Psychological Medicine

cambridge.org/psm

Original Article

Cite this article: Hodsoll J *et al* (2024). Predictors of outcome following psychological therapy for depression and anxiety in an urban primary care service: a naturalistic Bayesian prediction modeling approach. *Psychological Medicine* **54**, 4503–4517. https://doi.org/10.1017/S0033291724001582

Received: 25 April 2023
Revised: 17 October 2023
Accepted: 10 June 2024
First published online: 16 December 2024

Keywords:
anxiety; Bayesian prediction modeling; depression; psychological therapy; recovery

Corresponding author:
Anthony J. Cleare;
Email: anthony.cleare@kcl.ac.uk

Predictors of outcome following psychological therapy for depression and anxiety in an urban primary care service: a naturalistic Bayesian prediction modeling approach

John Hodsoll[1], Rebecca Strawbridge[2], Sinead King[2], Rachael W. Taylor[2], Gerome Breen[3], Nina Grant[4], Nick Grey[5], Nilay Hepgul[2], Matthew Hotopf[2,6], Viryanaga Kitsune[2], Paul Moran[7], André Tylee[2], Janet Wingrove[8], Allan H. Young[2,6] and Anthony J. Cleare[2,6]

[1]Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK; [2]Department of Psychological Medicine, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK; [3]MRC Social, Genetic, and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, London, UK; [4]Sussex Partnership NHS Foundation Trust, and Department of Psychology, University of Sussex, Brighton, UK; [5]Centre for Anxiety Disorders and Trauma, South London & Maudsley NHS Foundation Trust, London, UK; [6]South London and Maudsley NHS Foundation Trust, London, UK; [7]Centre for Academic Mental Health, Population Health Sciences Department, Bristol Medical School, University of Bristol, Bristol, UK and [8]Southwark Psychological Therapies Service, South London & Maudsley NHS Foundation Trust, London, UK

Abstract

Background. England's primary care service for psychological therapy (Improving Access to Psychological Therapies [IAPT]) treats anxiety and depression, with a target recovery rate of 50%. Identifying the characteristics of patients who achieve recovery may assist in optimizing future treatment. This naturalistic cohort study investigated pre-therapy characteristics as predictors of recovery and improvement after IAPT therapy.

Methods. In a cohort of patients attending an IAPT service in South London, we recruited 263 participants and conducted a baseline interview to gather extensive pre-therapy characteristics. Bayesian prediction models and variable selection were used to identify baseline variables prognostic of good clinical outcomes. Recovery (primary outcome) was defined using (IAPT) service-defined score thresholds for both depression (Patient Health Questionnaire [PHQ-9]) and anxiety (Generalized Anxiety Disorder [GAD-7]). Depression and anxiety outcomes were also evaluated as standalone (PHQ-9/GAD-7) scores after therapy. Prediction model performance metrics were estimated using cross-validation.

Results. Predictor variables explained 26% (recovery), 37% (depression), and 31% (anxiety) of the variance in outcomes, respectively. Variables prognostic of recovery were lower pre-treatment depression severity and not meeting criteria for obsessive compulsive disorder. Post-therapy depression and anxiety severity scores were predicted by lower symptom severity and higher ratings of health-related quality of life (EuroQol questionnaire [EQ5D]) at baseline.

Conclusion. Almost a third of the variance in clinical outcomes was explained by pre-treatment symptom severity scores. These constructs benefit from being rapidly accessible in healthcare services. If replicated in external samples, the early identification of patients who are less likely to recover may facilitate earlier triage to alternative interventions.

Introduction

Current first-line treatments for depression and anxiety disorders include both psychological and pharmacological interventions. Whilst these have roughly equivalent efficacy (Cleare et al., 2015), patients report a preference for psychological therapies (McHugh, Whitton, Peckham, Welge, & Otto, 2013), although these are more costly to deliver (Koeser, Donisi, Goldberg, & McCrone, 2015). In the UK, the National Institute for Health and Care Excellence (NICE) recommends a range of psychological therapies for depression (National Institute for Health and Clinical Excellence (NICE), 2022), including cognitive behavioral therapy (CBT) and interpersonal psychotherapy (IPT), and CBT for generalized anxiety disorders (NICE, 2011). Although people with depression and anxiety disorders may respond to first-line psychological and pharmacological treatments, this is not the case for all patients. Response rates are typically reported to be between 50% and 60% (Cuijpers, Stringaris, & Wolpert, 2020; Roy-Byrne, 2015).

The Improving Access to Psychological Therapies (IAPT) program was initially designed as a talking therapy service to implement evidence-based psychological interventions for people with mild-to-moderate symptoms of anxiety and depression. The service, based in England, has been relatively successful, with delivery scaled up and current recovery rates at 51.9% in August 2020 (NHS Digital, 2021). This is consistent with other evidence of psychological treatment outcomes for mood and anxiety disorders (Cuijpers et al., 2014), where rates are calculated from treatment 'completers'; defined as people who get at least two therapy sessions, excluding those who are signposted elsewhere on assessment or those who do not attend any sessions.

Given this variability in treatment outcomes, one important question is whether it is possible to identify baseline variables prognostic of therapy working well (for depression and anxiety outcomes). Some demographic and clinical variables are routinely available and may help in tailoring effective treatment choices. Previous research has shown mixed results in different settings using differing methodologies. For example, systematic reviews of factors predicting treatment response for depression (Taylor, Marwood, Greer, Strawbridge, & Cleare, 2019) found inconsistent evidence for a wide range of predictors across studies. Severe symptoms seemed to be the most consistent predictor. Similar findings are apparent with anxiety disorders (Mululo, de Menezes, Vigne, & Fontenelle, 2012). The majority of these studies have taken place in settings which were not representative of real-world care (Doorn, Kamsteeg, Bate, & Aafjes, 2021). Where real-world care data have been used, Coley, Boggs, Beck, and Simon (2021) found little predictive information in baseline variables, although only a limited range of variables were assessed.

Given these mixed results and limited settings, our PROMPT study (Grant et al., 2014) purposefully combined a naturalistic approach (by recruiting participants within an IAPT service) with a collection of a rich set of baseline or pre-therapy variables. Initial findings suggested that people within IAPT commonly have a range of psychiatric comorbidities (Hepgul et al., 2016; Strawbridge, Alexander, Richardson, Young, & Cleare, 2023). The ultimate aim was to comprehensively characterize the predictive information in pre-treatment variables associated with post-treatment outcomes and take the first step toward development of a prediction model for the IAPT service outcomes (Kent, Cancelliere, Boyle, Cassidy, & Kongsted, 2020). The set of variables covered demographic, clinical, diagnostic, and psychosocial domains, using measures intended to be feasibly administered in clinical practice and alignment with the extant literature (Grant et al., 2014). We define this study as exploratory within the Prognosis Research Strategy (PROGRESS) framework (Hemingway et al., 2013; Riley et al., 2013).

This paper addresses the primary aim of the PROMPT study in two parts. Firstly, we describe and summarize key baseline variables and use these as univariate predictors of therapy outcome status (recovered or not/improved or not). Secondly, we use a robust Bayesian prediction model strategy to assess multivariate predictors of outcome. Defining a set of useful predictors is a variable selection problem to which the Bayesian projective prediction approach is particularly well suited (Piironen, Paasiniemi, & Vehtari, 2020). Firstly a 'reference' model is developed which predicts the data (and associated uncertainties) as well as possible. Secondly, information from the reference model is 'projected' to sub-models to find the most parsimonious or simplest model which gives similar predictive performance to the reference model. This method is attractive both clinically and statistically. Clinically, it provides simple, interpretable models. Statistically, projective prediction models incorporate (rather than ignore) the uncertainty associated with the variable selection process. In previously used methods such as frequentist forward or backward selection, this selection uncertainty is ignored leading to over-optimistic models which do not generalize well to other samples. Other methods such as the LASSO contain little information about the uncertainty associated with model parameters. In short, this projective prediction method compares well with other methods of variable selection, both Bayesian (Piironen & Vehtari, 2017a) and non-Bayesian (Bartonicek, Wickham, Pat, & Conner, 2021) with the advantage of allowing valid estimates of the uncertainty of model parameters.

Method

Participants

The study was a naturalistic cohort study of patients referred to the Southwark IAPT Psychological Therapies Service in South London. Participants were recruited by PROMPT study researchers accessing electronic patient records. Lists of individuals who had agreed to take part in research when registered for therapy were obtained. Patients were contacted from this list and asked to take part in PROMPT. If willing, they were invited to a baseline interview with a trained postgraduate researcher (prior to therapy initiation) where assessments selected to measure variables predictive of subsequent therapy outcomes were collected. Subsequently, on starting therapy, patients' progress through therapy sessions was monitored using the treatment outcome data collected routinely within IAPT electronic patient records. The study team accessed and used this follow-up data as outcomes for PROMPT. Retention of blindness in relation to the collection of outcome data was not possible. All research visits took place at the National Institute for Health Research King's Wellcome Clinical Research Facility. The first patient was enrolled on 5 February 2014 and recruitment continued until 29 July 2016.

Participants were eligible to participate if they were aged 18 years or over, had provided informed consent, and were on a waiting list to receive psychological therapy. Exclusions from the research study (but not the IAPT service) were made if participants had already begun therapy or if they could not speak English well enough to carry out the baseline interview.

Treatments

Southwark IAPT talking therapies service implements a stepped care provision in which most individuals are initially offered low-intensity therapy. If individuals do not respond to low-intensity treatment, they are 'stepped up' to high-intensity therapy. Step 1 involves registering the individual with the service and a pre-treatment assessment. Individuals are then assigned to either a step 2 or 3 level treatment. Step 2 (low-intensity therapy) entails interventions such as four to six sessions of guided self-help based on CBT principles, and/or workshops such as CBT for sleep with a trained psychological well-being practitioner. Step 3 (high-intensity therapy) offers a course of, e.g. individual CBT, counseling, or IPT with a trained psychological therapist. Individuals can

https://doi.org/10.1017/S0033291724001582 Published online by Cambridge University Press

be stepped up from step 2 to 3, and/or stepped down (National Collaborating Centre for Mental Health, 2018).

Outcomes

We utilized four outcomes in this analysis: the IAPT recovery index (primary outcome), IAPT reliable change, and total severity scores from the depression and anxiety self-report scales. In fact, outcomes 3 and 4 were the self-report scales from which composite IAPT outcomes are derived; the Patient Health Questionnaire for depression (PHQ-9; Kroenke, Spitzer, & Williams, 2001) and the Generalized Anxiety Disorder assessment for anxiety (GAD-7; Spitzer, Kroenke, Williams, & Löwe, 2006). The PHQ-9 has nine Likert scale items and the GAD-7 has seven items. Patients rate each symptom from 0 'not at all' to 3 'nearly every day' with total scores (maximum 27 on the PHQ-9 and 21 on the GAD-7) used as indicators of symptom severity. These scales have recently been shown to be unidimensional with temporal measurement invariance (Stochl et al., 2020).

The recovery index is a joint consideration of the clinical thresholds of these two scales. Patients are considered a 'clinical case' if they score 10 or more on the PHQ-9 or score 8 or more on GAD-7. Recovery means moving below these clinical cut-offs on both questionnaires at discharge from the IAPT service. The recovered YES/NO classification reflects individuals who have recovered from case-ness at the end of treatment *v.* those who have not. Here we relax the condition of pre-treatment case-ness as including non-cases, in alignment with the naturalistic approach of our study. Additionally, patients who are symptomatic but not cases may still have a clinically meaningful change in outcomes. However, given the IAPT definition of recovery, we include a sensitivity analysis in which recovery is predicted for cases only.

The second outcome, IAPT reliable improvement, considers whether individuals undergo a positive 'reliable change' (Jacobson & Truax, 1991) in *either* PHQ-9 *or* GAD-7 scores. To count as reliable improvement for one outcome there must be no negative change in the other. This outcome putatively identifies patients whose symptoms improve sufficiently beyond that expected by measurement error, even if they do not meet IAPT definition of recovery. Note that the *reliable recovery* index also exists, a further composite of both the recovery index and change score (Gyani, Shafran, Layard, & Clark, 2013). As a composite of a composite, we consider reliable recovery to be particularly difficult to interpret, and accordingly, we only report this in supplementary information also as preliminary analysis showed this outcome to have little association with predictors.

To improve prediction performance and stratify for differences in severity at the start of treatment, PHQ-9 and GAD-7 scores at the start of treatment (first therapy session) were included in present analyses.

Baseline predictors

Measures included as predictors were collected at the baseline visit and were divided into three domains: demographic, clinical/diagnostic, and psychosocial. Demographic variables were age, gender, ethnicity, education, employment status, income, and relationship status. Clinical status was indicated by patient age (in years) from the first onset of psychiatric symptoms, the duration of the current episode in months, and current antidepressant use. Relevant co-occurring diagnoses selected from the MINI diagnostic interview (Sheehan et al., 1998) were bipolarity, diagnoses of social phobia, panic disorder, agoraphobia, obsessive compulsive disorder (OCD), post-traumatic stress disorder (PTSD), alcohol abuse, substance abuse/dependence, anorexia, and bulimia. Other clinical assessments included were personality disorder traits (Standardized Assessment of Personality – Abbreviated Scale [SAPAS]; Moran et al., 2003) and physical health (Cumulative Illness Rating Scale [CIRS]; Linn, Linn, & Gurel, 1968).

Psychosocial predictors were: current beliefs about mental illness (Brief Illness Perceptions questionnaire [B-IPQ]; Broadbent, Petrie, Main, & Weinman, 2006); self-criticism by negative cognitions and self-reassurance (Forms of Self Reassurance and Self Criticism scale [FSRC]; Gilbert, Clarke, Hempel, Miles, & Irons, 2004), self-efficacy (General self-efficacy scale; Schwarzer & Jerusalem, 1995), quality of life (EuroQol questionnaire [EQ5D]; Herdman et al., 2011), social support (Oslo Social Support scale [OSS]; Dalgard et al., 2006), stressful life events (List of Threatening Events Questionnaire [LTE]; Brugha, Bebbington, Tennant, & Hurry, 1985), and adverse events during childhood (Childhood Trauma Questionnaire [CTQ]; Bernstein et al., 2003). Further description of the measures and variables are detailed in the supplementary information.

Statistical analysis

Descriptive statistics

To report our findings we use the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines (Collins, Reitsma, Altman, & Moons, 2015). Table 1 summarizes patient characteristics by recovery condition. Median and interquartile range (IQR) are reported for continuous measures, and frequency and proportion for categorical. Group differences in continuous measures were assessed with the Wilcoxon rank-sum test and for categorical variables using the c^2 test (or Fisher's exact test if expected counts were less than 5 in table cells).

Pre-processing of predictors and univariate associations

Prior to developing the prognostic models and to reduce the opportunity for overfitting, a set of 30 candidate variables were identified from those collected at baseline (Grant et al., 2014) by authors A. J. C., R. S., A. H. Y., and J. H. These were intended to cover as wide a range of variable domains as possible whilst excluding potentially redundant information; for example, relationship support would be captured better by the social support scale than categorical relationship status. For the prognostic modeling, pre-processing of predictors was made to ease both computation and interpretation, i.e. categorical variables simplified (cf. SI methods). Due to the low frequency of occurrence, eating disorder diagnoses were not included, leaving 28 variables. Following this, independent associations between the selected predictor variables and binary outcomes (recovery and reliable improvement) were assessed using Firth penalized logistic regression, adjusted for age, gender, and pre-therapy PHQ-9 and GAD-7 scores.

Prediction modeling

To build a robust and interpretable prognostic model we used Bayesian projective prediction for variable selection. This is a recently developed method with two distinct phases (Piironen et al., 2020). Firstly, a reference model is constructed including all candidate variables as the best model of the data (Pavone,

https://doi.org/10.1017/S0033291724001582 Published online by Cambridge University Press

Table 1. Potential predictor variables by recovery status at the end of treatment

(a): Demographic and sociodemographic variables					
			Recovery		
Characteristic	*N*	Overall, *N* = 263[a]	No, *n* = 124[a]	Yes, *n* = 139[a]	*p* value[b]
Age	263	39 (28, 51)	39 (27, 49)	39 (30, 52)	0.14
Gender – Female	263	168 (64%)	76 (61%)	92 (66%)	0.4
Education	249				0.5
No qualifications		23 (9.2%)	14 (12%)	9 (6.9%)	
GCSE/O' levels/NVQ		37 (15%)	19 (16%)	18 (14%)	
A' levels/GNVQ		55 (22%)	26 (22%)	29 (22%)	
Higher degree or above		134 (54%)	59 (50%)	75 (57%)	
Missing		14	6	8	
Relationship status	208				0.4
Single		100 (48%)	51 (54%)	49 (43%)	
Steady		54 (26%)	23 (24%)	31 (27%)	
Married		36 (17%)	15 (16%)	21 (19%)	
Other		18 (8.7%)	6 (6.3%)	12 (11%)	
Missing		55	29	26	
Ethnicity	206				0.13
White		161 (78%)	70 (71%)	91 (84%)	
Asian		6 (2.9%)	5 (5.1%)	1 (0.9%)	
Black		15 (7.3%)	8 (8.2%)	7 (6.5%)	
Mixed		8 (3.9%)	6 (6.1%)	2 (1.9%)	
Other		16 (7.8%)	9 (9.2%)	7 (6.5%)	
Missing		57	26	31	
Income	247				<0.001
0–£5475		57 (23%)	40 (34%)	17 (13%)	
£5.476–£12 097		34 (14%)	14 (12%)	20 (15%)	
£12 098–£20 753		37 (15%)	20 (17%)	17 (13%)	
£20 754–£31 494		38 (15%)	17 (15%)	21 (16%)	
£31 495 or more		81 (33%)	26 (22%)	55 (42%)	
Missing		16	7	9	
Employment	247				0.031
Employed/student		163 (66%)	68 (58%)	95 (73%)	
Unemployed		48 (19%)	26 (22%)	22 (17%)	
Long term sick		36 (15%)	23 (20%)	13 (10%)	
Missing		16	7	9	
Quality of Life (EQ-5D)	249	0.74 (0.58, 0.85)	0.69 (0.38, 0.81)	0.85 (0.69, 0.85)	<0.001
Missing		14	6	8	
OSS (social support)	249	10 (8, 12)	9 (8, 11)	10 (9, 12)	0.003
Missing		14	6	8	
LTE (life events)	249	5 (3, 7)	5 (3, 7)	5 (4, 6.50)	0.4
Missing		14	6	8	
CTQ (childhood trauma)	243	37 (29, 52)	39 (30, 57)	35 (27, 47)	0.010

(*Continued*)

https://doi.org/10.1017/S0033291724001582 Published online by Cambridge University Press

Table 1. (*Continued.*)

(a): Demographic and sociodemographic variables					
			Recovery		
Characteristic	*N*	Overall, *N* = 263[a]	No, *n* = 124[a]	Yes, *n* = 139[a]	*p* value[b]
Missing		20	8	12	

EQ-5D, EuroQol Quality of Life Instrument 5D (3 level) Utility Index; LTE, List of Threatening Events; CTQ, Childhood Trauma Questionnaire.

(b): Clinical variables					
			Recovery		
Characteristic	*N*	Overall, *N* = 263[a]	No, *n* = 124[a]	Yes, *n* = 139[a]	*p* value[b]
PHQ 9 (depression)	263	14 (9, 19)	18 (13, 21)	10 (5, 14)	<0.001
GAD 7 (dnxiety)	263	12 (7, 16)	15 (11, 18)	9 (5, 14)	<0.001
IAPT case (yes/no)	263	211 (80%)	120 (97%)	91 (65%)	<0.001
Lifetime age of onset (years)	214	16 (12, 25)	15 (11, 22)	18 (13, 25)	0.056
Missing		49	28	21	
Current episode (in months)	203	11 (4, 35)	18 (5, 38)	7 (4, 24)	0.013
Missing		60	31	29	
Current anti-depressants – prescribed (yes/no)	250	104 (42%)	53 (45%)	51 (39%)	0.3
Missing		13	6	7	
Bipolar (*v.* unipolar)	262	75 (29%)	40 (32%)	35 (25%)	0.2
Missing		1	0	1	
SAPAS (personality disorder trait) score	249	3 (2, 5)	4 (3, 5)	3 (2, 4)	<0.001
Missing		14	6	8	
CIRS (physical health)	213	15 (13, 16)	15 (13, 18)	14 (13, 16)	0.074
Missing		50	29	21	
Panic disorder	256				0.021
None		134 (52%)	53 (44%)	81 (60%)	
Current		49 (19%)	30 (25%)	19 (14%)	
Lifetime		73 (29%)	38 (31%)	35 (26%)	
Missing		7	3	4	
Agoraphobia – yes	255	113 (44%)	69 (57%)	44 (33%)	<0.001
Missing		8	3	5	
Social phobia – yes	261	92 (35%)	55 (45%)	37 (27%)	0.003
Missing		2	1	1	
Obsessive-compulsive-disorder – yes	262	52 (20%)	37 (30%)	15 (11%)	<0.001
Missing		1	0	1	
Post-traumatic-stress-disorder – yes	261	33 (13%)	24 (20%)	9 (6.5%)	0.002
Missing		2	1	1	
Alcohol abuse – yes	263	60 (23%)	29 (23%)	31 (22%)	0.8
Substance abuse – yes	263	19 (7.2%)	14 (11%)	5 (3.6%)	0.016
Bulimia – yes	261	8 (3.1%)	5 (4.1%)	3 (2.2%)	0.5
Missing		2	1	1	
Anorexia – yes	259	0 (0%)	0 (0%)	0 (0%)	>0.9
Missing		4	3	1	

(*Continued*)

https://doi.org/10.1017/S0033291724001582 Published online by Cambridge University Press

Table 1. (*Continued.*)

(b): Clinical variables					
			Recovery		
Characteristic	*N*	Overall, *N* = 263[a]	No, *n* = 124[a]	Yes, *n* = 139[a]	*p* value[b]
IPQ (illness perception score)	242	48 (43, 54)	51 (46, 57)	45 (39, 51)	<0.001
Missing		21	12	9	
FSCRS (self-criticism score)	248	28 (20, 37)	32 (21, 41)	26 (17, 35)	0.001
Missing		15	6	9	
FSCRS (self-reassurance score)	248	15.0 (10.0, 19.0)	13.0 (9.0, 18.0)	16.0 (12.0, 19.8)	0.008
Missing		15	6	9	
Self-efficacy score	245	27.0 (23.0, 30.0)	26.0 (21.2, 28.8)	28.0 (24.8, 30.0)	0.004
Missing		18	7	11	

PHQ, Patient Health Questionnaire; GAD, Generalized Anxiety Disorder scale; SAPAS, Standard assessment of Personality – Abbreviated Scale; CIRS, Cumulative Illness Rating Scale; IPQ, Illness Perceptions Questionnaire; FSCRS, Forms of Self-Criticizing/Attacking and Self-Reassuring Scale.
[a]Median (IQR); *n* (%).
[b]Wilcoxon rank-sum test; Pearson's c^2 test; Fisher's exact test.

Piironen, Bürkner, & Vehtari, 2022). Best predictive performance is evaluated as the predicted out-of-sample performance (how well the model would predict new data), estimated via leave-one-out cross-validation (LOO-CV). Prediction performance was measured by the expected log predicted density (ELPD; Gelman, Hill, & Vehtari, 2020). In constructing the reference model, we addressed potential overfitting by using regularized horseshoe priors, which have been shown to perform well in comparison to other informative priors (Piironen & Vehtari, 2017a). These function by weighting model coefficients toward zero but allowing larger coefficients with more information to escape this weighting. In setting the prior, the number of expected non-zero variables is defined. Here, an expectation of five non-zero variables seemed reasonable with the aim of producing a parsimonious model.

Having defined a reference model, the second phase variable selection was undertaken via projective prediction (Piironen et al., 2020). The latter involves both a search and evaluation component. In the search, variables are ranked for inclusion by the degree to which they reduce the Kullback–Leiber divergence (a measure of information) between the reference and null model. Sub-models (with variables input by rank) are compared to the reference models using the ELPD. Non-inferiority in predictive performance was set at least 1 standard error of the ELPD difference and determined the number of variables in the model. Cross-validation was used for both search and evaluation procedures to avoid overfitting, but this also meant it was possible to evaluate model stability. Model stability was estimated by how often variables were included in cross-validation sets for each different size sub-model, i.e. the number of times a particular variable was included in a sub-model of size 2, 3, 4, and so on up to 30. The coefficients are interpretable as associations between the predictor and outcome conditional on the model selection process (and as is clear for prediction models, do not represent estimates of causal associations).

Bayesian models were specified with the R package rstanarm (Goodrich, Gabry, Ali, & Brilleman, 2022) which fits models using Hamilton Monte Carlo Markov chains (HMC). HMC chains work by repeatedly sampling from a combination of prior information and data (posterior distribution) to reach a best (and stable) solution. Models were run with four chains of 1000 warm up and 2000 sampling iterations and checked for adequacy with diagnostic tests for convergence: visual, Gelman and Rubin statistics (Rhat) and effective sample size (ESS). All models reported below passed these checks with credible results (Rhat near 1 and ESS > 2000). Model coefficients are reported as median and 95% credible intervals of the posterior distributions.

Model performance

To assess logistic regression model performance for recovery and reliable improvement outcomes, we generated predicted probabilities adjusted for overfitting by cross-validation. Using these predicted probabilities, metrics for model validation and calibration were calculated. Firstly, the C statistic or area under the curve (AUC) indicates how well the model can discriminate between recovered and non-recovered cases. The AUC is the proportion of pair-wise comparisons between participants with an event (recovery) with to those without. To address calibration (how well outcome probability matches the true underlying probability), the calibration intercept and slope are estimated by a logistic regression of the outcomes by predicted probabilities. For the calibration slope, deviations from 1 indicate the degree to which predicted probabilities do not agree with those expected across the range of risk of the outcome. The calibration intercept gives a measure of overall calibration with 0 good, and less or greater than 0 reflecting over or underestimation of risk, respectively. For continuous outcomes (PHQ-9 depression and GAD-7 anxiety severity scores), the root mean squared error of the fitted linear models was reported. Finally, for all models, the Bayesian R^2 (Gelman, Goodrich, Gabry, & Vehtari, 2019) was calculated as a measure of explained variance.

Missing data

There were missing data in both the outcome and predictor variables. Participants with missing outcomes were removed from further analysis (45/308, 14.6%) as they had not had therapy (cf section 'Sample size and candidate variables'). In total, there

https://doi.org/10.1017/S0033291724001582 Published online by Cambridge University Press

were 446 missing predictor values (6.3%) which varied according to predictor. For 20 predictors, the total number of missing values was low (<6%) with only BME status (22%), current episode duration (23%), age of onset (19%), and CIRS (19%) above this. In total, 132 (50%) of participants had a full set of predictor variable data and 248 (94%) with three or less predictor values missing. We generated a single imputation of missing data prior to modeling using a non-parametric random forest algorithm available in the MICE package. This algorithm can handle datasets with both continuous and categorical variables and makes few assumptions about the multivariate structure of the data (Waljee et al., 2013).

Sample size and candidate variables

This was a naturalistic, observational study. *A priori* sample size estimates were based on patient throughput, and it was estimated that 600 patients could be recruited into the baseline PROMPT interview across the recruitment period. This was an overestimation, and 371 were ultimately recruited. It was anticipated that 28% of patients would drop-out of treatment (Grant et al., 2014). Ultimately, drop-out and enrolment of patients who had already started therapy led to the exclusion of 108 (29%) individuals (see Fig. 1) which gave an event-to-predictor ratio of 4.1 (124 non-recovered/263). In fact, this reduced sample size supported the adoption of a Bayesian prediction modeling approach with informative priors. Regularized horseshoe priors were used as these have been shown to perform well (i.e. produce stable and reliable models) when the ratio of events to predictors is quite low (Piironen & Vehtari, 2017b).

Results

Participant flow

Of 1806 patients identified and 922 assessed for eligibility, 371 (40.2% of those deemed eligible) were enrolled into the study. In total, 263 patients completed therapy, defined by attendance at least two therapy sessions, Fig. 1. The median number of therapy sessions was 8 (IQR: 6, 14; minimum 2, maximum 32).

Outcomes

Of the 263 eligible patients who completed the study, 139 (52.9%) were classed as recovered. Median PHQ-9 scores were 4 (IQR: 2, 6) in the recovered group and 15 (IQR: 11, 19) in the non-recovered group. Median GAD-7 scores were 3 (IQR: 2, 5) in the recovered group and 13 (IQR: 9, 16) in the non-recovered group. In total, 53.6% of patients (141/263) were classified as having made a reliable improvement. In total, 99/141 (70.2%, or 37.6% of the whole sample) met both response definitions of recovery and reliable improvement.

Therapy allocation

Of the 263 patients enrolled in the longitudinal study, 99 (37.6%) were allocated to step 2 therapy (commonly computerized CBT or guided self-help) and the remaining 164 (62.4%) to step 3 therapy (largely face-to-face counseling or CBT). Patients had up to three separate treatment courses with the majority receiving one course of treatment. Recovery was more likely after one treatment (123/220, 57% recovered) relative to two (12/30, 40% recovered) or 3 treatments (3/12, 25% recovered). Patients who had recovered/responded after step 2 treatment were discharged whereas those who did not respond were likely to move to step 3.

Sample characteristics for recovered and non-recovered groups

Table 1 shows differences in baseline characteristics between recovered and non-recovered patients. These group differences were observed for several variables. In terms of demographics there were differences in employment and income with the non-recovered group having larger proportions of people in the unemployed or long-term sick categories relative to recovered. The recovered group had higher household income than non-recovered. Clinically, for recovered *v.* non-recovered patients, their lifetime age of onset was higher (median 18 *v.* 15) and current episode duration was shorter (median 7 *v.* 18 months). Further, recovered patients were less likely to have diagnoses of panic disorder (40% *v.* 56%), agoraphobia (33% *v.* 57%), social phobia (27% *v.* 45%), OCD (11% *v.* 30%), PTSD (6.5% *v.* 20%), and reported substance abuse (3.6% *v.* 11%). Recovered patients also had fewer personality disorder traits indicated (SAPAS score 3 *v.* 4), and better scores on psychosocial measures: they had more positive illness perception scores, higher self-efficacy ratings, were less self-critical (lower negative cognition and higher self-reassurance), had higher QoL scores, better social support, and lower reports of childhood trauma.

At the start of therapy, recovered patients had lower PHQ-9 and GAD-7 scores. For the PHQ-9, the median score was 10 in the recovered group *v.* 18 in the non-recovered group ($p < 0.001$). Similarly, median GAD-7 scores were 9 in the recovered *v.* 15 in the non-recovered group ($p < 0.001$). Not all patients were cases at pre-treatment (treatment session 1): 211 of the 263 (80.2%) patients were IAPT cases. Of the 52 non-cases, 48 (92.3%) were recovered at the end of therapy. Four patients who were non-cases at the start of treatment were cases at outcome.

Univariate predictors of recovery and reliable improvement

Univariate logistic regressions with recovery and reliable improvement as outcomes showed associations with several variables, but a more limited range than the group differences described in section 'Sample characteristics for recovered and non-recovered groups' when adjusted for age, gender, and pre-treatment depression (PHQ-9) and anxiety (GAD-7); cf. Table 2. For recovery, unemployment through long-term sickness, higher personality disorder trait scores, and diagnoses of agoraphobia and OCD and lower QoL were associated with lower odds of recovery. Only agoraphobia and OCD diagnoses remained statistically significant after FDR adjustment for multiple comparisons. Patients were less likely to indicate reliable improvement if they had lower educational qualifications, higher personality disorder scores, were taking antidepressant medication, met criteria for agoraphobia, social phobia, and OCD, and had lower self-efficacy and QoL ratings. Following adjustment for multiple comparisons, associations with personality disorder, OCD diagnosis, self-efficacy, and QoL remained statistically significant.

Bayesian prediction modeling

Recovery

For the recovery outcome, the reference model had an AUC of 0.80 and Bayesian R^2 of 0.26 (95% Bayesian credible intervals

https://doi.org/10.1017/S0033291724001582 Published online by Cambridge University Press

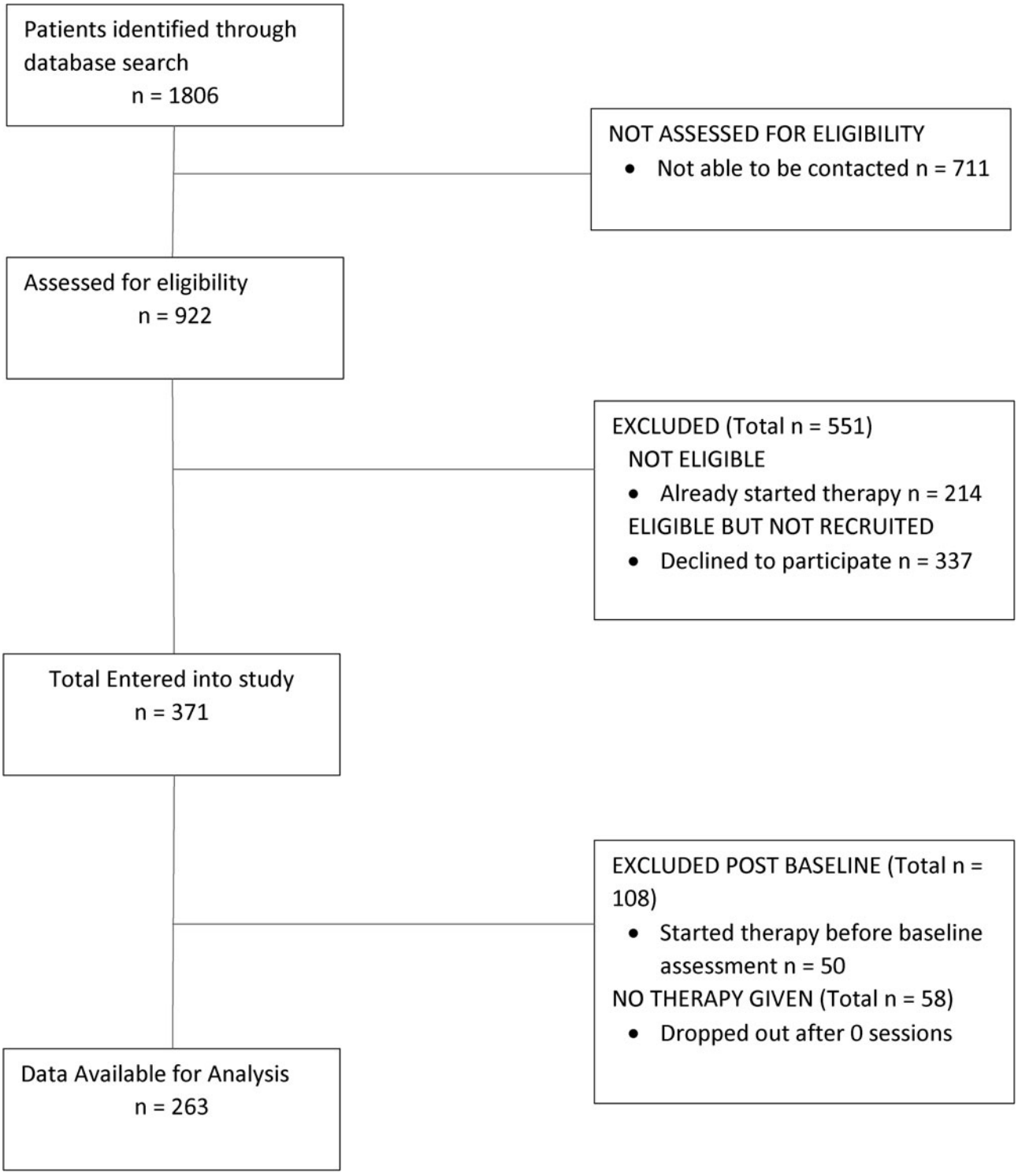

Figure 1. Flowchart of participants through the study.

[bCI] 0.16–0.36). Internal calibration was good with slope and intercept at 0 and 1 (Table 3). Regularization via horseshoe priors meant most variables had an OR of 1 with only pre-treatment PHQ-9, OCD, and agoraphobia diagnoses showing a discrepancy from 0 of greater than 0.01 (cf. SI Table 2.2.1). Projective inference resulted in a parsimonious model with two predictors for performance (Fig. 2). Recovery was negatively associated with higher PHQ-9 scores (OR 0.30, 95% bCI 0.20–0.41) and OCD diagnosis (OR 0.75, 95% bCI 0.52–0.98). Using the divide-by-4 rule for the log-odds-ratio (Gelman et al., 2020), a 1 s.d. increase in baseline PHQ-9 (6.5 points in this sample) decreases the probability of recovery by 32%. A diagnosis of OCD reduces the probability of recovery by a maximum of 7.2%. In terms of model stability, both variables were included in the two-variable model for 100% of the cross-validation sets (SI Table 2).

https://doi.org/10.1017/S0033291724001582 Published online by Cambridge University Press

Table 2. Univariate predictors of Recovery and Reliable Improvement outcomes, adjusted for age, gender, and pretreatment depression (PHQ 9) and anxiety (GAD 7) scores

	Recovery			Reliable improvement		
Variable	OR (95% CI)	*p* value	*q* value[a]	OR (95% CI)	*p* value	*q* value[a]
PHQ 9 (depression)	0.33 (0.21–0.5)	0.000	0.000	0.73 (0.5–1.06)	0.097	0.224
GAD 7 (anxiety)	0.8 (0.54–1.19)	0.270	0.403	2.43 (1.66–3.65)	0.000	0.000
Demographics						
Age	1.28 (0.97–1.72)	0.081	0.221	1.1 (0.85–1.43)	0.464	0.581
Gender	0.82 (0.62–1.09)	0.177	0.332	0.84 (0.65–1.08)	0.176	0.330
BAME	0.83 (0.41–1.66)	0.599	0.642	0.98 (0.51–1.87)	0.947	0.973
Education	1.17 (0.86–1.59)	0.324	0.403	1.41 (1.06–1.88)	0.018	0.077
Employed: unemployed	0.59 (0.28–1.23)	0.160	0.320	0.74 (0.38–1.45)	0.381	0.520
Employed: long-term sick	0.42 (0.17–0.98)	0.045	0.193	0.52 (0.24–1.11)	0.093	0.224
Income	1.42 (1.06–1.9)	0.019	0.105	1.26 (0.96–1.67)	0.093	0.224
Clinical/diagnostic						
Age of onset (lifetime)	1.13 (0.83–1.54)	0.434	0.501	1.03 (0.79–1.36)	0.815	0.905
Current episode duration	0.82 (0.61–1.1)	0.191	0.337	0.91 (0.7–1.18)	0.465	0.581
Current anti-depressants	0.85 (0.63–1.13)	0.254	0.401	0.76 (0.58–0.98)	0.035	0.117
Bipolar	0.93 (0.7–1.24)	0.625	0.647	0.97 (0.75–1.27)	0.845	0.905
SAPAS (personality)	0.7 (0.52–0.95)	0.021	0.105	0.65 (0.49–0.85)	0.002	0.030
CIRS (physical health)	0.83 (0.61–1.12)	0.219	0.365	0.77 (0.58–1.02)	0.064	0.192
Panic disorder	0.77 (0.58–1.03)	0.075	0.221	0.86 (0.66–1.11)	0.249	0.401
Agoraphobia	0.66 (0.49–0.88)	0.005	0.050	0.73 (0.55–0.97)	0.028	0.105
Social phobia	0.76 (0.57–1.02)	0.067	0.221	0.72 (0.55–0.95)	0.018	0.077
OCD	0.65 (0.47–0.88)	0.005	0.050	0.69 (0.52–0.9)	0.007	0.042
PTSD	0.76 (0.56–1.01)	0.059	0.221	0.81 (0.63–1.05)	0.114	0.244
Alcohol abuse	1.11 (0.83–1.48)	0.488	0.542	1 (0.77–1.31)	0.973	0.973
Substance abuse	0.8 (0.58–1.08)	0.149	0.319	0.89 (0.69–1.14)	0.352	0.503
Psychosocial						
IPQ (illness perception)	0.75 (0.53–1.05)	0.097	0.242	0.79 (0.59–1.07)	0.128	0.256
FSCRS (self-criticism)	0.96 (0.7–1.31)	0.780	0.780	0.85 (0.64–1.13)	0.269	0.404
FSCRS (self-reassurance)	1.15 (0.86–1.54)	0.336	0.403	1.07 (0.82–1.39)	0.638	0.766
Self-efficacy	1.17 (0.88–1.56)	0.285	0.403	1.46 (1.11–1.94)	0.006	0.042
QoL (EQ-5D)	1.46 (1.06–2.03)	0.019	0.105	1.54 (1.15–2.07)	0.003	0.030
OSS (social support)	1.16 (0.87–1.55)	0.318	0.403	1.17 (0.9–1.52)	0.254	0.401
LTE (life events)	0.86 (0.64–1.15)	0.304	0.403	0.84 (0.65–1.09)	0.195	0.344
CTQ (childhood trauma)	0.79 (0.59–1.06)	0.112	0.258	0.95 (0.73–1.23)	0.680	0.785

PHQ, Patient Health Questionnaire; GAD, Generalized Anxiety Disorder scale; SAPAS, Standard assessment of Personality – Abbreviated Scale; CIRS, Cumulative Illness Rating Scale; IPQ, Illness Perceptions Questionnaire; FSCRS, Forms of Self-Criticizing/Attacking and Self-Reassuring Scale; EQ-5D, EuroQol Quality of Life Instrument 5D (3 level) Utility Index; OSS, Oslo Social Support scale; LTE, List of Threatening Events; CTQ, Childhood Trauma Questionnaire.
[a]False discovery rate correction for multiple testing.

The model showed similar results when including only cases at baseline; only PHQ-9 and OCD diagnosis were included in the projection model. However, there was some concomitant reduction in model performance (AUC = 0.72, Bayesian R^2: 0.14, bCI 0.06–0.23). The association with pre-treatment PHQ-9 was weaker (OR 0.45, 95% bCI 0.30–0.62) but the OCD association was near identical (OR 0.73, 95% bCI 0.48–0.99). PHQ-9 and OCD were in 100% of the cross-validation sets (SI Table 4).

Reliable improvement

The reference model for reliable improvement had a Bayesian R^2 of 0.12 (95% bCI 0.04–0.19) and AUC of 0.69; predictive

https://doi.org/10.1017/S0033291724001582 Published online by Cambridge University Press

Table 3. Performance statistics for reference and projection models for recovery, reliable improvement, depression (PHQ 9), and anxiety (GAD 7) scores

	Recovery		Reliable improvement		Depression (PHQ-9)		Anxiety (GAD-7)	
Statistic	Reference	Projection	Reference	Projection	Reference	Projection	Reference	Projection
C (ROC)	0.79	0.79	0.69	0.7				
Brier	1.19	1.19	1.22	1.22				
Intercept	0	−0.01	0.01	0				
Slope	0.98	1.01	0.97	1.02				
R^2	0.26 (0.16, 0.36)	0.27 (0.17, 0.36)	0.12 (0.04, 0.19)	0.12 (0.05, 0.19)	0.37 (0.25, 0.46)	0.37 (0.25, 0.46)	0.31 (0.21, 0.4)	0.31 (0.21, 0.4)
ELPD (s.e.)	−144 (7.8)	−145 (7.8)	−167 (5.5)	−165 (5.5)	−823 (11.3)	−824 (11)	−323 (11.3)	−322 (11.1)

C (ROC), C statistic (area under the curve); ELPD (s.e.), expected log predictive density (standard error); PHQ, Patient Health Questionnaire; GAD, Generalized Anxiety Disorder.

performance was notably lower than that for recovery. The projection model included four terms (Fig. 2). The odds of reliable improvement increased for higher baseline GAD-7 scores (OR 2.27, 95% bCI 1.59–3.14), and for higher baseline QoL (OR 1.21, 95% bCI 0.99–1.69). The odds of reliable improvement decreased with an OCD diagnosis (OR 0.86, 95% bCI 0.59–1.02) and increased SAPAS scores (OR 0.85, 95% bCI 0.59–1.02). In terms of variable stability, these four variables were included in 100% of cross-validation folds (SI Table 6).

Continuous depression outcome (PHQ-9)

For PHQ-9, the reference model (SI Table 7) had a Bayesian R^2 of 0.37 (95% bCI 0.25–0.46). Projective prediction resulted in a sub-model with two variables, baseline depression (PHQ-9) and QoL scores; see Fig. 3a. More severe depressive symptoms (higher pre-treatment PHQ-9) were associated with higher PHQ-9 scores at outcome ($b = 3.45$; 95% bCI 2.67–4.18), as were lower QoL scores ($b = -1.33$; 95% CI −2.12 to −0.39). Note that a sub-model additionally including OCD resulted in equivalent performance to the reference model ($b = 0.77$; 95% CI 0.04–1.51). The selected variables were included in the 100% of cross-validation folds (SI Table 8).

Continuous anxiety outcome (GAD-7)

For the anxiety reference model (SI Table 9) Bayesian R^2 was 0.31 (95% CI 0.21–0.40). Using projective prediction, a sub-model with four variables gave performance equivalent to the reference model. More severe anxiety symptoms (higher GAD-7 scores) at outcome were associated with higher pre-treatment GAD-7 ($b = 1.39$, 95% bCI 0.21–2.42) and PHQ-9 scores ($b = 1.42$, 95% bCI 0.14–2.45), lower QoL ($b = -0.76$, 95% bCI −1.49 to −0.10), and meeting criteria for agoraphobia ($b = 0.83$, 95% bCI 0.18–1.53). Again, model stability was good with variables appearing in 100% of cross-validation sets (SI Table 10).

Discussion

Summary of findings

This study set out to assess whether clinically useful baseline predictors of psychotherapy outcomes could be identified in a naturalistic setting (an English psychotherapy 'IAPT' service). Including a rich set of baseline predictor variables, we showed in a descriptive analysis that many of these predictors carry information predictive of outcomes (cf. Tables 1 and 2). However, combining these predictors in a multivariable prediction model showed that baseline predictor information adds little beyond that contained in pre-treatment depression and anxiety severity scores. This held for all versions of the clinical outcome. Not only were potentially useful baseline predictors few in number (projection models contained a maximum of 3 baseline variables in addition to pre-treatment depression and anxiety), but the strength of the associations between these baseline predictors and outcomes was relatively small (Figs 2 and 3). The usefulness of these baseline predictors in developing prognostic models for treatment outcome may be only modest.

Selection of predictor variables

There was some variation in the variables included in the prediction models by outcome. Considering baseline symptoms, pre-treatment depression severity (PHQ-9) predicted both recovery and PHQ-9 severity outcomes; both pre-treatment anxiety and depression severity predicted GAD-7 outcome and only pre-treatment anxiety for reliable improvement. These differences are due to the correlation structure between PHQ-9 and GAD-7 scores at pre-treatment and outcomes. For example, the partial correlations between post-treatment PHQ-9 and pre-treatment GAD-7 score are independent given (conditioned on) pre-treatment PHQ-9 scores. In the case of reliable improvement, since it is a change score, the effect of pre-treatment scores may be due to residual confounding. In fact, higher GAD-7 pre-treatment scores predicted greater improvement consistent with regression to the mean.

Although contributions to prediction performance were relatively modest, there were some variables providing additional information, although this varied by outcome. Not meeting the criteria for OCD was associated with better odds of recovery and reliable improvement. Secondly, higher QoL was associated with higher odds of reliable improvement, and with less severe depression and anxiety scores at the end of treatment. Thirdly, fewer personality disorder traits (lower SAPAS scores) were associated with higher odds of reliable improvement. Finally, not meeting criteria for agoraphobia was associated with lower endpoint anxiety (GAD-7). Whilst there were some discrepancies between the variables included per outcome, it should be noted that these four variables were identified for all models in the top 5 or 6 ranked items in the search algorithm and all these

https://doi.org/10.1017/S0033291724001582 Published online by Cambridge University Press

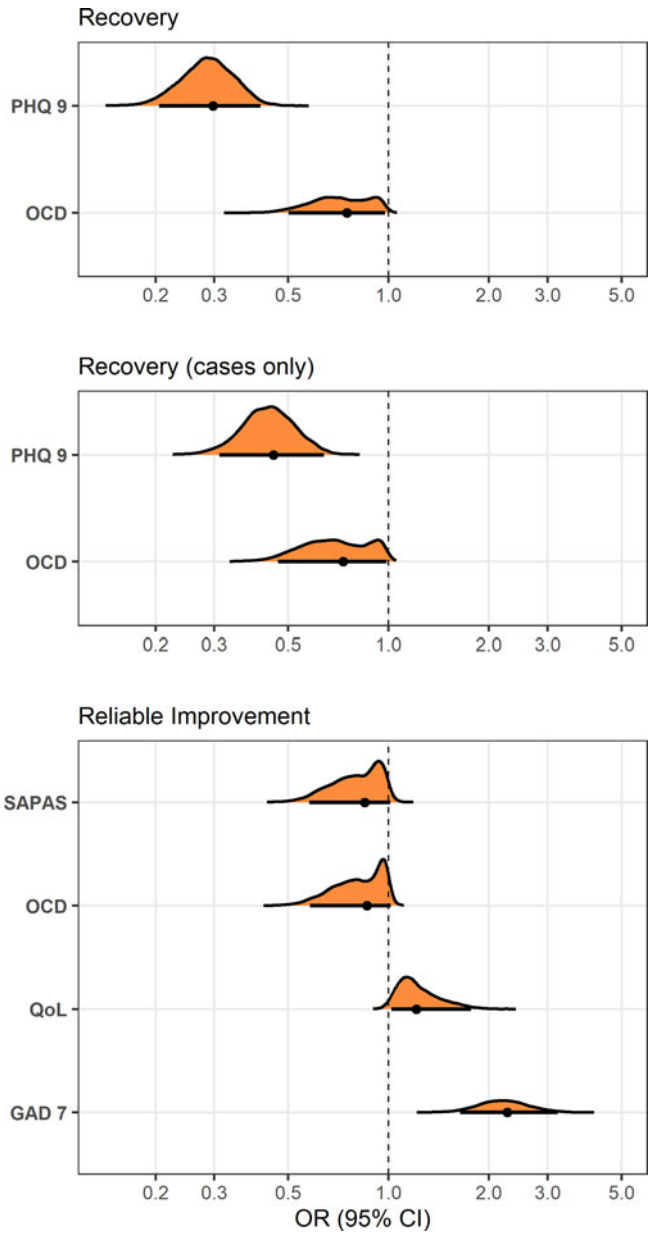

Figure 2. Odds ratios (OR) for posterior distributions from projection models for recovery and reliable improvement outcomes, visualizing the posterior distribution with the black dot and bar representing the median and 95% Bayesian credible intervals (CI). The pre-treatment predictors are: PHQ, Patient Health Questionnaire; GAD, Generalized Anxiety Disorder scale; SAPAS, Standard assessment of Personality; QoL: EQ-5D, EuroQol Quality of Life Instrument 5D (3 level) Utility Index; OCD, diagnosis of Obsessive Compulsive disorder.

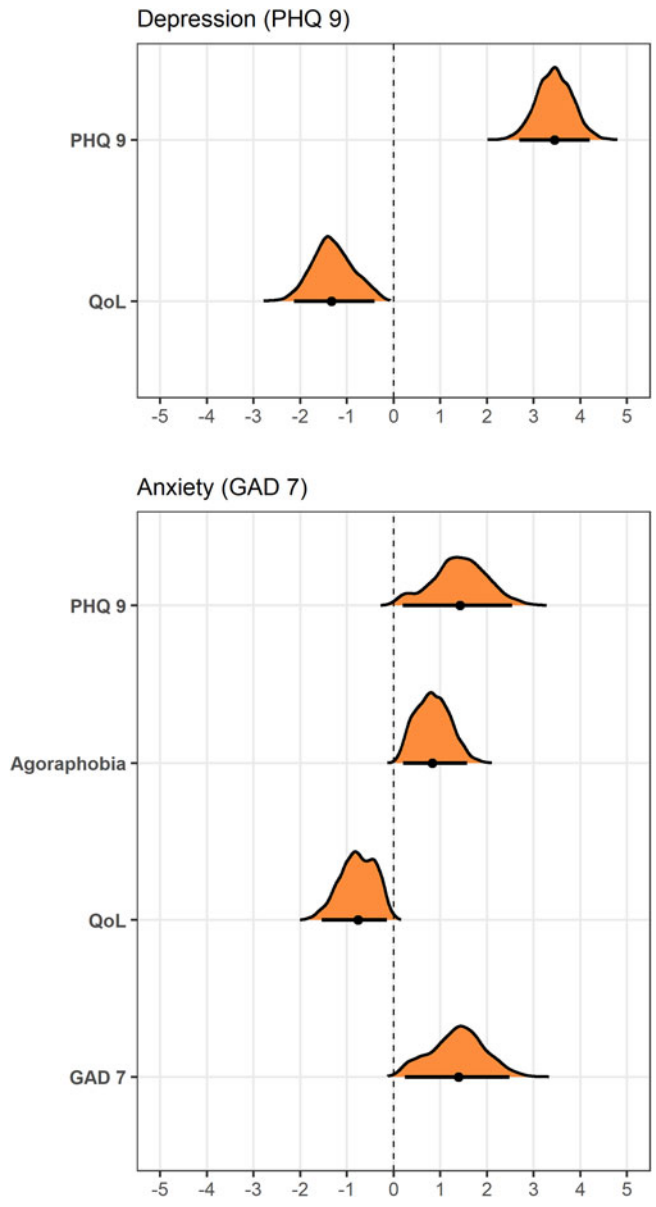

Figure 3. Coefficient posterior distributions from projection models for depression (PHQ-9) and anxiety (GAD-7) continuous outcomes, visualizing the posterior distribution with the black dot and bar representing the median and 95% Bayesian credible intervals (CI). The pre-treatment predictors are: PHQ, Patient Health Questionnaire; GAD, Generalized Anxiety Disorder scale; QoL: EQ-5D, EuroQol Quality of Life Instrument 5D (3 level) Utility Index; Agoraphobia diagnosis.

variables could contribute positively to the prediction of outcomes. For example, for the PHQ-9 severity outcome, if QoL was not included in variable selection, a combination of OCD, agoraphobia, and SAPAS would give match reference model performance (Bayesian $R^2 = 0.37$, SI Table 12). These other variables, although their predictive information is relatively modest, may suggest targets for treatment. For example, Simkin, Hodsoll, and Veale (2022) showed the interdependence of change in OCD and depression symptoms through therapy (in an observational cohort of OCD patients).

As indicated in the descriptive analysis in Tables 1 and 2, this does not mean that variables not selected in the projection prediction model cannot be informative for outcomes. There were several univariate associations for both recovery and reliable improvement as shown in Table 2. All these variables contain predictive information about the outcome which overlaps with the pre-treatment variables in the projection models. To illustrate this point, we re-ran the Bayesian variable selection process excluding the PHQ-9 and GAD-7 pre-treatment scores from the projection pathway for the PHQ 9 outcome. In estimating the predictive performance of the remaining variables (relative to the reference model), whilst not able to match the reference model performance for the recovery outcome, a three- or four-variable model with QoL (EQ5D), illness perceptions, OCD diagnosis, and income reached within 1 S.E. of maximum performance at Bayesian $R^2 = 0.24$ (SI

https://doi.org/10.1017/S0033291724001582 Published online by Cambridge University Press

Table 13). Similarly for GAD-7, a four- or five-variable model with QoL (EQ5D), illness perceptions, OCD, agoraphobia diagnoses, and income had an R^2 of 0.24 (SI Table 14).

Comparisons of outcomes

Overall, model performance differed between the outcomes, with recovery and dimensional continuous scores performing better than reliable improvement (as Bayesian R^2 can be calculated for all models, Table 3). One point is that reliable improvement is based on change scores which in terms of predictive performance are less statistically efficient than endpoint scores as they contain two sources of variability (pre and post treatment). A further loss of power results from the dichotomization of the continuous measure of change and the threshold for change being conservative. Indeed, reliable change assumes homogeneity of measurement error and is liable to recording false positives when measurement error is high, missing true smaller change when measurement error is low (McAleavey, 2024). Whilst recovery showed better performance than reliable improvement it should be noted that excluding non-cases (as per the service definition of recovery) reduced performance considerably ($R^2 = 14\%$ for cases only *v.* 26% including non-cases). The best predictive performance was for continuous depression and anxiety outcomes. Arguably these outcomes are the best choice for predicting patient outcomes. Whilst the recovery index is typically preferred by stakeholders for purposes of interpretation, classification of patient response could be made using predicted scores from the projective prediction models for PHQ-9 and GAD-7. Note that the mean absolute difference between observed and predicted scores from the projection models here was smaller than the criteria for reliable improvement; 4.5 for the PHQ-9 and 3.8 for the GAD-7 relative to the range of the scales (0–27 PHQ-9 and 0–21 GAD-7).

Relation to literature

The finding that a simple model based primarily on pre-treatment depression or anxiety severity predicts outcomes after psychological therapy mirrors other recent attempts to identify predictors of anxiety and depression treatment outcomes. Several studies are also relevant from the psychiatric prediction modeling literature. In a naturalistic psychiatric hospital setting, Webb et al. (2020) predicted 2-week outcomes (PHQ-9 scores) for 484 inpatient depression patients undergoing various multimodal therapies, predominantly intensive group CBT over a 2-week period with adjunct individual CBT therapy and medication if needed. As here, the most important predictor for the trained model was PHQ-9 baseline scores but the inclusion of 13 other demographic and clinical variables (out of a possible 51) significantly improved prediction performance (increasing R^2 by 8%). Interestingly, there were similarities and differences to our results with some reflecting the different set of included variables (e.g. QoL). OCD was included in Webb et al.'s elastic net model, along with other anxiety diagnoses (social anxiety, PTSD, and GAD scores); whilst our reference model for PHQ-9 indicated predictive information in the social anxiety, agoraphobia, and PTSD variables (SI Table 8), the projection model showed that these were not needed for equivalent prediction performance. One advantage of the Bayesian projection approach here is that a minimal set of predictive variables with associated uncertainty can be found in addition to standard regularization methods.

Conversely, other studies have been more negative about the information value of prognostic variables other than baseline outcome. The NESDA study (Dinga et al., 2018) predicted MDD diagnosis (2 years post baseline, $n = 804$) for unipolar depression patients recruited from the general population, undergoing psychotherapy, pharmacotherapy or no treatments. Prediction performance was modest overall (AUC = 0.66) with other variables from psychological, clinical, and biological domains adding only 0.01 to that prognostic value. Similarly, two anti-depressant treatment studies found that baseline depression severity items were the strongest predictor of remission, namely the GENDEP study (Iniesta et al., 2016), and Chekroud et al. (2016), the latter using data from the STAR*D trial. Whilst Iniesta et al. showed clinically useful prediction performance (AUC = 0.72), the prediction model for the Chekroud study showed poor performance with external data from other trials (max AUC = 59.7%). The predominance of baseline severity symptom score appears to generalize over both therapy and anti-depressant treatments.

Finally, two recent studies focused on psychotherapy outcomes using routine data are particularly relevant. Coley et al. (2021) developed and evaluated prediction models for psychotherapy (typically CBT) outcomes in depression. The strongest predictor for PHQ-9 outcomes was baseline PHQ-9 scores with only anxiety and medication further included in the model. In contrast to the results here, model performance both for this outcome and dichotomized 50% reduction was poor. The model only explained about 12% of PHQ-9 outcome variance whilst the AUC for 50% reduction was 57% and calibration poor. Similarly, an examination of outcomes to outpatient CBT (Hilbert et al., 2020) found that performance of the prediction models was robust, but not at the level of clinical utility (for remission AUC < 0.6 and dimensional scores $R^2 < 0.05$). Both these studies used routine care databases with separate validation sets rather than estimating out-of-sample performance using cross-validation. Given the contrast in these findings to the relatively good performance of our prognostic models identified, the translation of findings from smaller research-led studies such as this one to routine care will need well-designed validation.

Strengths

The PROMPT naturalistic study was designed as a minimal adjunct to the IAPT service delivery and as such reflects care and outcomes in real-world service provision. Consequently, our prognostic factors and model identified have potential applicability to clinical practice. Secondly, the use of modern Bayesian methods and the use of a reference model allowed an evaluation of variable selection which was statistically valid. In estimating of out-of-sample performance through cross-validation we were able to evaluate the stability of variable selection and uncertainty of the final model parameters. Thirdly, we evaluated a wide range of outcomes, adding dimensional scores in addition to the composite IAPT recovery and reliable improvement outcomes to the analysis. As continuous measures (as well as being the outcomes the IAPT recovery index was derived from), they hold more information and likely support better prediction of patient outcomes.

Limitations

Arguably the most important limitation of this study is the fact there was no control group, meaning it is not possible to move beyond prognosis, that is, distinguish predictors of recovery due

https://doi.org/10.1017/S0033291724001582 Published online by Cambridge University Press

to treatment from those associated with recovery but not specifically related to treatment. A randomized design, with repeated baseline measures of predictors to account for measurement error, would help to resolve this. This study would also have been improved with a new test dataset to externally validate the prediction model. Whilst the out-of-sample performance can be estimated using LOO-CV, a more robust test of performance would be using data collected as part of a different study. Thirdly, the PROMPT sample may not be representative in terms of all IAPT service-user settings and population. Recruitment was only undertaken in South London, whilst IAPT is a nationwide service. There may also be differences in the recruited sample (who both engaged in the IAPT process and consented to take part in additional research interviews) compared with non-engagers and non-consenters. Including non-cases at baseline, in our primary analysis, brings challenges although the exclusion of non-cases did not strongly affect our results. Another issue is measurement error, a pervasive problem in clinical research and it is unclear how measurement bias in predictor variables may affect variable selection. Finally, recruitment did not reach the intended target meaning the sample size was smaller than planned. Because of this we implemented a clinician-led pre-analysis selection of candidate variables from all those measured. Whilst this was a change to the planned protocol, recent studies have shown selection by clinician to perform as well as machine learning or data-driven methods (e.g. Fusar-Poli et al., 2019). On the other hand, we may have excluded relevant predictor variables in this analysis. It may further be the case that symptom-level predictors are important, and these questions will be the subject of further work. There may also be factors unmeasured in our study, e.g. biomarkers which may have had predictive utility.

Implications and conclusion

The present study showed that Bayesian prediction models performed well in predicting depression and anxiety (and to a lesser extent recovery) at outcome for a cohort from a South London IAPT service. Variables prognostic of outcome were predominantly pre-treatment depression and anxiety symptoms, but with some more modest contribution from other predictors (poor outcomes predicted by lower QoL, OCD, and agoraphobia diagnoses and higher personality disorder traits). These findings have potential implications for IAPT treatment. Specifically, patients with more severe symptoms may require earlier stepping up of care, longer treatments or adjunctive/combination therapy approaches. Patients with particular comorbidities (such as OCD) may be less likely to respond to more generic therapy for depression and anxiety symptoms, raising the possibility of screening for those and offering targeted therapy instead of, or in addition to, standard IAPT approaches. The poorer outcomes in those with agoraphobia diagnoses may be indicative of a need to address avoidance behaviors. Similarly, higher scores on SAPAS screening could inform treatment choices. Clearly this work requires replication, and potential interventions based on screening would need further evaluation. Nevertheless, if sufficiently replicated in further development and validation studies, such an approach could have implications for improving future clinical outcomes by tailoring the stepped care program to better serve patients identified by those factors prognostic of worse outcome.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291724001582.

Funding statement. This study was part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care.

Competing interests. J. H. is supported by the National Institute for Health and Care Research (NIHR) Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust. In the last 3 years: R. S. declares honoraria from Janssen. A. H. Y. declares honoraria for speaking from Astra Zeneca, Lundbeck, Eli Lilly, Sunovion; honoraria for consulting from Allergan, Livanova, Lundbeck, Sunovion, and Janssen; and research grant support from Janssen. A. J. C. has received honoraria for presentations from Janssen, Otsuka, COMPASS Pathways Plc., Viatris and Medscape, honoraria for consulting from Janssen, Otsuka and COMPASS Pathways Plc., research grant support from ADM Protexin Ltd. and Beckley Psytech Ltd., and is President of the International Society for Affective Disorders (unpaid). M. H. has received funding from the Innovative Medicines Initiative for the RADAR-CNS program, a public–private pre-competitive consortium in mHealth, and his university received research funding from Janssen, Biogen, UCB, MSD, and Lundbeck. P. M. is supported by the NIHR Applied Research Collaboration (ARC; West) and the NIHR Biomedical Research Centre at University Hospitals Bristol, Weston NHS Foundation Trust, and the University of Bristol. All other authors declare no other conflicting interests.

References

Bartonicek, A., Wickham, S. R., Pat, N., & Conner, T. S. (2021). The value of Bayesian predictive projection for variable selection: An example of selecting lifestyle predictors of young adult well-being. *BMC Public Health*, *21*(1), 695. https://doi.org/10.1186/s12889-021-10690-3

Bernstein, D. P., Stein, J. A., Newcomb, M. D., Walker, E., Pogge, D., Ahluvalia, T., … Zule, W. (2003). Development and validation of a brief screening version of the Childhood Trauma Questionnaire. *Child Abuse and Neglect*, *27*(2), 169–190. https://doi.org/10.1016/S0145-2134(02)00541-0

Broadbent, E., Petrie, K. J., Main, J., & Weinman, J. (2006). The brief illness perception questionnaire. *Journal of Psychosomatic Research*, *60*(6), 631–637. https://doi.org/10.1016/j.jpsychores.2005.10.020

Brugha, T., Bebbington, P., Tennant, C., & Hurry, J. (1985). The list of threatening experiences: A subset of 12 life event categories with considerable long-term contextual threat. *Psychological Medicine*, *15*(1), 189–194. https://doi.org/10.1017/S003329170002105X

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., … Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet Psychiatry*, *3*(3), 243–250. https://doi.org/10.1016/s2215-0366(15)00471-x

Cleare, A., Pariante, C., Young, A., Anderson, I., Christmas, D., Cowen, P., … Uher, R. (2015). Evidence-based guidelines for treating depressive disorders with antidepressants: A revision of the 2008 British Association for Psychopharmacology guidelines. *Journal of Psychopharmacology*, *29*(5), 459–525. https://doi.org/10.1177/0269881115581093

Coley, R. Y., Boggs, J. M., Beck, A., & Simon, G. E. (2021). Predicting outcomes of psychotherapy for depression with electronic health record data. *Journal of Affective Disorders Reports*, *6*, 100198. https://doi.org/10.1016/j.jadr.2021.100198

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 Medicine*, *13*(1), 1. https://doi.org/10.1186/s12916-014-0241-z

Cuijpers, P., Karyotaki, E., Weitz, E., Andersson, G., Hollon, S. D., & van Straten, A. (2014). The effects of psychotherapies for major depression in adults on remission, recovery and improvement: A meta-analysis. *Journal of Affective Disorders*, *159*, 118–126. https://doi.org/10.1016/j.jad.2014.02.026

Cuijpers, P., Stringaris, A., & Wolpert, M. (2020). Treatment outcomes for depression: Challenges and opportunities. *The Lancet Psychiatry*, *7*(11), 925–927. https://doi.org/10.1016/S2215-0366(20)30036-5

Dalgard, O. S., Dowrick, C., Lehtinen, V., Vazquez-Barquero, J. L., Casey, P., Wilkinson, G., … Dunn, G. (2006). Negative life events, social support and gender difference in depression. *Social Psychiatry and Psychiatric Epidemiology*, *41*(6), 444–451. https://doi.org/10.1007/s00127-006-0051-5

Dinga, R., Marquand, A. F., Veltman, D. J., Beekman, A. T. F., Schoevers, R. A., van Hemert, A. M., … Schmaal, L. (2018). Predicting the naturalistic course of depression from a wide range of clinical, psychological, and biological data: A machine learning approach. *Translational Psychiatry*, *8*(1), 1–11. https://doi.org/10.1038/s41398-018-0289-1

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

Fusar-Poli, P., Stringer, D., M. S. Durieux, A., Rutigliano, G., Bonoldi, I., De Micheli, A., & Stahl, D. (2019). Clinical-learning versus machine-learning for transdiagnostic prediction of psychosis onset in individuals at-risk. *Translational Psychiatry*, *9*(1), Article 1. https://doi.org/10.1038/s41398-019-0600-9

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for Bayesian regression models. *The American Statistician*, *73*(3), 307–309. https://doi.org/10.1080/00031305.2018.1549100

Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge, UK: Cambridge University Press.

Gilbert, P., Clarke, M., Hempel, S., Miles, J. N. V., & Irons, C. (2004). Criticizing and reassuring oneself: An exploration of forms, styles and reasons in female students. *British Journal of Clinical Psychology*, *43*(1), 31–50. https://doi.org/10.1348/014466504772812959

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2022). *rstanarm: Bayesian applied regression modeling via Stan*. https://mc-stan.org/rstanarm/

Grant, N., Hotopf, M., Breen, G., Cleare, A., Grey, N., Hepgul, N., … Tylee, A. (2014). Predicting outcome following psychological therapy in IAPT (PROMPT): A naturalistic project protocol. *BMC Psychiatry*, *14*(1), 170. https://doi.org/10.1186/1471-244X-14-170

Gyani, A., Shafran, R., Layard, R., & Clark, D. M. (2013). Enhancing recovery rates: Lessons from year one of IAPT. *Behaviour Research and Therapy*, *51*(9), 597–606. https://doi.org/10.1016/j.brat.2013.06.004

Hemingway, H., Croft, P., Perel, P., Hayden, J. A., Abrams, K., Timmis, A., … Riley, R. D. (2013). Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ*, *346*, e5595. https://doi.org/10.1136/bmj.e5595

Hepgul, N., King, S., Amarasinghe, M., Breen, G., Grant, N., & Grey, N. (2016). Clinical characteristics of patients assessed within an improving access to psychological therapies (IAPT) service: Results from a naturalistic cohort study (predicting outcome following psychological therapy; PROMPT). *BMC Psychiatry*, *16*, 52. https://doi.org/10.1186/s12888-016-0736-6

Herdman, M., Gudex, C., Lloyd, A., Janssen, MF., Kind, P., Parkin, D., … Badia, X. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research*, *20*(10), 1727–1736. https://doi.org/10.1007/s11136-011-9903-x

Hilbert, K., Kunas, S. L., Lueken, U., Kathmann, N., Fydrich, T., & Fehm, L. (2020). Predicting cognitive behavioral therapy outcome in the outpatient sector based on clinical routine data: A machine learning approach. *Behaviour Research and Therapy*, *124*, 103530. https://doi.org/10.1016/j.brat.2019.103530

Iniesta, R., Malki, K., Maier, W., Rietschel, M., Mors, O., Hauser, J., … Uher, R. (2016). Combining clinical variables to optimize prediction of antidepressant treatment outcomes. *Journal of Psychiatric Research*, *78*, 94–102. https://doi.org/10.1016/j.jpsychires.2016.03.016

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19. https://doi.org/10.1037/0022-006X.59.1.12

Kent, P., Cancelliere, C., Boyle, E., Cassidy, J. D., & Kongsted, A. (2020). A conceptual framework for prognostic research. *BMC Medical Research Methodology*, *20*(1), 172. https://doi.org/10.1186/s12874-020-01050-7

Koeser, L., Donisi, V., Goldberg, D. P., & McCrone, P. (2015). Modelling the cost-effectiveness of pharmacotherapy compared with cognitive–behavioural therapy and combination therapy for the treatment of moderate to severe depression in the UK. *Psychological Medicine*, *45*(14), 3019–3031. https://doi.org/10.1017/S0033291715000951

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Linn, B. S., Linn, M. W., & Gurel, L. (1968). Cumulative illness rating scale. *Journal of the American Geriatrics Society*, *16*(5), 622–626. https://doi.org/10.1111/j.1532-5415.1968.tb02103.x

McAleavey, A. A. (2024). When (not) to rely on the reliable change index: A critical appraisal and alternatives to consider in clinical psychology. *Clinical Psychology: Science and Practice*, *31*(3), 351–366. https://doi.org/10.1037/cps0000203

McHugh, R. K., Whitton, S. W., Peckham, A. D., Welge, J. A., & Otto, M. W. (2013). Patient preference for psychological vs pharmacologic treatment of psychiatric disorders: A meta-analytic review. *The Journal of Clinical Psychiatry*, *74*(6), 13979. https://doi.org/10.4088/JCP.12r07757

Moran, P., Leese, M., Lee, T., Walters, P., Thornicroft, G., & Mann, A. (2003). Standardised assessment of personality – abbreviated scale (SAPAS): Preliminary validation of a brief screen for personality disorder. *British Journal of Psychiatry*, *183*(Sept), 228–232. https://doi.org/10.1192/bjp.183.3.228

Mululo, S. C. C., de Menezes, G. B., Vigne, P., & Fontenelle, L. F. (2012). A review on predictors of treatment outcome in social anxiety disorder. *Brazilian Journal of Psychiatry*, *34*, 92–100. https://doi.org/10.1590/S1516-44462012000100016

National Collaborating Centre for Mental Health. (2018). *THE IAPT manual version 5*. https://www.rcpsych.ac.uk/improving-care/nccmh/service-design-and-development/iapt

National Institute for Health and Clinical Excellence (Nice). (2011). *Generalised anxiety disorder and panic disorder in adults: Management [NG113]*. (NG113).

National Institute for Health and Clinical Excellence. (2022). *Depression in adults: Treatment and management [NG222]*. (NG222). https://www.nice.org.uk/guidance/ng222

NHS Digital. (2021). *Psychological therapies: Reports on the use of IAPT services, England – June 2021 Final including reports on the IAPT pilots and Quarter 1 data 2021–22*. https://www.gov.uk/government/statistics/psychological-therapies-reports-on-the-use-of-iapt-services-england-june-2021-final-including-reports-on-the-iapt-pilots-and-quarter-1-data-2021-2

Pavone, F., Piironen, J., Bürkner, P.-C., & Vehtari, A. (2022). Using reference models in variable selection. *Computational Statistics*, *38*(1), 349–371. https://doi.org/10.1007/s00180-022-01231-6

Piironen, J., Paasiniemi, M., & Vehtari, A. (2020). Projective inference in high-dimensional problems: Prediction and feature selection. *Electronic Journal of Statistics*, *14*(1), 2155–2197. https://doi.org/10.1214/20-EJS1711

Piironen, J., & Vehtari, A. (2017a). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, *27*(3), 711–735. https://doi.org/10.1007/s11222-016-9649-y

Piironen, J., & Vehtari, A. (2017b). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, *11*(2), 5018–5051. https://doi.org/10.1214/17-EJS1337SI

Riley, R. D., Hayden, J. A., Steyerberg, E. W., Moons, K. G. M., Abrams, K., Kyzas, P. A., … PROGRESS Group. (2013). Prognosis research strategy (PROGRESS) 2: Prognostic factor research. *PLOS Medicine*, *10*(2), e1001380. https://doi.org/10.1371/journal.pmed.1001380

Roy-Byrne, P. (2015). Treatment-refractory anxiety; definition, risk factors, and treatment challenges. *Dialogues in Clinical Neuroscience*, *17*(2), 191–206. https://doi.org/10.31887/DCNS.2015.17.2/proybyrne

Schwarzer, R., & Jerusalem, M. (1995). Generalized self-efficacy scale. In J. Weimann, S. Wright, & M. Johnson (Eds.), *Measures in health psychology: A user's portfolio. Causal and control beliefs* (pp. 35–37). Windsor, UK: NFER-NELSON.

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., … Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry*, *59*(Suppl. 20), 22–33.

Simkin, V., Hodsoll, J., & Veale, D. (2022). The relationship between symptoms of obsessive compulsive disorder and depression during therapy: A

https://doi.org/10.1017/S0033291724001582 Published online by Cambridge University Press

random intercept cross-lagged panel model. *Journal of Behavior Therapy and Experimental Psychiatry*, *76*, 101748.

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

Stochl, J., Fried, E. I., Fritz, J., Croudace, T. J., Russo, D. A., Knight, C., … Perez, J. (2020). On dimensionality, measurement invariance, and suitability of sum scores for the PHQ-9 and the GAD-7. *Assessment*, *29*(3), 107319112097686. https://doi.org/10.1177/1073191120976863

Strawbridge, R., Alexander, L., Richardson, T., Young, A. H., & Cleare, A. J. (2023). Is there a 'bipolar iceberg' in UK primary care psychological therapy services?. *Psychological Medicine*, *53*(12), 5385–5394.

Taylor, R. W., Marwood, L., Greer, B., Strawbridge, R., & Cleare, A. J. (2019). Predictors of response to augmentation treatment in patients with treatment-resistant depression: A systematic review. *Journal of Psychopharmacology*, *33*(11), 1323–1339. https://doi.org/10.1177/0269881119872194

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., … Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, *3*(8), e002847. https://doi.org/10.1136/bmjopen-2013-002847

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

https://doi.org/10.1017/S0033291724001582 Published online by Cambridge University Press