frontiers | Frontiers in Psychiatry

TYPE Original Research
PUBLISHED 25 September 2023
DOI 10.3389/fpsyt.2023.1236551

OPEN ACCESS

EDITED BY
Matthias Jaeger,
Psychiatrie Baselland, Switzerland

REVIEWED BY
Yannik Terhorst,
University of Ulm, Germany
Seyed-Ali Sadegh-Zadeh,
Staffordshire University, United Kingdom

*CORRESPONDENCE
Katinka Franken
✉ c.p.m.franken@utwente.nl

RECEIVED 11 June 2023
ACCEPTED 11 September 2023
PUBLISHED 25 September 2023

CITATION
Franken K, ten Klooster P, Bohlmeijer E, Westerhof G and Kraiss J (2023) Predicting non-improvement of symptoms in daily mental healthcare practice using routinely collected patient-level data: a machine learning approach.
Front. Psychiatry 14:1236551.
doi: 10.3389/fpsyt.2023.1236551

Predicting non-improvement of symptoms in daily mental healthcare practice using routinely collected patient-level data: a machine learning approach

Katinka Franken*, Peter ten Klooster, Ernst Bohlmeijer, Gerben Westerhof and Jannis Kraiss

Department of Psychology, Health and Technology, Faculty of Behavioural, Management and Social Sciences, University of Twente, Enschede, Netherlands

Objectives: Anxiety and mood disorders greatly affect the quality of life for individuals worldwide. A substantial proportion of patients do not sufficiently improve during evidence-based treatments in mental healthcare. It remains challenging to predict which patients will or will not benefit. Moreover, the limited research available on predictors of treatment outcomes comes from efficacy RCTs with strict selection criteria which may limit generalizability to a real-world context. The current study evaluates the performance of different machine learning (ML) models in predicting non-improvement in an observational sample of patients treated in routine specialized mental healthcare.

Methods: In the current longitudinal exploratory prediction study diagnosis-related, sociodemographic, clinical and routinely collected patient-reported quantitative outcome measures were acquired during treatment as usual of 755 patients with a primary anxiety, depressive, obsessive compulsive or trauma-related disorder in a specialized outpatient mental healthcare center. ML algorithms were trained to predict non-response (< 0.5 standard deviation improvement) in symptomatic distress 6 months after baseline. Different models were trained, including models with and without early change scores in psychopathology and well-being and models with a trimmed set of predictor variables. Performance of trained models was evaluated in a hold-out sample (30%) as a proxy for unseen data.

Results: ML models without early change scores performed poorly in predicting six-month non-response in the hold-out sample with Area Under the Curves (AUCs) < 0.63. Including early change scores slightly improved the models' performance (AUC range: 0.68–0.73). Computationally-intensive ML models did not significantly outperform logistic regression (AUC: 0.69). Reduced prediction models performed similar to the full prediction models in both the models without (AUC: 0.58–0.62 vs. 0.58–0.63) and models with early change scores (AUC: 0.69–0.73 vs. 0.68–0.71). Across different ML algorithms, early change scores in psychopathology and well-being consistently emerged as important predictors for non-improvement.

Conclusion: Accurately predicting treatment outcomes in a mental healthcare context remains challenging. While advanced ML algorithms offer flexibility, they showed limited additional value compared to traditional logistic regression in this study. The current study confirmed the importance of taking early change scores in both psychopathology and well-being into account for predicting longer-term outcomes in symptomatic distress.

KEYWORDS

machine learning, mental disorder, well-being, psychopathology, prediction, non-improvement

1. Introduction

1.1. Prevalence and impact of psychiatric disorders

Worldwide, around one in eight people has one or more mental disorders (1). Mental disorders are the leading cause of years lived with disability (YLDs), accounting for one in every six YLDs globally (1). They contribute significantly to a lack of quality of life (2) and the direct and indirect economic and societal costs are substantial (1). Depression and anxiety alone result in the loss of nearly US$ 1 trillion and 12 billion working days every year (3). The increasing demand for care in combination with limited treatment effects puts pressure on waiting lists in mental healthcare (4). Insights in predicting who is less likely to improve early in treatment would be helpful to make treatment more efficient, reduce waste of financial and human resources and tailor treatment to the individual (5–7).

1.2. Predicting treatment effects

Studies show that 60% of patients with a mental disorder do not benefit from evidence-based treatments (8–12). At present, no convincing evidence has been found for a difference in treatment effect for any specific treatment, neither for mood disorders (13, 14) nor anxiety disorders (14, 15). Norcross and Lambert (16) argue that fitting psychotherapy to patient characteristics is necessary for treatment success. Clinical practice, however, shows that a DSM-classification alone does not give sufficient direction to appropriate treatment (17–19). This underlines the relevance of adopting a more transdiagnostic approach in clinical practice and searching for predictors across the main diagnoses. Early identification of non-responders can increase treatment effectiveness as it may support personalized treatment recommendations (20).

Mental disorders are complex and trajectories of treatment can depend on many factors, making the prediction of treatment outcomes challenging. Previous studies have incidentally found several predictors for treatment outcomes in various populations, including sociodemographic features (age, gender, employment status), symptom severity, emotion regulation abilities, problem duration, level of functioning, interpersonal problems, prior treatments, comorbidity of personality disorders or medical conditions, treatment non-adherence and alliance [e.g., (6, 7, 21–33)]. However, no consistent pre-treatment characteristics have been identified that reliably predict treatment outcomes (34, 35).

1.3. Importance of analyzing longitudinal data from the real-world psychiatric context

Findings about predictors for treatment outcomes often stem from data from randomized controlled trials (RCTs), which may be problematic for several reasons. First, only a selective and limited number of potential predictors are usually included in randomized controlled trials (RCTs), only allowing for limited conclusions about what predictors are relevant for treatment outcomes. Second, RCTs often do not meet the required sample size needed for detecting significant predictors (36–41). Third, many RCTs, especially efficacy trials, tend to have rather strict in- and exclusion criteria and controlled study procedures. While the use of such criteria leads to relatively high internal validity, it may decrease external validity and limit generalizability to patient populations treated in daily clinical practice (42–44).

Considering that most patients are not treated in RCTs, but in naturalistic clinical institutions, using real-world clinical data to identify predictors of treatment outcomes is likely to be more externally valid (45). Large observational studies using data collected in the real-world context may be a valuable alternative to develop more generalizable prediction models (28). Longitudinal routinely collected patient-reported outcome data of psychopathology and well-being are increasingly available that provide information about treatment outcomes [e.g., (46, 47)]. For instance, electronic health records (EHRs) of psychiatric patients contain large amounts of potentially useful clinical information. However, despite increased external validity, such routinely collected data presents challenges as well. Predictive features are heterogeneous and may interact with each other in ways that traditional statistical models may not be able to capture. By including a larger number of features there is also a risk for overfitting. In addition, using real-world data is often challenging, especially due to high attrition and missing data rates.

1.4. The potential of machine learning

Recent improvements in computational power and the refinement of the applications of machine learning (ML) technologies have been suggested to offer possibilities to develop robust and generalizable prediction models for treatment response using real-world data (18, 48, 49). ML has shown promise within clinical psychology in helping to understand large-scale health data (50–55). ML is a subfield of artificial intelligence that involves the development of algorithms and statistical models that enable computers to learn and make predictions or decisions based on data without being explicitly programmed to do so (56).

ML can predict treatment effects using high-quality data such as patient characteristics and questionnaire scores over time [e.g., (55, 57, 58)]. The techniques used in building ML models depend on the type of data and can be based on supervised learning, unsupervised learning, and reinforcement learning. Supervised learning, as applied in current study, involves training a model on labeled data, where the desired output is already known. The ultimate goal is to build a model that can accurately predict future outcomes (59). Aafjes-van Doorn, Kamsteeg, Bate, and Aafjes (60) systematically reviewed 51 studies of ML in psychotherapy and concluded that most model development

studies used supervised learning techniques to classify or predict labeled treatment process or outcome data, whereas some used unsupervised techniques to identify clusters in the unlabeled patient or treatment data.

In ML models, the main statistic of interest is the prediction accuracy of the algorithm in a hold-out sample. The hold-out sample is a random subset of the original dataset that is held back and not used during training. For categorical outcomes the accuracy is usually reported as the accuracy, sensitivity (or recall) and specificity, and area under the curve (AUC) computed from the confusion matrix of the predicted against the observed labels of the observations.

Application of ML has various potential advantages above traditional statistical methods. First, by employing robust statistical and probabilistic techniques, ML has the ability to make predictions regarding treatment effects, enabling the comprehension of complex, integrated datasets consisting of heterogeneous features (57, 60). Second, ML methods require less restrictive assumptions regarding the non-linear relationship of high-dimensional data and the skewed distribution of features (61). The potential of ML has been demonstrated by improved accuracy compared to regular methods such as regression (62, 63). Third, the application of cross-validation techniques, which are common in ML methods but usually not applied in traditional prediction analyzes such as significance-based regression, reduces the risk for overfitting (64). Fourth, ML increases the generalizability of the predictions since some ML algorithms might perform better than traditional analysis techniques in complex datasets involving many features (65).

1.5. Predicting non-improvement by ML using outcome data of psychopathology and well-being

Real-world mental health data have been used in various ML applications, such as modeling disease progression (66), predicting disease deterioration (67), predicting risk factors for adverse outcomes, such as mortality, readmission or prolonged length of stay (68) as well as predicting treatment outcomes (69). However, research predicting outcomes using real-world clinical data is still scarce. Some studies have shown that compared to traditional research methods ML can increase prediction accuracy using sociodemographic, clinical and biological data (19, 63, 64, 70–74). However, ML has not often been applied to the routine collection of patient-level outcome data in combination with sociodemographic and clinical data.

Hence, the objective of the current study is to evaluate and compare the performance of different ML models in predicting treatment outcomes in an observational sample of patients treated in routine specialized mental healthcare. This will be done by predicting non-improvement in psychopathology 6 months after start of treatment in a group of patients with anxiety and mood disorders. A range of routinely available clinical, demographic and self-reported outcome features will be used to predict treatment outcomes. Several models are explored, such as those involving the incorporation of change scores early in treatment as supplementary predictors, and models that are trained on a reduced set of features using feature reduction techniques.

2. Methods

2.1. Study design and data collection

The present study concerned an exploratory machine learning based prediction analysis of routinely collected observational longitudinal quantitative data. The recommendations for reporting machine learning analyzes in clinical research (75) were followed. We used data collected in the context of routinely collected patient-level outcome data of psychopathology and well-being, a standardized service to measure treatment effects. Patients in a mental healthcare center in the Netherlands completed online questionnaires every 3 months from the initial interview to end of treatment. Data were collected before start of treatment (T0), and three (T1), six (T2), nine (T3), and 12 (T4) months after treatment commenced. Invitations to complete the questionnaires were sent automatically and data from the completed questionnaires were stored anonymously by an independent data controller in a database generated for this longitudinal study. The data were gathered between March 2015 and November 2019. About 19% ($n = 145$) were lost to 3-month follow-up, 34% ($n = 254$) did not complete the six-month follow-up assessment, and about 58% ($n = 439$) did not complete the 12-month follow-up.

Patients provided passive informed consent for their anonymized data to be used for scientific research. As data were collected in the context of regular care and only anonymized data were analyzed, the study did not require medical ethical approval according to Dutch law. Inclusion criteria were: (1) aged between 18 to 65 years, (2) full completion of the questionnaires on the same day, and (3) diagnosed by depressive, bipolar, anxiety, trauma related or obsessive-compulsive disorder. The diagnosis was based on an extensive interview by a licensed clinical psychologist or psychiatrist. The diagnosis and related (evidence- and practice-based) treatment options were discussed and confirmed in a multidisciplinary team.

2.2. Baseline features

An overview of all available baseline features that were included in the models can be found in Table 1. These include sociodemographic (e.g., gender, age), diagnostic (e.g., main diagnosis, comorbidity), and clinical characteristics of patients (e.g., number of treatments in the past, social problems). One additional clinical feature was created that was labeled as treatment intensity. This feature represents the ratio of number of treatments in the past and total duration of past treatments. Routinely collected self-reported psychological features included the total and subscales scores of the Outcome Questionnaire [OQ-45; (76)] and the Mental Health Continuum-Short Form [MHC-SF; (77, 78)]. The OQ-45 is a 45-item self-report measure of psychopathology and includes four subscales, namely symptomatic distress (e.g., "I'm anxious"), interpersonal relations (e.g., "Often I have fights"), somatic complaints (e.g., "I tire quickly"), and social roles performance (e.g., "I feel like I'm not doing well with my work"). Items are answered on a five-point Likert scale ranging from 0 (*never*) to 4 (*almost always*). Previous studies have shown that the OQ-45 is a reliable and valid instrument across different cultural contexts (76, 79, 80). The 14-item MHC-SF measures the presence of different well-being dimensions during the past month on three subscales: emotional (e.g., "Feeling satisfied with

TABLE 1 Overview of baseline features.

Sociodemographic	Psychological
Age Gender (male/female) Education (low/moderate/high) Marital status (no partner/partner/other)	OQ-45 Total score OQ-45 Symptomatic distress OQ-45 Anxiety and somatic distress OQ-45 Interpersonal relationships OQ-45 Social role adjustment MHC-SF Total score MHC-SF Emotional well-being MHC-SF Social well-being MHC-SF Psychological well-being GAF score
Diagnostic	**Clinical**
Main diagnosis (depressive disorder, anxiety disorder, bipolar disorder, OCD, traumatic disorder) First comorbidity (no/yes) Second comorbidity (no/yes) Somatic comorbidity (no/yes)	Axis II problem (no/yes) Axis IV financial problem (no/yes) Axis IV relationship problems (no/yes) Axis IV social problems (no/yes) Axis IV work problems (no/yes) Number of treatments in the past (0–4/5–10/10+) Years since first time enrolled (0–3/3–10/10+) Sum of previous enrollments in years (0–2/2–5/5+) Log-transformed treatment intensity[a]

[a]Treatment intensity was calculated as the ratio of number of treatments in the past and total duration of past treatments.

life"), social (e.g., "Feeling that you belong to a community"), and psychological well-being (e.g., "Feeling that your life as a sense of direction or meaning to it"). Items are answered on a six-point Likert scale ranging from 0 (*never*) to 5 (*every day*). The MHC-SF has shown good psychometric properties in the general population [e.g., (77, 78)] and in clinical groups (81). In total, 41 baseline features were included in the models.

2.3. Response variable

Non-improvement on the OQ-45 total scores at six-month follow-up was used as binary response variable. Cases were labeled as 'not improved' if the change from baseline in the symptomatic distress scale of the OQ-45 6 months after baseline was smaller than half a standard deviation (0.5 SD). The choice of this cut-off is motivated by a previous systematic review of 38 studies, suggesting that half a standard deviation consistently reflected a minimally important difference for health-related quality of life instruments across studies (82). Half a standard deviation also corresponds with a medium effect size according to Cohen's conventional rule of thumb (83). The reason to use improvement at six-month follow-up as response variable, was that missing data become too high at later follow-up points and because 6 months was considered a time period long enough to be clinically relevant. Besides, hardly any additional average treatment effects were observed after that time in the dataset.

2.4. Preprocessing

Descriptive analyzes were done in the statistical package for social sciences (SPSS) version 27 (84). All other ML analyzes were conducted in R (85) using the caret R-package (86). Data, syntax and output files can be found on the Open Science Framework website (https:osf.io/xwme4/).

All categorical features were dummy coded and continuous features were visually checked for approximate normal distribution. The feature 'treatment intensity' was log-transformed, since it was not normally distributed and right-skewed. Cases that did not complete the OQ-45 at 6 months after baseline were removed. Only complete cases were used, since imputing the response variable might overestimate the performance of the ML algorithms, as common imputation techniques (e.g., random forest) would be similar to what ML algorithms would use to predict non-improvement at follow-up. After data preprocessing and cleaning, the remaining data was randomly split into a training (70%) and hold-out sample (30%). Next, missing baseline data (0.8%) was globally imputed (before conducting k-fold cross-validations) and separately for training and hold-out data, using random forest imputation (87).

2.5. Machine learning models and model performance

The goal of ML is to identify patterns in observed high complex data in high dimensional settings, make accurate predictions or classifications, and improve their performance over time by learning from new data [e.g., (57, 58, 63, 88–90)]. ML algorithms involve three main components, which are (1) a model, (2) data for training, testing and validation, and (3) an optimization algorithm. The model represents the data and relationships between features. The training data is used to optimize model weights using cross-validation (CV) to minimize error or loss, while the optimization algorithm finds the optimal values of the model weights. ML algorithms are conducted in two steps: training and testing. During training, the objective is to find a balance between identifying specific patterns in the patient data and preventing overfitting (training data so well that it negatively affects its performance on new data, which occurs when the algorithm fits too closely to the random noise in the data). In the test phase, the accuracy of the predictions made by the algorithm is computed by comparing the predictions made for new data with the actual values observed in the new sample. CV optimizes the ML model by assessing skills of the ML model and testing its performance (or accuracy) in new data later.

Different ML algorithms were compared to predict non-improvement at six-month follow-up. The following algorithms were used: Logistic regression (LR), random forest (RF), support vector machine (SVM) with linear, radial and polynomial kernels, and gradient boosting machine (GBM). These algorithms differ in their underlying principles and modeling techniques. LR focuses on estimating probabilities based on linear relationships, RF combines decision trees for predictions, SVM find optimal hyperplanes for classification, and GBM sequentially build models to minimize prediction error. The rationale for choosing these algorithms was to be able to compare this study with previous studies that used similar algorithms [e.g., (19, 74)]. Furthermore, we not only wanted to

include flexible and less interpretable algorithms (e.g., GBM or SVM), but also techniques that are easier to interpret, while being less flexible (91).

All models were trained on the training set using repeated k-fold cross-validation with 10 folds and 10 repetitions (90). As the response variable was imbalanced, up-sampling was used for training purposes, which randomly replicates instances of the minority class. We explored the effect of class imbalance before applying up-sampling. If no up-sampling was used models performed comparably well in terms of overall accuracy, but were not useful because the sensitivity was extremely high (often higher than 90%), while the specificity was often extremely low (often about 10–20%). We therefore decided to use up-sampling techniques for training the model, in order to create models that are more balanced in terms of sensitivity and specificity.

Using class weights (i.e., imposing a heavier cost for errors made in the minority class) was tested as an alternative to up-sampling, but did not lead to a substantially different performance.

Depending on the model, different hyperparameters were tuned for training the models. For RF models, the number of features used at each split was tuned. For linear SVM, the C hyperparameter was tuned, for SVM with radial basis function kernel the C and sigma parameters were tuned, for SVM with polynomial basis function the C, degree, and scale parameters were tuned, and for GBM number of iterations and complexity of the tree were tuned, while shrinkage and minimum number of training set samples in a node to commence splitting was held constant at 0.1 and 10, respectively. Model training was done in different settings. First, models were fitted that only included baseline features (T0). Second, models were fitted that additionally included three-month change scores in OQ-45 (psychopathology) and MHC-SF (well-being) subscales and total scores. Change scores were included in the second setting, because early improvements in treatment have been shown to be a strong and unique indicator for ongoing improvement at a later moment across a range of psychiatric disorders (92–94). If such a model would perform substantially better, it would be of added value for practice to (additionally) use this model some months after the treatment started to make more accurate predictions.

Third, additional feature reduction was used in both settings, because this might avoid overfitting and lead to better generalizability and increased performance on the test set. The practical usefulness of a model would increase if a reduced set of features yields comparable or even superior performance in predicting non-improvement. Least absolute shrinkage and selection operator regression (LASSO) was used to reduce the number of features. LASSO has the advantage of shrinking less relevant weights to zero, allowing to use it to reduce the number of features (90, 95, 96). In total, this resulted in four settings used for training the models: (1) no change scores and not reduced, (2) no change scores and reduced, (3) change scores and not reduced, and (4) change scores and reduced.

The trained models were then validated in the hold-out sample using a default probability cut-off of 0.5 (82). This means that every case that had a probability higher than 50% of not being improved, was classified as 'not improved'. Performance of all models was evaluated using balanced accuracy, sensitivity, specificity, and area under the curve (AUC). Sensitivity, also known as True Positive Rate (TPR) or recall, focuses on the model's ability to correctly detect positive instances whereas specificity, also known as True Negative Rate (TNR), assesses the model's ability to correctly identify negative instances. Both sensitivity and specificity refer to a specific prediction threshold of the outcome. The AUC, on the other hand, provides a global evaluation, capturing the model's performance across the entire range of threshold choices. AUC thus provides a holistic view of performance, independent of thresholds, making it a valuable measure to assess the overall discriminatory power of our binary classification model (improvement versus non-improvement). Therefore, the AUC was used as the primary evaluation measure in this study. Guidelines for interpreting AUC scores suggest that scores from 0.5 to 0.59 can be seen as extremely poor, from 0.60 to 0.69 as poor, 0.70 to 0.79 as fair, 0.80 to 0.89 as good and > 0.90 as excellent (97).

To be better able to interpret the models and for reasons of conciseness, we additionally determined the top five most important features in the hold-out sample of each model in the four different settings. Feature importance was determined using the varImp evaluation function from the caret package, a generic calculation method and analysis technique for statistical modeling. It evaluates the impact of each predictor feature by assessing how much the model's performance deteriorates when a particular feature is removed. By measuring the relative contribution of the features, it helps in understanding the ranking of influence on the prediction of non-improvement, ensuring further model optimization. Depending on the type of model, different metrics are used to determine feature importance [for an overview, see Kuhn, (86)].

3. Results

3.1. Sample

At baseline, 755 patients receiving outpatient treatments within multidisciplinary teams consisting of psychologists, psychiatrists, nurses and art therapists, were included in the dataset. Most patients were female, followed lower (43%), intermediate (37.1%) or higher (19.9%) vocational education, and lived with a partner and children (see Table 2). Almost one third had social, relation and/or work problems. The respondents were classified into five common psychopathological groups based on their primary diagnosis: depressive disorder ($n = 417$; 55.2%), bipolar disorder ($n = 79$; 10.5%), anxiety disorder ($n = 114$; 15.1%), trauma related disorder ($n = 115$; 15.2%) or obsessive-compulsive disorder ($n = 30$; 4.0%). Most patients had comorbid disorders ranging from attention deficit hyperactivity disorder (ADHD), depression, anxiety, trauma or addiction, and/or had personality problems respectively: depressive disorder (5.0%; 31.3%), bipolar disorder (6.3%; 1.3%), anxiety disorder (8.8%; 29.8%), trauma related disorder (14.8%; 32.2%) or obsessive-compulsive disorder (OCD; 10.0%; 26.7%).

3.2. Psychopathology and well-being per diagnosis over time

For descriptive purposes, Figure 1 shows the average OQ-45 symptomatic distress scale scores over the 12-month time span for

10.3389/fpsyt.2023.1236551

TABLE 2 Major characteristics of respondents ($N = 755$).

	Depression ($n = 417$) (55.2%)		Bipolar ($n = 79$) (10.5%)		Anxiety ($n = 114$) (15.1%)		Trauma ($n = 115$) (15.2%)		OCD ($n = 30$) (4.0%)		Total ($N = 755$)	
Gender *n* (%)												
Male	190	(45.6)	32	(40.5)	45	(39.5)	35	(30.4)	8	(26.7)	310	(41.1)
Female	227	(54.4)	47	(59.5)	69	(60.5)	80	(69.6)	22	(73.3)	445	(58.9)
Age												
Mean	46.0		45.6		39.3		41.0		36.8		43.8	
Range	20–65		25–64		21–62		19–63		21–65		19–65	
SD	10.8		10.0		10.4		10.6		12.0		11.1	
Level of education *n* (%)[a]												
Low	182	(46.8)	13	(19.1)	43	(39.4)	60	(53.6)	6	(20.7)	304	(43.0)
Moderate	143	(36.8)	29	(42.6)	45	(41.3)	33	(29.5)	12	(41.4)	262	(37.1)
High	64	(16.5)	26	(38.2)	21	(19.3)	19	(17.0)	11	(37.9)	141	(19.9)
Marital status *n* (%)												
Single without children	77	(18.9)	14	(19.7)	17	(14.9)	30	(26.5)	2	(6.7)	140	(19.0)
Single with children	30	(7.4)	6	(8.5)	13	(11.4)	16	(14.2)	1	(3.3)	66	(9.0)
Married without children	93	(22.9)	12	(15.2)	22	(19.3)	17	(15.0)	9	(30.0)	153	(20.8)
Married with children	161	(39.6)	33	(46.5)	36	(31.6)	33	(29.2)	11	(36.7)	274	(37.3)
Other	46	(11.3)	6	(8.5)	26	(22.8)	17	(15.0)	7	(23.3)	102	(13.9)
Comorbid society problems *n* (%)												
House problem	18	(4.3)	0	(0)	2	(1.8)	2	(1.7)	2	(6.7)	24	(3.2)
Work problem	112	(27.0)	14	(17.7)	31	(27.4)	29	(25.2)	6	(20.0)	192	(25.3)
Relation problem	109	(26.3)	3	(3.8)	21	(18.6)	28	(24.3)	3	(10.0)	164	(21.8)
Social problem	126	(30.4)	10	(12.7)	30	(26.5)	33	(28.7)	6	(20.0)	205	(27.3)
Financial problem	59	(14.2)	1	(1.3)	11	(9.7)	13	(11.3)	3	(10.0)	87	(11.6)
Somatic problem	61	(14.6)	3	(3.8)	21	(18.4)	8	(7.0)	2	(6.7)	95	(12.6)
Comorbid diagnosis *n* (%)												
None	273	(65.5)	58	(73.4)	67	(58.8)	42	(36.5)	19	(63.3)	459	(60.8)
Two or more	21	(5.0)	5	(6.3)	10	(8.8)	17	(14.8)	3	(10.0)	56	(7.4)
Personality problems	130	(31.3)	1	(1.3)	34	(29.8)	37	(32.2)	8	(26.7)	225	(29.8)
Nature all comorbid diagnoses *n* (%)												
ADHD	24	(5.8)	12	(15.2)	6	(5.3)	16	(13.9)	1	(3.3)	59	(7.8)
Depression	–	–	–	–	25	(21.9)	27	(23.5)	6	(20.0)	58	(7.7)
Anxiety	32	(7.7)	0	(0)	–	–	4	(3.5)	1	(3.3)	37	(4.9)
Trauma	34	(8.2)	5	(6.3)	6	(5.3)	–	–	0	(0)	45	(6.0)
Addiction	19	(4.6)	2	(2.5)	2	(1.8)	6	(5.2)	1	(3.3)	30	(4.0)
Other	35	(8.4)	2	(2.5)	8	(7.0)	20	(17.4)	2	(6.7)	67	(8.9)

[a]Low = primary school, lower vocational education; moderate = secondary school, intermediate vocational education; high = higher vocational education, university.

the different diagnostic categories, as well as the percentages of patients who did or did not improve by more than half an SD compared to baseline. For patients with depressive disorder, a continuous improvement from baseline to 12-month follow-up seemed to be present in the total OQ-45 scores. For patients with anxiety disorder, it seemed that on average no improvement was present after six-month follow-up. The binary improvement data suggests that the largest proportion of improvement happened within the first 3 months. The increase in percentage improved after this point seemed very small for all diagnostic groups. The percentage of improved patients in the trauma-related disorder group seemed especially small.

Figure 2 summarizes the course of total well-being scores over the period of 12 months and the proportion of patients that improved

10.3389/fpsyt.2023.1236551

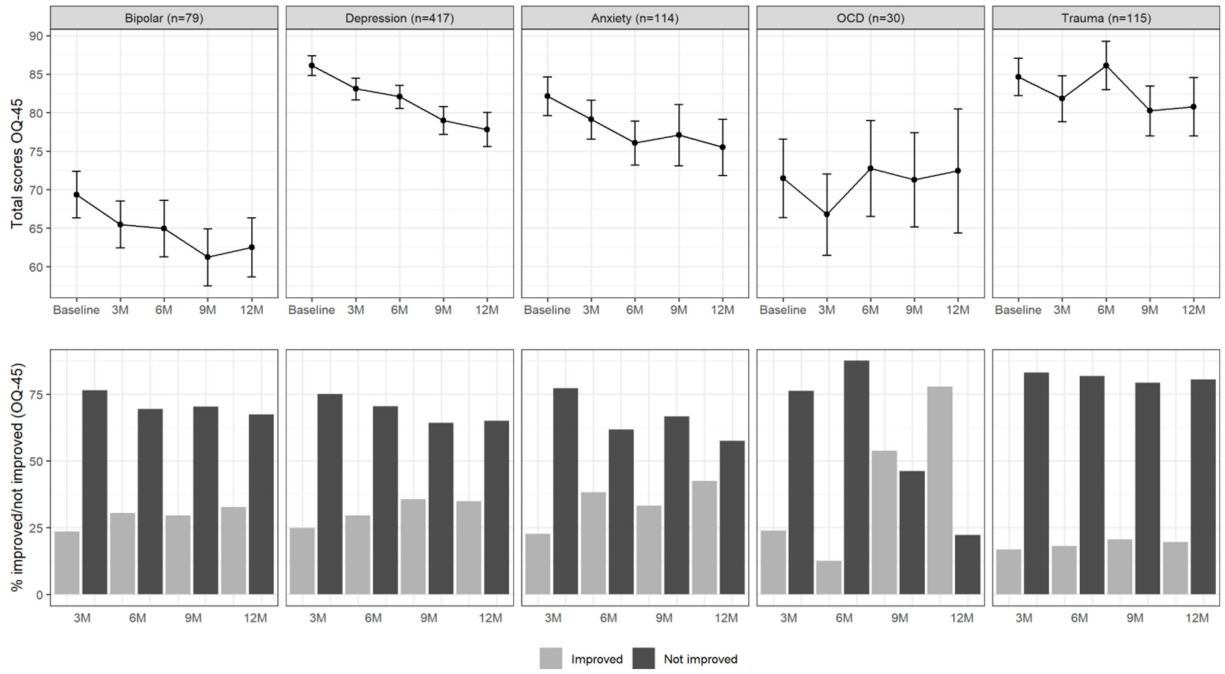

FIGURE 1
Total OQ-45 scores (upper panel) and percentage of improved and not improved patients (lower panel) per diagnosis group and over time. The error bars in the upper panel represent 95% confidence intervals.

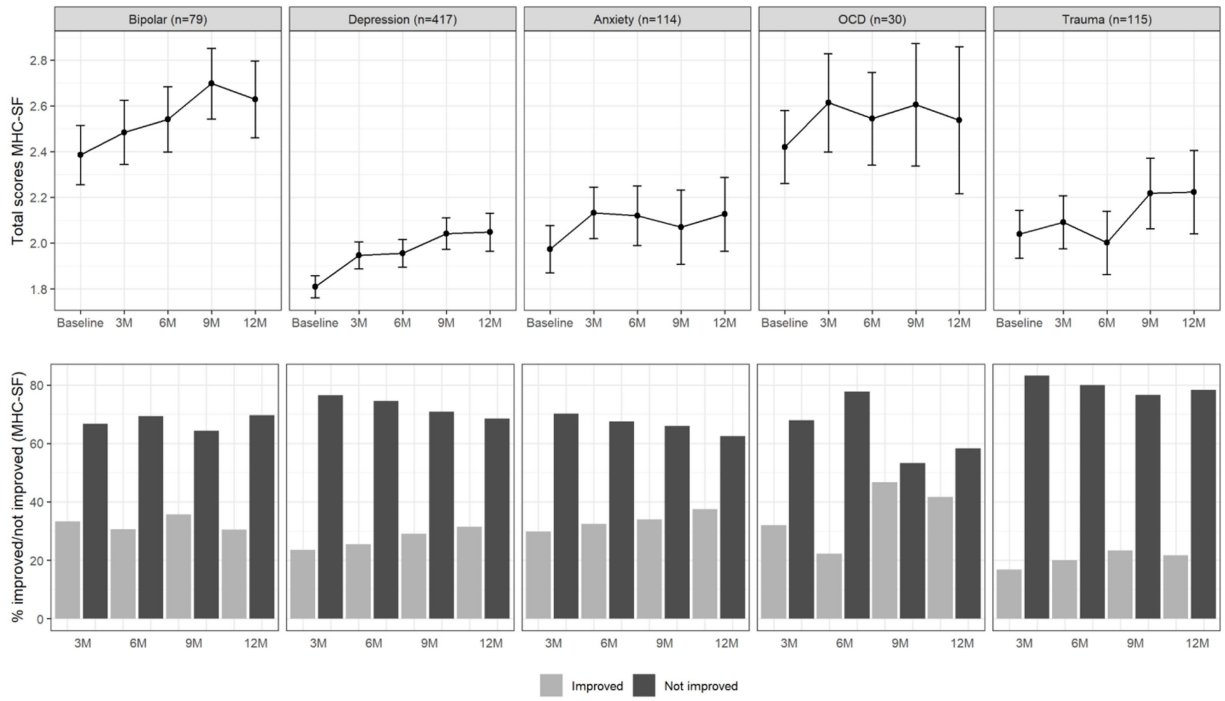

FIGURE 2
Total MHC-SF scores (upper panel) and percentage of improved and not improved patients (lower panel) per diagnosis group and over time. The error bars in the upper panel represent 95% confidence intervals.

(> 0.5 SD) in well-being. Overall, a similar picture emerged. Improvements in well-being appeared to happen mainly within the first 3 months, while the increase in improvements after this point remained rather small.

3.3. Feature reduction

Table 3 gives an overview of features that were included in the models without change scores and with change scores after LASSO regression was applied as additional preparatory step. In models without change scores, the only psychological feature of the 16 remaining features after feature reduction was the baseline total score of the OQ-45, while all others were demographic, diagnostic and clinical features. In models with change scores of the remained 13, psychological features of both the OQ-45 and the MHC-SF turned out to be of interest.

3.4. Predicting improvement at 6 months

In the training set, 70% of cases did not improve and in the hold-out sample 71% of cases did not improve. An overview of the performance of all models under the four different settings can be found in Table 4. Overall, the models performed best when change scores were included. In settings in which early change scores were included (from 0 to 3 months), the highest overall performance on the training set was obtained (AUC range: 0.79–0.84). The models in this setting also performed best on the hold-out sample (AUC range: 0.69–0.73). The best performing overall model in the hold-out sample in settings with change scores included was gradient boosting (AUC = 0.73). The models performed relatively poor in settings without change scores. In the training set, modest AUC values were found in these settings, ranging from 0.67 to 0.73. The best performance in the hold-out sample when no change scores were included was found for logistic regression (AUC = 0.63). Overall, these findings suggest that including change score substantially improves model performance in this dataset. An overview of all final hyperparameters after model training can be found in Table 5.

Another important comparison included settings in which reduced sets of features were used versus settings in which no reduced sets were used. Overall, the findings suggest that using a reduced set of features seemed to somewhat improve the performance in the training set. Yet, when validating the models on the hold-out sample it seems that using a reduced set of features does not substantially contribute the performance of the models. This indicates that using a reduced set of features does not decrease performance of the models to a relevant degree, suggesting that a reduced set of features might have a similar predictive ability compared with the full set of baseline features. The confusion matrices of the best performing models in the hold-out sample within each setting can be found in Table 6.

TABLE 3 Overview of features that were included after feature reduction was applied using LASSO regression.

Model without change scores ($k = 16$)	Model with change scores ($k = 13$)
OQ-45 symptomatic distress	OQ-45 symptomatic distress
Gender	Change score OQ-45 interpersonal relations
Working problems	Change score OQ-45 somatic complaints
Living problems	Change score OQ-45 symptomatic distress
Log-transformed treatment intensity[a]	Change score MHC-SF total score
Education: moderate	Change score MHC-SF emotional well-being
Living situation: no partner	Main diagnosis: trauma
Living situation: other	Main diagnosis: anxiety
Comorbidity	Second comorbidity
Second comorbidity	Living situation: other
Main diagnosis: trauma	Working problems
Main diagnosis: anxiety	Living problems
Sum of previous enrollments in years: 0–2	Social problems
Sum of previous enrollments in years: 5+	–
Number of treatments in the past: 1–4	–
Number of treatments in the past: 5–10	–

MHC-SF, Mental Health Continuum-Short Form; OQ-45, Outcome Questionnaire. All change scores refer to change from baseline to 3-month follow-up.

3.5. Feature importance

To allow for some interpretation of the models, one last step was to identify the most important features from the models that showed the best performance on the hold-out sample in each setting. An overview of these five most important features can be found in Table 7. It is noteworthy that change scores seem to play a crucial role in the models that include change scores. This, again, suggests that including information about change within the beginning of treatment seems to be valuable when aiming to improve model accuracy. Furthermore, in all settings, except the second setting, both psychopathology and well-being are among the most important features. This indicates that not only psychopathology seems to be of importance when predicting improvement in symptoms, but also well-being.

4. Discussion

4.1. Main findings

The goal of the current study was to evaluate and compare the performance of different machine learning (ML) models in predicting non-improvement in an observational sample of patients treated in routine specialized mental healthcare. Below, the results are critically discussed in the light of previous research and opportunities for future research.

First, the ML models applied in the current study showed only modest performance in predicting treatment outcomes. Although some previous prediction studies show relatively good predictive results [e.g., (98–100)], most previous studies also indicate modest performance [e.g., (30, 53, 57, 70, 73, 101, 102)]. Some explanations for the modest performance in the current study should be considered. Firstly, 'confounding by indication' could have introduced a bias into the observed association of observed features and non-improvement (103). The decision to assign (intensity of) treatment or adjustments along the way can be influenced by various factors, such as disease severity, previous

10.3389/fpsyt.2023.1236551

TABLE 4 Model performance metrics of the six algorithms under different conditions in the training and hold-out sample.

		Training sample (n = 344)				Hold-out sample (n = 146)			
Setting	Algorithm	ACC_{Bal}	Sens	Spec	AUC	ACC_{Bal}	Sens	Spec	AUC
No change scores, not reduced	Logistic regression	0.61	0.66	0.56	0.68	0.59	0.64	0.54	0.63
	Random forest	0.62	0.67	0.57	0.67	0.52	0.59	0.44	0.58
	SVM (linear)	0.63	0.66	0.60	0.68	0.58	0.65	0.51	0.60
	SVM (radial)	0.62	0.70	0.54	0.69	0.56	0.65	0.47	0.62
	SVM (polynomial)	0.62	0.69	0.56	0.69	0.54	0.59	0.49	0.58
	Gradient boosting	0.61	0.68	0.54	0.67	0.54	0.71	0.37	0.58
No change scores, reduced	Logistic regression	0.66	0.69	0.64	0.73	0.59	0.63	0.56	0.62
	Random forest	0.65	0.68	0.62	0.71	0.56	0.61	0.51	0.58
	SVM (linear)	0.66	0.67	0.65	0.73	0.58	0.58	0.58	0.61
	SVM (radial)	0.66	0.70	0.63	0.73	0.56	0.61	0.51	0.59
	SVM (polynomial)	0.67	0.67	0.67	0.73	0.61	0.60	0.63	0.60
	Gradient boosting	0.65	0.67	0.63	0.72	0.59	0.67	0.51	0.62
Change scores, not reduced	Logistic regression	0.70	0.76	0.64	0.79	0.65	0.77	0.53	0.69
	Random forest	0.68	0.88	0.47	0.80	0.65	0.93	0.37	0.71
	SVM (linear)	0.72	0.76	0.67	0.80	0.67	0.76	0.58	0.69
	SVM (radial)	0.72	0.77	0.67	0.80	0.63	0.70	0.56	0.71
	SVM (polynomial)	0.72	0.75	0.68	0.81	0.63	0.73	0.53	0.68
	Gradient boosting	0.73	0.77	0.69	0.81	0.66	0.77	0.56	0.71
Change scores, reduced	Logistic regression	0.74	0.78	0.70	0.83	0.65	0.74	0.56	0.69
	Random forest	0.74	0.81	0.66	0.83	0.64	0.74	0.54	0.69
	SVM (linear)	0.74	0.77	0.71	0.84	0.66	0.74	0.58	0.69
	SVM (radial)	0.73	0.76	0.69	0.81	0.63	0.72	0.54	0.70
	SVM (polynomial)	0.74	0.76	0.71	0.84	0.67	0.76	0.58	0.70
	Gradient boosting	0.74	0.78	0.71	0.83	0.64	0.77	0.52	0.73

treatments, or patient preferences. It is possible that the predictors that drive treatment assignment, in this case confounding features, could have effected the treatment outcome and have made it difficult to assess the true predictive nature of the features considered in this study (103). Secondly, in real-world scenarios, external factors or sources of noise could have affected the outcome and introduced unpredictability. These factors may not be captured by the available features. Accounting for such factors or acquiring additional relevant data might help improve performance. Feature selection, domain expertise, or acquiring additional relevant features can potentially enhance the model's performance. The challenge remains to add the right features predicting treatment success (104). Thirdly, in the current study treatment success is assessed based on subjective self-reported measures. The patient's responses to outcome measures might be influenced by their desire to align their responses with the clinician's expectations. This can result in inflated self-reported outcomes, leading to reduced accuracy in predicting treatment success. People respond inconsistently over time, but algorithms assume no response bias (105). These potential errors undermine prediction. ML techniques *per se* aren't a panacea for higher accuracy without a quality dataset of informative and relative features and domain-specific considerations (106, 107).

Second, more complicated and flexible ML models did not perform substantially better than logistic regression. This is in line with a review of 71 clinical prediction modeling studies (108) and with a recent prediction study of eating disorder treatment response by Espel-Huynh et al. (98). One explanation for this finding might be that the feature set in the current study was not large enough for the more complex models to have an advantage over logistic regression. ML algorithms lead to better performance including in the prevention of the risk of overfit with a greater number of predictors than traditional statistical methods (109). More studies have to be conducted to investigate which model works best in which circumstances (60, 108, 110). Further research into the possibilities of ML methods is still warranted since traditional regression-related approaches have various potential limitations, such as the assumption of straightforward linearity, which may render them less suitable for investigating the complex relational patterns between varied predictors for treatment success in mental healthcare (58, 111).

Third, although still modest, models that included change scores showed the highest overall performance in the hold-out sample, with the gradient boosting model achieving the best overall performance. Models without change scores performed poorly overall. These findings suggest that including change scores substantially improves

10.3389/fpsyt.2023.1236551

TABLE 5 Final hyperparameters used for prediction in the hold-out sample after model training.

Setting	Algorithm	Hyperparameter
No change scores, not reduced	Logistic regression	NA
	Random forest	mtry = 1
	SVM (linear)	C = 0.01
	SVM (radial)	C = 0.5, sigma = 0.02
	SVM (polynomial)	C = 0.25, degree = 3, scale = 0.01
	Gradient boosting	nTrees = 150, ID = 1, shrinkage = 0.1, NT = 10
No change scores, reduced	Logistic regression	NA
	Random forest	mtry = 1
	SVM (linear)	C = 0.01
	SVM (radial)	C = 0.25, sigma = 0.04
	SVM (polynomial)	C = 0.25, degree = 2, scale = 0.01
	Gradient boosting	nTrees = 150, ID = 1, shrinkage = 0.1, NT = 10
Change scores, not reduced	Logistic regression	NA
	Random forest	mtry = 2
	SVM (linear)	C = 0.01
	SVM (radial)	C = 0.25, sigma = 0.01
	SVM (polynomial)	C = 0.25, degree = 1, scale = 0.01
	Gradient boosting	nTrees = 50, ID = 1, shrinkage = 0.1, NT = 10
Change scores, reduced	Logistic regression	NA
	Random forest	mtry = 1
	SVM (linear)	C = 0.01
	SVM (radial)	C = 0.5, sigma = 0.06
	SVM (polynomial)	C = 0.5, degree = 1, scale = 0.01
	Gradient boosting	nTrees = 100, ID = 1, shrinkage = 0.1, NT = 10

C, C-parameter; ID, Interaction depth; mtry, number of features used at each split; NA, Not applicable; nTrees, Number of trees; NT, number of training set samples in a node to commence splitting. For all gradient boosting models, shrinkage and NT were held constant at 0.1 and 10, respectively.

TABLE 6 Confusion matrices of the best performing models in the hold-out sample within each setting.

		Reference	
Setting 1: Logistic regression		Non-improvement	Improvement
Predicted	Non-improvement	66	20
	Improvement	37	23
Setting 2: Gradient boosting			
Predicted	Non-improvement	69	2
	Improvement	34	22
Setting 3: Gradient boosting			
Predicted	Non-improvement	79	19
	Improvement	24	24
Setting 4: Gradient boosting			
Predicted	Non-improvement	79	21
	Improvement	24	22

prediction performance in this setting. Improvement in the first months has often been found to be related to later treatment success in other studies as well (93, 94) and early change predicts outcome even better than patient characteristics (92, 112, 113). This underscores the relevance of continuous treatment effect monitoring and treatment adjustments in clinical practice.

TABLE 7 Five most important features of the best performing models in each setting.

	Setting 1: Logistic regression	Setting 2: Gradient boosting	Setting 3: Gradient boosting	Setting 4: Gradient boosting
Feature 1	OQ-45 symptomatic distress	OQ-45 symptomatic distress	Change score OQ-45 symptomatic distress	Change score OQ-45 symptomatic distress
Feature 2	Treatment intensity	Treatment intensity	Change score OQ-45 somatic complaints	Change score OQ-45 somatic complaints
Feature 3	OQ-45 social role performance	Number of previous treatments: 5–10	OQ-45 symptomatic distress	Change score MHC-SF total score
Feature 4	GAF score	Working problems	Change OQ-45 interpersonal relations	OQ-45 symptomatic distress
Feature 5	MHC-SF total score	Main diagnosis: anxiety	Change score MHC-SF total score	Living problems

Fourth, the feature-reduced models demonstrated no relevant decrease in performance for predicting treatment outcomes at 6 months in the hold-out sample. Feature-reduced models potentially prevent overfitting and increase generalizability. A trade-off exists between interpretability and accuracy when choosing algorithms. Reducing features also improves the explainability of ML based prediction models. Additional, if a reduced set of features performs equally well (or even better) in predicting non-improvement, it would also increase the practical value and implementability of such a model in daily clinical practice.

Finally, analysis of the feature importance across the different model settings suggested that the most relevant features were the 0–3 month change scores in symptomatic distress, somatic complaints, and well-being, as well as baseline symptomatic distress. The importance of monitoring both the level of psychopathology and well-being in patients with mental health problems has been demonstrated more often (81, 114–118). Crucial predictors found in prior research, including chronicity, comorbidity, interpersonal functioning and familial problems (119), seemed less relevant for predicting non-response in the current study.

For practice, past and present findings underline the importance of searching for additional features to better predict treatment effect in real-world treatment context. Hilbert et al. (73) argued previously that prediction models developed within a diagnostically homogeneous sample are not necessarily superior to a more diverse sample that includes different diagnostic groups. The current study shows that the specific main diagnosis has less predictive value than, for example, early change in treatment effect. After all, where psychiatric patients differ enormously in severity, duration or symptoms of psychopathology and in risk of recurrence, treatments in daily care differ in used methods, assumed mechanisms and appointment frequency. Even within a specific diagnostic group, tailoring psychotherapeutic interventions specifically to the circumstances and characteristics of the patient can improve treatment outcomes (16, 120, 121). Depending on the context and goal of a ML model, one might want to adjust the probability cut-off for predicting non-improvement. We decided to use a probability cut-off of 50% for predicting non-improvement, because we did assume the cost of misspecification to be equal for the positive and negative class. For example, if one wants to aim for a model that has higher sensitivity, lowering the threshold could be desirable.

4.2. Strengths and limitations

The current study is one of the first to explore the potential of different machine learning models to predict treatment outcomes in a real-world mental healthcare context using a wide range of routinely available sociodemographic, clinical and patient-reported outcome data. There are however some limitations to the current study that need to be considered.

First, although the current study used a cross-validation approach by randomly splitting the dataset into a training and a test sample, which is the common approach in ML, it should be noted that the study is still exploratory in nature, Although common practice in ML, the test set consisted of a random subset from the same overall patient sample and therefore the study was still limited in its ability to test the generalizability of the final models. Confirmatory studies in independent datasets from different contexts are still necessary to further examine the robustness of the prediction models (122).

Second, in the context of the routine collection of patient-reported outcome data, data is often missing during the course of the treatment process because patients have already improved sufficiently or, on the contrary, have not improved. This missing data is not at random, resulting in the ML algorithms to ultimately relate to a select and biased subpopulation that continues to receive treatment for at least a certain period of time.

Third, the features available in this study consisted largely of self-report data. For the future it would be interesting to incorporate more objective features such as psychological measurements into ML models (123, 124). Future ML studies could improve mental health predictions by adding a unique source of high-frequent and continuous data collecting using multi-modal assessment tools during the period of treatment. mHealth (mobile health) provides individuals real-time biofeedback via sensor apps in everyday devices such as smartphones or wearables on physiological or self-reported behavioral and state parameters, such as heart rate, sleep patterns, physical activity or stress levels (124–126). The combination of ML and mHealth, despite challenges in dimensionality, ethics, privacy and security, shows promise as a clinical tool for monitoring populations at risk and forms the basis for the next generation of mHealth interventions (124, 125).

Finally, though the chosen criterion of 0.5 SD for non-improvement is often used [e.g., (127–131)], a disadvantage is

that this cutoff is sample-dependent. Also, an improvement of 0.5 SD does not necessarily mean that a patient has recovered in such a way that (s)he no longer has clinically relevant complaints. Future research could consider to use the Jacobson-Truax concept of the Reliable Change Index (RCI), which considers the reliability of the improvement in the context of the overall distribution that the patient is likely to belong to post-treatment (132). Patients moving reliably into the functional distribution are *recovered*. Patients are considered to have *improved* if they have made a reliable change but remain in the dysfunctional population, *unchanged* if they have not made a reliable change, and *deteriorated* if they have reliably worsened (132).

4.3. Clinical implications and recommendations for future research

Some recommendations can be made for future research. On the one hand, the use of sophisticated psychological data with relevant features according to the latest theoretical models may increase predictions and thereby improve decision-making on therapy indication. This could include the therapeutic relationship as a known predictor of interest (133) diagnosis specific questionnaires in addition to generics, which could mean that the case for a transdiagnostic approach may not yet have been settled, or program-specific questionnaires, appropriate to the therapy offered. On the other hand, the development of more advanced tools is necessary to detect predictors for treatment response based on high-dimensional patient data (134). Based on current research, practitioners might decide to stop or adjust a treatment. In the future, it is desirable that patients can be indicated in a more targeted manner. After all, at present ML approaches cannot yet contribute to specific individualized clinical judgments (135). We would encourage future studies to develop predictors over rather broad diagnostic patient groups and not exclude features in advance, but use the full potential of information available in patient EHRs (136). Interestingly, ML techniques offer the opportunity to study patients who are underrepresented in RCTs.

Additionally, ML has the potential to benefit mental healthcare as it can account for the interaction between many features (137). The ML techniques are suitable to detect features with the strongest predictive influence in different contexts and mutual interactions, thereby providing a combined measure of both individual and multivariate impact of each feature (138). Subsequently, based on findings, the number of features to be implemented in daily care can be substantially reduced.

To reduce response bias, improve the predictive performance of the model, and provide a more comprehensive picture of treatment success, it may be helpful to consider multiple perspectives and assessment sources. In addition, it is important to recognize and address the potential discrepancies between the assessments of different stakeholders (e.g., clinician and patient) when defining the criterion for treatment success in predictive studies.

As change scores in both psychopathology and well-being proved relevant, implementing change measurements in ML applications could be more standardized. Therefore, for future studies, we recommend that in addition to predicting changes in psychopathology, algorithms to predict non-improvements in well-being and other domain/construction should be included. Also, adding multiple change scores, such as living conditions in daily activities and social relationships, or compliance with homework-related adherence could be relevant (139). Adding other data modalities, such as the relationship with the patient's life story, or test data could also improve prediction performance (140, 141). In any case, it is advisable to closely monitor changes in psychopathology and well-being in clinical practice and decision making from the very beginning, so that timely adjustments can be made in the therapy of non-responders. Tiemens et al. (142) recommend doing this at least 4 weeks after starting treatment. The measurement of change scores is also important because the use of feedback based on these evaluations in itself has a positive effect on complaint reduction and it can shorten the duration of treatment (143, 144).

Finally, applying both ML and traditional statistical approaches in the same study allows for comparisons (109, 145). By learning from unique strengths and limitations of different ML algorithms, future ML research can contribute to increasingly accurate predictions (146).

5. Conclusion

In the current study we applied ML techniques in a real-world mental healthcare patient population to predict non-improvement using sociodemographic, psychological, diagnostic and clinical data. The overall conclusion is that working with a reduced set of data, and implementing early change scores and relatively simple models gives the best results, both in terms of accuracy and broader in interpretability and applicability. Our results show that ML can be used as a step to indicate treatment change in an early stage of treatment, where it seems to be important to use psychopathology and well-being as important features. The results are encouraging and provide an important step to use patient specific and routine collected patient-level outcome data in clinical practice to help individual patients and clinicians select the right treatments. ML may help to bridge the gap between science and practice. None of the ML applications were developed to replace the clinician, but instead were designed to advance the clinicians' skills and treatment outcome (147). ML might become part of evidence-based practice, as a source of valuable information in addition to clinical knowledge and existing research evidence.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: Data, syntax and output files can be found on the Open Science Framework website (https://osf.io/xwme4/).

Ethics statement

Ethical approval was not required for the studies involving humans because as data were collected in the context of regular care and only anonymized data were analyzed, the study did not require medical ethical approval according to Dutch law. The studies were conducted in accordance with the local legislation and institutional

requirements. The participants provided their written informed consent to participate in this study.

Author contributions

KF collected the data, organized the database, and wrote the first draft of the manuscript. KF, JK, and PK contributed substantial to the conception, design and manuscript draft, and ensuring that the work was appropriately investigated and resolved. JK implemented machine learning algorithms and statistical analysis and wrote the first method section of the manuscript. EB and GW supervised and critically reviewed the manuscript. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. World Health Organization. *World mental health report: Transforming mental health for all*. Geneva: World Health Organization (2022).

2. Whiteford HA, Ferrari AJ, Degenhardt L, Feigin V, Vos T. The global burden of mental, neurological and substance use disorders: an analysis from the global burden of disease study 2010. *PLoS One*. (2015) 10:e0116820. doi: 10.1371/journal.pone.0116820

3. Chisholm D, Sweeny K, Sheehan P, Rasmussen B, Smit F, Cuijpers P, et al. Scaling-up of treatment of depression and anxiety–Authors' reply. *Lancet Psychiatry*. (2016) 3:603–4. doi: 10.1016/S2215-0366(16)30131-6

4. Jennings C, Singh B, Oni H, Mazzacano A, Maher C. A needs assessment for self-management services for adults awaiting community-based mental health services. *BMC Public Health*. (2023) 23:1–10. doi: 10.1186/s12889-023-15382-8

5. Deisenhofer AK, Delgadillo J, Rubel JA, Boehnke JR, Zimmermann D, Schwartz B, et al. Individual treatment selection for patients with posttraumatic stress disorder. *Depress Anxiety*. (2018) 35:541–50. doi: 10.1002/da.22755

6. Delgadillo J, Huey D, Bennett H, McMillan D. Case complexity as a guide for psychological treatment selection. *J Consult Clin Psychol*. (2017) 85:835–53. doi: 10.1037/ccp0000231

7. Lambert MJ, Hansen NB, Finch AE. Patient-focused research: using patient outcome data to enhance treatment effects. *J Consult Clin Psychol*. (2001) 69:159–72. doi: 10.1037/0022-006X.69.2.159

8. Ægisdóttir S, White MJ, Spengler PM, Maugherman AS, Anderson LA, Cook RS, et al. The meta-analysis of clinical judgment project: fifty-six years of accumulated research on clinical versus statistical prediction. *Couns Psychol*. (2006) 34:341–82. doi: 10.1177/0011000005285875

9. Carvalho A, McIntyre R. *Treatment-resistant mood disorders*. Oxford: Oxford Psychiatry Library (2015).

10. Cuijpers P, Karyotaki E, Weitz E, Andersson G, Hollon SD, van Straten A. The effects of psychotherapies for major depression in adults on remission, recovery and improvement: a meta-analysis. *J Affect Disord*. (2014) 159:118–26. doi: 10.1016/j.jad.2014.02.026

11. De Vos JA, LaMarre A, Radstaak M, Bijkerk CA, Bohlmeijer ET, Westerhof GJ. Identifying fundamental criteria for eating disorder recovery: a systematic review and qualitative meta-analysis. *J Eat Disord*. (2017) 5:1–14. doi: 10.1186/s40337-017-0164-0

12. Driessen E, Van HL, Don FJ, Peen J, Kool S, Westra D, et al. The efficacy of cognitive-behavioral therapy and psychodynamic therapy in the outpatient treatment of major depression: a randomized clinical trial. *Am J Psychiatr*. (2013) 170:1041–50. doi: 10.1176/appi.ajp.2013.12070899

13. Barth J, Munder T, Gerger H, Nuesch E, Trelle S, Znoj H, et al. Comparative efficacy of seven psychotherapeutic interventions for patients with depression: a network meta-analysis. *PLoSMedicine*. (2013) 10:e1001454. doi: 10.1371/journal.pmed.1001454

14. Cuijpers P, Cristea IA, Karyotaki E, Reijnders M, Huibers MJ. How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence. *World Psychiatry*. (2016) 15:245–58. doi: 10.1002/wps.20346

15. Mangolini VI, Andrade LH, Lotufo-Neto F, Wang Y-P. Treatment of anxiety disorders in clinical practice: a critical overview of recent systematic evidence. *Clinics*. (2019) 74:e1316. doi: 10.6061/clinics/2019/e1316

16. Norcross JC, Lambert MJ. Psychotherapy relationships that work III. *Psychotherapy*. (2018) 55:303–15. doi: 10.1037/pst0000193

17. Bolton D. Overdiagnosis problems in the DSM-IV and the new DSM-5: can they be resolved by the distress–impairment criterion? *Can J Psychiatry*. (2013) 58:612–7. doi: 10.1177/070674371305801106

18. Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. *Biol Psychiatry*. (2018) 3:223–30. doi: 10.1016/j.bpsc.2017.11.007

19. Hilbert K, Jacobi T, Kunas SL, Elsner B, Reuter B, Lueken U, et al. Identifying CBT non-response among OCD outpatients: a machine-learning approach. *Psychother Res*. (2021) 31:52–62. doi: 10.1080/10503307.2020.1839140

20. Haug T, Nordgreen T, Öst L-G, Kvale G, Tangen T, Andersson G, et al. Stepped care versus face-to–face cognitive behavior therapy for panic disorder and social anxiety disorder: predictors and moderators of outcome. *Behav Res Ther*. (2015) 71:76–89. doi: 10.1016/j.brat.2015.06.002

21. Cloitre M, Petkova E, Su Z, Weiss BJ. Patient characteristics as a moderator of posttraumatic stress disorder treatment outcome: combining symptom burden and strengths. *BJPsych open*. (2016) 2:101–6. doi: 10.1192/bjpo.bp.115.000745

22. Cohen M, Beard C, Björgvinsson T. Examining patient characteristics as predictors of patient beliefs about treatment credibility and expectancies for treatment outcome. *J Psychother Integr*. (2015) 25:90–9. doi: 10.1037/a0038878

23. Flückiger C, Del Re A, Wlodasch D, Horvath AO, Solomonov N, Wampold BE. Assessing the alliance–outcome association adjusted for patient characteristics and treatment processes: a meta-analytic summary of direct comparisons. *J Couns Psychol*. (2020) 67:706–11. doi: 10.1037/cou0000424

24. Goddard E, Wingrove J, Moran P. The impact of comorbid personality difficulties on response to IAPT treatment for depression and anxiety. *Behav Res Ther*. (2015) 73:1–7. doi: 10.1016/j.brat.2015.07.006

25. Gregertsen EC, Mandy W, Kanakam N, Armstrong S, Serpell L. Pre-treatment patient characteristics as predictors of drop-out and treatment outcome in individual and family therapy for adolescents and adults with anorexia nervosa: a systematic review and meta-analysis. *Psychiatry Res*. (2019) 271:484–501. doi: 10.1016/j.psychres.2018.11.068

26. Hamilton KE, Dobson KS. Cognitive therapy of depression: pretreatment patient predictors of outcome. *Clin Psychol Rev*. (2002) 22:875–93. doi: 10.1016/S0272-7358(02)00106-X

27. Hoyer J, Wiltink J, Hiller W, Miller R, Salzer S, Sarnowsky S, et al. Baseline patient characteristics predicting outcome and attrition in cognitive therapy for social phobia: results from a large multicentre trial. *Clin Psychol Psychother*. (2016) 23:35–46. doi: 10.1002/cpp.1936

28. Kessler RC, van Loo HM, Wardenaar KJ, Bossarte RM, Brenner L, Ebert D, et al. Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiol Psychiatr Sci*. (2017) 26:22–36. doi: 10.1017/S2045796016000020

29. Knopp J, Knowles S, Bee P, Lovell K, Bower P. A systematic review of predictors and moderators of response to psychological therapies in OCD: do we have enough empirical evidence to target treatment? *Clin Psychol Rev*. (2013) 33:1067–81. doi: 10.1016/j.cpr.2013.08.008

30. Lutz W, Lambert MJ, Harmon SC, Tschitsaz A, Schürch E, Stulz N. The probability of treatment success, failure and duration—what can be learned from empirical data to support decision making in clinical practice? *Clin Psychol Psychother*. (2006) 13:223–32. doi: 10.1002/cpp.496

31. Mululo SCC, de Menezes GB, Vigne P, Fontenelle LF. A review on predictors of treatment outcome in social anxiety disorder. *Braz J Psychiatry*. (2012) 34:92–100. doi: 10.1590/S1516-44462012000100016

32. Salomonsson S, Santoft F, Lindsäter E, Ejeby K, Ingvar M, Öst L-G, et al. Predictors of outcome in guided self-help cognitive behavioural therapy for common mental disorders in primary care. *Cogn Behav Ther*. (2020) 49:455–74. doi: 10.1080/16506073.2019.1669701

33. Sarter L, Heider J, Witthöft M, Rief W, Kleinstäuber M. Using clinical patient characteristics to predict treatment outcome of cognitive behavior therapies for individuals with medically unexplained symptoms: a systematic review and meta-analysis. *Gen Hosp Psychiatry*. (2022) 77:11–20. doi: 10.1016/j.genhosppsych.2022.03.001

34. Eskildsen A, Hougaard E, Rosenberg NK. Pre-treatment patient variables as predictors of drop-out and treatment outcome in cognitive behavioural therapy for social phobia: a systematic review. *Nord J Psychiatry*. (2010) 64:94–105. doi: 10.3109/08039480903426929

35. Haby MM, Donnelly M, Corry J, Vos T. Cognitive behavioural therapy for depression, panic disorder and generalized anxiety disorder: a meta-regression of factors that may predict outcome. *Aust N Z J Psychiatry*. (2006) 40:9–19. doi: 10.1080/j.1440-1614.2006.01736.x

36. Aguinis H, Beaty JC, Boik RJ, Pierce CA. Effect size and power in assessing moderating effects of categorical variables using multiple regression: a 30-year review. *J Appl Psychol*. (2005) 90:94–107. doi: 10.1037/0021-9010.90.1.94

37. Luedtke A, Sadikova E, Kessler RC. Sample size requirements for multivariate models to predict between-patient differences in best treatments of major depressive disorder. *Clin Psychol Sci*. (2019) 7:445–61. doi: 10.1177/2167702618815466

38. Ribeiro DC, Milosavljevic S, Abbott JH. Sample size estimation for cluster randomized controlled trials. *Musculoskeletal Sci Pract*. (2018) 34:108–11. doi: 10.1016/j.msksp.2017.10.002

39. Rothwell JC, Julious SA, Cooper CL. A study of target effect sizes in randomised controlled trials published in the health technology assessment journal. *Trials*. (2018) 19:1–13. doi: 10.1186/s13063-018-2886-y

40. Tam W, Lo K, Woo B. Reporting sample size calculations for randomized controlled trials published in nursing journals: a cross-sectional study. *Int J Nurs Stud*. (2020) 102:103450. doi: 10.1016/j.ijnurstu.2019.103450

41. Whitehead AL, Julious SA, Cooper CL, Campbell MJ. Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. *Stat Methods Med Res*. (2016) 25:1057–73. doi: 10.1177/0962280215588241

42. Bothwell LE, Greene JA, Podolsky SH, Jones DS. Assessing the gold standard–lessons from the history of RCTs. *Mass Medical Soc*. (2016) 374:2175–81. doi: 10.1056/NEJMms1604593

43. Frieden TR. Evidence for health decision making—beyond randomized, controlled trials. *N Engl J Med*. (2017) 377:465–75. doi: 10.1056/NEJMra1614394

44. Rothwell PM. External validity of randomised controlled trials:"to whom do the results of this trial apply?". *Lancet*. (2005) 365:82–93. doi: 10.1016/S0140-6736(04)17670-8

45. Stuart EA, Bradshaw CP, Leaf PJ. Assessing the generalizability of randomized trial results to target populations. *Prev Sci*. (2015) 16:475–85. doi: 10.1007/s11121-014-0513-z

46. Saunders R, Cape J, Fearon P, Pilling S. Predicting treatment outcome in psychological treatment services by identifying latent profiles of patients. *J Affect Disord*. (2016) 197:107–15. doi: 10.1016/j.jad.2016.03.011

47. Stochl J, Soneson E, Stuart F, Fritz J, Walsh AE, Croudace T, et al. Determinants of patient-reported outcome trajectories and symptomatic recovery in improving access to psychological therapies (IAPT) services. *Psychol Med*. (2022) 52:3231–40. doi: 10.1017/S0033291720005395

48. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. (2019) 380:1347–58. doi: 10.1056/NEJMra1814259

49. Theobald O. *Machine learning for absolute beginners: A plain English introduction*, vol. *157*. London, UK: Scatterplot Press (2017).

50. Coppersmith G, Ngo K, Leary R, Wood A. *Exploratory analysis of social media prior to a suicide attempt*. Paper presented at the Proceedings of the third workshop on computational linguistics and clinical psychology. (2016).

51. De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M. *Discovering shifts to suicidal ideation from mental health content in social media*. Paper presented at the Proceedings of the 2016 CHI conference on human factors in computing systems. (2016).

52. Galatzer-Levy IR, Karstoft K-I, Statnikov A, Shalev AY. Quantitative forecasting of PTSD from early trauma responses: a machine learning application. *J Psychiatr Res*. (2014) 59:68–76. doi: 10.1016/j.jpsychires.2014.08.017

53. Haynos AF, Wang SB, Lipson S, Peterson CB, Mitchell JE, Halmi KA, et al. Machine learning enhances prediction of illness course: a longitudinal study in eating disorders. *Psychol Med*. (2021) 51:1392–402. doi: 10.1017/S0033291720000227

54. Jaques N, Taylor S, Nosakhare E, Sano A, Picard R. *Multi-task learning for predicting health, stress, and happiness*. Paper presented at the NIPS Workshop on Machine Learning for Healthcare. (2016).

55. Shatte AB, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. *Psychol Med*. (2019) 49:1426–48. doi: 10.1017/S0033291719000151

56. Mitchell TM. Does machine learning really work? *AI Mag*. (1997) 18:11–1.

57. Iniesta R, Malki K, Maier W, Rietschel M, Mors O, Hauser J, et al. Combining clinical variables to optimize prediction of antidepressant treatment outcomes. *J Psychiatr Res*. (2016) 78:94–102. doi: 10.1016/j.jpsychires.2016.03.016

58. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science*. (2015) 349:255–60. doi: 10.1126/science.aaa8415

59. Chen EE, Wojcik SP. A practical guide to big data research in psychology. *Psychol Methods*. (2016) 21:458–74. doi: 10.1037/met0000111

60. Aafjes-van Doorn K, Kamsteeg C, Bate J, Aafjes M. A scoping review of machine learning in psychotherapy research. *Psychother Res*. (2021) 31:92–116. doi: 10.1080/10503307.2020.1808729

61. Malley JD, Kruppa J, Dasgupta A, Malley KG, Ziegler A. Probability machines. *Methods Inf Med*. (2012) 51:74–81. doi: 10.3414/ME00-01-0052

62. Barak-Corren Y, Castro VM, Javitt S, Hoffnagle AG, Dai Y, Perlis RH, et al. Predicting suicidal behavior from longitudinal electronic health records. *Am J Psychiatr*. (2017) 174:154–62. doi: 10.1176/appi.ajp.2016.16010077

63. Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. *Clin Psychol Sci*. (2017) 5:457–69. doi: 10.1177/2167702617691560

64. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. *Annu Rev Clin Psychol*. (2018) 14:91–118. doi: 10.1146/annurev-clinpsy-032816-045037

65. Yarkoni T, Westfall J. Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect Psychol Sci*. (2017) 12:1100–22. doi: 10.1177/1745691617693393

66. Lim B, Van der Schaar M. *Disease-atlas: Navigating disease trajectories using deep learning*. Paper presented at the Machine Learning for Healthcare Conference. (2018).

67. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*. (2019) 572:116–9. doi: 10.1038/s41586-019-1390-1

68. Alaa AM, van der Schaar M. Prognostication and risk factors for cystic fibrosis via automated machine learning. *Sci Rep*. (2018) 8:11242. doi: 10.1038/s41598-018-29523-2

69. Athreya AP, Neavin D, Carrillo-Roa T, Skime M, Biernacka J, Frye MA, et al. Pharmacogenomics-driven prediction of antidepressant treatment outcomes: a machine-learning approach with multi-trial replication. *Clin Pharmacol Ther*. (2019) 106:855–65. doi: 10.1002/cpt.1482

70. Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry*. (2016) 3:243–50. doi: 10.1016/S2215-0366(15)00471-X

71. Chernozhukov V, Demirer M, Duflo E, Fernandez-Val I. *Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India*. (2018).

72. Gong X, Hu M, Basu M, Zhao L. Heterogeneous treatment effect analysis based on machine-learning methodology. *CPT Pharmacometrics Syst Pharmacol*. (2021) 10:1433–43. doi: 10.1002/psp4.12715

73. Hilbert K, Kunas SL, Lueken U, Kathmann N, Fydrich T, Fehm L. Predicting cognitive behavioral therapy outcome in the outpatient sector based on clinical routine data: a machine learning approach. *Behav Res Ther*. (2020) 124:103530. doi: 10.1016/j.brat.2019.103530

74. Van Mens K, De Schepper C, Wijnen B, Koldijk SJ, Schnack H, De Looff P, et al. Predicting future suicidal behaviour in young adults, with different machine learning techniques: a population-based longitudinal study. *J Affect Disord*. (2020) 271:169–77. doi: 10.1016/j.jad.2020.03.081

75. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for reporting machine learning analyses in clinical research. *Circ Cardiovasc Qual Outcomes*. (2020) 13:e006556. doi: 10.1161/CIRCOUTCOMES.120.006556

76. De Jong K, Nugter MA, Polak MG, Wagenborg JE, Spinhoven P, Heiser WJ. The outcome questionnaire (OQ-45) in a Dutch population: a cross-cultural validation. *Clin Psychol Psychother*. (2007) 14:288–301. doi: 10.1002/cpp.529

77. Keyes CL, Wissing M, Potgieter JP, Temane M, Kruger A, Van Rooy S. Evaluation of the mental health continuum–short form (MHC–SF) in setswana-speaking south Africans. *Clin Psychol Psychother*. (2008) 15:181–92. doi: 10.1002/cpp.572

78. Lamers SMA, Westerhof GJ, Bohlmeijer ET, ten Klooster PM, Keyes CL. Evaluating the psychometric properties of the mental health continuum-short form (MHC-SF). *J Clin Psychol*. (2011) 67:99–110. doi: 10.1002/jclp.20741

79. De Jong K, Nugter A. De Outcome Questionnaire: psychometrische kenmerken van de Nederlandse vertaling. *Ned Tijdschr Psychol Grensgebieden*. (2004) 59:77–80. doi: 10.1007/BF03062326

80. De Jong K, Spinhoven P. De Nederlandse versie van de Outcome Questionnaire (OQ-45): een crossculturele validatie. *Psychol Gezond*. (2008) 36:35–45. doi: 10.1007/BF03077465

81. Franken K, Lamers SM, ten Klooster PM, Bohlmeijer ET, Westerhof GJ. Validation of the mental health continuum-short form and the dual continua model of well-being and psychopathology in an adult mental health setting. *J Clin Psychol*. (2018) 74:2187–202. doi: 10.1002/jclp.22659

82. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care*. (2003) 41:582–92. doi: 10.1097/01.MLR.0000062554.74615.4C

83. Cohen J. Statistical power analysis for the behavioral sciences: Jacob Cohen. *J Am Stat Assoc*. (1988) 84:19–74.

84. IBM Corp. *Released 2013. IBM SPSS statistics for windows, Version 27.0*. Armonk, NY: IBM Corp. (2020).

85. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing (2020).

86. Kuhn M. *Caret: Classification and regression training Version 6.0–86*. (2020). Available at: https://CRAN.R-project.org/package=caret.

87. Stekhoven DJ, Bühlmann P. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*. (2012) 28:112–8. doi: 10.1093/bioinformatics/btr597

88. Bishop CM, Nasrabadi NM. *Pattern recognition and machine learning*, vol. *4*. Berlin: Springer (2006).

89. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychol Bull*. (2017) 143:187–232. doi: 10.1037/bul0000084

90. Kuhn M, Johnson K. *Applied predictive modeling*, vol. *26*. Berlin: Springer (2013).

91. James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning*, vol. *103*. Berlin: Springer (2013).

92. Rubel J, Lutz W, Schulte D. Patterns of change in different phases of outpatient psychotherapy: a stage-sequential pattern analysis of change in session reports. *Clin Psychol Psychother*. (2015) 22:1–14. doi: 10.1002/cpp.1868

93. Schibbye P, Ghaderi A, Ljótsson B, Hedman E, Lindefors N, Rück C, et al. Using early change to predict outcome in cognitive behaviour therapy: exploring timeframe, calculation method, and differences of disorder-specific versus general measures. *PLoS One*. (2014) 9:e100614. doi: 10.1371/journal.pone.0100614

94. Wilson GT. Rapid response to cognitive behavior therapy. *Clin Gastroenterol Hepatol*. (1999) 6:289–92. doi: 10.1093/clipsy.6.3.289

95. Fonti V, Belitser E. *Feature selection using lasso*. VU Amsterdam research paper in business analytics, pp. 1–25. (2017).

96. Muthukrishnan R, Rohini R. *LASSO: a feature selection technique in predictive modeling for machine learning*. Paper presented at the 2016 IEEE international conference on advances in computer applications (ICACA). (2016).

97. Hosmer DW, Lemeshow S, Sturdivant RX. *Applied logistic regression*, vol. *398*. New York: John Wiley and Sons (2013).

98. Espel-Huynh H, Zhang F, Thomas JG, Boswell JF, Thompson-Brenner H, Juarascio AS, et al. Prediction of eating disorder treatment response trajectories via machine learning does not improve performance versus a simpler regression approach. *Int J Eat Disord*. (2021) 54:1250–9. doi: 10.1002/eat.23510

99. Lenhard F, Sauer S, Andersson E, Månsson KN, Mataix-Cols D, Rück C, et al. Prediction of outcome in internet-delivered cognitive behaviour therapy for paediatric obsessive-compulsive disorder: a machine learning approach. *Int J Methods Psychiatr Res*. (2018) 27:e1576. doi: 10.1002/mpr.1576

100. Tymofiyeva O, Yuan JP, Huang C-Y, Connolly CG, Blom EH, Xu D, et al. Application of machine learning to structural connectome to predict symptom reduction in depressed adolescents with cognitive behavioral therapy (CBT). *NeuroImage*. (2019) 23:101914. doi: 10.1016/j.nicl.2019.101914

101. Ball TM, Stein MB, Ramsawh HJ, Campbell-Sills L, Paulus MP. Single-subject anxiety treatment outcome prediction using functional neuroimaging. *Neuropsychopharmacology*. (2014) 39:1254–61. doi: 10.1038/npp.2013.328

102. Lutz W, Saunders SM, Leon SC, Martinovich Z, Kosfelder J, Schulte D, et al. Empirically and clinically useful decision making in psychotherapy: differential predictions with treatment response models. *Psychol Assess*. (2006) 18:133–41. doi: 10.1037/1040-3590.18.2.133

103. Kyriacou DN, Lewis RJ. Confounding by indication in clinical research. *JAMA*. (2016) 316:1818–9. doi: 10.1001/jama.2016.16435

104. Flach P. *Machine learning: The art and science of algorithms that make sense of data*. Cambridge, UK: Cambridge University Press (2012).

105. Eikelenboom M, Smit JH, Beekman AT, Kerkhof AJ, Penninx BW. Reporting suicide attempts: consistency and its determinants in a large mental health study. *Int J Methods Psychiatr Res*. (2014) 23:257–66. doi: 10.1002/mpr.1423

106. Makridakis S, Spiliotis E, Assimakopoulos V. Statistical and machine learning forecasting methods: concerns and ways forward. *PLoS One*. (2018) 13:e0194889. doi: 10.1371/journal.pone.0194889

107. Zhang GP. Avoiding pitfalls in neural network research. *IEEE Trans Syst Man Cybern Part C Appl Rev*. (2006) 37:3–16. doi: 10.1109/TSMCC.2006.876059

108. 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. doi: 10.1016/j.jclinepi.2019.02.004

109. Schwartz B, Cohen ZD, Rubel JA, Zimmermann D, Wittmann WW, Lutz W. Personalized treatment selection in routine care: integrating machine learning and statistical algorithms to recommend cognitive behavioral or psychodynamic therapy. *Psychother Res*. (2021) 31:33–51. doi: 10.1080/10503307.2020.1769219

110. Jacobucci R, Littlefield AK, Millner AJ, Kleiman E, Steinley D. *Pairing machine learning and clinical psychology: How you evaluate predictive performance matters*. United States: OSF Storage (2020).

111. Linthicum KP, Schafer KM, Ribeiro JD. Machine learning in suicide science: applications and ethics. *Behav Sci Law*. (2019) 37:214–22. doi: 10.1002/bsl.2392

112. Cooper M. *Essential research findings in counselling and psychotherapy*. Thousand Oaks, CA: Sage Publication, pp. 1–256. (2008).

113. Haas E, Hill RD, Lambert MJ, Morrell B. Do early responders to psychotherapy maintain treatment gains? *J Clin Psychol*. (2002) 58:1157–72. doi: 10.1002/jclp.10044

114. Bolier L, Haverman M, Westerhof G, Riper H, Smit F. Positive psychology interventions: a meta-analysis of randomized controlled studies. *BMC Public Health*. (2013) 13:20. doi: 10.1186/1471-2458-13-119

115. Iasiello M, Van Agteren J, Schotanus-Dijkstra M, Lo L, Fassnacht DB, Westerhof GJ. Assessing mental wellbeing using the mental health continuum—short form: a systematic review and meta-analytic structural equation modelling. *Clin Psychol Sci Pract*. (2022) 29:442–56. doi: 10.1037/cps0000074

116. Kraiss JT, Peter M, Moskowitz JT, Bohlmeijer ET. The relationship between emotion regulation and well-being in patients with mental disorders: a meta-analysis. *Compr Psychiatry*. (2020) 102:152189. doi: 10.1016/j.comppsych.2020.152189

117. Slade M. Mental illness and well-being: the central importance of positive psychology and recovery approaches. *BMC Health Serv Res*. (2010) 10:1–14. doi: 10.1186/1472-6963-10-26

118. Trompetter H, Lamers S, Westerhof GJ, Fledderus M, Bohlmeijer ET. Both positive mental health and psychopathology should be monitored in psychotherapy: confirmation for the dual-factor model in acceptance and commitment therapy. *Behav Res Ther*. (2017) 91:58–63. doi: 10.1016/j.brat.2017.01.008

119. Vall E, Wade TD. Predictors of treatment outcome in individuals with eating disorders: a systematic review and meta-analysis. *Int J Eat Disord*. (2015) 48:946–71. doi: 10.1002/eat.22411

120. Norcross JC, Wampold BE. What works for whom: tailoring psychotherapy to the person. *J Clin Psychol*. (2011) 67:127–32. doi: 10.1002/jclp.20764

121. Vermote R, Lowyck B, Luyten P, Verhaest Y, Vertommen H, Vandeneede B, et al. Patterns of inner change and their relation with patient characteristics and outcome in a psychoanalytic hospitalization-based treatment for personality disordered patients. *Clin Psychol Psychother*. (2011) 18:303–13. doi: 10.1002/cpp.713

122. Sammut C, Webb GI. *Encyclopedia of machine learning*. Berlin: Springer Science and Business Media (2011).

123. Derks YP. *Alexithymia in borderline personality pathology: From theory to a biosensor application*. Netherlands: University of Twente (2022).

124. Terhorst Y, Knauer J, Baumeister H. Smart sensing enhanced diagnostic expert systems In: C Montag and H Baumeister, editors. *Digital phenotyping and Mobile sensing: New developments in Psychoinformatics*. Berlin: Springer (2022). 413–25.

125. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annu Rev Clin Psychol*. (2017) 13:23–47. doi: 10.1146/annurev-clinpsy-032816-044949

126. Moshe I, Terhorst Y, Opoku Asare K, Sander LB, Ferreira D, Baumeister H, et al. Predicting symptoms of depression and anxiety using smartphone and wearable data. *Front Psych*. (2021) 12:625247. doi: 10.3389/fpsyt.2021.625247

127. Azizoddin DR, Jolly M, Arora S, Yelin E, Katz P. Longitudinal study of fatigue, stress, and depression: role of reduction in stress toward improvement in fatigue. *Arthritis Care Res*. (2020) 72:1440–8. doi: 10.1002/acr.24052

128. Chan KS, Friedman LA, Bienvenu OJ, Dinglas VD, Cuthbertson BH, Porter R, et al. Distribution-based estimates of minimal important difference for hospital anxiety and depression scale and impact of event scale-revised in survivors of acute respiratory failure. *Gen Hosp Psychiatry*. (2016) 42:32–5. doi: 10.1016/j.genhosppsych.2016.07.004

129. Mao F, Sun Y, Wang J, Huang Y, Lu Y, Cao F. Sensitivity to change and minimal clinically important difference of Edinburgh postnatal depression scale. *Asian J Psychiatr*. (2021) 66:102873. doi: 10.1016/j.ajp.2021.102873

130. Mauskopf JA, Simon GE, Kalsekar A, Nimsch C, Dunayevich E, Cameron A. Nonresponse, partial response, and failure to achieve remission: humanistic and cost burden in major depressive disorder. *Depress Anxiety*. (2009) 26:83–97. doi: 10.1002/da.20505

131. Strandberg RB, Graue M, Wentzel-Larsen T, Peyrot M, Rokne B. Relationships of diabetes-specific emotional distress, depression, anxiety, and overall well-being with HbA1c in adult persons with type 1 diabetes. *J Psychosom Res*. (2014) 77:174–9. doi: 10.1016/j.jpsychores.2014.06.015

132. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Cons Clin Psychol*. (1991) 59:12–9. doi: 10.1037/0022-006X.59.1.12

133. Flückiger C, Del Re A, Wampold BE, Horvath AO. Alliance in adult psychotherapy In: JC Norcross and MJ Lambert, editors. *Psychotherapy relationships that work: Evidence-based therapist contributions*. Oxford: Oxford University Press (2019)

134. Bica I, Alaa AM, Lambert C, Van Der Schaar M. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clin Pharmacol Ther*. (2021) 109:87–100. doi: 10.1002/cpt.1907

135. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*. (2019) 28:231–7. doi: 10.1136/bmjqs-2018-008370

136. Selby JV, Fireman BH. Building predictive models for clinical care—where to build and what to predict? *JAMA Netw Open*. (2021) 4:e2032539–9. doi: 10.1001/jamanetworkopen.2020.32539

137. Yao L, Wang Z, Gu H, Zhao X, Chen Y, Liu L. Prediction of Chinese clients' satisfaction with psychotherapy by machine learning. *Front Psych*. (2023) 14:947081. doi: 10.3389/fpsyt.2023.947081

138. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol Methods*. (2009) 14:323–48. doi: 10.1037/a0016973

139. Simpson HB, Marcus SM, Zuckoff A, Franklin M, Foa EB. Patient adherence to cognitive-behavioral therapy predicts long-term outcome in obsessive-compulsive disorder. *J Clin Psychiatry*. (2012) 73:1265–6. doi: 10.4088/JCP.12l07879

140. Smink WA. *What works when for whom?: A methodological reflection on therapeutic change process research*. Netherlands: University of Twente (2021).

141. Smink WA, Sools AM, Postel MG, Tjong Kim Sang E, Elfrink A, Libbertz-Mohr LB, et al. Analysis of the emails from the Dutch web-based intervention "alcohol de baas": assessment of early indications of drop-out in an online alcohol abuse intervention. *Front Psych*. (2021) 12:575931. doi: 10.3389/fpsyt.2021.575931

142. Tiemens B, Kloos M, Spijker J, Ingenhoven T, Kampman M, Hendriks G-J. Lower versus higher frequency of sessions in starting outpatient mental health care and the risk of a chronic course; a naturalistic cohort study. *BMC Psychiatry*. (2019) 19:1–12. doi: 10.1186/s12888-019-2214-4

143. De Jong K, Conijn JM, Gallagher RA, Reshetnikova AS, Heij M, Lutz MC. Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: a multilevel meta-analysis. *Clin Psychol Rev*. (2021) 85:102002. doi: 10.1016/j.cpr.2021.102002

144. Rognstad K, Wentzel-Larsen T, Neumer S-P, Kjøbli J. A systematic review and meta-analysis of measurement feedback systems in treatment for common mental health disorders. *Adm Policy Ment Health Ment Health Serv Res*. (2023) 50:269–82. doi: 10.1007/s10488-022-01236-9

145. Cohen ZD, Kim TT, Van HL, Dekker JJ, Driessen E. A demonstration of a multi-method variable selection approach for treatment selection: recommending cognitive–behavioral versus psychodynamic therapy for mild to moderate adult depression. *Psychother Res*. (2020) 30:137–50. doi: 10.1080/10503307.2018.1563312

146. Cho G, Yim J, Choi Y, Ko J, Lee S-H. Review of machine learning algorithms for diagnosing mental illness. *Psychiatry Investig*. (2019) 16:262–9. doi: 10.30773/pi.2018.12.21.2

147. Beaulieu-Jones BK, Finlayson SG, Yuan W, Altman RB, Kohane IS, Prasad V, et al. Examining the use of real-world evidence in the regulatory process. *Clin Pharmacol Ther*. (2020) 107:843–52. doi: 10.1002/cpt.1658