Article:
Delgadillo, J. and 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). pp. 14-24. ISSN: 0022-006X

https://doi.org/10.1037/ccp0000476

Targeted Prescription of Cognitive-Behavioral Therapy Versus Person-Centered Counseling for Depression Using a Machine Learning Approach

Jaime Delgadillo and Paulina Gonzalez Salas Duhne

Clinical Psychology Unit, Department of Psychology, University of Sheffield, Sheffield, UK

Conflicts of interest: None.

Correspondence: jaime.delgadillo@nhs.net

Abstract

Objective: Depression is a highly common mental disorder and a major cause of disability worldwide. Several psychological interventions are available, but there is a lack of evidence to decide which treatment works best for whom. This study aimed to identify subgroups of patients who respond differentially to cognitive behavioural therapy (CBT) or person-centred counselling for depression (CfD).

Methods: This was a retrospective analysis of archival routine practice data for 1435 patients who received either CBT (N=1104) or CfD (N=331) in primary care. The main outcome was post-treatment reliable and clinically significant improvement (RCSI) in the PHQ-9 depression measure. A targeted prescription algorithm was developed in a training sample (N=1085) using a supervised machine learning approach (elastic net with optimal scaling). The clinical utility of the algorithm was examined in a statistically independent test sample (N=350) using chi-square analysis and odds ratios.

Results: Cases in the test sample that received their model-indicated "optimal" treatment had a significantly higher RCSI rate (62.5%) compared to those who received the "suboptimal" treatment (41.7%); x^2 (DF = 1) = 4.79, p = .03, OR = 2.33 (95% CI = 1.09, 5.02).

Conclusions: Targeted prescription has the potential to make best use of currently available evidence-based treatments, improving outcomes for patients at no additional cost to psychological services.

Keywords: depression; cognitive behavioural therapy; counselling; precision medicine; personalized treatment selection

Public health significance: Reviews of clinical trials often conclude that different types of psychological interventions for depression are equally efficacious, and differential effects are rarely statistically or clinically significant. Recent studies suggest that subgroups of patients with specific characteristics respond more favourably to specific psychotherapies, and therefore targeted prescription could potentially improve their treatment outcomes. Using data from a large (N=1435) naturalistic cohort of patients who accessed either cognitive behavioural therapy or person-centred counselling for depression, the present study found evidence of differential response to treatments in a subsample of ~30%. These results were cross-validated in an adequately powered external test sample, which considerably strengthens the reliability of findings. Targeted prescription of these widely available and well-established psychological treatments could potentially improve clinical outcomes at no additional cost to services, representing a major advance in precision mental healthcare.

1. Introduction

Current guidelines for the management of depression in adults (National Institute for Health and Care Excellence; NICE, 2018) recommend psychological interventions such as cognitive behavioural therapy, interpersonal psychotherapy, brief psychodynamic therapy and person-centred counselling. These recommendations are informed by meta-analyses of randomized controlled trials which generally conclude that these psychological treatments are equally efficacious, and differential effects are rarely statistically or clinically significant (Cuijpers et al., 2008; Cuijpers et al., 2013). Practice-based studies comparing the effectiveness of psychological interventions reach similar conclusions. For example, a recent large-scale comparison of cognitive behavioural therapy (CBT) and person-centred counselling for depression (CfD) in primary care suggested that these treatments are equally effective (Pybis, Saxon, Hill, & Barkham, 2017).

Evidence such as this has led some to argue that established therapies may work through similar mechanisms or so-called *common factors* (Frank and Frank 1991; Wampold and Imel, 2015), thus justifying the need to consider patients' preferences through a shared-decision making process to guide treatment selection (Swift et al., 2011). Although there is evidence that accommodating patients' treatment preferences leads to better outcomes (Swift et al., 2018), there is ongoing debate in the field about the *common factors* and *treatment equivalence* hypotheses in psychotherapy (Mulder et al., 2017; Cuijpers et al., 2018). One line of argument is that different therapies may work better for different subgroups of patients with specific attributes, and that such *aptitude-treatment interactions* (ATI) are yet to be successfully identified (Cronbach and Snow, 1977). Despite its long history spanning over 40 years, ATI research has mostly yielded inconclusive evidence of differential treatment effects, possibly due to methodological problems including inadequately powered sample sizes, lack of replication in independent samples, and a predominance of studies investigating single variables as potential moderators of treatment effects (Dance and Neufeld, 1988; Smith and Sechrest, 1991; Snow, 1991).

A renewed interest in ATI research is evident in the field of psychotherapy, spurred by the emergence of studies that apply a multivariable approach to investigate how to optimally match patients to treatments (Cohen and DeRubeis, 2018). Unlike traditional single-moderator studies, contemporary studies investigate the combined influence of multiple patient-attributes (e.g., demographic, diagnostic, personality features) to profile patients into subgroups that respond to treatment in similar ways. For example, multivariable prediction algorithms have been developed to identify patients who respond differentially to CBT vs. antidepressant medication (DeRubeis et al., 2014), CBT vs. interpersonal psychotherapy (Huibers et al., 2015), CBT vs. eye-movement desensitization and reprocessing (Deisenhofer et al., 2018), prolonged exposure vs. cognitive processing therapy (Keefe et al., 2018), and CBT vs. psychodynamic therapy (Cohen et al., 2019). Recent studies are also leveraging the advantages of machine learning methods in order to optimize feature selection, to discover nonlinear associations and interactions between variables, and to yield prediction models that are more likely to generalise to new samples (e.g., Cohen et al., 2019; Deisenhofer et al., 2018; Delgadillo, Huey, Bennett, & McMillan, 2017; Keefe et al., 2018; Lorenzo-Luaces et al., 2017). Despite the increased sophistication of new studies, this emerging literature is still limited by old problems. Most of these studies use samples from randomized controlled trials, which are often statistically underpowered for multivariable analyses. It is also likely that some studies that seek to discover so-called *prescriptive variables* (treatment x moderator interactions) are likely to be underpowered to achieve this and may well be treating spurious findings as clinically informative (Luedtke et al., 2019). Furthermore, in spite of their use of *internal* cross-validation procedures, the majority of these studies are limited by a lack of replication of the apparent advantage of personalized treatment selection in an adequately powered *external* (hold-out) test sample that was not used to train the prediction algorithms. These limitations mean that the generalisability, replicability and clinical value of emerging treatment selection algorithms are yet to be demonstrated. For these reasons, it has been suggested that large practice-based and observational datasets could help to overcome some of the limitations concerning sample size and might have the additional advantage of

yielding treatment selection models that generalise to ordinary healthcare populations (Kessler, Bossarte, Luedtke, Zaslavsky, & Zubizarreta, 2019).

Informed by contemporary ATI research, the present study used a large practice-based dataset to investigate if certain subgroups of patients might respond differentially to CBT or CfD for depression. Based on prior studies, we hypothesised that a subsample of cases would show differential response to these treatments, and that cases assigned to their optimal treatment would have significantly higher rates of depression symptom remission. In order to support this hypothesis, the advantage of assignment to an optimal treatment would have to be replicated in an external validation sample which was not used to train the targeted prescription model.

2. Method

2.1. Setting and interventions

This study is based on the analysis of fully anonymized routine care data from patients treated for depression in a primary care service in the North of England. The study was registered with the National Health Service as a service evaluation (Reference: SEP_0517a), and was exempt from research ethical approval.

The participating service was part of the *Improving Access to Psychological Therapies* (IAPT) programme, a national treatment system which offers psychological interventions following a stepped care model (Clark, 2018). The initial step of treatment in this system involves brief (<8 sessions) guided self-help interventions offered to patients with mild-to-moderate symptoms. Patients who remain symptomatic after accessing guided self-help, or those deemed to be more severe cases at initial assessments, are stepped-up to high intensity therapies. All cases included in the present study sample accessed high intensity therapies, either CBT or CfD, which were prescribed using a shared-decision making process following clinical guidelines (NICE, 2018). This process typically requires the assessing clinician to provide a brief description to the patient regarding available treatments recommended for depression, after which an agreement is reached about a suitable treatment option. Such decisions

were often influenced by the assessor's clinical judgment, consultation with other clinicians in the service, patients' preferences, and contextual factors such as waiting times.

High intensity CBT in this service involved highly standardised, disorder-specific interventions listed in the Roth and Pilling (2008) competency framework. Patients with depression received protocol-driven Beckian CBT (Beck et al., 1979) or Behavioural Activation (Martell et al., 2001). The present study sample only selected cases that accessed Beckian CBT with accredited therapists, whose training was based on a national curriculum (Department of Health, 2011). CfD for depression is a person-centred, non-directive, experiential form of therapy broadly aligned to Rogers' school of humanistic counselling (Sanders and Hill, 2014). CfD was delivered by accredited counsellors who completed a *counselling for depression* training course guided by a national curriculum (Hill, 2011a) and competency framework (Hill, 2011b).

According to national guidelines, patients should access up to 20 sessions of CBT or up to 12 sessions of CfD (NICE, 2018). In practice, the mean number of sessions accessed by CBT (M=9.56, SD=7.02) and CfD cases (M=8.31, SD=4.29) in the study sample was not significantly different; Mann-Whitney U = 175647.00, p = .29. All therapists (CBT, CfD) in the service practiced under regular (weekly or every two-weeks) clinical supervision with accredited and experienced supervisors specialising in their treatment modality.

2.2. Measures

The primary outcome measure in this study was the Patient Health Questionnaire (PHQ-9; Kroenke, Spitzer, & Williams, 2001), which is a nine-item screening tool for depression. Each item is rated on a four-point Likert scale ranging from "0" (not at all) to "3" (nearly every day). Total scores range from 0 to 27, where lower scores indicate less severe symptoms. A cut-off ≥10 has been recommended identify clinically significant symptoms of major depressive disorder, with adequate sensitivity (88%) and specificity (88%) (Kroenke et al., 2001). A change ≥6 points has been recommended to assess statistically reliable improvement or deterioration (Richards & Borglin, 2011).

Patients treated in the participating service completed the PHQ-9 at an initial (pre-treatment) assessment, and before each weekly therapy session. In addition to the PHQ-9 measure, patients also completed weekly measures of anxiety symptoms (GAD-7; Spitzer, Kroenke, Williams, & Löwe, 2006) and functional impairment (WSAS; Mundt, Marks, Shear, & Greist, 2002) using standardised questionnaires. A last-observation-carried-forward method was programmed into the clinical service database, which prevented loss of data due to early treatment dropout and enabled an intention-to-treat approach to data analysis. Only pre- and post-treatment outcome measures were included in the study dataset.

At initial assessment, a standardised set of demographic and clinical information was gathered. Demographics included: gender, age, ethnicity, employment status and disability. The index of multiple deprivation (IMD) was also collected, which is an area-level index of socioeconomic deprivation for each patient's neighbourhood (Department for Communities and Local Government, 2011). Clinical information included primary diagnosis, chronicity (duration in years/months) of the primary problem, number of previous treatment episodes, comorbidity of long-term conditions (e.g., diabetes, asthma, etc.), antidepressant medication status, and outcome expectancy using a standardised measure (Lutz, Leon, Martinovich, Lyons & Stiles, 2007).

2.3. Sample characteristics

The study sample (N = 1435) only included patients who accessed ≥2 sessions of high intensity CBT or CfD, who had a primary diagnosis of major depressive disorder or recurrent depression, and who had case-level symptoms at the first session of high intensity therapy (PHQ-9 ≥10). At least two sessions were necessary to take the first as a baseline measure, and the last attended session as a post-treatment measure. More patients accessed CBT (N = 1104) compared to CfD (N = 331), since there was a larger number of qualified CBT therapists (N = 67) compared to counsellors (N = 12). Less than half (40.6%) of patients had previously accessed low intensity guided self-help before being stepped-up to either CBT or CfD. The average duration of a treatment episode, from initial assessment to end of treatment, was 7.12 months (SD=4.40).

Table 1 presents a detailed summary of sample characteristics. Statistical comparisons between CBT and CfD cases revealed significant case-mix differences in demographic and clinical features, showing that CBT cases tended to have higher indices of symptom severity (PHQ-9, GAD-7), functional impairment (WSAS) and use of antidepressant medication.

[Table 1]

2.4. Data analysis

The primary objective of the analysis was to develop an algorithm that would prescribe CBT vs. CfD based on the identification of subgroups of patients who tend to respond better to one or the other treatment. To achieve this, the analysis was conducted in four steps: (1) partitioning of the dataset into training and test samples; (2) development of prognostic indices for each form of treatment; (3) development of a treatment prescription algorithm; (4) external cross-validation of the algorithm in the test sample.

Our first challenge was to construct a test sample that (a) was adequately powered and (b) balanced patients' characteristics across groups in a similar way to randomised experiments – so as to minimise confounding due to the natural selection of cases into CBT or CfD through shared decision-making. A sample size calculation based on Cohen's rule (1992) indicated that 87 cases were required to detect a medium effect size comparing cases that received their optimal vs. suboptimal treatment, using a chi-square analysis with DF = 1, $\alpha = .05$ and 80% power. Informed by recent studies focusing on depression (DeRubeis et al., 2014, Cohen et al., 2019), we expected that only a subgroup of cases would be classified as having an optimal treatment. Therefore, we quadrupled the sample size requirement (N = 350) to ensure the sample would be sufficient to carry out hypothesis testing in a subset of cases as small as 25%. The sample size calculation was carried out specifically to undertake one primary hypothesis test in an external validation sample, and to avoid multiple tests in the same sample which is a common pitfall and limitation in prior treatment selection studies.

On this basis, we selected a random sample of 175 CBT cases from the full dataset and matched them to 175 CfD cases using propensity score matching (Rosenbaum & Rubin, 1983). Propensity score matching is an appropriate statistical approach to balance baseline differences in observed covariates across groups of patients exposed to different treatments in observational studies (Rosenbaum, 2017), and is recommended to support the development of treatment selection rules using naturalistic datasets (Kessler et al., 2019). This matching procedure was based on a logistic regression predicting CBT group membership, entering all demographic and pre-treatment clinical measures described in Table 1 as predictors, using a one-to-one nearest neighbours approach with a conservative tolerance level (calliper = 0.2) specified a priori, and allowing replacement to maximize matching precision. This procedure resulted in a test sample of 350 cases with balanced characteristics across groups (shown in online supplemental material), and a larger training sample of 1085 cases.

In the second step, each case in the full dataset was labelled with a binary clinical outcome (RCSI = 0 or 1) applying conventional criteria for reliable and clinically significant improvement (Jacobson and Truax, 1991), which participating therapists used to assess depression treatment response in routine care. RCSI was attained if post-treatment PHQ-9 scores were below the diagnostic cut-off (<10) and ≥6 points lower than the PHQ-9 measure at the first session of high intensity treatment. Missing data (< 10% in this sample) for demographic and clinical features were imputed using an expectation-maximization method (Schafer & Olsden, 1998).

Next, prognostic indices for each treatment were developed by separately analysing CBT (N = 929) and CfD (N = 156) cases in the training sample (N = 1085), enabling the identification of common and unique predictors for two alternative treatments. The prognostic index was expressed as a predicted probability of attaining post-treatment RCSI, based on the combined weight of multiple patient-characteristics that were available before the start of treatment. Variable selection and weight setting to construct each prognostic index were performed using a supervised machine learning approach, with the RCSI label as a dependent variable. This statistical model combined Elastic Net regularization, optimal scaling, and nested internal cross-validation loops. Elastic Net regularization

combines the Ridge regression and LASSO approaches, which are part of the *penalized regression* family of machine learning algorithms developed to enhance prediction accuracy and generalizability (Zou & Hastie, 2005). Ridge regression minimizes the residual sum of squares through the application of L2 regularization, imposing the "sum of squared values of all coefficients" as a penalty term that shrinks coefficients towards zero to minimize overfitting to a training sample. This shrinkage procedure differentially weights variables according to their predictive importance, allowing multiple and inter-correlated variables in the same model. The LASSO (*Least Absolute Shrinkage and Selection Operator*) applies L1 regularization, imposing the "sum of absolute values of all coefficients" as a penalty term that shrinks coefficients exactly to zero if they have no predictive value or if they are highly collinear with stronger predictors in the model. As such, the LASSO additionally performs feature selection by shrinking some coefficients to zero, yielding sparse and conservative models (Tibshirani, 1996). The Elastic Net approach leverages the advantages of both L1 and L2 regularization, attaining feature selection via the LASSO and allowing for the inclusion of "weak predictors" via Ridge regression in a model that is robust to multicollinearity. Optimal scaling enables the discovery of non-linear relationships between dependent and independent variables by fitting splines (Gifi, 1990). An advantage of optimal scaling is that it enables the discovery of non-linear trends that are not pre-specified, without a need to introduce (or penalize) exponential terms (e.g., quadratic, cubic) into the model. In order to determine the model with minimum *expected prediction error*, an internal cross-validation loop was applied 1000 times to each iteration of the Elastic Net solution increasing the penalty terms in 0.01 units (also known as a *grid search*), using the .632+ bootstrap resampling method (Efron & Tibshirani, 1997). The optimal model was chosen using the 1 standard error rule. The grid search space was constrained by the alpha hyperparameter ($\alpha = 0.10$), which was found to optimize predictive accuracy with this set of features, in parameter tuning tests run within the training sample. As recommended by previous researchers, we ran a minimal number of parameter tuning tests to minimize overfitting due to multiple testing (Pearson, Pisner, Meyer, Shumake, & Beevers, 2018), comparing three alpha values selected a priori: $\alpha = 0.50$, $\alpha = 0.25$, $\alpha = 0.10$. The resulting treatment-

specific equations could be applied to new cases (whose demographic and clinical features are known) in order to calculate the predicted probability of symptom remission (RCSI) with CBT or with CfD.

In order to test the robustness of the above prognostic model, we also generated an aptitude-treatment-interaction (ATI) model using the full training sample pooling CBT and CfD cases. To achieve this, we trained an ensemble of 1000 decision trees using a random forest approach (Breiman, 2001) which yielded predicted probabilities of response to CBT and CfD for each patient based on their individual features.[1] We examined the accuracy of the prognostic versus ATI models in the test sample (N = 350), comparing the predicted probabilities yielded by each model using the area under the curve (AUC). The most accurate model was then used to develop a targeted prescription algorithm described below.

In the third step of analysis, we used the optimal equations constructed in the training sample to develop a *personalized advantage index* (PAI; DeRubeis et al., 2014). The PAI was calculated by taking the difference between the two prognostic indices (CBT prognosis – CfD prognosis) for each case. It was expressed in percentage (probability) units and was centred at zero; where a positive score favoured prescribing CBT and a negative score favoured CfD. We then classified cases into three subgroups based on the PAI cut-offs that were found to be 1 standard deviation above/below the mean: cases for whom no optimal treatment was strongly indicated (PAI close to the mean), and cases for whom the (model-indicated) optimal prescription was CBT or CfD.

In the final step of analysis, we performed chi-square analyses and estimated odds ratios to compare observed RCSI rates between cases in the test sample that received their (model-indicated) optimal treatment vs. cases that received the suboptimal treatment. A series of sensitivity analyses were performed to assess the robustness of these results; adjusting the odds ratios after controlling for (a) prior low intensity treatment and (b) propensity scores. We calculated a *mean prognostic index* (MPI; mean of the two treatment-specific prognoses) for each case in the test sample, and examined

[1] Further details about this analytic approach are provided in supplementary appendix A.

its correlation with the PAI. Finally, in order to assess if the primary results were not simply an artefact of the RCSI binary outcome metric, we compared differences in post-treatment PHQ-9 scores between cases assigned to optimal vs. suboptimal treatments using effect sizes (Hedges' g, adjusting for small and unbalanced sample sizes).

3. Results

3.1. General effectiveness of therapy

The base rate of cases meeting RCSI criteria in the full sample was 40.2%, with no significant differences between the CBT (39.3%) and CfD (43.2%) groups; x^2 (DF = 1) = 1.60, p = .21.

[Table 2]

3.2. Comparison of prognostic vs. ATI models

Table 2 compares the relative performance of alternative outcome prediction models in the training and test samples. Overall, the performance of the prognostic model was more stable (lower cross-validation shrinkage) and marginally better than the ATI model in the test sample (average AUC 0.62 vs. 0.58). Therefore, subsequent analyses applied the prognostic model equations.

[Table 3]

3.3. Predictors of treatment outcomes

Table 2 shows the variables and regularized coefficients selected into the optimal prognostic models for each treatment group. Beta coefficients that were shrunk to exactly 0 were indicative of variables that did not have prognostic value. Six common prognostic variables featured in both prediction models: age, employment status, disability, baseline PHQ-9 and WSAS, and IMD. Disability predicted poorer outcomes in CBT but better outcomes in CfD. Patients living in more deprived areas

tended to have poor outcomes in CBT, but better outcomes in CfD. Ethnicity was only relevant to CBT, where people from minority ethnic groups tended to have poorer outcomes. Four variables were only relevant to CfD: better treatment outcomes were predicted for patients with higher baseline anxiety, longer chronicity, lower outcome expectancy and those not taking antidepressant medication. The relative magnitude of regularized coefficients provides an indication of the prognostic importance of each variable.[2] Baseline depression severity (PHQ-9) was the most important predictor for CBT, whereas baseline functional impairment (WSAS) was most important for CfD.

[Figure 1]

3.4. External cross-validation of targeted prescription model

Using PAI scores derived from the above prognostic models enabled us to classify test-sample cases into three groups: those for whom no optimal treatment was selected (N = 238, 68.0%), and those for whom the optimal treatment was CBT (N = 48, 13.7%) or CfD (N = 64, 18.3%). Within the subset of cases that had differential response to treatment, 57.1% did in fact receive their optimal treatment, although agreement between the model-indicated prescription and the actual prescription was weak (Kappa = .13, SE = .09). Figure 1 shows that cases that received their model-indicated treatment had a significantly higher RCSI rate (62.5%) compared to those who received the suboptimal treatment (41.7%); x^2 (DF = 1) = 4.79, p = .03; OR = 2.33 (95% CI = 1.09, 5.02), Nagelkerke R^2 = .06. These findings were robust to sensitivity analyses. The odds ratios remained statistically significant (p < .05) after controlling for prior low intensity treatment in the stepped care system (OR = 2.31 [95% CI = 1.08, 4.98]) and after controlling for propensity scores (OR = 2.18 [95% CI = 1.01, 4.74]).

[Figure 2]

[2] For comparison, a predictor importance chart for the random forest approach is available in supplemental appendix B.

Figure 2 plots the distribution of PAI scores against MPI scores, which were moderately correlated; r = -.35, p <.001. The most responsive cases (MPI >60%, less severe and complex cases) tended to have a model-indicated advantage in CfD; whereas the least responsive cases (MPI <20%) tended to have no model-indicated treatment.

Figure 3 presents between-group effect sizes for cases in the test sample that were assigned to their optimal vs. suboptimal treatment. The overall effect size was g = .26, favouring assignment to the optimal treatment. We examined effect sizes for subsamples that excluded cases with the best and poorest expected prognoses along the MPI scale. Mean post-treatment PHQ-9 scores were consistently lower in cases assigned to their optimal treatment across all subgroups (see online supplement). However, as shown in Figure 3, optimal treatment assignment resulted in a greater advantage (g = .41) for cases in the middle range of the MPI scale.

[Figure 3]

4. Discussion

4.1. Main findings

Consistent with previous studies (e.g., Pybis et al., 2017), we found no significant differences in treatment outcomes when results were aggregated and broadly compared between CBT and CfD. However, using a machine learning analysis and external cross-validation procedure, we identified a subgroup of patients (~30%) with differential response to these treatments. Within this subgroup, patients who accessed their model-indicated treatment had significantly higher response rates (~60%) compared to those who received the suboptimal treatment (~40%). The odds ratio for this comparison (2.33) indicates that patients assigned to their optimal treatment were 2 times more likely to attain reliable and clinically significant improvement. This finding remained significant after controlling for cases that accessed prior low intensity treatment (41%), which was not associated with response rates.

The main result was also robust to sensitivity analyses adjusting for the non-random assignment of cases to treatments, and examining between-group differences in post-treatment PHQ-9 scores.

4.2. Implications for practice and theory

We found that the shared decision-making process which guided treatment selection in this service tended to prescribe treatments appropriately for only 57% of cases where targeted prescription really matters. Clearly, treatment outcomes could be improved if more patients within this subgroup were consistently matched to their optimal treatment. The most responsive and less impaired cases (MPI > .60) tended to be better candidates for CfD. Moreover, no optimal treatment was indicated for the subgroup of cases who had extremely poor expected prognoses (MPI < .10). This observation is consistent with the *intractable cases hypothesis* (DeRubeis et al., 2014b), which suggests that no differential treatment response would be found in the cases that are most difficult to treat.

These findings challenge the assumption of *treatment equivalence*, since evidently some patients respond better to CBT and others to CfD. It is also difficult to reconcile these findings with the *common factors* model (Wampold & Imel, 2015). If CBT and CfD worked through the same mechanisms of action, no differential response would be observed in cases that are equally responsive to treatment (e.g., cases close to the mean of the MPI distribution, who in fact benefited the most from optimal treatment assignment). Furthermore, patients in this cohort accessed treatment based on shared decision-making, which considers patients' preferences (NICE, 2018). Even after controlling for this allocation process (using propensity scores), differential treatment effects were still observed, so this cannot be simply explained by preference accommodation (Swift et al., 2018).

We identified several variables that predict differential effects in each treatment, and which could be thought of as moderators. Patients from minority ethnic groups, living in poverty, and with disabilities tended to have a poorer response to CBT, whereas they tended to have better outcomes in CfD. There is a growing literature on the need for cultural adaptation of CBT to enhance its acceptability to minority ethnic populations (e.g., Hinton, Rivera, Hofmann, Barlow, & Otto, 2012;

Naeem, Waheed, Gobbi, Ayub, & Kingdon, 2011). Emerging evidence also suggests that psychological treatments such as CBT might not be adequately meeting the needs of patients from socioeconomically deprived backgrounds (see review by Finegan, Firth, Wojnarowski, & Delgadillo, 2018), and may thus require some adaptation. On the other hand, CfD was less effective for patients who had a shorter chronicity of depression, high outcome expectations, and who were taking antidepressant medications. It may be that such cases have a greater degree of self-efficacy and readiness to change, and benefit from a change-oriented and skills-based intervention like CBT. Although our interpretations of the findings are speculative in the absence of therapy process data, it is plausible that CBT and CfD have specific mechanisms of action (over and above common factors) that are necessary to attain remission of symptoms in subgroups of cases with specific characteristics: a *phenotype-mechanism match*. Alternatively, it might be that these treatments do not work well in patients with certain characteristics, and targeted prescription might actually rectify a *phenotype-treatment mismatch*. Limitations of the present data preclude us from drawing firm conclusions about the explanation for the observed differential effects; but future research may help to elucidate if this may be due to *phenotype-mechanism match*, *phenotype-treatment mismatch*, or both.

4.3. Methodological considerations

Overall, our findings concur with recent studies showing differential treatment effects when comparing different psychotherapies for depression (Huibers et al., 2015; Cohen et al., 2019). Unlike most prior studies, the present results were cross-validated in an external test sample that was not used to train the prediction model. In addition, the sample was adequately powered for the intended analyses, using a robust method to identify common and unique prognostic indicators for each treatment. There is a current debate in this field about alternative approaches to develop treatment selection models (e.g., prognostic models vs. ATI models) and to select and weight variables for such models. In this study, we have shown that developing regularized prognostic indices separately for each treatment sample yielded more stable and accurate predictions compared to an ATI model trained using a random forest approach. Furthermore, we found that the cross-validation shrinkage

of the ATI model was large (~ 0.29 training-to-test AUC difference), which is indicative of considerable overfitting in the training sample. The prognostic model is likely to be less vulnerable to selecting spurious *prescriptive variables*, compared to ATI approaches that attempt to model interactions in relatively small samples such as the CfD group in this study (<200 cases). These features of the study sample and design increase our confidence in the generalisability of these findings, although there are also several caveats and methodological limitations that should be considered.

Firstly, the reliance on routine practice data raises problems with internal validity, since treatment allocation was not randomised. We applied appropriate methods to adjust for non-random allocation (propensity score matching, and adjustment for propensity scores), although these methods also have limitations and do not rule out the possibility that relevant unmeasured factors may not be balanced across groups (Smith & Todd, 2005). Other rigorous causal inference methods for observational samples are available, such as G-computation (Snowden, Rose, & Mortimer, 2011) or targeted maximum likelihood estimation (Schuler & Rose, 2017), which may yield different results and could be examined in combination with machine learning analyses in future studies. On the other hand, the naturalistic study design enhances external validity, particularly since the treatments were highly standardised, closely supervised, and consistent with routine practice in IAPT stepped care services. Still, the dataset was drawn from a single service, and the extent to which this targeted prescription model may generalize to other services and groups of therapists is unknown.

We note that the accuracy of treatment-specific prognostic indices was modest (test-sample AUC = 0.59 to 0.65) by conventional standards. The AUC statistic is commonly used in diagnostic accuracy studies to assess the performance of a screening test relative to a diagnostic standard, where values above 0.70 are deemed clinically useful (Swets, 1988). Large AUC values are to be expected in diagnostic accuracy studies that measure the test (predictor) and diagnostic standard (outcome) contiguously in time (i.e., on the same day). From a signal detection perspective, we might say that a valid screening test is measuring a *proximal* outcome: an underlying condition which is present at the time of screening. However, the prognostic models detailed above are in fact predicting *distal*

outcomes that occur approximately 7 months after initial (pre-treatment) assessments. From this viewpoint, it is remarkable that a relatively sparse set of features gathered before the start of treatment can yield clinically useful prognostic models which generalize to a new sample, and which could potentially double the probability of improvement for patients who are likely to benefit from targeted prescription. The above AUC values, therefore, should be interpreted as small-to-moderate effect sizes for the prediction of *distal* outcomes (approximate Cohen's *d* = .32 to .54). These data-driven methods are imperfect, but they still represent a considerable advance over the well-known limitations of prognostic assessments derived from clinical judgment (Ægisdóttir et al., 2006).

Another caveat is that we chose to apply a binary outcome (RCSI) rather than a continuous outcome measure (e.g., post-treatment PHQ-9) in our main analyses. Loss of power and information can be a disadvantage of dichotomizing continuous measures. As explained above, we have taken great care to work out a sample size calculation for a binary outcome, and to construct a test sample that would be adequately powered, using highly conservative assumptions (expecting to detect a medium effect in a subgroup as small as 25% of the sample). Any costs in terms of statistical loss of information are outweighed by the fact that the RCSI definition gives us a better grasp of individualised outcomes, since we can count and estimate the proportion of individuals who attained full remission of symptoms. As shown in Figure 3, assessing outcomes using aggregated mean differences between groups obscures the fact that some cases benefit more from optimal treatment assignment compared to others. In addition, the RCSI definition is routinely applied to assess treatment response in the target services that could implement this treatment selection algorithm, and it is also highly cited and applied in the psychotherapy literature (Jacobson & Truax, 1991). Furthermore, this stringent outcome definition prioritizes the attainment of remission of symptoms to maximize maintenance of treatment gains, rather than simply a reduction of symptomology, and is consistent with evidence that partial remission of depression symptoms is a well-known risk factor for short-term relapse (Paykel, 2008; Wojnarowski, Firth, Finegan, & Delgadillo, 2019). Another reason to apply a binary outcome definition that prioritises remission of symptoms is to overcome classification problems that

can occur when developing treatment selection algorithms using a continuous outcome measure. For example, a PAI model based on predicted post-treatment scores could classify “treatment A” as optimal even if no significant change is expected, but if “treatment B” is expected to result in deterioration. Similarly, in a scenario where deterioration is predicted in both treatments, the treatment with the least expected deterioration could be selected as optimal. Such recommendations would be unacceptable in clinical practice, especially in the context of a stepped care system where other intensive treatment options could be offered if little progress is expected with first-line psychological treatments. The algorithms that we produced would enable clinicians to determine which of these two treatment options may be advantageous for specific individuals, but also to estimate a general expected prognosis, which might in fact help to fast-track some cases to other more intensive treatments.

On balance, the evidence outlined in this study makes a persuasive case to move towards testing the efficacy of targeted prescription in prospective, experimental studies. At worst, this would be no different than the current shared decision-making strategy, with minimal associated risks. However, if the potential benefits of targeted prescription are demonstrated in an adequately powered controlled trial, this could represent a major advance in our ability to make best use of currently available evidence-based treatments, improving outcomes for patients at no additional cost to psychological services.

References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., ... & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34*(3), 341-382. https://doi.org/10.1177/0011000005285875

Beck, A. T., Rush, J. A., Shaw, B. F., Emery, G. (1979). *Cognitive therapy of depression*. New York, NY: The Guilford press.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5-32. https://doi.org/10.1023/A:1010933404324

Clark, D. M. (2018). Realizing the Mass Public Benefit of Evidence-Based Psychological Therapies: The IAPT Program. *Annual Review of Clinical Psychology, 14*, 159-183. https://doi.org/10.1146/annurev-clinpsy-050817-084833

Cohen, J. (1992). A power primer. *Psychological Bulletin, 12*, 155-159.

Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment selection in depression. *Annual Review of Clinical Psychology, 14*, 209-236. https://doi.org/10.1146/annurev-clinpsy-050817-084746

Cohen, Z. D., Kim, T. T., Van, H. L., Dekker, J. J., & Driessen, E. (2019). A demonstration of a multi-method variable selection approach for treatment selection: Recommending cognitive–behavioral versus psychodynamic therapy for mild to moderate adult depression. *Psychotherapy Research*, in press. https://doi.org/10.1080/10503307.2018.1563312

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. Irvington.

Cuijpers, P., Van Straten, A., Andersson, G., & Van Oppen, P. (2008). Psychotherapy for depression in adults: a meta-analysis of comparative outcome studies. *Journal of Consulting and Clinical Psychology, 76*(6), 909. http://dx.doi.org/10.1037/a0013075

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

Cuijpers, P., Reijnders, M., & Huibers, M. J. (2018). The role of common factors in psychotherapy outcomes. *Annual Review of Clinical Psychology, 15*. https://doi.org/10.1146/annurev-clinpsy-050718-095424

Dance, K. A., & Neufeld, R. W. (1988). Aptitude-treatment interaction research in the clinical setting: A review of attempts to dispel the "patient uniformity" myth. *Psychological Bulletin, 104*(2), 192-213.

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., Huey, D., Bennett, H., & McMillan, D. (2017). Case complexity as a guide for psychological treatment selection. *Journal of Consulting and Clinical Psychology, 85*, 835-853. http://dx.doi.org/10.1037/ccp0000231

Department for Communities and Local Government. (2011). *The English indices of deprivation 2010.* Retrieved from https://www.gov.uk/government/statistics/english-indices-of-deprivation-2010

Department of Health. *National curriculum for Cognitive Behavioural Therapy courses*, Second edition, updated and revised March 2011. Kings College London. 2011.

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*, e83875. https://doi.org/10.1371/journal.pone.0083875

DeRubeis, R. J., Gelfand, L. A., German, R. E., Fournier, J. C., & Forand, N. R. (2014b). Understanding processes of change: How some patients reveal more than others—and some groups of

therapists less—about what matters in psychotherapy. *Psychotherapy Research, 24*(3), 419-428. https://doi.org/10.1080/10503307.2013.838654

Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: The 632+ bootstrap method. *Journal of the American Statistical Association, 92*, 548-560. https://doi.org/10.1080/01621459.1997.10474007

Finegan, M., Firth, N., Wojnarowski, C., & Delgadillo, J. (2018). Associations between socioeconomic 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

Frank, J. D., & Frank, J. B. (1991). *Persuasion and healing: A comparative study of psychotherapy*, 3rd edn. Baltimore, MD: Johns Hopkins University Press.

Gifi, A. (1990). *Nonlinear multivariate analysis.* Chichester, England: Wiley.

Hill, A. (2011a). *Curriculum for counseling for depression: Continuing professional development for qualified therapists delivering high intensity interventions*. Lutterworth: 2011. BACP.

Hill, A. (2011b). *The competences required to deliver effective Counselling for Depression (CfD)*. Lutterworth: BACP. 2011. Available at https://www.ucl.ac.uk/pals/research/cehp/research-groups/core/pdfs/Counselling_for_Depression/Depression_Counselling_for_depression_clinician_s_guide.pdf.

Hinton, D. E., Rivera, E. I., Hofmann, S. G., Barlow, D. H., & Otto, M. W. (2012). Adapting CBT for traumatized refugees and ethnic minority patients: Examples from culturally adapted CBT (CA-CBT). *Transcultural Psychiatry, 49*(2), 340-365.

Huibers, M. J., Cohen, Z. D., Lemmens, L. H. J. M., Arntz, A., Peeters, F. P. M. L., Cuijpers, P., & DeRubeis, R. J. (2015). Predicting optimal outcomes in cognitive therapy or interpersonal psychotherapy for depressed individuals using the personalized advantage index approach. *PloS ONE, 10*, e0140771. https://doi.org/10.1371/journal.pone.0140771

Jacobson, N., and Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12-19.

Keefe, J. R., Wiltsey Stirman, S., Cohen, Z. D., DeRubeis, R. J., Smith, B. N., & Resick, P. A. (2018). In rape trauma PTSD, patient characteristics indicate which trauma-focused treatment they are most likely to complete. *Depression and Anxiety, 35*(4), 330-338. https://doi.org/10.1002/da.22731

Kessler, R. C., Bossarte, R. M., Luedtke, A., Zaslavsky, A. M., & Zubizarreta, J. R. (2019). Machine learning methods for developing precision treatment rules with observational data. *Behaviour Research and Therapy, 120*, 103412. https://doi.org/10.1016/j.brat.2019.103412

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,* 606-613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Lorenzo-Luaces, L., DeRubeis, R. J., van Straten, A., & Tiemens, B. (2017). A prognostic index (PI) as a moderator of outcomes in the treatment of depression: A proof of concept combining multiple variables to inform risk-stratified stepped care models. *Journal of Affective Disorders, 213*, 78-85. https://doi.org/10.1016/j.jad.2017.02.010

Luedtke, A., Sadikova, E., & Kessler, R. C. (2018). Sample size requirements for multivariate models to predict between-patient differences in best treatments of major depressive disorder. *Clinical Psychological Science*, *in press*. https://doi.org/10.1177/2167702618815466

Lutz, W., Leon, S. C., Martinovich, Z., Lyons, J. S., & Stiles, W. B. (2007). Therapist effects in outpatient psychotherapy: A three level growth curve approach. *Journal of Counseling Psychology, 54*, 32-39. http://dx.doi.org/10.1037/0022-0167.54.1.32

Martell, C. R., Addis, M. E., & Jacobson, N. S. (2001). *Depression in context: Strategies for guided action*. New York: Norton and Co.

Mulder, R., Murray, G., & Rucklidge, J. (2017). Common versus specific factors in psychotherapy: Opening the black box. *The Lancet Psychiatry, 4*(12), 953-962. https://doi.org/10.1016/S2215-0366(17)30100-1

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*, 461-464. https://doi.org/10.1192/bjp.180.5.461

Naeem, F., Waheed, W., Gobbi, M., Ayub, M., & Kingdon, D. (2011). Preliminary evaluation of culturally sensitive CBT for depression in Pakistan: findings from Developing Culturally-sensitive CBT Project (DCCP). *Behavioural and Cognitive Psychotherapy, 39*(2), 165-173.

National Institute for Health and Care Excellence, 2018. *Depression in Adults: Recognition and Management: Clinical Guideline* [CG90]. Retrieved from: https://www.nice.org.uk/guidance/cg90

Paykel, E. S. (2008). Partial remission, residual symptoms, and relapse in depression. *Dialogues in Clinical Neuroscience, 10*(4), 431.

Pearson, R., Pisner, D., Meyer, B., Shumake, J., & Beevers, C. G. (2018). A machine learning ensemble to predict treatment outcomes following an Internet intervention for depression. *Psychological Medicine*. https://doi.org/10.1017/S003329171800315X

Pybis, J., Saxon, D., Hill, A., & Barkham, M. (2017). The comparative effectiveness and efficiency of cognitive behaviour therapy and generic counselling in the treatment of depression: evidence from the 2nd UK National Audit of psychological therapies. *BMC Psychiatry, 17*(1), 215. doi: 10.1186/s12888-017-1370-7

Richards, D. A. & Borglin, G. (2011). Implementation of psychological therapies for anxiety and depression in routine practice: Two year prospective cohort study. *Journal of Affective Disorders, 133*, 51-60. https://doi.org/10.1016/j.jad.2011.03.024

Rosenbaum, P. R. (2017). *Observation and experiment: An introduction to causal inference*. Cambridge, Massachusetts: Harvard University Press.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–55. https://doi.org/10.1093/biomet/70.1.41

Roth, A. D., & Pilling, S. (2008). Using an Evidence-Based Methodology to Identify the Competences Required to Deliver Effective Cognitive and Behavioural Therapy for Depression and Anxiety Disorders. *Behavioural and Cognitive Psychotherapy, 36*(02), 129–147. https://doi.org/10.1017/S1352465808004141

Sanders, P. & Hill, A. (2014). *Counselling for Depression: A person centred and experiential approach to practice*. London: Sage.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*, 545-571. https://doi.org/10.1207/s15327906mbr3304_5

Schuler, M. S., & Rose, S. (2017). Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology, 185*(1), 65-73. https://doi.org/10.1093/aje/kww165

Smith, B., & Sechrest, L. (1991). Treatment of Aptitude × Treatment Interactions. *Journal of Consulting and Clinical Psychology, 59*(2), 233.

Smith, J., and Todd, P. (2005). Does matching overcome LaLonde's critique of non-experimental estimators? *Journal of Econometrics, 125*, 305–353. https://doi.org/10.1016/j.jeconom.2004.04.011

Snow, R. E. (1991). Aptitude-treatment interaction as a framework for research on individual differences in psychotherapy. *Journal of Consulting and Clinical Psychology, 59*(2), 205. http://dx.doi.org/10.1037/0022-006X.59.2.205

Snowden, J. M., Rose, S., & Mortimer, K. M. (2011). Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. American Journal of Epidemiology, 173(7), 731-738. https://doi.org/10.1093/aje/kwq472

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

Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*, 1285–1293. doi: 10.1126/science.3287615

Swift, J. K., Callahan, J. L., & Vollmer, B. M. (2011). Preferences. *Journal of Clinical Psychology, 67*, 155–165. https://doi.org/10.1002/jclp.20759

Swift, J. K., Callahan, J. L., Cooper, M., & Parkin, S. R. (2018). The impact of accommodating client preference in psychotherapy: A meta-analysis. *Journal of Clinical Psychology, 74*(11), 1924-1937. https://doi.org/10.1002/jclp.22680

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological), 58*, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Wampold, B. E., & Imel, Z. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work* (2nd ed.). Mahwah, NJ: Erlbaum.

Wojnarowski, C., Firth, N., Finegan, M., & Delgadillo, J. (2019). Predictors of depression relapse and recurrence after cognitive behavioural therapy: a systematic review and meta-analysis. *Behavioural and Cognitive Psychotherapy*. https://doi.org/10.1017/S1352465819000080

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

Table 1. Sample characteristics and comparisons between therapy groups

	Full sample N = 1435 (100%)	**CBT** N = 1104 (76.93%)	**CfD** N = 331 (23.07%)	**test statistic**	***p***
Demographics					
Females [2]	64.4%	63.2%	68.3%	x^2 (1)=2.84	.092
Age [1]	39.64 (12.87)	39.18 (12.73)	41.19 (13.23)	*t* (1,433)=-2.49	.013
Ethnicity [2]					
White British	88.2%	87.9%	89.4%	x^2 (1)=0.60	.439
Other	11.8%	12.1%	10.6%		
IMD decile [1]					
1 = Poorest	4.51 (2.91)	4.51 (2.91)	4.53 (2.91)	*t* (1,433)=-0.12	.906
10 = Affluent					
Unemployed [2]	29.5%	31.4%	23.3%	x^2 (1)=8.16	.004
Disabled [2]	16.9%	16.4%	18.7%	x^2 (1)=0.99	.320
Baseline severity measures					
PHQ-9 [1]	17.50 (4.50)	17.67 (4.57)	16.94 (4.25)	*t* (1433)=2.59	.010
GAD-7 [1]	13.79 (4.53)	13.92 (4.56)	13.34 (4.40)	*t* (1433)=2.07	.038
WSAS [1]	22.81 (7.97)	23.42 (7.93)	20.79 (7.74)	*t* (1433)=5.32	<.001
Diagnosis [2]					
Depressive Episode	74.9%	71.5%	86.4%	x^2 (1)=30.23	<.001
Recurrent Depression	25.1%	28.5%	13.6%		
Long term condition[2]	26.8%	25.3%	32.0%	x^2 (1)=5.92	.015
Chronicity (deciles) [1]	5.48 (2.94)	5.63 (3.00)	4.98 (2.71)	*t* (593)=3.71	<.001
Expectancy [1]	7.13 (1.82)	7.08 (1.84)	7.28 (1.79)	*t* (1433)=-1.75	.080
Previous treatment [2]	72.8%	73.8%	69.5%	x^2 (1)=2.42	.120
Taking medication [2]	65.6%	68.0%	57.4%	x^2 (1)= 12.73	<.001

CBT = cognitive-behavioural therapy; CfD = Counselling for Depression; IMD Decile = Index of multiple deprivation in deciles; PHQ-9 = Patient Health Questionnaire; GAD-7 = Generalized Anxiety Disorder Questionnaire; WSAS = Work and Social Adjustment Scale

[1] = Mean and Standard Deviation; [2] = Percentages

Table 2. Predictive accuracy of prognostic model vs. aptitude-treatment-interaction (ATI) model

	Training sample	Test sample	Cross-validation shrinkage
	AUC	**AUC**	**AUC difference**
Prognostic model trained using elastic net with optimal scaling			
CBT outcomes	.67	.59	.08
CfD outcomes	.71	.65	.06
ATI model trained using a random forest ensemble of 1000 decision trees			
CBT outcomes	.86	.58	.28
CfD outcomes	.87	.57	.30

Notes: CBT = cognitive behavioural therapy; CfD = Person-centred counselling for depression; outcomes = post-treatment reliable and clinically significant improvement in depression symptoms measured using the PHQ-9 questionnaire; AUC = area under the curve; cross-validation shrinkage = an index of cross-validation performance, where lower values are indicative of more stable predictions and adequate generalization to an external test sample

Table 3. Elastic Net: variable selection and regularized coefficients for each prognostic model

	CBT training sample (N=929) R-square = .09		CFD training sample (N=156) R-square = .13	
Variables	**B**	**SE**	**B**	**SE**
Gender	.000	.011	.000	.040
Age decade*	**.010**	.026	**.065**	.048
Minority ethnic group[1]	**-.002**	.019	.000	.032
Unemployed*	**-.078**	.031	**-.045**	.062
Disability*	**-.051**	.028	**.036**	.063
Recurrent depression	.000	.009	.000	.038
Comorbid long-term condition	.000	.017	.000	.042
Previous treatment	.000	.007	.000	.038
ADM[2]	.000	.010	**-.109**	.072
Baseline PHQ-9*	**-.166**	.036	**-.054**	.133
Baseline GAD-7[2]	.000	.026	**.004**	.087
Baseline WSAS*	**-.025**	.027	**-.158**	.079
IMD decile*	**.010**	.024	**-.026**	.081
Chronicity decile[2]	.000	.016	**.025**	.104
Expectancy[2]	.000	.018	**-.039**	.071

CBT = cognitive behavioural therapy; CFD = person-centred counselling; * Variables that had prognostic value for both CBT and CFD treatments; [1] variables that only predicted CBT outcomes; [2] variables that only predicted CFD outcomes; Gender reference category = male; Age was categorized into decade groups; Minority ethnic group reference category = white British; Recurrent depression diagnosis reference category = major depressive disorder; ADM = taking antidepressant medication; all baseline scores were taken before the start of the first high intensity therapy session; IMD = index of multiple deprivation, where higher values indicate more affluent (less deprived) neighbourhoods; chronicity = duration of primary diagnosis

Figure 1. Cross-validation of targeted prescription model in an external test sample

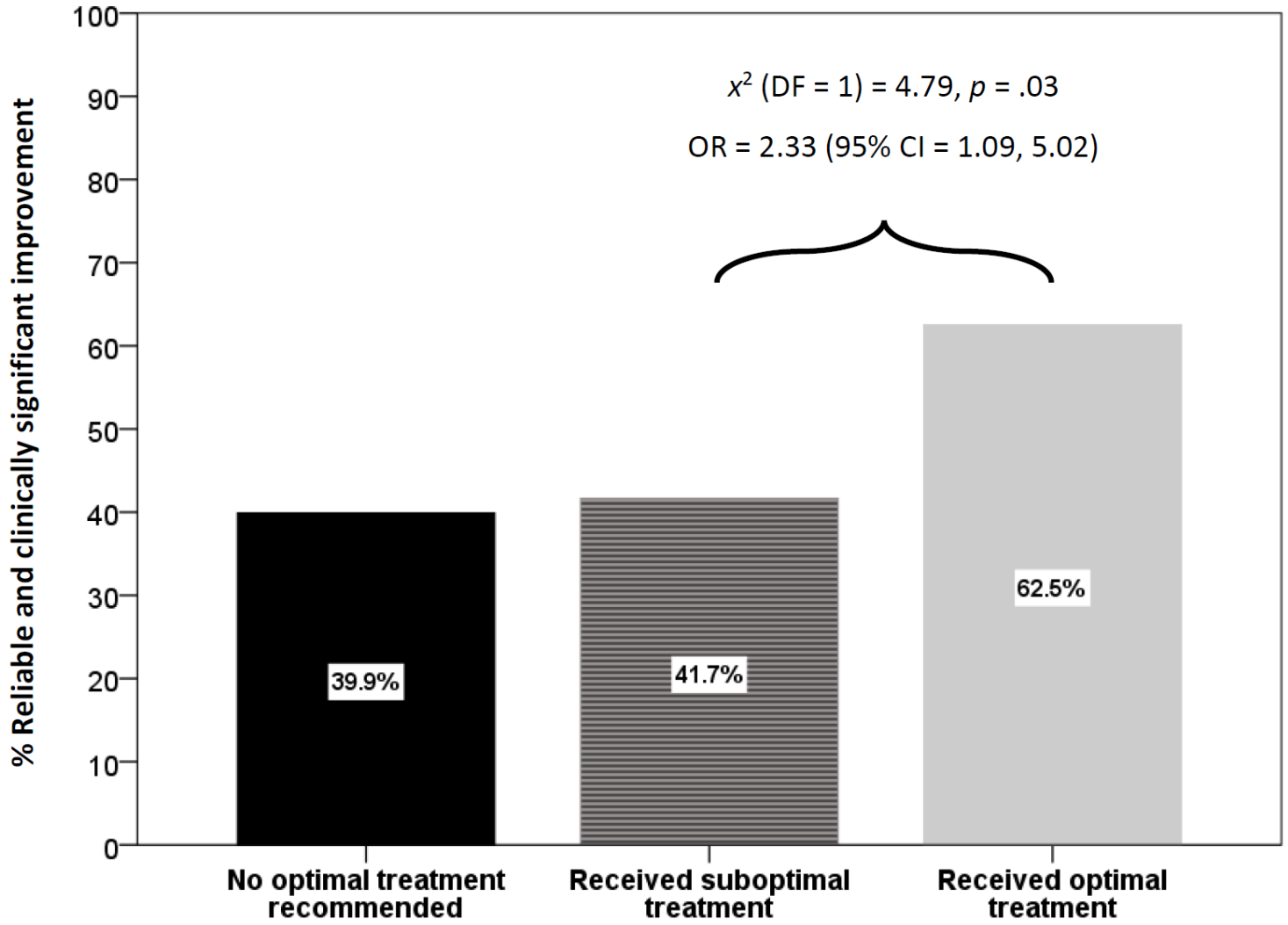

Figure 2. Distribution of personalized advantage index scores according to expected prognosis

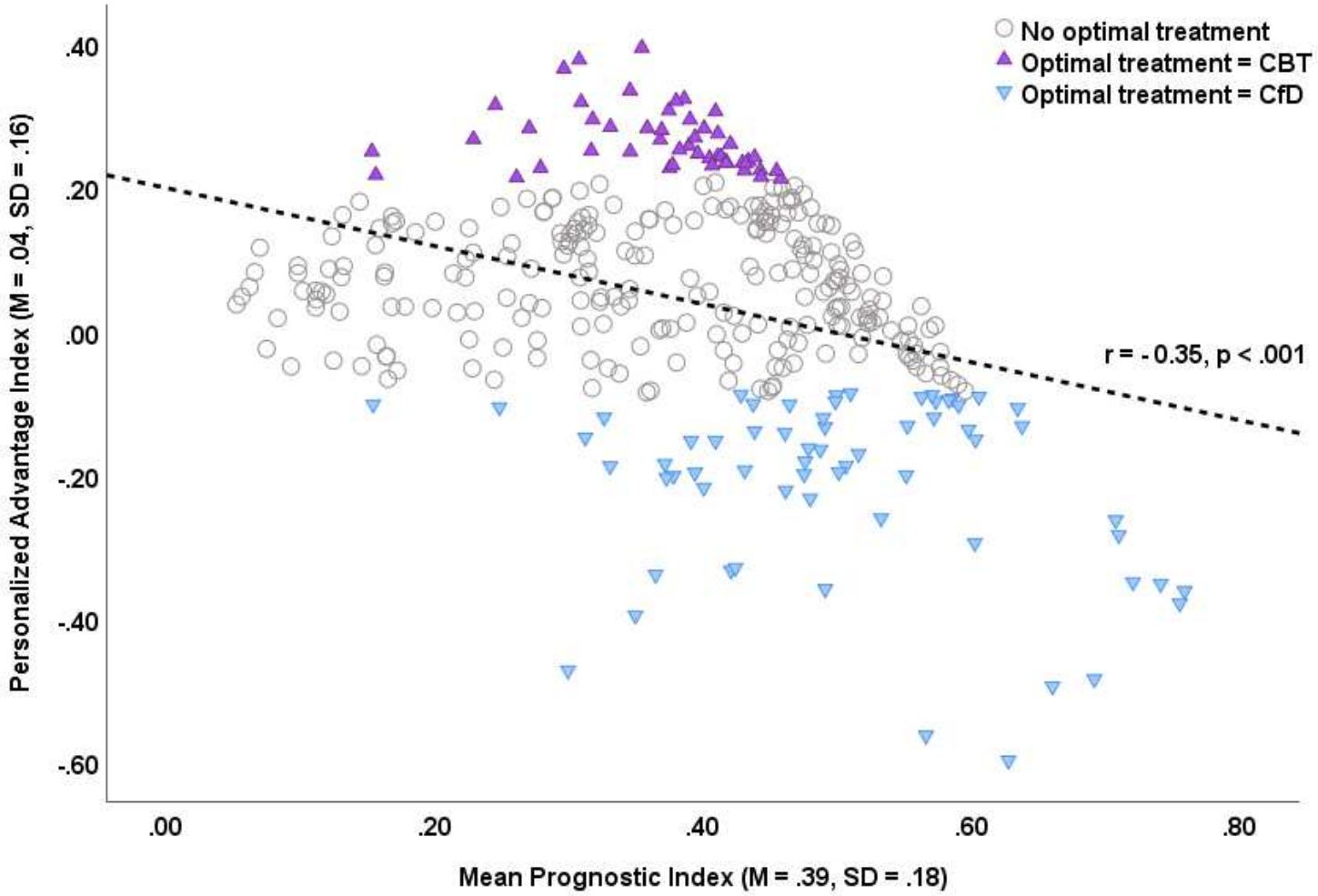

Figure 3. Between-group effect sizes (Hedges' g) for optimal vs. suboptimal treatment assignment along the MPI gradient