



Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/128568/>

Version: Accepted Version

Article:

Deisenhofer, A.K., Delgadillo, J., Rubel, J. et al. (2018) Individual treatment selection for patients with posttraumatic stress disorder. *Depression and Anxiety*, 35 (6). pp. 541-550. ISSN: 1091-4269

<https://doi.org/10.1002/da.22755>

© 2018 Wiley Periodicals, Inc. This is an author produced version of a paper subsequently published in *Depression and Anxiety*. Uploaded in accordance with the publisher's self-archiving policy.

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Individual treatment selection for patients with post-traumatic stress disorder

Anne-Katharina Deisenhofer, University of Trier

Jaime Delgado, Clinical Psychology Unit, Department of Psychology, University of
Sheffield

Julian Rubel, University of Trier

Jan R. Böhnke

Dundee Centre for Health and Related Research, School of Nursing and Health Sciences,
University of Dundee, Dundee DD1 4HJ; Department of Health Sciences, University of York

YO10 5DD

Dirk Zimmermann, University of Trier

Brian Schwartz, University of Trier

Wolfgang Lutz, University of Trier

Contact Information :

Anne-Katharina Deisenhofer

Clinical Psychology and Psychotherapy

Department of Psychology

University of Trier

D-54296 Trier, Germany

Phone: +49-651-201-2882

Fax: +49-651-201-2886

E-mail: deisenhofer@uni-trier.de

This work was supported by the German Research Foundation (W.L., grant numbers LU
660/10-1, LU 660/8-1).

Abstract

Background: Trauma focused cognitive behavioral therapy (Tf-CBT) and eye movement desensitization and reprocessing (EMDR) are two highly effective treatment options for post-traumatic stress disorder (PTSD). Yet, on an individual level, PTSD patients vary substantially in treatment response. The aim of the paper is to test the application of a treatment selection method based on a personalized advantage index (PAI).

Method: The study used clinical data for patients accessing treatment for PTSD in a primary care mental health service in the north of England. PTSD patients received either EMDR ($N = 75$) or Tf-CBT ($N = 242$). The Patient Health Questionnaire (PHQ-9) was used as an outcome measure for depressive symptoms associated with PTSD. Variables predicting differential treatment response were identified using an automated variable selection approach (genetic algorithm) and afterwards included in regression models, allowing the calculation of each patient's personalized advantage index (PAI).

Results: Age, employment status, gender and functional impairment were identified as relevant variables for Tf-CBT. For EMDR, baseline depressive symptoms as well as prescribed antidepressant medication were selected as predictor variables. Fifty-six percent of the patients ($n = 125$) had a PAI equal or higher than one standard deviation. From those patients, 62 (50%) did not receive their model-predicted treatment and could have benefited from a treatment assignment based on the PAI.

Conclusions: Using a PAI-based algorithm has the potential to improve clinical decision-making and to enhance individual patient outcomes, although further replication is necessary before such an approach can be implemented in prospective studies.

Introduction

Within the context of evidence-based medicine, a series of clinical practice guidelines for post-traumatic stress disorder (PTSD) have been developed and published internationally (e.g. the National Institute for Health and Clinical Excellence, 2005; Foa, Keane, Friedman, & Cohen, 2008). In a review, Forbes and colleagues (2010) examined existing guidelines and concluded that trauma-focused cognitive behavioral therapy (Tf-CBT) is consistently recommended as first-line psychological treatment, whereas eye movement desensitization and reprocessing (EMDR) is not always endorsed as equivalent. Despite these differences, CBT as well as EMDR have been shown to be highly effective for the treatment of PTSD in several randomized controlled trials and systematic reviews (Benish, Imel, & Wampold, 2008; Bisson et al., 2007; Bisson & Andrew, 2009; Bradley, Greene, Russ, Dutra, & Westen, 2005; Ehlers et al., 2010; Seidler & Wagner, 2006; Watts et al., 2013).

Although both interventions can be effective in the treatment of PTSD, patients with this condition vary substantially in their treatment response and illness course. Dropout rates in PTSD treatments vary between 14% and 32% (Van Minnen, Arntz, & Keijsers, 2002; Hembree et al., 2003; Van Etten & Taylor, 1998; Resick, Nishith, Weaver, Astin, & Feuer, 2002; Schottenbauer, Glass, Arnkoff, Tendick, & Gray, 2008). Furthermore, several studies show that between 41% and 58% remain clinically distressed after treatment (Tarrier et al., 1999; Resick et al., 2002). In a review, Schottenbauer and colleagues (2008) report non-response rates as high as 50%. Despite research results which show equivalent outcomes on average for Tf-CBT and EMDR, individual patients might respond differently to each of the two treatments. It is possible that a treatment that is highly effective for one patient might be ineffective or even harmful for another patient. Based on this argument, Schnyder (2005), for example, suggests a need to develop new treatments for PTSD. In contrast, Seidler and

Wagner (2006) argue that future research should focus on understanding which patients are more likely to benefit from one treatment or the other.

In this sense, clinical research is gradually shifting to an increased focus on the individual patient. Tailoring treatments to the specific characteristics of a patient is an approach that has been described as *precision medicine*. This approach has a long tradition in medicine, exemplified by attempts to identify genetic and neuroimaging markers that predict differential treatment response to pharmacological treatments (see for example: Hamburg & Collins, 2010). In the context of psychotherapy research, different strategies have been applied to empirically define the most promising treatment for a particular patient (e.g., Cuijpers, 2014; Cohen & DeRubeis, 2017; DeRubeis et al., 2014; Kessler et al., 2017; Lutz et al., 2006; Ng & Weisz, 2016).

Thus far, in the context of psychotherapy research, some studies have focused on the task of developing prognostic and/or prescriptive models for differential treatment selection. A model is called prognostic if the variables included predict response in a single treatment whereas prescriptive variables are often referred to as moderators that affect the direction or strength of the differences in outcome between two or more treatment conditions (Cohen & DeRubeis, 2017). In this sense, DeRubeis and colleagues (2014) developed a prescriptive model, which they called Personalized Advantage Index (PAI). Implemented in the context of two highly effective treatments, the PAI predicted a clinically meaningful advantage for 60% of depressed patients if they had been assigned to their predicted optimal treatment. Huibers et al. (2015) applied the same approach to a sample where depressed patients were randomized to either cognitive therapy or interpersonal psychotherapy ($N = 134$). The PAI was able to predict a clinically meaningful advantage in one of the therapies for 63% of the sample. More recently, Delgado, Huey, Bennett and McMillan (2017) used data from a large naturalistic cohort ($N = 1512$) of patients to construct a prognostic index. They found

that a subgroup of patients referred to as “complex cases” tended to attain significantly better outcomes if they were initially assigned to high intensity psychotherapies, rather than low intensity CBT. Similarly, Kessler et al. (2017) recommended using prognostic models to estimate predictions for each treatment type, ideally using large naturalistic clinical samples to identify reliable predictors that could be applied in future clinical trials.

In summary, the application of treatment selection based on prognostic and prescriptive models has mainly focused on samples of patients suffering from depression, where they were most often randomly assigned to different treatments. However, heterogeneity in treatment response is not limited to patients with major depression. As highlighted above, PTSD is heterogeneous in presentation and prognosis, and therefore this diagnosis is not in itself sufficient for making an optimal treatment decision. On this basis, the objective of the present study was to develop a treatment selection method using data from a naturalistic PTSD sample where patients were treated either with Tf-CBT or EMDR, in order to identify the optimal treatment for each patient.

Methods

This study was based on the analysis of anonymous clinical case records for $N = 317$ patients accessing treatment for PTSD in a primary care mental health service in the north of England. This service was part of the national Improving Access to Psychological Therapies (IAPT) programme in England, which offers evidence-based psychological interventions organized in a stepped care model (Clark, 2011). Patients with PTSD received either trauma-focused cognitive behavioral therapy (Tf-CBT; Ehlers, Clark, Hackmann, McManus, & Fennell, 2005) or eye-movement desensitization and reprocessing (EMDR; Shapiro, 2001). Following clinical guidelines (NICE, 2005), these treatment options were explained at the time of initial assessments and patients made joint decisions with the assessing therapists about their treatment of choice. In the analysis sample, 242 patients received Tf-CBT

whereas 75 patients received EMDR. This distribution of cases across treatments was also influenced by the larger number of CBT therapists (N=43) relative to EMDR therapists in this sample (N=4). These interventions were delivered by qualified therapists who practiced under regular clinical supervisions, in weekly appointments for up to 20 sessions. De-identified clinical assessment records were available, including demographic and clinical information (primary diagnosis, session-to-session outcome measures) (see Table 1).

Measures

Patient Health Questionnaire (PHQ-9). The PHQ-9 is a validated nine-item screening tool, which is routinely used in IAPT services to measure symptom improvement (Kroenke, Spitzer, & Williams, 2001). The questionnaire has been developed to measure depression. Patients rate each item on a 0 to 3 scale (0 = 'not at all' to 3 = 'nearly every day'), yielding a total score between 0 and 27. A difference of six points on the PHQ-9 was taken as a reliable change index (Richards & Borglin, 2011). Classification of PHQ-9 scores following Kroenke, Spitzer and Williams (2001): mild depression ≥ 5 , moderate depression ≥ 10 , moderately severe depression ≥ 15 , severe depression ≥ 20 . The mean level of baseline depression severity in this sample was PHQ-9 = 16.32. We did not have a PTSD-specific outcome measure available in this sample, however the PHQ-9 has been shown to be strongly associated ($r = 0.59$) with PTSD symptoms (Gerrity, Corson, & Dobcha, 2007), and therefore using it as a primary outcome could be justified. The PHQ-9 has also been extensively validated in primary care populations (Kroenke, Spitzer, Williams, & Löwe, 2010; Kroenke et al., 2001).

Generalized Anxiety Disorder 7 (GAD-7). The GAD-7 is a seven-item measure developed to screen for anxiety disorders (Spitzer, Kroenke, Williams, & Löwe, 2006). Each item is also rated on a 0 to 3 scale, yielding a total anxiety severity score between 0 and 21. In the following, the instrument served as potential predictor variable.

Work and Social Adjustment Scale (WSAS). Functional impairment was assessed using the WSAS (Mundt, Mark, Shear, & Griest, 2002). The WSAS measures the extent to which mental health problems impair daily functioning across five domains (work, home chores, social leisure, private leisure, and relationships). The internal consistency of the WSAS has been found to range from $\alpha = .70$ to $\alpha = .94$ with a test-retest reliability of $r = .73$ (Mundt et al., 2002). Initial WSAS scores were included in the process of identifying treatment-specific predictors.

Data Analysis Strategy

Missing data. To address missing data, including missing outcome assessments, we implemented a non-parametric missing value imputation procedure using random forests with the R package *missForest* (Stekhoven & Bühlmann, 2012; Stekhoven, 2013). Imputation via *missForest* has been shown to yield a lower imputation error than other common imputation approaches (Waljee et al., 2013). Table 1 gives an overview of missing values in the analysis sample. A good performance of *missForest* for continuous variables is assessed with the normalized root mean squared error (NRMSE) and for categorical variables using a proportion of falsely classified entries (PFC) close to zero (Stekhoven & Bühlmann, 2012). Imputation was successful with a NRMSE of 0.30 and a PFC of 0.12. All further depictions and analyses are based on the imputed dataset.

Propensity Score Matching. Since the present data stem from a naturalistic treatment context, we had to adjust for confounding by indication. Therefore, we implemented propensity score matching (PSM; Rosenbaum & Rubin, 1983; Lutz, Schiefele, Wucherpfennig, Rubel, & Stulz, 2016). PSM can be used to match or to equate groups based on comparable baseline characteristics (West et al., 2014). We performed PSM in R based on the *MatchIT* package (Ho et al., 2011) using *optimal matching*. Optimal matching finds the matched samples with the smallest average absolute distance across all the matched pairs

based on all available baseline characteristics. The process was iterated until all optimal matches between the EMDR sample and the Tf-CBT sample were found. For this study, we implemented the standardized mean difference (*SMD*) technique, a widely used method to check covariate balance between samples (e.g., Guo & Fraser, 2014). The *SMD* method is similar to Cohen's *d* and allows the comparison of differences in matched and unmatched conditions for each covariate (with the *SD* of the unmatched condition being used as the denominator). A *SMD* < .25 indicates acceptable match between samples on the respective covariates (Rubin, 2001). For further analyses, all predictors were centered within the respective treatment condition: continuous variables were grand mean-centered whereas dichotomous variables were dummy coded with values set to -0.5 and 0.5 (Kraemer & Blasey, 2004).

Selection of variables via a genetic algorithm. Traditional methods for model selection are based on null hypothesis testing (e.g., Anderson, 2007; Anderson & Burnham, 2002). This often includes a stepwise procedure where the researcher fits the full model and looks for terms that are not statistically significant, i.e., whose removal does not significantly reduce the fit of the model. The procedure can be repeated until all effects in the formula are found significant. This approach is often called 'backward elimination'. A similar strategy is to start from the simplest model and to sequentially add the most significant effects (a 'forward selection' approach). Hence, for each step a significance test is needed to determine whether the removal or addition of a given term is useful. Since the number of tests is typically high, this poses the problem of choosing a relevant significance level (Harrell, 2001).

We used a genetic algorithm (GA) for model selection, which overcomes the above described problems of traditional methods and is considered to be robust and efficient (Calcagno & Mazancourt, 2010). The basic idea of the algorithm is that there is not one

perfect model, but a set of candidate models that are satisfactory, which results in a set of variables that are important for further analyses. The importance of a predictor variable is calculated by summing up the weights for the models in which the variable appears. We used an 80% *importance threshold* to differentiate between important and not so important variables (Calcagno & Mazancourt, 2010).

In this paper we implemented GA in R based on *glmulti*, a package for automated model selection (Calcagno, 2013). The package produces model formulas, which are passed through a fitting function. In our case, this fitting function is a linear regression model with treatment outcome (PHQ-9 final score) as dependent variable. We included initial symptom measures (PHQ-9, WSAS, GAD-7) as well as other available variables (comorbid long-term medical condition, disability, antidepressant medication, gender, age, employment status) as candidate predictors in the model selection process. Medication and employment status were included in the analysis as binary variables. The medication variable included cases that were “prescribed antidepressant medication” (52.17%) and those who were “prescribed but not taking antidepressant medication” (1.34%) into a single category (coded 0.5), with “no prescribed medication” (46.49%) as the reference category (coded -0.5). Employment status was collapsed merging “employed” (50%) and “student” (3.47%) into a single category (coded 0.5), versus “unemployed” (“unemployed” (2.43%), “longterm sick” (11.46%) and “other” (32.64%)) which was the reference category (coded -0.5). See Table 1 for a more detailed description concerning treatment groups. Since interactions of variables with treatment conditions require large sample sizes (Calcagno & Mazancourt, 2010), we computed separate models for each treatment (Kessler et al., 2017).

Personalized Advantage Index (PAI). We used a leave-one-out approach (Efron & Gong, 1983; DeRubeis et al., 2014), in which regression models were estimated according to the sample size of each treatment condition (Tf-CBT $N = 150$ & EMDR $N = 75$). Each of the

models excluded the target patient for whom the PAI prediction was estimated to avoid overfitting. For each patient, two separate predictions were made, one based on the Tf-CBT ($N = 150$) treatment-specific predictors established in the preceding step and one for EMDR ($N = 75$). Since patients had received one of the two treatments, one prediction represented the patient's factual (predicted score for the treatment the patient actually received) prediction and the other represented the patient's counterfactual (predicted score for the treatment the patient did not receive) treatment prediction.

In a next step, the aim was to test the within-sample utility of the model for enhancing the treatment outcome through treatment assignment. First, we examined the *true error* as the average of the absolute difference between the observed scores and factual predictions. Second, we examined the observed change scores of the patients. For that we compared the predictions from the two regression models for every patient and we refer to the treatment condition predicted to have a greater benefit as the *optimal treatment*, whereas the other is referred to as a *sub-optimal treatment*. We then classified patients accordingly as having received the optimal or sub-optimal treatment and compared the rates of clinically significant improvement achieved in both groups between pre- and post-assessments (difference of 6 points on the PHQ-9; Richards & Borglin, 2011). To aid interpretation from a clinical perspective, we calculated the number-needed-to-treat (NNT) using the formulae provided by Kraemer and Kupfer (2006). Finally, we calculated for each patient the predicted absolute difference in outcome between receiving their optimal versus sub-optimal treatment. Following DeRubeis et al. (2014), we referred to the index of the predicted advantage as the Personalized Advantage Index (PAI).

Results

Comparison of Tf-CBT and EMDR

Comparing the effect sizes (Cohen's d) between treatment conditions, there were no significant differences between Tf-CBT (PHQ-9: $d = 0.81$, 95% CI [0.63,1.00]) and EMDR (PHQ-9: $d = 0.89$, 95% CI [0.55, 1.22]) concerning average treatment effectiveness (see Table 2).

Independent-samples t tests and χ^2 tests were calculated to further compare baseline variables (see Table 1) between the two treatment conditions. In summary, the two naturalistic samples differed systematically in a number of the investigated baseline variables, as was to be expected in observational data. Consequently, PSM was necessary before selecting variables for model building.

Propensity Score Matching

The matching process via *optimal matching* (Ho et al., 2011) with all baseline variables resulted in a subsample of 150 patients of the Tf-CBT sample while all EMDR ($N = 75$) cases were included (see Table 1). After applying PSM, all baseline variables were sufficiently well balanced, as none of the SMD scores exceeded .25. The sample selection process is presented in a flow chart in Figure 1. For correlations of baseline variables, see Appendix Table A1.

Separate variable selection for Tf-CBT and EMDR

Relevant Variables for Tf-CBT. The genetic algorithm selected four variables for Tf-CBT (see Table 3 and Table 4). Higher functional impairment (WSAS) at the beginning of therapy significantly predicted higher PHQ-9 end scores ($B = 0.24$, $t(147) = 4.30$, $p < 0.001$) whereas female patients showed better therapy outcomes ($B = -2.09$, $t(147) = -1.93$, $p = 0.06$). Furthermore, being employed and being older led to significantly better therapy outcomes ($B = -4.99$, $t(147) = -4.21$, $p = 0.001$; $B = -0.10$, $t(147) = -2.51$, $p = 0.01$)

Relevant Variables for EMDR. For EMDR, the variable selection process identified two variables (see Table 3). First, higher PHQ-9 scores at the beginning of treatment significantly predicted higher end scores in the same instrument ($B = 0.44$, $t(72) = 2.77$, $p < 0.001$). Furthermore, patients who were prescribed antidepressants tended to have poorer treatment outcomes compared to patients that were not prescribed medications ($B = 4.40$, $t(72) = 3.77$, $p < 0.01$; see Table 4).

Personalized Advantage Index (PAI)

The prediction of an individual's final PHQ-9 score was developed separately for each treatment condition based on treatment-specific predictors described above. The true error of the PHQ-9 post-treatment score predictions was 5.07, representing the average absolute difference between the predicted and actual scores across the 225 patients. The true error for the EMDR model ($N = 75$) and the Tf-CBT model ($N = 150$) was 5.37 and 4.92, respectively. In Figure 2, we present the frequency of predicted PHQ-9 post-treatment scores in both the optimal and sub-optimal treatment, a lower PHQ-9 score representing better therapy outcome. Patients who were classified as having received their optimal treatment had an observed mean PHQ-9 post score of 7.86 ($SD = 6.77$; $n = 124$) which would be classified as mild depression (Kroenke, Spitzer, & Williams, 2001). Patients who received their predicted sub-optimal treatment had a PHQ-9 post-treatment score of 10.89 ($SD = 8.17$; $n = 101$) which can be classified within the range of moderate depression. The standardized difference in observed PHQ-9 post-treatment scores between patients who received their predicted optimal treatment and patients who received their sub-optimal treatment corresponds to Cohen's $d = 0.40$ (95% CI [0.13, 0.67]). This finding can be translated into a NNT = 4.49, meaning that it is necessary to treat between four and five patients in their predicted optimal treatment to attain one additional case with reliable improvement, by comparison to the sub-optimal treatment.

To further test the utility of our approach, we compared the rates of reliable improvement between patients who had their optimal or their sub-optimal treatment based on the treatment-specific predictors. Table 5 shows that sixty-three percent ($n = 78$) of the patients receiving their optimal treatment ($n = 124$) had a reliable improvement after treatment, while in the group of patients receiving their sub-optimal treatment this was only true for 33.66% ($n = 34$). In total the model prediction was true for 64.44% ($n = 145$) percent of the total sample ($n = 225$). The presented frequencies in Table 5 were significantly different, $\chi^2(1, n = 225) = 19.54, p < .001$.

Based on the predictions, the PAI for each individual patient was calculated by subtracting the predicted outcomes for the model-determined optimal treatment from the sub-optimal treatment. The average PAI was 2.49 ($SD = 1.92, range = 0.01 - 9.34$), representing an expected average of 2.49 point difference in PHQ-9 post-treatment scores between the predicted optimal treatment (predicted $M = 7.97, SD = 3.47, range = 0.62 - 14.56$) versus the sub-optimal (predicted $M = 10.47, SD = 3.68, range = 2.36 - 18.07$) treatment. The predicted benefit of treatment selection is displayed in Figure 3 based on the frequencies of PAI scores. Note that the PAI can be as low as zero, which occurs when the same outcome is predicted for both treatments. For 32 patients (14.22%) the PAI was close to zero ($PAI \leq 0.5$) meaning that for those patients there was no predicted difference between the two treatment conditions. Helping to classify the importance of the difference between predicted optimal and sub-optimal treatment, the PAI was further inspected regarding its standard deviation (SD; see Figure 3). Hence, a PAI of 0.96 corresponds to half a SD. In our sample, 171 patients (76%) had a PAI of this size or larger. Furthermore, 55.56% ($n = 125$) of the patients had a PAI that corresponds to one SD. From those, 63 patients (49.6%) did not receive their model predicted optimal treatment and could have benefited from an algorithm-based

treatment assignment. A PAI equal or higher than two SDs was true for 22.67% ($n = 51$) of the patients of the sample.

Discussion

This is the first study using the PAI approach in a naturalistic PTSD sample treated by either Tf-CBT or EMDR. We explored how patients' pretreatment characteristics may guide optimal treatment assignment to enhance therapy outcome for PTSD patients in the context of otherwise highly effective treatments.

Main findings

In the matched naturalistic treatment conditions, we identified treatment-specific outcome predictors. For Tf-CBT, functional impairment, age, gender and employment status were significant predictors. For EMDR initial impairment as well as prescribed antidepressant medications were significant variables for outcome prediction.

As expected, outcomes for patients who received their model determined optimal treatment were better than those of patients who received their sub-optimal treatment. Patients in their model-predicted optimal treatment finished treatment with PHQ-9 scores in the range of mild depression, whereas observed scores for those who received their sub-optimal treatment were still in the range of moderate depression. Although the difference in absolute terms between those two groups could be classified as small (3.03 points difference in PHQ-9 units) it corresponds to a NNT close to four. Bearing in mind that in the present study two highly effective treatments for PTSD were compared with each other, a NNT of four is an impressive effect size difference. In comparison, a meta-analysis which focused on the relative efficacy of psychotherapies for PTSD reported an effect size of 0.16 which translates into $NNT = 12$ (Benish, Imel, & Wampold, 2008). Concerning the prediction models, 56% ($n = 125$) of the sample had a PAI equal or higher than one SD. From those,

50% ($n = 63$) did not receive their model predicted optimal treatment and could have benefited from an algorithm-based treatment assignment.

Our results are consistent with existing literature where the PAI predicted a meaningful advantage for ~60% of cases receiving their optimal treatment (DeRubeis et al., 2014; Huibers et al., 2015). Despite these similarities, there are important differences concerning diagnoses, outcome measures and definitions of clinically important change. Nevertheless, as in previous studies, we found that the PAI approach is promising enough in retrospective samples to justify its application prospectively in applied research studies.

It is still unclear if it is better to build treatment selection models by searching for moderators that are interactions of predictors with treatment in the whole sample (prescriptive models) or by developing separate prognostic models within each treatment condition. Until now, most of the PAI studies used a prescriptive approach (for example DeRubeis et al. 2014; Huibers et al., 2015). Others suggested using prognostic models in large observational treatment studies to answer questions concerning personalized predictions (e.g., Delgado et al., 2017; Kessler et al., 2017). Up to this point, there is no empirical evidence to guide us on the relative strengths, weaknesses or differences between these modelling strategies. We decided to use a prognostic approach since finding interactions in small samples has low statistical power. Nevertheless, future research needs to address this empirical question.

Strengths and limitations

The current study is the first to apply the PAI approach in a naturalistic PTSD sample. So far, research investigated the PAI approach focusing solely on depression RCT samples. RCT samples are highly selective since only patients can be included who agree to be treated with each of the treatments investigated. The present study used a patient preference design where patients can choose between two available treatments within the context of routine care. As a consequence, in a RCT sample model selection is compared to randomization

which is far from being an approach used in routine care. In contrast to that, due to the present study design, model prediction is compared to patient preference which has more meaning for clinical practice. That is, our results show that with a machine learning based approach for treatment selection better outcomes can be achieved than with a selection based on patient preferences. Of course, this assumes that patients who have a preference for a certain treatment can be convinced by the data to select a different treatment than their preferred one if this promises better outcomes for them. However, even if patients agree to undergo a treatment that is not their first choice, it is unclear whether this preference per se might influence treatment outcomes. Future applications of the PAI in RCTs and naturalistic contexts explicitly need to take patient preferences into account in order to get a better understanding of the role of this potentially important variable.

Apart from this, the use of retrospective data to determine the potential utility of the PAI still raises some questions about the viability, acceptability, and potential effects of using this strategy to prospectively assign patients to treatments. Nonetheless, a genetic algorithm as well as a leave-one-out approach was implemented to enhance the probability that the identified predictor variables will be replicated in other samples. Furthermore, we have demonstrated and replicated the usefulness of the PAI approach, which justifies the effort to conduct prospective studies. To our knowledge, there is only one project that is applying prospective personalized treatment recommendations in an outpatient clinic (Lutz, Zimmermann, Müller, Deisenhofer, & Rubel, 2017). Instead of the PAI, personalized predictions are based on a nearest neighbor approach (NN). In our view, this is the first step to bring personalized treatments finally within reach (Cuijpers & Christensen, 2017).

A further limitation concerns the relatively small sample size in which these models were evaluated. There is a potential risk of overconfidence due to the fact that variable selection was performed on the full sample. Unfortunately, due to the sample size we could

not test the validation of our analyses in an external validation sample. There is a clear need for replications in bigger samples that allow external validation.

Another limitation refers to the main outcome measure PHQ-9 which is an instrument to assess depressive symptoms, but which in English IAPT services is one of the few mandatory instruments in routine data collections. Consequently, our results may not generalize to PTSD symptoms assessed with disorder-specific measures. Nevertheless, we still found meaningful differences between optimal and sub-optimal treatments and they therefore encourage future investigations using PTSD-specific instruments (Brewin, 2005). One possible explanation of our finding is that depression is one of the most common comorbid diagnoses in PTSD (Bradley et al., 2005).

PSM has its advantages and limits, and this study is no exception (Shadish, 2013). Although the method is able to account for observed potential confounders, there is still the possibility that unobserved variables might have had an influence on treatments, treatment selection, or therapy outcome. Nevertheless, we think that complete elimination of bias is unrealistic and the advantage of implementing the PAI in routine care data outweighs the potential shortcomings of the matching method.

Conclusions

Acknowledging the above limitations, this article adds to the growing body of research on personalized treatments in psychotherapy research. Results suggest that it is possible to enhance individual treatment outcomes in assigning PTSD patients to their optimal treatment. This kind of study promotes the development of meaningful decision trees, which will result in supported clinical decision-making (see: Lutz, de Jong, & Rubel, 2015). In the future, PAIs could be integrated in the diagnostic process at the beginning of psychotherapy to find the optimal treatment for a patient in advance. Matching patients to

their model-determined optimal treatment could potentially further enhance treatment outcomes.

Acknowledgements

None.

Financial support

This work was supported by the German Research Foundation (W.L., grant numbers LU 660/10-1, LU 660/8-1).

Conflict of interest

None.

Ethical Standard

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

References

- Anderson, D. R. (2007). *Model based inference in the life sciences: a primer on evidence*. Springer Science & Business Media.
- Anderson, D. R., & Burnham, K. P. (2002). Avoiding pitfalls when using information-theoretic methods. *The Journal of Wildlife Management*, 912-918.
- Benish, S. G., Imel, Z. E., & Wampold, B. E. (2008). The relative efficacy of bona fide psychotherapies for treating post-traumatic stress disorder: A meta-analysis of direct comparisons. *Clinical psychology review*, 28(5), 746-758.
- Bisson, J., & Andrew, M. (2009). Psychological treatment of post-traumatic stress disorder (PTSD). *The Cochrane Library, Issue 1*. JohnWiley & Sons, Ltd.
- Bisson, J. I., Ehlers, A., Matthews, R., Pilling, S., Richards, D., & Turner, S. (2007). Psychological treatments for chronic post-traumatic stress disorder. *The British journal of psychiatry*, 190(2), 97-104.
- Bradley, R., Greene, J., Russ, E., Dutra, L., & Westen, D. (2005). A multidimensional meta-analysis of psychotherapy for PTSD. *American journal of psychiatry*, 162(2), 214-227.
- Brewin, C. R. (2005). Systematic review of screening instruments for adults at risk of PTSD. *Journal of traumatic stress*, 18(1), 53-62.
- Calcagno, V. (2013). *glmulti: Model selection and multimodel inference made easy*. R package version 1.0.7. <https://CRAN.R-project.org/package=glmulti>
- Calcagno, V., & de Mazancourt, C. (2010). glmulti: an R package for easy automated model selection with (generalized) linear models. *Journal of statistical software*, 34(12), 1-29.

- Clark, D. M. (2011). Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: the IAPT experience. *International Review of Psychiatry*, 23, 318-327.
- Cuijpers, P. (2014). Personalized treatment for functional outcome in depression. *Medicographia*, 36(4).
- Cuijpers, P., & Christensen, H. (2017). Are personalised treatments of adult depression finally within reach?. *Epidemiology and psychiatric sciences*, 26(1), 40-42.
- Cohen, J. (1992). A power primer. *Psychological bulletin*, 112(1), 155-159.
- Cohen, Z. D., & DeRubeis, R. J. (2017). Treatment selection in depression. Manuscript submitted for publication.
- Delgadillo, J., Huey, D., Bennett, H., & McMillan, D. (2017). Case complexity as a guide for psychological treatment selection. *Journal of Consulting and Clinical Psychology*. 85(9), 835-853.
- 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.
- Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1), 36-48.
- Ehlers, A., Bisson, J., Clark, D. M., Creamer, M., Pilling, S., Richards, D., ... & Yule, W. (2010). Do all psychological treatments really work the same in posttraumatic stress disorder? *Clinical psychology review*, 30(2), 269-276.
- Ehlers, A., Clark, D. M., Hackmann, A., McManus, F., & Fennell, M. (2005). Cognitive therapy for PTSD: development and evaluation. *Behaviour Research and Therapy*, 43, 413-431.

- Foa, E. B., Keane, T. M., Friedman, M. J., & Cohen, J. A. (Eds.). (2008). *Effective treatments for PTSD: practice guidelines from the International Society for Traumatic Stress Studies*. Guilford Press.
- Forbes, D., Creamer, M., Bisson, J. I., Cohen, J. A., Crow, B. E., Foa, E. B., ... & Ursano, R. J. (2010). A guide to guidelines for the treatment of PTSD and related conditions. *Journal of traumatic stress, 23*(5), 537-552.
- Gerrity, M. S., Corson, K., & Dobscha, S. K. (2007). Screening for posttraumatic stress disorder in VA primary care patients with depression symptoms. *Journal of general internal medicine, 22*(9), 1321-1324.
- Guo, S., & Fraser, M. W. (2014). *Propensity score analysis: Statistical methods and Applications*. Sage Publications, Thousand Oaks, CA.
- Hamburg, M. A., & Collins, F. S. (2010). The path to personalized medicine. *New England Journal of Medicine, 363*(4), 301-304.
- Harrell, F. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer.
- Hembree, E. A., Foa, E. B., Dorfan, N. M., Street, G. P., Kowalski, J., & Tu, X. (2003). Do patients drop out prematurely from exposure therapy for PTSD? *Journal of traumatic stress, 16*(6), 555-562.
- Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric casual inference. *Journal of Statistical Software, 42*, 1-28.
- 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.
- Kessler, R. C., Van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Ebert, D.

- D., ... & Schoevers, R. A. (2017). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiology and psychiatric sciences*, 26(1), 22-36.
- Kraemer, H. C., & Blasey, C. M. (2004). Centering in regression analyses: a strategy to prevent errors in statistical inference. *International journal of methods in psychiatric research*, 13(3), 141-151.
- Kraemer, H. C., & Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological psychiatry*, 59(11), 990-996.
- Kroenke, K, Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine* 16, 606-613.
- Kroenke, K, Spitzer, R. L., Williams, J. B., & Löwe, B. (2010). The patient health questionnaire somatic anxiety and depressive symptoms scales: a systematic review. *General hospital psychiatry*, 32(4), 345-359.
- Löwe, B., Unützer, J., Callahan, C. M., Perkins, A. J., & Kroenke, K. (2004). Monitoring depression treatment outcomes with the patient health questionnaire-9. *Medical care*, 42(12), 1194-1201.
- Lutz, W., De Jong, K., & Rubel, J. (2014). Patient-focused and feedback research in psychotherapy: Where are we and where do we want to go?. *Psychotherapy research: journal of the Society for Psychotherapy Research*, 25(6), 625-632.
- 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.
- Lutz, W., Schiefele, A. K., Wucherpfennig, F., Rubel, J., & Stulz, N. (2016). Clinical

effectiveness of cognitive behavioral therapy for depression in routine care: A propensity score based comparison between randomized controlled trials and clinical practice.

Journal of affective disorders, 189, 150-158.

Lutz, W., Zimmermann, D., Müller, V., Deisenhofer, A.-K., & Rubel, J. (2017, June 28).

Randomized controlled trial to evaluate the effects of personalized prediction and adaptation tools on treatment outcome in outpatient psychotherapy. Retrieved from osf.io/3gr8j.

Mundt, J. C., Mark, I. M., Shear, M. K., & Griest, J. M. (2002). The Work and Social Adjustment Scale: a simple measure of impairment in functioning. *British Journal of Psychiatry* 180, 461-464.

National Institute for Health and Clinical Excellence. (2005). *Post-traumatic stress disorder*. London: Royal College of Psychiatrists and The British Psychological Society.

Ng, M. Y., & Weisz, J. R. (2016). Annual research review: building a science of personalized intervention for youth mental health. *Journal of Child Psychology and Psychiatry*, 57(3), 216-236.

Resick, P. A., Nishith, P., Weaver, T. L., Astin, M. C., & Feuer, C. A. (2002). A comparison of cognitive-processing therapy with prolonged exposure and a waiting condition for the treatment of chronic posttraumatic stress disorder in female rape victims. *Journal of consulting and clinical psychology*, 70(4), 867.

Richards, D. A., & Borglin, G. (2011). Implementation of psychological therapies for anxiety and depression in routine practice: two year prospective cohort study. *Journal of Affective Disorders*, 133(1), 51-60.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41-55.

- Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health services & outcomes research methodology* 2, 169-188.
- Schottenbauer, M. A., Glass, C. R., Arnkoff, D. B., Tendick, V., & Gray, S. H. (2008). Nonresponse and dropout rates in outcome studies on PTSD: Review and methodological considerations. *Psychiatry: Interpersonal and biological processes*, 71(2), 134-168.
- Schnyder, U. (2005). Why new psychotherapies for posttraumatic stress disorder? *Psychotherapy and psychosomatics*, 74(4), 199-201.
- Seidler, G. H., & Wagner, F. E. (2006). Comparing the efficacy of EMDR and trauma-focused cognitive-behavioral therapy in the treatment of PTSD: a meta-analytic study. *Psychological medicine*, 36(11), 1515-1522.
- Shadish, W.R. (2013). Propensity score analysis: promise, reality and irrational exuberance. *Journal of Experimental Criminology* 9, 129-144.
- Shapiro, F. (2001). *Eye Movement desensitization and reprocessing: Basic principles, protocols, and procedures*. New York: Guilford Press.
- Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of internal medicine*, 166(10), 1092-1097.
- Stekhoven, D. J. (2013). missForest: *Nonparametric Missing Value Imputation using Random Forest*. R package version 1.4.
- Stekhoven, D. J., & Bühlmann, P. (2012). MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112-118.
- Tarrier, N., Pilgrim, H., Sommerfield, C., Faragher, B., Reynolds, M., Graham, E., &

- Barrowclough, C. (1999). A randomized trial of cognitive therapy and imaginal exposure in the treatment of chronic posttraumatic stress disorder. *Journal of consulting and clinical psychology*, 67(1), 13.
- Van Etten, M. L., & Taylor, S. (1998). Comparative efficacy of treatments for post-traumatic stress disorder: a meta-analysis. *Clinical psychology and psychotherapy*, 5, 126-144.
- Van Minnen, A., Arntz, A., & Keijsers, G. P. J. (2002). Prolonged exposure in patients with chronic PTSD: Predictors of treatment outcome and dropout. *Behaviour research and therapy*, 40(4), 439-457.
- Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., ... & Higgins, P.D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ open*, 3(8), e002847.
- Watts, B. V., Schnurr, P. P., Mayo, L., Young-Xu, Y., Weeks, W. B., & Friedman, M. J. (2013). Meta-analysis of the efficacy of treatments for posttraumatic stress disorder. *The journal of clinical psychiatry*, 74(6), 541-550.
- West, S.G., Cham, H., Thoemmes, F., Renneberg, B., Schulze, J., Weiler, M., 2014. Propensity scores as a basis for equating groups: Basic principles and application in clinical treatment outcome research. *Journal of Consulting and Clinical Psychology* 82, 906-919.

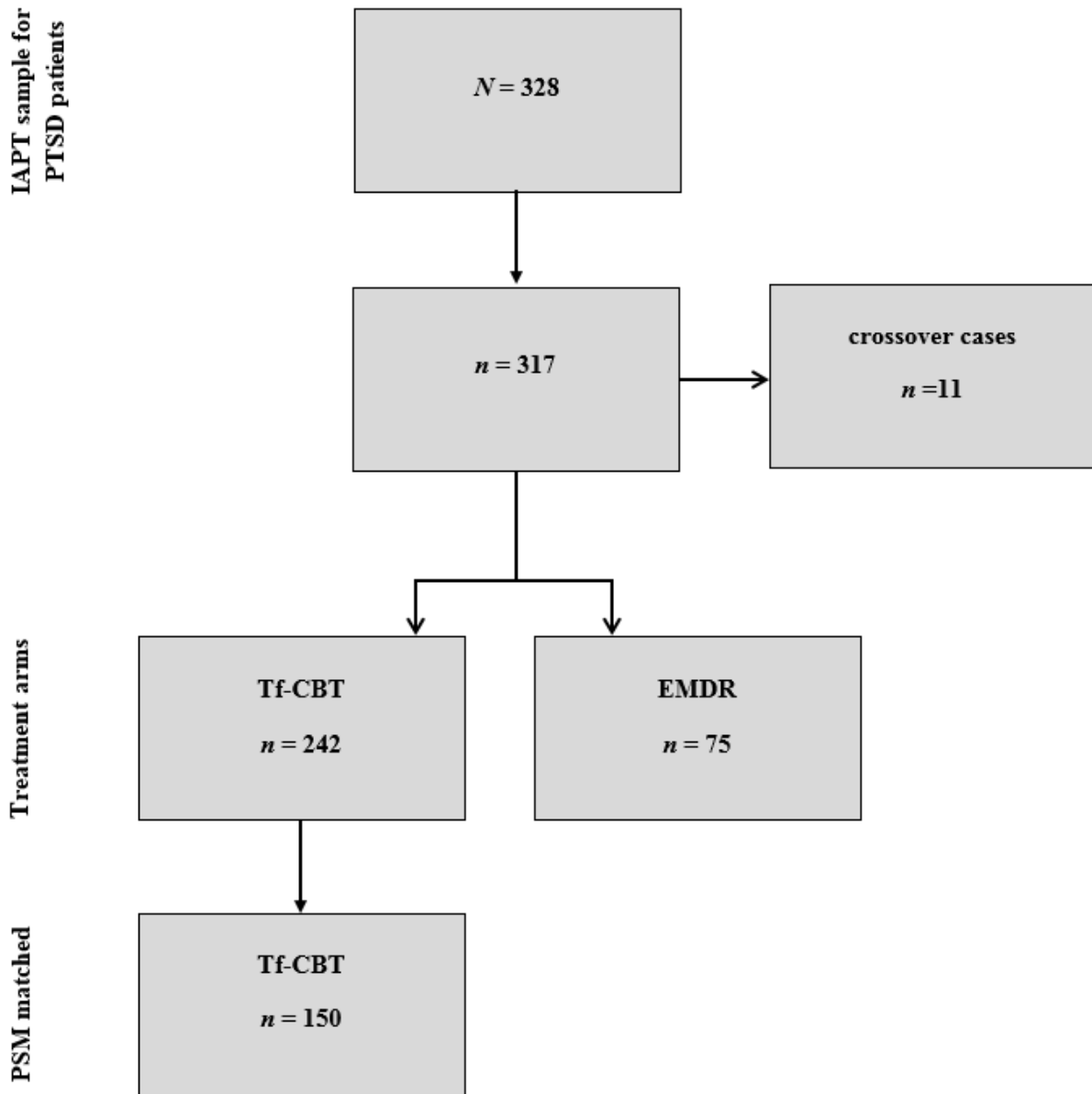


Figure 1. Flow chart of sample preparation and selection. IAPT = Improving Access to Psychological Therapies; PTSD = post-traumatic stress disorder; crossover cases = patients treated by therapists that offered both treatments; Tf-CBT = trauma-focused cognitive behavioral therapy; EMDR = eye movement desensitization and reprocessing; PSM = Propensity Score Matching.

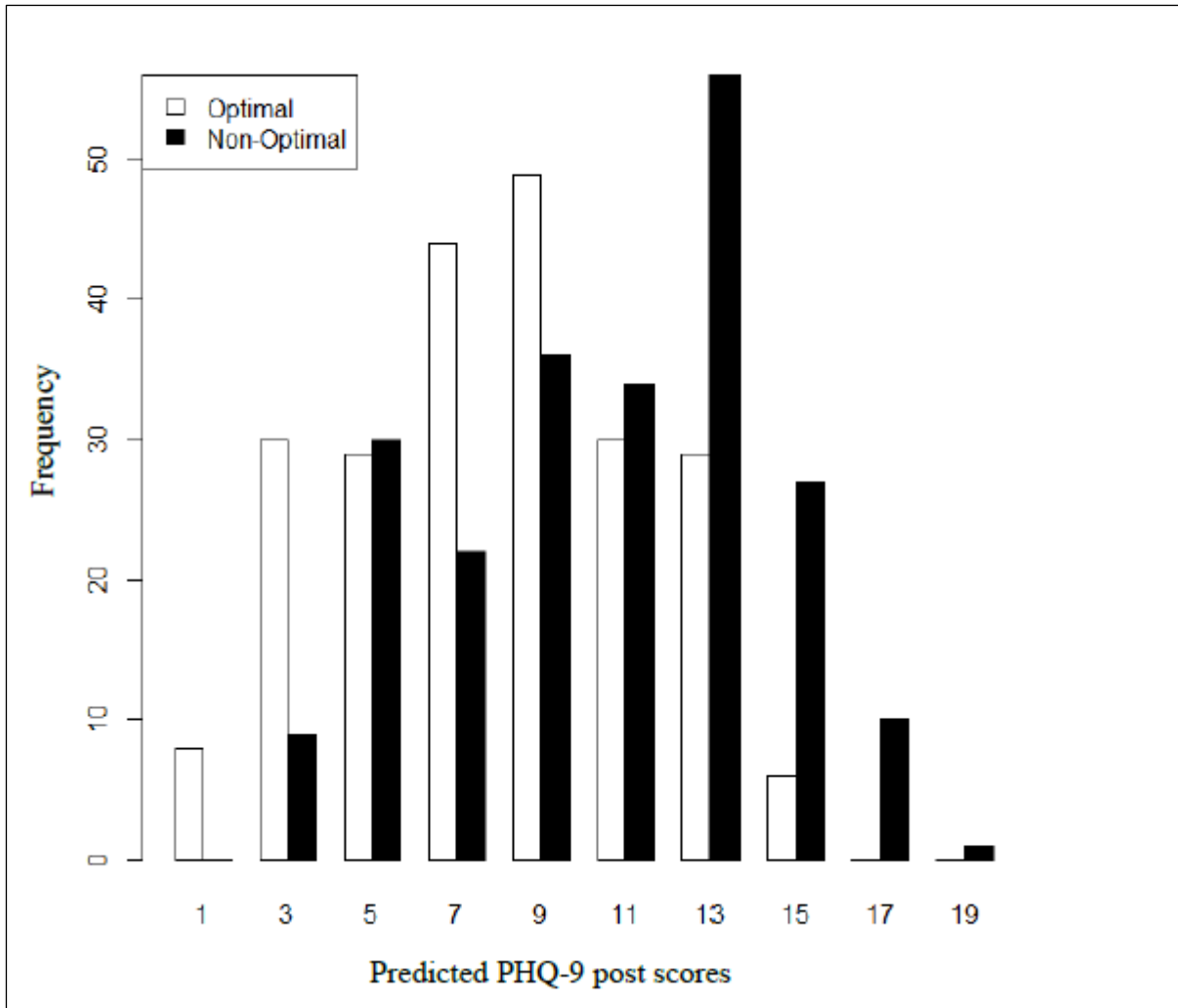


Figure 2. Distribution of predicted PHQ-9 post scores in the optimal and sub-optimal treatment. Optimal = treatment condition with a predicted benefit for the individual patient; Sub-optimal = treatment condition which is predicted as being less effective for the individual patient.

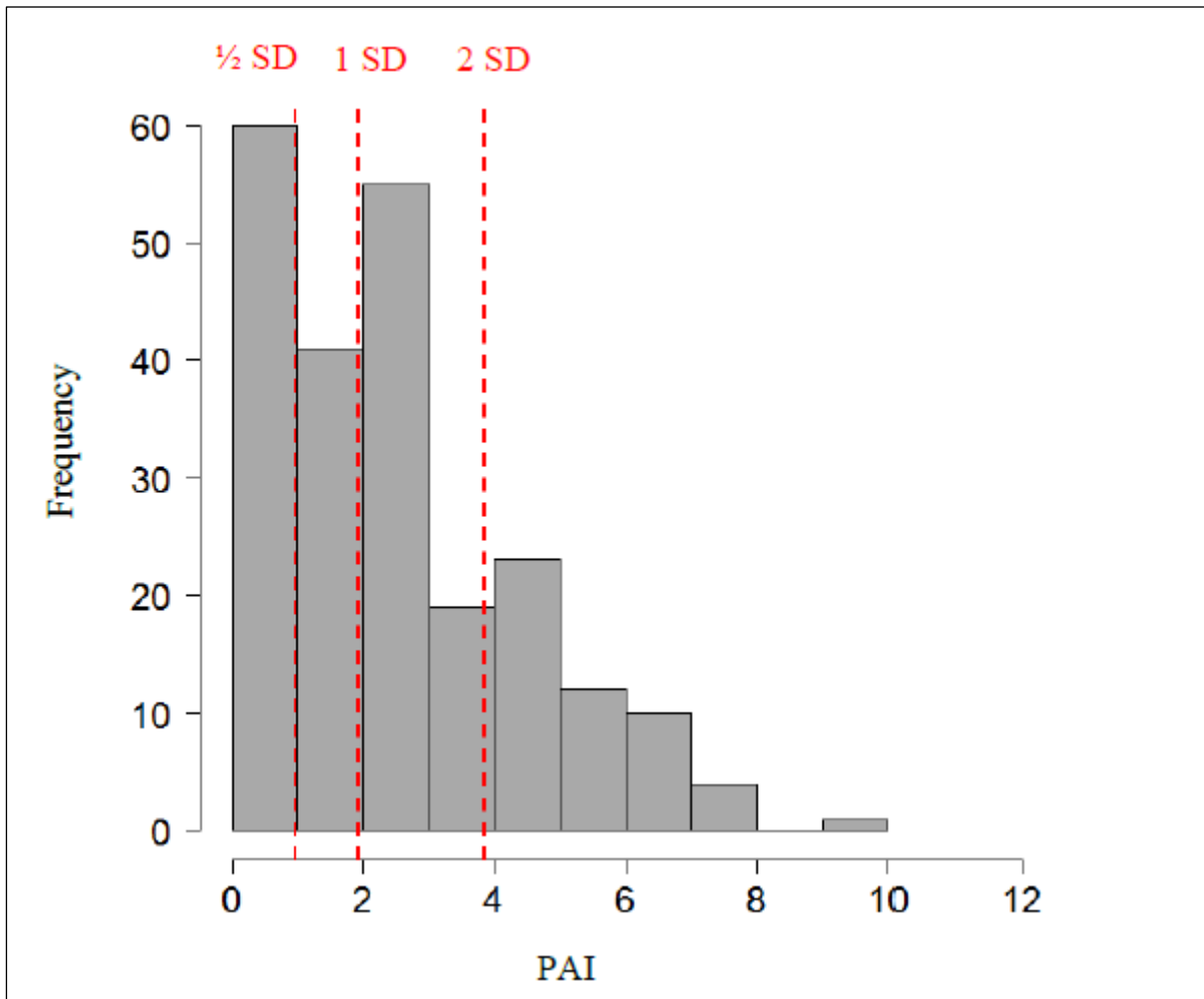


Figure 3. Distribution of PAI scores. PAI = Personalized advantage index (Magnitude of the predicted difference between optimal and sub-optimal treatment); *SD* = Standard Deviation; $\frac{1}{2}$ *SD* = 0.96; 1 *SD* = 1.92; 2 *SD* = 3.85.

Table 1. Sample baseline characteristics (after imputation)

	<u>Tf-CBT</u> (<i>N</i> = 242)	<u>Tf-CBT</u> (<i>N</i> = 150)	<u>EMDR</u> (<i>N</i> = 75)	Missing data (before imputation)
	Mean (SD) or %	Mean (SD) or %	Mean (SD) or %	(%)
PHQ-9 _{pre}	17.49 (6.42)	15.49 (6.91)	15.22 (6.92)	5.40
PHQ-9 _{post}	11.46 (8.11)	9.42 (7.59)	8.81 (7.55)	0.30
GAD-7 _{pre}	15.37 (5.06)	14.13 (5.67)	14.11 (5.71)	5.40
WSAS _{pre}	22.11 (9.53)	21.11 (10.02)	21.18 (10.85)	15.10
Gender (female)	44.63	56.67	60.00	0
Age	38.33 (12.20)	39.39 (12.40)	41.85 (12.52)	0
LTC	27.69	26.00	25.33	0
Disability	68.60	50.67	42.67	0
Employment				
- Employed	47.93	52.00	46.67	9.10
- Student	1.65	3.33	2.67	
- Unemployed	3.31	2.67	4.00	
- Longterm sick	10.33	12.00	16.00	
- Other ^A	36.78	30.00	30.67	
Medication ^B				
- Prescribed	54.13	50.67	46.67	
- Prescribed not taking	0.41 45.45	0.67 48.67	4.00 49.33	5.70
- Not prescribed				

Note. Tf-CBT = trauma-focused cognitive behavioral therapy; EMDR = eye-movement desensitization and reprocessing; PHQ-9 = patient health questionnaire; GAD-7 = Generalized Anxiety Disorder; WSAS = work and social adjustment scale; LTC = lifelong medical condition (e.g., diabetes, heart disease, chronic pain, arthritis, etc.).

^A Other = voluntary work, homemaker, carer, or retired

^B Medication = Antidepressants

Table 2. Treatment outcome in different samples after imputing and matching

Measure	Sample	d ^A	Confidence Interval	
			Lower Limit	Upper Limit
PHQ-9	IAPT (N = 225)	0.84	0.64	1.03
	Tf-CBT (n = 150)	0.82	0.58	1.05
	EMDR (n = 75)	0.89	0.55	1.22
GAD-7	IAPT (N = 225)	0.90	0.7	1.09
	Tf-CBT (n = 150)	0.91	0.67	1.15
	EMDR (n = 75)	0.86	0.53	1.20
WSAS	IAPT (N = 225)	0.73	0.54	1.92
	Tf-CBT (n = 150)	0.67	0.43	0.90
	EMDR (n = 75)	0.85	0.51	1.18

Note. Tf-CBT = trauma-focused cognitive behavioral therapy; EMDR = eye-movement desensitization and reprocessing; PHQ-9 = patient health questionnaire 9; GAD-7 = Generalized Anxiety Disorder 7; WSAS = work and social adjustment scale.

^A Effect size d was computed using the pooled standard deviation of pre and post treatment scores.

Table 3. Importance values for predictor variables for Tf-CBT and EMDR.

	Importance	Importance
	Tf-CBT ($N = 150$)	EMDR ($N = 75$)
PHQ-9 _{pre}	0.60	0.95
GAD-7 _{pre}	0.40	0.30
WSAS _{pre}	0.81	0.28
Gender (female)	0.82	0.51
Age	0.91	0.30
LTC	0.30	0.53
Disability	0.27	0.28
Employment	1.00	0.28
Medication	0.34	0.95

Note. Importance = sum of model weights. An importance of 0.80 is used as cutoff to differentiate between important and not important variables. Tf-CBT = trauma-focused cognitive behavioral therapy; EMDR = eye-movement desensitization and reprocessing; PHQ-9 = patient health questionnaire; GAD-7 = Generalized Anxiety Disorder; WSAS = work and social adjustment scale; LTC = lifelong medical condition (e.g., diabetes, heart disease, chronic pain, arthritis, etc.); Medication = “prescribed antidepressant medication” coded as 0.5 with “no prescribed antidepressant medication” as reference category; Employed = “employed” coded as 0.5 with “unemployed” as reference category. WSAS = work and social adjustment scale; Gender = female coded as 0.5.

Table 4. Separate regression models for EMDR and Tf-CBT with treatment specific predictors and PHQ-9 as outcome variable.

		<i>B</i>	<i>SE B</i>	β	<i>t</i>
EMDR (<i>N</i> = 75)	Intercept	8.78	0.71	0.00	12.42***
	PHQ-9	0.44	0.12	0.40	2.77**
	Medication	4.40	1.59	0.29	3.77***
Tf-CBT (<i>N</i> = 150)	Intercept	9.83	0.50	0.00	19.41***
	WSAS	0.24	0.06	0.32	4.30***
	Employed	- 4.99	1.19	- 0.33	- 4.21***
	Age	- 0.10	0.04	- 0.17	- 2.51*
	Gender	- 2.09	1.08	- 0.14	- 1.93

Note. Tf-CBT = trauma-focused cognitive behavioral therapy; EMDR = eye-movement desensitization and reprocessing; PHQ-9 = Patient Health Questionnaire; Medication = “prescribed antidepressant medication” coded as 0.5 with “no prescribed antidepressant medication” as reference category; Employed = “employed” coded as 0.5 with “unemployed” as reference category. WSAS = work and social adjustment scale; Gender = female coded as 0.5.

^A *N* = 75. ^B *N* = 150.

*** *p* < 0.001, * *p* < 0.05.

Table 5. Testing the utility of the PAI approach with regard to observed scores and reliable change

		Optimal	Sub-Optimal
		<i>n</i> (%)	<i>n</i> (%)
Reliable change	yes	78 (62.90)	34 (33.66)
	no	46 (37.10)	67 (66.34)

Note. PAI = Personalized advantage index; Reliable change index ≥ 6 for the PHQ-9 (Richards & Borglin, 2011); Optimal = treatment condition with a predicted benefit for the individual patient; Sub-optimal = treatment condition which is predicted as being less effective for the individual patient.

Appendix

Table A1. Correlations of baseline variables

Variables	1	2	3	4	5	6	7	8	9
1. PHQ-9 _{pre}	-								
2. GAD-7 _{pre}	0.83***	-							
3. WSAS _{pre}	0.73***	0.66***	-						
4. Age	0.07	0.06	0.08	-					
5. Gender	-0.02	-0.01	-0.08	-0.09	-				
6. LTC	0.00	0.05	0.10	0.04	-0.03	-			
7. Disability	-0.01	0.04	0.03	0.12	-0.03	-0.02	-		
8. Employment	-0.40***	-0.40***	-0.41***	-0.11	0.28***	-0.10	-0.12	-	
9. Medication	0.44***	0.36***	0.46***	0.16*	-0.08	0.01	-0.09	-0.24***	-

Note. PHQ-9 = patient health questionnaire; GAD-7 = Generalized Anxiety Disorder; WSAS = work and social adjustment scale; LTC = lifelong medical condition (e.g., diabetes, heart disease, chronic pain, arthritis, etc.); Medication = “prescribed antidepressant medication” coded as 0.5 with “no prescribed antidepressant medication” as reference category; Employed = “employed” coded as 0.5 with “unemployed” as reference category; Gender = female coded as 0.5.

*** $p < 0.001$, * $p < 0.05$.