Dynamic prediction of psychological treatment outcomes: development and validation of a prediction model using routinely collected symptom data

Claire Bone*, Melanie Simmonds-Buckley*, Richard Thwaites, David Sandford, Mariia Merzhvynska, Julian Rubel, Anne-Katharina Deisenhofer, Wolfgang Lutz, Jaime Delgadillo

Summary

Background Common mental disorders can be effectively treated with psychotherapy, but some patients do not respond well and require timely identification to prevent treatment failure. We aimed to develop and validate a dynamic model to predict psychological treatment outcomes, and to compare the model with currently used methods, including expected treatment response models and machine learning models.

Methods In this prediction model development and validation study, we obtained data from two UK studies including patients who had accessed therapy via Improving Access to Psychological Therapies (IAPT) services managed by ten UK National Health Service (NHS) Trusts between March, 2012, and June, 2018, to predict treatment outcomes. In study 1, we used data on patient-reported depression (Patient Health Questionnaire 9 [PHQ-9]) and anxiety (Generalised Anxiety Disorder 7 [GAD-7]) symptom measures obtained on a session-by-session basis (Leeds Community Healthcare NHS Trust dataset; n=2317) to train the Oracle dynamic prediction model using iterative logistic regression analysis. The outcome of interest was reliable and clinically significant improvement in depression (PHQ-9) and anxiety (GAD-7) symptoms. The predictive accuracy of the model was assessed in an external test sample (Cumbria Northumberland Tyne and Wear NHS Foundation Trust dataset; n=2036) using the area under the curve (AUC), positive predictive values (PPVs), and negative predictive values (NPVs). In study 2, we retrained the Oracle algorithm using a multiservice sample (South West Yorkshire Partnership NHS Foundation Trust, North East London NHS Foundation Trust, Cheshire and Wirral Partnership NHS Foundation Trust, and Cambridgeshire and Peterborough NHS Foundation Trust; n=42 992) and compared its performance with an expected treatment response model and five machine learning models (Bayesian updating algorithm, elastic net regularisation, extreme gradient boosting, support vector machine, and neural networks based on a multilayer perceptron algorithm) in an external test sample (Whittington Health NHS Trust; Barnet Enfield and Haringey Mental Health Trust; Pennine Care NHS Foundation Trust; and Humber NHS Foundation Trust; n=30 026).

Findings The Oracle algorithm trained using iterative logistic regressions generalised well to external test samples, explaining up to 47·3% of variability in treatment outcomes. Prediction accuracy was modest at session one (AUC 0·59 [95% CI 0·55–0·62], PPV 0·63, NPV 0·61), but improved over time, reaching high prediction accuracy (AUC 0·81 [0·77–0·86], PPV 0·79, NPV 0·69) as early as session seven. The performance of the Oracle model was similar to complex (eg, including patient profiling variables) and computationally intensive machine learning models (eg, neural networks based on a multilayer perceptron algorithm, extreme gradient boosting). Furthermore, the predictive accuracy of a more simple dynamic algorithm including only baseline and index-session scores was comparable to more complex algorithms that included additional predictors modelling sample-level and individual-level variability. Overall, the Oracle algorithm significantly outperformed the expected treatment response model (mean AUC 0·80 *vs* 0·70, p<0·0001]).

Interpretation Dynamic prediction models using sparse and readily available symptom measures are capable of predicting psychotherapy outcomes with high accuracy.

Funding University of Sheffield.

Lancet Digit Health 2021; 3: e231–40

See **Comment** page e207

*Joint first authors

Department of Psychology, Clinical Psychology Unit, University of Sheffield, Sheffield, UK (C Bone DClinPsy, M Simmonds-Buckley PhD, J Delgadillo PhD); **Cumbria, Northumberland, Tyne and Wear NHS Foundation Trust, Penrith, UK** (R Thwaites DClinPsy); **Lancashire and South Cumbria NHS Foundation Trust, Preston, UK** (D Sandford MSc); **Department of Psychology, University of Zurich, Zurich, Switzerland** (M Merzhvynska MSc); **Department of Psychology, Justus Liebig University Giessen, Giessen, Germany** (J Rubel PhD); **Department of Psychology, University of Trier, Trier, Germany** (A-K Deisenhofer PhD, Prof W Lutz PhD)

Correspondence to: Dr Jaime Delgadillo, Department of Psychology, Clinical Psychology Unit, University of Sheffield, Sheffield S1 2LT, UK **jaime.delgadillo@nhs.net**

Introduction

Precision mental health care is an emerging field that employs data-driven methods to monitor patients' treatment response, to model their prognosis, and to personalise their treatment accordingly.[1] Two types of data-driven methods include expected treatment response models and patient profiling models.

The expected treatment response concept was first introduced by Lutz and colleagues[2] and further developed by Finch and colleagues[3] into a practical

Research in context

Evidence before this study

We searched PubMed and PsycINFO from inception to Jan 8, 2021, for meta-analyses of feedback-informed treatment studies in the field of psychotherapy using the search terms “therapy”, “feedback”, and “routine outcome monitoring”, with no language restrictions. We also hand-searched the reference lists of eligible studies. Overall, eight meta-analyses, including more than 20 controlled studies, were published in peer-reviewed journals (excluding unpublished studies and educational theses). According to these meta-analyses, the evidence was generally of moderate quality. Findings indicated that treatment outcomes can be improved when therapists use clinical prediction and feedback models, with small effect sizes in general clinical samples (Cohen’s *d* 0·07–0·28) and small to moderate effects in patients classified as not on track (Cohen’s *d* 0·21–0·53). Such models enable the timely identification of patients who have a poor expected prognosis, enabling therapists to resolve problems. Contemporary systems use expected treatment response models to provide feedback to therapists about their patients’ treatment response.

Added value of this study

This study demonstrates that clinical prediction models can generalise with high accuracy to multiple psychological services, in different geographical regions, with different therapists. Furthermore, the dynamic prediction model was significantly more accurate than the expected treatment response model, which is the current methodological standard in the field of psychotherapy.

Implications of all the available evidence

The development of dynamic clinical prediction models represents an important methodological advance in the field of precision mental health care because it improves the accuracy and potential clinical utility of such models.

clinical decision making tool. To develop expected treatment response models, repeated patient-reported outcome measures from large clinical samples are analysed using growth curve modelling. Cases are clustered into subgroups who have the same baseline severity in the relevant outcome measure, and a growth curve is plotted following the curvilinear dose-response effect observed by Howard and colleagues.[4] Finch and colleagues[3] proposed modelling 80% prediction intervals around the expected treatment response curve to enable clinicians to assess if a patient’s response to treatment is on track or not on track by comparison to clinical norms. Although expected treatment response models offer a practical method of monitoring treatment response, response patterns for individuals can substantially deviate from aggregated group trends.[5]

It has long been recognised that clinical populations are often heterogeneous in terms of their psychometric, demographic, and interpersonal characteristics, which might partly explain why some patients respond better or more quickly than others.[6] Patient profiling focuses on understanding the patient’s unique combination of characteristics and their association with clinical outcomes. Several patient profiling methods have been applied in psychotherapy, including case-mix adjusted expected treatment response models,[2] *k*-nearest neighbours analysis,[7] latent profile analysis,[8] and risk stratification approaches.[9,10] Despite their methodological differences, these studies converge in the observation that different subgroups of patients respond differently to treatment.

To date, expected treatment response models are considered the methodological standard in psychotherapy services that apply data-driven models to guide clinical decision making. Expected treatment response models have been integrated into a number of computerised feedback systems in different countries.[11–14] Such systems have been tested in clinical trials whereby therapists receive computerised feedback about their patients’ treatment response, with a particular focus on patients who are not on track and might require some personalised adjustments to their treatment plan. Meta-analyses of these trials have indicated that using feedback systems helps to improve treatment outcomes,[15,16] especially for patients classified as not on track.[16]

Despite their proven effects, expected treatment response-based feedback models have some limitations. First, they model expected symptom trajectories that are adjusted for baseline severity, but not other relevant patient characteristics.[15] Some researchers have argued that this is a parsimonious approach to generate clinical norms,[17] but others propose that combining patient profiling and expected treatment response methods could enhance predictive accuracy.[7] Second, conventional expected treatment response models use data smoothing techniques (ie, growth curves), which result in some loss of information that might be contained in raw time-series data that typically show non-linear patterns of change. Third, expected treatment response models are designed to identify an atypical and relatively small subsample of patients at risk of elevated symptoms (around 10%),[3] but they are not sensitive to the large proportion of patients who do not attain clinically significant improvement after therapy (around 40–50%). Fourth, expected treatment response curves are so-called fixed predictions calculated after the initial assessment, which do not change over the course of treatment. This approach places considerable emphasis on a single baseline severity measure, which could be influenced by measurement error. So-called dynamic expected treatment response models could be designed to have the capability of learning from new incoming data collected during treatment.[14] Furthermore, the prognostic accuracy of such models could potentially be enhanced using modern machine learning techniques such as regularisation, resampling, and cross-validation.[10]

On the basis of the available literature, we developed and tested the generalisability of a dynamic progress feedback model using data from two linked studies, with the aim of overcoming some of the limitations of previous expected treatment response systems.

Methods

Study design and data sources

For this prediction model development and validation study, we obtained data from two linked studies of adult patients (aged ≥16 years) with common mental disorders who had accessed therapy via Improving Access to Psychological Therapies (IAPT) services managed by UK National Health Service (NHS) Trusts. In study 1, fully anonymised clinical data were collected by the IAPT services in two NHS Trusts in the north of England: Leeds Community Healthcare NHS Trust and Cumbria Northumberland Tyne and Wear NHS Foundation Trust. Data from Leeds Community Healthcare NHS Trust were collected between March, 2012, and December, 2014, and data from Cumbria Northumberland Tyne and Wear NHS Foundation Trust were collected between April, 2017, and June, 2018. In study 2, data were collected contemporaneously by IAPT services in eight NHS Trusts covering diverse regions of England: South West Yorkshire Partnership NHS Foundation Trust, North East London NHS Foundation Trust, Whittington Health NHS Trust, Barnet Enfield and Haringey Mental Health Trust, Pennine Care NHS Foundation Trust, Cheshire and Wirral Partnership NHS Foundation Trust, Cambridgeshire and Peterborough NHS Foundation Trust, and Humber NHS Foundation Trust. These IAPT services opted-in to participate in the study, which was promoted via email to IAPT services in all regions of England between January, 2014 and May, 2017. Together, the participating NHS Trusts across both studies managed 25 teams that were part of the national IAPT programme. Datasets collected in both studies included anonymised electronic health records for adolescents (aged ≥16 years) and adults accessing psychological treatments for common mental disorders across participating IAPT services.

Data collection for study 1 was approved by the West Midlands—Coventry and Warwickshire Research Ethics Committee (ref 18/WM/0012) and data collection for study 2 was approved by the London—City and East NHS Research Ethics Committee (ref 15/LO/2200). The need for written informed consent was waived because patient-level data were fully anonymised.

Procedures

IAPT is a treatment system that delivers psychological interventions for depression and anxiety disorders organised in a stepped care model.[18] Many patients initially access low intensity guided self-help and later have the option to access high intensity psychological therapies if their symptoms persist. Low intensity interventions are based on principles of cognitive behavioural therapy, and involve learning coping skills with the support of a qualified practitioner for up to eight sessions. High intensity interventions are psychotherapies of longer duration (up to 20 sessions), including cognitive behavioural therapy, interpersonal psychotherapy, person-centred counselling, and other empirically-supported treatments. These interventions are delivered by practitioners qualified to a postgraduate level, following structured treatment protocols endorsed by national guidelines,[19,20] and under regular supervision from experienced therapists.

We collected patient-reported depression (Patient Health Questionnaire 9 [PHQ-9]) and anxiety (Generalised Anxiety Disorder 7 [GAD-7]) symptom measures on a session-by-session basis to monitor treatment progress. The PHQ-9 is a nine-item measure of depression symptoms, in which each item is rated on a 0–3 Likert scale, yielding an overall depression severity score between 0 and 27.[21] A cutoff score of 10 or higher is recommended to screen for clinically significant depression symptoms, with adequate sensitivity (88%) and specificity (88%), and a difference of 6 points or more between assessments is indicative of statistically reliable change.[22] The GAD-7 questionnaire is a seven-item measure used to identify anxiety disorders; each item is also rated on a 0–3 Likert scale, yielding a total anxiety severity score between 0 and 21.[23] A cutoff score of 8 or higher is recommended to identify clinically significant anxiety symptoms, with adequate sensitivity (77%) and specificity (82%),[23] and a difference of 5 points or more is indicative of statistically reliable change.[22]

Treatment response was defined according to Jacobson and Truax's concept of reliable and clinically significant improvement,[24] which requires post-treatment scores to be below the relevant cutoff score for each outcome measure, and to have improved by a magnitude greater or equal to the reliable change index relative to the baseline measure. Separate prediction models were calculated for depression (PHQ-9) and anxiety (GAD-7) outcome measures. This approach supported the development of feedback models that prioritise the attainment of full remission of symptoms, and which provide therapists with highly specific feedback about each symptom domain (depression or anxiety).

The datasets from study 1 included Leeds Risk Index scores for each patient. The Leeds Risk Index is a patient profiling method that classifies patients into those with low, moderate, or high risk of treatment failure.[9] The Leeds Risk Index determines the patient's profile using combined weights from six features (age, employment, disability, depression severity, functional impairment, and initial outcome expectancy).

Descriptive demographic (age, gender, ethnicity, employment) and clinical (primary diagnosis recorded in clinical records) data were obtained for participants in both studies.

Model cross-validation strategy

Research on trajectories of change in psychotherapy indicates that at least three timepoints are necessary to model non-linear trends, and an additional timepoint is necessary to ensure that treatment outcomes (ie, dependent variable) are not confounded with predictors (ie, independent variables) in time-series analysis.[25] On this basis, we only included patients who had pretreatment symptoms above the cutoff score for one or both measures (PHQ-9, GAD-7), and who attended at least four treatment sessions. Thus, data for 4353 patients from study 1 and 73 018 patients from study 2 were used to develop our model. Overall, the sample of patients included in studies 1 and 2 represents 66% of all patients who accessed therapy in IAPT services managed by participating NHS Trusts during the data collection period.

We used the dataset from Leeds Community Healthcare NHS Trust (n=2317) from study 1 to train prediction models and the dataset from Cumbria Northumberland Tyne and Wear NHS Foundation Trust (n=2036) was used as an external test sample. The full dataset from study 2 was split into a training sample (n=42 992; South West Yorkshire Partnership NHS Foundation Trust, North East London NHS Foundation Trust, Cheshire and Wirral Partnership NHS Foundation Trust, and Cambridgeshire and Peterborough NHS Foundation Trust) and test sample (n=30 026; Whittington Health NHS Trust, Barnet Enfield and Haringey Mental Health Trust, Pennine Care NHS Foundation Trust, and Humber NHS Foundation Trust).

Development and validation of dynamic prediction models

We used the study 1 datasets to develop a parsimonious prediction model based solely on session-by-session outcome measures. The predictive accuracy of the model was also compared with a model that included patient profiling data.

We developed separate prediction models for depression (PHQ-9) and anxiety (GAD-7) outcomes, only using data from the training sample (n=2317). Session-by-session outcome data were entered into an iterative logistic regression analysis. The output of the prediction model was a predicted probability (range 0–100%) of attaining reliable and clinically significant improvement by the end of therapy. The model was repeatedly retrained by the addition of information collected at each sequential therapy session, which recalibrated the output (predicted probability) over time. Since treatment duration varies considerably across patients, we trained each subsequent regression model using smaller subsamples of patients who remained in treatment at each modelling step. For example, the regression model for session six included all patients who attended at least seven sessions, so that predictive information collected up to session six was not confounded with the post-treatment outcome that was measured at session seven or later. Previous research indicates that IAPT patients who attend a similar number of treatment sessions have similar outcomes;[25] therefore, this modelling strategy enabled the leverage of prognostic information from similar patients clustered by treatment duration in a way that reflects the fact that therapists do not know in advance exactly how many sessions each patient will attend.

Four predictors were entered into each regression model: a baseline outcome measure (eg, PHQ-9 at session one), the measure recorded at the index session (eg, PHQ-9 at session five for the fifth model), a risk sum, and a within-person SD. The risk sum was calculated by comparing the patient's current outcome measure with that of the sample mean for that session. If the patient's measure was 1 SD higher than the sample mean (ie, more severe than expected), the risk sum increased by 1 point. If the patient's symptoms were consistently more severe than the sample mean, the risk sum increased cumulatively over time. The SD was calculated from each patient's session-by-session outcome measures, representing their degree of variability in change over time. Patients with extreme symptom fluctuations had a higher SD than did patients with less intense symptom fluctuations. Thus, the prediction model learned from sample-level and patient-level changes over time. Regression models started by entering all four predictors, and non-significant predictors ($p>0\cdot05$) were removed using backward elimination at each new iteration, to attain the most simple and most parsimonious model for each therapy session.

We did an a priori sample size calculation using Hsieh's criteria[26] for logistic regression and effect size estimates from previous literature,[25] which indicated a minimum requirement of 132 patients for each modelling step (ie, for each subsample selected over the iterative models from session-to-session). Therefore, we applied the dynamic regression modelling approach up to session seven for low intensity interventions, and up to session 12 for high intensity interventions, at which point sample sizes reduced below the minimum sample size requirement. Separate models for low and high intensity treatments were estimated to meet the assumption of independent observations (because some patients accessed both treatments), and previous research indicates different dose–response patterns for these types of interventions.[27] This strategy enabled us to train a dynamic prediction model using iterative logistic regressions, referred to as the Oracle algorithm hereafter, and which is mainly characterised by the inclusion of the four predictors described.

We also developed a second version of the model, referred to as Oracle2 hereafter. A baseline model included patient-profiling information (Leeds Risk Index classification scores) from a pretreatment assessment and symptom severity data from treatment session one. Subsequent models were retrained by additionally entering Leeds Risk Index classification scores for each patient as a predictor at each modelling step, so that the output was adjusted for different patient profiles. Once both versions of the algorithm were developed in the training sample, we applied them in the external test sample. The out-of-sample performance of these models was examined using conventional indices of explained variance (Nagelkerke's pseudo R^2 from logistic

	Study 1		**Study 2**	
	Training sample (n=2317)	Test sample (n=2036)	Training sample (n=42 992)	Test sample (n=30 026)
Demographic characteristics				
Age, years	36·59 (13·39)	41·19 (15·43)	40·01 (14·47)	40·68 (14·51)
Sex				
Male	764 (33·0%)	765 (37·6%)	14 692 (34·2%)	10 266 (34·2%)
Female	1553 (67·0%)	1271 (62·4%)	28 300 (65·8%)	19 760 (65·8%)
Unemployed	414 (17·9%)	195 (9·6%)	9886 (23·0%)	7629 (25·4%)
Ethnicity				
White British	1933/2228 (86·8%)	1962 (96·4%)	34 838/40 320 (86·4%)	23 144/27 683 (83·6%)
Other	295/2228 (13·2%)	74 (3·6%)	5482/40 320 (13·6%)	4539/27 683 (16·4%)
Clinical characteristics				
Baseline PHQ-9 score	15·38 (5·81)	15·77 (5·27)	16·10 (5·62)	16·35 (5·70)
Baseline GAD-7 score	14·22 (4·44)	14·75 (4·28)	14·60 (4·38)	14·68 (4·42)
Leeds Risk Index				
Low risk	714 (30·8%)	457 (22·4%)	NA	NA
Moderate risk	1161 (50·1%)	1156 (56·8%)	NA	NA
High risk	442 (19·1%)	423 (20·8%)	NA	NA
Primary diagnosis				
Affective disorder*	520/2196 (23·7%)	711/1898 (37·5%)	16 677/39 157 (42·6%)	7822/25 509 (30·7%)
Generalised anxiety disorder	229/2196 (10·4%)	735/1898 (38·7%)	5820/39 157 (14·9%)	3323/25 509 (13·0%)
Mixed depression and anxiety disorder	787/2196 (35·8%)	18/1898 (0·9%)	8973/39 157 (22·9%)	9583/25 509 (37·6%)
Panic disorder with or without agoraphobia	102/2196 (4·6%)	93/1898 (4·9%)	1049/39 157 (2·7%)	1089/25 509 (4·3%)
Social anxiety disorder	59/2196 (2·7%)	0	724/39 157 (1·8%)	650/25 509 (2·5%)
Specific phobia	18/2196 (0·8%)	16/1898 (0·8%)	299/39 157 (0·8%)	209/25 509 (0·8%)
Obsessive-compulsive disorder	65/2196 (3·0%)	50/1898 (2·6%)	1043/39 157 (2·7%)	632/25 509 (2·5%)
Post-traumatic stress disorder	42/2196 (1·9%)	120/1898 (6·3%)	1280/39 157 (3·3%)	1186/25 509 (4·6%)
Other	374/2196 (17·0%)	155/1898 (8·2%)	3292/39 157 (8·4%)	1015/25 509 (4·0%)
Treatment pathway				
Completed treatment after low intensity treatment	1293 (55·8%)	875 (43·0%)	17 283 (40·2%)	11 440 (38·1%)
Completed treatment after high intensity treatment	1024 (44·2%)	1161 (57·0%)	25 709 (59·8%)	18 586 (61·9%)
Treatment sessions	12 (6·5)	8 (4·4)	10·39 (4·48)	9·26 (4·23)
RSCI after low intensity treatment				
PHQ-9	1205 (52·0%)	1405 (69·0%)	23 216 (54·0%)	15 313 (51·0%)
GAD-7	1251 (54·0%)	1405 (69·0%)	23 216 (54·0%)	15 313 (51·0%)
RCSI after high intensity treatment				
PHQ-9	1321 (57·0%)	1242 (61·0%)	21 496 (50·0%)	14 412 (48·0%)
GAD-7	1321 (57·0%)	1201 (59·0%)	21 496 (50·0%)	14 412 (48·0%)

Data are n (%), mean (SD), or n/N (%). Percentages were calculated excluding cases with missing data. Sample characteristics by low intensity and high intensity treatment are available in the appendix (pp 1–2). PHQ-9=Patient Health Questionnaire 9. GAD-7=Generalised Anxiety Disorder 7. NA=not available. RCSI=reliable and clinically significant improvement after treatment. *Includes major depressive disorder, recurrent depression, and dysthymia; around 85% of patients in this category presented with major depressive disorder.

Table 1: **Sample characteristics**

Session	Test sample, n	Overall sessions (SD)*	Oracle (R^2 0·031–0·382†)			Oracle2 (R^2 0·054–0·382†)		
			AUC (95% CI)	PPV	NPV	AUC (95% CI)	PPV	NPV
Session 1	1040	8·20 (3·92)	0·585 (0·548–0·622)	0·630	0·607	0·614 (0·578–0·650)	0·660	0·554
Session 2	1000	8·14 (3·90)	0·669 (0·635–0·702)	0·668	0·581	0·678 (0·645–0·711)	0·668	0·585
Session 3	997	8·21 (3·93)	0·737 (0·706–0·768)	0·711	0·608	0·739 (0·709–0·770)	0·706	0·604
Session 4	853	8·94 (3·80)	0·769 (0·736–0·801)	0·756	0·646	0·769 (0·736–0·802)	0·750	0·627
Session 5	716	9·71 (3·67)	0·780 (0·744–0·816)	0·772	0·670	0·783 (0·748–0·819)	0·771	0·672
Session 6	589	10·52 (3·63)	0·788 (0·748–0·827)	0·780	0·647	0·783 (0·744–0·823)	0·770	0·636
Session 7	479	11·33 (3·57)	0·814 (0·773–0·855)	0·788	0·690	0·814 (0·773–0·855)	0·788	0·690
Session 8	356	12·37 (3·51)	0·826 (0·779–0·873)	0·809	0·708	0·826 (0·779–0·873)	0·809	0·708
Session 9	283	13·27 (3·39)	0·811 (0·760–0·862)	0·783	0·676	0·811 (0·760–0·862)	0·783	0·676
Session 10	220	14·29 (3·30)	0·830 (0·775–0·884)	0·790	0·667	0·830 (0·775–0·884)	0·790	0·667
Session 11	176	15·01 (3·12)	0·816 (0·754–0·877)	0·764	0·660	0·816 (0·754–0·877)	0·764	0·660
Session 12	125	16·22 (2·94)	0·824 (0·746–0·902)	0·795	0·730	0·824 (0·746–0·902)	0·795	0·730

The outcome of interest was reliable and clinically significant improvement in depression after treatment, measured using the Patient Health Questionnaire 9. AUC=area under the curve. PPV=positive predictive value. NPV=negative predictive value. *Mean number of total therapy sessions attended by the subsample of patients used to train each of the prediction models. †Proportion of variance explained for the target disorder (estimated by Nagelkerke's pseudo R^2).

Table 2: **Out-of-sample performance indices for the Oracle and Oracle2 dynamic prediction models for depression treatment outcome in study 1**

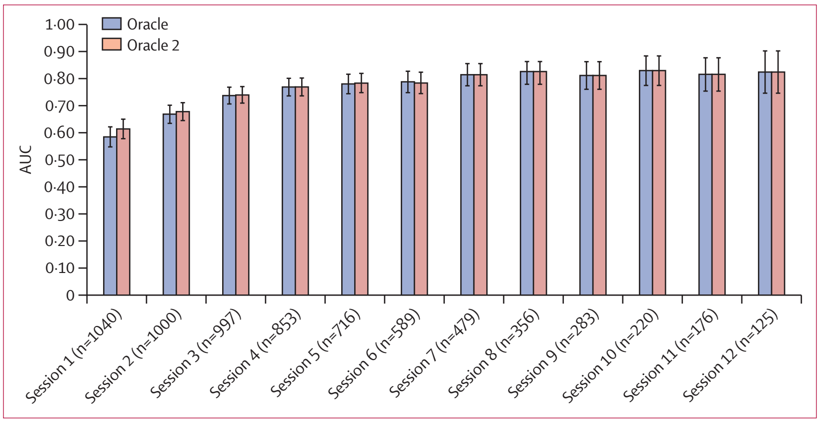

Figure 1**: Out-of-sample predictive accuracy of Oracle and Oracle2 models over time in study 1**
Error bars show 95% CIs. AUC=area under the curve.

See Online for appendix

regression[28]) and predictive accuracy (area under the curve [AUC], positive predictive values [PPVs], and negative predictive values [NPVs]). The AUC was the primary index of performance accuracy. We interpreted values of 0·70 and higher to indicate fair accuracy and values of 0·80 and higher to indicate optimal accuracy, according to conventional standards in clinical medicine.[29] PPV and NPV help to assess the extent to which the model can correctly classify cases that do and do not attain reliable and clinically significant improvement of symptoms after treatment. We also developed two simpler models: a baseline model that only used baseline symptom severity as the sole predictor and a model that only used baseline severity and the symptom score of each index therapy session.

We used the data from study 2 to compare the prediction accuracy of the Oracle algorithm with alternative modelling approaches. As a first step, we retrained the Oracle algorithm using a multiservice training sample to enhance generalisability. Next, using the same training sample, we developed six alternative models (Bayesian updating algorithm, elastic net regularisation, extreme gradient boosting, support vector machine, neural networks, and expected treatment response model; appendix pp 11–13). All modelling approaches used the same four predictors (baseline symptom score, symptom score at the index session, sample-based risk sum, and within-person SD), and followed the same sample selection and dynamic iterative modelling pipeline described in study 1.

The second model was similar to Oracle, but applied a Bayesian updating framework,[30] in which the predicted probability from model k (a posterior probability) was entered as an additional predictor (now a prior probability) in model $k+1$ along the dynamic modelling pipeline. Four additional models performed variable selection, variable weighting (eg, the magnitude assigned to each regression term), and computation of predicted probabilities using different machine learning approaches that included: elastic net regularisation;[31] extreme gradient boosting;[32] support vector machines with calibrated posterior probabilities based on Platt's method;[33] and a multilayer perceptron algorithm,[34] which is a feed-forward, supervised neural network approach. Further details about the machine learning models are included in the appendix (pp 11–13).

We also developed a conventional expected treatment response model, based on the method proposed by Finch and colleagues,[3] and adapted for outcome measures used in IAPT services.[13] This method was based on clustering patients according to baseline severity subgroups in each outcome measure (PHQ-9, GAD-7), applying growth curve modelling using a log-linear trend, and computing

	Test sample (n)	Logistic regression (R^2 0·037–0·473*)			Bayesian updating algorithm (R^2 0·037–0·472*)			Elastic net regularisation (R^2 0·037–0·473*)			Extreme gradient boosting (R^2 0·063–0·490*)			Support vector machine (R^2 0·053–0·477*)			NeuralNet (R^2 0·052–0·513*)			Expected treatment response model (R^2 0·000–0·313*)		
		AUC†	PPV	NPV	AUC†	PPV	NPV	AUC†	PPV	NPV	AUC†	PPV	NPV	AUC†	PPV	NPV	AUC†	PPV	NPV	AUC†	PPV	NPV
Session 1	16 641	0·596	0·563	0·600	0·596	0·563	0·601	0·596	0·563	0·601	0·623	0·577	0·600	0·599	0·559	0·621	0·609	0·539	0·656	NA‡	NA‡	NA‡
Session 2	15 922	0·679	0·612	0·642	0·679	0·612	0·642	0·679	0·612	0·642	0·691	0·639,	0·646	0·682	0·614	0·655	0·688	0·626	0·653	0·599	0·646	0·573
Session 3	15 943	0·726	0·649	0·680	0·729	0·652	0·678	0·726	0·649	0·680	0·735	0·669	0·675	0·727	0·662	0·673	0·733	0·673	0·671	0·651	0·641	0·630
Session 4	14 399	0·757	0·671	0·701	0·760	0·677	0·701	0·757	0·672	0·701	0·764	0·695	0·700	0·761	0·696	0·696	0·757	0·693	0·688	0·685	0·612	0·685
Session 5	12 729	0·775	0·683	0·717	0·780	0·688	0·719	0·775	0·683	0·718	0·783	0·722	0·701	0·780	0·706	0·711	0·786	0·705	0·716	0·706	0·589	0·728
Session 6	10 815	0·789	0·691	0·738	0·792	0·699	0·733	0·789	0·691	0·739	0·800	0·719	0·729	0·793	0·712	0·723	0·792	0·711	0·725	0·724	0·568	0·763
Session 7	8823	0·801	0·695	0·748	0·803	0·700	0·739	0·801	0·695	0·748	0·811	0·720	0·747	0·804	0·723	0·736	0·800	0·729	0·723	0·744	0·552	0·810
Session 8	7016	0·811	0·710	0·761	0·814	0·714	0·749	0·811	0·711	0·761	0·820	0·744	0·748	0·816	0·736	0·741	0·816	0·719	0·762	0·752	0·526	0·828
Session 9	5681	0·815	0·710	0·758	0·819	0·720	0·754	0·815	0·710	0·758	0·819	0·733	0·745	0·818	0·731	0·742	0·815	0·736	0·731	0·762	0·516	0·845
Session 10	4543	0·818	0·719	0·760	0·822	0·723	0·750	0·818	0·720	0·761	0·823	0·740	0·746	0·822	0·739	0·744	0·821	0·759	0·728	0·776	0·510	0·875
Session 11	3622	0·828	0·725	0·774	0·831	0·737	0·766	0·828	0·726	0·774	0·833	0·758	0·758	0·830	0·746	0·755	0·825	0·752	0·735	0·782	0·503	0·875
Session 12	2724	0·838	0·733	0·795	0·840	0·742	0·784	0·838	0·734	0·796	0·842	0·766	0·769	0·841	0·755	0·781	0·836	0·739	0·796	0·776	0·501	0·870
Session 13	2080	0·843	0·728	0·796	0·849	0·746	0·804	0·843	0·726	0·801	0·850	0·760	0·787	0·845	0·746	0·784	0·841	0·749	0·784	0·779	0·507	0·871
Session 14	1542	0·841	0·701	0·796	0·845	0·714	0·791	0·841	0·702	0·797	0·835	0·742	0·775	0·840	0·717	0·789	0·844	0·711	0·797	0·785	0·501	0·881
Session 15	1136	0·841	0·725	0·806	0·846	0·716	0·796	0·841	0·725	0·806	0·837	0·728	0·770	0·840	0·734	0·798	0·840	0·715	0·785	0·786	0·501	0·879
Session 16	855	0·841	0·715	0·801	0·851	0·726	0·796	0·841	0·715	0·801	0·841	0·751	0·764	0·842	0·722	0·800	0·841	0·725	0·795	0·770	0·506	0·900
Session 17	661	0·860	0·713	0·807	0·860	0·726	0·797	0·860	0·714	0·807	0·867	0·741	0·806	0·861	0·716	0·804	0·874	0·769	0·787	0·766	0·515	0·888
Session 18	494	0·853	0·693	0·828	0·856	0·706	0·803	0·853	0·692	0·825	0·842	0·684	0·817	0·854	0·711	0·823	0·851	0·726	0·810	0·759	0·532	0·885
Session 19	382	0·857	0·700	0·816	0·861	0·690	0·827	0·856	0·700	0·816	0·797	0·687	0·790	0·860	0·709	0·811	0·853	0·688	0·813	0·749	0·535	0·888
Session 20	280	0·851	0·676	0·826	0·846	0·694	0·821	0·851	0·677	0·827	0·794	0·661	0·804	0·855	0·688	0·819	0·847	0·630	0·836	0·738	0·519	0·900

The outcome of interest was post-treatment reliable and clinically significant improvement in depression, measured using the Patient Health Questionnaire (PHQ-9). NeuralNet=neural networks based on a multilayer perceptron algorithm. AUC=area under the curve. PPV=positive predictive value. NPV=negative predictive value. NA=not applicable. *Proportion of variance explained for the target disorder (estimated by Nagelkerke's pseudo R^2). †95% CIs for AUC values are shown in the appendix (p 7). ‡The expected treatment response model starts to calculate predictions after session 1.

Table 3: **Out-of-sample performance indices for alternative prediction models for depression treatment outcome in study 2**

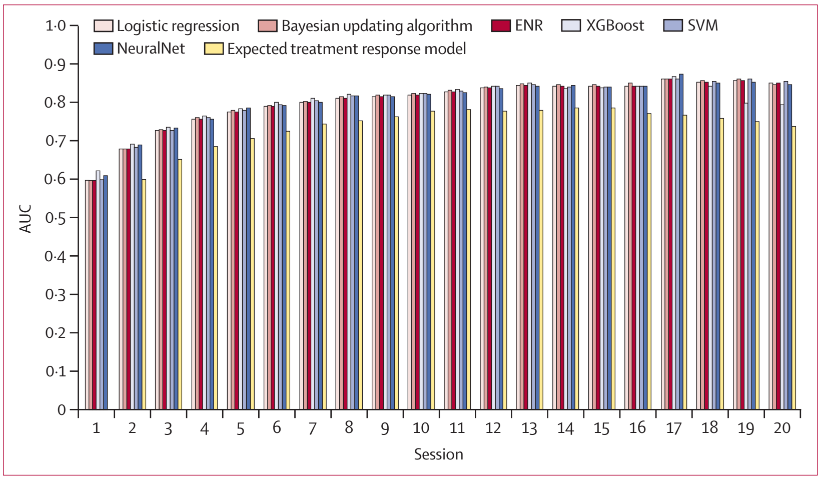

Figure 2: **Out-of-sample predictive accuracy indices for prediction models in study 2**
95% CIs for AUC values are shown in the appendix (p 7). AUC=area under the curve. ENR=elastic net regularisation. XGBoost=extreme gradient boosting. SVM=support vector machines. NeuralNet=neural networks based on a multilayer perceptron algorithm.

95% CIs to classify patients as on track or not on track. Once all models were developed in the training sample, their performance was assessed in an external test sample using the same indices as study 1. All analyses were done using SPSS (version 26) and IBM Modeler (version 18.2.1).

Role of the funding source

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

Results

A summary of baseline and clinical characteristics of the patients included in datasets is provided in table 1. The proportion of patients who achieved reliable and clinically significant improvement and socioeconomic features were balanced across the training and test samples of both studies. For simplicity, here we describe the results of the depression (PHQ-9) dynamic prediction models for high intensity psychotherapies. Results for the anxiety (GAD-7) models, and those for low intensity interventions are in the appendix (pp 3–10).

In study 1, the predictive accuracy of the Oracle algorithm was modest for session one (AUC 0·59 [95% CI 0·55–0·62], PPV 0·63, NPV 0·61) but improved over time, reaching clinically useful accuracy as early as session three (AUC 0·74, PPV 0·71, NPV 0·61), and high accuracy (AUC 0·81, PPV 0·79, NPV 0·69) from session seven onwards (table 2). Similarly, PPV and NPV improved over time, but PPV was generally higher than NPV (table 2). Oracle explained up to 38·2% of variability in treatment outcomes. The performance indices for Oracle2 were very similar. No significant differences in predictive accuracy were identified between the Oracle algorithm and Oracle2, as shown by the overlap in 95% CIs (figure 1).

A simpler model that only used baseline severity had poorer performance indices, never exceeding an AUC of 0·65, and the addition of a session score as an additional predictor improved the AUC to an accuracy level comparable with the Oracle algorithm, but with marginally lower explained variance (R^2 0·031–0·379; appendix p 5). The same findings were observed for anxiety treatment outcomes (appendix p 3), and for low intensity treatment prediction models (appendix p 4).

In study 2, the predictive accuracy of the Oracle algorithm trained using logistic regression improved over time, with modest accuracy observed at session one (AUC 0·60, PPV 0·56, NPV 0·60), which reached an optimal threshold (AUC ≥0·80) by session seven (AUC 0·80, PPV 0·70, NPV 0·75; table 3). PPV and NPV improved over time, and NPV was generally higher than PPV (table 3). The Oracle algorithm explained up to 47·3% of variability in treatment outcomes. The performance indices for all alternative models were very similar, with the exception of the expected treatment response model, which was consistently less accurate (figure 2). The expected treatment response model had high NPV indices (>0·80 by session seven), but low PPV (around 0·50) relative to the other models, and the lowest explained variance (upper bound of 31·3%). The results also showed that the performance of the extreme gradient boosting model marginally outperformed other models up to session six, with very similar performance to the other models observed up until session 17, after which its performance was lower than other models, but better than the expected treatment response model.

The same pattern of results was obtained for anxiety treatment outcomes (appendix p 8), and for low intensity treatment prediction models (appendix pp 9–10).

Discussion

The results of this prediction model development and validation study demonstrate that it is possible to predict psychological treatment outcomes with high accuracy, using routinely available outcome measures collected on a session-by-session basis. Our results indicate that a dynamic clinical prediction model generalises to multiple services in different geographical regions, with different therapists, and future cases relative to the time at which patients included in the training sample were treated. Prediction accuracy was modest for the earliest sessions of therapy, but reached a fair level (AUC ≥0·70) as early as session three and an optimal level (AUC ≥0·80) by session seven. The explained variance of the model (but not predictive accuracy) was marginally improved by the inclusion of sample-level (risk sum) and patient-level (within-subject SD) information, beyond baseline severity and session-by-session symptom scores. However, supplementing the prediction model with more detailed patient profiling data (Leeds Risk Index) did not significantly improve explained variance or prediction accuracy. This finding suggests that simple, easy to collect, session-by-session symptom scores can yield clinically useful prognostic information, which minimises problems resulting from missing data and the burden, cost, and complexity associated with multivariable models.

The results of our study demonstrate that dynamic prediction models outperform the expected treatment response models, which are considered the methodological standard in the field of feedback-informed treatment at present. The difference in predictive accuracy (AUC) for the first seven sessions between the expected treatment response model and all other dynamic models was 0·07–0·09. In this regard, the development of dynamic prediction models is an important advance in the field of progress feedback. The expected treatment response model had fairly high NPV, which is important because this ensures that patients at high risk of treatment failure are accurately identified in a timely manner to rectify problems that might be impeding symptom improvement.

The improved accuracy of dynamic models was particularly evident in the early phases of therapy, where information from early response to treatment is leveraged. Previous evidence indicates that early symptomatic response to psychotherapy is a reliable prognostic indicator.[35] Similarly, non-linear patterns of change such as sudden improvements in symptoms are also known to predict treatment outcomes, and these changes mostly occur during the early phase of treatment.[36] Furthermore, dose–response studies in IAPT settings consistently show that most treatment responders show signs of symptomatic improvement during the first 2 months of treatment.[25,27]

Advanced machine learning analyses did not enhance prediction accuracy. Overall, the performance of the most simple and most parsimonious model based on logistic regression was similar to that of more computationally intensive approaches. These results are consistent with previous research showing similar out-of-sample performance for different machine learning approaches applied in various clinical settings.[37,38] Similarly, a systematic review of 71 studies concluded that clinical prediction models trained using machine learning analyses do not significantly outperform those trained using logistic regression.[39] In our study, however, we only had access to a relatively sparse set of patient-level features. In theory, machine learning approaches are thought to be advantageous in datasets with a much larger number of predictors or in situations where there is a larger ratio of features-to-cases (ie, many predictors across a restricted number of individuals).[33,34] The performance of the extreme gradient boosting model was marginally more accurate in the early sessions of treatment (1–6), and less accurate than other models after session 17. This difference in performance is likely to indicate that the variable weighting approach used by the extreme gradient boosting model might help to marginally optimise accuracy in large samples, but the accuracy deteriorates when the sample size reduces below 500, as observed in the present study.

In the future, dynamic prediction algorithms could be integrated into computerised symptom-monitoring and feedback systems that could enable psychotherapists to adjust treatment in real-time for patients with a poor expected prognosis.[14] The improved prediction accuracy of this dynamic model compared with the expected treatment response model used in routine care is likely to minimise classification error by correctly identifying patients at risk of a poor prognosis, thus prompting their timely evaluation as part of clinical supervision, which is a well-established approach to improve outcomes for patients at risk of deterioration.[13] To implement the new Oracle algorithm in routine care, the model could be integrated into clinical data management systems used in psychological services to process routine outcomes data to alert clinicians to patients who are not on track, thus prompting their timely prioritisation for clinical supervision.

The results of our study should be considered in the context of several limitations. Outcomes were defined using patient-reported symptom measures, and no formal post-treatment diagnoses or observer-rated outcomes were available. Furthermore, the post-treatment outcomes were defined using the same measures used to monitor symptoms during treatment; therefore, we could not test prediction accuracy relative to an independent outcome measure. Treatment outcomes were defined at the end of psychological treatment, and therefore the prediction accuracy of these models over a longer timeframe could not be assessed. Furthermore, the majority of patients included in this study were of White British ethnicity; therefore, the generalisability to patients from other ethnic groups remains unclear. Datasets for this study were aggregated at an NHS Trust level, and thus more granular distinctions between different IAPT teams working within these organisations could not be made. It remains unclear whether implementation of this dynamic prediction system would improve outcomes compared with available expected treatment response systems, and evidence from randomised controlled trials would be required to assess this question. Overall, future studies should seek to overcome these limitations by validating the predictive accuracy of dynamic predictions based on self-reported questionnaires relative to observer-rated outcomes or diagnostic interviews over a longer period of time after the end of therapy, and should include more ethnically diverse populations.

In conclusion, dynamic clinical prediction models considerably outperform so-called static predictions made at a single timepoint, and predictions from expected treatment response models used in contemporary feedback systems.

Contributors

CB led the analyses for study 1, MS-B led the analyses for study 2. JD was the principal investigator of the research programme, research grant manager, and data custodian. JD, CB, MS-B, MM, and WL directly accessed and verified the data reported in the manuscript. JD, WL, CB, JR, A-KD, RT, DS, and MS-B conceptualised the study and methods. MM supported data preparation and analysis. All authors contributed to

the interpretation of results and the writing of the manuscript. JD, CB, MS-B, MM, and WL had full access to the study data, and all authors had final responsibility for the decision to submit for publication.

Declarations of interests

We declare no competing interests.

Data sharing

In line with the requirements of the ethics committees who approved this research, requests for access to data are to be made in writing to the corresponding author. Only de-identified participant data can be made available, along with a data dictionary, to suitably qualified researchers who obtain ethical approval for their proposed analysis; pre-register their statistical analysis plan; and provide a signed data-sharing contract, which enables data storage and analysis for a time-limited period. Interested readers and researchers who wish to replicate the analyses described in this study can receive the full record of all regression models and corresponding receiver operation characteristics curve plots in a spreadsheet, the full stream of modelling ensembles, algorithms, and output files from the IBM Modeler software (version 18.2.1), which can be made available free of charge to researchers who request this in writing to the corresponding author.

Acknowledgments

This research was partly funded by a Knowledge Exchange grant from the University of Sheffield (Sheffield, UK; reference X/160853).

References

1 Delgadillo J, Lutz W. A development pathway towards precision mental health care. *JAMA Psychiatry* 2020; **77:** 889–90.

2 Lutz W, Martinovich Z, Howard KI. Patient profiling: an application of random coefficient regression models to depicting the response of a patient to outpatient psychotherapy. *J Consult Clin Psychol* 1999; **67:** 571–77.

3 Finch AE, Lambert MJ, Schaalje BG. Psychotherapy quality control: the statistical generation of expected recovery curves for integration into an early warning system. *Clin Psychol Psychother* 2001; **8:** 231–42.

4 Howard KI, Moras K, Brill PL, Martinovich Z, Lutz W. Evaluation of psychotherapy. Efficacy, effectiveness, and patient progress. *Am Psychol* 1996; **51:** 1059–64.

5 Krause MS, Howard KI, Lutz W. Exploring individual change. *J Consult Clin Psychol* 1998; **66:** 838–45.

6 Garfield SL. Some problems associated with "validated" forms of psychotherapy. *Clin Psychol Sci Pract* 1996; **3:** 218–29.

7 Lutz W, Leach C, Barkham M, et al. Predicting change for individual psychotherapy clients on the basis of their nearest neighbors. *J Consult Clin Psychol* 2005; **73:** 904–13.

8 Saunders R, Cape J, Fearon P, Pilling S. Predicting treatment outcome in psychological treatment services by identifying latent profiles of patients. *J Affect Disord* 2016; **197:** 107–15.

9 Delgadillo J, Moreea O, Lutz W. Different people respond differently to therapy: a demonstration using patient profiling and risk stratification. *Behav Res Ther* 2016; **79:** 15–22.

10 Delgadillo J, Huey D, Bennett H, McMillan D. Case complexity as a guide for psychological treatment selection. *J Consult Clin Psychol* 2017; **85:** 835–53.

11 Lambert MJ. Progress feedback and the OQ-system: the past and the future. *Psychotherapy* 2015; **52:** 381–90.

12 Duncan BL. The Partners for Change Outcome Management System (PCOMS): the heart and soul of change project. *Can Psychol* 2012; **53:** 93–104.

13 Delgadillo J, de Jong K, Lucock M, et al. Feedback-informed treatment versus usual psychological treatment for depression and anxiety: a multisite, open-label, cluster randomised controlled trial. *Lancet Psychiatry* 2018; **5:** 564–72.

14 Lutz W, Rubel JA, Schwartz B, Schilling V, Deisenhofer A-K. Towards integrating personalized feedback research into clinical practice: development of the Trier Treatment Navigator (TTN). *Behav Res Ther* 2019; **120:** 103438.

15 Shimokawa K, Lambert MJ, Smart DW. Enhancing treatment outcome of patients at risk of treatment failure: meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *J Consult Clin Psychol* 2010; **78:** 298–311.

16 Kendrick T, El-Gohary M, Stuart B, et al. Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults. *Cochrane Database Syst Rev* 2016; **7:** CD011119.

17 Lambert MJ, Hansen NB, Finch AE. Patient-focused research: using patient outcome data to enhance treatment effects. *J Consult Clin Psychol* 2001; **69:** 159–72.

18 Clark DM. Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: the IAPT experience. *Int Rev Psychiatry* 2011; **23:** 318–27.

19 Richards DA, Whyte M. Reach out. National programme educator materials to support the delivery of training for psychological wellbeing practitioners delivering low intensity interventions, 2nd edn. London: Rethink, 2009.

20 Roth AD, Pilling S. Using an evidence-based methodology to identify the competences required to deliver effective cognitive and behavioural therapy for depression and anxiety disorders. *Behav Cogn Psychother* 2008; **36:** 129–47.

21 Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001; **16:** 606–13.

22 Richards DA, Borglin G. Implementation of psychological therapies for anxiety and depression in routine practice: two year prospective cohort study. *J Affect Disord* 2011; **133:** 51–60.

23 Kroenke K, Spitzer RL, Williams JBW, Monahan PO, Löwe B. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. *Ann Intern Med* 2007; **146:** 317–25.

24 Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* 1991; **59:** 12–19.

25 Delgadillo J, McMillan D, Lucock M, Leach C, Ali S, Gilbody S. Early changes, attrition, and dose-response in low intensity psychological interventions. *Br J Clin Psychol* 2014; **53:** 114–30.

26 Hsieh FY. Sample size tables for logistic regression. *Stat Med* 1989; **8:** 795–802.

27 Robinson L, Kellett S, Delgadillo J. Dose-response patterns in low and high intensity cognitive behavioral therapy for common mental health problems. *Depress Anxiety* 2020; **37:** 285–94.

28 Nagelkerke NJ. A note on a general definition of the coefficient of determination. *Biometrika* 1991; **78:** 691–92.

29 Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988; **240:** 1285–93.

30 West M, Harrison J. Bayesian forecasting and dynamic models, 2nd edn. New York, NY: Springer, 2006.

31 Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc B* 2005; **67:** 301–20.

32 Chen T, Guestrin C. Xgboost: a scalable tree boosting system. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Fransisco, CA; Aug 13–17, 2016.

33 Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers* 1999; **10:** 61–74.

34 Fine TL. Feedforward neural network methodology, 3rd edn. New York, NY: Springer-Verlag, 1999.

35 Beard JIL, Delgadillo J. Early response to psychological therapy as a predictor of depression and anxiety treatment outcomes: a systematic review and meta-analysis. *Depress Anxiety* 2019; **36:** 866–78.

36 Aderka IM, Nickerson A, Bøe HJ, Hofmann SG. Sudden gains during psychological treatments of anxiety and depression: a meta-analysis. *J Consult Clin Psychol* 2012; **80:** 93–101.

37 Colombet I, Ruelland A, Chatellier G, Gueyffier F, Degoulet P, Jaulent MC. Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression. *Proc AMIA Symp* 2000; 156–60.

38 Lynam AL, Dennis JM, Owen KR, et al. Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults. *Diagn Progn Res* 2020; **4:** 6.

39 Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019; **110:** 12–22.