Psychiatry Research 336 (2024) 115910

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

Predicting post-treatment symptom severity for adults receiving psychological therapy in routine care for generalised anxiety disorder: a machine learning approach

H. Delamain [a,*], J.E.J. Buckman [a,b], C. O'Driscoll [a], J.W. Suh [a], J. Stott [c], S. Singh [d], S.A. Naqvi [e], J. Leibowitz [b], S. Pilling [a,f], R. Saunders [a]

[a] CORE Data Lab, Centre for Outcomes Research and Effectiveness (CORE), Research Department of Clinical, Educational and Health Psychology, UCL, London, United Kingdom
[b] iCope - Camden and Islington Psychological Therapies Services, Camden & Islington NHS Foundation Trust, London, United Kingdom
[c] ADAPT Lab, Research Department of Clinical, Educational and Health Psychology, UCL, London, United Kingdom
[d] Waltham Forest Talking Therapies, North East London NHS Foundation Trust, London, United Kingdom
[e] Barking and Dagenham and Havering IAPT Services, North East London NHS Foundation Trust, London, United Kingdom
[f] Camden and Islington NHS Foundation Trust, London, United Kingdom

ARTICLE INFO

Keywords:
Personalised treatment
Prognosis prediction
Outcome monitoring
Ensemble modelling
TRIPOD

ABSTRACT

Approximately half of generalised anxiety disorder (GAD) patients do not recover from first-line treatments, and no validated prediction models exist to inform individuals or clinicians of potential treatment benefits. This study aimed to develop and validate an accurate and explainable prediction model of post-treatment GAD symptom severity. Data from adults receiving treatment for GAD in eight Improving Access to Psychological Therapies (IAPT) services (n=15,859) were separated into training, validation and holdout datasets. Thirteen machine learning algorithms were compared using 10-fold cross-validation, against two simple clinically relevant comparison models. The best-performing model was tested on the holdout dataset and model-specific explainability measures identified the most important predictors. A Bayesian Additive Regression Trees model out-performed all comparison models (MSE=16.54 [95 % CI=15.58; 17.51]; MAE=3.19; R^2=0.33, including a single predictor linear regression model: MSE=20.70 [95 % CI=19.58; 21.82]; MAE=3.94; R^2=0.14). The five most important predictors were: PHQ-9 anhedonia, GAD-7 annoyance/irritability, restlessness and fear items, then the referral-assessment waiting time. The best-performing model accurately predicted post-treatment GAD symptom severity using only pre-treatment data, outperforming comparison models that approximated clinical judgement and remaining within the GAD-7 error of measurement and minimal clinically important differences. This model could inform treatment decision-making and provide desired information to clinicians and patients receiving treatment for GAD.

1. Introduction

Generalised anxiety disorder (GAD) is one of the most commonly occurring and burdensome mental disorders (Ferrari et al., 2022). Efficacious pharmacological and psychotherapeutic treatments exist (Bandelow et al., 2017), but even with the first-line recommended treatments only around half of patients in clinical trials or routine care achieve symptomatic recovery post-treatment (Clark, 2018; Loerinc et al., 2015). There is, therefore, a pressing need to improve patient treatment outcomes.

Knowledge of individual treatment prognosis can inform clinical planning, potentially improving patient outcomes and healthcare cost-effectiveness (Chekroud et al., 2021). Such knowledge is wanted by both patients and clinicians (Hayden et al., 2013), with a wider societal demand for precision medicine (Fernandes et al., 2017). Mental health resources are constrained (Clark, 2018) and so the need for accurate patient prognostic prediction models is clear. Despite this, there is a notable lack of such research with patients with GAD.

Clinical prediction models are often constrained by the limited use of routinely collected data in their development (Dwyer et al., 2018). This

* Corresponding author.
E-mail address: henry.delamain@ucl.ac.uk (H. Delamain).

https://doi.org/10.1016/j.psychres.2024.115910
Received 3 August 2023; Received in revised form 3 April 2024; Accepted 8 April 2024
Available online 9 April 2024

contributes to reduced generalisability and a lack of ecologically valid ways of explaining how predictions are obtained (Bouwmeester et al., 2012). Such models are therefore not readily implementable, do not perform well in routine care and lack transparency in how predictions are obtained, limiting their value and associated trust (Obermeyer and Emanuel, 2016). To increase the potential utility, prediction models need to be developed using naturalistic samples, leverage explainability measures (to improve model interpretability), and adhere to accepted development guidelines (Collins et al., 2015). This study aims to develop and validate an accurate and explainable prognostic prediction model for adults receiving psychological treatment for GAD, using routinely collected data.

2. Methods

The methods for this study were preregistered and were adhered to without deviation (see: https://osf.io/s23hc/).

2.1. Participants

Routinely collected healthcare data were analysed for adults (aged ≥18) treated for GAD in eight Improving Access to Psychological Therapies (IAPT) services (grouped together in four NHS Trusts) in the North and Central East London IAPT Service Improvement and Research Network (NCEL IAPT SIRN; see Supplementary Material 1 Table 3; Saunders et al., 2020). The national IAPT programme delivers evidence-based psychological therapies using a stepped care model for depression and anxiety disorders, including GAD (Clark, 2018). Included individuals within this study had a primary diagnosis of GAD (referred to as 'problem descriptor' in IAPT services; see Supplementary Material 1 Table 2 for further details), had received two or more treatment sessions and had completed treatment by August 2020.

2.2. Predictors and outcomes

IAPT services are mandated by NHS England to collect a standardised set of sociodemographic data, as well as symptom and functioning measures: the IAPT Minimum Dataset (MDS). The predictors used in this study, that were those available to inform clinical decision making prior to starting treatment and were all collected routinely as part of the MDS at the initial assessment, are the individual items from the Generalised Anxiety Disorder Assessment (GAD-7; Spitzer et al., 2006) used to measure symptoms of generalised anxiety disorder, the Patient Health Questionnaire (PHQ-9; Kroenke et al., 2001) for depressive symptoms, the Work and Social Adjustment Scale (WSAS; Mundt et al., 2002) for functional impairment, and the IAPT Phobia Scale Improving Access to Psychological Therapies (2011) to screen for phobic anxiety disorders. The GAD-7 and PHQ-9 total scores were also included. Only items two-to-five of the WSAS were used as the first item is not applicable for unemployed individuals. Other predictors were: age, ethnicity, employment status, gender, long-term health condition status, and psychotropic medication prescription and usage. Local Layer Super Output Areas (LSOAs) were converted to Indices of Multiple Deprivation (IMD) as a measure of local area deprivation. Finally, the number of weeks between referral and initial assessment, and between assessment and first therapy session were included. The primary outcome was the GAD-7 score at the last attended therapy session. The full description of baseline variables is included in Supplementary Material 1 (Tables 1 and 2).

The Transparent Reporting of a Multivariable Prediction model for Individual Prognosis Or Diagnosis (TRIPOD) Statement (Collins et al., 2015) guidelines were followed to ensure model development best practice; see Supplementary Material 2.

2.3. Data processing

Analyses were undertaken in R (R Core Team, 2022). Missing baseline data were single imputed using the 'missForest' package (Stekhoven and Bühlmann, 2012), following conventions from other prediction model development and validation studies (Buckman et al., 2021a; Webb et al., 2020), with missingness varying by predictor from nil to 22.7 % (long-term health condition). Continuous variables were centred and scaled. Categorical variables were dummy encoded post-imputation, with the first dummy encoded variable removed to avoid issues regarding multicollinearity.

The dataset was then divided into three for training, validation and testing, as detailed in the study protocol. Division was determined by service locations within NHS Trusts and undertaken to create an accurate representation of the model's ability to generalise to services beyond those whose data were used to develop and validate the models (Chekroud and Koutsouleris, 2018). The variables collected and delivery of interventions within these service locations are standardised across England (National Collaborating Centre for Mental Health, 2023). The training dataset services were three from Trust 1 and three from Trust 2 (n=8,064; 50.8 %). The validation dataset services were two from Trust 3 (n=5,021; 31.7 %). The holdout dataset service was a single one from Trust 4 (n=2,774; 17.5 %). To identify statistically significant group differences between the three independent samples, chi-square tests were used for the categorical variables, and ANOVA with Welch t-tests for continuous variables. Detection of such differences did not influence the submission of a given predictor to the prediction models.

2.4. Machine learning algorithms

Thirteen machine learning algorithms were selected to provide a broad comparison (of a variety of modelling techniques) and to be consistent with research that has used similar methods (Webb et al., 2020). To provide a benchmark against the machine learning models, an ordinary least squares regression (OLS) model with all the available same predictors (in the same manner as the machine learning models) was included. The multiple OLS acts as a comparator with penalised regression models by fitting the data without any regularisation, making it prone to overfitting compared to models with regularisation. Additionally, two other comparison models akin to those used in similar research were fitted (Buckman et al., 2021a). The first was an OLS regression model with only the pre-treatment GAD-7 score included, as initial symptom severity is one of the most important factors considered by clinicians in predicting treatment outcomes (and allocating treatment) for their patients (Amati et al., 2018; Buckman et al., 2021b; O'Driscoll et al., 2021). Separately, a null model was created using the mean post-treatment GAD-7 score in the training and validation dataset as the prediction for all test dataset cases. Five models based upon elastic net regularised regression (ENR; Friedman et al., 2010) were included (alpha parameters were 0.0; full ridge, 0.25, 0.50, 0.75 and 1.0; full lasso), using the ridge and least absolute shrinkage and selection operator (lasso) penalisations to reduce overfitting. Two spline regression models were used: adaptive splines (Friedman, 1991) and adaptive polynomial splines (Stone et al., 1997). In addition, two decision-tree-based algorithms were used: random forest (Breiman, 2001) and Bayesian Additive Regression Trees (BART; Chipman et al., 2010). Lastly, three support vector machines (SVM) were included with linear, polynomial and radial kernels (Dimitriadou et al., 2009). Full details of the models are presented in Supplementary Material 1 (Table 3).

SuperLearner (van der Laan et al., 2007), the chosen ensemble modelling tool, uses cross-validation to estimate the predictive performance of individual machine learning algorithms, and ultimately determines the most accurate model within the ensemble. The primary evaluation metric was the mean squared error (MSE) and the secondary metrices were variance explained (coefficient of determination; R^2) and

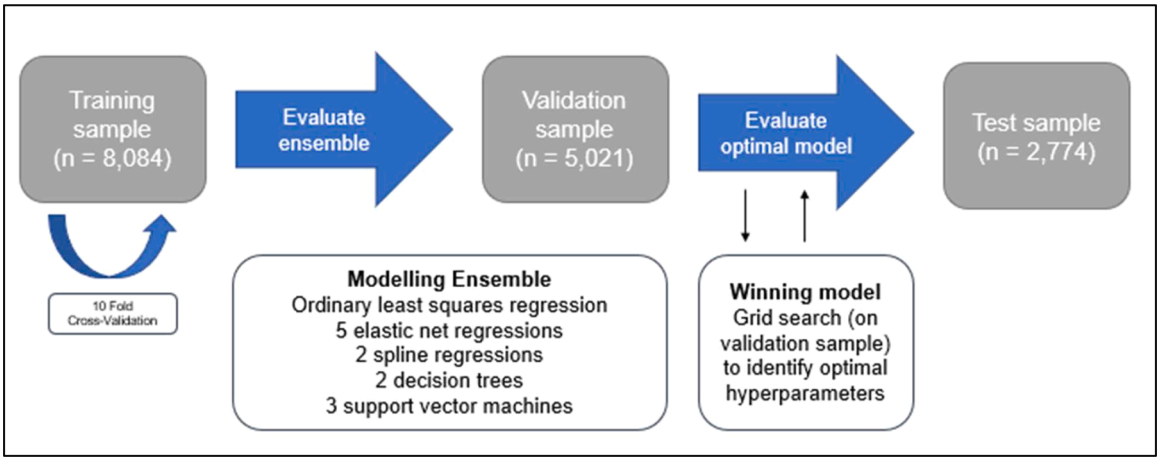

Fig. 1. Analysis pipeline workflow.

the mean absolute error (as an average prediction error; MAE).

Initially, the ensemble was trained within the training dataset using 10-fold cross-validation. The validation dataset was used as an initial test of each model's ability to generalise, with the model with the lowest MSE selected (taken as an average over the ten repeats). Once the best performing algorithm had been established within the validation dataset, the exact same training and validation process was repeated to tune its hyperparameters, using the grid search method of the 'caret' package (see Supplementary Material 1; Max, 2008). The tuning of hyperparameters was only undertaken for the best performing algorithm to ensure an equal comparison (as hyperparameters vary across models) of the baseline models within the validation dataset.

Subsequently, only the model with tuned hyperparameters was trained on the combined training and validation dataset, then finally tested on the test dataset. The analysis pipeline is shown in Fig. 1.

2.5. Explainability

In line with the study protocol, model-specific explainability measures were preferred where the winning model was an inherently interpretable algorithm (such as a decision tree) as they leverage individual model's distinct features, thereby improving fidelity (Belle and Papantonis, 2021). Additionally, model-specific measures bypass known issues (such as an inability to isolate individual contributions of individual features) with global and local model-agnostic explainability methods that use input perturbation within highly correlated variable sets, such as SHAP or LIME (Slack et al., 2020). Where model-specific measures did not exist or were not available, accumulated local explanation plots (Apley and Zhu, 2019) were used as a global model-agnostic explainability measure, due to their ability to handle correlated features.

2.6. Ethics

NHS Ethical approval was not required for this study (confirmed by the Health Research Authority July 2020 #81/81). The IAPT services provided data as part of a wider service improvement project. This research followed procedures outlined by the respective data hosting providers and was registered with the individual NHS Trusts operating the IAPT services (project reference: 00,519-IAPT).

3. Results

There were 30,833 patients treated for GAD. Of these, 14,974 did not meet inclusion criteria for reasons such as having <2 treatment sessions or having missing PHQ-9 or GAD-7 item level data (see Supplementary Material 1 Figure 1). This resulted in an analytic sample of n=15,859 participants which was split into n=8,064 in the training dataset, n=5,021 in the validation dataset, and n=2,774 in the test dataset. Out of the sample of 15,958 participants, 1,397 (8.81 %) had only two sessions and 7,959 (50.2 %) completed at least six sessions. Demographic and clinical characteristics for participants in each dataset are shown and compared in Table 1, with statistically significant differences observed on multiple variables between the datasets.

3.1. Exploratory ensemble modelling

The mean of the post-treatment GAD-7 scores in the validation sample was 7.4 (±5.2) and 37.1 % individuals were in caseness at the end of treatment (their GAD-7 score was ≥8). Through training the ensemble using 10-fold internal cross-validation within the training sample (repeated ten times), the BART algorithm produced the lowest prediction error when generalising to the validation dataset (MSE [95 % CI]=16.72 [16.12; 17.32]; MAE=3.16; R^2=0.38; see Table 2). The BART model also had the combined lowest prediction error in terms of the secondary outcome metric (MAE) along with the adaptive polynomial spline model. Both models' MAE indicated that on average the predicted post-treatment GAD-7 scores differed from the true values by 3.16 points. The linear regression model (with all predictors) had the greatest R^2 but was more error prone to error, with higher MSE and MAE values. As MSE was the primary evaluation metric, the BART model was determined to be the winning model within the ensemble.

3.2. Optimal BART prognosis prediction model

The available hyperparameters of the winning model (BART) were tuned using 'caret' (Max, 2008) and were determined to be: number of trees=50, k=2, α=0.95, β=0.5 and ν=3; see Supplementary Material 1 for further details. Consequently, the obtained performance metrics for the tuned model within the validation dataset were: MSE=16.59 [95 % CI=15.92; 17.27]; MAE=3.26; R^2=0.40.

The mean of the post-treatment GAD-7 scores in the test sample was 7.3 (±4.9) and 36.0 % of individuals were in caseness at the end of treatment. The hyperparameter tuned BART model was tested on the holdout dataset to predict post-treatment GAD-7 scores (MSE=16.54 [95 % CI=15.58; 17.51]; MAE=3.19; R^2=0.33). The single predictor OLS comparison model achieved the following performance metrics: MSE=20.70 [95 %CI = 19.58; 21.82]; MAE=3.94; R^2=0.14).

Table 1
Demographics and group comparison.

Sample Characteristics	Training (%)	Validation (%)	Testing (%)	Group Comparison (Chi Square or ANOVA)
Overall	8064 (50.8)	5021 (31.7)	2774 (17.5)	
Employment status				X^2 (14, N=15574)=754.1, p<0.001
Employed	5300 (65.7)	3674 (73.2)	2066 (74.5)	
Unemployed but seeking work	338 (4.2)	147 (2.9)	242 (8.7)	
Student	548 (6.8)	257 (5.1)	301 (10.9)	
Long-term sick or disabled	463 (5.7)	117 (2.3)	31 (1.1)	
Homemaker	368 (4.6)	157 (3.1)	49 (1.8)	
Not working	466 (5.8)	339 (6.8)	3 (0.1)	
Voluntary work	48 (0.6)	14 (0.3)	5 (0.2)	
Retired	461 (5.7)	114 (2.3)	66 (2.4)	
Missing	72 (0.9)	202 (4.0)	11 (0.4)	
Ethnicity				X^2 (10, N=15323)=221.0, p<0.001
Asian	736 (9.1)	218 (4.3)	153 (5.5)	
Black	686 (8.5)	428 (8.5)	115 (4.1)	
Chinese	51 (0.6)	37 (0.7)	31 (1.1)	
Mixed	442 (5.5)	227 (4.5)	149 (5.4)	
Other	346 (4.3)	134 (2.7)	81 (2.9)	
White	5638 (69.9)	3718 (74.0)	2133 (76.9)	
Missing	442 (5.5)	259 (5.2)	112 (4.0)	
Gender[†]				X^2 (2, N=15824)=3.7, p=0.159
Female	5809 (72.0)	3534 (70.4)	1992 (71.8)	
Male	2239 (27.8)	1469 (29.3)	781 (28.2	
Missing	16 (0.2)	18 (0.4)	1 (0.0)	
Long term health condition				X^2 (2, N=12257)=155.9, p<0.001
No	4969 (61.6)	2233 (44.5)	1957 (70.5)	
Yes	1993 (24.7)	434 (8.6)	671 (24.2)	
Missing	1102 (13.7)	2354 (46.9)	146 (5.3)	
Medication status				X^2 (4, N=14850)=81.8, p<0.001
Not prescribed	4913 (60.9)	3190 (63.5)	1956 (70.5)	
Prescribed and taking	2338 (29.0)	1119 (22.3)	680 (24.5)	
Prescribed not taking	378 (4.7)	161 (3.2)	115 (4.1)	
Missing	435 (5.4)	23 (0.5)	551 (19.9)	
		Mean (± SD)		
Age	38.5 (14.2)	35.1 (11.2)	33.9 (11.9)	**F(2, 15856)**=185.8 (p<0.001)
IMD Decile	4.6 (2.4)	2.8 (1.3)	3.6 (1.9)	**F(2, 15708)**=1206 (p<0.001)
missing (n; %)	*66; 0.8*	*59; 1.2*	*23; 0.8*	
Agoraphobia score	*2.4 (2.6)*	2.2 (2.5)	2.2 (2.3)	**F(2, 15527)**=12.5 (p<0.001)
missing (n; %)	*23; 0.3*	*290; 5.8*	*16; 0.6*	
Social phobia score	2.6 (2.4)	2.5 (2.2)	2.5 (2.1)	**F(2, 15527)**=2.7 (p=0.065)
missing (n; %)	*23; 0.3*	*290; 5.8*	*16; 0.6*	
Specific phobia score	2.1 (2.6)	1.8 (2.5)	1.9 (2.3)	**F(2, 15527)**=20.1 (p<0.001)
missing (n; %)	*23; 0.3*	*290; 5.8*	*16; 0.6*	
GAD-7 Nervousness	2.3 (0.8)	2.2 (0.9)	2.3 (0.8)	**F(2, 15856)**=18.7 (p<0.001)
GAD-7 Worry (control)	2.3 (0.9)	2.2 (0.9)	2.2 (0.9)	**F(2, 15856)**=12.3 (p<0.001)
GAD-7 Worry (excessiveness)	2.3 (0.9)	2.2 (0.9)	2.3 (0.8)	**F(2, 15856)**=18.7 (p<0.001)
GAD-7 Trouble relaxation	2.0 (0.9)	2.0 (0.9)	2.0 (0.9)	**F(2, 15856)**=3.3 (p=0.036)
GAD-7 Restlessness	1.2 (1.1)	1.1 (1.0)	1.0 (1.0)	**F(2, 15856)**=17.9 (p<0.001)
GAD-7 Annoyance/Irritability	1.7 (1.0)	1.7 (1.0)	1.6 (1.0)	**F(2, 15856)**=14.0 (p<0.001)
GAD-7 Apprehensive expectation	1.7 (1.1)	1.6 (1.1)	1.6 (1.1)	**F(2, 15856)**=17.6 (p<0.001)
GAD-7 (pre-treatment) total	13.9 (4.7)	13.4 (4.7)	13.2 (4.5)	**F(2, 15856)**=27.2 (p<0.001)
GAD-7 (post-treatment) total	7.5 (5.6)	7.4 (5.2)	7.3 (4.9)	**F(2, 15856)**=2.8 (p=0.064)
PHQ-9 Anhedonia	1.0 (0.9)	0.9 (0.9)	0.9 (0.8)	**F(2, 15856)**=21.13 (p<0.001)
PHQ-9 Depressed mood	1.6 (1.0)	1.5 (0.9)	1.5 (0.9)	**F(2, 15856)**=57.6 (p<0.001)
PHQ-9 Sleep	1.8 (1.1)	1.7 (1.1)	1.7 (1.0)	**F(2, 15856)**=25.5 (p<0.001)
PHQ-9 Energy	1.9 (1.0)	1.8 (1.0)	1.8 (0.9)	**F(2, 15856)**=32.5 (p<0.001)
PHQ-9 Appetite	1.3 (1.1)	1.1 (1.1)	1.1 (1.0)	**F(2, 15856)**=56.2 (p<0.001)
PHQ-9 Failure	1.7 (1.1)	1.6 (1.0)	1.6 (1.0)	**F(2, 15856)**=9.0 (p<0.001)
PHQ-9 Concentration	1.5 (1.1)	1.4 (1.0)	1.4 (1.0)	**F(2, 15856)**=4.2 (p=0.015)
PHQ-9 Psychomotor symptoms	0.8 (1.0)	0.7 (1.0)	0.7 (0.9)	**F(2, 15856)**=32.5 (p<0.001)
PHQ-9 Suicidal ideation	0.3 (0.7)	0.3 (0.6)	0.3 (0.6)	**F(2, 15856)**=7.0 (p<0.001)
PHQ-9 (pre-treatment) total	12.9 (5.9)	11.9 (5.7)	11.4 (5.4)	**F(2, 15856)**=81.1 (p<0.001)
Number of sessions	7.38 (4.37)	7.44 (4.70)	7.04 (4.11)	**F(2, 15856)**=8.1 (p<0.001)
Weeks from ref. to assess.	3.8 (4.7)	3.7 (3.7)	3.7 (2.5)	**F(2, 15854)**=2.0(p=0.136)
missing (n; %)	*2; 0.0*			
Weeks from assess. to treatment	9.8 (9.2)	6.7 (7.7)	8.0 (6.4)	**F(2, 15129)**=203.9 (p<0.001)
missing (n; %)	*502; 6.2*	*85; 1.7*	*140; 5.0*	
WSAS question 2	3.1 (2.3)	2.9 (2.3)	2.7 (2.0)	**F(2, 15502)**=30.0 (p<0.001)
missing (n; %)	*23; 0.3*	*313; 6.2*	*18; 0.6*	
WSAS question 3	3.7 (2.4)	3.6 (2.3)	3.3 (2.1)	**F(2, 15497)**=31.5 (p<0.001)
missing (n; %)	*24; 0.3*	*317; 6.3*	*18; 0.6*	
WSAS question 4	3.1 (2.5)	3.1 (2.4)	2.8 (2.2)	**F(2, 15495)**=19.7 (p<0.001)
missing (n; %)	*24; 0.3*	*319; 6.4*	*18; 0.6*	
WSAS question 5	3.5 (2.4)	3.5 (2.3)	3.1 (2.2)	**F(2, 15493)**=40.9 (p<0.001)
missing (n; %)	*24; 0.3*	*321; 6.4*	*18; 0.6*	

Note. [†]Service users were asked about their 'gender' and were provided with these options; we note that this does not consider the full range of experience but are the only categories available within IAPT services. Welch *t*-tests are reported for significant ANOVA (<0.001) in Supplementary Material 1 (Table 4).

Table 2
Ensemble validation dataset results.

Category of algorithm	Algorithm	MSE [95 % CI]	SE	MAE	R^2
Multiple predictor OLS	Linear regression	16.86 [16.27 17.44]	0.38	3.20	0.38
Penalised regression	Ridge regression	16.81 [16.23 17.39]	0.37	3.20	0.38
	Elastic net (alpha = 0.25)	16.83 [16.24 17.41]	0.38	3.20	0.38
	Elastic net (alpha = 0.50)	16.83 [16.25 17.42]	0.38	3.20	0.38
	Elastic net (alpha = 0.75)	16.84 [16.25 17.42]	0.38	3.20	0.38
	LASSO regression	16.84 [16.25 17.42]	0.38	3.20	0.38
Spline regression	Adaptive splines	16.94 [16.32 17.56]	0.40	3.17	0.37
	Adaptive polynomial splines	16.77 [16.17 17.37]	0.39	3.16	0.38
Decision tree	Random forest	16.78 [16.20 17.36]	0.38	3.19	0.38
	Bayesian additive regression trees	16.72 [16.12 17.32]	0.39	3.16	0.38
Support vector regression	SVR (polynomial)	19.61 [18.89 20.33]	0.47	3.38	0.30
	SVR (linear)	17.10 [16.48 17.71]	0.40	3.19	0.37
	SVR (radial)	17.65 [17.01 18.29]	0.41	3.22	0.36
Single predictor OLS	Linear regression	23.52 [22.63 24.40]	0.45	3.90	0.40
Null model	Post-treatment mean	26.99 [25.98 28.01]	0.52	4.21	N/A

Note. MSE = mean squared error; CI = confidence interval; SE = standard error of MSE; MAE = mean absolute error; R^2 = coefficient of determination; OLS = ordinary least squares; SVR = support vector regression.

3.3. Explainability

Using the model-specific explainability measures available, the inclusion proportion for each individual feature used as a splitting rule within the BART tree was calculated (for further details: see Supplementary Material 1 Table 8). The PHQ-9 item on anhedonia was the most important variable for prediction, chosen twice as many times as the next most important predictor. Next, chosen around 3.5 % of times were three GAD-7 items (annoyance/irritability, restlessness and fear) and the number of weeks from referral to assessment. The remaining variables performed similarly and those included ≥2.5 % of times are shown in Fig. 2.

4. Discussion

The aim of this study was to develop and validate an accurate and explainable prognosis prediction model for patients receiving psychological treatment for GAD. The winning BART model can predict an accurate index score within the error of measurement for the GAD-7 (±4; Spitzer et al., 2006), as well as within a minimal clinically important difference (MCID) score (±3.3; Bauer-Staeb et al., 2021) between the true and observed values. The obtained accuracies were better than the base comparison models, that on average made predictions outside the error of measurement and the MCID from the true GAD-7 scores. The model was developed using only routinely collected pre-treatment data, highlighting the potential value of this prediction model. Several key variables that all had face-validity were identified, and improve the explainability of the model. These variables provide some insight to the underlying behaviour of the model to clinicians and patients. Anhedonia (from the PHQ-9) was the most important predictor, followed by three GAD-7 items. This is consistent with other research showing symptoms of GAD and depression to be highly comorbid and influential in treatment outcomes (O'Driscoll et al., 2021) and anhedonia substantially contributing to psychological therapy outcomes (Khazanov et al., 2020). Waiting time was similarly evidenced as a key indicator of psychological treatment outcomes (Clark, 2018).

4.1. Strengths and limitations

This study had a limited inclusion criteria, used a large naturalistic sample to develop the models, and used routinely collected pre-treatment data, improving external validity and addressing generalisability concerns (Meehan et al., 2022). However, data came from services in similar geographical areas, so replication in other primary or community care mental health services that treat GAD, both within the UK and internationally (Cromarty et al., 2016; Knapstad et al., 2018) would provide a more thorough test of generalisability. The study compared a varied range of machine learning algorithms and presents a potential baseline for future research. In addition, the winning model is only usable in services that collect the same data and that provide the same psychological therapies as those used in this study. However, in-line with the study protocol, only the most important variables for prediction were explored for the winning model as we considered this to be the most likely to be used in clinical practice. The use of explainability measures follows recommendations for the development of machine learning prediction models (Dwyer et al., 2018) to build confidence and trust in predictions, increasing the potential for utility. However, within the context of the study, the included simple comparison models were only an approximation of how clinicians might make prognostic predictions. For a more complete evaluation, a comparative test of the prospective prediction performance of the model relative to clinicians is required. It would also be valuable to explore the important

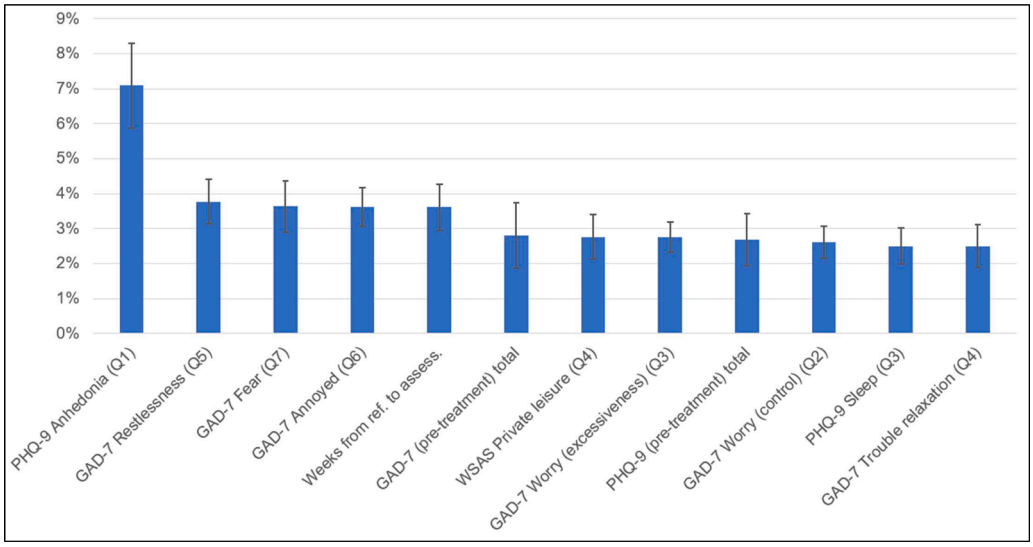

Fig. 2. Percentage of times each predictor was chosen as a splitting rule within BART
Note. Only predictors with ≥2.5 % inclusion proportion shown; see Supplementary Material 1 Table 8 for the remaining predictors.

variables for prediction for the other models within the ensemble. The study was conducted and reported in accordance with TRIPOD guidelines that might help to improve the potential replicability and interpretability (Burke et al., 2019). This is particularly important given the paucity of prognostic GAD treatment models.

Separate prediction models could have been produced for those receiving low and high-intensity treatments, given their distinction in clinical guidelines (National Institute for Health and Care Excellence, 2020). However, as patients can be stepped up or down between intensities, or receive pharmacotherapy concurrent to their psychological therapy intensity level, separating the models does not best represent the clinical settings that models might be used within. Treatment options for GAD within IAPT are limited to low and high intensity CBT, thus limiting the generalisability of the model to alternative therapies. The dataset could have also been partitioned by sociodemographic features (such as age or ethnicity), or severity of symptoms, to provide additional tests of the robustness of the findings.

There was a large amount of unexplained variance for individual models, although this could be partly attributed to measurement error (Bone et al., 2021). Models might have been more accurate with a broader range of predictors, such as those collected through passive measurement of disorder-specific behaviours (De Angel et al., 2022). However, this study only used routinely collected data to ensure greater potential for clinical utility. The hyperparameters were only tuned for the winning model from the first test of generalisation (i.e. the BART model), so an alternative potentially more accurate model could have been obtained if all models were tuned prior. However, this method provided an equal baseline comparison of all models within the ensemble and provides a potential basis for future research.

4.2. Implications and conclusions

The winning model could be embedded within healthcare systems to provide a more accurate treatment prognosis for individuals following an initial assessment, providing patients and clinicians with desire knowledge and informing their joint clinical decision-making. Prior to embedding and being used to help guide treatment decisions, a more robust test of the value of the model in clinical practice would be needed. This might include randomising clinicians and patients to receive (or not receive) predictions from the model prospectively, then investigating the effects of this on treatment outcomes. This process has been found to be successful with other data-informed treatment predictions in similar settings (Delgadillo et al., 2018). If there is utility in the routine use of the model, improved treatment outcomes and potentially reduced costs of care would be expected. Those facing poor treatment prognoses would be recommended to: start at high-intensity, have more regular reviews to closely monitor progress throughout treatment, consider combined pharmacotherapy and psychotherapy, or be given augmented treatment options. Those with particularly good prognoses could be recommended to start with lower intensity therapy which might make more efficient use of clinical resources. Further, were causality to be demonstrated, the identified important variables for prediction might help clinicians and services use targeted interventions. For example, offering interventions that target anhedonia (Khazanov et al., 2020) or reducing the number of weeks waited (fast-tracking) for individuals with poor predicted prognoses. This might further improve outcomes and reduce the long-term cost of care. Future research would seek to comparatively test the model against clinicians' predictions, obtain a parsimonious model and further demonstrate the generalisability of the winning model.

CRediT authorship contribution statement

H. Delamain: Writing – original draft, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **J.E.J. Buckman:** Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **C. O'Driscoll:** Writing – review & editing, Methodology. **J.W. Suh:** Writing – review & editing, Methodology. **J. Stott:** . **S. Singh:** Writing – review & editing, Data curation. **S.A. Naqvi:** Writing – review & editing, Data curation. **J. Leibowitz:** Writing – review & editing, Data curation. **S. Pilling:** Writing – review & editing, Funding acquisition, Data curation. **R. Saunders:** Writing – original draft, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

All authors declare that there are no conflicts of interest.

Role of the funding source

The funder of the study had no role in the study design, data collection, data analysis, data interpretation, or writing of the report.

Acknowledgements

The North and Central East London (NCEL) IAPT Service Improvement and Research Network (SIRN) includes: Ana Cardoso, Andre Lynam-Smith, Catherine Simpson, Evi Aresti, Helmi Van Leur, John Cape, Jon Wheatley, Joshua Cane, Joshua E J Buckman, Judy Leibowitz, Laura Fontaine, Maria Perez, Mina Spatha, Mirko Cirkovic, Renuka Jena, Rob Saunders, Sarah Ellard, Satwant Singh, Stephen Pilling, Syed Ali Naqvi, Tania Knight & Tina Cross. We would like to thank all clinicians and patients from NCEL IAPT services. We are grateful to the service leads for their support with the NCEL project and to the local data managers for their time and dedication. In addition, we would like to thank Adam Kapelner for his support and advice.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.psychres.2024.115910.

References

Amati, F., Banks, C., Greenfield, G., Green, J., 2018. Predictors of outcomes for patients with common mental health disorders receiving psychological therapies in community settings: a systematic review. J. Public Health 40 (3), Article 3. https://doi.org/10.1093/pubmed/fdx168.

Apley, D.W., & Zhu, J. (2019). Visualizing the effects of predictor variables in black box supervised learning models (arXiv:1612.08468). arXiv. 10.48550/arXiv.1612.08468.

Bandelow, B., Michaelis, S., Wedekind, D., 2017. Treatment of anxiety disorders. Dialogues Clin. Neurosci. 19 (2), Article 2.

Bauer-Staeb, C., Kounali, D.Z., Welton, N.J., Griffith, E., Wiles, N.J., Lewis, G., Faraway, J.J., Button, K.S., 2021. Effective dose 50 method as the minimal clinically important difference: evidence from depression trials. J. Clin. Epidemiol. 137, 200–208. https://doi.org/10.1016/j.jclinepi.2021.04.002.

Belle, V., Papantonis, I., 2021. Principles and practice of explainable machine learning. Front. Big Data 4, 688969. https://doi.org/10.3389/fdata.2021.688969.

Bone, C., Simmonds-Buckley, M., Thwaites, R., Sandford, D., Merzhvynska, M., Rubel, J., Deisenhofer, A.K., Lutz, W., Delgadillo, J., 2021. Dynamic prediction of psychological treatment outcomes: development and validation of a prediction model using routinely collected symptom data. Lancet Digit. Health 3 (4), e231–e240. https://doi.org/10.1016/S2589-7500(21)00018-2.

Bouwmeester, W., Zuithoff, N.P.A., Mallett, S., Geerlings, M.I., Vergouwe, Y., Steyerberg, E.W., Altman, D.G., Moons, K.G.M., 2012. Reporting and methods in clinical prediction research: a systematic review. PLoS Med. 9 (5), Article 5. https://doi.org/10.1371/journal.pmed.1001221.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Buckman, J.E.J., Cohen, Z.D., O'Driscoll, C., Fried, E.I., Saunders, R., Ambler, G., DeRubeis, R.J., Gilbody, S., Hollon, S.D., Kendrick, T., Watkins, E., Eley, T.C., Peel, A.J., Rayner, C., Kessler, D., Wiles, N., Lewis, G., Pilling, S, 2021a. Predicting prognosis for adults with depression using individual symptom data: a comparison of modelling approaches. Psychol. Med. 1–11. https://doi.org/10.1017/S0033291721001616.

Buckman, J.E.J., Saunders, R., Cohen, Z.D., Barnett, P., Clarke, K., Ambler, G., DeRubeis, R.J., Gilbody, S., Hollon, S.D., Kendrick, T., Watkins, E., Wiles, N., Kessler, D., Richards, D., Sharp, D., Brabyn, S., Littlewood, E., Salisbury, C., White, I. R., Pilling, S., 2021b. The contribution of depressive 'disorder characteristics' to determinations of prognosis for adults with depression: an individual patient data meta-analysis. Psychol. Med. 51 (7), Article 7. https://doi.org/10.1017/S0033291721001367.

Burke, T.A., Ammerman, B.A., Jacobucci, R., 2019. The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: a systematic review. J. Affect. Disord. 245, 869–884. https://doi.org/10.1016/j.jad.2018.11.073.

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), Article 2. https://doi.org/10.1002/wps.20882.

Chekroud, A.M., Koutsouleris, N., 2018. The perilous path from publication to practice. Mol. Psychiatry 23 (1), Article 1. https://doi.org/10.1038/mp.2017.227.

Chipman, H.A., George, E.I., McCulloch, R.E., 2010. BART: bayesian additive regression trees. Ann. Appl. Stat. 4 (1), Article 1. https://doi.org/10.1214/09-AOAS285.

Clark, D.M., 2018. Realizing the mass public benefit of evidence-based psychological therapies: the IAPT program. Annu Rev. Clin. Psychol. 14 (1), Article 1. https://doi.org/10.1146/annurev-clinpsy-050817-084833.

Collins, G.S., Reitsma, J.B., Altman, D.G., Moons, K.G., 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC Med. 13 (1), 1. https://doi.org/10.1186/s12916-014-0241-z.

Cromarty, P., Drummond, A., Francis, T., Watson, J., Battersby, M., 2016. NewAccess for depression and anxiety: adapting the UK improving access to psychological therapies program across Australia. Australas. Psychiatry. https://doi.org/10.1177/1039856216641310 in press.

De Angel, V., Lewis, S., White, K., Oetzmann, C., Leightley, D., Oprea, E., Lavelle, G., Matcham, F., Pace, A., Mohr, D.C., Dobson, R., Hotopf, M., 2022. Digital health tools for the passive monitoring of depression: a systematic review of methods. NPJ Digit. Med. 5 (1), Article 1. https://doi.org/10.1038/s41746-021-00548-8.

Delgadillo, J., Jong, K.de, Lucock, M., Lutz, W., Rubel, J.A., Gilbody, S., Ali, S., Aguirre, E., Appleton, M., Nevin, J., O'Hayon, H., Patel, U., Sainty, A., Spencer, P., McMillan, D, 2018. Feedback-informed treatment versus usual psychological treatment for depression and anxiety: a multisite, open-label, cluster randomised controlled trial. Lancet Psychiatry. https://doi.org/10.1016/S2215-0366(18)30162-7.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., & Weingessel, A. (2009). E1071: misc functions of the department of statistics (E1071), TU Wien. In R package version 1.5–24 (Vol. 1).

Dwyer, D.B., Falkai, P., Koutsouleris, N., 2018. Machine learning approaches for clinical psychology and psychiatry. Annu Rev. Clin. Psychol. 14 (1), 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037.

Fernandes, B.S., Williams, L.M., Steiner, J., Leboyer, M., Carvalho, A.F., Berk, M., 2017. The new field of 'precision psychiatry. BMC Med. 15 (1), 80. https://doi.org/10.1186/s12916-017-0849-x.

Ferrari, A., Santomauro, D., Herrera, A., Shadid, J., Ashbaugh, C., Erskine, H., Charlson, F., Degenhardt, L., Scott, J., McGrath, J., Allebeck, P., Benjet, C., Breitborde, N., Brugha, T., Dai, X., Dandona, L., Dandona, R., Fischer, F., Haagsma, J., Whiteford, H., 2022. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry. https://doi.org/10.1016/S2215-0366(21)00395-3.

Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Stat. 19 (1), 1–67. https://doi.org/10.1214/aos/1176347963.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), 1–22.

Hayden, J.A., van der Windt, D.A., Cartwright, J.L., Côté, P., Bombardier, C., 2013. Assessing bias in studies of prognostic factors. Ann. Intern. Med. 158 (4), Article 4. https://doi.org/10.7326/0003-4819-158-4-201302190-00009.

Improving Access to Psychological Therapies. (2011). The IAPT data handbook guidance on recording and monitoring outcomes to support local evidence-based practice. Version 2.0. http://webarchive.nationalarchives.gov.uk/20160302160058/http://www.iapt.nhs.uk/silo/files/iapt-data-handbook-v2.pdf.

Khazanov, G.K., Xu, C., Dunn, B.D., Cohen, Z.D., DeRubeis, R.J., Hollon, S.D., 2020. Distress and anhedonia as predictors of depression treatment outcome: a secondary analysis of a randomized clinical trial. Behav. Res. Ther. 125, 103507 https://doi.org/10.1016/j.brat.2019.103507.

Knapstad, M., Nordgreen, T., Smith, O.R.F., 2018. Prompt mental health care, the Norwegian version of IAPT: clinical outcomes and predictors of change in a multicenter cohort study. BMC Psychiatry 18 (1), Article 1. https://doi.org/10.1186/s12888-018-1838-0.

Kroenke, K., Spitzer, R.L., Williams, J.B., 2001. The PHQ-9: validity of a brief depression severity measure. J. Gen. Intern. Med. 16 (9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x.

Loerinc, A.G., Meuret, A.E., Twohig, M.P., Rosenfield, D., Bluett, E.J., Craske, M.G., 2015. Response rates for CBT for anxiety disorders: need for standardized criteria. Clin. Psychol. Rev. 42, 72–82. https://doi.org/10.1016/j.cpr.2015.08.004.

Max, K., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28 https://doi.org/10.18637/jss.v028.i05.

Meehan, A.J., Lewis, S.J., Fazel, S., Fusar-Poli, P., Steyerberg, E.W., Stahl, D., Danese, A., 2022. Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. Mol. Psychiatry 27 (6), Article 6. https://doi.org/10.1038/s41380-022-01528-4.

Mundt, J.C., Marks, I.M., Shear, M.K., Greist, J.H., 2002. The work and social adjustment scale: a simple measure of impairment in functioning. Br. J. Psychiatry J. Ment. Sci. 180, 461–464. https://doi.org/10.1192/bjp.180.5.461.

National Collaborating Centre for Mental Health. (2023). The NHS Talking Therapies manual. https://www.england.nhs.uk/publication/the-improving-access-to-psychological-therapies-manual/.

National Institute for Health and Care Excellence. (2020). What treatments should I be offered for GAD? Clinical guideline [CG113]. https://www.nice.org.uk/guidance/cg113/ifp/chapter/what-treatments-should-i-be-offered-for-gad.

Obermeyer, Z., Emanuel, E.J., 2016. Predicting the future—big data, machine learning, and clinical medicine. N. Engl. J. Med. 375 (13), 1216–1219. https://doi.org/10.1056/NEJMp1606181.

O'Driscoll, C., Buckman, J.E.J., Fried, E.I., Saunders, R., Cohen, Z.D., Ambler, G., DeRubeis, R.J., Gilbody, S., Hollon, S.D., Kendrick, T., Kessler, D., Lewis, G.,

Watkins, E., Wiles, N., Pilling, S., 2021. The importance of transdiagnostic symptom level assessment to understanding prognosis for depressed adults: analysis of data from six randomised control trials. BMC Med. 19 (1), 109. https://doi.org/10.1186/s12916-021-01971-0.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org/.

Saunders, R., Cape, J., Leibowitz, J., Aguirre, E., Jena, R., Cirkovic, M., Wheatley, J., Main, N., Pilling, S., Buckman, J.E.J., 2020. Improvement in IAPT outcomes over time: are they driven by changes in clinical practice? Cogn. Behav. Therap. 13, e16. https://doi.org/10.1017/S1754470×20000173.

Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H., 2020. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 180–186. https://doi.org/10.1145/3375627.3375830.

Spitzer, R.L., Kroenke, K., Williams, J.B.W., Löwe, B., 2006. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch. Intern. Med. 166 (10), Article 10. https://doi.org/10.1001/archinte.166.10.1092.

Stekhoven, D.J., Bühlmann, P., 2012. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics 28 (1), 112–118. https://doi.org/10.1093/bioinformatics/btr597.

Stone, C.J., Hansen, M.H., Kooperberg, C., Truong, Y.K., 1997. Polynomial splines and their tensor products in extended linear modeling. Ann. Stat. 25 (4), 1371–1425.

van der Laan, M.J., Polley, E.C., Hubbard, A.E., 2007. Super learner. Stat. Appl. Genet. Mol. Biol. 6, Article25. https://doi.org/10.2202/1544-6115.1309.

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