Article:
Moggia, D. orcid.org/0000-0001-6321-4450, Saxon, D. orcid.org/0000-0002-9753-8477, Lutz, W. orcid.org/0000-0002-5141-3847 et al. (2 more authors) (2023) Applying precision methods to treatment selection for moderate/severe depression in person-centered experiential therapy or cognitive behavioral therapy. Psychotherapy Research, 34 (8). pp. 1035-1050. ISSN 1050-3307

https://doi.org/10.1080/10503307.2023.2269297

Routledge
Taylor & Francis Group

Psychotherapy Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tpsr20

Applying precision methods to treatment selection for moderate/severe depression in person-centered experiential therapy or cognitive behavioral therapy

Danilo Moggia, David Saxon, Wolfgang Lutz, Gillian E. Hardy & Michael Barkham

To cite this article: Danilo Moggia, David Saxon, Wolfgang Lutz, Gillian E. Hardy & Michael Barkham (02 Nov 2023): Applying precision methods to treatment selection for moderate/severe depression in person-centered experiential therapy or cognitive behavioral therapy, Psychotherapy Research, DOI: 10.1080/10503307.2023.2269297

To link to this article: https://doi.org/10.1080/10503307.2023.2269297

Psychotherapy Research, 2023
https://doi.org/10.1080/10503307.2023.2269297

RESEARCH ARTICLE

Applying precision methods to treatment selection for moderate/severe depression in person-centered experiential therapy or cognitive behavioral therapy

DANILO MOGGIA [1], DAVID SAXON [2], WOLFGANG LUTZ [1], GILLIAN E. HARDY [2], & MICHAEL BARKHAM [2]

[1]*University of Trier, Trier, Germany &* [2]*Clinical and Applied Psychology Unit, Department of Psychology, University of Sheffield, Sheffield, UK*

(Received 8 March 2023; revised 30 September 2023; accepted 3 October 2023)

Abstract

Objective To develop two prediction algorithms recommending person-centered experiential therapy (PCET) or cognitive–behavioral therapy (CBT) for patients with depression: (1) a *full data model* using multiple trial-based and routine variables, and (2) a *routine data model* using only variables available in the English NHS Talking Therapies program.

Method Data was used from the PRaCTICED trial comparing PCET vs. CBT for 255 patients meeting a diagnosis of moderate or severe depression. Separate full and routine data models were derived and the latter tested in an external data sample.

Results The *full data model* provided the better prediction, yielding a significant difference in outcome between patients receiving their optimal vs. non-optimal treatment at 6- (Cohen's $d = .65$ [.40, .91]) and 12 months ($d = .85$ [.59, 1.10]) post-randomization. The *routine data model* performed similarly in the training and test samples with non-significant effect sizes, $d = .19$ [−.05, .44] and $d = .21$ [−.00, .43], respectively. For patients with the strongest treatment matching ($d \geq 0.3$), the resulting effect size was significant, $d = .38$ [.11, 64].

Conclusion A treatment selection algorithm might be used to recommend PCET or CBT. Although the overall effects were small, targeted matching yielded somewhat larger effects.

Keywords: precision methods; personalized mental health; machine learning; intersectionality; depression; person-centered experiential therapy; cognitive behavioral therapy

Clinical or Methodological Significance of this Article: The results provide insight into sociodemographic and clinical characteristics that can be used to target patients' profiles to recommend either PCET or CBT for depression. Patients benefiting from PCET were more likely to be employed females or unemployed males, characterized by a more *punitive self* (i.e., feelings of guilt and being criticized by others), and impairment in close relationships but also *personal meaning* (i.e., a sense of purpose in life; greater expectancy of improvement from PCET). Patients benefiting from CBT were more likely to be unemployed females or employed males, characterized by *letting oneself down* (i.e., feelings of worthlessness), no impairment in close relationships but *being resilient* (i.e., feelings of being in control of life; being able to bounce back). Optimal assignment is more important the greater the predicted difference between treatments.

Depression is one of the most prevalent mental health problems (Liu et al., 2020), associated with high costs in healthcare systems (Kessler, 2012), with psychological therapies being one of the first-line treatments recommended by several clinical guidelines (Zafra-Tanaka et al., 2019). For example,

Correspondence concerning this article should be addressed to Michael Barkham Clinical and Applied Psychology Unit, Department of Psychology, University of Sheffield, Cathedral Court, 1 Vicar Lane, Western Bank, Sheffield, S1 2LT, UK. Email: m.barkham@sheffield.ac.uk

guidelines from the English National Institute for Health and Care Excellence (NICE, 2022), which are also applicable in Wales, recommend a number of psychological therapies for depression, including cognitive–behavioral therapy (CBT), behavioral activation, interpersonal psychotherapy, short-term psychodynamic psychotherapy and person-centered experiential therapy (PCET; also known as person-centered experiential counseling for depression [PCE-CfD]; Murphy, 2019), among other psychosocial and pharmacological interventions. All these psychological therapies are considered first-line treatments for subthreshold, mild, moderate, and severe depression. However, based on the interpretation of cost-effectiveness studies and implementation factors (e.g., therapists' availability), CBT is the first-choice intervention in the English NHS Talking Therapies for Anxiety and Depression program, previously known as the Improving Access to Psychological Therapies (IAPT) program (National Collaborating Centre for Mental Health, 2021).

NHS Talking Therapies is a national-level program across England that offers psychological treatments based on a stepped care model (Clark, 2018) in which Step two corresponds to the entry point into the system involving low-intensity psychoeducational interventions delivered by psychological wellbeing practitioners (PWPs). Step three comprises high-intensity treatments, namely approved psychological therapies as listed previously, delivered by appropriately trained therapists or counselors (i.e., high-intensity practitioners). These therapies can be offered to patients with depression based on the principle of shared decision making following a discussion of a patient's preferences (NICE, 2022). All patients in both steps complete a battery of routine outcome measures at each attended session. These measures are used by PWPs in Step 2 to identify patients who are not showing a significant response to Step 2 treatment and require stepping up to a Step 3 high-intensity treatment.

Although a choice of Step 3 therapies may be available, CBT is the most frequently delivered therapy, accounting for 58% of referrals for therapy in 2018–2019. The second most frequently delivered therapy, PCET, accounted for 36% of referrals in the same time period (Moller et al., 2019). These differences between referral rates are due to variability in the personnel available to deliver each therapy. As patients are seen within their local NHS Talking Therapy service, their choice is limited by the available resources. Hence, while PCET receives lower numbers of referrals, it plays a significant role in delivering on the overarching agenda of providing patient choice and increasing throughput of patients. However, various studies suggested that CBT might not be superior to other modalities of psychological therapies for the treatment of depression (e.g., Cuijpers et al., 2013; King et al., 2014). These findings prompted a formal examination of their comparative efficacy through a pragmatic randomized non-inferiority trial embedded within an NHS Talking Therapy service.

The PRaCTICED trial assessed the non-inferiority of PCET vs. CBT in the treatment of moderate or severe depression, assessed at six- and 12-months post-randomization, for patients attending NHS Talking Therapies services (Barkham et al., 2021). A total of 510 participants were recruited and randomly assigned to PCET ($n = 254$) or CBT ($n = 256$). Results showed PCET to be non-inferior to CBT at six months post-randomization in the intent-to-treat (Cohen's $d = -.03$ [−.23, .17]) and per-protocol samples ($d = .09$ [−.14, .32]). However, there was a significant difference at 12 months post-randomization favoring CBT in both the intent-to-treat ($d = .27$ [.05, .49]) and per-protocol analyses ($d = .34$ [.09, .60]). Hence, findings yielded differential outcomes dependent on the time point of comparison, with non-inferiority of PCET compared with CBT supported in the short-term, but results favoring CBT when measured more distally from end of therapy (at 12 months post-randomization). Therapist effects were 0.2%, implying that the variability among therapists had negligible influence on the overall efficacy of the therapeutic interventions provided.

Given the trial findings showed either CBT or PCET to be viable treatment options for patients in the short term, a significant advance in delivering improved outcomes might be achieved by adopting initiatives derived from the field of precision mental health care (Cohen et al., 2021; Lutz, Schwartz, et al., 2022). In brief, precision mental health aims to offer or allocate patients to the treatment that best fits their presenting profile and individual characteristics. It works by identifying variables that function as moderators or predictors of treatment outcome (Kraemer et al., 2001), thereby identifying subpopulations of differential treatment responders. The application of precision research methods to RCT designs is additionally informative as RCTs only compare averages of patients, thereby yielding heterogeneous treatment effects in specific patient subpopulations, meaning that differential treatment responses (i.e., individual differences) remain hidden.

One method developed to assess which of two interventions would be expected to result in a better outcome for a specific patient is the personalized advantage index (PAI; DeRubeis et al., 2014). The PAI is the difference between predicted

outcomes of two different models (one for each treatment), yielding a score representing the magnitude by which one treatment is predicted to outperform the other and thereby be recommended as the optimal treatment. The PAI has been evaluated in retrospective studies showing that better outcomes were obtained by patients receiving their optimal rather than non-optimal treatment (Deisenhofer et al., 2018; Schwartz et al., 2021).

Specifically focusing on PCET, Delgadillo and Gonzalez Salas Duhne (2020) used the PAI to develop a model recommending either PCET or CBT for depression using a naturalistic sample from the NHS Talking Therapies program. They predicted a categorical outcome (reliable and clinically significant improvement [RCSI] in the PHQ-9) using routinely collected variables as predictors (i.e., sociodemographic and clinical background variables, depression and anxiety levels, as well as social and work impairment). Analyzed retrospectively, patients who received their optimal treatment obtained better outcomes (RCSI rate of 62.5%) than those who received their non-optimal treatment (RCSI rate of 41.7%). However, relying on a naturalistic dataset raises issues with internal validity as treatment allocation was not randomized. In addition, transforming a continuous outcome variable (PHQ-9) into a dichotomous variable (RCSI) likely loses information and decreases the sensitivity to detect meaningful associations, reduces statistical power and thereby affecting the reliability and generalizability of the predictive model.

In reviewing the body of literature implementing the PAI, we noted three features that could yield a more refined application of this method. First, diverse studies have tended to use sociodemographic variables which are tested simultaneously in prediction models and the statistically significant effects become the focus of analyses (e.g., Finegan et al., 2018). Concerning social determinants, this approach is problematic because it does not consider that each individual's social positions may interact in complex ways (Cairney et al., 2014). Less common has been the exploration of the interaction between sociodemographic variables. Second, measures within prediction models have invariably used total scores and, while not necessarily problematic, these fail to take account of variability in response to individual items in that some items carry more weight than others (e.g., Fried & Nesse, 2015). Less frequent has been the inclusion of item-level data. Third, in most cases, the predicted outcome is the most proximal score to the end of treatment, while follow-up (i.e., longer-term) assessment scores are rarely examined.

In considering these three features, we proposed a revised research strategy to maximize the sensitivity of our prediction modeling. We reasoned that the inclusion of interaction terms of socio-economic variables and individual items from instruments when machine learning algorithms are implemented, combined with data drawn from longer-term outcomes, would allow the identification of more complex patterns and non-linear associations using many variables with potentially greater predictive strength in terms of enduring (i.e., longer-term) outcomes.

Informed by these methodological issues, the current research aimed to develop a treatment selection algorithm for depression (recommending PCET or CBT). As many of the studies published applying the PAI method are based on RCTs, and the translation of these models into clinical practice is difficult (due to differences in variables collected, number of predictors used), we aimed to test two sets of predictive models: the first utilized the whole set of variables, comprising measures collected exclusively for the purposes of the PRaCTICED trial *and* including variables and measures routinely collected by NHS Talking Therapies (termed *full data model*; Table I, Panel A), and a second using *only* the routinely collected variables and measures collected as standard (termed *routine data model*; Table I, Panel B). Only this latter routine battery of variables and measures is available in the wider national delivery program of the NHS Talking Therapies.

The *full data model* was considered a model of "maximum potential," because it was developed using a training sample that contained a comprehensive set of variables related to patients' intake characteristics collected as part of the PRaCTICED trial and served as a reference point. In contrast, the *routine data model* was developed for the same sample of patients using only those variables collected in the national implementation with the aim of assessing the performance and generalizability of the predictive models in real-world scenarios, thereby enabling a direct comparison between the full and routine models and serving as a validation step for the practicality and effectiveness of the *routine data model.*

Table I. Summary of relationship between clinical samples and datasets.

	Clinical samples	
	Training samples from PRaCTICED trial	External test sample from routine practice
Datasets	A. Full data model: Comprehensive trial data including routine data	[Not applicable]
	B. Routine data model: Only variables from routine dataset within the trial	C. Only variables from non-trial routine dataset

Each model was used to test the advantage of matching patients to their optimal treatment compared with their non-optimal treatment. We expected the *full data model* would result in a more precise classification of patients; therefore, in more precise prognoses. However, it was only possible, and appropriate, to check the generalization of the predictive model based on the routine data (i.e., the r*outine data model*) as this is the only data collected at a national level. Accordingly, we conducted such a test on a routine dataset from an external independent test sample within the same NHS Talking Therapies service (Table I, Panel C).

Method

PRaCTICED Trial Design

The design was a pragmatic non-inferiority trial embedded in the Sheffield NHS Talking Therapies service in England. Participants were older than 17 years, meeting moderately-severe or severe depression criteria on the Clinical Interview Schedule-Revised (CIS-R; Lewis, 1994). Participants were excluded if they presented with an organic condition, a previous diagnosis of personality disorder, bipolar disorder, schizophrenia, drug or alcohol dependency, an elevated clinical risk of suicide, or a long-term physical condition. Eligible participants were randomly assigned to receive PCET or CBT. Patients were nested in therapists within treatments, with 18 counselors delivering PCET and 32 therapists delivering CBT. The intervention professionals were trained PCET counselors and CBT therapists who followed the NHS Talking Therapies service delivery model. They received special training for the trial and regular individual and group supervision sessions led by experienced qualified counselors and senior therapists. Additionally, treatment adherence was assessed by experienced and independent trainers using standard adherence scales.

The primary outcome measure was the Patient Health Questionnaire–9 (PHQ-9; Kroenke et al., 2001), while the General Anxiety Disorder–7 (GAD-7; Spitzer et al., 2006), the Beck Depression Inventory–II (BDI-II; Beck et al., 1996), the Clinical Outcomes in Routine Evaluation–Outcome Measure (CORE-OM; Evans et al., 2002) and the Work and Social Adjustment Scale (WSAS; Mundt et al., 2002) were used as secondary outcomes. Participants were assessed at screening, six and 12-months post-randomization. At screening they were also assessed with the Connor-Davidson Resilience Scale (CD-RISC; Connor & Davidson, 2003; Davidson, 2018) and a credibility/expectancy questionnaire for each treatment approach (Devilly & Brokovec, 2000). Because the trial was embedded in the routine service, patients also completed the NHS Talking Therapies-mandated outcome measures at each attended therapy session (i.e., PHQ-9, GAD-7, and WSAS). Patients could receive a maximum of 20 sessions in either intervention. Ethics approval was granted by the English Health Research Authority (Research Ethics Committee 14/YH/0001).

Current Study Design

The current study was structured in two phases. The first phase employed a training sample derived from the PRaCTICED trial dataset re-analyzed to estimate a prediction model for treatment allocation (recommending PCET or CBT) based on the PAI. This training sample comprised 255 patients from the original trial sample who had completed their scheduled treatment and had all the instruments administered at screening and, crucially, completed assessments at 12-months post randomization. There were 46 therapists (29 therapists delivering CBT and 17 counselors delivering PCET) with a mean (*SD*) of 5.67 (7.16) patients per therapist (range 1-35). The sample was split into two subsamples: patients who received PCET ($n = 140$) and those who received CBT ($n = 115$). Based on these two subsamples, data analyses for selecting predictors, prediction models of treatment outcome, and the PAI were estimated. In this phase, we wanted to assess two models. First, the *full data model*, including all variables and measures collected for the PRaCTICED trial, variables from the routine service data, and, following the methodological considerations previously explained, including item-level information, sociodemographic variables and their interactions. Second, the *routine data model* based only on the NHS Talking Therapies routinely collected measures and variables for later testing in an external test sample.

The second phase determined the generalizability of the *routine data model* in an external test sample from NHS Talking Therapies. This sample ($n = 255$) was selected with propensity score matching (PSM) procedures drawn from a larger patient sample of patients ($n = 6{,}049$) receiving high-intensity treatment, either PCET or CBT, during the same time period of the PRaCTICED trial and provided by the Sheffield Health and Social Care NHS Foundation Trust (see data analysis section below).

We followed the guidelines of the "Transparent Reporting of Multivariable Prediction Model for Individual Prognosis or Diagnosis" (TRIPOD; Collins et al., 2015; Moons et al., 2015) to conduct and report the current study.

Treatments

Person-centered experiential therapy (PCET). PCET (also called person-centered experiential counseling for depression; Hill, 2011; Murphy, 2019; Sanders & Hill, 2014) is a form of treatment derived from the humanistic-experiential approaches to psychological therapies (Duffy et al., 2023; Elliott et al., 2021). It integrates Rogers' person-centered model (Rogers, 2003) with emotion-focused therapy components, particularly that of process-guiding (Greenberg & Watson, 2006; Murphy, 2019). It aims to help patients access underlying feelings, make sense of them, and draw on the new meanings that emerge to make positive changes in their lives. A therapeutic attitude based on empathy, therapists' authenticity, unconditional positive reward, and working with experiential techniques are the prominent features (Sanders & Hill, 2014). A manual was designed and adopted for the trial (PRaCTICED Trial Team, 2014b).

Cognitive–behavioral therapy (CBT). Beck's cognitive therapy for depression (Beck et al., 1979) was provided by high-intensity CBT practitioners. CBT delivery was standardized by adopting a treatment manual (PRaCTICED Trial Team, 2014a). The main characteristics are a therapeutic attitude based on collaborative empiricism and interventions based on psychoeducation, behavioral activation, and identification and change of maladaptive patterns of thinking.

Measures, Sociodemographic, and Clinical Background Variables

Routinely collected NHS talking therapies measures and variables

Sociodemographic and clinical background variables. We used available information about age, gender (female, male, or other), ethnicity (white or non-white), employment status (employed full-time, employed part-time, homemaker, long-term sick, or disabled receiving incapacity benefits, long-term sick, or disabled without receiving incapacity benefits, retired, student, unpaid voluntary work, unemployed) and index of multiple deprivations (IMD) of the patients. The IMD is an index of socio-economic deprivation used by the UK government, which is derived from a patient's address and neighborhood (Noble et al., 2019). It produces deciles ranging from 1 to 10, with 1 indicating extreme deprivation and 10 less deprived areas. Additionally, we counted the number of low-intensity treatment sessions the patients received before the screening.

Patient Health Questionnaire–9 (PHQ-9; Kroenke et al., 2001). The PHQ-9 is a self-administered questionnaire comprising 9 items which assess components of major depressive disorder (MDD) diagnosis criteria. The items are answered on a four-point Likert scale ranging from 0 ("not at all") to 3 ("nearly all day"). PHQ-9 total score is determined by summing up item ratings, with scores of 5, 10, 15, and 20 representing mild, moderate, moderately severe, and severe depression, respectively.

Generalized Anxiety Disorder–7 (GAD-7; Spitzer et al., 2006). The GAD-7 is a seven-item self-administered questionnaire designed to assess generalized anxiety disorder (GAD) symptoms according to the DSM-IV diagnosis criteria. Patients answer the items on a four-point Likert scale ranging from 0 ("not at all") to 3 ("nearly every day"). Total scores are calculated by summing all item ratings. Total scores of 5, 10, and 15 correspond to mild, moderate, and severe levels of anxiety, respectively.

Work and Social Adjustment Scale (WSAS; Mundt et al., 2002). The WSAS is a five-item self-administered scale developed to measure social and work impairment. Its items are answered on a nine-point Likert scale ranging from 0 ("not at all") to 8 ("very severely"). The total score is calculated by summing up all item ratings. The maximum score is 40, with higher scores indicating higher impairment.

Practiced trial additional measures

Clinical Interview Scheduled–Revised (CIS-R; Lewis, 1994). The CIS-R is a standardized and computer-delivered clinical interview developed for use in general practice and community settings which delivers a diagnosis, thereby ensuring standardization of responses to self-reported items. It assesses 14 common mental health disorders and items are scored on a 0–4 scale according to the frequency and severity of the symptoms presented during the week before the interview. The scores obtained from each of the reported symptoms are averaged to obtain a total score. Due to the computerized automation of this diagnostic tool, item-level information was not available.

Beck Depression Inventory–II (BDI-II; Beck et al., 1996). The BDI-II is a self-report questionnaire, containing 21 items answered on a four-point

Likert scale with a range from 0 to 3. It measures depressive symptoms and their severity according to the DSM-IV diagnostic criteria for dysthymia and MDD. The higher the scores, the higher the intensity of depressive symptoms. A score of 13 or less denotes minimal levels of depression.

Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM; Evans et al., 2002). The CORE-OM is a self-reported questionnaire comprising 34 items answered on a five-point Likert scale ranging from "not at all" to "most or all the time." The questionnaire is designed to assess psychological distress, the higher the scores, the higher the problems reported. The items are grouped into four domains: subjective well-being, symptoms and problems, life functioning, and risk of self-harm or harm to others. The higher the scores, the higher the psychological distress. A clinical cut-off of 10 is suggested to differentiate between clinical and non-clinical populations (Connell et al., 2007).

Connor-Davidson Resilience Scale (CD-RISC; Connor & Davidson, 2003). The CD-RISC is a self-reported questionnaire aimed to assess personal resilience. It comprises 25 items answered on a Likert scale from 0 ("not true at all") to 4 ("true nearly all the time"). The item ratings are summed up to obtain a total score. The higher the score, the higher the adaptability and resilience. A mean of 80.4 (SD = 12.8) has been reported for the general population, and a mean of 57.1 (SD = 13.3) in a major depressive disorder outpatient sample (Davidson, 2018).

Treatment Credibility/Expectancy Questionnaire (CEQ; Devilly & Brokovec, 2000). The CEQ is a six-item self-reported questionnaire designed to assess cognitively based credibility (four items) and affectively based expectancy (two items) of treatment. Four items (from the credibility and expectancy items-subgroup) are answered on a Likert scale from 1 ("not at all") to 9 ("very"). Two items about how much improvement the patient thinks and feels he/she will obtain are answered on a percentile scale. In the PRaCTICED trial, all patients answered the questionnaire prior to randomization, assessing the credibility and their expectancy of both CBT and PCET.

Data Analyses

First, the characteristics of the sample selected were explored with descriptive statistics. As some patients omitted certain items of the questionnaires and some of the sociodemographic information and PHQ-9 total scores at six and 12 months were missing for some cases, we performed missing completely at random (MCAR) analyses to determine that missing values were produced randomly. Once checked, data imputation was conducted with a random forest (RF) algorithm (Stekhoven & Buehlmann, 2012). Afterwards, the sample was split into the PCET and CBT subsamples and predictive models of outcomes were derived for each therapy modality.

Predictors for the *full* and *routine data model* were selected using RF algorithms (Liaw & Wiener, 2002) and a branch-and-bounds algorithm (Furnival & Wilson, 1974), respectively (see below). After selecting the variables, they were introduced into linear regression analyses using a stepwise algorithm for variable retention and controlling for PHQ-9 total score at screening. Additionally, a leave-one-out cross-validation (LOOCV) procedure was implemented to prevent overfitting. At this point, two sets of models were computed, one comprising item scores from trial measures, including routine measures and sociodemographic variables and their interactions (*full data model*), and another one with only the NHS Talking Therapies total scores from the mandated routine outcome measures and sociodemographic variables (*routine data model*) collected within the same trial.

Once the models were estimated, they were applied in each treatment modality to obtain a factual and a counterfactual prediction. The predicted outcomes from the PCET and CBT models were subtracted to obtain the PAI and classify the patients as those who received their *optimal* and *non-optimal* treatment. We assessed the differential prediction (PAI) performance with a repeated-measures analysis of variance (RM-ANOVA) and checked the differences between patients who received their optimal and non-optimal treatment on the PHQ-9 scores at six and 12-months. Additionally, we applied the *routine data model* to the routine data from the trial patients (PHQ-9 from first and last session) as a transition toward testing on an external routine data sample.

Finally, to test the *routine data model* with the external sample, we performed a PSM procedure to select patients with similar characteristics to the trial patients from the local NHS Talking Therapies service and applied the previously computed prediction models. We followed the same procedure to classify the patients based on PAI scores in this sample and assessed its performance. Since the PAI is a difference score, larger scores indicate larger predicted effect size differences between the optimal and

non-optimal group. It is expected that with an increasing effect size difference, the predictions become more clinically relevant (e.g., Cohen et al., 2021). Therefore, it is reasonable to hypothesize that patients with higher PAI scores will experience greater benefits from receiving their optimal treatment compared to patients with lower PAI scores. Following this, we evaluated the performance of the *routine data model* in a subgroup of patients with at least a predicted effect size difference of $d \geq .3$ (e.g., Lutz, Deisenhofer, et al., 2022).

Data pre-processing and data imputation. After performing MCAR analyses and verifying that missing values were produced randomly, missing values imputation was carried out for PHQ-9 total score at six (13% missingness) and 12 (15.3% missingness) months, as well as missing item scores of the instruments at screening (11.4% missingness) and some of the sociodemographic information (i.e., employment status [16.5% missingness] and ethnicity [4.3% missingness]). This was performed with the R package "missForest" 1.4 (Stekhoven, 2013) on R (R Core Team, 2021) and Rstudio (RStudio Team, 2021).

Once data were imputed, we re-coded the variable "employment status" in a four-level categorical variable called "remunerated activity status" (RAS: doing a daily remunerated activity, doing a daily non-remunerated activity, not doing a daily activity but receiving a remuneration, or not doing a daily activity and not receiving remuneration).

Afterwards, the sample was divided into the PCET and CBT datasets. In each dataset, all continuous variables were z-standardized, dichotomous variables were dummy coded at −½, +½, and categorical variables were dummy coded at $1 - 1/m$, $-1/m$ (m being the number of categories; Kraemer & Blasey, 2004). Additionally, two-way and three-way interactions were computed between the sociodemographic variables (i.e., gender, ethnicity, IMD, and RAS). All these transformations were performed with IBM SPSS 28 (IBM Corp, 2021).

Variables selection. Following Schwartz et al. (2021), to select predictors for the *full data model* (using item scores, sociodemographic variables, and their interactions), RF analyses were implemented in each treatment modality using JASP 0.16.1 (JASP Team, 2022). We computed seven RF models in each treatment modality, one for each questionnaire or group of variables as follows: (1) Sociodemographic, clinical background variables, and CIS-R total score; (2) GAD-7 items; (3) WSAS items; (4) BDI-II items; (5) CORE-OM items; (6) CD-RISC items; and (7) Credibility and Expectancy items.

We configured the RFs dividing each treatment modality into training (80%) and test (20%) subsamples. A total of 129 variables were included in the analyses (see the complete list of predictors in Supplementary Tables 1 and 2). Information from PHQ-9 was not included in the RF analyses because it was later included in the regression models (by forced entry) to control for depression levels at screening. In contrast to the criteria employed by Schwartz et al. (2021), where predictors accumulating 90% importance in the model were selected, we opted for a selection criterion based on positive scores in "mean decrease in accuracy." This criterion was chosen given its simplicity and robustness (Genuer et al., 2010; Strobl et al., 2007; see Supplemental Material for further details). Nevertheless, using RF to select a subset of features from a small number of variables (consistent with the routine data from NHS Talking Therapies program) may lead to problems of overfitting and biased importance because variables that are more strongly correlated with the target variable may dominate the importance rankings, while potentially important variables with weaker correlations may be overlooked (e.g., Tang et al., 2018). Thus, we used a branch-and-bounds algorithm with the R package "leaps" (Lumley, 2022) to select predictors for the routine data model (using total scale scores and sociodemographic variables). This algorithm was applied to each treatment modality (CBT and PCET). Branch-and-bounds allows the selection of the best subset of variables in linear regression. Its results do not depend on a penalty model for model size, which allows controlling for multicollinearity based on all possible combinations of variables from the original set. As with RF analyses, the PHQ-9 total score was not included in the branch-and-bounds algorithms. For further details on the variable selection procedures see the Supplemental Material.

Estimation of Regression Models

We estimated two sets of regression models, *full* and *routine data models*. Each set contained two regression models, one for each treatment modality. Each of these models was estimated regressing the variables previously selected on the PHQ-9 total score at 12 months, controlling for the PHQ-9 total score at screening. A stepwise algorithm for variable retention (based on Akaike's Information Criterion [AIC]) was implemented (Heinze et al., 2018). Additionally, a LOOCV procedure was used for cross-validation

purposes and over-fitting prevention (Efron, 1982). The LOOCV procedure creates n models in each treatment modality dataset (n being the sample size of the dataset) with a sample size of $n-1$. In that way, each model was estimated without any information about the patient whose scores were predicted. Thus, the predictions are considered to contain a small or null bias. In the end, the n model estimates are averaged to obtain a single model for each treatment modality. These procedures were performed with the R package "caret" (Kuhn, 2008).

Estimation of the PAI. With the previously estimated models in each set (*full* and *routine data models*), we obtained a factual prediction for each patient in each treatment modality (CBT and PCET). Next, we applied the regression model obtained with one treatment modality to the other to obtain a counterfactual prediction for each patient. In that way, we generated two predictions for each patient, one for each treatment. The PAI was then estimated as the difference between the two predictions (the predicted outcome for PCET was subtracted from the predicted outcome for CBT). Due to lower scores on the outcome variable (PHQ-9) indicating better treatment outcomes, a PAI > 0 indicated that PCET was recommended while a PAI < 0 indicated a recommendation of CBT. Afterwards, each patient was classified as having received their optimal or non-optimal treatment if they were actually treated with the treatment recommended by the PAI or not.

Assessment of the differential prediction performance. With RM-ANOVA, we tested the differences in PHQ-9 total scores at six and 12 months between patients who received their optimal and non-optimal treatment, controlling for PHQ-9 total scores at screening. Effect sizes were computed as standardized differences between scores (Cohen's d).

Test Sample Selection

From the external NHS Talking Therapies dataset, we selected all patients who attended at least two high-intensity sessions ($n = 6{,}049$). We then selected all patients with no missing values in either of their variables (listwise deletion), yielding a sample of 4,084 patients. We transformed the variables of this sample following the same procedures previously described for the training sample. With this test sample pool and the training sample, we applied a PSM algorithm to select a study test sample with similar characteristics to the training trial (matching ratio 1:1). To allow heterogeneity in the matching procedure (increasing the external validity of the PAI), a caliper was set in .20. The co-variates considered were age, gender, ethnicity, IMD, RAS, PHQ-9, GAD-7, and WSAS. The PSM was set without replacement (which increases the heterogeneity of the selection). Once the test sample was selected ($n = 255$), its characteristics were compared with the training sample characteristics using multivariate analysis of variance (MANOVA) applying Pillai's trace, and Chi-squared (χ^2) test.

Routine Data Model Evaluation

Finally, the regression models previously computed with the training sample (PRaCTICED trial) were applied to the test sample to compute the PAI. An analysis of covariance (ANCOVA) was applied to evaluate the differences in PHQ-9 total scores at last therapy session between patients who received their optimal and non-optimal treatment, controlling for potential differences on pre-treatment scores that might exist between groups. Adjusted effect sizes for ANCOVA were computed following the method proposed by Hedges et al. (2023). The same procedure was applied to evaluate the performance of the *routine data model* in the subgroup of patients with at least a predicted effect size difference of $d \geq .3$.

Results

Samples Characteristics

Training sample. The training sample comprised 255 patients who completed the scheduled treatment and all instruments administered at screening: 149 (58.4%) were women, 106 (41.6%) were men, and 14 (5.5%) were non-White. Their age mean was 40.51 ($SD = 13.03$) years, with an IMD mean of 5.81 ($SD = 3.28$). Regarding their RAS, 144 (56.5%) were conducting a paid activity as their primary occupation, 28 (11%) were not engaged in a primary activity but were receiving some income, 15 (5.9%) were engaged in a daily activity without receiving a salary, and 26 (10.2%) had neither activity nor income. The patients received 1.39 ($SD = 1.14$) sessions of low-intensity treatment on average before the screening. At intake they had a mean score of 31.42 ($SD = 8.05$) on the CIS-R, 18.62 ($SD = 4.09$) on the PHQ-9, 12.89 ($SD = 4.38$) on the GAD-7, 25.0 ($SD = 7.48$)

on the WSAS, 36.21 ($SD = 8.60$) on the BDI-II, 21.84 ($SD = 4.74$) on the CORE-OM, and 40.10 ($SD = 13.89$) on the CD-RISC.

Test sample. The test sample comprised 255 patients selected with PSM from the NHS Talking Therapies external dataset. There were no differences between this sample and the training sample in the proportion of women ($n = 158$, 62%) and men ($n = 97$, 38%), $\chi^2(1) = .66$, $p = .41$, the proportion of non-White ($n = 15$, 5.9%) and White ($n = 240$, 94.1%) patients, $\chi^2(1) = .04$, $p = .85$, and patients' age ($M = 39.87$, $SD = 14.22$), $V = .17$, $F(1;\ 508) = .28$, $p = .60$. However, there were statistically significant differences in RAS between both samples, with the test sample having a higher proportion of patients who were not engaged in an activity nor received an income ($n = 53$, 20.8%), $\chi^2(3) = 11.95$, $p < .01$ (Cramer's $V = .15$). There were also statistically significant differences in the number of low intensity sessions ($M = 1.82$, $SD = 1.48$), IMD ($M = 4.51$, $SD = 3.12$), PHQ-9 ($M = 16.54$, $SD = 5.10$), GAD-7 ($M = 14.16$, $SD = 4.22$) and WSAS ($M = 21.62$, $SD = 8.74$) scores at intake. Patients from the test sample presented a lower IMD on average, Pillai's $V = .17$, $F(1,\ 508) = 20.85$, $p < .001$, $\eta_p^2 = .04$, scored lower on the PHQ-9, $F(1,\ 508) = 25.59$, $p < .001$, $\eta_p^2 = .05$, and WSAS, $F(1,\ 508) = 21.99$, $p < .001$, $\eta_p^2 = .04$, but higher on the GAD-7, $F(1,\ 508) = 205.83$, $p < .001$, $\eta_p^2 = .02$, and attended more low intensity sessions, $F(1,\ 508) = 25.59$, $p < .001$, $\eta_p^2 = .05$. Nevertheless, all these differences were small with weak effect sizes (Cramer's $V < .20$ and $\eta_p^2 < .05$).

Variables Selection

For the *full data model*, the results from the seven RF algorithms are shown in Supplementary Tables 1 and 2. They present each model's variables ranked from the highest to the lowest importance (according to the mean decrease in accuracy) for each group of variables. For the PCET dataset (Supplementary Table 1), 85 variables (65.89%) were selected as potential predictors while for the CBT dataset (Supplementary Table 2), 64 variables (49.61%) were selected.

For the *routine data model*, the branch-and-bounds algorithm suggested gender and WSAS as predictors for the PCET dataset (Mallow's $Cp = 2.66$), and number of low intensity sessions, WSAS, and RAS for the CBT dataset (Mallow's $Cp = 1.75$).

Estimation of Regression Models

For the *full data model*, 85 variables were introduced into the regression analysis for the PCET dataset and 11 were retained by the stepwise algorithm, obtaining a model with an $R^2 = .35$ (with PHQ-9 total score at screening accounting for $R^2 = .068$). For the CBT dataset, 64 variables were introduced into the regression analysis and 13 were retained by the stepwise algorithm, obtaining a model with an $R^2 = .47$ (with PHQ-9 total score at screening accounting for $R^2 = .061$). The parameters of both models are shown in Supplementary Table 3.

For the *routine data model*, three variables were introduced into the regression analysis for the PCET dataset and all were retained by the stepwise algorithm, obtaining a model with an $R^2 = .16$ (with PHQ-9 total score at screening accounting for $R^2 = .09$). For the CBT dataset, 4 variables were introduced into the regression analysis and all were retained by the stepwise algorithm, obtaining a model with an $R^2 = .24$ (with PHQ-9 total score at screening accounting for $R^2 = .08$). The parameters of both models are presented in Supplementary Table 4.

Estimation of the PAI

For the *full data model*, the PAI mean was -.014 ($Mdn = -.18$, $SD = 4.23$, range = −13.15–11.24) indicating a slight advantage to CBT in predicted outcomes. PCET was recommended to 127 (49.8%) patients, while CBT was recommended to 128 (50.2%). Of the 255 patients, 132 (51.8%) received their optimal treatment (72 in PCET and 60 in CBT) and 123 (48.2%) their non-optimal treatment (55 in PCET and 68 in CBT). For the *routine data model*, the PAI mean was -.003 ($Mdn = -.11$, $SD = 1.72$, range = −3.70–9.71) indicating a small advantage to CBT in predicted outcomes. PCET was recommended to 127 (49.8%) patients, while CBT was recommended to 128 (50.2%) patients. Of the 255 patients, 130 (51%) received their optimal treatment (71 in PCET and 59 in CBT) and 125 (49%) their non-optimal treatment (56 in PCET and 69 in CBT).

Assessment of the Differential Prediction Performance

Based on the *full data model*, we obtained a significant effect of the differential prediction (i.e., having received the optimal or non-optimal treatment) on PHQ-9 total scores at six and 12 months: Greenhouse–Geisser's $\varepsilon = 1.00$, $F(1.99,\ 504.38) = 29.37$,

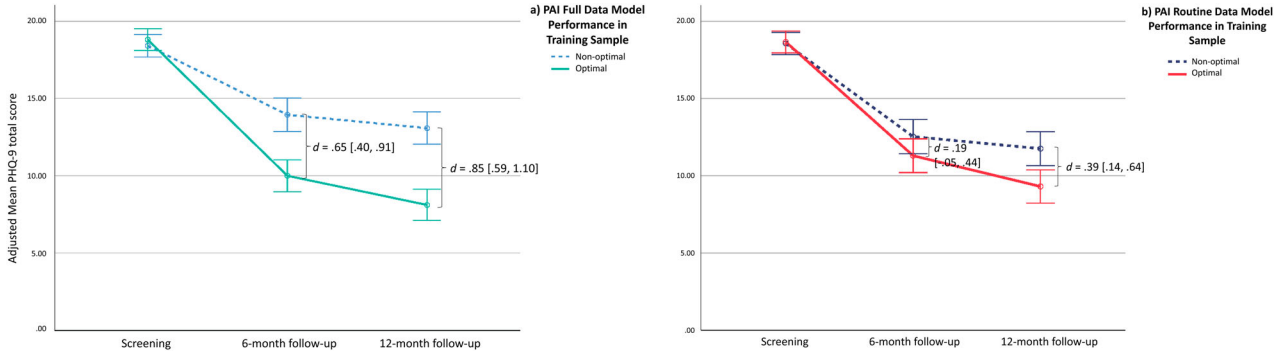

Figure 1. Estimated marginal means of Patient Health Questionnaire-9 (PHQ-9) total score at screening, 6-month and 12-month follow-up, 95% confidence intervals and effect sizes for patients who received their optimal and non-optimal treatment according to the Personalized Advantage Index (PAI) for the trained models. Note: (a) PAI classification according to full data model; (b) PAI classification according to routine data model; d = Cohen's effect size; In brackets = 95% confidence intervals for the effect sizes between groups; PAI = Personalized Advantage Index; PHQ-9 = Patient Health Questionnaire-9.

$p < .001$, with contrasts revealing statistically significant differences between those patients who received their optimal vs. non-optimal treatment at the two-time points: $F(1, 253) = 27.20$, $p < .001$, $d = .65$ [.40, .91] at six months, and $F(1, 253) = 45.61$, $p < .001$, $d = .85$ [.59, 1.10] at 12-months (see Figure 1a). In terms of effect sizes within patients, those who received their optimal treatment obtained a pre–post effect size of $d = 1.43$ [1.19, 1.67] at six-, and $d = 1.90$ [1.61, 2.18] at 12 months. By contrast, patients who received their non-optimal treatment obtained $d = .75$ [.55, .95] at six-, and $d = .85$ [.64, 1.05] at 12 months.

Based on the *routine data model*, with RM-ANOVA we obtained a significant effect of the differential prediction on PHQ-9 total scores at six and 12 months: Greenhouse–Geisser's $\varepsilon = .98$, $F(1.96, 496.65) = 5.39$, $p < .01$, with contrasts revealing statistically significant differences between those patients who received their optimal vs. non-optimal treatment at 12 months: $F(1, 253) = 9.77$, $p < .01$, $d = .39$ [.14, .64], although no differences were found at six months: $F(1, 253) = 2.43$, $p = .12$, $d = .19$ [-.05, .44] (see Figure 1b). In terms of within-patient effect sizes, receipt of optimal treatment yielded values of $d = 1.38$ [1.12, 1.65] at six-, and $d = 1.77$ [1.50, 2.06] at 12 months, while receipt of non-optimal treatment obtained $d = 1.09$ [.84, 1.34] at six-, and $d = 1.25$ [.98, 1.51] at 12 months.

When the *routine data model* was applied to routine data from the trial (first and last session) we obtained a significant effect of the differential prediction on PHQ-9 total scores at last session: Greenhouse–Geisser's $\varepsilon = 1.00$, $F(1, 253) = 11.30$, $p < .001$, with contrasts revealing statistically significant differences between those patients who received their optimal vs. non-optimal treatment: $F(1, 253) = 6.35$, $p < .05$, $d = .48$ [.23, .73]. In terms of within-patient effect sizes, the model yielded values of $d = 1.74$ [1.47, 2.01] for those who received the optimal treatment and $d = 1.12$ [.89, 1.34] for those who received the non-optimal treatment. These results are shown in Supplementary Figure 1b, together with the results of the full data model applied to the routine data from the trial (Supplementary Figure 1a).

Routine Data Model Evaluation Based on the Test Sample

Applying the *routine data model* to the test sample yielded a PAI mean of -.86 ($Mdn = -1.34$, $SD = 2.26$, range = −5.05–5.91) indicating an advantage of CBT in predicted outcomes. PCET was recommended to 98 (38.4%) patients, while CBT was recommended to 157 (61.6%) patients. Of the 255 patients, 127 (49.8%) received their optimal treatment (42 in PCET and 85 in CBT) and 128 (50.2%) their non-optimal treatment (72 in PCET and 56 in CBT). The difference in outcomes between the differential prediction groups was tested with ANCOVA and yielded a non-significant difference between PHQ-9 total scores at last session, $F(1, 252) = 3.89$, $p = .05$, with an adjusted effect size of $d = .21$ [−.00003, .43] (see Figure 2a). In terms of within-patient effect sizes, optimal and non-optimal pre–post effect sizes were $d = 1.07$ [.85, 1.28] and $d = .82$ [.62, 1.02], respectively.

Finally, when evaluating the differential prediction performance in the subgroup of patients with strongest indications ($n = 168$ with an effect size difference of $d \geq 0.3$, PAI score ≥ 1.53 or ≤ -1.53), we obtained a statistically significant difference between those patients who received their optimal vs. non-optimal treatment at their last session: $F(1, 165) = 7.96$, p

< .01, with an adjusted effect size of $d = .38$ [.11, .64] (see Figure 2b). In terms of within-patient effect sizes, optimal and non-optimal pre–post effect sizes were $d = .98$ [.73, 1.22] and $d = .56$ [.31, .81], respectively.

Discussion

The current study was aligned with the goals of precision mental health care. In this regard, it aimed to develop a treatment selection algorithm to provide psychological therapy services with an evidence-based means of better matching therapy, in this instance either PCET or CBT, to individual patients. To our knowledge, this is the first study that developed such an algorithm considering a distal treatment outcome (i.e., 12-months follow-up) focusing on PCET. Given the finding in the PRaCTICED trial that, at the group level, there was no difference in outcomes at the final session but that outcomes favored CBT at 12 months, the current study reframes a comparison between treatment modalities with one of matching individual patients with their optimal treatment.

Noteworthy is the finding of a significant effect ($d = .38$) between optimal and non-optimal when the threshold is set to focus solely on those patients with the strongest indications. The effect translates into a 21% advantage. Differential effects often occur toward the ends of a distribution with those patients toward the center being less swayed by preference or effect of one condition or the other. Hence, for many patients, receipt of CBT or PCET is not a crucial matter. But where the PAI score exceeds an effect size difference of $d = 0.3$, then matching matters. But it is important to emphasize that the subgroup analysis focuses on identifying those patients who *maximize* the difference between optimal and non-optimal (by showing the strongest indications for a particular treatment). They tend to be more severe patients and even though their outcomes are not as good as for all patients, the effect of receiving the optimal treatment is even more important for them.

The findings parallel those of the original DeRubeis et al. (2014) study who reported *d* values of .28 (whole sample) and .58 (higher PAI score sample) for the advantage to optimal over non-optimal treatments in a comparison of CBT vs. antidepressant medication. Clearly, greater effects are achieved by targeted prescription. Identifying those patients for whom treatment mismatch is greatest may likely improve overall outcomes and reduce dropout.

However, even without such targeted prescription, the adoption of a strategy for all patients that yielded a small additional effect (e.g., $d = .21$) would be beneficial. The small effect, in traditional terms, is to be expected as such models are based on interactions between treatments and patient variables. There is increasing recognition that smaller effects are relevant in psychological therapy research and that, because of the complexity of matching individuals to treatment modalities, the effects are naturally going to be small (Barkham, 2023). But, in the context of the large numbers of people referring for psychological support (e.g., >1.5 million patients per annum are referred to the NHS Talking Therapies program), the gain at a population level is considerable. As such, this area of work is consistent with calls to recognize the value of building a science on the cumulative yield of small effects (e.g., Götz et al., 2022).

In the training sample, the PAI provided a robust classification of patients who received their optimal and non-optimal treatment according to both the full and routine data models. While in the *full data model*, optimal and non-optimal groups showed statistically significant differences at both time points, in the *routine data model*, these groups did not differ statistically significantly in the PHQ-9 at six months but did so at 12 months. Bearing in mind that the regression models for estimating the PAI predicted treatment outcome at 12 months, it is noteworthy that at six months, the PAI for the *full data model* already differentiated between patients who received their optimal and non-optimal treatment with a .65 effect size. These results show, retrospectively, that an algorithm for treatment allocation (i.e., the PAI) may provide better outcomes than allocating patients randomly. As expected, the *full data model* included more predictors than the *routine data model*, with the former resulting in a better classification of patients (i.e., larger effects sizes within and between patients).

When applying the *routine data model* to the test sample, the resulting effects yielded similar within- and between- patients effect sizes to those obtained with this model in the training sample. This is important as the PSM algorithm was set to allow heterogeneity between samples, resulting in small differences in patients' features between the two samples and thereby increasing the generalizability and external validity of the model. Despite these differences, the *routine data model* produced similar results in both samples (training and test samples), which can be considered evidence of its replicability.

Considering methodological issues in the training sample, in terms of variable selection, we went beyond previous studies incorporating and analyzing

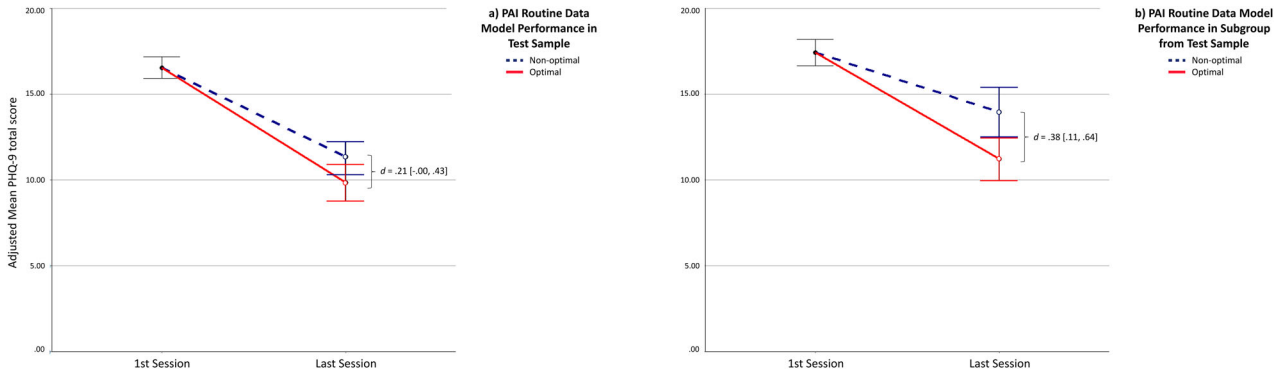

Figure 2. Estimated marginal means of Patient Health Questionnaire-9 (PHQ-9) total score at first and last sessions, 95% confidence intervals and adjusted effect sizes for patients who received their optimal and non-optimal treatment according to the Personalized Advantage Index (PAI) for the routine data model in the test sample. Note: (a) PAI routine data model performance in the external sample; (b) PAI routine data model performance in the subgroup of patients with strongest indications; *d* = Adjusted effect size; In brackets = 95% confidence intervals for the effect sizes between groups; PAI = Personalized Advantage Index; PHQ-9 = Patient Health Questionnaire-9.

item scores and the interaction between sociodemographic variables. When looking at the results from the RF algorithms for the *full data model*, more predictors were required for predicting outcome in PCET at 12 months (85) than for CBT (64). And in considering the variance explained by the regression analyses, a greater percentage of variance was explained for CBT in the *full* and *routine data models* sets (47% and 24%, respectively), compared with PCET (35% and 16%, respectively). With a slightly higher number of predictors entered into the CBT model yielding a higher contribution in terms of adjusted R^2 than for the PCET model, the outcome variability in CBT is better explained than the outcome variability in PCET in both set of models. However, the variance explained for both models in both sets exceeded those reported by Delgadillo and Gonzalez Salas Duhne (2020)—9% and 13% for PCET and CBT, respectively. These differences may be explained by the nature of the samples, the way the treatments were provided, and the predictors used in both studies (naturalistic sample and treatment conditions, and use of total scale scores by Delgadillo and Gonzalez Salas Duhne (2020)).

The difference in the amount of variance explained by the models suggests caution in the interpretation of difference in effectiveness found between PCET and CBT at 12 months. The CBT models may be classifying patients better than the PCET models. The characteristics of each approach might explain these differences. PCET is a non-directive approach, centered in emergent meanings that come about during the process, which can be idiosyncratic; thus, what each patient takes from the sessions might present a wider range of variability from one person to another. In contrast, CBT, which comprises a more structured approach centered on specific targets (i.e., specific goals and symptoms) might represent a more homogeneous treatment (vary less from one person to another) and thereby producing more homogeneous effects.

Clinical implications

Item-level information can contribute to a more precise patient profiling, describing the specific clinical characteristics that may benefit a patient from one treatment or another. Additionally, this information is useful to clinicians in targeting specific symptoms and adapting treatment to the particular patient profile beyond the information provided by a total abstract score (see O'Driscoll et al., 2023). Taking the predictors that were entered in the *full data model*, patients that obtained better outcomes in PCET at 12 months were characterized by being employed females or men not employed from more advantaged areas, who have a sense of purpose in life, greater expectations of improvement from PCET, together with feelings of being criticized by others, impairment in close relationships, and guilty feelings. By contrast, patients that obtained better outcomes in CBT at 12 months were characterized by having a main activity focus, particularly employed males from more advantageous neighborhoods and also females not in employment from these areas, together with greater feelings of being in control of life and viewing themselves as being able to bounce back, with a sense of current worthlessness but having felt warmth for someone at some time. In addition, the finding that sociodemographic variables and their interactions yield, at times, predictions in the opposite directions (e.g., gender, RAS and IMD) are consistent with studies showing the necessity to

adapt psychological treatments to the socio-economic background of the patients (i.e., to meet their socio-economic needs; Finegan et al., 2018) and the incorporation of a gender-sensitive perspective (Budge & Moradi, 2018).

Those patients who did not obtain a clear indication of the superiority of one treatment over the other represent a subgroup that can genuinely be offered a choice based on patient preferences. A personalized choice adds a critical nuance compared to personalization through empirical prescription alone because it recognizes that patients have unique preferences and values that should be considered.

Regarding the implementation of both predictive models, the *routine data model* can be directly implemented in clinical practice due to being based on the variables routinely collected by NHS Talking Therapies services in England. However, this model offers a less precise prediction than the *full data model*, thereby supporting the argument for selective additional information to yield a more accurate prognosis tool in routine care. Nevertheless, the effectiveness of providing psychological therapies under the recommendation of these algorithms should be tested in prognostic studies.

Limitations and future directions

We were not able to adopt a sample to test the results of the models at 12 months because the NHS Talking Therapies program does not assess patients beyond the end of treatment. Similarly, we did not test the *full data model* on an external sample due to the variables used to build the latter not being routinely collected in the NHS Talking Therapies program. Accordingly, the *full data model* should be viewed as an indication of the potential yield of such a model and of the approach in general.

Albeit a pragmatic trial, the intrinsic features and procedures of such a design may limit the ecological validity and real-life application of the algorithms developed. Also, to develop the predictive models, we did not count on a holdout sample. However, several methods for internal validation and overfitting prevention were implemented (e.g., sample splitting configuration for the RF algorithms, LOOCV procedure to compute regression models), and the models were tested in an external sample. The current study comprised mainly western White patients, which, while reflecting the general characteristics of the majority of users of the NHS Talking Therapies program, warrants being extended to be more inclusive of cultural and ethnic minorities.

Additionally, the subgroup of patients without benefits from PCET or CBT requires further investigation. In this sense, there is a need to explore alternative treatment options for this specific subgroup, as their lack of response to PCET or CBT suggests the existence of underlying factors or characteristics that require further investigation. By acknowledging the limitations of both PCET and CBT in these cases, we can better tailor interventions to individual patients' needs and refer them to more suitable alternatives.

Therapist variability was not included in the analysis and further studies are needed regarding the inclusion of therapist effects in predictive models like the PAI. Although in the PRaCTICED trial the therapist effect was insignificant at 0.2% and unlikely to have changed appreciably with only a difference of four therapists in the study sample, the effect is likely to be larger and significant in routine delivery where there may be a more diverse population and less consistent treatment delivery (Saxon et al., 2017).

Finally, there is a need for prospective studies in the field of precision mental health care. Future studies should design RCTs aimed at testing the validity and effectiveness of these algorithms specified a priori (e.g., Lutz, Deisenhofer, et al., 2022).

Conclusion

We provided two algorithms that could be applied to recommend PCET or CBT for depression based on the highest probability of improvement a patient may have. This recommendation is not only based on which treatment might yield a better outcome for a patient at six months, but also on which might sustain those gains at 12 months. Further research is required on how to implement and translate these developments into routine care, especially considering the use of additional measures to ensure better functioning of predictive models and long-term follow-up assessments.

Disclosure Statement

MB is a co-developer of the CORE-OM that was used in this study but receives no financial gain from its use. All other co-authors state they have no conflicts of interest.

Please see the publication in *The Lancet Psychiatry* (Barkham et al., 2021) for a full list of acknowledgments regarding the original trial.

Supplemental Data

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10503307.2023.2269297.

ORCID

Danilo Moggia http://orcid.org/0000-0001-6321-4450
David Saxon http://orcid.org/0000-0002-9753-8477
Wolfgang Lutz http://orcid.org/0000-0002-5141-3847
Gillian E. Hardy http://orcid.org/0000-0002-9637-815X
Michael Barkham http://orcid.org/0000-0003-1687-6376

References

Barkham, M. (2023). Smaller effects matter in the psychological therapies: 25 years on from Wampold et al. (1997). *Psychotherapy Research*, *33*(4), 530–532. https://doi.org/10.1080/10503307.2022.2141589

Barkham, M., Saxon, D., Hardy, G. E., Bradburn, M., Galloway, D., Wickramasekera, N., Keetharuth, A. D., Bower, P., King, M., Elliott, R., Gabriel, L., Kellett, S., Shaw, S., Wilkinson, T., Connell, J., Harrison, P., Ardern, K., Bishop-Edwards, L., Ashley, K., … Brazier, J. E. (2021). Person-centred experiential therapy versus cognitive behavioural therapy delivered in the English Improving Access to Psychological Therapies service for the treatment of moderate or severe depression (PRaCTICED): A pragmatic, randomised, non-inferiority trial. *The Lancet Psychiatry*, *8*(6), 487–499. https://doi.org/10.1016/S2215-0366(21)00083-3

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression.* Guildford.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the beck depression inventory-II*. Psychological Corporation.

Budge, S. L., & Moradi, B. (2018). Attending to gender in psychotherapy: Understanding and incorporating systems of power. *Journal of Clinical Psychology*, *74*(11), 2014-2027. https://doi.org/10.1002/jclp.22686

Cairney, J., Veldhuizen, S., Vigod, S., Streiner, D. L., Wade, T. J., & Kurdyak, P. (2014). Exploring the social determinants of mental health service use using intersectionality theory and CART analysis. *Journal of Epidemiology and Community Health*, *68*(2), 145–150. https://doi.org/10.1136/jech-2013-203120

Clark D. M. (2018). Realizing the mass public benefit of evidence-based psychological therapies: The IAPT Program. *Annual Review of Clinical Psychology*, *14*(1), 159–183. https://doi.org/10.1146/annurev-clinpsy-050817-084833

Cohen, Z., Delgadillo, J., & DeRubeis, R. (2021). Personalized treatment approaches. In M. Barkham, W. Lutz, & L. Castonguay (Eds.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (7th Ed., pp. 673–704). Wiley.

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

Connell, J., Barkham, M., Stiles, W. B., Twigg, E., Evans, O., & Miles, J. N. V. (2007). Distribution of CORE-OM scores in a general population, clinical cut-off points and comparison with the CIS-R. *British Journal of Psychiatry*, *190*(1), 69–74. https://doi.org/10.1192/bjp.bp.105.017657

Connor, K. M., & Davidson, J. R. T. (2003). Development of a new resilience scale: The Connor-Davidson Resilience Scale (CD-RISC). *Depression and Anxiety*, *18*(2), 71–82. https://doi.org/10.1002/da.10113

Cuijpers, P., Berking, M., Andersson, G., Quigley, L., Kleiboer, A., & Dobson, K. S. (2013). A meta-analysis of cognitive-behavioural therapy for adult depression, alone and in comparison with other treatments. *The Canadian Journal of Psychiatry*, *58*(7), 376–385. https://doi.org/10.1177/070674371305800702

Davidson, J.R.T. (2018). *Connor-Davidson Resilience Scale (CD-RISC) Manual* [Unpublished manuscript]. www.cd-risc.com.

Deisenhofer, A. -K., Delgadillo, J., Rubel, J. A., Böhnke, J. R., Zimmermann, D., Schwartz, B., & Lutz, W. (2018). Individual treatment selection for patients with posttraumatic stress disorder. *Depression and Anxiety*, *35*(6), 541–550. https://doi.org/10.1002/da.22755

Delgadillo, J., & Gonzalez Salas Duhne, P. (2020). Targeted prescription of cognitive-behavioral therapy versus person-centered counseling for depression using a machine learning approach. *Journal of Consulting and Clinical Psychology*, *88*(1), 14–24. https://doi.org/10.1037/ccp0000476

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014). The Personalized Advantage Index: Translating research on prediction into individualized treatment recommendations. A demonstration. *PloS One*, *9*(1), e83875. https://doi.org/10.1371/journal.pone.0083875

Devilly, G. J., & Brokovec, T. D. (2000). Psychometric properties of the credibility/expectancy questionnaire. *Journal of Behavior Therapy and Experimental Psychiatry*, *31*(2), 73–86. https://doi.org/10.1016/S0005-7916(00)00012-4

Duffy, K. E. M., Simmonds-Buckley, M., Delgadillo, J., & Barkham, M. (2023). The efficacy of individual humanistic-experiential therapies for the treatment of depression: A systematic review and meta-analysis of randomized controlled trials. *Psychotherapy Research.* Advanced online publication. https://doi.org/10.1080/10503307.2023.2227757

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Society for Industrial and Applied Mathematics (SIAM)

Elliott, R., Watson, J., Timulak, L., & Sharbanee, J. (2021). Research on humanistic-experiential psychotherapies: Updated review. In M. Barkham, W. Lutz, & L. G. Castonguay (Eds.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (pp. 421–468). Wiley.

Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, G., Mellor-Clark, J., & Audin, K. (2002). Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE-OM. *British Journal of Psychiatry*, *180*(1), 51–60. https://doi.org/10.1192/bjp.180.1.51

Finegan, M., Firth, N., Wojnarowski, C., & Delgadillo, J. (2018). Associations between socio-economic status and psychological therapy outcomes: A systematic review and meta-analysis. *Depression and Anxiety*, *35*(6), 560–573. https://doi.org/10.1002/da.22765

Fried, E. I., & Nesse, R. M. (2015). Depression sum-scores don't add up: Why analyzing specific depression symptoms is essential. *BMC Medicine*, *13*(72), 1–11. https://doi.org/10.1186/s12916-015-0325-4

Furnival, G. M., & Wilson, R. W. (1974). Regressions by leaps and bounds. *Technometrics*, *16*(4), 499–511. https://doi.org/10.1080/00401706.1974.10489231

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

Götz, F. M., Gosling, S. D., & Rentfrow, P. J. (2022). Small effects: The indispensable foundation for a cumulative psychological science. *Perspectives on Psychological Science*, *17*(1), 205–215. https://doi.org/10.1177/1745691620984483

Greenberg, L. S., & Watson, J. C. (2006). *Emotion-focused therapy for depression* (1st Ed.). American Psychological Association.

Hedges, L. V., Tipton, E., Zejnullahi, R., & Diaz, K. G. (2023). Effect sizes in ANCOVA and difference-in-differences designs. *The British Journal of Mathematical and Statistical Psychology*, *76*(2), 259–282. https://doi.org/10.1111/bmsp.12296

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection - A review and recommendations for the practicing statistician. *Biometrical Journal*, *60*(3), 431–449. https://doi.org/10.1002/bimj.201700067

Hill, A. (2011). *Curriculum for counselling for depression: Continuing professional development for qualified therapists delivering high intensity interventions*. British Association of Counselling and Psychotherapy (BACP). https://www.yumpu.com/en/document/read/46055498/curriculum-for-counselling-for-depression-iapt-it-shared-services

IBM Corp. (2021). *IBM SPSS statistics for windows* (Version 28) [Computer software]. https://www.ibm.com/analytics/spss-statistics-software

JASP Team. (2022). *JASP* (Version 0.16.1) [Computer software]. https://jasp-stats.org/

Kessler, R. C. (2012). The costs of depression. *Psychiatric Clinics of North America*, *35*(1), 1–14. https://doi.org/10.1016/j.psc.2011.11.005

King, M., Marston, L., & Bower, P. (2014). Comparison of non-directive counselling and cognitive behaviour therapy for patients presenting in general practice with an ICD-10 depressive episode: A randomized control trial. *Psychological Medicine*, *44*(9), 1835–1844. https://doi.org/10.1017/S0033291713002377

Kraemer, H. C. & Blasey, C.M. (2004). Centring in regression analyses: A strategy to prevent errors in statistical inference. *International Journal of Methods in Psychiatric Research.*, *13*(3), 141–151. https://doi.org/10.1002/mpr.170

Kraemer, H. C., Stice, E., Kazdin, A., Offord, D., & Kupfer, D. (2001). How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. *American Journal of Psychiatry*, *158*(6), 848–856. https://doi.org/10.1176/appi.ajp.158.6.848

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

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

Lewis, G. (1994). Assessing psychiatric disorder with a human interviewer or a computer. *Journal of Epidemiology and Community Health*, *48*(2), 207–210. https://doi.org/10.1136/jech.48.2.207

Liaw, A., & Wiener, M. (2002). *Classification and regression by randomForest*. *R News*, *2*(3), 18–22. http://CRAN.R-project.org/doc/Rnews/

Liu, Q., He, H., Yang, J., Feng, X., Zhao, F., & Lyu, J. (2020). Changes in the global burden of depression from 1990 to 2017: Findings from the Global Burden of Disease study. *Journal of Psychiatric Research*, *126*, 134–140. https://doi.org/10.1016/j.jpsychires.2019.08.002.

Lumley, T. (2022). *Leaps: Regression subset selection* (Version 3.1). R package. https://cran.r-project.org/web/packages/leaps/index.html

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

Lutz, W., Schwartz, B., & Delgadillo, J. (2022). Measurement-based and data-informed psychological therapy. *Annual Review of Clinical Psychology*, https://doi.org/10.1146/annurev-clinpsy-071720-014821

Moller, N. P., Ryan, G., Rollings, J., & Barkham, M. (2019). The 2018 UK NHS Digital annual report on the improving access to psychological therapies programme: A brief commentary. *BMC Psychiatry*, *19*(1), 252. https://doi.org/10.1186/s12888-019-2235-z

Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W., Vickers, A. J., Ransohoff, D. F., & Collins, G. S. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and Elaboration. *Annals of Internal Medicine*. https://doi.org/10.7326/M14-0698

Mundt, J. C., Marks, I. M., Shear, M. K., & Greist, J. H. (2002). The work and social adjustment scale: A simple measure of impairment in functioning. *British Journal of Psychiatry*, *180*(5), 461–464. https://doi.org/10.1192/bjp.180.5.461

Murphy, D. (2019). *Person-centred experiential counselling for depression: A manual for training and practice*. SAGE Publishing.

National Collaborating Centre for Mental Health. (2021). *The improving access to psychological therapies manual*. https://www.england.nhs.uk/publication/the-improving-access-to-psychological-therapies-manual/

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

Noble, S., McLennan, D., Noble, M., Plunkett, E., Gutacker, N., Silk, M., & Wright, G. (2019). *The English indices of deprivation* 2019. *Research Report*. Ministry of Housing, Communities & Local Government. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/833947/IoD2019_Research_Report.pdf

O'Driscoll, C., Buckman, J. E. J., Saunders, R., Ellard, S., Naqvi, S. A., Singh, S., Wheatley, J., & Pilling, S. (2023). Symptom-specific effects of counselling for depression compared to cognitive-behavioural therapy. *BMJ Mental Health*, *26*(1), e300621. https://doi.org/10.1136/bmjment-2022-300621

PRaCTICED Trial Team. (2014a). *Cognitive Behaviour Therapy (CBT) PRaCTICED manual* (Version 2.0).

PRaCTICED Trial Team. (2014b). *Counselling for Depression (CfD) PRaCTICED manual* (Version 2.0).

R Core Team. (2021). *R: A language and environment for statistical computing* [computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rogers, C. R. (2003/1951). *Client-centered therapy: Its current practice, implications and theory.* Constable.

RStudio Team. (2021). *Rstudio: Integrated development for R* [computer software]. RStudio, Inc. http://www.rstudio.com/

Sanders, P., & Hill, A. F. (2014). *Counselling for depression: A person-centred and experiential approach to practice*. SAGE.

Saxon, D., Firth, N., & Barkham, M. (2017). The relationship between therapist effects and therapy delivery factors: Therapy modality, dosage, and non-completion. *Administration and Policy in Mental Health and Mental Health Services Research*, *44*(5), 705–715. https://doi.org/10.1007/s10488-016-0750-5

Schwartz, B., Cohen, Z. D., Rubel, J. A., Zimmermann, D., Wittmann, W. W., & Lutz, W. (2021). Personalized treatment selection in routine care: Integrating machine learning and statistical algorithms to recommend cognitive behavioral or psychodynamic therapy. *Psychotherapy Research*, *31*(1), 33–51. https://doi.org/10.1080/10503307.2020.1769219

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

Stekhoven, D. J. (2013). *missForest: Nonparametric missing value imputation using random forest. R package version 1*.4. https://cran.r-project.org/web/packages/missForest/missForest.pdf

Stekhoven, D. J., & Buehlmann, P. (2012). MissForest - Nonparametric missing value imputation for mixed-type data. *Bioinformatics (oxford, England)*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8* (1), 25. https://doi.org/10.1186/1471-2105-8-25

Tang, C., Garreau, D., & von Luxburg, U. (2018). When do random forests fail? *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 2987–2997).

Zafra-Tanaka, J. H., Goicochea-Lugo, S., Villarreal-Zegarra, D., & Taype-Rondan, A. (2019). Characteristics and quality of clinical practice guidelines for depression in adults: A scoping review. *BMC Psychiatry*, *19*(1), 76. https://doi.org/10.1186/s12888-019-2057-z