Received: 18 October 2022 | Accepted: 27 November 2023

DOI: 10.1111/bjc.12452

REGISTERED REPORT STAGE 2

Predicting treatment response using machine learning: A registered report

Kristin Jankowsky[1] | **Lina Krakau**[2] | **Ulrich Schroeders**[1] | **Rüdiger Zwerenz**[2] | **Manfred E. Beutel**[2]

[1]Psychological Assessment, University of Kassel, Kassel, Germany

[2]Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Mainz, Mainz, Germany

Correspondence
Kristin Jankowsky, University of Kassel, Hollaendische Strasse 36-38, Kassel 34127, Germany.
Email: jankowsky@psychologie.uni-kassel.de

Abstract

Objective: Previous research on psychotherapy treatment response has mainly focused on outpatients or clinical trial data which may have low ecological validity regarding naturalistic inpatient samples. To reduce treatment failures by proactively screening for patients at risk of low treatment response, gain more knowledge about risk factors and to evaluate treatments, accurate insights about predictors of treatment response in naturalistic inpatient samples are needed.

Methods: We compared the performance of different machine learning algorithms in predicting treatment response, operationalized as a substantial reduction in symptom severity as expressed in the Patient Health Questionnaire Anxiety and Depression Scale. To achieve this goal, we used different sets of variables—(a) demographics, (b) physical indicators, (c) psychological indicators and (d) treatment-related variables—in a naturalistic inpatient sample ($N = 723$) to specify their joint and unique contribution to treatment success.

Results: There was a strong link between symptom severity at baseline and post-treatment ($R^2 = .32$). When using all available variables, both machine learning algorithms outperformed the linear regressions and led to an increment in predictive performance of $R^2 = .12$. Treatment-related variables were the most predictive, followed psychological indicators. Physical indicators and demographics were negligible.

Conclusions: Treatment response in naturalistic inpatient settings can be predicted to a considerable degree by using baseline indicators. Regularization via machine learning algorithms leads to higher predictive performances as opposed to including nonlinear and interaction effects. Heterogenous aspects of mental health have incremental predictive value and should be considered as prognostic markers when modelling treatment processes.

KEYWORDS

inpatients, machine learning, predictive modelling, prognostic markers, treatment response

Practitioner Points

- The present study shows that patients' characteristics at the start of a psychotherapy can be used to predict treatment response.
- Machine learning algorithms can help enhance predictive accuracy, however, not due to the incorporation of nonlinear or interaction effects but rather by reducing the models' overfit via regularization, stressing the need for high-quality data and reliable indicators rather than more complex models.
- Beyond baseline symptomatic, various indicators on mental health had incremental value for the prediction of treatment response and should therefore be focused on at baseline assessments (as opposed to demographics and indicators of physical health).
- Prediction models such as the one in this study could be implemented using routine baseline assessments and provide valuable information on the risk of treatment nonresponse at a time when intervention is still possible.

BACKGROUND

Multi-modal inpatient treatment is a valid and effective treatment option for patients with severe mental disorders (Liebherz & Rabung, 2014). In Germany, psychosomatic treatment is offered both in an inpatient and day-clinic setting with psychotherapy as its main treatment modality complemented with additional somatic, psychopharmacological and specialized therapies (e.g., creative therapy). While patient samples and response rates are comparable between inpatient and day-clinic settings (Zeeck et al., 2015), not all patients respond equally well to treatment. Heterogeneous treatment responses have been well documented for outpatient treatment of depression (Kaiser et al., 2022), but several studies showed that the phenomenon also translates to other mental disorders (Altmann et al., 2020; Senger et al., 2021) and inpatient treatment settings (Hartmann et al., 2018; Zeeck et al., 2020).

To improve response rates, reduce relapse rates and avoid exposing patients to multiple treatment failures, researchers and clinicians have been interested in learning about risk factors of treatment nonresponse, adapting treatments to patient needs and understanding which treatment is best suited to an individual patient (Delgadillo, 2021; Delgadillo & Lutz, 2020; Zeeck et al., 2013). Patient, therapist and process factors have all been established to contribute to therapy outcomes (Luborsky et al., 1971; Lutz et al., 2021). Identifying reliable patient pretreatment characteristics would enable practitioners to adapt the treatment to individuals prior to starting treatment, thereby avoiding suboptimal attempts as well as saving financial, time and personnel resources. Despite considerable research effort, findings so far have been mixed with most of the prognostic markers identified making only a minor contribution in explaining treatment response (Chekroud et al., 2021). The most robust finding pertains to the impact of symptom load (severity) at baseline (Cuijpers et al., 2022), yet studies modelling course trajectories have repeatedly found patient groups with high baseline load who either responded very well or did not change reliably (Altmann et al., 2015).

Other psychological and psychiatric variables that have shown associations with treatment response (for depression) are among others chronicity, psychosocial functioning, psychological and physical comorbidity, personality, childhood adversity and recent trauma, cognitive deficits and coping resources (Kessler et al., 2017; Maj et al., 2020). For inpatient treatment specifically, comorbid (mental) disorders, personality, chronicity and patient motivation have been found to impact response (Beutel & Bleichner, 2011; Zeeck et al., 2016, 2020). However, the number of studies empirically addressing this question in inpatient settings is sparse. Rather than a single, predominant factor leading to treatment response, multiple predictors '[outweigh] and [interact] with each other in so far incomprehensible ways' (Hilbert et al., 2021, p. 53). These predictors can be assigned to different variable groups, for example: (a) sociodemographic background variables (e.g., gender, age), (b) indicators of physical health (e.g., subjective health, BMI, smoking), (c) indicators of personality and mental health (e.g., maladaptive personality traits, anxiety or depression scores) and (d) treatment variables (e.g., number of treatments within the last 12 months). These groups of predictors differ in terms of reliability, assessment method (e.g., self-report questionnaire, clinical interview) and time- and content-related proximity to the criterion (proximal vs. distal).

Research on prognostic markers has heavily relied on re-analyses of clinical trial data. These individual studies are oftentimes underpowered, limiting the identification of reliable predictors and interaction effects (Fisher et al., 2017). However, even adequately powered individual participant data meta-analyses (IPD-MA) have mostly reported symptom severity as the single best predictor of treatment response for depression (see Cuijpers et al., 2022, for an overview). One problematic aspect of randomized controlled trials (RCT) which are considered gold-standard for therapy evaluation is their limited ecological validity (Philips & Falkenström, 2021). RCTs usually have strict in- and exclusion criteria (e.g., excluding patients with multiple comorbidities, see O'Hara et al., 2017), leading to more homogenous samples compared to the population. Patients presenting, for example, acute suicidal ideation, substance abuse, or specific personality disorders are excluded from RCTs although, from a clinical perspective, these factors likely interfere with treatment compliance or complicate treatment (Krause & Behn, 2021). Moreover, these more severe clinical characteristics are key reasons why patients seek more intensive care in inpatient and day hospital wards. As this specific group of patients is being precluded from participating in RCTs and their inclusion into RCTs is often not feasible due to ethical reasons, these trials have limited capacity to inform treatment prognosis for real-world intensive care settings (Webb et al., 2020).

Recent work has pointed to the potential benefits of machine learning (ML) techniques in large-scale observational data (Aafjes-van Doorn et al., 2021; Chekroud et al., 2021; Dwyer et al., 2018). As we do not have a clear theoretical model in which ways patients' sociodemographics interact with psychological and medical variables and how these translate to treatment (non)-response, the field embraces the possibilities of ML to examine a plethora of predictors and their potentially nonlinear and higher-order interaction effects. Rather than evaluating a specific, theoretically derived moderator of treatment response in a rather simplistic understanding of dependencies (see also the concept of *Flatland Fallacy*, Jolly & Chang, 2019), the goal in ML is to use all available information to establish connections between the variables in a data-driven way and to increase predictive power (Chekroud et al., 2021; Yarkoni & Westfall, 2017).

There have been few studies to date using ML algorithms to predict treatment response in naturalistic inpatient settings including patients with diverse diagnoses. For example, Webb et al. (2020) compared 14 different ML algorithms in the prediction of post-treatment depression scores in the *Patient Health Questionnaire-9* (PHQ-9). In doing so, the authors used a range of predictor variables that were routinely assessed at admission (including demographics, clinical measures, treatment history, or physical health variables). The best-performing algorithm (elastic net regressions) explained 38% of interindividual variance in the depression scores in a holdout sample, meaning a sample that was not used during model training. Particularly important variables for the prediction of treatment response were the patients' expectations of improvement, baseline symptom severity—that is, baseline PHQ-9 values and baseline Generalized Anxiety Disorder Scale-7 values—as well as whether patients took mood stabilizers.

Pros and cons of machine learning in psychotherapy research and a call for closer methodological scrutiny

Apart from the possibility of including complex interaction effects to enhance predictive performance, ML algorithms have several in-built features that are promising when trying to tackle methodological challenges usually encountered in the prediction of treatment response. For example, it is possible to reduce model overfit by using algorithms that employ some form of regularization. Overfit can be defined as the difference in predictive performance of a model using training data versus independent, unseen testing data (Urban & Gates, 2021). Especially in scenarios with small sample sizes and a large number of predictors—which is a realistic setting in many studies on treatment outcomes using baseline indicators (Chekroud et al., 2021)—unregularized regression models tend to overfit, hampering the generalizability and the clinical usefulness of the predictive models.

However, the use of ML in clinical psychology has also been viewed critically (Wilkinson et al., 2020). Typical criticisms include a lack of assessment of ML's benefits relative to its costs (Kessler et al., 2020) as well as its worse interpretability compared to simpler models (Siddaway et al., 2020) which could lead to lower clinical utility as well as lower acceptance and implementation rates by clinicians (see Lutz et al., 2022, for an example of the influence of therapists' attitudes towards and rated usefulness of machine learning-based digital decision support and feedback system on its overall effectiveness). On a much more fundamental level, there is also increasing and strong evidence across several research fields including psychiatry that ML analyses are often flawed. For example, many ML models are evaluated incorrectly, biasing model validation in favour of more complex and flexible algorithms as those are better equipped to recognize specific data patterns as well as exploiting any spillover of information between training and testing data (Jacobucci et al., 2021; Kapoor & Narayanan, 2022).

In our reading of the psychometric literature, many reproducibility issues underlying ML studies can be described by two main factors: they are often overhyped and underchecked. We refer to the term overhyped in the sense that similar to a general publication bias, that is, the tendency to publish innovative significant findings with large effects (Ferguson & Brannick, 2012; Ferguson & Heene, 2012), novel ML models that are seemingly highly predictive are more likely to gain traction and to get published. Thus, the incentive to follow a new methodological fashion and to employ new ML models is strong, especially when the outcome to be explained is multifactorially influenced and has steered a lot of inconclusive previous research such as what works for whom in therapy. At the same time, consolidated knowledge of using ML as a statistical tool is not widespread outside computational science and statistics. In a comprehensive survey, Kapoor and Narayanan (2022) showed across a wide range of disciplines (including medicine and psychiatry) that many ML models in the literature were not validated correctly, which could lead to the dissemination of false discoveries or the development of unsubstantiated theories. Consequently, these overoptimistic or biased ML models do not live up to their expectations if correctly validated (see Jacobucci et al., 2021). Thus, we propose a more open debate and culture of mutual scrutiny (Vazire, 2020) to enhance transparency and avoid common pitfalls in ML (see also Cearns et al., 2019; Kapoor & Narayanan, 2022). One way to achieve this is by employing registered reports (Scheel et al., 2021) which make methodological feedback prior to (running the actual study or) conducting the analyses the norm.

The present study

In this study, we predict treatment response defined by the post-treatment sum score of the Patient Health Questionnaire Anxiety and Depression Scale (PHQ-ADS; Kroenke et al., 2016) in an inpatient sample. The PHQ-ADS is a composite of the 9-item Patient Health Questionnaire and the 7-item Generalized Anxiety Disorder scale (GAD-7; Gräfe et al., 2004) which has been found to be a reliable (Cronbach's alpha between .88 and .92 in three different trials; Kroenke et al., 2016), valid and (sufficiently) unidimensional indicator of depressive symptoms (depression and anxiety). We deliberately decided against grouping patients (remission vs. no remission) according to specific cut-offs to avoid information loss (and grouping those who show no change together with patients whose symptoms deteriorate). We use routine outcome monitoring data from a clinic and polyclinic in Germany, including predictor variables on demographics, personality and indicators of mental health, as well as physical health, and treatment-related variables. This study has three major goals:

First, we further examine the incremental predictive performance of different ML approaches in predicting therapy response by comparing increasingly complex ML models to linear models. Simple linear regression models using either a naïve guessing approach, the PHQ-ADS scores of the baseline assessment as the sole predictor or all information available serve as benchmark models. Thus, we aim to quantify the increment of using all baseline variables beyond naïve guessing or baseline symptom severity. The linear regression model with all available variables was then compared to (a) elastic net regressions as an example of regularized linear regressions and (b) gradient boosting machines, which allows for nonlinear and higher-order interaction effects. This comparison aimed to quantify the incremental value of ML algorithms over and above traditional methods.

Second, we establishe the unique and joint contribution of all predictor groups in the prediction of treatment success by systematically rerunning the best-performing algorithm with all possible combinations of groups. The predictors are grouped as follows: (a) sociodemographic variables, (b) indicators of physical health, (c) indicators of mental health and (d) treatment variables. We examine constructs that often have been missing or range-restricted in previous research (e.g., The Personality Inventory for DSM-5) due to homogeneous person sampling in RCTs. Thus, we aim to further knowledge on predictors and moderators of treatment response, potentially screening for participants at risk of treatment nonresponse. To render our prediction models more interpretative, we provide importance measures for all variables of all models.

Third, we counter the objection that machine learning research is inevitably accompanied by increased researcher's degrees of freedom, forming the basis for another reproducibility and replication crisis (Hullman et al., 2022) by registering all analytical decisions beforehand. At first glance, this approach seems to counteract the empirically driven and flexible nature of ML algorithms. However, many aspects concerning data cleaning, variable transformation, handling of missing data, etc. can be registered in ML studies the same way as in every other study. Also, the settings for data-driven hyperparameter tuning can also be defined in advance. Surprisingly, we were not able to find any previous studies on the prediction of psychotherapy outcomes using machine learning that capitalize on the benefits of a registered report. Machine learning modelling and registered reports rarely have been combined in psychological research so far (for one of the few exceptions, see Costello et al., 2021). However, a consistent conclusion of several reviews and meta-analyses on machine learning models in clinical research is that the stark differences in implementation and the often non-transparent model evaluations hinder a useful aggregation of findings (Christodoulou et al., 2019; Lee et al., 2018). We strive to provide an example of a thorough registration of the proposed analysis pipeline that still allows for the analytical flexibility of the ML algorithms (e.g., through hyperparameter tuning).

METHOD

Sample

We used routine outcome monitoring data from 723 patients of a clinic and polyclinic in Rhineland-Palatinate collected between 2018 and 2021. The clinic comprises three inpatient and day hospital units offering multi-modal treatment consisting of two to three individual therapy sessions per week, two weekly sessions of art therapy, up to two sessions of body-oriented therapy and up to three sessions of group therapy. The duration of treatment is typically 4–12 weeks. Averaged treatment length in our sample was at 6 weeks. While the focus of the clinic is psychodynamic, treatment integrates different schools and modalities, including educational elements regarding the pathogenesis and maintenance of the disorder, and specific modules (e.g., relaxation training or physiotherapy) tied to the individual needs of the patients. In the group settings, a new member is admitted when a patient is discharged from the hospital. Hence, the groups comprise patients at different treatment stages. This 'slow-open' principle offers the possibility for peer learning, where new members can benefit from the perspectives of patients who have already undergone parts of their treatment and more experienced patients can become more aware of their change processes when confronted with attitudes and scepticism of the novices. The multi-professional team consists of psychosomatic medical specialists and residents, psychologists, creative therapists, specialized nurses and social workers. The nursing staff is constantly present, aiming at ensuring stability, holding and reassurance (Beutel et al., 2008).

Using patient data for research is regulated by the German State Hospital Act and was approved by the Rhineland-Palatinate Chamber of Physicians (nr. 837.191.16 (10510)). We provide a descriptive overview of patients' characteristics on all available measures of this study in Table S1 at https://osf.io/86zng. The patient data are not publicly available due to privacy restrictions, but we provide a correlation matrix for all variables and a synthetic version of the data in the supplemental materials at https://osf.io/jxst4/ to render our results as transparent and reproducible as possible. For data access upon request, please contact the second author.

Measures

In Table S1, we present an overview of all available measures that are included in the prediction models. We predicted the patients' treatment response operationalized as post-treatment PHQ-ADS scale sum scores (controlled for pre-treatment PHQ-ADS sum scores). All categorial variables were dummy-coded prior to the analysis with the first category as a reference. We excluded eight patients with more than 30% missingness on all variables. We also excluded categorical predictor variables with fewer than 10 events to avoid computational problems due to low variances (i.e., two response options regarding pensions due to reduced earning capacity). For our analysis, we used the standardized individual item scores to fully capture all potential effects since it has repeatedly been shown that individual items outperform scale scores in prediction tasks (McClure et al., 2021; Seeboth & Mõttus, 2018).

Statistical analyses

All our analyses were conducted using the R package *caret* (Kuhn, 2008) as an interface for modelling, prediction and evaluation. Irrespective of the modelling algorithm, we employed a nested cross-validation approach (Bischl et al., 2012; Pargent et al., 2023) which is often recommended to strictly separate any data pre-processing and hyperparameter tuning from the final model validation. Thus, nested cross-validation avoids information leakage between the training and the testing sample that is used for model evaluation. Nested cross-validation combines an outer and inner validation loop: First, in every iteration of the *outer validation loop*, the full data are split into training data (for our study, 80%

TABLE 1 Proposed models for the prediction of treatment response.

No.	Model/algorithm	Predictor variables	Tuning parameters
0	Linear Regression	Naïve guessing model	
1	Linear Regression	PHQ-ADS at baseline	
2	Linear Regression	All available variables	
2	Linear Regression	All available variables	
3	Elastic Net Regression	All available variables	α: 40 evenly distributed values between 0 and 1 λ: data-driven; sequence between minimum and maximum λ generated from a glmnet model, assuming that a sensible range of λ is provided using an α value of .50
4	Gradient Boosting Machines	All available variables	Interaction depth of 1, 2, 3, 4, 5 Minimum leaf size of 5, 10, 20, 50 Shrinkage as a sequence between .001 and .201 using steps of .02 Number of trees 50, 100, 150, 300, 500, 1000
4	Best-performing algorithm of Model 2–4 [= BPA]	All four predictor groups[a] (= All available variables)	Parameter set as in best-performing algorithm of Model 2–4
5–8	BPA	Only one predictor group	Parameters of BPA
9–14	BPA	Any pair of two predictor groups	Parameters of BPA
15–18	BPA	Any triple of three predictor groups	Parameters of BPA

Abbreviation: BPA, Best-performing algorithm.

[a] Demographics, physical health, mental health, treatment variables. For more information on the implementation of the glmnet tuning grid, please see also https://github.com/topepo/caret/blob/master/models/files/glmnet.R. The repeated model numbers (2 and 4) do not indicate that these models will be estimated multiple times but divide the table into different blocks/sequences of comparison.

of the full data set) and a holdout sample (the remaining 20%) as testing data. Any missing values will be imputed separately for the training and testing datasets (i.e., after the 80/20 split) using multiple imputations (k = per cent of missingness averaged across all predictor variables, but a minimum of 10) via the random forest algorithm implemented in the R package *mice* (van Buuren & Groothuis-Oudshoorn, 2011). Within the *inner validation loop*, we trained the respective models for each of the imputed training datasets using 10-fold cross-validation. Predictive performance of these models was then calculated as the average performance across the k-test datasets. Further, we averaged these results across 100 iterations of the *outer validation loop* to provide an accurate estimate of the expected prediction performance using unseen testing data.

Table 1 arranges the different models we compared in this study into three major blocks. As a first step, we quantified the incremental predictive validity of using all variables in comparison to more simple benchmark models. In more detail, the first comparison consists of three linear regression models, (a) a naïve guessing model (or null model, Model 0), (b) a model solely using the PHQ-ADS score at baseline to predict the post-treatment PHQ-ADS scores, because initial symptom severity has been demonstrated a strong predictor of treatment response (Model 1) and (c) a linear regression model with all available predictor variables (Model 2). As a second step, we completed this regression model using all variables (Model 2) to two ML algorithms—elastic net regressions and gradient boosting machines (Model 3 and 4). Finally, in a third step, we examined the unique and combined contribution of different variable groups in the prediction of treatment response (Model 5–18). To this end, we used the algorithm and tuning parameters that showed the highest predictive performance in the aforementioned methodological comparison.

In the following, we briefly describe the key characteristics of the two ML modelling approaches. Elastic net regressions are regularized regressions that lead to parsimonious models by penalizing the regression weights of certain predictors. They compromise between ridge regressions and least absolute shrinkage and selection operator (LASSO) regressions. By using the tunable shrinkage parameter λ and penalty parameter a (Zou & Hastie, 2005), they strike for an optimal balance between minimizing the sum of squared weights (assigning variables small, but non-zero weights) and the sum of absolute weights (leading to models with many variables given weights of zero), thereby aiming to maximize predictive performance.

Gradient boosting machine algorithms are tree-based algorithms that allow for the integration of nonlinear and higher-order interaction effects into the modelling without the need for specific assumptions on functions between predictor variables and the respective outcome (James et al., 2017). They sequentially combine multiple decision trees into an ensemble. Every new tree aims at fitting the residual error of the previous one, leading to potentially better predictive performance. Their complexity depends on hyperparameter settings (e.g., number of trees, minimum leaf size) which should be sensibly tuned to avoid overfitting due to overly complex models (McNamara et al., 2022).

Model evaluation

We use the following indices to evaluate predictive models: explained variance (R^2), the root mean squared error (RMSE) and the Mean Absolute Error (MAE). All indices are calculated for the training sample and also for the holdout sample across the 100 iterations of splitting the data into training and testing data. We present the results for all indices using box and jitter plots to illustrate (a) the overall predictive performance and (b) the amount of overfit for all modelling approaches. For all models, variable importances are presented using the varImp function in caret. However, we focus the discussion of important predictor variables on the model with the highest prediction performance.

To further the comprehensibility and accountability of the described analytical approach, we provide annotated R syntax of our analyses. These materials can be found at https://osf.io/jxst4/. The time-stamped Stage 1 of this registered report can be found at https://osf.io/tkm2h.

RESULTS

A unidimensional factor explained 42% of the variance of the PHQ-ADS items at baseline. Reliability was high ($\alpha = .92$). The averaged raw difference between baseline and outcome sum scores was −9.0 (empirical range: −37; 19). For 13% of the full sample ($N = 94$ patients, using multiple imputed data), the PHQ-ADS values, that is, symptom load did not change or even increased. This statement is not about statistical or clinical significance, but to describe the wide range of individuals' treatment successes.

Table 2 provides an overview of the averaged predictive performances of all 18 models for the 100 training and testing datasets, respectively. Overall, there was a strong link between the PHQ-ADS at baseline and the post-treatment PHQ-ADS ($R^2 = .319$; Model 1), which is used as the point of reference when making statements about the incremental predictive validity of predictor variable groups.

Figure 1 shows the distribution of the 100 different R^2 values for the three different model families using all available variables (Model 2–4), indicating a large variance across different data splits and thus, underlining the need to use outer cross-validation to obtain reliable estimates for model performance using unseen testing data. Whereas the linear regression models on average explained an increment of 4.8% of the outcome's variance over the baseline model, both machine learning algorithms outperformed the linear regressions with an increment of 12.0% (elastic net regression)

TABLE 2 Predictive performances in training and test data for the prediction of treatment response.

			Train			Test		
No	Algorithm	Predictors	R^2	RMSE	MAE	R^2	RMSE	MAE
0	LinReg	Naïve guessing model		.464	.377		.455	.365
1	LinReg	PHQ-ADS at baseline	.329	.380	.302	.319	.376	.298
2	LinReg	All available variables	.598	.294	.233	.367	.374	.296
3	Enet	All available variables	.535	.320	.252	.439	.342	.268
4	GBM	All available variables	.700	.259	.205	.441	.342	.269
5	Enet	Demographics	.351	.375	.299	.330	.373	.296
6	Enet	Physical	.333	.340	.303	.315	.377	.299
7	Enet	Mental	.449	.347	.275	.370	.362	.286
8	Enet	Treatment	.433	.350	.275	.401	.354	.277
9	Enet	Demographics, Physical	.356	.373	.298	.328	.374	.296
10	Enet	Demographics, Mental	.452	.346	.275	.371	.362	.286
11	Enet	Demographics, Treatment	.447	.346	.273	.403	.352	.277
12	Enet	Physical, Mental	.448	.347	.276	.368	.363	.286
13	Enet	Physical, Treatment	.438	.349	.274	.396	.355	.278
14	Enet	Mental, Treatment	.532	.320	.252	.441	.341	.268
15	Enet	Demographics, Physical, Mental	.451	.347	.275	.370	.362	.286
16	Enet	Demographics, Physical, Treatment	.452	.345	.272	.402	.353	.277
17	Enet	Demographics, Mental, Treatment	.533	.320	.252	.440	.342	.268
18	Enet	Physical, Mental, Treatment	.534	.319	.252	.440	.342	.268

Abbreviations: Enet, elastic net regression; GBM, gradient boosting machines; Linreg, linear regression; MAE, Mean Absolute Error; RMSE, Root Mean Square Error.

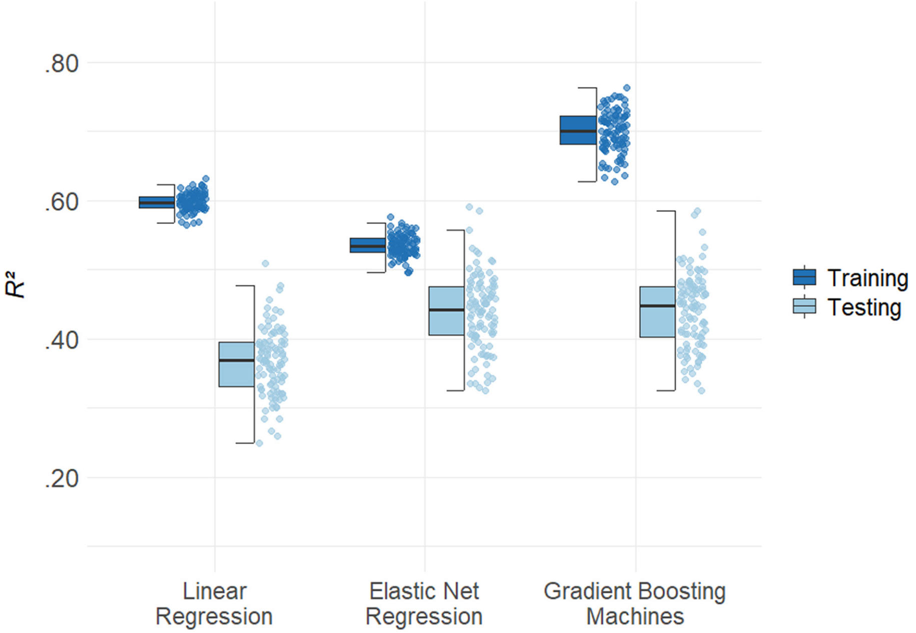

FIGURE 1 Explained variance for all 100 iterations of the outer validation loop for the three models using all available predictor variables. *Note*: The box reflects the interquartile range (IQR), the solid line the median and the whiskers 1.5 times the IQR across 100 iterations (i.e., outer loop data splits). R^2 values of the 100 models are displayed as a jittered distribution on the right.

and 12.2% (GBM) over the baseline model. Since the two machine learning models using all available predictor variables were nearly identical in predictive performances within the test data, we conducted all following variable group comparisons with both algorithms to provide interested readers with a full result set.

In Table 2, we present the results using elastic net regressions since those show a considerably less amount of overfit (i.e., difference between training and testing performance) across the different models (for the results of Model 5–18 using GBM, see Table S2). The rank order of the performances of the different models with respect to variables included in the models was similar across the two algorithms so that conclusions about the importance of variable groupings were identical. The main findings were as follows: First, demographics and indicators of physical health seem negligible (see Table S2). Second, out of the models using a single predictor set (Model 5–8), the model with the treatment variables performed best, followed by indicators of mental health. Third, out of all possible combinations, the model using treatment and mental health variables (Model 14) had the highest predictive performance in the test data. Even the models with additional variables (Models 15–18 and Model 3) showed no further improvement, which can be attributed to less overfit of the more parsimonious model.

As preregistered, we provide a detailed overview of variable importances for all models in the online supplement (see https://osf.io/gnzqs), but present and discuss only the most predictive model within the manuscript. Since variable importance measures are based on the training data and the GBM models overfitted significantly more than the elastic net regressions (at similar performance in the testing data), we will focus on the latter for the most accurate estimations. Figure 2 shows the 20 most important variables for Model 3 which overall confirm the results of the previous model comparisons: The baseline PHQ-ADS value was by far the most important variable, followed by whether the patients found the treatment helpful, treatment length and satisfaction with treatment. There was only one demographic variable (patients' age) and one indicator of physical health (healthy nutrition) among the 20 most important predictors. Importantly, the latter variable represents a self-reported indicator of health behaviour rather than an objective measure of physical health. Relevant indicators of mental health

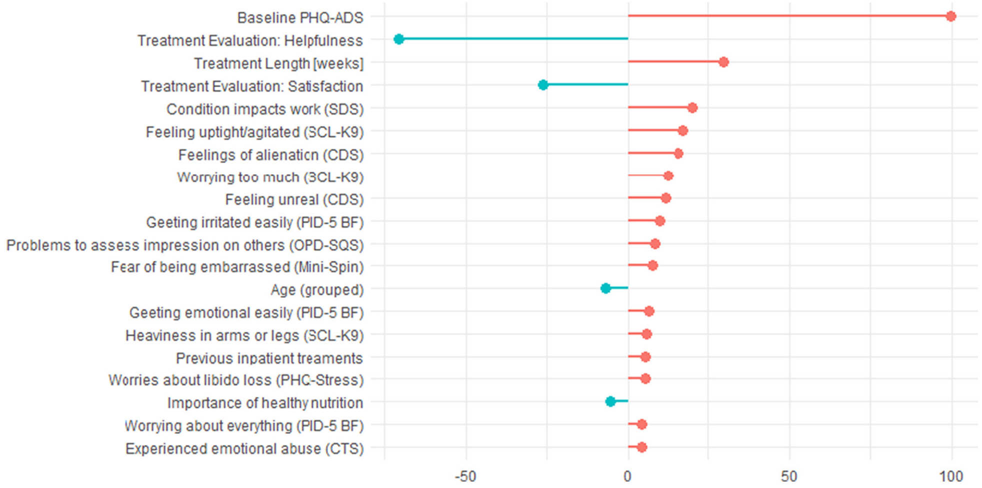

FIGURE 2 Twenty variables with the highest averaged variable importances for Model 3. *Note*: PHQ-ADS, Patient Health Questionnaire Anxiety and Depression Scale (Kroenke et al., 2016); SDS, Sheehan Disability Scale (Sheehan et al., 1983); SCL-K9, Symptom Checklist (Petrowski et al., 2019); CDS-2, Cambridge Depersonalization Scale 2 (Michal et al., 2010); PID-5, The Personality Inventory for DSM-5—Brief Form (Krueger et al., 2013); OPD-SQS, OPD-Structure Questionnaire Short (Ehrenthal et al., 2015); Mini-Spin, Mini-Social-Phobia-Inventory (Wiltink et al., 2017); PHQ-Stress, Patient Health Questionnaire Stress (Gräfe et al., 2004); CTS, Childhood Trauma Screener (Grabe et al., 2012).

included the self-rated level of functioning at work or school (measured by the Sheehan Disability Scale), depersonalization-derealization, items of the SCL9 (which is to be expected since they can be seen as an alternative measure of symptom load), a PID-5 item measuring whether patients are quickly annoyed by all sorts of things, aspects of social phobia (whether patients were afraid to be ashamed or feel dumb), level of structural integration (as indicated by an item of the OPD-SQS), remembered childhood emotional abuse and being worried about libido loss.

DISCUSSION

With this registered report, we aimed to further the knowledge on predictors of treatment response in ecologically valid naturalistic inpatient samples, ultimately working towards the admittedly ambitious goal of reducing treatment failures and relapse rates by proactively screening for patients at risk. Our methodological comparisons showed that the prediction of treatment response can be enhanced by using machine learning algorithms, however, not due to the incorporation of nonlinear or interaction effects but rather by reducing the models' overfit via regularization as indicated by the equally good performance of the elastic net regressions and the GBM models using all available predictor variables. The regularized elastic net regressions had a higher predictive performance in independent testing samples than the non-regularized version. Hence, the results provide yet another argument for focusing on collecting high-quality data in large samples with reliable indicators for clinical prediction models instead of on more and more complex modelling approaches when aiming for generalizability and in turn, clinical usability.

Important variables for the prediction of treatment response

With an R^2 of .44, the overall predictive performance of the best-performing model in this study is comparable to or even slightly higher than the results of similar previous research using elastic net regressions for the prediction of treatment response. For example, Webb et al. (2020) were able to explain 38% of variance in post-treatment depression scores. As it has been often found (Cuijpers et al., 2022), anxiety and depression symptoms at discharge (i.e., post-treatment PHQ-ADS values) were best explained by their baseline levels ($R^2 = .32$) with an increment of $\Delta R^2 = .12$ in predictive performance adding all other available variables. This increment has to be seen in light of the two-fold nature of symptom severity as a predictor of poorer prognosis and as a predictor of successful/positive change (Constantino et al., 2021). If modelled explicitly, the literature points to distinct response patterns associated with high baseline severity: Patients with high baseline severity who improve strongly are often additionally characterized by lower impairment in other domains or low risk-related behaviour (e.g., self-harm or externalizing symptoms; Uckelstam et al., 2019), underscoring the need of a multidimensional assessment of functioning. Several authors have called for assessing the complexity of mental disorders across interlinked domains of functioning to better explain the variability in their phenotypes and responses to treatment (Barton et al., 2017).

For example, patients' ratings of their condition impacting their workability emerged as an important predictor of treatment response. Occupational functioning is a relevant criterion typically rated alongside psychological and social functioning in the psychiatric global assessment of function (GAF; Aas, 2011). Interestingly, self-rated workability had a more pronounced impact compared to self-rated social functioning, and thus, might serve as a more important indicator of clinical severity. Patients reporting a higher number of previous inpatient treatments were also more likely to report higher symptom severity at discharge, with 'unsuccessful' treatments being an indicator of disorder chronicity (Fava et al., 1996; Taylor et al., 2012). High levels of depersonalization and derealization (DP/DR), which describe the phenomenon of feeling detached and alienated from the self and the environment, were also among the most predictive variables. DP/DR can be classified as a disorder but is also coded as a symptom of the dissociative subtype of posttraumatic stress disorder, the dissociative features of dissociative identity disorder, cannabis intoxication, borderline personality disorder and anxiety disorders according to DSM-5. DP/DR often takes a chronic course (Baker et al., 2003) and individuals with co-occurring DP/DR are also at higher risk for chronic courses of comorbid mental health disorders (Mula et al., 2007; Schlax et al., 2020). DP/DR is therefore understood as an indicator of disorder severity (Baker et al., 2003; Michal et al., 2011; Simeon et al., 2003) and has previously been associated with less favourable therapy courses across different mental health disorders (Bae et al., 2016; Kleindienst et al., 2016; Lyssenko et al., 2018). Our study shows that DP/DR is predictive of anxiety and depression severity at discharge in a sample of mixed psychosomatic inpatients. This is an important finding as DP/DR is often underdiagnosed and not likely to be part of the assessment in routine outcome monitoring (Michal & Beutel, 2009). Unfortunately, the literature on evidence-based treatment of DP/DR is still scarce (Wang et al., 2023).

The SCL-K9 is an alternative measure of symptom severity. In previous research, the general severity index (GSI) of the SCL-K-9 correlated highly with measures of anxiety and depression (Petrowski et al., 2019; Prinz et al., 2008). Within the context of the present investigation, items tapping into depression (worry), anxiety (tightness) and somatic symptoms (heaviness) were important predictor variables, pointing to the interrelatedness of somatic, anxious and depressive symptom experiences—also called the SAD triad (Löwe et al., 2008). Though typically captured as a symptom, worry is also understood as a trait component associated with proneness to experience negative emotions (Weiss & Deary, 2020). A tendency to worry 'about everything', was also among the predictive PID-5 items. Similar symptoms related to domains of cognitive-affective schemes and therefore overlapping with personality (e.g., shame, worry, irritability) represented another group of important predictor variables. Excessive worrying is a form of repetitive negative thinking (RNT). It is typically defined as uncontrollable negative thoughts regarding close or distant future events, while worries focused on the past are labelled rumination (Nolen-Hoeksema et al., 2008). RNT is thought to be implicated in the development and maintenance of anxiety and depressive disorders and pretreatment levels of RNT have previously been associated with worse treatment outcomes across different mental disorders (Bredemeier et al., 2020; Kertz et al., 2015; Sarter et al., 2021). Beyond worry, other relevant PID-5 items were high irritability and emotional instability. The three most relevant items from the PID-5 all stem from the negative affectivity domain which is associated with depressivity (Gonçalves et al., 2022) but also closely overlaps with the construct of borderline personality disorder in empirical research (Gutiérrez et al., 2023). Relatedly, one item assessing personality functioning, namely the capacity to have a correct impression on how others might perceive oneself emerged as predictive. Adequate self-other functioning is the hallmark of personality disorders closely relating to mentalizing capacities (Wendt et al., 2023).

Only one demographic variable was important: Higher age was associated with better treatment response. We are unaware of inpatient studies with a similar finding. Some studies have pointed to comparable benefits across the age range (Cuijpers et al., 2018; Haigh et al., 2018). In our sample there was a preponderance of young patients. Population-based studies (Beutel et al., 2016) have shown that procrastination is particularly frequent in young people. This characteristic is likely to counteract symptom improvement but has so far been neglected in outcome studies of inpatient psychotherapy.

The group of variables that were most strongly associated with post-treatment PHQ-ADS values dealt with the treatment itself. This finding might seem rather straightforward because the self-report variables about the treatment are a direct, albeit subjective patients' evaluation of the overall treatment process and also the most temporal proximal predictors to the post-treatment outcome measure. Interestingly, whether patients found the treatment helpful was more predictive than mere satisfaction with the treatment, possibly underlining that wording matters in assessments of such subjective patients' evaluations (see also Ammerman et al., 2021 for wording effects in the context of self-harm). Also, this might hint to the fact that even momentarily dissatisfying (e.g., physically or mentally demanding) treatments do not necessarily lead to an impression of unhelpfulness. Treatment length and the number of previously undergone treatments were predictive of outcome in the sense that longer treatment and a higher number of treatment attempts were associated with more severe symptoms at discharge. Previous research indicates that the relationship between treatment duration and outcome is complex. On average, improvements of relationship patterns and personality functioning need more time than symptom improvement (Haase et al., 2008). On the contrary, the rate and magnitude of further change decline with an increasing duration of inpatient therapy (Liebherz & Rabung, 2014). Particularly complex cases unresponsive to previous outpatient or shorter inpatient treatments may require lengthy inpatient treatments followed by day hospital treatments which achieve overall comparably small benefits. In the case of the present study, it might be that highly complex cases received longer stays but were less likely to show vast change rates within their stay.

Strengths and limitations

This study represents one of the first registered reports in the field of psychotherapy research using machine learning algorithms for predictive modelling. Open science practices are by no means a sign of high-quality research by themselves, but rather a prerequisite (Bakker et al., 2020). By using two-stage approaches of publication, the risk for a posteriori modification of the research rationale or the analyses based on the study results can be reduced or at least be made transparent. Thinking of a larger context, this could also be highly relevant to address any presumptions about or effects of researcher allegiance biases in psychotherapy research. Also, transparent methods and openly available data are necessary to enable meaningful aggregations of research findings across studies in the form of reviews or meta-analyses. We acknowledge that sharing raw data is not always possible in clinical science due to privacy concerns (as it was the case in the current study). However, the provision of a synthetic version of the data seems a useful compromise between reproducibility of the analyses as well as reuse of data on the one hand, and data protection and privacy concerns on the other hand. To provide a good practice example, we also created a synthetic dataset using the convenient R package *synthpop* (Nowok et al., 2016).

In this study, we used a large naturalistic inpatient sample with mixed diagnoses and modelled treatment response using easily available baseline indicators which can be equally seen as a strength and limitation. On the one hand, we aim for validity for inpatient treatments without exclusion of patients, for example, those with recent suicide attempts or with multiple comorbid disorders, on the other hand, our results may not be immediately applied to the treatment of certain specific disorders. Additionally, using only baseline indicators can be seen at odds with findings on treatment response being affected by multiple, time-variant factors underlining the need for multi-modal and intensive longitudinal data (Chekroud et al., 2021). Nonetheless, prediction models such as the one in this study could be more easily implemented than models requiring further data acquisition since baseline assessments are carried out in German psychotherapy clinics on a routine basis and they provide valuable information on the risk of treatment nonresponse at a time when intervention is still possible.

One limiting factor that our study shares with similar research is a specific, fixed set of available predictor variables. In any case, there are further variables (e.g., patients' motivation to participate in the therapy at baseline; Jankowsky et al., 2023) that might incrementally explain treatment response. The present results have to be seen against this background: For example, we found that demographics had a negligible role in our models. However, there have been studies providing the first evidence for tailored treatments for minority patients, for example, queer patients (Bochicchio et al., 2022). Related information was not systematically assessed in our sample. Generally, the awareness of the topics of diversity and inclusion has increased in Germany only in recent years (Kluge et al., 2020) and studies investigating the experiences and needs of minority groups in the mental health care system are needed.

Due to our study design, we cannot make statements about patients' long-term treatment responses since we predicted post-treatment scores that were assessed directly at discharge. Previous research has shown that symptom severity at follow-ups can differ strongly from these assessments (Steinert et al., 2014). Thus, a worthwhile endeavour for further research would be to examine to what extent predictive models using outcomes at discharge still hold when tested at a later point in time. If this were not the case, one could argue that long-term response is the clinically more relevant outcome and should be used to train prediction models, thereby additionally providing more information on who relapses and why, which could then inform clinical decisions about relapse prevention.

CONCLUSION

In this registered report, we demonstrated that it is possible and worthwhile to combine rigorous open science practices with the analytical flexibility of complex machine learning algorithms for the prediction of treatment response. It was possible to predict treatment response to a considerable degree, taking advantage of regularization approaches inherent to the algorithms that were used. Our results again underline the large association between baseline and post-treatment symptoms; however, they also show the importance of a multidimensional assessment of functioning and identify possible prognostic markers. Our results highlight the importance of negative affectivity and self-other regulatory capacities related to depression and anxiety symptoms but also of symptoms such as depersonalization and derealization that have not been focused on in previous research.

AUTHOR CONTRIBUTIONS

Kristin Jankowsky: Conceptualization; formal analysis; methodology; visualization; writing – original draft; writing – review and editing. **Lina Krakau:** Conceptualization; data curation; writing – original draft; writing – review and editing. **Ulrich Schroeders:** Methodology; supervision; validation; writing – review and editing. **Rüdiger Zwerenz:** Data curation; project administration; resources; writing – review and editing. **Manfred E. Beutel:** Data curation; project administration; resources; writing – review and editing.

ACKNOWLEDGEMENTS

Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

The patient data are not publicly available due to privacy restrictions. For data access upon request, please contact Lina Krakau at lina.krakau@unimedizin-mainz.de. We present all predictor variables, sample descriptives as well as a correlation matrix and a synthetic dataset at https://osf.io/jxst4/.

ORCID

Kristin Jankowsky https://orcid.org/0000-0002-4847-0760

REFERENCES

Aafjes-van Doorn, K., Kamsteeg, C., Bate, J., & Aafjes, M. (2021). A scoping review of machine learning in psychotherapy research. *Psychotherapy Research*, *31*(1), 92–116. https://doi.org/10.1080/10503307.2020.1808729

Aas, I. M. (2011). Guidelines for rating global assessment of functioning (GAF). *Annals of General Psychiatry*, *10*(1), 2. https://doi.org/10.1186/1744-859X-10-2

Altmann, U., Gawlytta, R., Hoyer, J., Leichsenring, F., Leibing, E., Beutel, M., Willutzki, U., Herpertz, S., & Strauss, B. (2020). Typical symptom change patterns and their predictors in patients with social anxiety disorder: A latent class analysis. *Journal of Anxiety Disorders*, *71*, 102200. https://doi.org/10.1016/j.janxdis.2020.102200

Altmann, U., Steyer, R., Kramer, D., Steffanowski, A., Wittmann, W. W., von Heymann, F., Auch-Dorsch, E., Bruckmayer, E., Pfaffinger, I., Fembacher, A., & Strauß, B. (2015). Verlaufsmuster depressiver Störungen bei ambulanten psychotherapeutischen Behandlungen und deren Vorhersage [typical patterns of depressive disorders during outpatient psychotherapy and their prediction]. *Zeitschrift für Psychosomatische Medizin und Psychotherapie*, *61*(2), 156–172. https://doi.org/10.13109/zptm.2015.61.2.156

Ammerman, B. A., Burke, T. A., Jacobucci, R., & McClure, K. (2021). How we ask matters: The impact of question wording in single-item measurement of suicidal thoughts and behaviors. *Preventive Medicine*, *152*, Article 106472. https://doi.org/10.1016/j.ypmed.2021.106472

Bae, H., Kim, D., & Park, Y. C. (2016). Dissociation predicts treatment response in eye-movement desensitization and reprocessing for posttraumatic stress disorder. *Journal of Trauma & Dissociation*, *17*(1), 112–130. https://doi.org/10.1080/15299732.2015.1037039

Baker, D., Hunter, E., Lawrence, E., Medford, N., Patel, M., Senior, C., Sierra, M., Lambert, M. V., Phillips, M. L., & David, A. S. (2003). Depersonalisation disorder: Clinical features of 204 cases. *The British Journal of Psychiatry: the Journal of Mental Science*, *182*, 428–433.

Bakker, M., Veldkamp, C. L. S., van Assen, M. A. L. M., Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D., & Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLoS Biology*, *18*(12), e3000937. https://doi.org/10.1371/journal.pbio.3000937

Barton, S., Armstrong, P., Wicks, L., Freeman, E., & Meyer, T. D. (2017). Treating complex depression with cognitive behavioural therapy. *The Cognitive Behaviour Therapist*, *10*, e17. https://doi.org/10.1017/S1754470X17000149

Beutel, M. E., & Bleichner, F. (2011). Inpatient psychosomatic treatment of anxiety disorders: Comorbidities, predictors, and outcomes. *International Journal of Clinical and Health Psychology*, *11*(3), 443–457.

Beutel, M. E., Klein, E. M., Aufenanger, S., Brähler, E., Dreier, M., Müller, K. W., Quiring, O., Reinecke, L., Schmutzer, G., Stark, B., & Wölfling, K. (2016). Procrastination, distress and life satisfaction across the age range – A German representative community study. *PLoS One*, *11*(2), e0148054. https://doi.org/10.1371/journal.pone.0148054

Beutel, M. E., Michal, M., & Subic-Wrana, C. (2008). Psychoanalytically-oriented inpatient psychotherapy of somatoform disorders. *The Journal of the American Academy of Psychoanalysis and Dynamic Psychiatry*, *36*(1), 125–142. https://doi.org/10.1521/jaap.2008.36.1.125

Bischl, B., Mersmann, O., Trautmann, H., & Weihs, C. (2012). Resampling methods for meta-model validation with recommendations for evolutionary computation. *Evolutionary Computation*, *20*(2), 249–275. https://doi.org/10.1162/EVCO_a_00069

Bochicchio, L., Reeder, K., Ivanoff, A., Pope, H., & Stefancic, A. (2022). Psychotherapeutic interventions for LGBTQ + youth: A systematic review. *Journal of LGBT Youth*, *19*(2), 152–179. https://doi.org/10.1080/19361653.2020.1766393

Bredemeier, K., Lieblich, S., & Foa, E. B. (2020). Pretreatment levels of rumination predict cognitive-behavioral therapy outcomes in a transdiagnostic sample of adults with anxiety-related disorders. *Journal of Anxiety Disorders*, *75*, 102277. https://doi.org/10.1016/j.janxdis.2020.102277

Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry*, *9*(1), 271. https://doi.org/10.1038/s41398-019-0607-2

Chekroud, A. M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z., Belgrave, D., DeRubeis, R., Iniesta, R., Dwyer, D., & Choi, K. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*, *20*(2), 154–170. https://doi.org/10.1002/wps.20882

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Constantino, M. J., Boswell, J. F., & Coyne, A. E. (2021). Patient, therapist, and relational factors. In M. Barkham, W. Lutz, & L. G. Castonguay (Eds.), *Handbook of psychotherapy and behavior change* (7th ed., pp. 225–262). Wiley.

Costello, C., Srivastava, S., Rejaie, R., & Zalewski, M. (2021). Predicting mental health from followed accounts on twitter. *Collabra: Psychology*, *7*(1), Article 18731. https://doi.org/10.1525/collabra.18731

Cuijpers, P., Ciharova, M., Quero, S., Miguel, C., Driessen, E., Harrer, M., Purgato, M., Ebert, D., & Karyotaki, E. (2022). The contribution of "individual participant data" meta-analyses of psychotherapies for depression to the development of personalized treatments: A systematic review. *Journal of Personalized Medicine*, *12*(1), 93. https://doi.org/10.3390/jpm12010093

Cuijpers, P., Karyotaki, E., Reijnders, M., & Huibers, M. J. H. (2018). Who benefits from psychotherapies for adult depression? A meta-analytic update of the evidence. *Cognitive Behaviour Therapy*, *47*(2), 91–106. https://doi.org/10.1080/16506073.2017.1420098

Delgadillo, J. (2021). Machine learning: A primer for psychotherapy researchers. *Psychotherapy Research*, *31*(1), 1–4. https://doi.org/10.1080/10503307.2020.1859638

Delgadillo, J., & Lutz, W. (2020). A development pathway towards precision mental health care. *JAMA Psychiatry*, *77*(9), 889–890. https://doi.org/10.1001/jamapsychiatry.2020.1048

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*(1), 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Ehrenthal, J. C., Dinger, U., Schauenburg, H., Horsch, L., Dahlbender, R. W., & Gierk, B. (2015). Entwicklung einer Zwölf-item-version des OPD-Strukturfragebogens (OPD-SFK) [development of a 12-item version of the OPD-structure questionnaire (OPD-SQS)]. *Zeitschrift für Psychosomatische Medizin und Psychotherapie*, *61*(3), 262–274. https://doi.org/10.13109/zptm.2015.61.3.262

Fava, M., Alpert, J. E., Borus, J. F., Nierenberg, A. A., Pava, J. A., & Rosenbaum, J. F. (1996). Patterns of personality disorder comorbidity in early-onset versus late- onset major depression. *American Journal of Psychiatry*, *153*(10), 1308–1312. https://doi.org/10.1176/ajp.153.10.1308

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*(1), 120–128. https://doi.org/10.1037/a0024445

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561. https://doi.org/10.1177/1745691612459059

Fisher, D. J., Carpenter, J. R., Morris, T. P., Freeman, S. C., & Tierney, J. F. (2017). Meta-analytical methods to identify who benefits most from treatments: Daft, deluded, or deft approach? *BMJ*, *356*, Article j573. https://doi.org/10.1136/bmj.j573

Gonçalves, B., Pires, R., Henriques-Calado, J., & Sousa Ferreira, A. (2022). Evaluation of the PID-5 depressivity personality dimensions and depressive symptomatology in a community sample. *Anales de Psicología*, *38*(3), 409–418. https://doi.org/10.6018/analesps.486921

Grabe, H., Schulz, A., Schmidt, C., Appel, K., Driessen, M., Wingenfeld, K., Barnow, S., Spitzer, C., John, U., Berger, K., Wersching, H., & Freyberger, H. (2012). Ein Screeninginstrument für Missbrauch und Vernachlässigung in der Kindheit: Der Childhood Trauma Screener (CTS). *Psychiatrische Praxis*, *39*(3), 109–115. https://doi.org/10.1055/s-0031-1298984

Gräfe, K., Zipfel, S., Herzog, W., & Löwe, B. (2004). Screening psychischer Störungen mit dem "Gesundheitsfragebogen für Patienten (PHQ-D)". *Diagnostica*, *50*(4), 171–181. https://doi.org/10.1026/0012-1924.50.4.171

Gutiérrez, F., Aluja, A., Ruiz Rodríguez, J., Peri, J. M., Gárriz, M., Garcia, L. F., Sorrel, M. A., Sureda, B., Vall, G., Ferrer, M., & Calvo, N. (2023). Borderline, where are you? A psychometric approach to the personality domains in the international classification of diseases, 11th revision (ICD-11). *Personality Disorders, Theory, Research, and Treatment*, *14*(3), 355–359. https://doi.org/10.1037/per0000592

Haase, M., Frommer, J., Franke, G. H., Hoffmann, T., Schulze-Muetzel, J., Jäger, S., Grabe, H. J., Spitzer, C., & Schmitz, N. (2008). From symptom relief to interpersonal change: Treatment outcome and effectiveness in inpatient psychotherapy. *Psychotherapy Research*, *18*(5), 615–624. https://doi.org/10.1080/10503300802192158

Haigh, E. A. P., Bogucki, O. E., Sigmon, S. T., & Blazer, D. G. (2018). Depression among older adults: A 20-year update on five common myths and misconceptions. *The American Journal of Geriatric Psychiatry*, *26*(1), 107–122. https://doi.org/10.1016/j.jagp.2017.06.011

Hartmann, A., von Wietersheim, J., Weiss, H., & Zeeck, A. (2018). Patterns of symptom change in major depression: Classification and clustering of long term courses. *Psychiatry Research*, *267*, 480–489. https://doi.org/10.1016/j.psychres.2018.03.086

Hilbert, K., Jacobi, T., Kunas, S. L., Elsner, B., Reuter, B., Lueken, U., & Kathmann, N. (2021). Identifying CBT non-response among OCD outpatients: A machine-learning approach. *Psychotherapy Research*, *31*(1), 52–62. https://doi.org/10.1080/10503307.2020.1839140

Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A., & Narayanan, A. (2022). The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning. https://doi.org/10.1145/3514094.3534196

Jacobucci, R., Littlefield, A. K., Millner, A. J., Kleiman, E. M., & Steinley, D. (2021). Evidence of inflated prediction performance: A commentary on machine learning and suicide research. *Clinical Psychological Science*, *9*(1), 129–134. https://doi.org/10.1177/2167702620954216

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning*. Springer.

Jankowsky, K., Zimmermann, J., Jaeger, U., Mestel, R., & Schroeders, U. (2023). First impressions count: Therapists' impression on patients' motivation and helping alliance predicts psychotherapy dropout. https://doi.org/10.31234/osf.io/nhs6c

Jolly, E., & Chang, L. J. (2019). The flatland fallacy: Moving beyond low-dimensional thinking. *Topics in Cognitive Science*, *11*(2), 433–454. https://doi.org/10.1111/tops.12404

Kaiser, T., Volkmann, C., Volkmann, A., Karyotaki, E., Cuijpers, P., & Brakemeier, E.-L. (2022). Heterogeneity of treatment effects in trials on psychotherapy of depression. *Clinical Psychology: Science and Practice*, *29*, 294–303. https://doi.org/10.1037/cps0000079

Kapoor, S., & Narayanan, A. (2022). Leakage and the reproducibility crisis in ML-based science (arXiv:2207.07048). arXiv. http://arxiv.org/abs/2207.07048

Kertz, S. J., Koran, J., Stevens, K. T., & Björgvinsson, T. (2015). Repetitive negative thinking predicts depression and anxiety symptom improvement during brief cognitive behavioral therapy. *Behaviour Research and Therapy*, *68*, 54–63. https://doi.org/10.1016/j.brat.2015.03.006

Kessler, R. C., Bossarte, R. M., Luedtke, A., Zaslavsky, A. M., & Zubizarreta, J. R. (2020). Suicide prediction models: A critical review of recent research with recommendations for the way forward. *Molecular Psychiatry*, *25*(1), 168–179. https://doi.org/10.1038/s41380-019-0531-0

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Ebert, D. D., de Jonge, P., Nierenberg, A. A., Rosellini, A. J., Sampson, N. A., Schoevers, R. A., Wilcox, M. A., & Zaslavsky, A. M. (2017). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiology and Psychiatric Sciences*, *26*(1), 22–36. https://doi.org/10.1017/S2045796016000020

Kleindienst, N., Priebe, K., Görg, N., Dyer, A., Steil, R., Lyssenko, L., Winter, D., Schmahl, C., & Bohus, M. (2016). State dissociation moderates response to dialectical behavior therapy for posttraumatic stress disorder in women with and without borderline personality disorder. *European Journal of Psychotraumatology*, *7*(1), 30375. https://doi.org/10.3402/ejpt.v7.30375

Kluge, U., Aichberger, M. C., Heinz, E., Udeogu-Gözalan, C., & Abdel-Fatah, D. (2020). Rassismus und psychische Gesundheit. *Der Nervenarzt*, *91*(11), 1017–1024. https://doi.org/10.1007/s00115-020-00990-1

Krause, M., & Behn, A. (2021). Depression and personality dysfunction: Towards the understanding of complex depression. In G. de la Parra, P. Dagnino, & A. Behn (Eds.), *Depression and personality dysfunction* (pp. 1–13). Springer International Publishing. https://doi.org/10.1007/978-3-030-70699-9_1

Kroenke, K., Wu, J., Yu, Z., Bair, M. J., Kean, J., Stump, T., & Monahan, P. O. (2016). Patient health questionnaire anxiety and depression scale: Initial validation in three clinical trials. *Psychosomatic Medicine*, *78*(6), 716–727. https://doi.org/10.1097/PSY.0000000000000322

Krueger, R., Derringer, J., Markon, K., Watson, D., & Skodol, A. (2013). *The personality inventory for DSM-5—Brief form (PID-5-BF) adult*. American Psychiatric Association.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Lee, Y., Ragguett, R.-M., Mansur, R. B., Boutilier, J. J., Rosenblat, J. D., Trevizol, A., Brietzke, E., Lin, K., Pan, Z., Subramaniapillai, M., Chan, T. C. Y., Fus, D., Park, C., Musial, N., Zuckerman, H., Chen, V. C.-H., Ho, R., Rong, C., & McIntyre, R. S. (2018). Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *Journal of Affective Disorders*, *241*, 519–532. https://doi.org/10.1016/j.jad.2018.08.073

Liebherz, S., & Rabung, S. (2014). Do patients' symptoms and interpersonal problems improve in psychotherapeutic hospital treatment in Germany? – A systematic review and meta-analysis. *PLoS One*, *9*(8), e105329. https://doi.org/10.1371/journal.pone.0105329

Löwe, B., Spitzer, R. L., Williams, J. B. W., Mussell, M., Schellberg, D., & Kroenke, K. (2008). Depression, anxiety and somatization in primary care: Syndrome overlap and functional impairment. *General Hospital Psychiatry*, *30*(3), 191–199. https://doi.org/10.1016/j.genhosppsych.2008.01.001

Luborsky, L., Auerbach, A. H., Chandler, M., Cohen, J., & Bachrach, H. M. (1971). Factors influencing the outcome of psychotherapy: A review of quantitative research. *Psychological Bulletin*, *75*(3), 145–185. https://doi.org/10.1037/h0030480

Lutz, W., de Jong, K., Rubel, J. A., & Delgadillo, J. (2021). Mesuring, predicting, and tracking change in psychotherapy. In M. Barkham, W. Lutz, & L. G. Castonguay (Eds.), *Handbook of psychotherapy and behavior change* (7. Aufl., S. 89–134). Wiley.

Lutz, W., Deisenhofer, A.-K., Rubel, J., Bennemann, B., Giesemann, J., Poster, K., & Schwartz, B. (2022). Prospective evaluation of a clinical decision support system in psychological therapy. *Journal of Consulting and Clinical Psychology*, *90*(1), 90–106. https://doi.org/10.1037/ccp0000642

Lyssenko, L., Schmahl, C., Bockhacker, L., Vonderlin, R., Bohus, M., & Kleindienst, N. (2018). Dissociation in psychiatric disorders: A meta-analysis of studies using the dissociative experiences scale. *American Journal of Psychiatry*, *175*(1), 37–46. https://doi.org/10.1176/appi.ajp.2017.17010025

Maj, M., Stein, D. J., Parker, G., Zimmerman, M., Fava, G. A., De Hert, M., Demyttenaere, K., McIntyre, R. S., Widiger, T., & Wittchen, H. (2020). The clinical characterization of the adult patient with depression aimed at personalization of management. *World Psychiatry*, *19*(3), 269–293. https://doi.org/10.1002/wps.20771

McClure, K., Jacobucci, R., & Ammerman, B. A. (2021). Are items more than indicators? An examination of psychometric homogeneity, item-specific effects, and consequences for structural equation models. https://doi.org/10.31234/osf.io/n4mxv

McNamara, M. E., Zisser, M., Beevers, C. G., & Shumake, J. (2022). Not just "big" data: Importance of sample size, measurement error, and uninformative predictors for developing prognostic models for digital interventions. *Behaviour Research and Therapy*, *153*, 104086. https://doi.org/10.1016/j.brat.2022.104086

Michal, M., & Beutel, M. E. (2009). Weiterbildung CME: Depersonalisation/Derealisation – Krankheitsbild, Diagnostik und Therapie. *Zeitschrift für Psychosomatische Medizin und Psychotherapie*, *55*(2), 113–140. https://doi.org/10.13109/zptm.2009.55.2.113

Michal, M., Wiltink, J., Till, Y., Wild, P. S., Blettner, M., & Beutel, M. E. (2011). Distinctiveness and overlap of depersonalization with anxiety and depression in a community sample: Results from the Gutenberg heart study. *Psychiatry Research*, *188*(2), 264–268. https://doi.org/10.1016/j.psychres.2010.11.004

Michal, M., Zwerenz, R., Tschan, R., Edinger, J., Lichy, M., Knebel, A., Tuin, I., & Beutel, M. (2010). Screening nach depersonalisation-Derealisation mittels zweier items der Cambridge depersonalisation scale. *Psychotherapie · Psychosomatik · Medizinische Psychologie*, *60*(5), 175–179. https://doi.org/10.1055/s-0029-1224098

Mula, M., Pini, S., & Cassano, G. B. (2007). The neurobiology and clinical significance of depersonalization in mood and anxiety disorders: A critical reappraisal. *Journal of Affective Disorders*, *99*(1–3), 91–99. https://doi.org/10.1016/j.jad.2006.08.025

Nolen-Hoeksema, S., Wisco, B. E., & Lyubomirsky, S. (2008). Rethinking rumination. *Perspectives on Psychological Science*, *3*(5), 400–424. https://doi.org/10.1111/j.1745-6924.2008.00088.x

Nowok, B., Raab, G. M., & Dibben, C. (2016). Synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software*, *74*(11), 1–26. https://doi.org/10.18637/jss.v074.i11

O'Hara, R., Beaudreau, S. A., Gould, C. E., Froehlich, W., & Kraemer, H. C. (2017). Handling clinical comorbidity in randomized clinical trials in psychiatry. *Journal of Psychiatric Research*, *86*, 26–33. https://doi.org/10.1016/j.jpsychires.2016.11.006

Pargent, F., Schoedel, R., & Stachl, C. (2023). Best practices in supervised machine learning: A tutorial for psychologists. *Advances in Methods and Practices in Psychological Science*, *6*(3). https://doi.org/10.1177/25152459231162559

Petrowski, K., Schmalbach, B., Kliem, S., Hinz, A., & Brähler, E. (2019). Symptom-checklist-K-9: Norm values and factorial structure in a representative German sample. *PLoS One*, *14*(4), e0213490. https://doi.org/10.1371/journal.pone.0213490

Philips, B., & Falkenström, F. (2021). What research evidence is valid for psychotherapy research? *Frontiers in Psychiatry*, *11*, 625380. https://doi.org/10.3389/fpsyt.2020.625380

Prinz, U., Nutzinger, D., Schulz, H., Petermann, F., Braukhaus, C., & Andreas, S. (2008). Die Symptom-Checkliste-90-R und ihre Kurzversionen: Psychometrische Analysen bei Patienten mit psychischen Erkrankungen. *Physikalische Medizin, Rehabilitationsmedizin, Kurortmedizin*, *18*(6), 337–343. https://doi.org/10.1055/s-0028-1093323

Sarter, L., Heider, J., Kirchner, L., Schenkel, S., Witthöft, M., Rief, W., & Kleinstäuber, M. (2021). Cognitive and emotional variables predicting treatment outcome of cognitive behavior therapies for patients with medically unexplained symptoms: A meta-analysis. *Journal of Psychosomatic Research*, *146*, 110486. https://doi.org/10.1016/j.jpsychores.2021.110486

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*, *4*(2), 1–12. https://doi.org/10.1177/25152459211007467

Schlax, J., Wiltink, J., Beutel, M. E., Münzel, T., Pfeiffer, N., Wild, P., Blettner, M., Ghaemi Kerahrodi, J., & Michal, M. (2020). Symptoms of depersonalization/derealization are independent risk factors for the development or persistence of psychological distress in the general population: Results from the Gutenberg health study. *Journal of Affective Disorders*, *273*, 41–47. https://doi.org/10.1016/j.jad.2020.04.018

Seeboth, A., & Mõttus, R. (2018). Successful explanations start with accurate descriptions: Questionnaire items as personality markers for more accurate predictions. *European Journal of Personality*, *32*(3), 186–201. https://doi.org/10.1002/per.2147

Senger, K., Rubel, J. A., Kleinstäuber, M., Schröder, A., Köck, K., Lambert, M. J., Lutz, W., & Heider, J. (2021). Symptom change trajectories in patients with persistent somatic symptoms and their association to long-term treatment outcome. *Psychotherapy Research*, *1–16*, 624–639. https://doi.org/10.1080/10503307.2021.1993376

Sheehan, D. V., Giddens, M. J. M., & Lisensi, P. P. (1983). Sheehan disability scale (SDS). *International Clinical Psychopharmacology*, *11*, 89–95.

Siddaway, A. P., Quinlivan, L., Kapur, N., O'Connor, R. C., & de Beurs, D. (2020). Cautions, concerns, and future directions for using machine learning in relation to mental health problems and clinical and forensic risks: A brief comment on "model complexity improves the prediction of nonsuicidal self-injury" (Fox et al., 2019). *Journal of Consulting and Clinical Psychology*, *88*(4), 384–387. https://doi.org/10.1037/ccp0000485

Simeon, D., Knutelska, M., Nelson, D., & Guralnik, O. (2003). Feeling unreal: A depersonalization disorder update of 117 cases. *The Journal of Clinical Psychiatry*, *64*(9), 990–997. https://doi.org/10.4088/JCP.v64n0903

Steinert, C., Hofmann, M., Kruse, J., & Leichsenring, F. (2014). Relapse rates after psychotherapy for depression – Stable long-term effects? A meta-analysis. *Journal of Affective Disorders*, *168*, 107–118. https://doi.org/10.1016/j.jad.2014.06.043

Taylor, D., Carlyle, J., McPherson, S., Rost, F., Thomas, R., & Fonagy, P. (2012). Tavistock adult depression study (TADS): A randomised controlled trial of psychoanalytic psychotherapy for treatment-resistant/treatment-refractory forms of depression. *BMC Psychiatry*, *12*(1), 60. https://doi.org/10.1186/1471-244X-12-60

Uckelstam, C.-J., Philips, B., Holmqvist, R., & Falkenström, F. (2019). Prediction of treatment outcome in psychotherapy by patient initial symptom distress profiles. *Journal of Counseling Psychology*, *66*(6), 736–746. https://doi.org/10.1037/cou0000345

Urban, C. J., & Gates, K. M. (2021). Deep learning: A primer for psychologists. *Psychological Methods*, *26*(6), 743–773. https://doi.org/10.1037/met0000374

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Vazire, S. (2020). A toast to the error detectors. *Nature*, *577*(7788), 9. https://doi.org/10.1038/d41586-019-03909-2

Wang, S., Zheng, S., Zhang, X., Ma, R., Feng, S., Song, M., Zhu, H., & Jia, H. (2023). The treatment of depersonalization-derealization disorder: A systematic review. *Journal of Trauma & Dissociation*, 1–24. https://doi.org/10.1080/15299732.2023.2231920

Webb, C. A., Cohen, Z. D., Beard, C., Forgeard, M., Peckham, A. D., & Björgvinsson, T. (2020). Personalized prognostic prediction of treatment outcome for depressed patients in a naturalistic psychiatric hospital setting: A comparison of machine learning approaches. *Journal of Consulting and Clinical Psychology*, *88*(1), 25–38. https://doi.org/10.1037/ccp0000451

Weiss, A., & Deary, I. J. (2020). A new look at neuroticism: Should we worry so much about worrying? *Current Directions in Psychological Science*, *29*(1), 92–101. https://doi.org/10.1177/0963721419887184

Wendt, L. P., Müller, S., & Zimmermann, J. (2023). Development and validation of the certainty about mental states questionnaire (CAMSQ): A self-report measure of mentalizing oneself and others. *Assessment*, *30*(3), 651–674. https://doi.org/10.1177/10731911211061280

Wilkinson, J., Arnold, K. F., Murray, E. J., van Smeden, M., Carr, K., Sippy, R., de Kamps, M., Beam, A., Konigorski, S., Lippert, C., Gilthorpe, M. S., & Tennant, P. W. G. (2020). Time to reality check the promises of machine learning-powered precision medicine. *The Lancet Digital Health*, *2*(12), e677–e680. https://doi.org/10.1016/S2589-7500(20)30200-4

Wiltink, J., Kliem, S., Michal, M., Subic-Wrana, C., Reiner, I., Beutel, M. E., Brähler, E., & Zwerenz, R. (2017). Mini – social phobia inventory (mini-SPIN): Psychometric properties and population based norms of the German version. *BMC Psychiatry*, *17*(1), 377. https://doi.org/10.1186/s12888-017-1545-2

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zeeck, A., von Wietersheim, J., Weiss, H., Beutel, M., & Hartmann, A. (2013). The INDDEP study: Inpatient and day hospital treatment for depression – Symptom course and predictors of change. *BMC Psychiatry*, *13*(1), 100. https://doi.org/10.1186/1471-244X-13-100

Zeeck, A., von Wietersheim, J., Weiß, H., Eduard Scheidt, C., Völker, A., Helesic, A., Eckhardt-Henn, A., Beutel, M., Endorf, K., Knoblauch, J., Rochlitz, P., & Hartmann, A. (2015). Symptom course in inpatient and day clinic treatment of depression: Results from the INDDEP-study. *Journal of Affective Disorders*, *187*, 35–44. https://doi.org/10.1016/j.jad.2015.07.025

Zeeck, A., von Wietersheim, J., Weiss, H., Hermann, S., Endorf, K., Lau, I., & Hartmann, A. (2020). Self-criticism and personality functioning predict patterns of symptom change in major depressive disorder. *Frontiers in Psychiatry*, *11*, 147. https://doi.org/10.3389/fpsyt.2020.00147

Zeeck, A., von Wietersheim, J., Weiss, H., Scheidt, C. E., Völker, A., Helesic, A., Eckhardt-Henn, A., Beutel, M., Endorf, K., Treiber, F., Rochlitz, P., & Hartmann, A. (2016). Prognostic and prescriptive predictors of improvement in a naturalistic study on inpatient and day hospital treatment of depression. *Journal of Affective Disorders*, *197*, 205–214. https://doi.org/10.1016/j.jad.2016.03.039

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Jankowsky, K., Krakau, L., Schroeders, U., Zwerenz, R., & Beutel, M. E. (2024). Predicting treatment response using machine learning: A registered report. *British Journal of Clinical Psychology*, *63*, 137–155. https://doi.org/10.1111/bjc.12452