for Recording Interpersonal Problems - German Version]* 2nd ed. Beltz Test GmbH, 2000.
- Grosse Holtforth M, Grawe K. Der Inkongruenzfragebogen (INK): Ein Messinstrument zur Analyse motivationaler Inkongruenz [The Incongruity Questionnaire (INK): A Measuring tool for analyzing motivational incongruity]. *Z Klin Psychol Psychopathol Psychother* 2003; **32**: 315–23.
- Hautzinger M, Joormann J, Keller F. *Die Skala dysfunktionaler Einstellungen (DAS) [The Scale of Dysfunctional Attitudes (DAS)]*. Hogrefe, 2005.
- Siegrist J, Geyer S. Inventar lebensverändernder Ereignisse. In *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS) [Compilation of Social Science Items and Scales (ZIS)]*. gesis, 1997 (<https://doi.org/10.6102/zis92> [cited 20 Sept 2021]).
- Horowitz K, Kazén M. *PSSI: Persönlichkeits-Stil- und Störungs-Inventar [PSSI: Personality Style and Disorder Inventory]* 2nd ed. Hogrefe, 2009.
- Cohen ZD, Wiley JF, Lutz W, Fisher AJ, Kim T, Saunders R, et al. *SMART Mental Health Prediction Tournament*. OSF, 2018 (osf.io/wxgzu [cited 23 Sept 2021]).
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2021.
- Stekhoven DJ. *missForest: Nonparametric Missing Value Imputation using Random Forest*. Astrophysics Source Code Library, 2015.
- Kuhn M. Building Predictive Models in R using the caret Package. *J Stat Soft* 2008; **28**: 1–26.
- Deane-Mayer ZA, Knowles JE. *caretEnsemble: Ensembles of Caret Models. R package version 2.0.1*. CRAN.R-project, 2019 (<https://cran.r-project.org/web/packages/caretEnsemble/index.html> [cited 2021 Sept 20]).

- 37 Kuhn M. *The caret Package: 6 Available Models*. No publisher, 2019 (<https://topepo.github.io/caret/index.html>) [cited 23 Sept 2021].
- 38 Brownlee J. *Machine Learning Mastery: Nested Cross-Validation for Machine Learning with Python*. Machine Learning Mastery, 2020 (<https://machinelearningmastery.com/nested-cross-validation-for-machine-learning-with-python/>) [cited 22 Sept 2021].
- 39 Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res* 2010; **11**: 2079–107.
- 40 Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002; **16**: 321–57.
- 41 Hand DJ, Vinciotti V. Choosing k for two-class nearest neighbour classifiers with unbalanced classes. *Pattern Recognit Lett* 2003; **24**: 1555–62.
- 42 Japkowicz N, Stephen S. The class imbalance problem: a systematic study. *Intell Data Anal* 2002; **6**: 429–449.
- 43 Brier GW. Verification of forecasts expressed in terms of probability. *Monthly Weather Rev* 1950; **78**: 1–3.
- 44 Green DM, Swets JA. *Signal Detection Theory and Psychophysics*. Wiley, 1966.
- 45 Zhang C, Ma Y. *Ensemble Machine Learning: Methods and Applications*. Springer 2012.
- 46 Mayer Z. *A Brief Introduction to caretEnsemble*. CRAN.R-project, 2019 (<https://cran.r-project.org/web/packages/caretEnsemble/vignettes/caretEnsemble-intro.html>) [cited 20 Sept 2021].
- 47 Kuhn M, Johnson K. *Applied Predictive Modeling*. Springer, 2013.
- 48 Liaw A, Wiener M. Classification and regression by randomForest. *R news* 2002; **2**: 18–22.
- 49 Ramzai J. *Simple Guide for Ensemble Learning Methods*. towards data science, 2019 (<https://towardsdatascience.com/simple-guide-for-ensemble-learning-methods-d87cc68705a2>) [cited 20 Sept 2021].
- 50 Chowdhury MZI, Turin TC. Variable selection strategies and its importance in clinical prediction modelling. *Fam Med Community Health* 2020; **8**: e000262.
- 51 Lutz W, de Jong K, Rubel JA, Delgadillo J. Measuring, predicting and tracking change in psychotherapy. In *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (7th edn) (eds M Barkham, W Lutz, LG Castonguay). Wiley, 2021.
- 52 Jacobucci R, Grimm KJ. Machine learning and psychological research: the unexplored effect of measurement. *Perspect Psychol Sci* 2020; **15**: 809–16.
- 53 Lueken U, Zierhut KC, Hahn T, Straube B, Kircher T, Reif A, et al. Neurobiological markers predicting treatment response in anxiety disorders: a systematic review and implications for clinical application. *Neurosci Biobehav Rev* 2016; **66**: 143–62.

