



## Cross-trial prediction in psychotherapy: External validation of the Personalized Advantage Index using machine learning in two Dutch randomized trials comparing CBT versus IPT for depression

Suzanne C. van Bronswijk, Sanne J. E. Bruijniks, Lorenzo Lorenzo-Luaces, Robert J. Derubeis, Lotte H. J. M. Lemmens, Frenk P. M. L. Peeters & Marcus J. H. Huibers

To cite this article: Suzanne C. van Bronswijk, Sanne J. E. Bruijniks, Lorenzo Lorenzo-Luaces, Robert J. Derubeis, Lotte H. J. M. Lemmens, Frenk P. M. L. Peeters & Marcus J. H. Huibers (2021) Cross-trial prediction in psychotherapy: External validation of the Personalized Advantage Index using machine learning in two Dutch randomized trials comparing CBT versus IPT for depression, *Psychotherapy Research*, 31:1, 78-91, DOI: [10.1080/10503307.2020.1823029](https://doi.org/10.1080/10503307.2020.1823029)

To link to this article: <https://doi.org/10.1080/10503307.2020.1823029>



© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 23 Sep 2020.



[Submit your article to this journal](#)



Article views: 2799



[View related articles](#)




[View Crossmark data](#)



Citing articles: 4 [View citing articles](#)

EMPIRICAL PAPER

## Cross-trial prediction in psychotherapy: External validation of the Personalized Advantage Index using machine learning in two Dutch randomized trials comparing CBT versus IPT for depression

SUZANNE C. VAN BRONSWIJK <sup>1†</sup>, SANNE J. E. BRUIJNIKS<sup>2,3†</sup>,  
LORENZO LORENZO-LUACES<sup>4</sup>, ROBERT J. DERUBEIS<sup>5</sup>, LOTTE H. J. M. LEMMENS<sup>1</sup>,  
FRENK P. M. L. PEETERS<sup>1</sup>, & MARCUS. J. H. HUIBERS<sup>3,5</sup>

<sup>1</sup>Department of Clinical Psychological Science, Maastricht University, Maastricht, The Netherlands; <sup>2</sup>Department of Clinical Psychology and Psychotherapy, University of Freiburg, Freiburg, Germany; <sup>3</sup>Department of Clinical Psychology, Amsterdam Public Health research institute, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands; <sup>4</sup>Department of Psychological and Brain Sciences, Indiana University, Bloomington, USA & <sup>5</sup>Department of Psychology, University of Pennsylvania, Philadelphia, USA

(Received 23 December 2019; revised 7 September 2020; accepted 8 September 2020)

### Abstract

**Objective:** Optimizing treatment selection may improve treatment outcomes in depression. A promising approach is the Personalized Advantage Index (PAI), which predicts the optimal treatment for a given individual. To determine the generalizability of the PAI, models need to be externally validated, which has rarely been done.

**Method:** PAI models were developed within each of two independent trials, with substantial between-study differences, that both compared CBT and IPT for depression (STEPd:  $n = 151$  and FreqMech:  $n = 200$ ). Subsequently, both PAI models were tested in the other dataset.

**Results:** In the STEPd study, post-treatment depression was significantly different between individuals assigned to their PAI-indicated treatment versus those assigned to their non-indicated treatment ( $d = .57$ ). In the FreqMech study, post-treatment depression was not significantly different between patients receiving their indicated treatment versus those receiving their non-indicated treatment ( $d = .20$ ). Cross-trial predictions indicated that post-treatment depression was not significantly different between those receiving their indicated treatment and those receiving their non-indicated treatment ( $d = .16$  and  $d = .27$ ). Sensitivity analyses indicated that cross-trial prediction based on only overlapping variables didn't improve the results.

**Conclusion:** External validation of the PAI has modest results and emphasizes between-study differences and many other challenges.

**Keywords:** depression; cognitive behavioural therapy; interpersonal psychotherapy; precision medicine; prediction; external validation

**Clinical or methodological significance of this article:** This study demonstrates the development and external validation of the Personalized Advantage Index (PAI), a prediction algorithm that aims for targeted prescription of different types of treatment. Two PAI models were developed using independent datasets of two Dutch randomized studies comparing cognitive behavioural therapy and interpersonal psychotherapy for depression. Although results of this study indicate that treatment recommendations based on the PAI can be promising, results concerning external validation are modest and

---

<sup>†</sup>These authors contributed equally to the work

Correspondence concerning this article should be addressed to Suzanne C. van Bronswijk Department of Clinical Psychological Science, Maastricht University, P.O. Box 616, 6200 MD Maastricht, Maastricht, MD 6200, The Netherlands. Email: [suzanne.vanbronswijk@maastrichtuniversity.nl](mailto:suzanne.vanbronswijk@maastrichtuniversity.nl)

This article has been republished with minor changes. These changes do not impact the academic content of the article.

emphasize the many challenges of external validation including heterogeneity in data collection, statistical methods, study populations and treatments.

## Introduction

Although there are multiple treatments available for major depressive disorder (MDD), only about half of the individuals recover after treatment (National Health Service, 2018). Current guidelines prescribe antidepressants, cognitive behavioural therapy (CBT) or interpersonal psychotherapy (IPT) among the first choices of treatment for depression, and this choice is often based on the severity of the depression and the number and effects of earlier treatments (American Psychiatric Association, 2009; Lorenzo-Luaces et al., 2015; National Collaborating Centre for Mental Health, 2010). However, individuals may respond differentially to different treatments, and these responses are difficult to predict. Matching individuals to their optimal treatment, also referred to as personalized or precision medicine, seems to be a promising way to improve treatment outcomes for depression (Cohen & DeRubeis, 2018).

In order to match patients to their optimal treatment, it is necessary to predict which treatment works best for whom. DeRubeis et al. (2014) introduced and demonstrated the Personalized Advantage Index (PAI) which generates actionable individual treatment recommendations. The PAI represents, for a certain individual, the difference in predicted outcomes between two or more treatments. In order to compute this difference, DeRubeis et al. built a multivariable regression model based on pre-treatment variables that were found to be prognostic (i.e., predicting treatment outcome irrespective of treatment, also known as predictors) and prescriptive (i.e., predicting differential treatment outcome, also known as moderators). In that study, which compared CBT with antidepressants, 60% of the patients had a statistically significant and clinically relevant difference between their predicted optimal treatment and their predicted non-optimal treatment. Similar studies have been conducted since, investigating individual advantages in CBT versus IPT for MDD (Huibers et al., 2015), CBT versus psychodynamic treatment for MDD (Cohen et al., 2019), and CBT versus CBT with integrated exposure and emotion-focused elements for MDD (Friedl et al., 2020). The PAI has also been studied in the contexts of supportive-expressive therapy, antidepressants and placebo for MDD (Zilcha-Mano et al., 2016), eye movement desensitization and reprocessing versus cognitive therapy for posttraumatic stress disorder (PTSD, Deisenhofer et al., 2018), prolonged

exposure versus cognitive processing therapy for PTSD (Keefe et al., 2018), antidepressants versus placebo for MDD (Webb et al., 2019), and CBT versus a positive psychology intervention for MDD (Lopez-Gomez et al., 2019). Overall, these studies indicated that different treatments may have different clinically relevant effects for a given individual, and that use of the PAI may improve outcomes by optimizing treatment selection.

However, to determine the generalizability of PAI models in actual clinical practice, the predictive accuracy of these models needs to be externally validated (Bleeker et al., 2003; Cohen & DeRubeis, 2018; Gillan & Whelan, 2017). External validation is considered a second phase in multivariable prognostic research, following model development and preceding impact studies (Moons et al., 2009). Although it is widely acknowledged that the performance of prediction models should be examined by external validation studies, this has been done very infrequently in medicine in general and mental health care specifically (Siontis et al., 2015). Most published studies examining PAI predictions that have focused on validation, focus on the issue of *internal* validation using techniques such as bootstrapping (random sampling with replacement), cross-validation, and split-sampling (Efron, 1983; Efron & Tibshirani, 1993; Efron & Tibshirani, 1997). Internal validation methods, in particular bootstrapping (Steyerberg et al., 2001), can provide bias-corrected estimates of model performance. However, models tend to have better results in data on which the model was developed than in new data, which is often referred to as “overfitting” (i.e., high variance in the bias-variance trade-off: models perform well on training data but have high error rates across testing dataset; Bleeker et al., 2003). Internal validation alone is unlikely to be sufficient for PAI models which have typically been based on relatively small datasets, and are considered to be prone to the risk of overfitting (Cohen & DeRubeis, 2018; Luedtke et al., 2019). One PAI study that used an external validation sample successfully was recently presented by Delgadillo and Gonzalez Salas Duhne (2020). In this study, individual advantages in CBT versus person centred counselling were investigated for depressed individuals in primary care. In order to develop a generalizable PAI model, the data was split into a training sample and a test sample. The training sample was used to develop two prognostic models to predict outcome, one model for each treatment. The test sample was subsequently

used to test the accuracy of these models. Interestingly, the prognostic models that were developed in the training sample showed significant accuracy in the test sample, indicating proof for generalizability. As expected, the patients who received their indicated treatment showed better response compared to the patients who received their non-indicated treatment.

The present study used a machine learning approach to compute PAI models in data from two Dutch randomized trials comparing CBT and IPT for MDD. Subsequently, the PAI models that were developed and tested within each sample were externally validated in the other sample to determine the generalizability of these models to future groups of patients. To our knowledge, this study is the first that attempts to externally validate the generalizability of PAI models between two randomized trials.

## Methods

### Study Design and Clinical Trial Data

PAI models were developed within each dataset of two Dutch randomized trials comparing acute-phase CBT and IPT for MDD. The first trial is called the STEPd trial and was conducted between February 2007 and April 2012, randomizing depressed patients into CBT ( $n = 76$ ), IPT ( $n = 75$ ) or a waitlist condition followed by a treatment of choice ( $n = 31$ ). Main results were reported in 2015 and 2019, indicating no significant acute and long-term outcome difference between CT and IPT on average (Lemmens et al., 2015; Lemmens et al., 2019). The second trial is the FreqMech trial ( $n = 200$ ), which was conducted between November 2014 and January 2018, randomizing depressed patients into different session frequencies of CBT and IPT (i.e., CBT weekly ( $n = 49$ ), CBT twice weekly ( $n = 49$ ), IPT weekly ( $n = 55$ ), IPT twice weekly ( $n = 47$ )). A higher session frequency led to more reduction of depression, but there were no average group differences between CBT and IPT (Bruijniks et al., 2020). Although the studies used the same treatment protocols for CBT and IPT (i.e., only the session frequency differed), the studies differed in a number of other important aspects. First, while the STEPd trial was a single centre study conducted in one specific area in the Netherlands (Maastricht), the FreqMech trial was a multi-centre study conducted in multiple areas in the Netherlands (Maastricht, Amsterdam, The Hague, Leiden, Utrecht, Nijmegen, Oss, Haarlem). Second, the two studies had a different number of therapists (10 in the STEPd trial and 76 in the FreqMech trial). Third, the average levels of pre- and post-

treatment depression differed: patients in the FreqMech trial started with higher levels of depression compared to patients in the STEPd trial (average Beck Depression Inventory – version II (BDI-II) scores of 29.8 versus 34.7 for STEPd and FreqMech respectively). After treatment, participants in the FreqMech study ended up with levels of moderate depression (average BDI-II scores between 20.0 and 24.2) that were still higher compared to the patients in the STEPd trial who showed mild levels of depression (average BDI-II scores between 13.3 and 15.8). This difference could be explained by different inclusion criteria: for the STEPd study a minimum BDI-II score of 10 was required, whereas for FreqMech this was a minimum of 20. Fourth, as described below, in- and exclusion criteria were slightly different. Details about the protocols of both trials can be found in Lemmens et al. (2011) and Bruijniks et al. (2015).

### Participants

Participants were adult outpatients (18–65 years) of the Academic Community Mental Health Centre Maastricht (both FreqMech and STEPd), and for FreqMech of PsyQ (The Hague, Leiden, Amsterdam or Haarlem), Altrecht GGZ (Utrecht), GGZ inGeest (Amsterdam), GGZ Oost-Brabant (Oss), and Pro Persona (Nijmegen). Participants had a primary diagnosis of MDD, confirmed by the Structured Clinical Interview for DSM-IV Axis I disorders (SCID-I; First et al., 1995) or Mini International Neuropsychiatric Interview-Plus (MINI-Plus; Van Vliet et al., 2000). Joint inclusion criteria were internet access and sufficient knowledge of the Dutch language. Joint exclusion criteria were a high suicide risk, or a diagnosis of abuse disorders (according to DSM-IV or V), concomitant psychological treatment, or mental retardation (IQ < 80). Inclusion and exclusion criteria between trials differed in that STEPd excluded bipolar or highly chronic (current episode > 5 years) depression, and current use of antidepressant medication, while in the FreqMech trial patients were only excluded when they started antidepressants or changed the dosage in the past three months. The FreqMech trial additionally only included patients with a pre-treatment score  $\geq 20$  on the Beck Depression Inventory II (BDI-II) and excluded presence of a diagnosis of a cluster A or B personality disorders (according to DSM-IV or V). In addition, patients that had received more than five sessions of adequate CBT or IPT in the previous year were excluded. The Medical Ethics Committee (MEC) approved both trial protocols (i.e., MEC of Maastricht University for the STEPd trial, MEC of VU Medical Centre for the FreqMech trial), and all participants provided written informed consent. The

trials are registered at the Netherlands Trial Register, part of the Dutch Cochrane Centre (NTR838 and NTR4856).

## Treatments

The same treatment protocols were used for both studies: CBT was based on the manual by Beck and colleagues (Beck et al., 1979) and IPT was based on the manual by Klerman and colleagues (Klerman et al., 1984). In both trials, patients received 12 up to 20 individual sessions of 45 min, but session frequency differed per trial: in the Freq-Mech trial half of the patients received 16 weekly sessions and the other half received 16 twice weekly sessions. The last four sessions were scheduled weekly for both conditions. In the STEPd trial patients received weekly sessions with some flexibility to schedule appointments less often than weekly. Treatment competence was rated good to excellent in 83–90% of the videotapes from the STEPd trial (Lemmens et al., 2015) and poor to excellent in the FreqMech trial (only 12–16% of the video-tapes were rated good-excellent; Bruijniks et al., 2020) using the Cognitive Therapy Scale for CBT (Dobson et al., 1985), and the short version of the IPT Adherence and Quality Scale for IPT (Stuart, 2011).<sup>1</sup> In both trials, there were significant differences in therapy-specific behaviour, with higher CBT-specific behaviour in CBT as compared to IPT, and higher IPT-specific behaviour in IPT as compared to CBT (rated with the Collaborative Study Psychotherapy Rating Scale version 6; Hollon et al., 1988).

## Outcome

In both trials, the BDI-II (Beck et al., 1996) was used as the primary outcome measure of post-treatment depression symptom severity. The BDI-II has 21 items with higher scores indicating higher severity. Reliability and validity of the BDI-II has been established (Wang & Gorenstein, 2013). The post-treatment time point was 7 and 6 months after the start of therapy in the STEPd study and the FreqMech study respectively.

## Variables

Pre-treatment variables covered the following domains: demographics (11 variables), depression and other symptoms (18 variables), present and previous health care use (7 variables), general functioning (14 variables), psychological processes (14 variables) and life and family history (7 variables).

An overview of the available pre-treatment variables can be found in Data Supplement 1.

## Statistical Analyses

### Data preparation

**Pre-selection of pre-treatment variables.** Pre-treatment variables were pre-selected for each dataset by examining their correlation matrices to remove highly correlated and redundant variables. This was done in a step-wise approach by first removing the predictors with the most correlated relationship (Kuhn & Johnson, 2013). In this way, a minimum number of variables was removed below the threshold of .7 corrected for attenuation. A few exceptions to this rule were made when variables were considered to relate to different underlying information. These decisions were achieved via consensus between the study authors. An overview of the included pre-treatment variables can be found in Data Supplement 1.

### Transformation of pre-treatment variables.

Pre-treatment variables were prepared in the following ways: 1. Categorical variables with limited observations in some of the categories were merged, since previous research recommends at least 10% of the sample in each category (Kuhn & Johnson, 2013), 2. Continuous predictors were checked on normality and were, in case of non-normality, log transformed (rightly skewed distributions) or squared (left skewed distributions) to obtain a normal distribution, 3. Normal distributed continuous variables were standardized, and dichotomous and discrete variables were centred by recoding the variables to prevent potential errors in statistical inference (Kraemer & Blasey, 2004). The applied transformation for each pre-treatment variable is described in Data Supplement 1.

### Transformation of the outcome variable.

Because residuals of the outcome variable in both datasets were not homoscedastic, outcomes were transformed using a squared root transformation. The transformed outcome variable had residuals that were not significantly different from a normal distribution.

**Data imputation.** Missing values of the pre-treatment and outcome variables were imputed using the “MissForest” package in R (Stekhoven, 2011; Stekhoven & Bühlmann, 2011; Tang & Ishwaran, 2017). The following information was used as input for the imputation procedure: 1) change scores from baseline of all non-missing post-treatment BDI-II outcomes (at 3 and 7 months for the

STEPd trial, at 6 months for the FreqMech trial; Moons et al., 2006); 2) all values of non-missing pre-treatment variables; 3) the received treatment (CBT versus IPT; twice weekly versus once weekly). The accuracy of the imputation method was tested by comparing imputed values with actual values using artificially produced missing data in the non-missing dataset. These comparisons were done by calculating the normalized root mean squared error (NRMSE) for continuous data and the proportion of falsely classified entries (PFC) for categorical data.

**Building the Personalized Advantage Index Model.** Earlier studies have shown that building two separate prognostic models (one for the patients who received CBT and one for the patients who received IPT) might lead to a better PAI model compared to a treatment-modality interaction model, in particular with small samplesizes (Delgadillo & Gonzalez Salas Duhne, 2020). To decide on the right approach, we tested the predictive accuracy of both the treatment-modality interaction model and the prognostic models.

Following (Keefe et al., 2018; van Bronswijk et al., 2019; van Bronswijk et al., 2019), two machine learning techniques were used to select pre-treatment variables in each dataset to build a treatment-modality interaction PAI model. Although the STEPd dataset was used before to develop a PAI model (Huibers et al., 2015), in that study less advanced methods were applied to build a prediction model (i.e., no imputation of missing data, stepwise variable selection approach and leave-one-out cross validation). Therefore, in the present study, we decided to build a PAI model for the STEPd dataset and the FreqMech dataset using state-of-the-art machine learning techniques. In the first variable selection step, a random forest algorithm was applied with the package “mobforest” in R (Garge et al., 2013). In this algorithm, a predetermined model is used, and multiple trees are created by splitting bootstrapped samples into subgroups based on pre-treatment variables that lead to significantly different model behaviour on either side of the split. The predetermined model for the current analyses was a regression model with post-treatment BDI-II scores as the dependent variable and treatment as the independent variable ( $y = \text{treatment}$ ). A split variable therefore indicates a variable within which there are treatment differences (i.e., a potential moderator). At each split, a random subset of variables was selected, to prevent that variables with smaller effects are dominated by the stronger variables (Strobl et al., 2008). Variables

were subsequently selected if the difference between predictive accuracy of a variable versus the predictive accuracy of a randomly permuted variable on the outcome (i.e., the variable importance score) was higher than the absolute value of the lowest ranking predictor. Parameters were set as follows: 10,000 trees were computed with a minimum level of 0.10 for splits and a minimum node size for splitting of 15 individuals. In the second variable selection step, the variables that resulted from random forest analyses were tested using backwards elimination on 1000 bootstrapped samples using the package “bootstepAIC” in R (Rizopoulos & Rizopoulos, 2009). A regression model with post-treatment BDI-II as the dependent variable and the selected variables with their interaction with treatment as independent variables was fitted and tested. Predictors and moderators that were included in at least 60% of the bootstrapped samples were included in the final model (Austin & Tu, 2004).

After computing the treatment-modality interaction model, two separate prognostic models using elastic net regression were developed for CBT and IPT (with R-package glmnet; Friedman et al., 2010; Zou & Hastie, 2005). In the elastic net regression, the following parameters were estimated in subsequent order: (1) alpha indicating the balance between Lasso ( $\alpha = 1$ ) and Ridge regression ( $\alpha = 0$ ), (2) lambda, reflecting the amount of shrinkage for the chosen alpha. The optimal alpha was chosen by estimating the cross-validated error for each potential .05 alpha between 0 and 1 (i.e., 0, .05, .10 etc.). For each possible alpha 10-fold cross-validation was conducted 25 times (i.e., 25 iterations) and the cross-validated error was computed. The alpha with the minimal mean cross-validated error (mcve) was chosen. Subsequently, the lambda's with the lowest mean cross-validated error was computed using 10-fold cross validation with the optimal alpha included in the model and repeating this process a 1000 times (i.e., 1000 iterations).

For both treatment-modality interaction models and the prognostic models, predictions were computed using a 5-fold cross-validation (in sample predictions). For each fold, BDI-II post-treatment was predicted using the weights of the individuals from the remaining four folds. In addition, BDI-II post-treatment scores were predicted for individuals in the other dataset (out-of-sample predictions). Model performance was compared using the R-squared (i.e., the explained variance corrected for the number of included pre-treatment variables, the higher the better) and the root mean squared error (RMSE, i.e., the root of the sum of the squared residuals, which are the observed values minus the

Table I. Sample description for CBT and IPT per trial dataset.

	STEPd		FreqMech	
	CBT ( <i>n</i> = 76)	IPT ( <i>n</i> = 75)	CBT ( <i>n</i> = 98)	IPT ( <i>n</i> = 102)
<b>Demographics</b>				
Female, <i>n</i> (%)	54 (71.1)	46 (61.3)	62 (63.3)	61 (59.8)
Age, <i>M</i> ( <i>SD</i> )	41.2 (12.4)	41.31 (11.8)	38.18 (12.4)	37.53 (12.1)
<b>Highest completed education</b>				
16 (21.1)	13 (17.3)	9 (9.2)	12 (11.8)	13 (17.3)
48 (63.2)	41 (54.7)	50 (51)	52 (51)	41 (54.7)
12 (15.8)	21 (28.0)	39 (39.8)	38 (37.3)	21 (28.0)
Partner, yes, <i>n</i> (%)	43 (56.6)	51 (68.0)	34 (34.7)	38 (37.3)
Current job, yes, <i>n</i> (%)	43 (56.6)	47 (62.7)	69 (50.7)	67 (49.3)
Born in the Netherlands, yes, <i>n</i> (%)	73 (96.1)	6 (90.8)	77 (78.6)	82 (80.4)
<b>Depression</b>				
BDI-II baseline, <i>M</i> ( <i>SD</i> )	28.4 (9)	31.2 (8.9)	35.83 (9)	33.61 (10.7)
BDI-II post-treatment, <i>M</i> ( <i>SD</i> )	13.8 (10.7)	16.0 (13.4)	22.75 (14)	21.36 (15.4)
Recurrent depression, <i>n</i> (%)	38 (50)	36 (48)	35 (35.7)	39 (17.5)
<b>Clinical variables</b>				
Number of comorbid Axis I disorders, <i>M</i> ( <i>SD</i> )	.6 (.8)	.7 (.7)	1.01 (1.1)	1.07 (1.3)
Treatment expectation, <i>M</i> ( <i>SD</i> ) (0 = not successful —10 = very successful)	6.8 (1.2)	6.5 (1.3)	6.59 (1.5)	6.63 (1.6)
RAND-36, physical functioning, <i>M</i> ( <i>SD</i> )	74 (23)	74.1 (20.5)	70.51 (26.9)	70.63 (25.6)
RAND-36, social functioning, <i>M</i> ( <i>SD</i> )	42.3 (20.2)	40.8 (20.2)	30.99 (21.3)	35.90 (22.7)
RAND-36, role limitations (physical problems), <i>M</i> ( <i>SD</i> )	37.8 (41.3)	33.7 (38.4)	33.41 (40.6)	37.50 (39.3)
RAND-36, role limitations (emotional problems), <i>M</i> ( <i>SD</i> )	18.8 (33.2)	13.3 (25.1)	9.86 (20.4)	10.78 (21.5)
RAND-36, general health experience, <i>M</i> ( <i>SD</i> )	46.7 (16.5)	43.8 (14.1)	43.41 (19.4)	43.33 (17.5)
RAND-36, perceived health change during past year, <i>M</i> ( <i>SD</i> )	32.2 (27.0)	26 (24.8)	33.92 (25.7)	31.61 (23.8)

Note. BDI-II = Beck Depression Inventory II; *M* = mean; *SD* = standard deviation; RAND-36 = Rand 36 Health Survey. None of the shown pre-treatment variables are transformed. In the STEPd study there was 1 missing on current job (in CBT), 17 missings in the BDI-II post-treatment scores (7 in CBT; 10 in IPT), 5 missing on recurrent depression (3 in CBT; 2 in IPT), 1 missing on the RAND-36 (in CBT). In the FreqMech dataset there was 1 missing on treatment expectation (in CBT), 55 missing on the BDI-II post-treatment scores (26 in CBT; 29 in IPT), 29 missing on the number of previous episodes (15 in CBT; 14 in IPT), 13 missing on the number of comorbid Axis I disorders (6 in CBT; 7 in IPT).

model predictions, the lower the better). All treatment-modality interaction models had superior or comparable model performance results as compared to the prognostic models. It was therefore decided to use the treatment-modality interaction models to compute all PAI scores. In the remainder of this paper, only the results of the treatment-modality interaction models are reported. Comparisons between treatment-modality interaction models and prognostic models can be found in Data Supplement 4.

**Computing the Personalized Advantage Index Scores.** For each dataset separately, PAI scores were computed by combining all final predictors and moderators in a multivariable regression model with post-treatment BDI-II as the outcome using 5-fold cross validation. For each fold, BDI-II post-treatment was predicted as if this individual would have received CBT (or IPT) using the weights of the individuals from the remaining four folds. For each individual, PAI scores were calculated as the difference between his or her predicted outcomes in CBT and IPT. Differences in observed BDI-II post-treatment scores between patients that received versus not received their PAI-indicated treatment were tested using a two-sample t-test for the total sample, and for CBT and IPT separately. In addition, these differences were evaluated when controlling for BDI-II baseline differences between the indicated and non-indicated groups using ANCOVA analyses (Data Supplement 5). Following DeRubeis et al. (2014), we also compared the observed post-treatment BDI-II scores of individuals with the highest 60% PAI-scores (absolute values). The effect sizes (Cohen’s *d*) for the difference in post-treatment scores between patients that were randomized versus not randomized to their optimal treatment were reported.

**Cross-Trial Prediction.** First, predictors and moderators from the PAI models of each independent dataset were compared to determine whether these predictors and moderators were also present in the other dataset. Second, PAI scores were generated in the other dataset using the PAI model and predictors and moderators’ weights of the original dataset. To facilitate this, similar transformations of predictors and moderators in the original datasets were applied to the other variables of the other dataset. Differences in observed BDI-II post-treatment scores between patients that received versus not received their optimal treatment were presented for both datasets.

**Sensitivity Analysis.** The aim of the present paper was to investigate how independent PAI models would translate to other study samples. Another way to investigate the potential of cross-trial prediction is to start with only overlapping variables from each dataset, develop the PAI models and investigate whether these models can be externally validated. This approach would avoid the problem of differences in measurements between studies and provide insight on how PAI models would externally validate if, in a hypothetical situation, they would have been developed and tested in mental health care centres with similar data collection procedures. To test this, we conducted a sensitivity analyses were we only included variables that were overlapping between the datasets. An overview of the overlapping variables between the datasets and the applied transformations can be found in Data Supplement 2.

## Results

### Variable Description and Missing Data Imputation

Table I shows the sample description of the STEPd and FreqMech trial. Variables with missing data in the STEPd trial had missing values between 0.007 and 0.03% of the cases. A total of 17 individuals had a missing value on the post-treatment BDI-II (11.3%). Variables with missing data in the FreqMech dataset had between .5 and 32.5% missing values. A total of 55 values were missing for the post-treatment BDI-II scores (27.5%). In both datasets, imputation appeared to be accurate when applied to the complete (non-missing) data with artificially produced missing data with estimated NRMSE's of 0.008 (STEPd) and 0.20 (FreqMech), and estimated PFC's of 0.10 and 0.32 (as compared to previous study examples; Stekhoven & Bühlmann, 2011; Waljee et al., 2013).

### The Personalized Advantage Index in STEPd

**Building the Personalized Advantage Index model.** Nine variables were selected using the mobforest algorithm. Of these variables, two predictors and five moderators were selected in at least 60% of the bootstrap samples using the backwards elimination technique. Having a job and less anxiety symptoms predicted lower post-treatment BDI-II scores irrespective of the type of treatment received. Paranoid symptoms, pre-treatment depression severity, and the number of life events in the past year was related to higher BDI-II

scores post-treatment in IPT compared to CBT. Cognitive problems and lower overall social, occupational, and psychological functioning were related to higher BDI-II scores post-treatment in CBT compared to IPT.

**Computing the Personalized Advantage Index scores.** A total of 67 patients (44.4%) received their PAI-indicated treatment (22.5% CBT and 21.9% IPT). Post-treatment BDI-II scores were significantly lower for individuals that received their PAI-indicated treatment ( $M = 11.64$ ,  $SD = 10.11$ ) compared to those that received their PAI non-indicated treatment ( $M = 18.17$ ,  $SD = 12.27$ ,  $t(149) = 3.51$ ,  $p < 0.001$ , effect size Cohen's  $d = .57$ ). As shown in Figure 1a, for individuals who had a PAI indicating CBT, post-treatment BDI-II scores were significantly lower for individuals that were actually randomized to CBT ( $M = 11.19$ ,  $SD = 9.79$ ) compared to those randomized to IPT ( $M = 19.85$ ,  $SD = 13.69$ ,  $t(74) = -3.10$ ,  $p = 0.003$ , effect size Cohen's  $d = .71$ ). Controlling for BDI-II baseline did not change the results of these comparisons (Data Supplement 5). For individuals who had a PAI indicating IPT, post-treatment BDI-II was non-significantly different for those receiving IPT ( $M = 12.09$ ,  $SD = 10.56$ ) and CBT ( $M = 16.48$ ,  $SD = 10.55$ ,  $t(73) = 1.79$ ,  $p = 0.08$ , effect size Cohen's  $d = .18$ ). After controlling for BDI-II baseline this comparison became significant ( $p = .01$ ; Data Supplement 5). Among individuals with the highest 60% PAI scores, mean post-treatment BDI-II scores differed significantly between these individuals that received the PAI-indicated treatment versus those that received the PAI non-indicated treatment (indicated treatment:  $M = 10.43$ , non-indicated treatment:  $M = 18.5$ ,  $t(89) = 3.35$ ,  $p = 0.001$ ), with a Cohen's  $d$  effect size estimate of 0.71.

### The Personalized Advantage Index in FreqMech

**Building the Personalized Advantage Index model.** A total of 15 variables were selected using the mobforest algorithm. Of these 15 variables, eight predictors and one moderator were selected in at least 60% of the bootstrap samples using the backwards elimination technique. A higher baseline depression, having received more previous treatments, higher levels of dysfunctional thinking, fewer physical problems, worse physical functioning, lower quality of life, worse emotional problems and less vitality were related to higher scores on the BDI-II post-treatment, irrespective of the received treatment. In addition, being female was related to



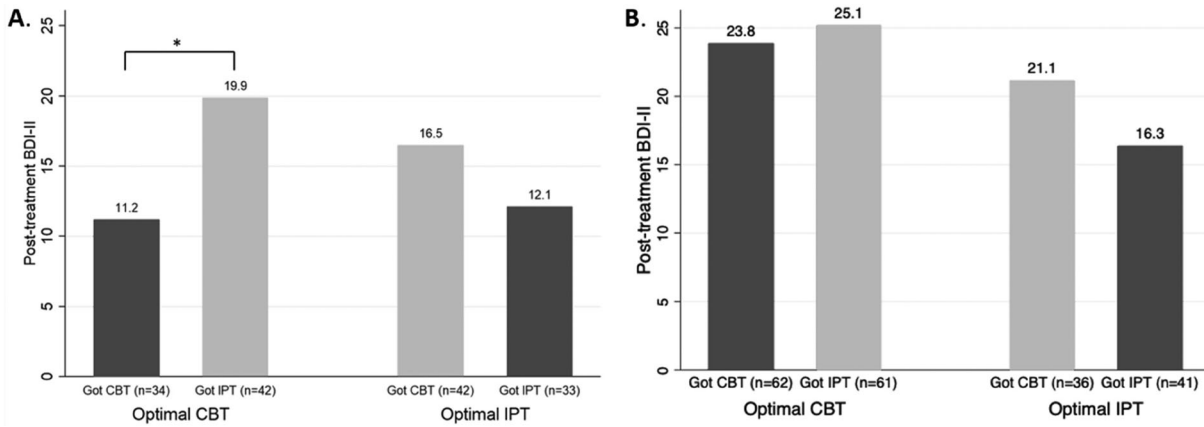


Figure 1. Comparisons of observed BDI-II post-treatment scores for patients randomly assigned to their PAI-indicated treatment versus those assigned to their PAI non-indicated treatment, by psychotherapy type in the STEPd study (A) and the FreqMech study (B). *Note.* CBT = cognitive behavioural therapy; IPT = interpersonal psychotherapy; BDI-II = Beck Depression Inventory, second edition. \* indicates a significant difference between observed BDI-II scores.

higher BDI-II scores post-treatment in IPT compared to CBT.

**Computing the Personalized Advantage Index scores.** A total of 103 patients (51.5%) received their PAI-indicated treatment (31% CBT and 20.5% IPT). Post-treatment BDI-II scores were lower for individuals who received their PAI-indicated treatment ( $M = 20.87$ ,  $SD = 13.23$ ) compared to those that received their PAI non-indicated treatment, but this result was not significant ( $M = 23.68$ ,  $SD = 13.65$ ,  $t(198) = 1.47$ ,  $p = .14$ , effect size Cohen's  $d = .20$ ). As shown in Figure 1b, for individuals who had a PAI indicating CBT, post-treatment BDI-II scores were lower for individuals that were actually randomized to CBT ( $M = 23.86$ ,  $SD = 13.86$ ) compared to those randomized to IPT ( $M = 25.17$ ,  $SD = 14.71$ ), but this difference was not significant ( $t(121) = -.50$ ,  $p = .61$ , effect size Cohen's  $d = .09$ ). For individuals who had a PAI indicating IPT, the post-treatment BDI-II was lower but not significantly different for those receiving IPT ( $M = 16.35$ ,  $SD = 10.89$ ) and CBT ( $M = 21.15$ ,  $SD = 11.40$ ,  $t(75) = 1.88$ ,  $p = .06$ , effect size Cohen's  $d = .43$ ). Controlling for BDI-II baseline did not change the results of these comparisons (Data Supplement 5). Among those with the highest 60% PAI scores, mean post-treatment BDI-II scores were lower between these individuals that received the PAI-indicated treatment versus those that received the PAI non-indicated treatment (indicated treatment = 19.92 ( $SD = 12.72$ ), non-indicated treatment = 22.19 ( $SD = 12.99$ ),  $t(118) = .96$ ,  $p = .33$ ), but this difference was not significant (effect size Cohen's  $d = .17$ ).

A more detailed description of the selection process of the pre-treatment variables in both datasets can be found in Data Supplement 3.

### Cross-trial Prediction

**STEPd to FreqMech.** Part of the STEPd PAI model could be tested in the FreqMech dataset, since we could not use all predictors and moderator because of differences in data collection. The following predictors and moderators were available in the FreqMech dataset: employment status (predictor), anxiety symptoms (Remission of Depression Questionnaire subscale, predictor), overall social, occupational, and psychological functioning (Global Assessment of Functioning DSM-IV, moderator), and pre-treatment depression severity (BDI-II, moderator). Results of the STEPd PAI model in the FreqMech dataset indicated that post-treatment BDI-II scores were non-significantly lower for individuals that received their PAI-indicated treatment ( $n = 112$ ) ( $M = 21.28$ ,  $SD = 13.07$ ) compared to those that received their PAI non-indicated treatment ( $n = 88$ ;  $M = 23.38$ ,  $SD = 13.60$ ,  $t(198) = 1.11$ ,  $p = .27$ , effect size Cohen's  $d = .16$ ). For those that had a PAI indicating CBT, post-treatment BDI-II scores were non-significantly lower for individuals that received CBT ( $M = 25.07$ ,  $SD = 13.55$ ), as compared to individuals that received IPT ( $M = 26.59$ ,  $SD = 14.59$ ,  $t(112) = -0.58$ ,  $p = .57$ ). For individuals who had a PAI indicating IPT, post-treatment BDI-II scores were non-significantly lower for those randomized to IPT ( $M = 16.58$ ,  $SD = 10.85$ ) compared to those randomized to CBT ( $M = 18.75$ ,  $SD = 10.60$ ,  $t(84) = 0.92$ ,  $p = .36$ ). Controlling for BDI-II baseline did not change the results of these comparisons (Data Supplement 5).

**FreqMech to STEPd.** It was possible to test a part of the FreqMech PAI model in the STEPd dataset, however because of differences in data collection we could not use all predictors and moderators. The following predictors and moderators were available in the STEPd dataset: baseline depression (BDI-II, predictor), physical problems (RAND-36, predictor), emotional problems (RAND-36, predictor), physical functioning (RAND-36, predictor), vitality (RAND-36, predictor), gender (moderator) and quality of life utility score (EQ5D, moderator). Using the FreqMech PAI model, in the STEPd dataset post-treatment BDI-II scores were lower for individuals that received their PAI-indicated treatment ( $n = 83$ ) ( $M = 13.81$ ,  $SD = 11.26$ ) compared to those that received their PAI non-indicated treatment ( $n = 68$ ;  $M = 17.06$ ,  $SD = 12.55$ ,  $t(149) = 1.67$ ,  $p = .09$ , effect size Cohen's  $d = .27$ ), however this difference was not significant. For those who had a PAI indicating CBT, post-treatment BDI-II scores were non-significantly lower for individuals that were randomized to CBT ( $M = 14.36$ ,  $SD = 11.06$ ), as compared to individuals randomized to IPT ( $M = 18.81$ ,  $SD = 13.56$ ,  $t(98) = -1.80$ ,  $p = .07$ ). For individuals who had a PAI indicating IPT, there were no significant differences between those receiving IPT ( $M = 12.79$ ,  $SD = 11.75$ ) and for those receiving CBT ( $M = 13.39$ ,  $SD = 9.36$ ,  $t(49) = .19$ ,  $p = .84$ ). Controlling for BDI-II baseline did not change the results of these comparisons (Data Supplement 5).

### Sensitivity Analyses

We conducted a sensitivity analyses were we included only the 20 variables that were overlapping between the two datasets (see Data Supplement 2).

#### STEPd

**PAI model in the STEPd dataset.** Seven variables were selected using the mobforest algorithm. Of these variables, four predictors and one moderator were selected in at least 60% of the bootstrap samples using the backwards elimination technique. Lower BDI-II baseline scores, fewer social problems, fewer anxiety symptoms and having a job predicted lower BDI-II post-treatment scores irrespective of the received type of treatment. Childhood trauma was related to higher post-treatment BDI-II scores in IPT compared to CBT. There were no significant differences between individuals who received their PAI-indicated treatment and those that received their PAI non-indicated treatment ( $M = 14.92$ ,  $SD = 12.06$  versus  $M = 15.64$ ,  $SD = 11.87$ ,  $t(149) = .37$ ,  $p = .71$ , effect size Cohen's  $d = .06$ ).

Controlling for BDI-II baseline did not change the results of this comparison (Data Supplement 5).

**Cross trial prediction to the FreqMech dataset.** Using the STEPd PAI model in the FreqMech dataset, post-treatment BDI-II scores were non-significantly lower for individuals that received their PAI-indicated treatment ( $n = 83$ ;  $M = 21.69$ ,  $SD = 12.87$ ) compared to those that received their PAI non-indicated treatment ( $n = 117$ ;  $M = 22.88$ ,  $SD = 13.99$ ,  $t(198) = 0.61$ ,  $p = .54$ , effect size Cohen's  $d = .09$ ). Controlling for BDI-II baseline did not change the results of this comparison (Data Supplement 5).

#### FreqMech

**PAI model in FreqMech dataset.** Five variables were selected using the mobforest algorithm. Of these variables, three predictors and one moderator were selected in at least 60% of the bootstrap samples using the backwards elimination technique. Higher BDI-II baseline scores, more physical problems and less physical functioning predicted higher BDI-II scores post-treatment irrespective of the received type of treatment. Being female related to higher BDI-II scores post-treatment in IPT compared to CBT. Post-treatment BDI-II scores were lower for individuals who received their PAI-indicated treatment ( $M = 20.90$ ,  $SD = 13.06$ ) compared to those that received their PAI non-indicated treatment, however this difference was not significant ( $M = 23.96$ ,  $SD = 13.87$ ,  $t(198) = 1.60$ ,  $p = .11$ , effect size Cohen's  $d = .22$ ). After controlling for BDI-II baseline this comparison became significant ( $p = .01$ ; Data Supplement 5).

**Cross trial prediction to the STEPd dataset.** Using the FreqMech PAI model in the STEPd dataset, post-treatment BDI-II scores were non-significantly lower for individuals that received their PAI-indicated treatment ( $n = 83$ ) ( $M = 13.81$ ,  $SD = 11.26$ ) compared to those that received their PAI non-indicated treatment ( $n = 68$ ;  $M = 17.06$ ,  $SD = 12.55$ ,  $t(149) = 1.67$ ,  $p = .09$ , effect size Cohen's  $d = .27$ ). Controlling for BDI-II baseline did not change the results of this comparison (Data Supplement 5).

### Discussion

The overall aim of the current study was to determine the generalizability of the PAI to clinical practice by externally validating PAI models using data from two independent Dutch randomized trials comparing CBT and IPT for MDD (i.e., the STEPd trial and the

FreqMech trial). In the STEPd dataset, two predictors and five moderators were found. The STEPd PAI resulted in a significant difference in observed post-treatment depression severity when comparing individuals assigned to their PAI-indicated treatment versus those assigned to their PAI non-indicated treatment (Cohen's  $d = .57$ ). This mean difference was more pronounced for individuals with the top 60% PAI scores (Cohen's  $d = .71$ ), and for individuals that had a CBT recommendation (Cohen's  $d = .71$ ). In the FreqMech study one moderator and eight predictors were found. Here, the PAI showed a modest effect size (Cohen's  $d = .20$ ) and resulted in modest and non-significant differences in observed post-treatment BDI-II scores between patients who received their indicated versus patients who received their non-indicated treatment. This difference in observed post-treatment BDI-II scores was not more pronounced for the highest 60% PAI scores (Cohen's  $d = .17$ ). In order to externally validate the PAI models, cross-trial predictions from the STEPd model to the FreqMech dataset, and from the FreqMech model to the STEPd dataset were conducted. Cross-prediction from STEPd to FreqMech was limited because three of the five moderators from the STEPd model were not available in the FreqMech dataset. The resulting PAI recommendations of the STEPd model in the FreqMech dataset showed a small effect size (Cohen's  $d = .16$ ) with non-significant differences between individuals that received their indicated versus their non-indicated treatment. For the cross-trial prediction of the FreqMech PAI model to the STEPd dataset, two FreqMech predictors were not available in the STEPd dataset. The resulting treatment recommendations of the FreqMech model in the STEPd dataset had a moderate effect size (Cohen's  $d = .27$ ), and the differences between the individuals who were randomized to their indicated versus those who were randomized to their non-indicated treatment were not statistically significant. In addition, a sensitivity analysis was conducted using only the overlapping 20 variables from each dataset for cross-trial prediction. Treatment recommendations of the STEPd model resulted in poor effect sizes in the STEPd dataset (Cohen's  $d = .06$ ), and in the FreqMech dataset (Cohen's  $d = .09$ ). Results of the FreqMech PAI model indicated small to moderate effect sizes of treatment recommendation in the FreqMech dataset (Cohen's  $d = .22$ ) as well as in the STEPd dataset (Cohen's  $d = .27$ ).

This study demonstrates that the effects of the PAI in the context of cross-trial prediction are rather modest. In the STEPd study, within-study PAI scores seemed very promising with a high effect size, however the effect size of this PAI decreased substantially in the cross-trial prediction. This

decrease of effect size might be indicative for overfitting of the within-study PAI model, which could be explained by double-dipping when applying the variable selection procedure in the complete sample (Fiedler, 2011). In the FreqMech study, within-study PAI scores showed small to moderate effect sizes that were maintained during cross-trial prediction. These results indicate that effect sizes seemed more stable across analyses in the study with a larger sample size ( $n = 200$  vs.  $n = 151$  in the STEPd study) and more heterogeneous data from multiple sites (instead of one treatment centre in the STEPd study). The effect sizes of the FreqMech study are also in line with the out-of-sample predictions of previous PAI study of Delgado and Gonzalez Salas Duhne (2020, Hedges'  $g = .26$ ) that is based on a larger dataset from multiple study sites. Overall, the results of our study emphasize the many challenges of external validation of PAI models, which is a necessary step for implementation in clinical practice.

One major challenge for external validation of PAI models is that the concept of external validation is not as straightforward as one might expect. Validation studies may include temporal validation (same location, but later in time), geographical validation (different locations), validation in different settings (e.g., primary and secondary settings) and validation in different domains (e.g., in adults and children; Debray et al., 2015). One way to understand these different types of validation is by distinguishing between model reproducibility (accuracy across samples from the same target population) and model transportability (accuracy across samples from different but related populations; i.e., external validation; Debray et al., 2015; Justice et al., 1999). However, the optimal balance between different and related data is unclear. In the context of our paper, we could debate that the STEPd and FreqMech study samples were too different and therefore lead to different moderators and predictors, despite the fact that these studies were both from the Netherlands and conducted by some members of the same research group. Differences between the STEPd and the FreqMech study involved the single-centre vs. multi-centre design, a low vs. high number of therapists, medium vs. high pre- and post-treatment depression severity, low vs. medium dropout rates, differences in in- and exclusion criteria, and differences in quality of the performed treatments. In addition to these differences, when we consider the population with a diagnosis of MDD, this categorization alone already involves a high level of heterogeneity. Possibly, depressed individuals within one study already belong to multiple distinctive populations (Nandi et al., 2009).

Another major obstacle for external validation of PAI models are the different types of data collection and statistical methods used in the different studies. The problem of dissimilar data collection is illustrated by our own findings, since we could only partially validate the STEPd and the FreqMech model in the other dataset. Moreover, restricting the number of variables to the degree of overlap between the datasets (sensitivity analyses) negatively affected the PAI recommendations of the STEPd model. Development of generalizable PAI models requires unambiguous definitions of outcome, potential predictors, and potential moderators. In addition, these variables need to be based on reproducible measurements that are potentially feasible for application in clinical practice. Since the development of PAI models is typically a secondary analysis of data of effectiveness trials, thorough pre-trial evaluation on what variables to include for a reproducible and generalizable PAI model is needed. In addition to the data collection problem, there are also statistical difficulties. For example, one possible reason why we did not find significant cross-trial prediction effects might be because of the (relatively) small sample sizes of our studies, in particular the STEPd study (Luedtke et al., 2019). Besides the problem of small sample sizes, there is also a striking heterogeneity in statistical methods used for missing data imputation and model building (Cohen et al., 2019; Cohen & DeRubeis, 2018). This is of importance since there is evidence that different methods of data imputation and model building may lead to differences in clinical conclusions (Cohen et al., 2019; Stavseth et al., 2019; Webb et al., 2020). However, comparisons between the STEPd PAI model built in 2015 (Huibers et al., 2015) and the PAI model in the present paper that used different methods on the same dataset, indicate significant overlap between selected pre-treatment variables and similar effect sizes. A better understanding of how differences in statistical methods affect predictive accuracy is warranted when comparing and externally validating different PAI models.

The final major challenge for external validation of PAI models is the generalizability of the contrasts between treatments. One possible explanation for finding different moderators in the STEPd versus the FreqMech trial is that the contrasts between CBT and IPT were different for each study. Despite the fact that therapists were instructed to follow the same CBT and IPT protocols in both studies, the overall quality of therapy was considerably lower in the FreqMech trial compared to the STEPd trial. Treatments in the FreqMech study were delivered by a high number of therapists from multiple treatment centres of different mental health organizations

located at different parts of the Netherlands. In the STEPd study this was a small number of experienced therapists all working in the same clinical setting. It is possible that these differences have led to differences in the performance of the CBT and IPT protocols and that the better and more consistent therapy quality in the STEPd dataset is responsible for finding a stronger PAI model within the STEPd dataset compared to the FreqMech dataset. Even holding quality and adherence constant, between-therapist differences as well as differences in locations or organizations may make external validation of prescriptive models much more complex than external validation of prognostic models. The results of various PAI studies so far seem to support these hypotheses. The finding that the effects of the FreqMech PAI were modest and non-significant contrasts earlier findings (DeRubeis et al., 2014; Huibers et al., 2015; Webb et al., 2019; Zilcha-Mano et al., 2016), but is in line with some, most notably newer studies (Cohen et al., 2019; Delgadoillo & Gonzalez Salas Duhne, 2020; Eskildsen et al., 2020; Friedl et al., 2020; Lopez-Gomez et al., 2019). Like the FreqMech dataset, the studies that had modest and non-significant findings included a high number of treatment centres and/or therapists. Therefore, one explanation for these modest and non-significant findings is that large heterogeneity in the sample and treatment might make it difficult to find large PAIs compared to studies with less variation. However, this explanation contrasts with findings of Webb et al. (2019), who included multiple centres and therapists as well, and found a PAI with a large effect size. A possible explanation for this could be that Webb et al. (2019) focused on treatment with antidepressants instead of psychotherapy, in which heterogeneity in therapists may play a smaller role.

Although the many challenges in the external validation of the PAI in psychotherapy are not easy to solve, they need to be addressed in order to make findings in precision medicine relevant for clinical practice. How do we make progress from here? Previous authors have proposed to start with building prescriptive algorithms on large (combined) observational datasets followed by pragmatic trials that randomize clinicians to receive or not receive information from these algorithms (Kessler, 2018; Luedtke et al., 2019). We think that this approach is promising, especially for less complex prognostic models. For the more fine-grained prescriptive models, such as the PAI models, we suggest a more “zoomed in” or local method that occurs parallel to the approach that Kessler and colleagues propose. Since the heterogeneity of study populations and treatments is very high, we first need to establish to what *extent* external validation is realistic; maybe

some treatment selection decisions are local decisions. With clear agreements on data collection and data analyses, we could first start with the development of local prescriptive models. In this context “local” means within the same geographical area, mental health care organization, and with a low number of therapists that have joint supervision. From there, we could focus on temporal validation (same location, different time), and then work towards external validation by expanding the location in terms of other geographical locations, other mental health care organizations and by increasing the number of therapists. Model updating should be a central part of these “zooming out” steps, enhancing the performance of models in other samples using methods such as recalibration (adjustment of intercepts and the regression coefficients using the calibration intercept and calibration slope) or model revision (re-estimation of the intercept and the regression coefficients using the combined datasets; Janssen et al., 2008). In addition, Bayesian inference, that inherently updates probabilities with new data, could be integrated in these models, especially in the context of small study samples (Depaoli et al., 2017). Besides externally validating PAI models, simulation studies are needed to find out how different choices in statistical methods and parameters (for example: differences in imputing missing data, differences between different forms (and combinations) of machine learning) can lead to different (clinical) conclusions.

### In Conclusion

The present study investigated the generalizability of PAI models, by building PAI models using data from two independent Dutch randomized trials and externally validated each model in the other sample. In one dataset (STEPd), post-treatment BDI-II scores were significantly lower for individuals that received their PAI-indicated treatment compared to those that received their PAI non-indicated treatment, however treatment recommendations based on cross-trial predictions had low effect sizes. In the other dataset (FreqMech), small to moderate effect sizes and non-significant post-treatment differences were found and these outcomes could be maintained during cross-trial prediction in the STEPd dataset. Before implementing the use of PAI models into clinical practice, studies that address the external validation of the PAI are highly necessary.

### Acknowledgements

We would like to acknowledge the contribution of participants and therapists of the STEPd and the

FreqMech study. Furthermore, we thank Annie Raven, Annie Hendriks, Danielle Tilburgs, Nicole Billings, Kris Wijma and Sofie Jansen for their assistance during the two studies.

### Funding

This work was supported by ZonMw, the Netherlands: [837002401]; Stichting tot Steun of the Vereniging voor Christelijke Verzorging van Geestes- en Zenuwzieken, the Netherlands.: [Grant Number NA]; research institute of Experimental Psychopathology (EPP), the Netherlands: [Grant Number NA]; Academic Community Mental Health Centre (RIAGG, now METGGZ Maastricht) in Maastricht, the Netherlands: [Grant Number NA].

### Disclosure of Interest

The authors report no conflict of interest.

### Author contributions

SvB and SB made the analysis plan and conducted the analysis. SvB, SB and LL were involved in the interpretation of the analyses. SvB and SB wrote the manuscript. All authors read, contributed to and approved the final manuscript.

### Note

<sup>1</sup> Questions regarding treatment competence had different Likert scales in the two studies. In STEPd, good to excellent competence was defined as a score of 4 or more on a 6 point Likert scale (for CBT) and as a score of 3 or more on a 5 point Likert scale (for IPT). In the FreqMech study, good to excellent competence was defined as a score of 5 or more on a 7 point Likert scale.

### ORCID

Suzanne C. Van Bronswijk  <http://orcid.org/0000-0002-2983-1268>

### References

- American Psychiatric Association. (2009). Practice guideline for the treatment of patients with major depressive disorder (3rd.). <http://psychiatryonline.org/guidelines.aspx>.
- Austin, P. C., & Tu, J. V. (2004). Bootstrap methods for developing predictive models. *The American Statistician*, 58(2), 131–137. <https://doi.org/10.1198/0003130043277>
- Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression*. Guilford Press.

- Beck, A. T., Steer, R., & Brown, G. K. (1996). *Beck depression inventory II: Manual*. Harcourt Brace.
- Bleeker, S., Moll, H., Steyerberg, E., Donders, A., Derksen-Lubsen, G., Grobbee, D., & Moons, K. (2003). External validation is necessary in prediction research: A clinical example. *Journal of Clinical Epidemiology*, 56(9), 826–832. [https://doi.org/10.1016/S0895-4356\(03\)00207-5](https://doi.org/10.1016/S0895-4356(03)00207-5)
- Brujniks, S. J., Bosmans, J., Peeters, F. P., Hollon, S. D., van Oppen, P., van den Boogaard, M., Dingemans, P., Cuijpers, P., Arntz, A., Franx, G., & Huibers, M. J. H. (2015). Frequency and change mechanisms of psychotherapy among depressed patients: Study protocol for a multicenter randomized trial comparing twice-weekly versus once-weekly sessions of CBT and IPT. *BMC Psychiatry*, 15(1), 137. <https://doi.org/10.1186/s12888-015-0532-8>
- Brujniks, S., Lemmens, L., Hollon, S., Peeters, F., Cuijpers, P., Arntz, A., & Huibers, M. (2020). Seeing depressed patients twice weekly improves outcomes: Results from an RCT comparing cognitive behavior therapy and interpersonal psychotherapy. *British Journal of Psychiatry*, 216(4), 222–230. <https://doi.org/10.1192/bjp.2019.265>
- Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment selection in depression. *Annual Review of Clinical Psychology*, 14(1), 209–236. <https://doi.org/10.1146/annurev-clinpsy-050817-084746>
- Cohen, Z. D., Kim, T. T., Van, H. L., Dekker, J. J., & Driessen, E. (2019). A demonstration of a multi-method variable selection approach for treatment selection: Recommending cognitive-behavioral versus psychodynamic therapy for mild to moderate adult depression. *Psychotherapy Research*, 1–14. <https://doi.org/10.1080/10503307.2018.1563312>
- Debray, T. P., Vergouwe, Y., Koffijberg, H., Nieboer, D., Steyerberg, E. W., & Moons, K. G. (2015). A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of Clinical Epidemiology*, 68(3), 279–289. <https://doi.org/10.1016/j.jclinepi.2014.06.018>
- Deisenhofer, A. K., Delgadillo, J., Rubel, J. A., Böhne, J. R., Zimmermann, D., Schwartz, B., & Lutz, W. (2018). Individual treatment selection for patients with posttraumatic stress disorder. *Depression and Anxiety*, 35(6), 541–550. <https://doi.org/10.1002/da.22755>
- Delgadillo, J., & Gonzalez Salas Duhne, P. (2020). Targeted prescription of cognitive behavioral therapy vs. person-centered counseling for depression using a machine learning approach. *Journal of Consulting and Clinical Psychology*, 88(1), 14–24. <https://doi.org/10.1037/ccp0000476>
- Depaoli, S., Rus, H. M., Clifton, J. P., van de Schoot, R., & Tiemensma, J. (2017). An introduction to Bayesian statistics in health psychology. *Health Psychology Review*, 11(3), 248–264. <https://doi.org/10.1080/17437199.2017.1343676>
- DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014). The Personalized Advantage Index: Translating research on prediction into individualized treatment recommendations. A demonstration. *PloS one*, 9(1), e83875. <https://doi.org/10.1371/journal.pone.0083875>
- Dobson, K. S., Shaw, B. F., & Vallis, T. M. (1985). Reliability of a measure of the quality of cognitive therapy. *British Journal of Clinical Psychology*, 24(4), 295–300. <https://doi.org/10.1111/j.2044-8260.1985.tb00662.x>
- Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382), 316–331. <https://doi.org/10.1080/01621459.1983.10477973>
- Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: The 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438), 548–560.
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap (Monograph on statistics and applied probability)*. Chapman & Hall View Article.
- Eskildsen, A., Reinholdt, N., van Bronswijk, S., Brund, R. B., Christensen, A. B., Hvenegaard, M., ... Rosenberg, N. K. (2020). Personalized psychotherapy for outpatients with major depression and anxiety disorders: Transdiagnostic versus diagnosis-specific group cognitive behavioural therapy.
- Fiedler, K. (2011). Voodoo correlations are everywhere—not only in neuroscience. *Perspectives on Psychological Science*, 6(2), 163–171. <https://doi.org/10.1177/1745691611400237>
- First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1995). *Structured clinical interview for DSM-IV axis I disorders (SCID-I)*. Biometrics Research Department New York State Psychiatric Institute.
- Friedl, N., Berger, T., Krieger, T., Caspar, F., & Grosse Holtforth, M. (2020). Using the Personalized Advantage Index for individual treatment allocation to cognitive behavioral therapy (CBT) or a CBT with integrated exposure and emotion-focused elements (CBT-EE). *Psychotherapy Research*, 30(6), 763–775. <https://doi.org/10.1080/10503307.2019.1664782>
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1. <https://doi.org/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(1), 125. <https://doi.org/10.1186/1471-2105-14-125>
- Gillan, C. M., & Whelan, R. (2017). What big data can do for treatment in psychiatry. *Current Opinion in Behavioral Sciences*, 18, 34–42. <https://doi.org/10.1016/j.cobeha.2017.07.003>
- Hollon, S., Evans, M., Auerbach, A., DeRubeis, R., Elkin, I., Lowery, A., ... Piasecki, J. (1988). Development of a system for rating therapies for depression: Differentiating cognitive therapy, interpersonal psychotherapy, and clinical management pharmacotherapy. *Unpublished Manuscript*.
- 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>
- Janssen, K., Moons, K., Kalkman, C., Grobbee, D., & Vergouwe, Y. (2008). Updating methods improved the performance of a clinical prediction model in new patients. *Journal of Clinical Epidemiology*, 61(1), 76–86. <https://doi.org/10.1016/j.jclinepi.2007.04.018>
- Justice, A. C., Covinsky, K. E., & Berlin, J. A. (1999). Assessing the generalizability of prognostic information. *Annals of Internal Medicine*, 130(6), 515–524. <https://doi.org/10.7326/0003-4819-130-6-199903160-00016>
- Keefe, J. R., Wiltsey Stirman, S., Cohen, Z. D., DeRubeis, R. J., Smith, B. N., & Resick, P. A. (2018). In rape trauma PTSD, patient characteristics indicate which trauma-focused treatment they are most likely to complete. *Depression and Anxiety*, 35(4), 330–338. <https://doi.org/10.1002/da.22731>
- Kessler, R. C. (2018). The potential of predictive analytics to provide clinical decision support in depression treatment planning. *Current Opinion in Psychiatry*, 31(1), 32–39. <https://doi.org/10.1097/YCO.0000000000000377>
- Klerman, G. L., Weissman, M. M., Rounsaville, B. J., & Chevron, E. S. (1984). *Interpersonal psychotherapy for depression*. Basis Books.
- Kraemer, H. C., & Blasey, C. M. (2004). Centring in regression analyses: A strategy to prevent errors in statistical inference. *International Journal of Methods in Psychiatric Research*, 13(3), 141–151. <https://doi.org/10.1002/mpr.170>

- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). Springer.
- Lemmens, L., Arntz, A., Peeters, F., Hollon, S., Roefs, A., & Huibers, M. (2015). Clinical effectiveness of cognitive therapy v. interpersonal psychotherapy for depression: Results of a randomized controlled trial. *Psychological Medicine*, *45*(10), 2095–2110. <https://doi.org/10.1017/S0033291715000033>
- Lemmens, L. H., Arntz, A., Peeters, F. P., Hollon, S. D., Roefs, A., & Huibers, M. J. (2011). Effectiveness, relapse prevention and mechanisms of change of cognitive therapy vs. interpersonal therapy for depression: Study protocol for a randomised controlled trial. *Trials*, *12*(1), 150. <https://doi.org/10.1186/1745-6215-12-150>
- Lemmens, L., Van Bronswijk, S., Peeters, F., Arntz, A., Hollon, S., & Huibers, M. (2019). Long-term outcomes of acute treatment with cognitive therapy v. interpersonal psychotherapy for adult depression: Follow-up of a randomized controlled trial. *Psychological Medicine*, *49*(3), 465–473. <https://doi.org/10.1017/S0033291718001083>
- Lopez-Gomez, I., Lorenzo-Luaces, L., Chaves, C., Hervás, G., DeRubeis, R. J., & Vázquez, C. (2019). Predicting optimal interventions for clinical depression: Moderators of outcomes in a positive psychological intervention vs. Cognitive-behavioral therapy. *General Hospital Psychiatry*, *61*, 104–110. <https://doi.org/10.1016/j.genhosppsy.2019.07.004>
- Lorenzo-Luaces, L., DeRubeis, R. J., & Bennett, I. M. (2015). Primary care physicians' selection of low-intensity treatments for patients with depression. *Family Medicine*, *47*(7), 511–516.
- Luedtke, A., Sadikova, E., & Kessler, R. C. (2019). Sample size requirements for multivariate models to predict between-patient differences in best treatments of major depressive disorder. *Clinical Psychological Science*, *7*(3), 445–461. <https://doi.org/10.1177/2167702618815466>
- Moons, K. G., Donders, R. A., Stijnen, T., & Harrell, J. F. E. (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*, *59*(10), 1092–1101. <https://doi.org/10.1016/j.jclinepi.2006.01.009>
- Moons, K. G., Royston, P., Vergouwe, Y., Grobbee, D. E., & Altman, D. G. (2009). Prognosis and prognostic research: What, why, and how? *BMJ*, *338*(feb23 1), b375. <https://doi.org/10.1136/bmj.b375>
- Nandi, A., Beard, J. R., & Galea, S. (2009). Epidemiologic heterogeneity of common mood and anxiety disorders over the life-course in the general population: A systematic review. *BMC Psychiatry*, *9*(1), 31. <https://doi.org/10.1186/1471-244X-9-31>
- National Collaborating Centre for Mental Health. (2010). *Depression: the treatment and management of depression in adults (updated edition)*.
- National Health Service. (2018). *Psychological therapies, annual report on the use of IAPT services – England, 2017–18*. <https://digital.nhs.uk/data-and-information/publications/statistical/psychological-therapies-annual-reports-on-the-use-of-iapt-services/annual-report-2017-18>
- Rizopoulos, D., & Rizopoulos, M. D. (2009). Package 'bootStepAIC'.
- Siontis, G. C., Tzoulaki, I., Castaldi, P. J., & Ioannidis, J. P. (2015). External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *Journal of Clinical Epidemiology*, *68*(1), 25–34. <https://doi.org/10.1016/j.jclinepi.2014.09.007>
- Stavseth, M. R., Clausen, T., & Røislien, J. (2019). How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data. *SAGE Open Medicine*, *7*, 2050312118822912. <https://doi.org/10.1177/2050312118822912>
- Stekhoven, D. J. (2011). Using the missForest package. *R Package*, 1–11.
- Stekhoven, D. J., & Bühlmann, P. (2011). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118. <https://doi.org/10.1093/bioinformatics/btr597>
- Steyerberg, E. W., Harrell Jr, F. E., Borsboom, G. J., Eijkemans, M., Vergouwe, Y., & Habbema, J. D. F. (2001). Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, *54*(8), 774–781. [https://doi.org/10.1016/S0895-4356\(01\)00341-9](https://doi.org/10.1016/S0895-4356(01)00341-9)
- Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307. <https://doi.org/10.1186/1471-2105-9-307>
- Stuart, S. (2011). IPT adherence and quality scale. *Interpersonal Psychotherapy Institute, Iowa (Unpublished manuscript)*.
- Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *10*(6), 363–377. <https://doi.org/10.1002/sam.11348>
- van Bronswijk, S. C., DeRubeis, R. J., Lemmens, L. H., Peeters, F. P., Keefe, J. R., Cohen, Z. D., & Huibers, M. J. (2019). Precision medicine for long-term depression outcomes using the Personalized Advantage Index approach: Cognitive therapy or interpersonal psychotherapy? *Psychological Medicine*, 1–11. <https://doi.org/10.1017/S0033291719003192>
- van Bronswijk, S. C., Lemmens, L. H., Keefe, J. R., Huibers, M. J., DeRubeis, R. J., & Peeters, F. P. (2019). A prognostic index for long-term outcome after successful acute phase cognitive therapy and interpersonal psychotherapy for major depressive disorder. *Depression and Anxiety*, *36*(3), 252–261. <https://doi.org/10.1002/da.22868>
- Van Vliet, I., Leroy, H., & Van Megen, H. (2000). De MINI-Internationaal neuropsychiatrisch interview: Een kort gestructureerd diagnostisch interview voor DSM-IV en ICD-10 psychiatrische stoornissen. *Leiden: LUMC*.
- Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., & Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, *3*(8), e002847. <https://doi.org/10.1136/bmjopen-2013-002847>
- Wang, Y.-P., & Gorenstein, C. (2013). Psychometric properties of the Beck depression inventory-II: A comprehensive review. *Brazilian Journal of Psychiatry*, *35*(4), 416–431. <https://doi.org/10.1590/1516-4446-2012-1048>
- Webb, C. A., Cohen, Z. D., Beard, C., Forgeard, M., Peckham, A. D., & Björgvinsson, T. (2020). Personalized prognostic prediction of treatment outcome for depressed patients in a naturalistic psychiatric hospital setting: A comparison of machine learning approaches. *Journal of Consulting and Clinical Psychology*, *88*(1), 25–38. <https://doi.org/10.1037/ccp0000451>
- Webb, C. A., Trivedi, M. H., Cohen, Z. D., Dillon, D. G., Fournier, J. C., Goer, F., & Parsey, R. (2019). Personalized prediction of antidepressant v. placebo response: Evidence from the EMBARC study. *Psychological Medicine*, *49*(7), 1118–1127. <https://doi.org/10.1017/S0033291718001708>
- 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. <https://doi.org/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. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>