Article:
Delgadillo, J., Rubel, J. and Barkham, M. (2020) Towards personalized allocation of patients to therapists. Journal of Consulting and Clinical Psychology, 88 (9). pp. 799-808. ISSN: 0022-006X

https://doi.org/10.1037/ccp0000507

Citation: Delgadillo, J., Rubel, J., & Barkham, M. (*in press*). Towards Personalized Allocation of Patients to Therapists. *Journal of Consulting and Clinical Psychology*. doi: 10.1037/ccp0000507

Towards Personalized Allocation of Patients to Therapists

Jaime Delgadillo[1], Julian Rubel[2], & Michael Barkham[1]

1. Clinical Psychology Unit, Department of Psychology, University of Sheffield, United Kingdom

2. Department of Psychology, Justus-Liebig-University Giessen, Germany

Declarations of interest: None.

Correspondence: jaime.delgadillo@nhs.net

Abstract

Objective: Psychotherapy outcomes vary between therapists, but it is unclear how such information can be used for treatment planning or practice development. This proof-of-concept study aimed to develop a data-driven method to match patients to therapists.

Methods: We analyzed data from N=4,849 patients who accessed cognitive behavioral therapy in UK primary care services. The main outcome was post-treatment reliable and clinically significant improvement (RCSI) in the PHQ-9 depression measure. Machine learning analyses were applied in a training sample (N=2425 patients treated by 68 therapists in year 1), including a Chi-Squared Automatic Interaction Detector (CHAID) algorithm and a Random Forest (RF) algorithm. The predictive models were cross-validated in a statistically independent test sample (N=2,424 patients treated by the same therapists in year 2) and evaluated using odds ratios (OR) adjusted for baseline depression severity.

Results: We identified subgroups of therapists that were differentially effective for highly specific subgroups of patients, yielding 17 *classes* of patient-to-therapist matches. The overall base rate of RCSI in the sample was 40.4%, but this varied from 10.5% to 69.9% across *classes*. Cases classed by the prediction algorithms as expected responders in the test sample were ~60% more likely to attain post-treatment RCSI compared to those classed as non-responders [adjusted odds ratios = 1.59, 1.60; p<.001].

Conclusions: Machine learning approaches could help to improve treatment outcomes by enabling the strategic allocation of patients to therapists and therapists to supervisors.

Keywords: depression; cognitive-behavioral therapy; precision medicine; therapist effects

Public health significance: It is well known that, even when they apply the same treatment model, some therapists attain better clinical outcomes compared to others. Using data from a large (N=4,849) naturalistic cohort of patients who accessed highly standardized cognitive behavioral therapy for common mental disorders, the present study shows that specific therapists are more or less able to help specific subgroups of patients. We developed machine learning algorithms that pinpoint the profiles of patients that could be matched to specific therapists in order to improve treatment outcomes. An additional possibility is to use this model to support practice development, by matching therapists to peer supervisors who evidently attain better outcomes with specific profiles of patients.

Precision medicine describes an endeavour that aims to empirically identify the best available treatment option for an individual patient (Perlis, 2016). To this end, patient information on a number of different characteristics are routinely collected that go beyond patients' functional diagnoses (e.g., age, gender, treatment expectations, DNA sequences etc.). The resulting large databases can be used to identify relevant variables and profiles that are predictive of treatment outcomes. Using data from previously treated patients, precision medicine approaches can help to identify and offer the most promising treatment option for a new patient (e.g., Hamburg & Colins, 2010).

In mental healthcare, recovery rates have been stagnating for the last 40 years despite the continuous development of new treatment options (e.g., Johnsen & Friborg, 2015). Precision mental healthcare (or *personalized mental health*) has been proposed as a promising way to enhance psychological treatment outcomes, enabled by technological and methodological advancements in computing capacity and machine learning methods (Lutz, Rubel, Schwartz, Schilling, & Deisenhofer, 2019). So far, the translation of precision medicine approaches in mental healthcare have mostly focused on *personalized treatment selection*, involving the data-driven selection of specific treatment packages or techniques for specific patients (see review by Cohen & DeRubeis, 2019). This research has led to a growing body of evidence that shows that multivariable models can help to predict differential treatment response, potentially enabling a more precise allocation of patients to different treatment options. For example, recent studies have developed machine learning algorithms that enable the personalized selection of cognitive behavioral therapy (CBT) vs. psychodynamic therapy (Cohen, Kim, Van, Dekker, & Driessen, 2019) or CBT vs. person-centered counseling for depression (Delgadillo & Gonzalez Salas Duhne, 2020). Using such algorithms, a psychological service could –for example– prescribe CBT for a specific patient based on their expected likelihood of treatment response. But what if the patient ends up being treated by a CBT therapist who fails to implement therapy in a way that best fits with the patient's problems and circumstances?

An important drawback of emerging treatment selection methods is that they focus on predicting variability in treatment response *between treatment models*, but ignore the considerable variability in treatment outcomes *between therapists*. Numerous reviews of psychotherapy studies indicate that treatment outcomes vary considerably across therapists, even if they deliver the same model of psychological therapy in carefully controlled trials (Castonguay & Hill, 2017; Norcross & Lambert, 2019; Wampold & Imel, 2015). Meta-analyses report that about 5% to 8% of outcome variance is attributable to systematic differences between therapists – referred to as *therapist effects* (e.g., Baldwin & Imel, 2013; Johns, Barkham, Kellet, & Saxon, 2019). Ignoring the therapist variable in precision mental healthcare approaches leaves out an important determinant of psychological treatments and could thus lead to biased and non-generalizable prediction models. Given the importance of therapists in the therapeutic process, it is important to develop therapist-specific prediction models that tell us which therapist in a given treatment setting would be most effective for a patient with a specific set of characteristics. We consider that such patient-therapist matching models would be complementary to patient-treatment selection models, potentially enabling mental healthcare services to address a classic challenge in psychotherapy (Paul, 1967), which is to determine how to make evidence-based decisions about *what* treatment should be offered and *who* should deliver it for individuals with specific problems and circumstances.

While many authors have taken interest in understanding the features and practices of outlier therapists that attain above-average outcomes (Castonguay & Hill, 2017; Heinonen & Nissen-Lie, 2019; Norcross & Lambert, 2019), less attention has been devoted to investigating the extent to which therapist effects are stable over time and across patients with different characteristics. Golderg et al. (2016) observed a general tendency for therapists to become less effective over time, although this trend varied considerably between therapists. There is some evidence that therapists' effectiveness differs as a function of certain patient features and outcome domains. For example, Saxon and Barkham (2012) found that some therapists are more effective than others at helping patients with severe symptoms and suicidal risk, in a sample of 10,786 patients treated by 119

therapists. Kraus et al. (2011) found patterns of differential therapist effectiveness depending on their patients' problem domain in a sample of 6,960 patients treated by 696 therapists. While about 4% of therapists were not effective in any of the outcome domains (sexual functioning, work functioning, violence, social functioning, panic/anxiety, substance abuse, psychosis, quality of life, sleep, suicidality, depression, and mania), not a single therapist was effective in every domain. Interestingly, this domain-specific effectiveness has been shown to be relatively stable over time (Kraus et al., 2011, 2016). That is, if therapists at one time were successful in treating depression but less successful treating substance use disorder, this was also the case at a later time point. However, there is also some conflicting evidence, suggesting that therapist effects may be global rather than domain-specific. Using multilevel factor analysis in two samples (n_1 = 5.828 and n_2 = 616), Nissen-Lie and colleagues (2016) reported good fit for a model that included a latent global therapist variable that explained large parts of the covariation in the subscales of the OQ-45 (sample 1; subscales: symptom distress, interpersonal relationships, and social role performance) and the CORE outcome measure (sample 2; subscales: subjective well-being, depression, anxiety, trauma, physical symptoms, close relationships, general functioning, social functioning, risk to self, and risk to others). However, even in the presence of therapist effects as a rather global construct, there still may be patient characteristics or profiles, with which some therapists are more or less effective than others.

The present proof-of-concept study is the first to develop and test a prediction model that specifically includes therapists in the modeling procedure. By doing so, we aimed to empirically match patients to therapists in a way that might improve treatment outcomes. Specifically, we aimed to address the following research questions: First, are therapist effects stable over time? Second, can we identify therapists who are more or less effective for certain subgroups of patients in order to derive therapist-specific predictions? Third, would such a model trained using data at a given point in time make valid predictions for new patients treated in the future? To this end, we used patient-reported outcomes data collected over a one-year period to develop a prediction model and validated this model with data from the subsequent year with the same therapists. We

hypothesized that we would identify subgroups of therapists who attained differential outcomes with specific subgroups of patients.

Method

Setting and Interventions

This study was based on the analysis of a routine practice dataset collected during a two-year period (2013 – 2015) across five UK National Health Service (NHS) Trusts in the north of England. These services covered a large and socioeconomically diverse population covering West Yorkshire, South Yorkshire and Cumbria. Approval for the analysis of this multi-service dataset was obtained from an NHS research ethics committee and from the Health Research Authority (REC Reference: 15/NE/0062).

The participating services delivered evidence-based low and high intensity psychological interventions for depression and anxiety disorders, organised in a stepped care model (National Institute for Health and Care Excellence, 2011). Most patients with mild-to-moderate symptoms initially accessed low intensity (≤ 8 sessions) guided self-help delivered by qualified practitioners. Patients whose symptoms persisted after low intensity interventions, and those with specific conditions for whom specific psychotherapies are indicated (e.g., post-traumatic stress disorder), were offered high intensity interventions such as cognitive behavioral therapy (CBT), person-centered counseling, and eye-movement desensitization and reprocessing (EMDR). Further details about this stepped care model and available interventions are described by Clark (2018).

High intensity CBT in these services followed standardized disorder-specific interventions listed in the Roth and Pilling (2008) competency framework. Qualified CBT therapists were trained to apply evidence-based treatment protocols for depression, generalized anxiety disorder, post-traumatic stress disorder, obsessive-compulsive disorder, social anxiety disorder, panic disorder, phobias and other common mental health problems. Therapists received regular clinical supervision by experienced peers in their service, equivalent to 1 hour per week of full-time practice. The

present study sample only included cases that accessed high intensity CBT in the stepped care system.

Measures

The primary outcome measure was the Patient Health Questionnaire (PHQ-9), which patients completed on a weekly basis before each therapy session. The PHQ-9 is a nine-item screening tool for depression symptoms (Kroenke, Spitzer, & Williams, 2001). 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 to 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).

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). Furthermore, a standardised set of demographic and clinical information was gathered for all patients at initial assessments: age, gender, ethnicity, employment status, index of multiple deprivation (IMD), use of antidepressant medication, primary diagnosis, and prior access to a low intensity intervention before starting CBT. The IMD is an area-level index of socioeconomic deprivation, which ranks patients' neighbourhoods into decile groups (Department for Communities and Local Government, 2011). Primary diagnoses were established at the time of initial assessment interviews, which involve the screening of common mental disorders using a set of validated case-finding measures described in the Improving Access to Psychological Therapies Manual (National Collaborating Centre for Mental Health, 2018).

Sample Selection and Characteristics

Overall, 48,698 patients accessed stepped care psychological interventions across the participating services during a two-year period, of whom 13,158 (27.0%) received individual CBT. Approximately 30.6% of CBT cases had received prior low intensity guided self-help and were stepped

up for additional treatment, while the rest were directly assigned to CBT at the start of their treatment pathway. Around 77.6% of CBT patients completed treatment and 22.4% dropped out. Further details about the stepped care pathway and wider sample characteristics are available elsewhere (Finegan, Firth, & Delgadillo, 2019).

Informed by sample size guidelines for the investigation of therapist effects (Schiefele et al., 2017), the study sample only included a subset of cases treated by therapists who had a minimum caseload of 20 patients per year. This requirement was based on an expected therapist effect in the region of 4% to 5%, which is typical of practice-based studies (Johns et al., 2019). In addition, the study sample was restricted to cases that attended a minimum of two CBT sessions in order to derive separate baseline and post-treatment outcome measures. Furthermore, we only included cases that had case-level symptoms on the primary outcome measure (PHQ-9 ≥ 10) at their initial CBT session to minimise confounding due to floor effects and to estimate clinically significant response rates. These selection rules yielded a study sample of N=4,849 patients treated by 68 therapists across two years (year 1 N=2,425, year 2 N=2,424). Patients were allocated to the first available therapist as they progressed through a waiting list after initial assessments, or after completing low intensity guided self-help in the case of those who were stepped up. Sample characteristics are summarized in Table 1.

Data Analysis

The primary objective of the analysis was to develop a method to predict treatment outcomes for patients treated by specific therapists. To achieve this, the analysis was conducted in four steps: (1) pre-processing of available data; (2) preliminary examination of therapist effects; (3) model development in a training sample; and (4) model validation in a test sample. A step-by-step explanation of the analysis is outlined below, along with a detailed glossary of technical terms which is available as online-only supplemental material.

Pre-processing. The full dataset was partitioned into two subsets which grouped cases that completed treatment in year 1 (training sample) and year 2 (test sample). This enabled us to examine

the performance of a prediction model in a statistically independent test sample of patients treated one year later, and also to examine the temporal stability of therapist effects. Missing demographic and clinical features (<10% for each feature) were imputed separately in the training and test samples by averaging 25 iterations of a Monte Carlo Markov Chain model into a single imputed dataset, which performs well with continuous, categorical, and mixed-type variables (Gilks, Richardson, & Spiegelhalter, 1995).

Examination of therapist effects. We applied a conventional multilevel modeling approach recommended to assess therapist effects (Baldwin & Imel, 2013), enabling the calculation of an intra-cluster correlation coefficient (ICC) as a measure of variability attributable to therapists. In this analysis, cases (level 1) were nested within therapists (level 2), the dependent variable was post-treatment PHQ-9 severity, a random intercept was included for therapists, and significant prognostic features (all patient characteristics in Table 1, except treatment duration and dropout status) were selected by backward elimination using a threshold of $p < .05$. A caterpillar plot of therapist residuals and 95% confidence intervals was applied to rank therapists according to the expected vs. observed outcomes in their caseloads (Saxon & Barkham, 2012). Rank (Spearman's) correlations were computed to assess the temporal stability of therapist ranking and the Kappa statistic was applied to compare the therapists' classification (better than average, average, worse than average) between years 1 and 2. The therapists' identifier (a categorical variable) was entered in subsequent analyses, so that we could examine the features of patients that respond better to therapists who are ranked differently using conventional multilevel modeling.

Model development. Using the training sample, we developed decision trees that statistically model interactions between the therapist nesting variable and patients' features, in order to predict post-treatment outcomes for patients with specific profiles. Predictors entered into the decision tree models included all patient-features listed in Table 1, except for process (treatment duration) and

outcome variables (dropout, post-treatment symptoms).[1] The outcome of interest was reliable and clinically significant improvement (RCSI) in depression symptoms (Jacobson & Truax, 1991), which required an improvement of ≥6 points and post-treatment severity in the sub-clinical (PHQ-9 < 10) range. This binary outcome was preferred to a continuous PHQ-9 score for three reasons. First, this would ensure that the model is able to match patients to therapists that will lead to a statistically reliable improvement by comparison to other therapists, and not just a clinically trivial difference in post-treatment symptoms. Second, this model would prioritise full remission of symptoms to recommend a treatment-match that is likely to lead to the best possible long-term outcome, considering that residual depression is a well-established predictor of relapse (Wojnarowski, Firth, Finegan, & Delgadillo, 2019). Third, this outcome definition is widely used to assess the effectiveness of treatment in the services that participated in this study, hence having direct applicability in this treatment system.

First, we developed a decision tree using a Chi-squared Automatic Interaction Detector (CHAID) algorithm (Kass, 1980). CHAID discovers interactions between a set of potential predictor variables through a recursive partitioning process that finds an optimal cut-point in continuous variables and merges categorical variables into homogeneous subgroups of cases, with reference to a target outcome (RCSI). Splitting was determined using Pearson's chi-square statistic with a Bonferroni adjustment of significance values, with a maximum tree depth of 5 levels and a minimum parent-to-child node sample size of 100:50. Each partition was trained by selecting an optimal solution within 1,000 iterations, which maximized homogeneity within child nodes and minimized prediction error (measured by the *risk* estimate). By forcing the therapist identifier as the first variable entered into the tree, the model finds homogeneous groups of therapists with a similar base rate of RCSI and identifies patient features (child nodes) that interact with each therapist group. The CHAID algorithm can derive more than two child nodes per split, potentially yielding a wider tree than other binary

[1] Only variables available before the start of high intensity CBT were entered as predictors, since we aimed to develop a prediction model that could inform the selection of therapists a priori. This is why variables available after the start of treatment (i.e., process and outcomes data) were not included as predictors.

partitioning approaches, resulting in highly specific profiles of cases. Compared to other –more complex– machine learning approaches, CHAID has the advantage of yielding a fairly simple and highly interpretable decision tree model.

Informed by the above results, we trained an ensemble of 1,000 decision trees using a random forest (RF) approach (Breiman, 2001), including the five groups of therapists discovered by the CHAID algorithm as predictors.[2] We transformed categorical variables into a set of binary features using one-hot encoding. Each tree grows on a bootstrap sample, which was obtained by sampling cases with replacement. For each node, the best split variable was selected from a subset of randomly drawn variables from the full set of potential predictors. Optimization was applied using the RBFOpt technique, which automatically explores a search space of potential hyper-parameters (e.g., number of observations drawn randomly for each tree, number of variables drawn randomly for each split, the splitting rule, etc.), building a series of models and comparing the models to derive optimal settings (Costa & Nannicini, 2018). Cross-validation using out-of-bag samples was applied to estimate generalization error and to tune hyper-parameters. Once the model is trained, a predicted outcome (RCSI = 0 or 1) is computed for each node across all trees. To predict the target value for an incoming case, the RF model finds in which terminal nodes it falls, and then combines the classifications of these nodes for the final prediction using a "voting" method. Compared to simpler decision tree approaches, RF is designed to minimize overfitting and to maximize out-of-sample generalizability (Breiman, 2001). However, it is also criticized by being a computationally intensive "black box" model (Guidotti et al., 2018), yielding results that cannot be easily explained since it aggregates predictions across a large number of decision trees that vary in their structure (i.e., tree depth, selected predictors) and predicted classifications.

Model validation. We applied the CHAID and RF models to classify all cases in the statistically independent test sample according to their predicted outcome (RCSI = 0 or 1). We compared actual

[2] Preliminary model fitting in the training sample confirmed that entering the therapist groups identified via CHAID improved prediction accuracy compared to entering individual therapist identifiers as binary predictors.

(observed) RCSI rates between cases classified as responders vs. non-responders using odds ratios adjusted for baseline PHQ-9 severity. In addition, we calculated sensitivity and specificity indices for each model. This enabled us to compare the relative performance of a simpler (CHAID) versus a more complex (RF) prediction model.

Results

Therapist Effects

Figure 1 displays a caterpillar plot derived from the training sample, which ranks therapists from least to most effective during year 1. Eight outliers were identified at the extreme ends of the distribution; four labeled as better than average (BTA) and four labeled as worse than average (WTA). Approximately 4% of variability in post-treatment depression severity was attributable to therapist effects (ICC = .04), after adjusting for statistically significant patient features in a multilevel model: employment status, diagnosis, use of medication, age and baseline severity in PHQ-9, GAD-7 and WSAS. Following the same procedure in the test sample (year 2), the therapist effect was approximately 3% (ICC = .03) after adjusting for case-mix, and eight outliers were identified.

There was a moderate correlation between the rank order of therapists across years 1 and 2 (r = .37, p = .002). The classification of less effective therapists was not temporally stable, since all four WTA therapists in year 1 were ranked as average in year 2, and others previously ranked as average shifted into the WTA category. However, three out of four BTA therapists in year 1 remained in the BTA category in year 2.

Model Development

Figure 2 displays the CHAID decision tree developed in the training sample. The root node displays the overall base rate of cases meeting post-treatment RCSI criteria (40.4%). The first level of nodes that branch out of the root identify five groups of therapists with similar RCSI rates each. The three most effective therapists that remained in the BTA category across years 1 and 2 were clustered together (Node 5), attaining the highest RCSI rate (69.9%) and no interactions were identified with patient features. The four other groups of therapists interacted with patient features

that split into highly specific subgroups of cases with differential RCSI rates. Overall, the CHAID algorithm identified 17 *classes* of patient-to-therapist matches, which correspond to the terminal nodes of the decision tree (i.e., where each branch stops splitting) shown in Figure 2. The overall base rate of RCSI in the sample was 40.4%, but this varied from 10.5% to 69.9% across *classes*. Features selected into the model included employment status, use of antidepressant medication, and baseline PHQ-9, GAD-7 and WSAS.

The RF ensemble selected the following features and ranked them in order of importance, as illustrated in Figure 3: WSAS, PHQ-9, employment status, age, GAD-7, IMD, CHAID therapist clusters 5-3-1-4, medication, diagnosis: generalized anxiety disorder, diagnosis: affective disorder, gender, prior guided self-help, diagnosis: mixed anxiety and depression. Each of the features listed after the therapist clusters had a negligible importance weight in the model (< 0.02).

Model Validation in the Test Sample

Cases classed by the prediction algorithms as expected responders in the test sample were ~60% more likely to attain post-treatment RCSI compared to those classed as non-responders. The CHAID [adjusted OR = 1.59 (95% CI: 1.34, 1.90), $p < .001$] and RF algorithms had highly similar performance indices [adjusted OR = 1.60 (95% CI: 1.29, 1.98), $p < .001$]. Both models had low sensitivity to identify responders (CHAID: 38%, RF: 25%) but high specificity to identify non-responders (CHAID: 73%, RF: 85%).

Discussion

This proof-of-concept study developed and tested a therapist-patient matching method, utilizing a large routine-practice dataset collected over a 2-year period. As such, it builds on work investigating the temporal stability of therapist effects, and the extent to which the data from therapists can be used to predict their future patients' outcomes (Kraus et al., 2016). Our first research question concerned the stability of therapist effects. We found some consistency in therapists' performance over time, such that three of the four "better than average" (BTA) therapists remained in this category at year 2, suggesting that therapists achieving exceptional outcomes at one given time-

point are likely to do so again. This fits with prior studies examining the relative stability of therapist effects over time (e.g., Wampold & Brown, 2005). More recently, Owen et al. (2019) reported that therapists who consistently achieved better outcomes at one point in time also did so with their subsequent clients. However, the sample was drawn from university counselling centers and it is not clear whether findings from such a setting generalize to mainstream adult psychotherapy services. Taken together, these emerging studies indicate that highly effective therapists remain effective over time, and this is consistent in university and primary care mental health services.

In contrast, the finding of consistency did not hold for "worse than average" (WTA) therapists, all of whom were ranked as average in year 2. Furthermore, some "average" therapists in year 1 were classed as WTA in year 2. This evidence shows that the relative ranking of most therapists tends to shift over time. This is likely to be explained as a function of the interaction between therapist and patient-level features. Although the relative ranking of most therapists had moderate temporal stability (r = .37), the therapist-specific prediction models trained using data from year 1 generalized well to cases treated by these therapists in year 2. This suggests that the relative ranking of most therapists was partly a function of their patients' characteristics and the extent to which they had "adequately matched" cases assigned to them. However, this evidently does not apply to three therapists who were consistently "better than average". Unfortunately, this dataset did not contain any information about the therapists' features or practice, so it was not possible to derive any further insights about the characteristics of highly effective therapists. In this regard, we refer readers to recent reviews of this literature that discuss the interpersonal skills and attitudes of highly effective therapists (Castonguay & Hill, 2017; Heinonen & Nissen-Lie, 2019), which suggest that they are able to relate to patients with diverse features and often complex problems.

Regarding our second research question, we were able to identify discrete subgroups of therapists –five in all– with clearly differentiating patient variables that yielded substantially differing rates of change both between and within subgroups. Therapists in Node 5 (see Figure 2) included three "better than average" therapists (5.5%) treating 133 patients, attaining an RCSI rate of 70%. The

absence of any interaction with other variables indicated that these therapists were effective almost regardless of which patients were assigned to them. Or, more precisely, there were no patient variables available in this dataset that led to a differential response within this group of therapists. This does not preclude an as yet unmeasured variable yielding differential outcomes in a future cohort. The RCSI rate for this subgroup was only exceeded (74%) by the 19 therapists in Node 1 when assigned patients who were employed and had an intake PHQ-9 scores in the range 13 to 16, equating to moderate depression severity.

The largest group of therapists, in Node 4, comprised 26 therapists whose best RCSI rate of 44% was achieved with employed patients with a baseline GAD-7 score of less than 19. This accounted for 22% of the sample. Therapists classed in Node 2 did best with patients scoring less than 19 on the WSAS, or if their score was over 19, that they were not in receipt of medication. The five therapists classed in Node 3 included three who were classed as "worse than average", and their patients' RCSI rates ranged between 10% and 31%. Clearly, treatment outcomes varied considerably across patient subgroups, and this variability was partly a function of adequate or inadequate therapist-patient matches. The treatment allocation system in this setting was quasi-random (not strategic), driven by the haphazard nature of waiting lists and therapist availability, and so it is evident that strategic and personalized allocation could be a viable way to improve treatment outcomes.

Finally, regarding our third research question, the decision tree algorithms developed in the training sample (year 1) generalized to a statistically independent test sample (year 2), albeit with a small-to-moderate effect size. Our findings indicate that the prediction accuracy of both machine learning models (CHAID and RF) was highly comparable. However, CHAID is the least complex and most interpretable of the two approaches, since it outputs a fully explainable decision tree model – as exemplified by Figure 2. Nevertheless, the results of the RF model are instructive, since they reveal that specific patient-level variables are generally more important outcome predictors compared to the therapist, but interactions between the therapist nesting variable and patient-features can be leveraged to derive generalizable treatment matching rules.

Strengths and Limitations

Strengths of this study include the large, multi-service and adequately powered sample size to model therapist effects, the rigorous external cross-validation of prediction models in a new sample of patients treated one year later, and the comparison of more and less complex decision tree approaches. Furthermore, CBT in the participating services was highly standardized, protocol-driven and closely supervised, which minimizes the chances that therapist effects may be confounded with effects attributable to different treatment models. On the other hand, we cannot assume that the present results generalize to other settings or treatment modalities. The study was also limited by the availability of relatively few patient-features, and no therapist-level features.

Implications for Practice and Future Research

This proof-of-concept study demonstrates how psychological services could move towards matching individual patients to therapists, bringing us closer to delivering on the challenge of precision mental healthcare articulated by Gordon Paul over 50 years ago (Paul, 1967). These findings may pave the way for prospective, experimental studies testing the efficacy of targeted allocation of patients to therapists. Additional possibilities to leverage these kinds of information include the identification of therapist-specific training needs and the effective make-up of peer-supervision groups in which participating therapists complement each other regarding their specific strengths and weaknesses.

The decision tree model illustrated in Figure 2, for example, could be used to (a) identify the terminal nodes that best characterize a new patient presenting for treatment in this specific treatment context, and (b) to easily estimate the probability of improvement for that patient if they were allocated to therapists in each of the five groups. This method could enable services to make decisions about which therapists to allocate the patient to; or decisions about which therapists could mentor or advise other therapists regarding patients with specific features. Such a method could be combined with other empirically supported personalization tools in a sequential decision-making process, where

the first step involves selecting the optimal treatment model (e.g., CBT vs. interpersonal psychotherapy), the second step determines an optimal therapist-patient match within that treatment modality, and the third step involves the use of routine outcome monitoring and feedback methods (Shimokawa, Lambert, & Smart, 2010) to ensure the treatment is adequately adjusted if problems arise.

An interesting observation is that patients' diagnoses were not selected into the CHAID model and had negligible importance in the RF model. The training and clinical supervision of CBT therapists in this treatment setting strongly emphasize attention to disorder-specific interventions (National Collaborating Centre for Mental Health, 2018), yet variability in clinical outcomes between therapists seems to be better explained by other features such as symptom severity, functioning, and socioeconomic factors. A challenge raised by the data concerns patients who are unemployed, approximately one-third of the patient sample, as this appeared to be the patient-variable most indicative of poorer outcomes for most therapists. This finding fits with prior evidence concerning the detrimental effects of unemployment on mental health (Waddell & Burton, 2006), and evidence that unemployment and socioeconomic deprivation are associated with poorer psychological treatment outcomes (e.g., Delgadillo, Moreea, & Lutz, 2016; Delgadillo & Gonzalez Salas Duhne, 2019; Finegan et al., 2019). A systematic review on this topic (Finegan, Firth, Wojnarowski, & Delgadillo, 2018) suggests that people who are unemployed and living in socioeconomically deprived circumstances tend to benefit less from psychological treatment, owing to persistent stressors that may be unresolved through therapy (i.e., debt, material deprivation, exposure to crime in the neighbourhood, etc.) but also psychological factors that may not be adequately addressed (i.e., perceived low social status may contribute to demoralisation and a sense of lack of control over one's life and future). We note, however, that unemployed patients did not have a poorer prognosis if they were treated by therapists in Node 5. Assignment of such cases to these therapists might seem logical. However, such actions would likely be short-sighted in terms of creating a highly skewed caseload for those therapists. It might be more strategic to provide them with a greater role for clinical supervision of the other

therapists with a targeted focus disseminating strategies and techniques for working with patients who are unemployed.

Conclusions

Overall, the findings confirm that there is considerable variability in therapist effectiveness, despite the fact that all therapists in this sample delivered evidence-based and highly standardized CBT interventions. The RCSI rates of patient subgroups ranged from approximately 10% to 70%, which was discovered using a highly granular investigation of therapist variability in routine care. The methods described in this study could help to optimize treatment allocation processes and, therefore, attain the best likely outcome for patients and, as a consequence, improve the overall rates of change in a clinic or service. The application of modern machine learning analyses represents a major step-change in our ability to understand *therapist effects* observed in clinical practice – which are influenced at least partly by the adequacy of patient-to-therapist matching.

Acknowledgements

This study was supported by the Northern IAPT Practice Research Network (www.iaptprn.com), a collaboration between academic researchers and psychological services in the north of England. Ethical approval, access permissions and data sourcing were enabled by Jaime Delgadillo, Mike Lucock, Michael Barkham, Dean McMillan, Gillian Donohoe, Stephen Kellett, Sarah Mullaney, Richard Thwaites. The development of the IAPT PRN dataset was supported by NHS Research Capability Funding from the West Yorkshire Clinical Commissioning Groups (Reference: RCF 2014 010). The dataset was used for research purposes with the approval of the Health Research Authority (REC Reference: 15/NE/0062).

References

Baldwin, S. A., & Imel, Z. E. (2013). Therapist effects: Findings and methods. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 258–297). New York, NY: Wiley.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123–140. https://doi.org/10.1007/BF00058655

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

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984). *Classification and Regression Trees*. Boca Raton: CRC press.

Castonguay, L., & Hill, C. E. (Eds.). (2017). *How and why are some therapists better than others? Understanding therapist effects*. Washington, DC: American Psychological Association. https://doi.org/10.1037/0000034-000

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, 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. M., & 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*. Advance online publication. http://dx.doi.org/10.1080/10503307.2018.1563312

Costa, A., & Nannicini, G. (2018). RBFOpt: an open-source library for black-box optimization with costly function evaluations. *Mathematical Programming Computation, 10*, 597-629. https://doi.org/10.5281/zenodo.597767.

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*, 14–24. https://psycnet.apa.org/doi/10.1037/ccp0000476

Delgadillo, J., Moreea, O., & Lutz, W. (2016). Different people respond differently to therapy: A demonstration using patient profiling and risk stratification. *Behaviour Research and Therapy, 79*, 15-22. https://doi.org/10.1016/j.brat.2016.02.003

Department for Communities and Local Government (2015). English Indices of Deprivation. Retrieved from https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015.

Finegan, M., Firth, N., & Delgadillo, J. (2019). Adverse impact of neighbourhood socioeconomic deprivation on psychological treatment outcomes: the role of area-level income and crime. *Psychotherapy Resea*rch, in press. https://doi.org/10.1080/10503307.2019.1649500

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*, 560–573. http://dx.doi.org/10.1002/da.22765

Gilks, W.R., Richardson, S., Spiegelhalter, D. (1995). *Markov Chain Monte Carlo in practice*. New York: Chapman and Hall/CRC. https://doi.org/10.1201/b14835

Goldberg, S. B., Rousmaniere, T., Miller, S. D., Whipple, J., Nielsen, S. L., Hoyt, W. T., ... Wampold, B. E. (2016). Do psychotherapists improve with time and experience? A longitudinal analysis of outcomes in a clinical setting. *Journal of Counseling Psychology, 63*, 1-11. doi:10.1037/cou0000131

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys, 51(93)*, 1-42. https://doi.org/10.1145/3236009

Hamburg, M. A., & Collins, F. S. (2010). The path to personalized medicine. *New England Journal of Medicine, 363*, 301-304. https://doi.org/10.1056/NEJMp1006304

Heinonen, E. & Nissen-Lie, H.A. (2019). The professional and personal characteristics of effective psychotherapists: a systematic review. *Psychotherapy Research*. Advance online publication. https://doi.org/10.1080/10503307.2019.1620366

Huibers, M. J., Cohen, Z. D., Lemmens, L. H., Arntz, A., Peeters, F. P., 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*(11), e0140771. https://doi.org/10.1371/journal.pone.0140771

Jacobson, N., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12-19. http://dx.doi.org/10.1037/10109-042

Johns, R. G., Barkham, M., Kellett, S., & Saxon, D. (2019). A systematic review of therapist effects: A critical narrative update and refinement to review. *Clinical Psychology Review, 67*, 78-93. https://doi.org/10.1016/j.cpr.2018.08.004

Johnsen, T. J., & Friborg, O. (2015). The effects of cognitive behavioral therapy as an anti-depressive treatment is falling: *A meta-analysis. Psychological Bulletin, 141*, 747–768. https://doi.org/10.1037/bul0000015

Kraus, D., Castonguay,L. G., Boswell, J. F., Nordberg, S. S., & Hayes, J. A. (2011) Therapist effectiveness: Implications for accountability and patient care. *Psychotherapy Research, 21*, 267-276, DOI: 10.1080/10503307.2011.563249

Kraus, D. R., Bentley, J. H., Alexander, P. C., Boswell, J. F., Constantino, M. J., Baxter, E. E., & Castonguay, L. G. (2016). Predicting therapist effectiveness from their own practice-based evidence. *Journal of Consulting and Clinical Psychology, 84*, 473. https://doi.org/10.1037/ccp0000083

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

Lutz, W., Rubel, J. A., Schwartz, B., Schilling, V., & Deisenhofer, A. K. (2019). Towards integrating personalized feedback research into clinical practice: Development of the Trier Treatment Navigator (TTN). *Behaviour Research and Therapy, 120*, 103438. https://doi.org/10.1016/j.brat.2019.103438

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

National Institute for Health and Care Excellence (2011). Common mental health disorders: Identification and pathways to care. London: NICE. Retrieved from http://www.nice.org.uk/guidance/CG123

National Collaborating Centre for Mental Health (2018). The Improving Access to Psychological Therapies Manual. Available from: https://www.england.nhs.uk/wp-content/uploads/2018/06/the-iapt-manual.pdf

Nissen-Lie, H. A., Goldberg, S. B., Hoyt, W. T., Falkenström, F., Holmqvist, R., Nielsen, S. L., & Wampold, B. E. (2016). Are therapists uniformly effective across patient outcome domains? A study on therapist effectiveness in two different treatment contexts. *Journal of Counseling Psychology, 63*, 367-378. http://dx.doi.org/10.1037/cou0000151

Norcross, J. C., & Lambert, M. J. (Eds.). (2019). *Psychotherapy relationships that work, 3rd Edition. Volume 1: Evidence-based therapist contributions*. New York: Oxford University Press.

Owen, J., Drinane, J. M., Kivlighan III, M., Miller, S., Kopta. M., & Imel, Z. (2019). Are high performing therapists both effective and consistent? A test of therapist expertise. *Journal of Consulting and Clinical Psychology, 87*, 1149-1156. https://psycnet.apa.org/doi/10.1037/ccp0000437

Paul, G. L. (1967). Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology, 31*, 109–118. https://doi.org/10.1037/h0024436

Perlis, R. H. (2016). Abandoning personalization to get to precision in the pharmacotherapy of depression*. World Psychiatry, 15*, 228-235. https://doi.org/10.1002/wps.20345

Raudenbush, S. W. (1993). Hierarchical linear models and experimental design. In L. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 459-496). New York: Marcel Dekker.

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

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*, 129-147. https://doi.org/10.1017/S1352465808004141

Saxon, D., & Barkham, M. (2012). Patterns of therapist variability: Therapist effects and the contribution of patient severity and risk. *Journal of Consulting and Clinical Psychology, 80*, 535. http://dx.doi.org/10.1037/a0028898

Schiefele, A. K., Lutz, W., Barkham, M., Rubel, J., Böhnke, J., Delgadillo, J.… Lambert, M. J. (2017). Reliability of therapist effects in practice-based psychotherapy research: A guide for the planning of future studies. *Administration and Policy in Mental Health and Mental Health Services Research, 44*, 598-613. doi: 10.1007/s10488-016-0736-3

Shimokawa, K., Lambert, M. J., & Smart, D. W. (2010). Enhancing treatment outcome of patients at risk of treatment failure: Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *Journal of Consulting and Clinical Psychology, 78*, 298–311. https://doi.org/10.1037/a0019247

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

Waddell, G., & Burton, A. K. (2006). *Is work good for your health and well-being?* London, UK: The Stationery Office.

Wampold, B. E., & Brown, G. S. (2005). Estimating variability in outcomes attributable to therapists: A naturalistic study of outcomes in managed care. *Journal of Consulting and Clinical Psychology, 73*, 914–923. http://dx.doi.org/10.1037/0022-006X.73.5.914

Wampold, B. E., & Imel, Z. E. (2015). *The great psychotherapy debate: The research evidence for what works in psychotherapy*, 2nd ed. New York, NY: Routledge.

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, 47*, 514-529. https://doi.org/10.1017/S1352465819000080

Table 1. Sample characteristics

Characteristics	Full sample N = 4849	Year 1 N = 2425	Year 2 N = 2424
Demographics			
Mean age (SD)	41.10 (13.76)	42.00 (13.72)	40.19 (13.74)
Females (%)	3077 (63.5)	1518 (62.6)	1559 (64.3)
Unemployed* (%)	1682 (34.7)	855 (35.3)	827 (34.1)
Ethnicity			
White British (%)	4599 (94.8)	2298 (94.8)	2301 (94.9)
Other (%)	250 (5.2)	127 (5.2)	123 (5.1)
Mean IMD (SD)	4.57 (2.76)	4.57 (2.75)	4.56 (2.78)
Clinical characteristics			
Baseline PHQ-9 mean (SD)	17.86 (4.66)	17.93 (4.69)	17.79 (4.62)
Baseline GAD-7 mean (SD)	15.18 (4.30)	15.15 (4.31)	15.22 (4.28)
Baseline WSAS mean (SD)	22.23 (8.51)	22.26 (8.56)	22.20 (8.47)
Prescribed pharmacotherapy (%)	3238 (66.8)	1640 (67.6)	1598 (65.9)
Primary diagnosis			
Affective disorder (%)	2032 (41.9)	1063 (43.8)	969 (40.0)
GAD (%)	563 (11.6)	306 (12.6)	257 (10.6)
Mixed (%)	1033 (21.3)	494 (20.4)	539 (22.2)
Panic disorder / agoraphobia (%)	243 (5.0)	140 (5.8)	103 (4.2)
Social anxiety disorder (%)	252 (5.2)	101 (4.2)	151 (6.2)
Specific phobia (%)	37 (0.8)	13 (0.5)	24 (1.0)
OCD (%)	225 (4.6)	104 (4.3)	121 (5.0)
PTSD (%)	333 (6.9)	143 (5.9)	190 (7.8)
Other (%)	131 (2.7)	61 (2.5)	70 (2.9)
Prior low intensity GSH (%)	1338 (27.6)	594 (24.5)	744 (30.7)
Mean CBT sessions (SD)	8.52 (5.52)	8.19 (5.33)	8.84 (5.69)
Dropped out of CBT (%)	874 (18.0)	447 (18.4)	427 (17.6)

IMD = index of multiple deprivation (lower = more deprived); PHQ-9 = measure of depression symptoms; GAD-7 = measure of anxiety symptoms; WSAS = work and social adjustment scale; Mixed = mixed depression and anxiety; GAD = generalized anxiety disorder; OCD = obsessive-compulsive disorder; PTSD = post-traumatic stress disorder; GSH = guided self-help; CBT = cognitive-behavioral therapy

Figure 1. Caterpillar plot: ranking therapists according to their effectiveness in year 1

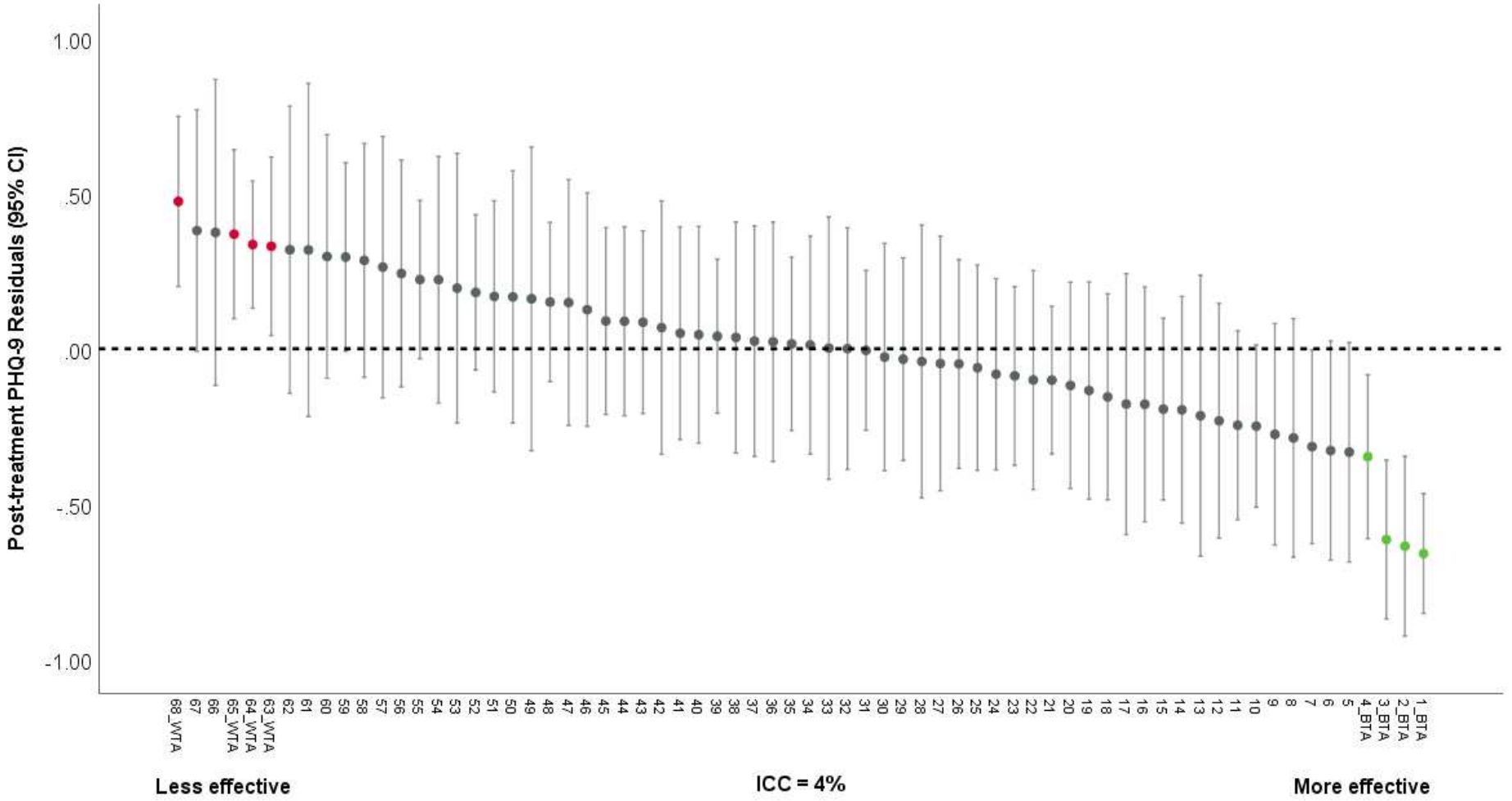


Figure 2. Decision tree trained using a Chi-squared Automatic Interaction Detector (CHAID) algorithm

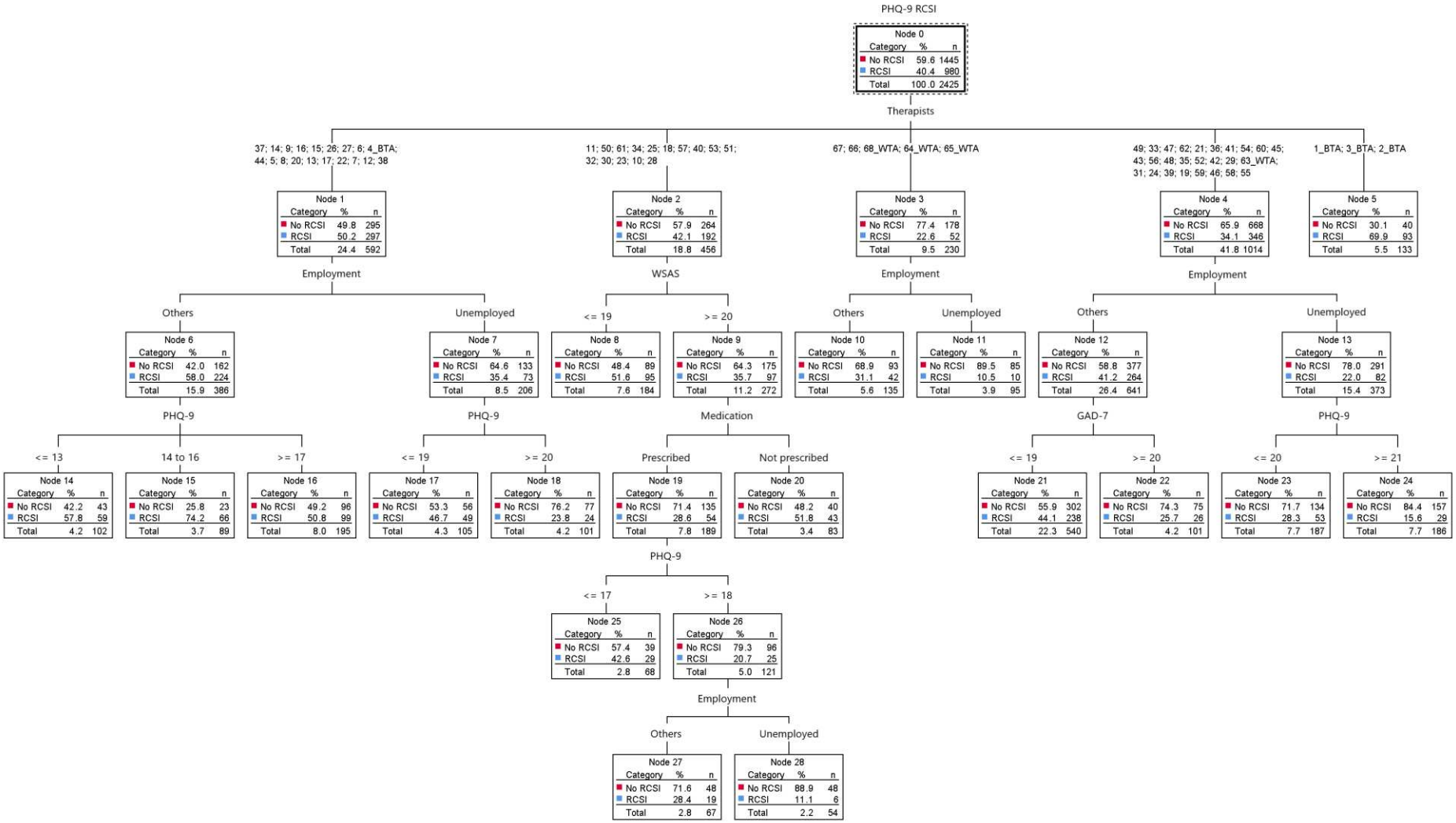
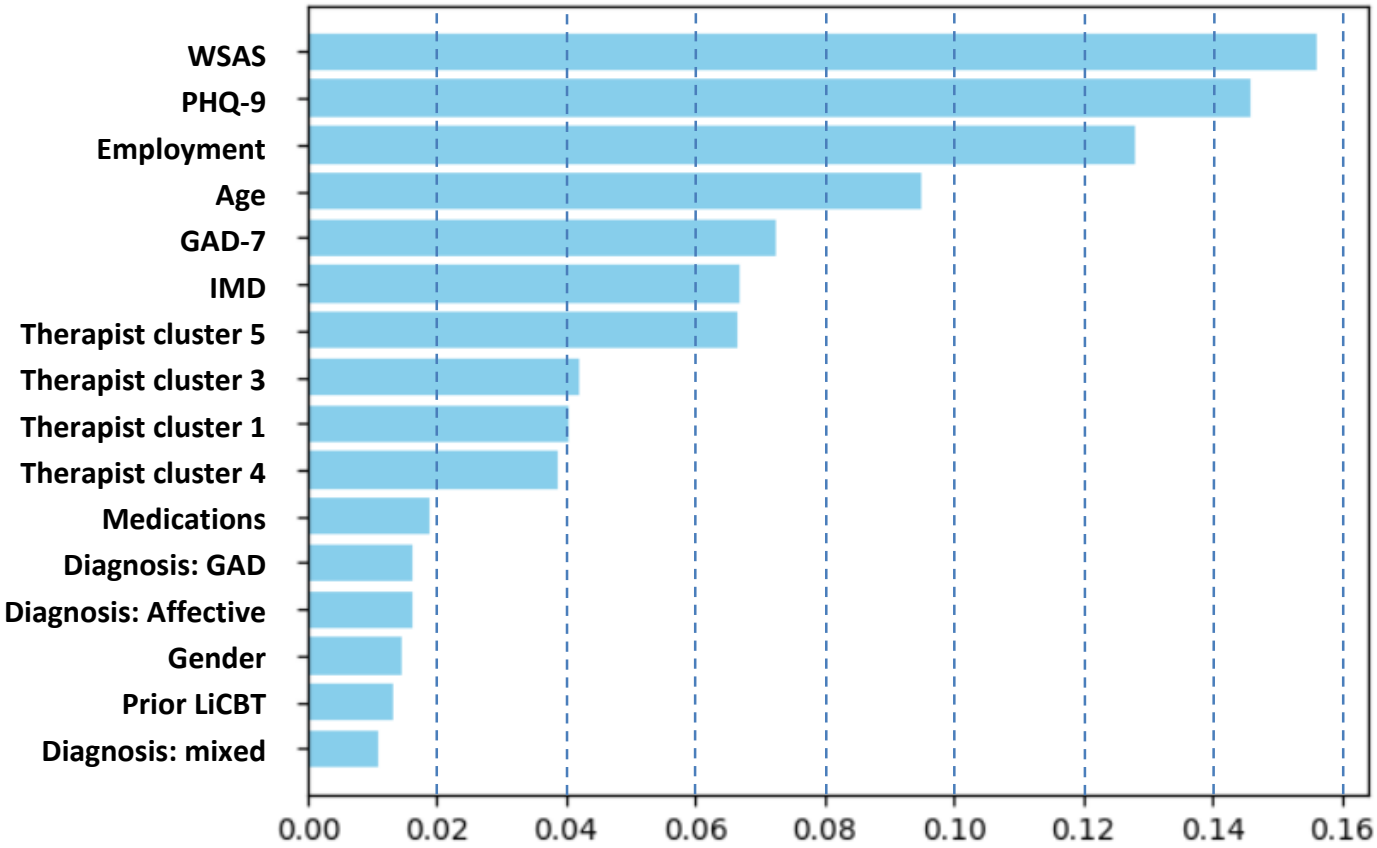


Figure 3. Random Forest: predictor importance plot



Towards Personalized Allocation of Patients to Therapists

Supplemental Material

Glossary of technical features of data analysis

Caterpillar plot. This method graphically displays the results of a multilevel or mixed-effects model. It is conventionally used in studies investigating therapist effects (see Baldwin & Imel, 2013; Johns, Barkham, Kellet, & Saxon, 2019), as it enables the graphical ranking of therapists according to their clinical outcomes. Each therapist is represented with a residual score and corresponding 95% confidence intervals, which represents the discrepancy between their observed outcomes relative to their expected (i.e. predicted) outcomes based on their patients' features. Outlier therapists with 95% confidence intervals that do not overlap with the *line of no difference* (plotted at zero using a horizontal line) are considered to be significantly above or below average in their clinical performance.

Cross-validation. This is a common feature of machine learning analysis, which involves developing a prediction model in a *training* sample and then examining its performance in a *test* sample. The test sample can be a statistically independent dataset or a partition / subset of the training sample which has been held-out (not used to train the prediction model).

Decision tree. These are prediction models that use features in a dataset to predict target outcomes (see Breiman et al., 1984). Models that predict discrete values (i.e. categories or labels) are often called *classification trees*, and models that predict continuous values are also called *regression trees*. These models typically output a tree-like flow diagram, which models interactions between features in the dataset, enabling a fine-grained profiling of subjects according to their characteristics and their predicted outcomes. The structure of the tree is comprised of a *root node* which branches out into *parent* and *child nodes* that represent subgroups of the fuller sample according to key features that are selected into the model. The nodes at the end of each branch are called *terminal nodes*. The total number of terminal nodes in a decision tree represents the number of subgroups or profiles identified in the dataset; each of these nodes has a corresponding prediction. Predictions for individual cases are made by identifying which terminal node they correspond to, based on their features.

Hyper-parameters. In machine learning, model hyper-parameters are specific instructions and functions that specify the learning process of a given algorithm. These have to be set before a prediction algorithm is trained. Relevant hyper-parameters for decision tree models include specifying the maximum *tree depth* (defined below), the minimum acceptable parent and child node sample sizes, the application of Bonferroni correction, the number of trees to model in order to select an optimal solution, the number of observations drawn randomly for each tree, number of variables drawn randomly for each split, the splitting rule to be used, etc.

Intra-cluster correlation coefficient (ICC). In multilevel modeling, the ICC provides a measure of variability in the dependent variable (e.g., patients' treatment outcomes) which can be attributable to differences between higher-level nesting variables (e.g., therapists, services). The ICC is typically reported as a measure of therapist effects, or variability in outcomes due to differential performance across therapists (Baldwin & Imel, 2013).

One-hot encoding. In machine learning, one-hot encoding refers to the transformation of categorical variables with multiple categories into a series of binary variables, where each category is either true (coded '1') or false (coded '0').

Out-of-bag samples. Bagging involves bootstrap resampling of a dataset and then aggregating the models learned on each bootstrap (Breiman, 1996). Out-of-bag samples are a set of bootstrap samples which are not contained in the original dataset, and which are used for cross-validation, typically during the training process to tune (i.e. select and optimize) hyper-parameters.

Random intercept. In multilevel modeling, intercepts can be allowed to vary randomly across instances of a higher-level clustering variable (e.g., therapists) so that the dependent variable for each individual observation can be predicted by the intercept that varies across clusters (Raudenbush, 1993).

Recursive partitioning. This statistical procedure aims to correctly classify samples (e.g., predict their label or outcome) by splitting them into similar subgroups based on a set of features (e.g., independent variables) (see Breiman et al., 1984). Each subgroup may be split or reclassified numerous times until the splitting process terminates after a particular stopping rule is reached. Stopping rules may be triggered because the minimum specified sample size has been reached, the maximum tree depth has been reached, or the subsequent splitting into smaller subgroups no longer adds predictive value to the model.

Risk estimate. Related to decision trees defined above; this statistic describes the risk of error in predicted values for specific nodes of the tree and for the tree as a whole.

Splitting. Related to the recursive partitioning process defined above; splitting refers to the assignment of samples into subgroups that are internally homogeneous (similar to each other in key features) but significantly different to other subgroups.

Tree depth. Related to decision trees defined above; tree depth refers to the maximum number of levels or layers (of parent and child nodes) that make up the structure of a tree.

Voting. In machine learning, *ensemble models* combine the predictions of several algorithms to make a final prediction for a specific case. Random forest is an example of an ensemble model, which combines the predictions of several decision trees. Voting is a method to combine predictions across models, where the final prediction is made based on the majority vote across all decision trees (see Breiman, 2001). *Weighted voting* involves a similar process, where the majority vote across trees informs the final prediction, except that the votes of some trees are given more or less weight based on a statistical estimate of *confidence* in the tree's accuracy.

References:

Baldwin, S. A., & Imel, Z. E. (2013). Therapist effects: Findings and methods. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 258–297). New York, NY: Wiley.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24,* 123–140. https://doi.org/10.1007/BF00058655

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

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984). *Classification and Regression Trees*. Boca Raton: CRC press.

Johns, R. G., Barkham, M., Kellett, S., & Saxon, D. (2019). A systematic review of therapist effects: A critical narrative update and refinement to review. *Clinical Psychology Review, 67*, 78-93. https://doi.org/10.1016/j.cpr.2018.08.004

Raudenbush, S. W. (1993). Hierarchical linear models and experimental design. In L. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 459-496). New York: Marcel Dekker.