




## A demonstration of a multi-method variable selection approach for treatment selection: Recommending cognitive-behavioral versus psychodynamic therapy for mild to moderate adult depression

Zachary D. Cohen, Thomas T. Kim, Henricus L. Van, Jack J. M. Dekker & Ellen Driessen


To cite this article: Zachary D. Cohen, Thomas T. Kim, Henricus L. Van, Jack J. M. Dekker & Ellen Driessen (2020) 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*, 30:2, 137-150, DOI: [10.1080/10503307.2018.1563312](https://doi.org/10.1080/10503307.2018.1563312)

To link to this article: <https://doi.org/10.1080/10503307.2018.1563312>

 [View supplementary material](#) 

 Published online: 11 Jan 2019.

 [Submit your article to this journal](#) 

 Article views: 461

 [View related articles](#) 

 [View Crossmark data](#) 

 Citing articles: 5 [View citing articles](#) 

# A demonstration of a multi-method variable selection approach for treatment selection: Recommending cognitive–behavioral versus psychodynamic therapy for mild to moderate adult depression

ZACHARY D. COHEN<sup>1</sup>, THOMAS T. KIM<sup>1</sup>, HENRICUS L. VAN<sup>2</sup>, JACK J. M. DEKKER<sup>2,3</sup>, & ELLEN DRIESSEN<sup>3</sup>

<sup>1</sup>Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA; <sup>2</sup>Arkin Mental Health Care, Amsterdam, The Netherlands & <sup>3</sup>Department of Clinical, Neuro and Developmental Psychology, Amsterdam Public Health research institute, Vrije Universiteit, Amsterdam, The Netherlands

(Received 19 July 2018; revised 23 November 2018; accepted 19 December 2018)

## ABSTRACT

**Objective:** We use a new variable selection procedure for treatment selection which generates treatment recommendations based on pre-treatment characteristics for adults with mild-to-moderate depression deciding between cognitive behavioral (CBT) versus psychodynamic therapy (PDT). **Method:** Data are drawn from a randomized comparison of CBT versus PDT for depression ( $N = 167$ , 71% female, mean-age = 39.6). The approach combines four different statistical techniques to identify patient characteristics associated consistently with differential treatment response. Variables are combined to generate predictions indicating each individual's optimal-treatment. The average outcomes for patients who received their indicated treatment versus those who did not were compared retrospectively to estimate model utility. **Results:** Of 49 predictors examined, depression severity, anxiety sensitivity, extraversion, and psychological treatment-needs were included in the final model. The average post-treatment Hamilton-Depression-Rating-Scale score was 1.6 points lower (95%CI = [0.5:2.8];  $d = 0.21$ ) for those who received their indicated-treatment compared to non-indicated. Among the 60% of patients with the strongest treatment recommendations, that advantage grew to 2.6 (95%CI = [1.4:3.7];  $d = 0.37$ ). **Conclusions:** Variable selection procedures differ in their characterization of the importance of predictive variables. Attending to consistently-indicated predictors may be sensible when constructing treatment selection models. The small  $N$  and lack of separate validation sample indicate a need for prospective tests before this model is used.

**Keywords:** precision medicine; depression; cognitive behavioral therapy; psychodynamic therapy; treatment selection; variable selection

**Clinical or Methodological Significance of this Article:** Adults seeking treatment for mild to moderate depression have a large variety of psychological and pharmacological treatment options available to them, and clinicians helping clients decide which treatment to pursue could use client-factors associated with differential response to improve their ability to determine which treatment, among the available options, would be most likely to result in a positive response. The process of determining which factors to use, and how to synthesize the available information into a clear, actionable recommendation could be improved through the use of treatment selection approaches based on statistical prediction models. Variable selection, an essential step in the construction of treatment selection models, can be stabilized by attending to those variables that are consistently indicated across several different variable selection approaches.

## Introduction

Major depressive disorder is a highly prevalent, debilitating mental disorder that is currently ranked as the single largest contributor to global disability (World

Health Organization, 2017). Among the most frequently utilized psychotherapies for depression are cognitive behavioral therapy (CBT) and psychodynamic therapy (PDT). Two randomized clinical

---

Correspondence concerning this article should be addressed to Zachary D. Cohen, Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA. Email: [zachary.d.cohen@gmail.com](mailto:zachary.d.cohen@gmail.com)

trials have found PDT noninferior to CBT in the outpatient treatment of depression (Driessen et al., 2013; Gibbons et al., 2016). These results are in line with meta-analytic findings reporting no significant differences between CBT and PDT for depression (Barth et al., 2013; Driessen et al., 2015). These minimal efficacy differences, along with differential therapeutic theories used in CBT and PDT (Hoffart & Johnson, 2017), raise the question whether individual patients can be identified that might benefit more from one of these treatments than the other. If so, treatment selection could improve outcomes in depression by helping individuals select the specific intervention that is most likely to be successful (Cohen & DeRubeis, 2018).

DeRubeis et al. (2014) developed a treatment selection approach that can be used to identify each individual's optimal treatment based on multiple patient characteristics, using data from a randomized clinical trial. This approach is called the Personalized Advantage Index (PAI). The core concept behind the PAI approach is to identify pre-treatment patient characteristics that are associated with differential response to treatment (so-called moderators) and, using these variables, to build a statistical model that can generate predictions for an individual in two (or more) treatments. This approach shares many common features with other approaches to treatment selection (Cohen & DeRubeis, 2018). For each individual, the treatment with the best predicted outcome is defined as the indicated treatment. In the case of a two-treatment comparison, an individual's PAI is a single number derived through the subtraction of their predictions in one treatment from the other. The PAI provides a directional indication of which treatment the individual should receive, as well as information about the strength of the recommendation, represented by the size (in absolute value) of the PAI.

In their initial demonstration, DeRubeis et al. (2014) analyzed data from a randomized comparison of antidepressant medication and CBT and found that, for patients with large predicted advantages in one treatment over the other (60% of sample), those who received their PAI-indicated treatment had superior outcomes relative to patients who received the non-indicated treatment, with an effect size (Cohen's  $d = 0.58$ ) larger than that reported in a recent systematic review of drug-placebo differences (Turner, Matthews, Linardatos, Tell, & Rosenthal, 2008). Huibers et al. (2015) published similar findings applying the PAI approach to a comparison of cognitive therapy versus interpersonal therapy for adult outpatient depression. Related analyses based on the PAI approach have generated models aimed at differentiating placebo and

antidepressants responders (Webb et al., 2018), minimizing risk of dropout (Zilcha-Mano et al., 2016) and relapse (Schweizer et al., submitted; Vittengl, Clark, Thase, & Jarrett, 2017). The principles on which the PAI is based have also been applied to treatment selection in post-traumatic stress disorder (Deisenhofer et al., 2018; Keefe et al., 2018). These studies represent but one strand of research on treatment selection. Other approaches include the  $M^*$  approach (Niles, Loerinc, et al., 2017; Niles, Wolitzky-Taylor, Arch, & Craske, 2017; Smagula et al., 2016; Wallace, Frank, & Kraemer, 2013) initially introduced by Kramer (Kraemer, 2013), the use of nearest neighbor modeling by Lutz and colleagues (Lutz et al., 2006), and a series of papers by Uher, Iniesta and colleagues (Iniesta, Hodgson et al., 2018; Iniesta, Malki et al., 2016; Uher et al., 2012). For a comprehensive review of this literature, we suggest recent reviews by Cohen and DeRubeis (2018), Gillan and Whelan (2017), and Kessler (2018).

Different statistical methods can be applied to select patient characteristics to be used in generating treatment recommendations. For instance, the initial applications (DeRubeis et al., 2014; Huibers et al., 2015) relied on a domain-based backwards stepwise-regression (Fournier et al., 2009) for variable selection, in which potential predictors were first grouped by domains (e.g., history of illness, demographic and life circumstances, cognitive dysfunction etc.) in each of which backwards step-wise variable selection was performed, and then the retained predictors from each domain were combined and a final backwards step-wise variable selection was performed to generate the final model. A recent PAI-based treatment selection paper by Vittengl et al. (2017) used a series of single-variable models to establish the statistical significance of independent moderators, and then used backwards and forwards step-wise variable selection procedures to reduce the set. Another recent PAI analysis relied upon a machine-learning approach called Random Forest (RF) for variable selection (Zilcha-Mano et al., 2016). Building on this work, Keefe et al. (2018) used a two-stage variable selection approach in which they used RF followed by a stepwise AIC-penalized bootstrapped method. This variability is discussed in Cohen and DeRubeis' (2018) review of treatment prediction reports in depression. They noted that there is very little consistency in the variable selection approaches that have been employed in this area.

This heterogeneity is problematic because different variable selection approaches applied to the same dataset can lead to different conclusions about variable importance (Bleich, Kapelner, George, &

Jensen, 2014), and treatment recommendations can vary based on which variables are included in the model. In their review, Cohen and DeRubeis (2018) identified 43 reviews (and meta-analyses) of predictors of treatment response in depression but, as many of these reviews noted, no coherent picture of which predictors are most important has emerged. The variability and lack of replicability in efforts to identify predictors of response in depression may be partially explained by the heterogeneity of the statistical approaches that are used to identify predictive variables.

This methodological heterogeneity also makes it difficult to determine *which* variable selection procedure should be used in the context of treatment selection. The strengths and weaknesses of different approaches might make them more or less attractive for a given purpose. For example, variable selection with elastic net regularization (ENR) can handle high numbers of potential predictors and can overcome issues of high correlations between baseline variables (Friedman, Hastie, & Tibshirani, 2010; Zou & Hastie, 2005). However, it does not have the capability to account for unspecified non-linear relationships in the way that is possible when using RF (Garge, Bobashev, & Eggleston, 2013). It is unlikely that any single variable selection procedure will be optimal for all situations. It will also be difficult to identify *which* single approach one should use without a sufficiently large dataset to allow for a training sample (in which one could try every approach and see which one appears to work best) and a held-out test sample (in which to show that the results hold, and are not due to chance findings). Unfortunately, RCT samples with relevant treatment comparisons are rarely large enough to support these efforts (Kessler, 2018), and the use of large non-randomized datasets risks potential confounds (e.g., selection effects) that could bias treatment selection efforts (Cohen & DeRubeis, 2018; c.f. Kessler, 2018).

Potentially useful predictors could be identified by observing which variables are consistently selected across multiple studies. However, due to the variability of statistical approaches employed across different papers (Cohen & DeRubeis, 2018) and the lack of relevant studies in which the collection of potential moderators in this study have been assessed and compared, it would be unwise to rely solely upon the consistency of findings from the existing literature. Another type of consistency that could inform variable selection is the consistency with which variables are selected across multiple methods within the same sample. Constructing models using variables that are selected consistently across multiple different feature selection

approaches within the same data can result in improved model performance (Kuhn & Johnson, 2013). Knowledge of the different methodologies could be used to understand whether those variables that are identified in some approaches but not in others are inconsistently identified due to weak or noisy effects, and thus should be considered poor predictors, or whether this pattern can be attributed to shortcomings of specific approaches (Kuhn & Johnson, 2013). For example, a variable might be selected by an ensemble-of-trees approach (e.g., RF) but excluded by another approach based on classic regression (e.g., ENR) because it involves a three-way interaction or non-linear interaction that was not considered in the latter classic approach (Kuhn & Johnson, 2013).

Using these principles, we aimed to demonstrate an improved PAI approach by generating individual treatment recommendations for adult outpatients with depression deciding between CBT versus PDT. We introduce a novel selection process that synthesizes the results of four different variable selection techniques (RF, ENR, Bayesian Additive Regression Trees and the AIC-penalized bootstrapped approach) by selecting the patient characteristics that are consistently identified as associated with differential response.

## Method

### Design and Participants

This paper draws on data from a randomized clinical trial comparing CBT and PDT in the outpatient treatment of depression (Driessen et al., 2013), which included 341 patients who met DSM-IV criteria for a major depressive episode and scored 14 or higher on the Hamilton Rating Scale for Depression (HAM-D; Hamilton, 1960). The Dutch Union of Medical-Ethical Trial Committees for mental health organizations approved the study design and the study protocol was published (Driessen et al., 2007). Efficacy results of this study are reported elsewhere, with no significant treatment differences found on any of the outcome measures (Driessen et al., 2013, 2015, 2017). Two prior papers examined which subgroups of patients in this trial might benefit more from one of the treatments than the other. One (Kikkert et al., 2016) was a replication study examining obsessive-compulsive and avoidant personality disorder traits as potential moderators of treatment efficacy that failed to replicate previous findings in that regard (Barber & Muenz, 1996), while the other applied model-based recursive partitioning to 23 potential moderators to identify subgroups of patients that might benefit

specifically from one of the two treatments (Driessen et al., 2016). However, these studies used different patient subsamples, examined only a subset of the potential predictors, and were not designed to produce a model that could be used to generate treatment recommendations for individual patients.

As part of the trial protocol, severely depressed (HAM-D >24) patients at baseline were offered adjunctive antidepressant medication ( $N = 129$ ). As the treatment effects observed in these individuals could have been a result of the psychotherapy, the medication, or both, they were excluded from the current analyses. Thus, this report relates to the patients with moderately severe depressive symptoms (baseline HAM-D = 14 to 24) who were treated with psychotherapy only ( $N = 212$ ). Of these 212 individuals, 17 were removed for having too much missing baseline data ( $\geq 50\%$  missing baseline predictors). Finally, an additional 28 individuals who dropped out before attending at least 4 sessions were excluded. As our goal was to build a model to answer how individuals who *received* a meaningful course of CBT or PDT fared, we felt that individuals who dropped out very early in treatment would not be informative for our models. In the extreme, the “outcome” for patients who dropped out prior to attending a single session does not reflect response to CBT or PDT. Additionally, we were less confident in our ability to impute valid week-16 outcomes for these early dropouts. We decided to remove from our analyses patients who attended 3 or fewer therapy sessions. This reduced our sample from 195 to 167. Thus, the final sample comprised 167 patients: 75 in the CBT and 92 in the PDT condition (see Supplemental Figure S1 for a Patient Flow Chart). Baseline sample demographic characteristics for the final sample are presented in Supplemental Table S1.

### Interventions

Both PDT and CBT encompassed 16 individual 45-minute sessions within 22 weeks and were conducted according to a published treatment manual (de Jonghe, 2005; Molenaar, Don, van den Bout, Sterk, & Dekker, 2009). CBT was based on the principles described by Beck, Rush, Shaw, and Emery (1979) and included behavioral activation and cognitive restructuring according to a session-by-session protocol with homework assignments. Short psychodynamic supportive psychotherapy (de Jonghe et al., 2013) represented the psychodynamic intervention. This modality involved an open patient-therapist dialogue that used supportive and insight-facilitating techniques to address the emotional background of

the depressive symptoms by discussing current relationships, internalized past relationships, and interpersonal patterns.

### Measures

HAM-D scores were used as the outcome measure for this study. Trained research assistants (master-level graduate students in clinical psychology) assessed the HAM-D according to the Dutch scoring manual (de Jonghe, 1994) at baseline, week-5, week-10, and week-22 (post-treatment). Assessors were not blind to treatment condition. Assessors engaged in one-hour peer supervision sessions bi-weekly, in which audiotaped interviews were discussed. The average intraclass correlation coefficient over 46 audiotaped assessments scored by multiple assessors was .97. Supplemental Table S2 lists the 49 patient characteristics considered during variable selection, all of which were assessed at pre-treatment.

### Building the Personalized Advantage Index Model

All analyses were performed in R (Team, 2000). Pre-processing and random forest-based imputation of missing data (Stekhoven & Buhlmann, 2012) was performed on baseline and outcome data prior to variable selection (details provided in Supplemental). Post-treatment HAM-D scores were missing for 20 of the 167 participants (12.0%). Future analyses in larger samples should perform these steps (especially imputation) separately for the training and test samples to avoid this form of double-dipping. Categorical variables (e.g., relationship status) were turned into binary variables and binary variables without sufficient variability (variables whose smallest category made up < 20% of the sample) were excluded (Kuhn & Johnson, 2013). Outliers for continuous variables were winsorized, and some variables with skewed distributions were log-transformed (see Supplemental Table S2 for more details).

### Variable Selection

To select the patient characteristics associated with treatment outcome, we applied a multi-phase selection procedure that combines four different variable selection methods, each of which has been used in recently published treatment selection analyses (Bleich et al., 2014; Iniesta et al., 2016; Keefe et al., 2018; Zilcha-Mano et al., 2016). The first step was to apply three approaches to identify predictors of (differential) treatment response: 1. mobForest

(Garge et al., 2013; random forest type approach for model-based recursive partitioning; referred to simply as random forest or RF), 2. glmnet (Friedman et al., 2010; Elastic Net Regularized Regression or ENR), and 3. bartMachine (Kapelner & Bleich, 2016; Bayesian Additive Regression Trees or BART). The second step was to reduce the variables identified by at least two of these three approaches using bootStepAIC (Rizopoulos, 2009; bootstrapped backwards stepwise AIC-penalized model selection; Austin & Tu, 2004). We will now describe each of these methods in more detail, and discuss their relative strengths and limitations. Supplemental Table S3 provides additional details of each method and contrasts different features.

Random Forest is a recursive partitioning approach that can accommodate large numbers of predictor variables as well as complex relationships including non-linear and higher order interactions (Kapelner & Bleich, 2016). RF builds upon recursive partitioning approaches like classification and regression trees and model-based recursive partitioning. It addresses model instability by randomly selecting features and creating many “tree models”, the predictions of which are aggregated to generate stable predictions (Austin & Tu, 2004). RF also allows for information from weaker predictors to be incorporated in situations where they might otherwise be dominated by stronger predictors, such as in bagging (Garge et al., 2013). The model function of RF can be specified as “ $y \sim tx$ ,” which forces the approach to select splits that maximize the difference in the treatment condition coefficient between subgroups, thus focusing on identifying moderators of the treatment effect. When RF is used for variable selection, the permuted variable importance is generated by comparing the mean square error (MSE) of the predictions in the held-out (out of bag) samples when the real values are used to the MSE when permuted values for a given predictor are used. The extent to which the MSE increases when permuted values are used indicated how “important” that variable is (Garge et al., 2013). Variables that surpass the recommended threshold (which is set based on the largest observed “noise” variable, for which the permuted data improves the MSE, relative to the real data) are selected.

Elastic Net Regularization can provide a hybrid of the Lasso and Ridge regression approaches, combining the L1 and L2 penalizations to allow for the selection of a parsimonious set of variables that predict outcome (Hastie, Tibshirani, & Friedman, 2009). We used the R package glmnet (Friedman et al., 2010) to implement ENR variable selection, and used Zou and Hastie’s (2005) recommended default value for the alpha parameter ( $\alpha = 0.5$ ).

Uses of ENR in the literature of variable selection and treatment selection have only investigated prognostic models in which a single treatment is modeled (Chekroud, Gueorguieva et al., 2017; Chekroud, Zotti et al., 2016; Iniesta et al., 2016). Current implementations of ENR in R do not accommodate variable selection for models in which moderators are of primary interest. In order to adapt ENR for the purpose of identifying moderators, we split the sample into each of the two treatment groups, and then constructed prognostic models within each group. Variables that were retained as predictors in only one condition, or that were selected in both but specified with differing coefficient values, were identified as potential moderators of treatment effects. We refer the reader to Cohen and DeRubeis’ (2018) review (specifically, their Figure 1) for a more in-depth discussion of why variables with these relationships are candidate moderators. As one example, consider a variable that is selected by ENR in one treatment condition and specified with a positive coefficient, and that is selected in the other treatment condition but specified with a negative coefficient. This information could suggest that a disordinal relationship exists between that variable and treatment, such that individuals with higher levels on that variable do worse relative to individuals with lower levels in one treatment, whereas individuals with lower levels on that variable do worse relative to individuals with higher levels in the other treatment.

Bayesian Additive Regression Trees builds on ensemble-of-tree methods such as RF by incorporating an underlying Bayesian probability model (Chipman, George, & McCulloch, 2010). BART and RF have similar strengths insofar as they both can handle large numbers of predictors, and can accommodate non-linear and higher order interactions. The inclusion of the Bayesian prior improves upon other tree-ensemble approaches by introducing regularization, which reduces the likelihood that the ensemble will become dominated by any single tree (Genuer, Poggi, & Tuleau-Malot, 2010). Kapelner and Bleich adapted the bart Machine R-package (2016) to help focus model building on moderators. To achieve this aim, they introduced a parameter that forces the search for variable splits to focus more on treatment than other variables, thus introducing more interactions between treatment and other variables. This is conceptually similar to when researchers only consider interactions between treatment and baseline variables (and not interactions between baseline variables themselves), or to how RF can specify the splitting criteria to evaluate the difference in the treatment coefficient for the model  $y \sim tx$ . Bleich et al. (2014) adapted BART to

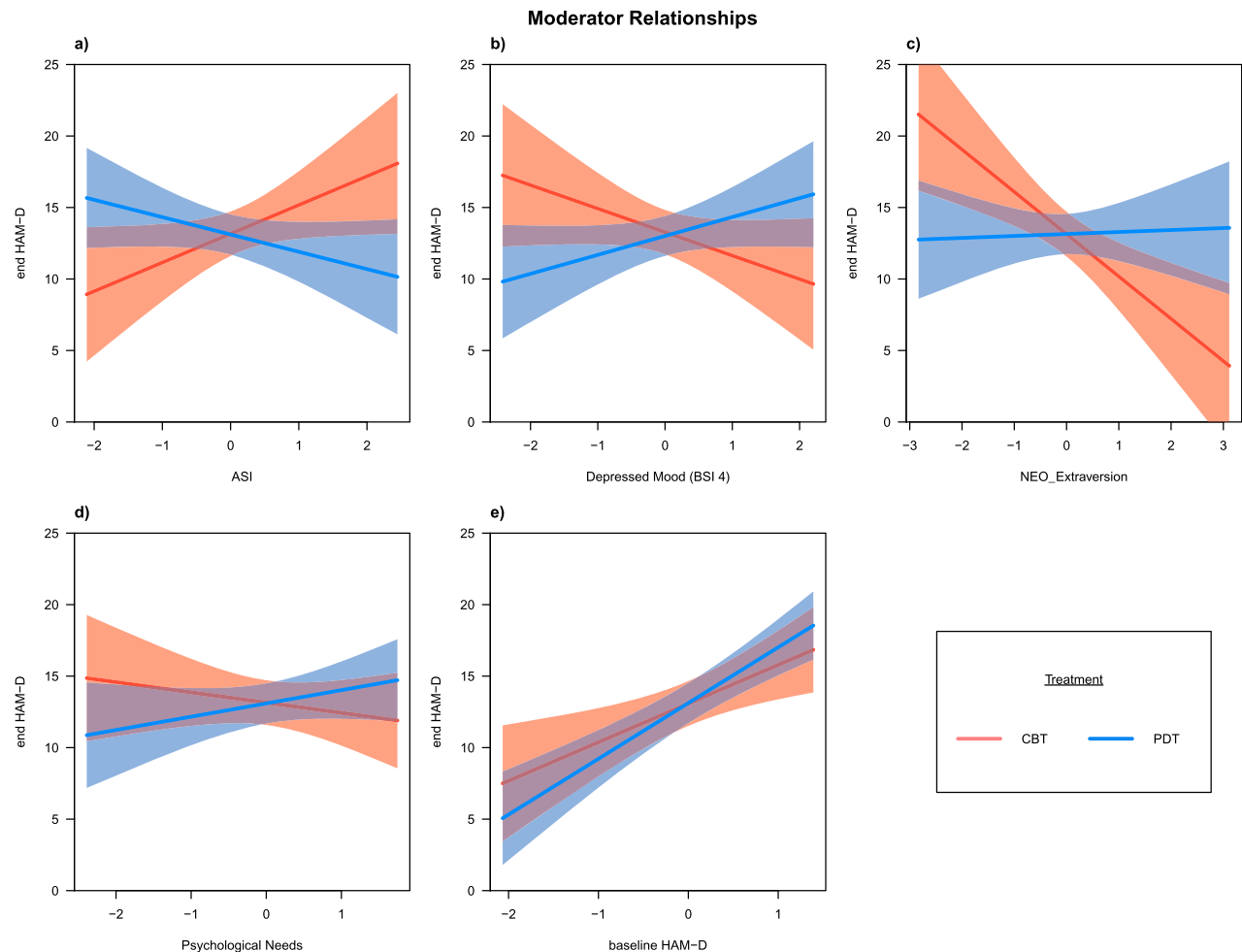


Figure 1. Visualization of the moderator relationships. Conditional plots with confidence bands for the conditional mean generated using R package visreg from the final model estimated in the complete sample. Conditioning for each plotted variable uses the mean value for all other variables. The X-axes represent the standardized/centered scores that were used during analysis.

extract informed prior information about variable importance, and provided an interaction plot feature that can be used to identify potential 3-way interactions. The ICEbox package in R (Goldstein, Kapelner, Bleich, & Pitkin, 2015) allows for the visualization of predictive relationships in BART models, including non-linear and higher-order interactions between variables and treatment. The  $N$  most important interactions identified by BART are retained, where  $N$  was decided based on the number of variables selected by Random Forest (which uses a permutation test to determine an importance threshold cutoff).<sup>1</sup> See the Supplemental Variable Selection, as well as Garge et al. (2013) for more details on how the threshold is determined.

We decided to reduce the variables consistently selected by the above three approaches using a specific fourth approach for the following reason: If the model that was used to generate predictions relied on linear or logistic regression, then the variables selected by

RF or BART could lead to a model with poor fit, if, for example, these variables relied on non-linear relationships or higher order interactions. The BootStepAIC package (Rizopoulos, 2009) performs variable selection using a stepwise AIC-penalized bootstrapped approach (Austin & Tu, 2004). By only including the moderator relationships identified in the other three approaches (and their corresponding main effects), this search generated a model emphasizing the prediction of differential treatment response, while reducing the chance that predictors that require unspecified linear or higher-order interactions were included. 10,000 bootstrapped training samples were drawn, and within each training sample backwards elimination was used to select variables that independently contributed to predicting outcome. Austin and Tu (2004) recommend selecting variables that are retained in at least 60% of bootstrapped samples, but this recommendation is specific to prognostic variables (main effects only). As we were

interested in interactions, we relied on the consistency of the direction of the coefficients across the 10,000 bootstrapped samples. By using a threshold of 95% consistency in sign of the moderator coefficient, variables with smaller effects that were consistent in the direction with which they predict differential response across treatments could be included. The primary goal of this step was to ensure that the variables selected will function properly and consistently, and increase the likelihood that the final model will replicate in future samples drawn from the same population.

### Generating PAIs

Based on the set of variables selected, outcome predictions were generated for each study participant in both of the treatments. To avoid the risk of overconfidence that could occur when evaluating model-performance on individuals whose data were used to set model-weights, these predictions were generated using ten-fold cross validation (CV). 10-fold CV is recommended based on its good bias and variance properties in small samples (Kuhn & Johnson, 2013). For each of the 10 folds, individuals in that fold were held out, and data from the patients in the other 9 folds were used to generate a linear regression model in which end-of-treatment HAM-D score was predicted by the set of selected predictor variables (main effects for each variable and terms representing their interactions with treatment). The data from the patients in the held-out fold were then used to generate predictions for those patients in each treatment. For each individual, the difference in the predicted HAM-D score in CBT and PDT is their PAI. Individuals with a lower predicted HAM-D score in CBT (and thus a better predicted outcome in CBT) were then classified as “CBT-indicated” and individuals who had a better predicted outcome in PDT were labeled “PDT-indicated”. The size of the PAI is taken to be an indication of the strength of the treatment recommendation (DeRubeis et al., 2014).

### Evaluating PAIs

To characterize the expected utility of the PAIs for guiding treatment selection, we compared the average end-of-treatment HAM-D scores of individuals who got their indicated treatment (based on their PAI scores) against that of participants who received their non-indicated treatment. Next, we looked within the subgroup indicated to need CBT, and compared HAM-D scores for those who received their indicated treatment (CBT) to those who received their non-indicated treatment (PDT). We then

performed the analogous comparison for those identified as “PDT-indicated.” In order to investigate the importance of the strength of these recommendations following earlier PAI analyses (DeRubeis et al., 2014; Huibers et al., 2015), we then evaluated the above comparisons within a subset of the sample using an a priori determined cutoff: the strongest 60% of PAIs (the 60% of the largest absolute value PAIs). The entire 10-fold cross-validation procedure and evaluation was repeated 1000 times to account for the influence of the selection of the 10 folds on the results (Kuhn & Johnson, 2013). Accuracy statistics for the predictions were assessed within each of the 1000 10-fold CVs. Findings, representing averaged results across all 1000 runs, are presented below.

## Results

### Variable Selection

Table I summarizes the results of our new variable selection approach at each stage (See Supplemental for a more detailed discussion of the results from each approach).

The final model including the variables selected by BootStepAIC was:

$$Y = tx*(HAM - D_{Baseline} + ASI + BSI\ 4\ Depressed\ Mood + NEO\ Extraversion + Psychological\ Needs)$$

Thus, five variables were selected as predictors of differential treatment response: HAM-D score, Brief Symptom Inventory (BSI; De Beurs & Zitman, 2005) Depressed Mood subscale, Anxiety Sensitivity Index total score (ASI; Reiss, Peterson, Gursky, & McNally, 1986), NEO Five Factor Inventory (NEO-FFI; Hoekstra, Ormel, & De Fruyt, 2003) Extraversion subscale, and Patient Request Form Psychological Needs subscale (Veeninga & Hafkenscheid, 2004). We note that although both the depressed mood subscale of the BSI and the HAM-D measure the construct of depression, the correlation between the two scales was low ( $r = .23$ ). One explanation for this might be that the HAM-D measures a more broad set of symptoms (e.g., sleep, libido, appetite, psychomotor retardation, etc.) than the 6 items that are captured by the BSI's depressed mood subscale: suicidal thoughts, loneliness, sad/depressed mood, lack of interest, hopelessness, and worthlessness. There were no significant differences between the treatment groups for any of the variables selected for the final model (see Supplemental Table S4). Table II presents the final model with weights set using the full sample.



Table I. Summary of variable selection results.

Variable	<— Step 1 —>			Step 2 Included in BootStep AIC	Result Selected by BootStep AIC
	Random Forest	Elastic Net	BART		
<b>Baseline HAM-D</b>	Yes	Yes	Yes	Yes	<b>Yes</b>
Age	Yes	No	No	No	N/A
<b>Anxiety Sensitivity (ASI)</b>	No	Yes	Yes	Yes	<b>Yes</b>
BAI	Yes	No	No	No	N/A
(BSI 2) Cognitive Problems	Yes	Yes	Yes	Yes	No
(BSI 3) Interpersonal Sensitivities	Yes	Yes	No	Yes	No
<b>(BSI 4) Depressed Mood</b>	Yes	No	Yes	Yes	<b>Yes</b>
(BSI 5) Fear	Yes	Yes	Yes	Yes	No
(BSI 7) Phobic Fears	Yes	Yes	No	Yes	No
(BSI 8) Paranoid Thoughts	Yes	Yes	Yes	Yes	<b>Yes*</b>
Contacted Physician	No	Yes	No	No	N/A
Dysthymia	No	Yes	No	No	N/A
Employed	Yes	Yes	Yes	Yes	No
Episode Duration	No	Yes	No	No	N/A
Inventory of Depressive Symptomatology (IDS)	Yes	Yes	Yes	Yes	No
LEIDS Acceptance	No	Yes	No	No	N/A
LEIDS Hopelessness	No	No	Yes	No	N/A
Mobility	No	Yes	No	No	N/A
<b>NEO Extraversion</b>	No	Yes	Yes	Yes	<b>Yes</b>
NEO Neuroticism	Yes	Yes	Yes	Yes	No
NVM Extraversion	Yes	Yes	Yes	Yes	No
NVM Somatization	Yes	Yes	Yes	Yes	No
Pain (VAS)	Yes	Yes	Yes	Yes	No
<b>Psychological Needs (PRF)</b>	Yes	No	Yes	Yes	<b>Yes</b>
NEO Neuroticism × Married	N/A	N/A	Yes	Yes	No

Table I. Summary of variable selection results for all variables selected by at least one approach. Three different variable selection approaches based on Random Forest, Elastic Net Regularization, and Bayesian Additive Regression Trees (BART) were applied to the full set of 49 potential baseline predictors. The 16 potential moderators that were selected by at least two of these three approaches were then submitted, along with one three-way interaction identified by BART (NEO Neuroticism × Married × Treatment), to a final variable selection stage with BootStepAIC. Bold text indicates the variables selected by BootStepAIC based on a criteria of at least 95% consistency of the coefficient sign for the interaction with treatment across 10,000 bootstrapped samples. \*although BSI 8 was selected by BootStepAIC, its *p*-value in the final model built in the full sample was .43, and so, following the recommendation of Kuhn and Johnson (2013) to favor simpler models, it was not included in the final model.

Baseline HAM-D score was included as both a main effect and as an interaction with treatment, but it did not appear to have a significant moderator relationship

Table II. Final regression model specified using the full sample.

Variable	<i>B</i>	SE	<i>p</i> value
(Intercept)	13.14	0.53	0.00**
Treatment	-0.50	1.05	0.63
ASI	0.40	0.64	0.53
Depressed Mood (BSI 4)	-0.16	0.63	0.80
HAM-D Baseline	3.30	0.58	0.00**
NEO Extraversion	-1.41	0.58	0.02*
Psychological Needs	0.11	0.56	0.85
Treatment × ASI	-3.22	1.28	0.01*
Treatment × Depressed Mood (BSI 4)	2.97	1.27	0.02*
Treatment × HAM-D Baseline	1.19	1.16	0.31
Treatment × NEO Extraversion	3.10	1.15	0.01**
Treatment × Psychological Needs	1.65	1.46	0.15

\**p* < .05.\*\**p* < .01.

in the context of the final model. The moderator relationships included in the final model are visualized in Figure 1. We refer the reader to a recent review on treatment selection by Cohen and DeRubeis (2018) for a detailed discussion of how to approach interpretation of moderator relationships in treatment selection.

### PAI Results

The mean absolute error of the predictions was 5.5 (SD = 7.1), and the associated R-squared was 0.18. Individuals who received their model-indicated treatment had better outcomes than those who received their non-indicated treatment (see Figure 2). The mean end of treatment HAM-D scores for individuals who received their PAI-indicated treatment (all results averaged across the 1,000 CVs), was 12.3 (SD = 7.6, *N* = 82.8); the average for those receiving their non-indicated treatment was 13.9 (SD = 7.9,

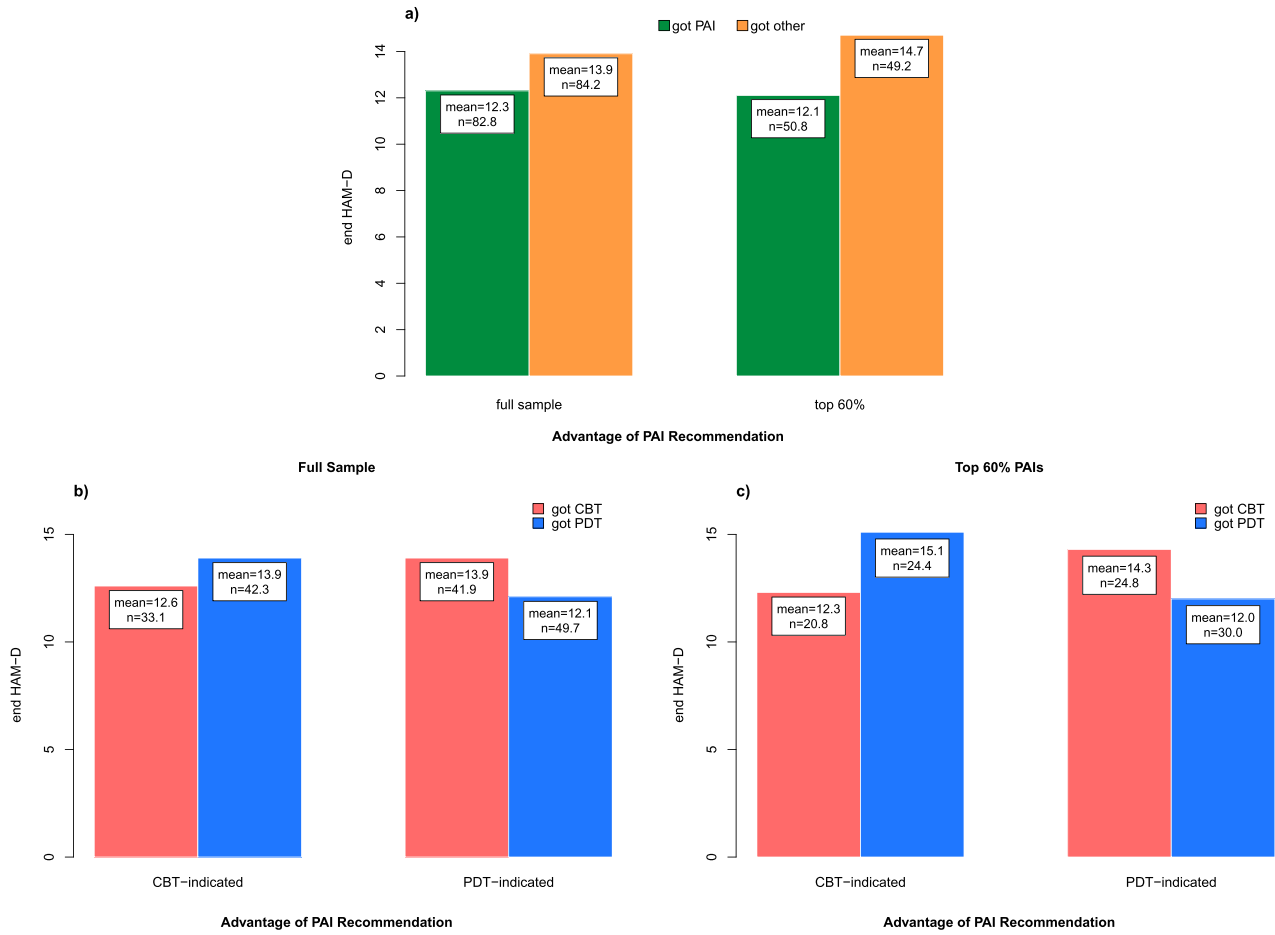


Figure 2. Comparison of end-of-treatment HAM-D scores for patients randomized to their PAI-indicated treatment with those who were randomized to their non-indicated treatment. (a) Shows this comparison with treatment conditions collapsed for the full sample, and for the 60% of patients with larger PAIs. (b) Decomposes the comparison by treatment for the full sample, with those indicated to need CBT represented by the left two bars, and those indicated to need PDT by the right two bars. (c) Presents the same breakdown as in Figure 2 (b), but for the 60% of patients with larger PAIs. Abbreviations used in manuscript: AIC = Akaike Information Criterion, ASI = Anxiety Sensitivity Index, BART = Bayesian Additive Regression Trees, BSI = Brief Symptom Inventory, CBT = Cognitive Behavioral Therapy, DSM-IV = Diagnostic and Statistical Manual of Mental Disorders 4th edition, ENR = Elastic Net Regularization, HAM-D = Hamilton Rating Scale for Depression, IDS = Inventory of Depressive Symptomatology, LEIDS = Leiden Index of Depression Sensitivity, MSE = Mean Squared Error, NEO = Neuroticism-Extraversion-Openness Personality Inventory, NVM = Shortened Dutch Adaptation of the Minnesota Multiphasic Personality Inventory, PAI = Personalized Advantage Index, PDT = Psychodynamic Therapy, PRF = Patient Request Form, RF = Random Forests, VAS = Visual Analog Scale for Pain.

$N = 84.2$ ). This reflected, on average, a 1.6 points advantage for those receiving their indicated treatment (95% CI = 0.5 to 2.8; Cohen’s  $d = 0.21$ , 95% CI = 0.07 to 0.37). When we restricted our evaluation to the largest (absolute value) 60% of PAIs (indicated mean = 12.1, SD = 6.4,  $N = 50.8$ ; non-indicated mean = 14.7, SD = 7.6,  $N = 49.2$ ), we found that the effect of treatment selection grew to 2.6 points (95% CI = 1.4 to 3.7; average Cohen’s  $d = 0.37$ , 95% CI = 0.19 to 0.54).

### Discussion

Helping service-users and clinicians make better-informed treatment decisions is one of the core

goals of precision medicine in mental health. In depression, treatment selection models can improve the ability to identify the best treatment among available options (Cohen & DeRubeis, 2018). Here, we have described a treatment selection model based on patient characteristics that could be used to decide between cognitive-behavioral and psychodynamic therapy for those with mild-to-moderate depression not taking antidepressants.

The differential prediction described here relied on four factors: anxiety sensitivity, depression symptom severity, extraversion, and psychological treatment needs. Although in this investigation the aim was to develop and provide a first test of a multivariable model that could inform treatment selection, it

nonetheless is important to attempt an understanding of the basis for each variable's contribution to the model. In the following we provide tentative, speculative interpretations of the findings, taking into account the directions of the relationships observed.

The ASI reflects a person's beliefs that anxiety experiences have negative somatic, psychological or social consequences. Higher scores on this measure were associated with superior outcomes in PDT relative to CBT, and vice versa for lower scores. Patients with higher baseline depressed mood, as measured by the BSI, tended to improve more in CBT, whereas those with lower BSI scores tended to respond better to PDT. Those with higher scores on a measure of extraversion (on the NEO) fared better in CBT than in PDT. The reverse was true for those low on extraversion. The Patient Request Form Psychological Needs scale assesses a patient's needs for a psychological treatment. Higher scores on this measure predicted better response to PDT, relative to CBT, and the reverse prediction was obtained for those with lower scores on this measure. Thus, in contrast to anxious, introverted patients, patients who were relatively more extraverted and who had low psychological treatments needs were better matched to CBT. We would speculate that these patients typically express themselves more and have already talked about their feelings and problems with others without much hesitation. They might be more in need of the structured approach of CBT, directed strongly at adapting behavior and changing cognitions through practical exercises. It may be that PDT is more efficacious for patients who search for a psychological solution to their depressive symptoms. Anxious and introverted patients, who have a tendency to avoid focusing on their problems despite a need to do so, may find that the supportive milieu of PDT fostered explorations of their feelings and problems in a way that was appropriate for their individual needs and capacities.

### Limitations

This study has a number of limitations. Treatment selection is likely to be most effective when the interventions under consideration differ substantively and substantially in their mechanisms and targets (Cohen & DeRubeis, 2018). Relative to comparisons between, for example, medications and psychotherapy, the similarity of these two psychotherapies likely resulted in a decreased potential to identify individuals who are strongly indicated to need one treatment over another. Nevertheless, among the 60% of patients with the strongest PAIs, a *d*-type

effect size of 0.37 was observed for receiving the indicated versus contraindicated treatment.

Patients seeking treatment for depression have many options, including other psychotherapies and medication; this model cannot inform the decision of whether or not to pursue treatments other than CBT and PDT. This model would at best be valid for use in similar populations. It cannot be known how the model would perform if it were applied to patients whose values on predictors were outside the observed range on the predictor measures, or if used in the context of a population of those with severe depression (defined in this trial as HAM-D >24) or in patients who are also taking antidepressant medications.

We know that therapist effects can account for significant variability in patient outcomes in psychotherapy (Lutz, Leon, Martinovich, Lyons, & Stiles, 2007). These analyses ignored potential therapist effects, but future work in larger samples could attempt to incorporate variables that predict which patients should be seen by which therapists. The generalizability of nomothetic data (upon which these analyses relied) to individuals has recently been called into question (Fisher, Medaglia, & Jeronimus, 2018), but recent work has demonstrated how ideographic approaches could be used to guide precision selection and ordering of therapeutic techniques or modules (Rubel, Fisher, Husen, & Lutz, 2018).

Despite our use of cross-validation during the weight-setting stage, our use of the full sample during variable selection and imputation could lead to model overfitting and inflated relationships (Fiedler, 2011), and as noted by Hastie et al. (2009), represents a form of double-dipping that can increase risk of overconfidence. Kessler et al. (2017) leveled a valid criticism of the early publications on treatment selection in depression (DeRubeis et al., 2014; Huibers et al., 2015), arguing that, because variable selection was performed within the sample on which the model was evaluated, "any attempt to use the coefficients in these models to predict differential treatment response in a new sample of patients would almost certainly yield less positive effects than those suggested by the results of studies" (p. 6). Kriegeskorte, Simmons, Bellgowan, and Baker (2009) also note that this approach can lead to distorted descriptive statistics and invalid statistical inference. The size of the sample used in these analyses, although drawn from the largest RCT comparing these two treatments to date, did not allow for a true hold-out. As described by Hastie et al. (2009), the CV scheme applied here has given an unfair advantage to the predictors as they were chosen on the basis of the full sample; therefore, we do not approximate an evaluation of

our models in a completely independent test set. Thus, we stress that this model would need to be validated in an independent sample before being considered for use as a clinical decision tool.

Finally, we must acknowledge that the sample size in which these analyses (and the vast majority of the existing literature on treatment select) were performed was likely insufficient for this work. Detecting group differences associated with small effect sizes ( $\sim d = .20 - .30$ ) using a multivariable regression model with just 6 predictors (with an alpha level of  $p < .05$  and 80% power) requires a sample of 686 per group (Cohen, 1992). The sample sizes needed for detecting moderator effects are significantly larger (Aguinis, Beaty, Boik, & Pierce, 2005). A recent simulation by Luedtke, Sadikova, and Kessler (2018) suggested that at least 300 patients per treatment arm are needed for adequate statistical power to detect clinically significant improvements in treatment response due to treatment selection. Very few RCTs meet these sample size requirements, leading Kessler to propose the use of large observational studies, pragmatic trial designs, or pooled datasets (Kessler, 2018).

Despite these shortcomings, we believe this work represents an important step towards the types of studies that would address many of these criticisms. These analyses propose potential moderators and present a specific model that could be investigated in future studies and identify an important issue that merits increased attention by those interested in precision medicine: methodological heterogeneity in variable selection.

### Future Directions

The treatment selection subfield of precision medicine in mental health is still in its developmental stage, and the statistical methods described in this and other similar papers are constantly evolving (Zilcha-Mano, 2018). Although replication and external validation are essential steps that should precede the implementation of any specific treatment selection model, the publication and discussion of candidate predictors, models and statistical approaches are equally important, as they set the foundation for future efforts.

A wide variety of feature selection techniques have been employed in recent efforts to construct treatment selection models in mental health, and no clear guidance exists as to which approach is best. We have presented an example of a new variable selection approach that incorporates several of the leading techniques in order to identify reliable predictors of differential response to treatment. We propose

this specific combination as a starting point and suggest that future efforts should explore different permutations, such as adding other methods (e.g., Support Vector Machines), reducing the number of approaches used, using different combinations, or adjusting the settings within each of these techniques. Examples of the latter include adjusting the thresholds for the inclusion of variables and specifying different tuning parameters (Kuhn & Johnson, 2013).

### Conclusion

We refined the Personalized Advantage Index approach to generate individual treatment recommendations for adults with mild to moderate depression not taking antidepressants who are deciding between CBT versus PDT. Our novel approach synthesized the results of four different variable selection techniques by selecting the patient characteristics that are consistently identified as associated with (differential) treatment outcome. Although no significant efficacy differences were found between CBT and PDT across the total sample, the resulting treatment recommendations suggested that for the majority of the individual patients, one of the treatments could be predicted to be more efficacious than the other based on a model including four pre-treatment patient characteristics (anxiety sensitivity, depression symptom severity, extraversion, and psychological treatment needs). The small sample and lack of a separate validation sample indicate the need for prospective tests (e.g., Lutz, Zimmermann, Müller, Deisenhofer, & Rubel, 2017) before using this model for treatment selection, but these findings add to a growing literature on the potential for model-guided treatment recommendations to improve patient outcomes for depression.

### Acknowledgments

We thank Robert DeRubeis for helpful comments on the manuscript and support of this work. We also wish to acknowledge the reviewers, whose feedback greatly improved this article and made us feel like we had gained three insightful coauthors. This work was supported by a grant from MQ: Transforming mental health MQ14PM\_27.

### Funding

Financial support for this work was provided by MQ Foundation to ZC [MQ: Transforming mental health MQ14PM\_27]. MQ had no role in the study design,

collection, analysis, or interpretation of the data, or in writing the manuscript or the decision to submit the article for publication. The randomized clinical trial from which data were drawn for this study was financed by an unrestricted research grant by Wyeth Pharmaceuticals, The Netherlands. Arkin Mental Health Care, The Netherlands, financially supported research logistics and the contributions of ED, HLV, and JJMD. Vrije Universiteit Amsterdam, Faculty of Behavioral and Movement Sciences, Section Clinical Psychology, The Netherlands, financially supported ED's contributions to the study. None of the sponsors had a role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; nor in the preparation, review, or approval of the manuscript. The opinions and assertions contained in this article should not be construed as reflecting the views of the sponsors.

### Supplemental data

Supplemental data for this article can be accessed [doi:10.1080/10503307.2018.1563312](https://doi.org/10.1080/10503307.2018.1563312).

### Note

<sup>1</sup> We decided to select the number of variables identified by RF, and not to use BART's built-in permutation test for thresholding variable importance because BART's test was created for use in contexts where the variable search was not biased to focus on treatment interactions.

### References

- Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology, 90*(1), 94–107. doi:10.1037/0021-9010.90.1.94
- Austin, P. C., & Tu, J. V. (2004). Bootstrap methods for developing predictive models. *The American Statistician, 58*(2), 131–137. doi:10.1198/0003130043277
- Barber, J. P., & Muenz, L. R. (1996). The role of avoidance and obsessiveness in matching patients to cognitive and interpersonal psychotherapy: Empirical findings from the treatment for depression collaborative research program. *Journal of Consulting and Clinical Psychology, 64*(5), 951–958. doi:10.1037/0022-006X.64.5.951
- Barth, J., Munder, T., Genger, H., Nüesch, E., Trelle, S., Znoj, H., ... Cuijpers, P. (2013). Comparative efficacy of seven psychotherapeutic interventions for patients with depression: A network meta-analysis. *PLoS Medicine, 10*(5), e1001454. doi:10.1371/journal.pmed.1001454
- Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression*. New York, NY: Guilford press.
- Bleich, J., Kapelner, A., George, E. I., & Jensen, S. T. (2014). Variable selection for bart: An application to gene regulation. *The Annals of Applied Statistics, 8*(3), 1750–1781. doi:10.1214/14-Aoas755
- Chekroud, A. M., Gueorguieva, R., Krumholz, H. M., Trivedi, M. H., Krystal, J. H., & McCarthy, G. (2017). Reevaluating the efficacy and predictability of antidepressant treatments: A symptom clustering approach. *JAMA Psychiatry, 74*(4), 370–378. doi:10.1001/jamapsychiatry.2017.0025
- Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., & Trivedi, M. H. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet Psychiatry, 3*, 243–250. doi:10.1016/S2215-0366(15)00471-x
- Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics, 4*(1), 266–298. doi:10.1214/09-AOAS285
- Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. doi:10.1037/0033-2909.112.1.155
- Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment selection in depression. *Annual Review of Clinical Psychology, 14*, 209–236. doi:10.1146/annurev-clinpsy-050817-084746
- De Beurs, E., & Zitman, F. (2005). *De Brief symptom Inventory (BSI). De betrouwbaarheid en validiteit van een handzaam alternatief voor de SCL-90*. Leiden: Leids universitair medisch centrum.
- 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. doi:10.1002/da.22755
- de Jonghe, F. (1994). *Leidraad voor het scoren van de Hamilton depression Rating Scale [Hamilton depression rating scale scoring manual]*. Amsterdam: Benecke.
- de Jonghe, F. (2005). *Kort en Krachtig: Kortdurende Psychoanalytische Steungevende Psychotherapie [Short and snappy: Short-term psychoanalytic supportive psychotherapy]*. Amsterdam: Benecke NI.
- de Jonghe, F., de Maat, S., Van, R., Hendriksen, M., Kool, S., van Aalst, G., & Dekker, J. (2013). Short-term psychoanalytic supportive psychotherapy for depressed patients. *Psychoanalytic Inquiry, 33*(6), 614–625. doi:10.1080/07351690.2013.835184
- DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014). The Personalized advantage index: Translating research on prediction into individualized treatment recommendations. A demonstration. *PLoS One, 9*(1), e83875. doi:10.1371/journal.pone.0083875
- Driessen, E., Smits, N., Dekker, J., Peen, J., Don, F., Kool, S., ... Van, H. (2016). Differential efficacy of cognitive behavioral therapy and psychodynamic therapy for major depression: A study of prescriptive factors. *Psychological Medicine, 46*(4), 731–744. doi:10.1017/S0033291715001853
- Driessen, E., Van, H. L., Don, F. J., Peen, J., Kool, S., Westra, D., ... Twisk, J. W. (2013). The efficacy of cognitive-behavioral therapy and psychodynamic therapy in the outpatient treatment of major depression: A randomized clinical trial. *American Journal of Psychiatry, 170*(9), 1041–1050. doi:10.1176/appi.ajp.2013.12070899
- Driessen, E., Van, H. L., Peen, J., Don, F. J., Kool, S., Westra, D., ... Dekker, J. J. (2015). Therapist-rated outcomes in a randomized clinical trial comparing cognitive behavioral therapy and psychodynamic therapy for major depression. *Journal of Affective Disorders, 170*, 112–118. doi:10.1016/j.jad.2014.08.023
- Driessen, E., Van, H. L., Peen, J., Don, F. J., Twisk, J. W., Cuijpers, P., & Dekker, J. J. (2017). Cognitive-behavioral versus psychodynamic therapy for major depression: Secondary outcomes of a randomized clinical trial. *Journal of*

- Consulting and Clinical Psychology*, 85(7), 653–663. doi:10.1037/ccp0000207
- Driessen, E., Van, H. L., Schoevers, R. A., Cuijpers, P., Van Aalst, G., Don, F. J., ... Peen, J. (2007). Cognitive behavioral therapy versus short psychodynamic supportive psychotherapy in the outpatient treatment of depression: A randomized controlled trial. *BMC Psychiatry*, 7(1), 58. doi:10.1186/1471-244X-7-58
- Fiedler, K. (2011). Voodoo correlations are everywhere—not only in neuroscience. *Perspectives on Psychological Science*, 6(2), 163–171. doi:10.1177/1745691611400237
- Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, 115(27), E6106–E6115.
- Fournier, J. C., DeRubeis, R. J., Shelton, R. C., Hollon, S. D., Amsterdam, J. D., & Gallop, R. (2009). Prediction of response to medication and cognitive therapy in the treatment of moderate to severe depression. *Journal of Consulting and Clinical Psychology*, 77(4), 775–787. doi:10.1037/a0015401
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1. doi:10.18637/jss.v033.i01
- Garge, N. R., Bobashev, G., & Eggleston, B. (2013). Random forest methodology for model-based recursive partitioning: The mobForest package for R. *BMC Bioinformatics*, 14, 125. doi:10.1186/1471-2105-14-125
- Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31(14), 2225–2236. doi:10.1016/j.patrec.2010.03.014
- Gibbons, M. B. C., Gallop, R., Thompson, D., Luther, D., Crits-Christoph, K., Jacobs, J., ... Crits-Christoph, P. (2016). Comparative effectiveness of cognitive therapy and dynamic psychotherapy for major depressive disorder in a community mental health setting: A randomized clinical noninferiority trial. *JAMA Psychiatry*, 73(9), 904–911. doi:10.1001/jamapsychiatry.2016.1720
- Gillan, C. M., & Whelan, R. (2017). What big data can do for treatment in psychiatry. *Current Opinion in Behavioral Sciences*, 18, 34–42. doi:10.1016/j.cobeha.2017.07.003
- Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1), 44–65. doi:10.1080/10618600.2014.907095
- Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, 23(1), 56–62. doi:10.1136/jnnp.23.1.56
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed). New York, NY: Springer.
- Hoekstra, H., Ormel, J., & De Fruyt, F. (2003). *NEO-FFI Big five persoonlijkheidsvragenlijst* [NEO-FFI Big Five personality questionnaire]. Lisse, The Netherlands: Swets & Zeitlinger BV.
- Hoffart, A., & Johnson, S. U. (2017). Psychodynamic and cognitive-behavioral therapies are more different than you think: Conceptualizations of mental problems and consequences for studying mechanisms of change. *Clinical Psychological Science*, 5(6), 1070–1086. doi:10.1177/2167702617727096
- 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. doi:10.1371/journal.pone.0140771
- Iniesta, R., Hodgson, K., Stahl, D., Malki, K., Maier, W., Rietschel, M., ... Uher, R. (2018). Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables. *Scientific Reports*, 8(1), 5530. doi:10.1038/s41598-018-23584-z
- Iniesta, R., Malki, K., Maier, W., Rietschel, M., Mors, O., Hauser, J., ... Uher, R. (2016). Combining clinical variables to optimize prediction of antidepressant treatment outcomes. *Journal of Psychiatric Research*, 78, 94–102. doi:10.1016/j.jpsychires.2016.03.016
- Kapelner, A., & Bleich, J. (2016). Bartmachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70(4), 1–40. doi:10.18637/jss.v070.i04
- 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. doi:10.1002/da.22731
- Kessler, R. C. (2018). The potential of predictive analytics to provide clinical decision support in depression treatment planning. *Current Opinion in Psychiatry*, 31(1), 32–39. doi:10.1097/YCO.0000000000000377
- Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Ebert, D. D., ... Zaslavsky, A. M. (2017). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiology and Psychiatric Sciences*, 26(1), 22–36. doi:10.1017/S2045796016000020
- Kikkert, M. J., Driessen, E., Peen, J., Barber, J. P., Bockting, C., Schalkwijk, F., ... Dekker, J. J. (2016). The role of avoidant and obsessive-compulsive personality disorder traits in matching patients with major depression to cognitive behavioral and psychodynamic therapy: A replication study. *Journal of Affective Disorders*, 205, 400–405. doi:10.1016/j.jad.2016.08.017
- Kraemer, H. C. (2013). Discovering, comparing, and combining moderators of treatment on outcome after randomized clinical trials: A parametric approach. *Statistics in Medicine*, 32(11), 1964–1973. doi:10.1002/sim.5734
- Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, 12(5), 535–540. doi:10.1038/nn.2303
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York, NY: Springer.
- Luedtke, A. R., 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*.
- 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(1), 32–39. doi:10.1037/0022-0167.54.1.32
- Lutz, W., Saunders, S. M., Leon, S. C., Martinovich, Z., Kosfelder, J., Schulte, D., ... Tholen, S. (2006). Empirically and clinically useful decision making in psychotherapy: Differential predictions with treatment response models. *Psychological Assessment*, 18(2), 133–141. doi:10.1037/1040-3590.18.2.133
- Lutz, W., Zimmermann, D., Müller, V. N., Deisenhofer, A.-K., & Rubel, J. A. (2017). Randomized controlled trial to evaluate the effects of personalized prediction and adaptation tools on treatment outcome in outpatient psychotherapy: Study protocol. *BMC Psychiatry*, 17(1), 306. doi:10.1186/s12888-017-1464-2
- Molenaar, P., Don, F., van den Bout, J., Sterk, F., & Dekker, J. (2009). *Cognitieve gedragstherapie bij depressie [Cognitive behavioral therapy for depression]*. The Netherlands: Bohn Stafleu van Loghum. Retrieved from <https://www.gerdiex.nl/wpcontent/uploads/2012/11/Cognitieve-gedragstherapie-bij-depressie.pdf>
- Niles, A. N., Loerinc, A. G., Krull, J. L., Roy-Byrne, P., Sullivan, G., Sherbourne, C. D., ... Craske, M. G. (2017). Advancing

- personalized medicine: Application of a novel statistical method to identify treatment moderators in the coordinated anxiety learning and management study. *Behavior Therapy*, 48(4), 490–500. doi:10.1016/j.beth.2017.02.001
- Niles, A. N., Wolitzky-Taylor, K. B., Arch, J. J., & Craske, M. G. (2017). Applying a novel statistical method to advance the personalized treatment of anxiety disorders: A composite moderator of comparative drop-out from CBT and ACT. *Behaviour Research and Therapy*, 91, 13–23. doi:10.1016/j.brat.2017.01.001
- Reiss, S., Peterson, R. A., Gursky, D. M., & McNally, R. J. (1986). Anxiety sensitivity, anxiety frequency and the prediction of fearfulness. *Behaviour Research and Therapy*, 24(1), 1–8. doi:10.1016/0005-7967(86)90143-9
- Rizopoulos, D. (2009). BootStepAIC: Bootstrap stepAIC. *R Package Version*, 1(0).
- Rubel, J. A., Fisher, A. J., Husen, K., & Lutz, W. (2018). Translating person-specific network models into personalized treatments: Development and demonstration of the dynamic assessment treatment algorithm for individual Networks (DATA-IN). *Psychotherapy and Psychosomatics*, 87, 249–251. doi:10.1159/000487769
- Schweizer, S., Cohen, Z., Hayes, R., DeRubeis, R., Crane, C., Kuyken, W., & Dalglish, T. (submitted). Relapse prevention for antidepressant medication (ADM) responders with recurrent depression: Using the personalized advantage index to decide between maintenance ADM and mindfulness-based cognitive therapy.
- Smagula, S. F., Wallace, M. L., Anderson, S. J., Karp, J. F., Lenze, E. J., Mulsant, B. H., ... Lotrich, F. E. (2016). Combining moderators to identify clinical profiles of patients who will, and will not, benefit from aripiprazole augmentation for treatment resistant late-life major depressive disorder. *Journal of Psychiatric Research*, 81, 112–118. doi:10.1016/j.jpsychires.2016.07.005
- Stekhoven, D. J., & Bühlmann, P. (2012). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics (Oxford, England)*, 28(1), 112–118. doi:10.1093/bioinformatics/btr597
- Team, R. C. (2000). *R language definition*. Vienna, Austria: R foundation for statistical computing.
- Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., & Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine*, 358(3), 252–260. doi:10.1056/NEJMsa065779
- Uher, R., Perlis, R., Henigsberg, N., Zobel, A., Rietschel, M., Mors, O., ... Bajs, M. (2012). Depression symptom dimensions as predictors of antidepressant treatment outcome: Replicable evidence for interest-activity symptoms. *Psychological Medicine*, 42(5), 967–980. doi:10.1017/S0033291711001905
- Veeninga, A., & Hafkenscheid, A. (2004). KORT INSTRUMENTEEL De Patienten Behoeften Vragenlijst (PBV). *Gedragstherapie*, 37, 197–204.
- Vittengl, J. R., Clark, L. A., Thase, M. E., & Jarrett, R. B. (2017). Initial steps to inform selection of continuation cognitive therapy or fluoxetine for higher risk responders to cognitive therapy for recurrent major depressive disorder. *Psychiatry Research*, 253, 174–181. doi:10.1016/j.psychres.2017.03.032
- Wallace, M. L., Frank, E., & Kraemer, H. C. (2013). A novel approach for developing and interpreting treatment moderator profiles in randomized clinical trials. *JAMA Psychiatry*, 70(11), 1241–1247. doi:10.1001/jamapsychiatry.2013.1960
- Webb, C. A., Trivedi, M. H., Cohen, Z. D., Dillon, D. G., Fournier, J. C., Goer, F., ... Parsey, R. (2018). Personalized prediction of antidepressant v. Placebo response: Evidence from the EMBARC study. *Psychological Medicine*, 1–10. doi:10.1017/S0033291718001708
- World Health Organization. (2017). *Depression and other common mental Disorders: Global health estimates* (Report No. CC BY-NC-SA 3.0 IGO). Geneva.
- Zilcha-Mano, S. (2018). Major developments in methods addressing for whom psychotherapy may work and why. *Psychotherapy Research*, 1–16. doi:10.1080/10503307.2018.1429691
- Zilcha-Mano, S., Keefe, J. R., Chui, H., Rubin, A., Barrett, M. S., & Barber, J. P. (2016). Reducing dropout in treatment for depression: Translating dropout predictors into individualized treatment recommendations. *The Journal of Clinical Psychiatry*, 77(12), e1584–e1590. doi:10.4088/JCP.15m10081
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320. doi:10.1111/j.1467-9868.2005.00503.x