Psychological Medicine

cambridge.org/psm

Advancing the personalized advantage index (PAI): a systematic review and application in two large multi-site samples in anxiety disorders

Charlotte Meinke[1], Silvan Hornstein[1], Johanna Schmidt[2], Volker Arolt[3], Udo Dannlowski[3], Jürgen Deckert[4], Katharina Domschke[5,6], Lydia Fehm[1], Thomas Fydrich[1], Alexander L. Gerlach[7,8], Alfons O. Hamm[9], Ingmar Heinig[10], Jürgen Hoyer[10], Tilo Kircher[11], Katja Koelkebeck[12,13], Thomas Lang[14,15], Jürgen Margraf[16], Peter Neudeck[17], Paul Pauli[18], Jan Richter[9,19], Winfried Rief[20], Silvia Schneider[16], Benjamin Straube[11], Andreas Ströhle[21], Hans-Ulrich Wittchen[22], Peter Zwanzger[22,23], Henrik Walter[24], Ulrike Lueken[1,4,6], Andre Pittig[25] and Kevin Hilbert[1,26]

Original Article

Cite this article: Meinke C *et al* (2024). Advancing the personalized advantage index (PAI): a systematic review and application in two large multi-site samples in anxiety disorders. *Psychological Medicine* 1–13. https://doi.org/10.1017/S0033291724003118

Received: 12 April 2024
Revised: 19 September 2024
Accepted: 30 October 2024

Keywords:
anxiety disorders; cognitive behavioral therapy; machine-learning; personalized advantage index; precision medicine; precision psychotherapy

Corresponding author:
Charlotte Meinke;
Email: charlotte.meinke@hu-berlin.de

Abstract

Background. The Personalized Advantage Index (PAI) shows promise as a method for identifying the most effective treatment for individual patients. Previous studies have demonstrated its utility in retrospective evaluations across various settings. In this study, we explored the effect of different methodological choices in predictive modelling underlying the PAI.

Methods. Our approach involved a two-step procedure. First, we conducted a review of prior studies utilizing the PAI, evaluating each study using the Prediction model study Risk Of Bias Assessment Tool (PROBAST). We specifically assessed whether the studies adhered to two standards of predictive modeling: refraining from using leave-one-out cross-validation (LOO CV) and preventing data leakage. Second, we examined the impact of deviating from these methodological standards in real data. We employed both a traditional approach violating these standards and an advanced approach implementing them in two large-scale datasets, PANIC-net ($n = 261$) and Protect-AD ($n = 614$).

Results. The PROBAST-rating revealed a substantial risk of bias across studies, primarily due to inappropriate methodological choices. Most studies did not adhere to the examined prediction modeling standards, employing LOO CV and allowing data leakage. The comparison between the traditional and advanced approach revealed that ignoring these standards could systematically overestimate the utility of the PAI.

Conclusion. Our study cautions that violating standards in predictive modeling may strongly influence the evaluation of the PAI's utility, possibly leading to false positive results. To support an unbiased evaluation, crucial for potential clinical application, we provide a low-bias, openly accessible, and meticulously annotated script implementing the PAI.

Introduction

A wide range of effective treatments exists for most mental disorders, encompassing various forms of psychotherapy, pharmacotherapy, and neuromodulation. Despite each of these treatments exhibiting medium to large effect sizes on average (e.g. Brunoni et al., 2017; Carpenter et al., 2018; Cipriani et al., 2018; Cuijpers et al., 2023), there is a considerable heterogeneity in treatment effects. This heterogeneity is most evident in substantial proportions of patients showing non-response across different treatment types and disorders (e.g. Fitzgerald, 2020; Loerinc et al., 2015; Papakostas & Fava, 2009). Additionally, direct evidence of heterogeneity in treatment effects has been observed in various mental disorders for both pharmacotherapy and psychotherapy (e.g. see Herzog & Kaiser, 2022; Kaiser et al., 2020; Plöderl & Hengartner, 2019). Following the concept of precision mental health care, the considerable heterogeneity in treatment effects demands individually tailored treatment selection strategies based on empirical evidence for patient stratification. Therefore, recent endeavors have been directed towards creating methods for personalized treatment selection. The Personalized Advantage Index (PAI), introduced by DeRubeis et al. (2014), is one such method for identifying the most

suitable treatment for an individual patient. The PAI is a single score indicating the more promising treatment option for an individual patient by comparing the expected post-treatment severity under each treatment. These predictions of post-treatment severity rely on predictive models predominantly utilizing sociodemographic and clinical predictor variables.

To date, several studies have evaluated the PAI retrospectively, examining whether the PAI would have been useful to guide treatment selection. This has been done by comparing the post-treatment severity of patients who received their optimal treatment (according to the PAI) with those of patients who received their nonoptimal treatment. However, inappropriate analytical choices which might bias the studies' results are common in predictive modelling (e.g. Meehan et al., 2022; Meinke, Lueken, Walter, & Hilbert, 2024), which serves as the foundation of the PAI. Therefore, we examined the validity of the predictive modelling approaches employed and their impact on the evaluation of the PAI.

We adopted a two-step methodology. In the first step, we conducted a systematic review to provide an overview of prior studies using the PAI. Each study was evaluated using the Prediction model study Risk Of Bias Assessment Tool (PROBAST), with a specific focus on adherence to two predictive modeling standards: refraining from leave-one-out cross-validation (LOO CV) and preventing data leakage. Both standards get relevant during internal cross-validation (CV), where a dataset is split into training and test set. LOO CV involves leaving one subject out as the test set while using the remaining subjects for training, repeating this process for each subject. It should be avoided as it is a less precise estimator of the model performance on unseen data compared to other CV-schemes (Varoquaux, 2018). Data leakage occurs when test set data are utilized for training, leading to an overestimation of model performance since this data would not be available in a real-world scenario.

Altogether, our expectation was that the PAI would prove beneficial in most studies, showing a small effect size for patients who received their predicted optimal treatment compared to those receiving their nonoptimal treatment (hypothesis 1). In terms of risk of bias, we anticipated that most studies would exhibit a high risk of bias according to PROBAST and a medium risk of bias concerning the applied CV-scheme and the occurrence of data leakage (hypothesis 2). Additionally, considering the impact of these two critical methodological characteristics on model performance (Moons et al., 2019; Varoquaux, 2018), we hypothesized that the effect size would increase as bias increases (hypothesis 3).

Subsequently, in the 2nd step, we explored the impact of adhering to the two standards described above in two original datasets. We employed both a traditional approach lacking these standards and an advanced approach implementing them. This analysis was carried out separately in two large-scale multicentric randomized controlled trials focusing on anxiety disorders (PANIC-net, Protect-AD). For both datasets, we expected to observe a higher effect of the PAI in the traditional approach compared to the advanced, less biased approach (hypothesis 4). For the advanced approach, we expected a significant but relatively small effect, as the treatment options were quite similar in both datasets. In such cases, the PAI's utility is expected to be lower (DeRubeis et al., 2014) but likely still present, as demonstrated in previous studies on similar treatment options (Bruijniks et al., 2022; Friedl, Berger, Krieger, Caspar, & Holtforth, 2020a). In addition, we expected that the effect of the PAI will increase when using a Random Forest Regressor, which is a more sophisticated algorithm compared to the originally applied ridge regression (hypothesis 5).

Systematic review

Methods

Search strategy and study selection

Our systematic review was preregistered in PROSPERO (CRD42022361290). The databases Scopus, PubMed and psycArticles were searched on August 1, 2024, using the search term 'personalized advantage index'. Reference lists were checked for additional relevant literature. Given the relatively small number of studies found, we submitted all findings to a full-text review (KH & CM). The following inclusion criteria were applied: (1) comparison of at least two treatment alternatives, (2) calculation of the PAI for these treatment alternatives, (3) evaluation of the PAI by comparing post-treatment severity values between patients who received their optimal *v.* nonoptimal treatment according to PAI recommendation, (4) empirical study with original data, (5) publication in a peer-reviewed journal. In (3), we focused on studies comparing post-treatment severity values to ensure comparability of effect sizes. Studies that did not meet these criteria or lacked sufficient information for judgement were excluded. In the case of disagreement about study inclusion, there was a discussion until consensus was reached. Despite our interest in the application of the PAI in mental disorders, we did not set any inclusion criteria regarding health conditions, as our focus was primarily on the PAI's methodological implementation.

Data extraction

As main outcome variables of interest, we extracted the mean PAI, the mean post-treatment severity difference between patients who received their PAI-indicated optimal *v.* nonoptimal treatment, and Cohen's *d* for this difference – both for the complete sample and a subsample of patients with the largest PAIs (e.g. top 50%, exact subsample definition depending on the original study). Additionally, we extracted authors and year of the study, sample size, diagnosis and treatment options, post-treatment severity measure, type of feature selection approach, type of outcome prediction approach, CV-scheme, and most relevant features (KH & CM).

Risk of bias assessment

We assessed risk of bias using PROBAST (Moons et al., 2019) which has been developed specifically for predictive modeling and has been applied in comparable studies (Navarro et al., 2021; Vieira, Liang, Guiomar, & Mechelli, 2022). Again, the rating was performed by two authors (CM & KH), with discrepancies resolved through discussion. PROBAST comprises 20 signaling questions, rated with yes, no, or no information, designed to evaluate bias in total and across four domains: participants, predictors, outcome, and analysis. In line with the PROBAST guidelines, we tailored the rating by adding two additional questions: 4.8.1 'What is the extent of risk of bias introduced by the cross-validation procedure?' (CV-scheme), 4.8.2 'What is the extent of risk of bias introduced by (not) integrating preprocessing steps into the CV? (data leakage)'. While question 4.8 rates the risk of bias introduced by not accounting for overfitting, our additional questions addressed the risk of bias from two specific sources. Moreover, we evaluated these questions on a more fine-grained 3-step scale ranging from low over moderate to high risk

https://doi.org/10.1017/S0033291724003118 Published online by Cambridge University Press

of bias. For instance, risk of bias introduced by data leakage (question 4.8.2) was rated as low risk of bias if all pipeline step were performed within CV, as medium if imputation and or scaling occurred outside CV, and as high if feature selection was done outside CV (see Table A1 in the supplement for more details).

Risk of bias and its relation to Cohen's d

We aimed to assess whether Cohen's *d* increases with the risk of bias introduced from the two procedures specifically assessed: CV-scheme and data leakage. As the number of subanalyses reporting Cohen's *d* in total ($n = 21$) and per group (e.g. 3 analyses with a medium bias rating regarding data leakage) was too low to conduct a statistical analysis such as an ANOVA, we decided to explore the association descriptively, comparing the mean and the distribution of effect sizes between levels of bias visually.

Results

Study characteristics and Cohen's d

We initially found 38 articles. After eliminating duplicates and adding studies from reference lists, a remainder of 36 articles was reviewed (see Fig. 1 for the PRISMA flow diagram). The final sample consisted of 19 articles with 25 analyses and $n = 5699$ patients. Table 1 provides an overview of the extracted outcomes and study characteristics. Overall, the PAI was most frequently calculated for patients with mental disorders, with only one study addressing patients with bodily constraints (urinary contingency, Loohuis et al., 2022). Most patients with mental disorders suffered from unipolar depressive disorder or depressive symptoms (10/19). Treatment options compared were primarily different types of psychotherapy or interventions, with cognitive behavioral therapy (CBT) as the most frequent type (Ahuvia, Mullarkey, Sung, Fox, & Schleider, 2023; Cohen, Kim, Van, Dekker, & Driessen, 2020; Deisenhofer et al., 2018; Hautmann et al., 2023; Huibers et al., 2015; Keefe et al., 2021; Lopez-Gomez et al., 2019; Schwartz et al., 2021; van Bronswijk et al., 2021). Furthermore, some studies compared variations of the same type of psychotherapy, differing in session frequency, thematic focus, or the integration of online treatment elements (Bremer et al., 2023; Bruijniks et al., 2022; Friedl et al., 2020a, 2020b; Held et al., 2023; Hoeboer et al., 2021; Senger et al., 2021), while others compared antidepressants to CBT or placebo (DeRubeis et al., 2014; Webb et al., 2019). Not all examined papers provided effect sizes quantifying the potential benefits of patients who received their PAI-indicated treatment. Those who did most often reported small (14/21 analyses) to medium (4/21 analyses) effect sizes, with a mean of Cohen's $d = 0.32$ and a range between 0.09–0.57. This was in accordance with hypothesis 1. For most studies, effect sizes were considerably larger in patients with high PAIs compared to the entire sample (see Fig. 2a).

Risk of bias and its relation to Cohen's d

As anticipated in hypothesis 2, all reviewed analyses exhibited a high overall risk of bias. The specific rating (Table B1) and a visual depiction (Fig. B1) can be found in supplement B. The high overall risk of bias was mainly based on a wide range of inappropriate choices in the analysis domain. First, most analyses had an insufficient sample size (20/25; question 4.1). Second, many analyses (14/25) dealt with missing values in an inappropriate way, for instance, imputing missing post-treatment severity values based on baseline data (11/25), thereby introducing label noise. More appropriate methods involve utilizing the last observation of symptom severity or, in its absence, excluding patients with missing outcome values. Third, many studies did not sufficiently account for model overfitting (4.8), using strongly biased CV-schemes (4.8.1) and/or allowing data leakage (4.8.2). Moreover, as anticipated in hypothesis 3, the descriptive analysis suggested that the effect size – indicating the potential utility of the PAI – diminishes as the risk of bias decreases, whether due to the CV-scheme or data leakage (see Fig. 2b).

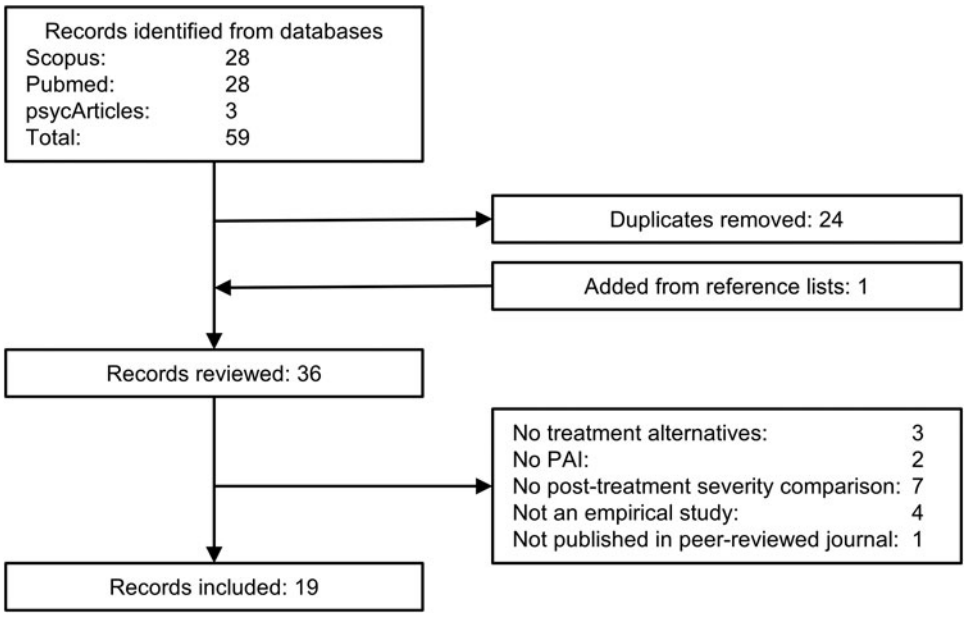

Figure 1. PRISMA-Flowchart.

https://doi.org/10.1017/S0033291724003118 Published online by Cambridge University Press

Table 1. Study characteristics

First Author, Year	Sample size	Diagnosis/Target Group	Treatment options	Post-treatment severity measure	Type of feature selection approach	Type of outcome prediction approach	CV	Mean absolute PAI	Mean diff in post-treatment severity (optimal *v.* nonoptimal)	Cohen's *d*
Ahuvia et al. (2023)	996	adolescents with depressive symptoms	single session intervention: Project Personality *v.* Action Brings Change Project	CDI-2-SF	none, but varying feature sets	linear regression, regularized linear regression, random forest, k-nearest neighbor model	holdout-set	0.16	0.06	
Bremer et al. (2023)	105	PTSD related to childhood abuse	STAIR-EMDR *v.* EMDR	CAPS-5	(random forest recursive partitioning OR Elastic Net) AND subsequent bootstrapped backward elimination	linear regression	10-fold	3.1	0.25	
Bruijniks et al. (2022)	200	depression	weekly *v.* twice-weekly CBT or IPT	BDI-II	random forest recursive partitioning ('mobforest') and subsequent bootstrapped backward elimination ('bootstepAIC')	linear regression	5-fold	4.93	0.37	
Cohen et al. (2020)	167	depression	CBT *v.* PDT	HAM-D	(random forest recursive partitioning ('mobforest') OR Elastic Net Regularized Regression ('glmnet') OR Bayesian Additive Regression Trees ('bartMachine')) AND subsequent bootstrapped backward elimination ('bootstepAIC')	linear regression	1000 * 10-fold	1.6	0.21	
Deisenhofer et al. (2018)	225	PTSD	Tf-Cbt *v.* EMDR	PHQ-9	Genetic Algorithm ('glmulti')	linear regression	leave-one-out	2.49	3.03	0.4
DeRubeis et al. (2014)	154	depression	Paroxetine *v.* CBT	HRSD	previous analyses on the same sample	generalized linear regression	leave-one-out	4.2	1.78	0.28
Friedl et al. (2020a)	123	depression	CBT *v.* CBT-EE	BDI-II	Bayesian Model Averaging	linear regression	leave-one-out	1.35		
Friedl et al. (2020b)	245	depression	blended treatment *v.* TAU	PHQ-9	Bayesian Model Averaging	linear regression	leave-one-out	2.33		
Hautmann et al. (2023)	110	parents of children with ADHD/ODD	self-help parent training: behavioural *v.* nondirective	FBB-ADHS, FBB-SSV	single moderator analyses and subsequent multiple moderator analyses	linear regression, regularized linear regression, random forest, k-nearest neighbor model	none	ADHD: 0.35; ODD: 0.54		
Held et al. (2023)	747	veterans with PTSD	3-week *v.* 2-week intensive PTSD program	PCL-5	random forest recursive partitioning	linear regression	leave-one-out	2.65	2.29	0.13
Hoeboer et al. (2021)	149	PTSD	PE & iPE *v.* STAIR	CAPS-5, PCL-5	random forest iterative comparison with random probes ('Boruta') and subsequent bootstrapped backward elimination ('bootstepAIC')	linear regression	leave-one-out	CAPS-5: 4.02; PCL-5: 4.69	CAPS-5: 0.55; PCL-5: 0.47	
Huibers et al. (2015)	134	depression	CT *v.* IPT	BDI-II	domain-wise hierarchical multiple linear regressions	linear regression	leave-one-out	8.9	6	0.51

Keefe et al. (2021)	156	borderline personality disorder	DBT *v.* GPM	GSI SQL	random forest recursive partitioning ('mobforest') and subsequent bootstrapped backward selection ('bootstepAIC')	linear regression	(1000 *) 10-fold	initial model (feature selection outside CV): 0.36; less biased model (feature selection inside CV): 0.22		
Loohuis et al. (2022)	262	urinary incontinence	App-based treatment *v.* care as usual	UISF	initial selection based on previous studies and sufficent variability; stepwise, backward elimination in multiple linear regressions	linear regression	5-fold	0.99	1.19	
Lopez-Gomez et al. (2019)	128	depression	Group-based IPPI-D *v.* Group-based CBT	BDI-II	random forest recursive partitioning ('mobforest') and Elastic Net	linear regression	10-fold	2.3	0.24	
Schwartz et al. (2021)	1379	transdiagnostic	CBT *v.* Psychodynamic therapy	BSI–GSI	random forest recursive partitioning ('mobforest') and subsequent bootstrapped backward elimination ('bootstepAIC')	linear regression	holdout-set	3.6%*	0.09	
Senger et al. (2021)	203	persistent somatic symptoms	CBT *v.* Encert	SOMS-7 T	random forest recursive partitioning ('mobforest') and Elastic Net Regularized Regression ('glmnet')	regression	leave-one-out	5	4.11	0.278
van Bronswijk et al. (2021); subanalysis STEPd	STEPd: 151, FreqMesh: 200, both: 351	depression	CT *v.* IPT	BDI-II	random forest recursive partitioning ('mobforest') and subsequent bootstrapped backward elimination ('bootstepAIC')	Elastic Net	5-fold, out-of-sample	STEPd: 6.53; FreqMech: 2.81; STEPd to FreqMech: 2.1; FreqMech to STEPd: 3.25	STEPd: 0.57; FreqMech: 0.2; STEPd to FreqMech: 0.16; FreqMech to STEPd: 0.27	
Webb (2019)	216	depression	Sertraline *v.* Placebo	HRSD	(random forest recursive partitioning ('mobforest') OR Elastic Net Regularized Regression ('glmnet') OR Bayesian Additive Regression Trees ('bartMachine')) AND subsequent bootstrapped backward elimination ('bootstepAIC')	regression	1000 * 10-fold	3.4	1.99	0.29

Abbreviations diagnosis: ADHD, attention-deficit/hyperactivity; ODD, oppositional defiant disorder; PTSD, post-traumatic stress disorder; Abbreviations treatment options: Blended-treatment = Face2Face CBT with internet-based CBT elements, CBT, Cognitive Behavioral Therapy; CBT – EE, Cognitive Behavioral Therapy with integrated exposure and emotion-focused elements; CFD, Person-centered counselling for depression; CT, Cognitive Therapy; DBT, dialectical behavior therapy; EMDR, Eye movement desensitization and reprocessing; Encert, CBT enriched with emotion regulation training; GPM, general psychiatric management; iPE, intensified Prolongued Exposure; IPPI-D, Integrative Positive Psychological Intervention for Depression; IPT, Interpersonal Psychotherapy; PDT, Psychodynamic Therapy; PE, Prolongued Exposure; STAIR, skills training; TAU, eatment as usual; tf-CBT, Trauma-focused Cognitive Behavioral Therapy; Abbreviations severity measures: BDI-II, Beck Depression Inventory II; BSI-GSI, Brief Symptom Inventory Global Severity Index; CAPS-5, Clinician-Administered PTSD Scale for DSM-5; CDI-2-SF, Children's Depression Inventory 2nd Edition Short Form; FBB-ADHS, Fremdbeurteilungsbogen für Aufmerksamkeitsdefizit-/Hyperaktivitätsstörung [rating scale for ADHD]; FBB-SSV, Fremdbeurteilungsbogen für Störungen des Sozialverhaltens [rating scale for ODD]; HAM-D, Hamilton Rating Scale for Depression; HRSD, Hamilton Rating Scale for Depression; PCL-5, PTSD checklist for DSM-5; PHQ-9, Patient Health Questionnaire 9; SOMS-7T, Screening for Somatoform Disorders-7T; UISF, Urinary Incontinence Short Form.

Note: Please note that the mean absolute PAI and the mean difference in post-treatment severity needs to be interpreted considering the study-specific severity measure. * = This study did not focus on raw severity but on reduction in severity from pre- to post-treatment in %.

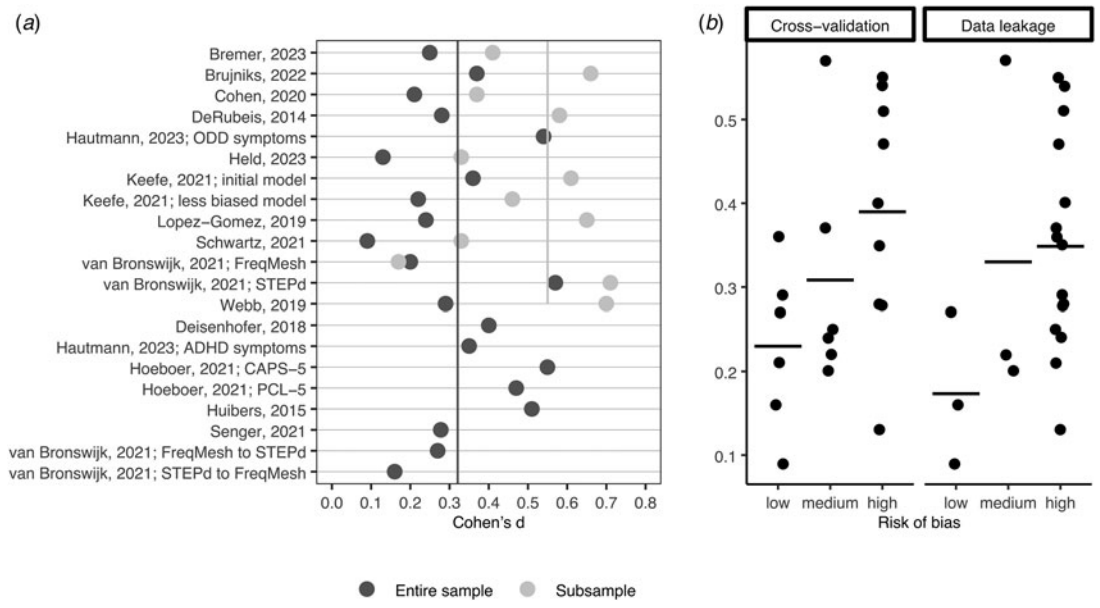

Figure 2. Cohen's *d* in relation to sample and risk of bias.
Note: Only analyses that reported Cohen's *d* for the difference in post-treatment symptom severity between patients who received their optimal *v.* nonoptimal treatment are depicted ($n = 21$). a: The vertical lines represent the mean Cohen's *d* per group (entire sample *v.* subsample). b: The horizontal lines represent the mean Cohen's *d* per each level of risk of bias.

Interim discussion

With a mean Cohen's *d* of 0.32, the systematic review indicates that the PAI might be a useful tool for treatment selection across many settings. However, a more thorough exploration of the methodological approaches using PROBAST revealed that all studies suffered from a high risk of bias. Furthermore, a more comprehensive examination of two characteristics that commonly contribute to significant bias, namely the employed CV-scheme and the occurrence of data leakage, suggests a potential association with the magnitude of Cohen's *d*. To explore the impact of these two characteristics further, we compared a traditional approach, suffering from these two characteristics, and an advanced approach, free from these pitfalls, in the subsequent empirical investigation.

Empirical study

Methods

Datasets

We analyzed data from two German multicenter randomized controlled trials focusing on anxiety disorders (PANIC-Net and Protect-AD). In PANIC-Net, patients with a diagnosis of panic disorder with agoraphobia received exposure-based CBT and were randomized to three treatment conditions: (i) therapist-guided exposure, (ii) self-guided exposure without therapist guidance, and (iii) wait-list control group. Both active treatment conditions exclusively differed in the way of performing the five exposure sessions integral to the therapy. In the therapist-guided exposures condition, patients completed one exposure with the therapist and were then asked to perform two additional exposures independently before the subsequent session. In contrast, patients in the therapist-unguided exposure condition conducted all exposures independently after thorough preparation by their therapist. In both active treatment conditions, patients improved significantly from pre to post with large effect sizes in terms of all primary outcomes, including the Hamilton Anxiety Rating Scale (HAM-A), which we will focus on here for comparability with Protect-AD. Full details on the trial and main results can be found elsewhere (Gloster et al., 2011). In Protect-AD, patients with a primary diagnosis of panic disorder, agoraphobia, social anxiety disorder, or multiple specific phobias received exposure-based CBT and were randomized to two treatment conditions: (i) temporally intensified exposure with six sessions delivered within 2 weeks, and (ii) standard non-intensified exposure with the same amount of exposure delivered as one session per week. Again, both treatment conditions were similar regarding all other treatment characteristics. Similar to PANIC-Net, patients in both treatment conditions showed substantial improvements with large effect sizes in the HAM-A, which was the primary outcome. Full details on the trial and main results can be found elsewhere (Heinig et al., 2017; Pittig et al., 2021).

Patients

In PANIC-net, data from the waiting-list condition were omitted. For both datasets, only patients with the primary outcome measure available at post-treatment were included in our analysis. This included patients who completed the treatment until post-assessment as well as those who underwent the post-assessment despite premature dropout, were included. This resulted in a final sample of $n = 261$ patients for PANIC-Net ($n = 119$ therapist-unguided and $n = 142$ therapist-guided) and $n = 614$ patients for Protect-AD ($n = 307$ intensified and $n = 307$ non- intensified).

Predictor and outcome variables

In both datasets, the PAI relied on post-treatment symptom severity as evaluated with the HAM-A and assessed through the Structured Interview Guide for the Hamilton Anxiety Scale (SIGH-A; Shear et al., 2001). Sociodemographic, diagnostic and clinical questionnaire data available at pre-treatment were used as predictor variables, including, for example, age, sex, HAM-A baseline severity, the clinical global impression scale (CGI;

https://doi.org/10.1017/S0033291724003118 Published online by Cambridge University Press

Busner & Targum, 2007), and the brief symptom inventory (BSI; Derogatis, 1993) with its global indices and subscales. Variables were overlapping between both datasets to a considerable degree but were not completely similar. All initially included variables and their sample statistics are available in Supplement C.

Machine-learning pipelines

We employed two different machine-learning approaches: a traditional approach which was very similar to early PAI implementations such as in DeRubeis et al. (2014) and Huibers et al. (2015), and an advanced approach which was based on more recent implementations (compare Schwartz et al., 2021) and was characterized by refraining from LOO CV and avoiding data leakage. A visualization of both approaches is presented in Fig. 3. Both approaches were separately applied on PANIC-Net and Protect-AD. They consisted in similar and partially equal steps but differed in the general architecture of their pipelines. In the traditional approach, all preprocessing steps, including dealing with missing values, excluding and selecting features, were performed on the entire dataset. Only afterwards, the dataset was split into training and test set within LOO CV. Thus, the procedure introduced data leakage as information from the test set was utilized to train the model. In the advanced approach, data leakage was avoided by conducting these steps only on the training set. Moreover, a 5-fold CV with 100 repetitions was employed, being more robust than the LOO CV used in the traditional approach (Varoquaux, 2018). The second key distinction was the way of generating predictions of symptom severity for the two treatment options. In the traditional approach, a single model was employed to predict outcomes for both treatment options by incorporating predictor × treatment interaction terms as independent variables in a linear regression. To predict outcomes for both treatment options, two distinct datasets were utilized, differing in the treatment as predictor variable and treatment-specific interaction terms. In the advanced approach, a distinct model was trained for each treatment, after having separately employed feature selection. Thus, predictions for each treatment option could easily be generated using these two models.

Dealing with missing values

The procedure was the same for both approaches. Initially, features with more than 30% missing values were excluded. Subsequently, missing data in binary and categorical features were imputed with their mode. Categorical features were then one-hot encoded and the resulting binary features were recoded to 0.5 and −0.5. Following this, missing values in dimensional features were imputed using Multivariate Imputation by Chained Equations (MICE; van Buuren & Groothuis-Oudshoorn, 2011).

Feature exclusion

The initial feature exclusion, taking place after dealing with missing values, was the same for both approaches. First, features with no variance and binary features with less than 10% percent of patients in one category were excluded. Then, the similarity between features was examined, calculating Pearson correlation and Jaccard similarity for dimensional and binary features, respectively. If two or multiple features had a correlation/similarity > 0.75, the one showing the highest mean correlation/similarity with the rest of the features was removed. The procedure was repeated until no correlation/similarity > 0.75 was observed.

Feature selection

Besides the embedding in the machine-learning pipeline, the approaches differed in the type of feature selection. In the traditional approach, a stepwise feature selection, similar to the one reported in Huibers et al. (2015) was employed on the whole dataset, consisting of three rounds of building a linear regression model. In each round, only those predictors whose beta coefficient p values underscored a certain threshold were kept and given to the next round, thereby iteratively reducing the number of predictors. The threshold applied from the first to the third round were 0.2, 0.1, 0.05. In the advanced approach, feature

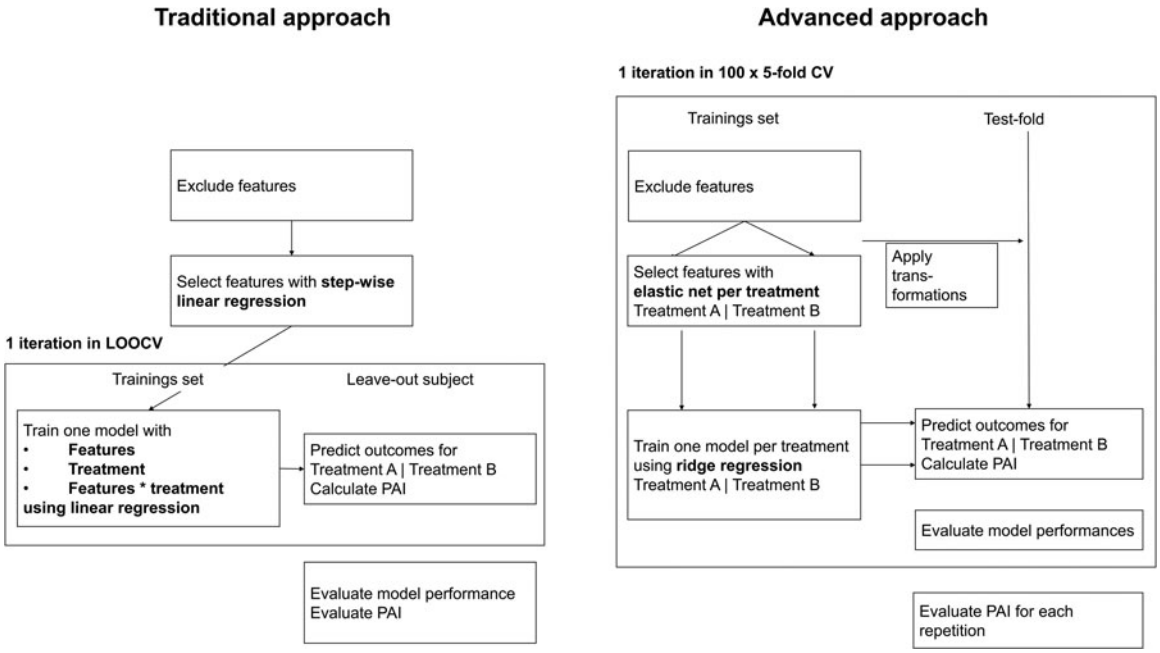

Figure 3. Graphical depiction of the traditional and the advanced approach.

https://doi.org/10.1017/S0033291724003118 Published online by Cambridge University Press

selection was implemented with Elastic Net (Zou & Hastie, 2005), which is a penalized linear regression.

PAI calculation and evaluation

The PAI is commonly computed based on the predictions of post-treatment severity for both treatment options compared. More specifically, here, the PAI was calculated as the prediction for the treatment factually received (factual prediction) minus the prediction for the treatment factually not received (counterfactual prediction; PAI = factual prediction – counterfactual prediction, DeRubeis et al., 2014; Huibers et al., 2015). Using this formula, a positive PAI indicated that a patient had received their nonoptimal treatment, as the predicted post-treatment severity was lower in the counterfactual treatment. In contrast, a negative PAI indicated that a patient had received their optimal treatment.

To evaluate whether the PAI would have been useful to guide treatment selection, we tested whether post-treatment severity scores of patients who received their optimal treatment were smaller than those of patients who received their nonoptimal treatment (DeRubeis et al., 2014; Huibers et al., 2015), using an independent one-sided t test. Cohen's d for the difference in mean severity was calculated. Moreover, similar to previous PAI analyses, this analysis was performed both for the entire sample and for the 50% of patients with the largest absolute PAI (Delgadillo & Duhne, 2020; DeRubeis et al., 2014; Huibers et al., 2015; Schwartz et al., 2021; van Bronswijk et al., 2021). In addition, to evaluate the validity of the prediction models underlying the PAI, correlations, mean absolute error (MAE) and root mean square error (RMSE) were calculated.

Further exploratory analyses

Given the lack of a notable difference between patients receiving their optimal *v.* nonoptimal treatment in the advanced approach, we conducted two further exploratory analyses. In exploratory analysis I, we used a Random Forest (Breiman, 2001) regressor instead of ridge regression to predict post-treatment severity. This approach was driven by Random Forests' ability to handle non-linear associations between predictor and outcome variables and their strength with tabular data (Grinsztajn, Oyallon, & Varoquaux, 2022). In exploratory analysis II, we used two composite scores as treatment outcomes instead of relying solely on the HAM-A: (1) a symptom index, based on HAM-A, CGI, DSM-5 Cross-D (Lebeau et al., 2012), and a symptom severity questionnaire score depending on the primary diagnosis, and (2) a functioning index based on the World Health Organization Disability Schedule (WHODAS 2.0; Üstün, Kostanjsek, Chatterji, & Rehm, 2010), the EuroQOL five-dimensional measure of health status (EQ-5D; Rabin & de Charro, 2001) and the global assessment of functioning (GAF; APA & Association, 2013). This approach was motivated by the assumption that a composite score could more accurately reflect treatment outcomes, as well as by prior studies that have utilized similar metrics (e.g. Pittig et al., 2023). Exploratory analysis II was restricted to Protect-AD, as the necessary variables had not been assessed in PANIC-net. Supplement D provides further details on the hyperparameters used in the random forest regressor and the calculation of the composite scores.

In addition, we conducted a further analysis to ensure that the different results of the traditional and advanced approach were truly due to the targeted methodological choices (LOO CV and data leakage) and not to other differences implemented, such as the choice of feature selector technique (stepwise linear regression *v.* elastic net). To address this, we implemented an extra pipeline, referred to as the 'mixed approach', which followed all the steps of the traditional approach but replaced LOO CV with 100 * 5-fold CV and avoided data leakage.

Results

Relevant metrics to evaluate the PAI`s utility and the performance of underlying models are reported in Table 2. In the traditional approach, patients receiving their optimal treatment had a significantly lower post-treatment severity than patients receiving their nonoptimal treatment, with a Cohen's d of a small to medium effect (PANIC-net: 0.41, Protect-AD: 0.25). However, this difference was not evident in the advanced approach. A Welch t test, chosen for its robustness against violations of equal variance assumptions (Rasch, Kubinger, & Moder, 2011) occurring in certain repetitions and approaches, yielded identical results.

Notably, no discernible group differences emerged even after implementing several modifications to the advanced approach. These adjustments included employing a random forest regressor, optimizing hyperparameters, and employing composite scores of severity and functioning as alternative outcome measures. This was also true when focusing solely on the top 50% of patients with the largest absolute PAI (see Supplement E). The results for the mixed approach, which differed from the traditional approach only in the type of CV and the avoidance of data leakage, were similar to those of the advanced approach (see Supplement F). Regarding model performance, the traditional approach outperformed the advanced approach descriptively across several metrics, as expected given the occurrence of data leakage.

Interim discussion

In both datasets, the traditional approach, characterized by a high risk of bias, exhibited a notable difference between patients receiving optimal and nonoptimal treatments. This difference was not observed in the advanced approach and its various modifications or when modifying the traditional approach only in terms of CV and data leakage ('mixed approach'). This pattern suggests that LOO CV and data leakage might indeed produce false positive results, corroborating the results of our systematic review.

General discussion

One crucial step towards precision mental healthcare is an evidence-based patient stratification for treatment. The PAI is an increasingly used method to identify the most promising treatment among various options for individual patients. Here, we examined the impact of critical methodological choices when calculating the PAI, conducting both a systematic review and empirical investigations on data from two large-scale multicenter clinical trials. Our review raised awareness that most previous studies employing the PAI did not follow current predictive modelling standards such as refraining from LOO CV and preventing data leakage, amplifying the risk of bias. Furthermore, our empirical investigations provided a clear illustration that an approach with these characteristics is likely to overestimate the PAI`s utility. Specifically, it demonstrated a more favorable outcome for patients receiving their optimal *v.* nonoptimal treatment, a pattern not observed in the unbiased advanced approach. Thus, it remains open, whether the retrospectively positive evaluation of the PAI in most studies would also hold true when using a less biased approach.

https://doi.org/10.1017/S0033291724003118 Published online by Cambridge University Press

Table 2. PAI and model performance evaluation metrics in different machine-learning approaches

	PANIC-net				Protect-AD				Protect-AD symptom	Protect-AD function
Metric	trad (linear)	adv (ridge)	adv (ridge w hp)	adv (rf)	trad (linear)	adv (ridge)	adv (ridge w hp)	adv (rf)	adv (ridge)	adv (ridge)
t-test	$p < 0.001$	1/100	2/100	0/100	$p < 0.01$	23/100	12/100	6/100	0/100	6/100
Cohen's *d*	0.41	0.01 (0.1)	−0.02 (0.11)	−0.07 (0.1)	0.25	0.08 (0.08)	0.05 (0.08)	0.03 (0.06)	−0.04 (0.06)	0.02 (0.07)
Mean PAI	2.42	4.53 (0.29)	4.65 (0.32)	2.91 (0.16)	3.12	3.31 (0.14)	3.45 (0.16)	3.13 (0.12)	0.77 (0.1)	0.3 (0.01)
MAE	4.71	7.12 (0.69)	7.31 (0.66)	5.38 (0.56)	5.88	7 (0.48)	6.97 (0.49)	6.38 (0.41)	0.9 (0.08)	0.38 (0.03)
RMSE	6.14	8.88 (0.78)	9.12 (0.8)	6.87 (0.72)	7.42	8.97 (0.65)	8.98 (0.67)	8.04 (0.54)	1.22 (0.13)	0.49 (0.04)
Correlation	0.58	0.28 (0.11)	0.25 (0.12)	0.43 (0.12)	0.45	0.25 (0.08)	0.25 (0.08)	0.32 (0.07)	0.11 (0.12)	0.27 (0.08)

PAI, Personalized advantage index; MAE, mean absolute error; RMSE, root mean square error; trad, traditional; adv, advanced; linear, linear regression; ridge, ridge regression; w hp, with hyperparameter tuning; rf, random forest.
Notes: Please note that the values given for the *t* test differ for both approaches. For the traditional approach, the *p* value is given. For the advanced approaches, the proportion of repetitions in which the *t* tests got significant are given. The complete statistics for the *t* test in the traditional approach are: PANIC-net: optimal ($M = 11.23$, $s.d. = 6.72$), nonoptimal ($M = 14.26$, $s.d. = 7.91$), $t(259) = 3.31$, $p < 0.001$; Protect-AD: optimal ($M = 11.78$, $s.d. = 7.71$), nonoptimal ($M = 13.83$, $s.d. = 8.69$), $t(612) = 3.09$. All other values presented in the advanced approaches represent the mean values across 100 repetitions of 5-fold CV, accompanied by their corresponding standard deviations in brackets. In contrast, the values in the traditional approach are single values without means.

It should be noted that the negative evaluation of the PAI in our unbiased approaches does not question the utility of the PAI framework *per se*. Instead, it underscores that the PAI may be unsuitable in the specific conditions we examined, which were characterized by a high similarity between treatments. As mentioned in DeRubeis et al. (2014), the utility of the PAI is likely limited if treatments build on similar mechanisms. Here, in both datasets, the treatment options exhibited considerable similarity, differing in therapist-accompaniment and frequency of exposure sessions in PANIC-net and Protect-AD, respectively. Furthermore, in both datasets, treatment options did not differ in their effect on symptom reduction as measured with the HAM-A. While this does not rule out the possibility of treatment heterogeneity effects *per se*, it emphasizes the high similarity. Ideally, the PAI framework should be robust enough to detect the equality of treatment options itself by generating PAIs around 0. However, current predictions of post-treatment symptom severity lack sufficient precision. Consequently, random differences between predictions of symptom severity will consistently occur, resulting in PAIs unequal to 0.

Unveiling numerous methodological choices that heighten the risk of bias, our study prompts several recommendations for researchers utilizing the PAI in future investigations. Although we acknowledge that there have been partially notable advancements in the PAI's methodology in recent years, we would like to highlight several important points. First, the two methodological weaknesses characterizing the biased traditional approach, namely data leakage and LOO CV, should be avoided. It should be noted that the identification of data leakage as a pervasive issue in various PAI calculation approaches is not novel; it has been raised in reviews on personalized treatment selection (Cohen & DeRubeis, 2018; Kessler et al., 2017) and as a limitation in some of the included studies (e.g. Huibers et al., 2015; Senger et al., 2021; Webb et al., 2019). Despite this recognition, the majority of recent studies had continued to use approaches plagued by data leakage. Consequently, our paper aims to further raise awareness that data leakage is not only a negligible side effect but might jeopardize the meaningfulness of the findings in the studies. Moreover, to facilitate the implementation of a state-of-the-art approach without data leakage, we provided a consistent, modularized, and extensively documented Python-script on GitHub (https://github.com/Charlotte-Marie/PAI_Advanced_Approach). Researchers are invited to use it for the calculation of the PAI in their datasets.

Second, given that only very few studies had a sufficient sample size according to the PROBAST rating (question 4.1), future studies should employ larger samples to train the models underlying the PAI and to test the PAI's utility. Even though there is no straightforward formula for determining an adequate sample size in predictive modelling, various rules of thumbs, based on simulation studies, exist. While the PROBAST criterium requires a sample size 20 times the number of candidate predictors, others (Luedtke, Sadikova, & Kessler, 2019; Varoquaux, 2018), including also a recent preprint (Zantvoort et al., 2024) suggest that datasets should at least include several hundreds of patients per model/treatment option. As pointed out in Luedtke et al. (2019), such sample sizes can be achieved through various means, such as implementing large multi-centric clinical trials, utilizing data from observational trials, which might also be beneficial in terms of ecological validity, or pooling data across various trials. Additionally, large sample sizes can easily be obtained by using electronic health care records, as done by Schwartz et al. (2021) and Bauer-Staeb, Griffith, Faraway, and Button (2023).

https://doi.org/10.1017/S0033291724003118 Published online by Cambridge University Press

Besides the methodological issues that might generate a more stable estimation of the PAI's utility, several other developments might improve the prediction performance of the models underlying the PAI. First, so far, mainly sociodemographic and clinical variables have been used as predictor variables. However, several meta-analyses suggest that a wide array of other types of variables such as EEG, (f)MRI or heart-rate variability might have a similar or even higher predictive ability (e.g. see Choi & Jeon, 2020; Vieira et al., 2022; Watts et al., 2022). Thus, including these variables might leverage the precision of the PAI.

Second, previous studies have mainly used models from traditional statistics that do not consider interactions between variables unless explicitly specified, such as multiple linear regression. In contrast, machine-learning algorithms, such as random forest-based algorithms or support vector machine, have the capacity to account for these interactions, thereby potentially enhancing predictive performance and leading to more precise PAIs. Indeed, in our exemplary analysis, the substitution of ridge regression (penalized multiple linear regression) with random forest resulted in an enhanced model performance and more stable PAIs across repetitions of CV. In the studies included in our review, employing machine-learning algorithms instead of multiple linear regression for post-treatment severity prediction might have been particularly useful because these algorithms were used for the preceding feature selection. Thus, to fully exploit the features' potential in the final models, machine-learning algorithms should as well be employed for this step. One barrier that might have deterred previous researchers from using machine-learning algorithms as final models could be the perceived lower explainability. However, there is a wide range of comprehensible model-agnostic ways to understand the contribution each feature makes to a prediction, such as SHAP (Shapley Additive exPlanations) values (Lundberg & Lee, 2017; see Molnar, 2022 for an introduction).

Both our systematic review and empirical study have certain limitations. Regarding our systematic review, we would like to emphasize that our three-level risk of bias rating for data leakage (question 4.8.2) provides only a rough estimate. It primarily focuses on the specific step (e.g. imputation *v.* selection) that was incorrectly applied on the entire dataset, but ignores other factors which might influence the risk of bias as well. For instance, the risk of bias introduced when applying data imputation incorrectly on the entire dataset is also affected by the numbers of missing cases, both per variable and across variables. However, since such detailed information was unavailable in most studies, we were unable to incorporate these aspects into our rating.

Regarding our empirical study, we would like to stress that our comparison of a traditional and an advanced approach in two datasets provides suggestive but inconclusive evidence about the traditional approach's risk to overestimate the utility of the PAI. To establish further evidence, a simulation study could complement the current results, varying the true difference between patients receiving their optimal *v.* nonoptimal treatment across simulated datasets of different sizes. To shed more light on the underlying mechanisms, this study should also systematically vary several machine-learning pipeline characteristics, including CV-scheme, data leakage, model building (separate model per treatment option *v.* common model) and feature selection.

Furthermore, we would like to emphasize that, despite its frequent use, the PAI is not the only approach to a personalized treatment selection. There are several other methods, often summarized under the term individualized treatment rule (ITR). Most of these approaches share a common logic in comparing predictions for different treatment options but differ in how they build the underlying predictive model(s) and/or conduct the retrospective evaluation of the ITR. For example, the targeted learning approach (e.g. Benjet et al., 2023; Kessler, 2022) is characterized by predicting the outcome difference scores ('PAIs') directly via a second-level classifier. In contrast, the approach of Kapelner et al. (2021) focuses on a statistically sound evaluation of an ITR-based application by employing a sophisticated bootstrap procedure. Moreover, even the PAI-logic of comparing predictions for different treatment options can be bypassed in specific scenarios. Delgadillo et al. (2022) and Delgadillo, Huey, Bennett, and McMillan (2017), for instance, developed and successfully validated an ITR that identified patients with a generally poorer model-based prognosis and assigned them to the more intense treatment of a 2-stepped care approach. These examples illustrate the diverse range of approaches to personalized treatment selection and show that the most suitable approach might also depend on the specific context. A closer systematic and methodological examination would be beyond the scope of this paper. However, in general, it is important to note that these approaches, incorporating predictive modelling, are similarly vulnerable to methodological choices that increase the risk of bias. Indeed, a scoping review across various types of ITRs, including the PAI, identified partially similar problems, such as a large heterogeneity of effect sizes and small sample sizes (Lorenzo-Luaces, Peipert, de Jesús Romero, Rutter, & Rodriguez-Quintana, 2021).

As pointed out above, our analysis aimed to show that violating predictive modelling standards, such as employing LOO CV and allowing data leakage, might lead to false positive results when retrospectively evaluating the utility of the PAI. However, even if such an analysis has only low bias, its results should always only be considered as an estimator of the PAI's utility on completely new data (external validation). Thus, any real-world application of the PAI would need to be preceded by a thorough external validation in different relevant settings, with the type of external validation (e.g. temporal, geographic, or spatial) depending on the context. For other factors that should be considered before a potential clinical application, please see Deisenhofer et al. (2024).

In summary, our study cautions that the pipeline design may strongly influence the evaluation of the PAI's utility. Therefore, future studies using and testing the PAI should adhere to established predictive modelling standards. Such an unbiased evaluation of the PAI's utility is essential before considering its potential clinical application, which could serve an evidence-based treatment selection. To facilitate this adherence, we contribute to the advancement of this field by providing an open Python script that implements a state-of-the-art pipeline.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291724003118 .

Acknowledgements. We would like to thank Rebecca Delfendahl and Till Adam for their support in data preparation as well as their contributions in crafting and presenting the Python script on GitHub.

Protect-AD:

We would like to thank the following individuals for their help: Jule Dehler, Dorte Westphal, Katrin Hummel, Jürgen Hoyer (Dresden), Verena Pflug, Dirk Adolph, Cornelia Mohr, Jan Cwik (Bochum), Maike Hollandt, Anne Pietzner, Jörg Neubert (Greifswald), Carsten Konrad, Yunbo Yang,

Isabelle Ridderbusch, Adrian Wroblewski, Hanna Christiansen, Anne Maenz, Sophia Tennie, Jean Thierschmidt (Marburg), Marcel Romanos, Kathrin Zierhut, Kristina Dickhöver, Markus Winkler, Maria Stefanescu, Christiane Ziegler (Würzburg), Nathalia Weber, Sebastian Schauenberg, Sophia Wriedt, Carina Heitmann (Münster) Caroline im Brahm, Annika Evers (Cologne), Isabel Alt, Sophie Bischoff, Jennifer Mumm, Jens Plag, Anne Schreiner, Sophie Meska (Berlin). Xina Grählert and Marko Käppler of the Coordinating Centre for Clinical Trials (KKS) data center (Dresden) provided support with the electronic data assessment and data banking. Eva Stolzenburg, Stanislav Bologov, and Karina Bley provided administrative support. A complete list of project publications can be found at www.fzpe.de.

Panic-Net

Centers: Principal investigators (PI) with respective areas of responsibility in the MAC study are V. Arolt (Münster: Overall MAC Program Coordination), H.U. Wittchen (Dresden: Principal Investigator (PI) for the Randomized Clinical Trial and Manual Development), A. Hamm (Greifswald: PI for Psychophysiology), A.L. Gerlach (Münster: PI for Psychophysiology and Panic subtypes), A. Ströhle (Berlin: PI for Experimental Pharmacology), T. Kircher (Marburg: PI for functional neuroimaging), and J. Deckert (Würzburg: PI for Genetics). Additional site directors in the RTC component of the program are G.W. Alpers (Würzburg), T. Fydrich and L. Fehm (Berlin-Adlershof), and T. Lang (Bremen).

Data access and responsibility: All principle investigators take responsibility for the integrity ofthe respective study data and their components. All authors and co-authors had full access to all study data. Data analysis and manuscript preparation were completed by the authors and co-authors of this article, who take responsibility for its accuracy and content.

Acknowledgments and staff members by site:

Greifswald (coordinating site for Psychophysiology): Christiane Melzig, Jan Richter, Susan Richter, Matthias von Rad; Berlin-Charite (coordinating Center for Experimental Pharmacology): Harald Bruhn, Anja Siegmund, Meline Stoy, Andre Wittmann; Berlin-Adlershof: Irene Schulz; Münster (Overall MAC Program Coordination, Genetics and Functional Neuroimaging): Andreas Behnken, Katharina Domschke, Adrianna Ewert, Carsten Konrad, Bettina Pfleiderer, Peter Zwanzger; Münster (coordinating site for psychophysiology and subtyping): Judith Eidecker, Swantje Koller, Fred Rist, Anna Vossbeck-Elsebusch; Marburg/Aachen (coordinating center for functional neuroimaging): Barbara Druke, Sonja Eskens, Thomas Forkmann, Siegfried Gauggel, Susan Gruber, Andreas Jansen, Thilo Kellermann, Isabelle Reinhardt, Nina Vercamer-Fabri; Dresden (coordinating site for data collection, analysis, and the RCT): Franziska Einsle, Christine Frohlich, Andrew T. Gloster, Christina Hauke, Simone Heinze, Michael Hofler, Ulrike Lueken, Peter Neudeck, Stephanie Preiß, Dorte Westphal; Würzburg Psychiatry Department (coordinating center for genetics): Andreas Reif; Würzburg Psychology Department: Julia Dürner, Hedwig Eisenbarth, Antje B. M. Gerdes, Harald Krebs, Paul Pauli, Silvia Schad, Nina Steinhäuser; Bremen: Veronika Bamann, Sylvia Helbig-Lang, Anne Kordt, Pia Ley, Franz Petermann, Eva-Maria Schröder. Additional support was provided by the coordinating center for clinical studies in Dresden (KKS Dresden): Xina Grahlert and Marko Kappler.

Funding statement. This work was funded by the Deutsche Forschungsgemeinschaft – FOR5187 (project number 442075332). PANIC-net is part of the German multicenter trial 'Mechanisms of Action in CBT (MAC)'. The MAC study is funded by the German Federal Ministry of Education and Research (BMBF; project no. 01GV0615) as part of the BMBF Psychotherapy Research Funding Initiative. PROTECT-AD (Providing Tools for Effective Care and Treatment of Anxiety Disorders) is one out of nine research consortia in the German federal research program Research Network on Mental Disorders, funded by the Federal Ministry of Education and Research (www.fzpe.de), PROTECT-AD P1 grant number: 01EE1402A. The presented work was derived from project P1.

Competing interests. The authors declare no competing interests.

[1]Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany; [2]Translational Psychotherapy, Department of Psychology, Friedrich–Alexander University Erlangen–Nürnberg, Erlangen/Nürnberg, Germany; [3]Institute for Translational Psychiatry, University of Münster, Münster, Germany; [4]Department of Psychiatry, Psychosomatics, and Psychotherapy, Center of Mental Health, University Hospital of Würzburg, Würzburg, Germany; [5]Department of Psychiatry and Psychotherapy, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany; [6]German Center for Mental Health (DZPG), partner site Berlin-Potsdam, Germany; [7]Department of Psychology, University of Münster, Münster, Germany; [8]Department of Clinical Psychology and Psychotherapy, Faculty of Psychology, University of Cologne, Cologne, Germany; [9]Department of Biological and Clinical Psychology/Psychotherapy, University of Greifswald, Greifswald, Germany; [10]Institute of Clinical Psychology and Psychotherapy, Technische Universität Dresden, Dresden, Germany; [11]Department of Psychiatry and Psychotherapy & Center for Mind, Brain and Behavior, Philipps-University Marburg, Marburg, Germany; [12]LVR-University Hospital Essen, Department of Psychiatry and Psychotherapy, Faculty of Medicine, University of Duisburg-Essen, Duisburg/Essen, Germany; [13]Center for Translational Neuro- and Behavioral Sciences (CTNBS), University of Duisburg-Essen, Duisburg/Essen, Germany; [14]Social & Decision Sciences, School of Business, Constructor University Bremen, Bremen, Germany; [15]Christoph-Donier Foundation for Clinical Psychology, Marburg, Germany; [16]Mental Health Research and Treatment Center, Ruhr-Universität Bochum, Bochum, Germany; [17]Protect-AD Study Site Cologne, Cologne, Germany; [18]Department of Psychology (Biological Psychology, Clinical Psychology, and Psychotherapy), University of Würzburg, Würzburg, Germany; [19]Department of Experimental Psychopathology, University of Hildesheim, Hildesheim, Germany; [20]Department of Clinical Psychology and Psychotherapy, Faculty of Psychology & Center for Mind, Brain and Behavior, Philipps-University Marburg, Marburg, Germany; [21]Department of Psychiatry and Psychotherapy, Campus Charité Mitte, Charité – Universitätsmedizin Berlin, Berlin, Germany; [22]Department of Psychiatry and Psychotherapy, University Hospital, Ludwig-Maximilians-University Munich, Munich, Germany; [23]kbo-Inn-Salzach-Klinikum, Clinical Center für Psychiatry, Psychotherapy, Geriatrics, Neurology, Gabersee Wasserburg, Germany; [24]Department of Psychiatry and Psychotherapy, CCM, Charité – Universitätsmedizin Berlin, corporate member of FU Berlin and Humboldt Universität zu Berlin, Berlin, Germany; [25]Translational Psychotherapy, Institute of Psychology, University of Göttingen, Göttingen, Germany and [26]Department of Psychology, Health and Medical University Erfurt, Erfurt, Germany

References

Ahuvia, I. L., Mullarkey, M. C., Sung, J. Y., Fox, K. R., & Schleider, J. L. (2023). Evaluating a treatment selection approach for online single-session interventions for adolescent depression. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *64*(12), 1679–1688. https://doi.org/10.1111/jcpp.13822

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). Arlington, VA: American Psychiatric Association. https://doi.org/10.1176/appi.books.9780890425596

Bauer-Staeb, C., Griffith, E., Faraway, J. J., & Button, K. S. (2023). Personalised psychotherapy in primary care: Evaluation of data-driven treatment allocation to cognitive-behavioural therapy versus counselling for depression. *BJPsych Open*, *9*(2), e46. https://doi.org/10.1192/bjo.2022.628

Benjet, C., Zainal, N. H., Albor, Y., Alvis-Barranco, L., Carrasco-Tapias, N., Contreras-Ibáñez, C. C., … Kessler, R. C. (2023). A precision treatment model for internet-delivered cognitive behavioral therapy for anxiety and depression among university students: A secondary analysis of a randomized clinical trial. *JAMA Psychiatry*, *80*(8), 768–777. https://doi.org/10.1001/jamapsychiatry.2023.1675

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

Bremer, S., van Vliet, N. I., van Bronswijk, S., Huntjens, R., de Jongh, A., & van Dijk, M. K. (2023). Predicting optimal treatment outcomes in phase-based treatment and direct trauma-focused treatment among patients with posttraumatic stress disorder stemming from childhood abuse. *Journal of Traumatic Stress*, *36*(6), 1044–1055. https://doi.org/10.1002/jts.22980

Bruijniks, S. J. E., van Bronswijk, S. C., DeRubeis, R. J., Delgadillo, J., Cuijpers, P., & Huibers, M. J. H. (2022). Individual differences in response to once

versus twice weekly sessions of CBT and IPT for depression. *Journal of Consulting and Clinical Psychology*, *90*(1), 5–17. https://doi.org/10.1037/ccp0000658

Brunoni, A. R., Chaimani, A., Moffa, A. H., Razza, L. B., Gattaz, W. F., Daskalakis, Z. J., & Carvalho, A. F. (2017). Repetitive transcranial magnetic stimulation for the acute treatment of major depressive episodes: A systematic review with network meta-analysis. *JAMA Psychiatry*, *74*(2), 143–152. https://doi.org/10.1001/jamapsychiatry.2016.3644

Busner, J., & Targum, S. D. (2007). The clinical global impressions scale: Applying a research tool in clinical practice. *Psychiatry (Edgmont)*, *4*(7), 28–37.

Carpenter, J. K., Andrews, L. A., Witcraft, S. M., Powers, M. B., Smits, J. A. J., & Hofmann, S. G. (2018). Cognitive behavioral therapy for anxiety and related disorders: A meta-analysis of randomized placebo-controlled trials. *Depression and Anxiety*, *35*(6), 502–514. https://doi.org/10.1002/da.22728

Choi, K. W., & Jeon, H. J. (2020). Heart rate variability for the prediction of treatment response in Major depressive disorder. *Frontiers in Psychiatry*, *11*, 607. https://doi.org/10.3389/fpsyt.2020.00607

Cipriani, A., Furukawa, T. A., Salanti, G., Chaimani, A., Atkinson, L. Z., Ogawa, Y., … Geddes, J. R. (2018). Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: A systematic review and network meta-analysis. *The Lancet*, *391*(10128), 1357–1366. https://doi.org/10.1016/S0140-6736(17)32802-7

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

Cohen, Z. D., Kim, T. T., Van, H. L., Dekker, J. J. M., & Driessen, E. (2020). A demonstration of a multi-method variable selection approach for treatment selection: Recommending cognitive–behavioral versus psychodynamic therapy for mild to moderate adult depression. *Psychotherapy Research*, *30*(2), 137–150. https://doi.org/10.1080/10503307.2018.1563312

Cuijpers, P., Miguel, C., Harrer, M., Plessen, C. Y., Ciharova, M., Ebert, D., & Karyotaki, E. (2023). Cognitive behavior therapy vs. Control conditions, other psychotherapies, pharmacotherapies and combined treatment for depression: A comprehensive meta-analysis including 409 trials with 52702 patients. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)*, *22*(1), 105–115. https://doi.org/10.1002/wps.21069

Deisenhofer, A-K, Delgadillo, J., Rubel, J. A., Böhnke, J. R., Zimmermann, D., Schwartz, B., & Lutz, W. (2018). Individual treatment selection for patients with posttraumatic stress disorder. *Depression and Anxiety*, *35*(6), 541–550. https://doi.org/10.1002/da.22755

Deisenhofer, A-K, Barkham, M., Beierl, E. T., Schwartz, B., Aafjes-van Doorn, K., Beevers, C. G., … Cohen, Z. D. (2024). Implementing precision methods in personalizing psychological therapies: Barriers and possible ways forward. *Behaviour Research and Therapy*, *172*, 104443. https://doi.org/10.1016/j.brat.2023.104443

Delgadillo, J., & Duhne, P. G. S. (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

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. https://doi.org/10.1037/ccp0000231

Delgadillo, J., Ali, S., Fleck, K., Agnew, C., Southgate, A., Parkhouse, L., … Barkham, M. (2022). Stratified care vs stepped care for depression: A cluster randomized clinical trial. *JAMA Psychiatry*, *79*(2), 101–108. https://doi.org/10.1001/jamapsychiatry.2021.3539

Derogatis, L. R. (1993). *Brief symptom inventory* (*BSI*). *Administration, scoring, and procedures manual* (3rd ed.). National Computer Systems.

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

Fitzgerald, P. B. (2020). An update on the clinical use of repetitive transcranial magnetic stimulation in the treatment of depression. *Journal of Affective Disorders*, *276*, 90–103. https://doi.org/10.1016/j.jad.2020.06.067

Friedl, N., Berger, T., Krieger, T., Caspar, F., & Holtforth, M. G. (2020a). 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

Friedl, N., Krieger, T., Chevreul, K., Hazo, J. B., Holtzmann, J., Hoogendoorn, M., … Berger, H. (2020b). Using the personalized advantage index for individual treatment allocation to blended treatment or treatment as usual for depression in secondary care. *Journal of Clinical Medicine*, *9*(2), Article 490. https://doi.org/10.3390/jcm9020490

Gloster, A. T., Wittchen, H-U, Einsle, F., Lang, T., Helbig-Lang, S., Fydrich, T., … Arolt, V. (2011). Psychological treatment for panic disorder with agoraphobia: A randomized controlled trial to examine the role of therapist-guided exposure *in situ* in CBT. *Journal of Consulting and Clinical Psychology*, *79*(3), 406–420. https://doi.org/10.1037/a0023584

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)* (pp. 507–520). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/0378c7692da36807bdec87ab043cdadc-Paper-Datasets_and_Benchmarks.pdf

Hautmann, C., Dose, C., Hellmich, M., Scholz, K., Katzmann, J., Pinior, J., … Döpfner, M. (2023). Behavioural and nondirective parent training for children with externalising disorders: First steps towards personalised treatment recommendations. *Behaviour Research and Therapy*, *163*, 104271. https://doi.org/10.1016/j.brat.2023.104271

Heinig, I., Pittig, A., Richter, J., Hummel, K., Alt, I., Dickhöver, K., … Wittchen, H-U (2017). Optimizing exposure-based CBT for anxiety disorders via enhanced extinction: Design and methods of a multicentre randomized clinical trial. *International Journal of Methods in Psychiatric Research*, *26*(2), Article e1560. https://doi.org/10.1002/mpr.1560

Held, P., Patton, E., Pridgen, S. A., Smith, D. L., Kaysen, D. L., & Klassen, B. J. (2023). Using the personalized advantage index to determine which veterans may benefit from more vs. less comprehensive intensive PTSD treatment programs. *European Journal of Psychotraumatology*, *14*(2), 2281757. https://doi.org/10.1080/20008066.2023.2281757

Herzog, P., & Kaiser, T. (2022). Is it worth it to personalize the treatment of PTSD? – a variance-ratio meta-analysis and estimation of treatment effect heterogeneity in RCTs of PTSD. *Journal of Anxiety Disorders*, *91*, 102611. https://doi.org/10.1016/j.janxdis.2022.102611

Hoeboer, C. M., Oprel, D. A. C., de Kleine, R. A., Schwartz, B., Deisenhofer, A-K, Schoorl, M., … Lutz, W. (2021). Personalization of treatment for patients with childhood-abuse-related posttraumatic stress disorder. *Journal of Clinical Medicine*, *10*(19), Article 4522. https://doi.org/10.3390/jcm10194522

Huibers, M. J. H., Cohen, Z. D., Lemmens, L. H. J. M., Arntz, A., Peeters, F. P. M. L., Cuijpers, P., & DeRubeis, R. J. (2015). Predicting optimal outcomes in cognitive therapy or interpersonal psychotherapy for depressed individuals using the personalized advantage index approach. *PLoS One*, *10*(11), Article e0140771. https://doi.org/10.1371/journal.pone.0140771

Kaiser, T., Volkmann, C., Volkmann, A., Karyotaki, E., Cuijpers, P., & Brakemeier, E.-L. (2020). Heterogeneity of treatment effects in trials on psychotherapy of depression. *Clinical Psychology: Science and Practice*, *29*(3), 102611. https://doi.org/10.31234/osf.io/mqzvy

Kapelner, A., Bleich, J., Levine, A., Cohen, Z. D., DeRubeis, R. J., & Berk, R. (2021). Evaluating the effectiveness of personalized medicine with software. *Frontiers in Big Data*, *4*, 572532. https://doi.org/10.3389/fdata.2021.572532

Keefe, J. R., Kim, T. T., DeRubeis, R. J., Streiner, D. L., Links, P. S., & McMain, S. F. (2021). Treatment selection in borderline personality disorder between dialectical behavior therapy and psychodynamic psychiatric management. *Psychological Medicine*, *51*(11), 1829–1837. https://doi.org/10.1017/s0033291720000550

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Ebert, D. D., … Zaslavsky, A. M. (2017). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiology and Psychiatric Sciences*, *26*(1), 22–36. https://doi.org/10.1017/s2045796016000020

Kessler, R. C., Furukawa, T. A., Kato, T., Luedtke, A., Petukhova, M., Sadikova, E., Sampson, N. A. (2022). An individualized treatment rule to optimize probability of remission by continuation, switching, or combining antidepressant medications after failing a first-line. *Psychological Medicine*, *52*, 3371–3380. https://doi.org/10.1017/S0033291721000027

Lebeau, R. T., Glenn, D. E., Hanover, L. N., Beesdo-Baum, K., Wittchen, H-U, & Craske, M. G. (2012). A dimensional approach to measuring anxiety for DSM-5. *International Journal of Methods in Psychiatric Research*, *21*(4), 258–272. https://doi.org/10.1002/mpr.1369

Loerinc, A. G., Meuret, A. E., Twohig, M. P., Rosenfield, D., Bluett, E. J., & Craske, M. G. (2015). Response rates for CBT for anxiety disorders: Need for standardized criteria. *Clinical Psychology Review*, *42*, 72–82. https://doi.org/10.1016/j.cpr.2015.08.004

Loohuis, A. M. M., Burger, H., Wessels, N. J., Dekker, J. H., Malmberg, A. G., Berger, M. Y., … van der Worp, H. (2022). Prediction model study focusing on eHealth in the management of urinary incontinence: The personalised advantage index as a decision – making aid. *BMJ Open*, *12*(7), Article e051827. https://doi.org/10.1136/bmjopen-2021-051827

Lopez-Gomez, I., Lorenzo-Luaces, L., Chaves, C., Hervas, G., DeRubeis, R. J., & Vazquez, 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.genhosppsych.2019.07.00

Lorenzo-Luaces, L., Peipert, A., de Jesús Romero, R., Rutter, L. A., & Rodriguez-Quintana, N. (2021). Personalized medicine and cognitive behavioral therapies for depression: Small effects, big problems, and bigger data. *International Journal of Cognitive Therapy*, *14*(1), 59–85. https://doi.org/10.1007/s41811-020-00094-3

Luedtke, A. R., 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

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30: 31st Annual Conference on Neural Information Processing Systems (NIPS 2017): Long Beach, California, USA, 4–9 December 2017* (pp. 4768–4777). Curran Associates Inc.

Meehan, A. J., Lewis, S. J., Fazel, S., Fusar-Poli, P., Steyerberg, E. W., Stahl, D., & Danese, A. (2022). Clinical prediction models in psychiatry: A systematic review of two decades of progress and challenges. *Molecular Psychiatry*, *27*, 2700–2708. https://doi.org/10.1038/s41380-022-01528-4

Meinke, C., Lueken, U., Walter, H., & Hilbert, K. (2024). Predicting treatment outcome based on resting-state functional connectivity in internalizing mental disorders: A systematic review and meta-analysis. *Neuroscience and Biobehavioral Reviews*, *160*, 105640. https://doi.org/10.1016/j.neubiorev.2024.105640

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models interpretable* (2nd ed.). https://christophm.github.io/interpretable-ml-book

Moons, K. G. M., Wolff, R. F., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., … Mallett, S. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, *170*(1), 51–58. https://doi.org/10.7326/M18-1376

Navarro, C. L. A., Damen, J. A. A., Takada, T., Nijman, S. W. J., Dhiman, P., Ma, J., … Hooft, L. (2021). Risk of bias in studies on prediction models developed using supervised machine learning techniques: Systematic review. *BMJ*, *375*, Article n2281. https://doi.org/10.1136/bmj.n2281

Papakostas, G. I., & Fava, M. (2009). Does the probability of receiving placebo influence clinical trial outcome? A meta-regression of double-blind, randomized clinical trials in MDD. *European Neuropsychopharmacology*, *19*(1), 34–40. https://doi.org/10.1016/j.euroneuro.2008.08.009

Pittig, A., Heinig, I., Goerigk, S., Thiel, F., Hummel, K., Scholl, L., … Wittchen, H-U (2021). Efficacy of temporally intensified exposure for anxiety disorders: A multicenter randomized clinical trial. *Depression and Anxiety*, *38*(11), 1169–1181. https://doi.org/10.1002/da.23204

Pittig, A., Heinig, I., Goerigk, S., Richter, J., Hollandt, M., Lueken, U., … Wittchen, H-U (2023). Change of threat expectancy as mechanism of exposure-based psychotherapy for anxiety disorders: Evidence from 8,484 exposure exercises of 605 patients. *Clinical Psychological Science*, *11*(2), 199–217. https://doi.org/10.1177/21677026221101379

Plöderl, M., & Hengartner, M. P. (2019). What are the chances for personalised treatment with antidepressants? Detection of patient-by-treatment interaction with a variance ratio meta-analysis. *BMJ Open*, *9*(12), e034816. https://doi.org/10.1136/bmjopen-2019-034816

Rabin, R., & de Charro, F. (2001). Eq-5D: A measure of health status from the EuroQol Group. *Annals of Medicine*, *33*(5), 337–343. https://doi.org/10.3109/07853890109002087

Rasch, D., Kubinger, K. D., & Moder, K. (2011). The two-sample *t* test: Pre-testing its assumptions does not pay off. *Statistical Papers*, *52*(1), 219–231. https://doi.org/10.1007/s00362-009-0224-x

Schwartz, B., Cohen, Z. D., Rubel, J. A., Zimmermann, D., Wittmann, W. W., & Lutz, W. (2021). Personalized treatment selection in routine care: Integrating machine learning and statistical algorithms to recommend cognitive behavioral or psychodynamic therapy. *Psychotherapy Research*, *31*(1), 33–51. https://doi.org/10.1080/10503307.2020.1769219

Senger, K., Schröder, A., Kleinstäuber, M., Rubel, J. A., Rief, W., & Heider, J. (2021). Predicting optimal treatment outcomes using the personalized advantage index for patients with persistent somatic symptoms. *Psychotherapy Research*, *32*(2), 165–178. https://doi.org/10.1080/10503307.2021.1916120

Shear, M. K., Vander Bilt, J., Rucci, P., Endicott, J., Lydiard, B., Otto, M. W., … Frank, D. M. (2001). Reliability and validity of a structured interview guide for the Hamilton anxiety rating scale (SIGH-A). *Depression and Anxiety*, *13*(4), 166–178. https://doi.org/10.1002/da.1033.abs

Üstün, T. B., Kostanjsek, N., Chatterji, S., & Rehm, J. (Eds.). (2010) . *Measuring health and disability: Manual for WHO disability assessment schedule* (*WHODAS 2.0*). Geneva, Switzerland: World Health Organization

van Bronswijk, S. C., Bruijniks, S. J. E., Lorenzo-Luaces, L., DeRubeis, R. J., Lemmens, L. H. J. M., Peeters, F. P. M. L., & Huibers, M. J. H. (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. https://doi.org/10.1080/10503307.2020.1823029

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, *180*(Part A), 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061

Vieira, S., Liang, X., Guiomar, R., & Mechelli, A. (2022). Can we predict who will benefit from cognitive-behavioural therapy? A systematic review and meta-analysis of machine learning studies. *Clinical Psychology Review*, *97*, Article 102193. https://doi.org/10.1016/j.cpr.2022.102193

Watts, D., Pulice, R. F., Reilly, J., Brunoni, A. R., Kapczinski, F., & Passos, I. C. (2022). Predicting treatment response using EEG in major depressive disorder: A machine-learning meta-analysis. *Translational Psychiatry*, *12*(1), 332. https://doi.org/10.1038/s41398-022-02064-z

Webb, C. A., Trivedi, M. H., Cohen, Z. D., Dillon, D. G., Fournier, J. C., Goer, F., … Pizzagalli, D. A. (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

Zantvoort, K., Nacke, B., Görlich, D., Hornstein, S., Jacobi, C., & Funk, B. (2024). *Predictive power, variance and generalizability – A machine learning case study on minimal necessary data sets sizes in digital mental health intervention predictions*. Research Square. https://doi.org/10.21203/rs.3.rs-4616728/v1

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

https://doi.org/10.1017/S0033291724003118 Published online by Cambridge University Press