NeuroImage 295 (2024) 120639

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Lack of evidence for predictive utility from resting state fMRI data for individual exposure-based cognitive behavioral therapy outcomes: A machine learning study in two large multi-site samples in anxiety disorders

Kevin Hilbert[a,b,1], Joscha Böhnlein[c,1,*], Charlotte Meinke[a], Alice V. Chavanne[a,d], Till Langhammer[a], Lara Stumpe[c], Nils Winter[c], Ramona Leenings[c], Dirk Adolph[e], Volker Arolt[c], Sophie Bischoff[f], Jan C. Cwik[g], Jürgen Deckert[h], Katharina Domschke[i], Thomas Fydrich[a], Bettina Gathmann[j], Alfons O. Hamm[k], Ingmar Heinig[l], Martin J. Herrmann[h], Maike Hollandt[k], Jürgen Hoyer[l], Markus Junghöfer[m], Tilo Kircher[n], Katja Koelkebeck[o], Martin Lotze[p], Jürgen Margraf[e], Jennifer L.M. Mumm[f], Peter Neudeck[q,r], Paul Pauli[s], Andre Pittig[t], Jens Plag[f,u], Jan Richter[k,v], Isabelle C. Ridderbusch[n], Winfried Rief[w], Silvia Schneider[x], Hanna Schwarzmeier[h], Fabian R. Seeger[h], Niklas Siminski[h], Benjamin Straube[n], Thomas Straube[y], Andreas Ströhle[f], Hans-Ulrich Wittchen[z], Adrian Wroblewski[n], Yunbo Yang[n], Kati Roesmann[m,y], Elisabeth J. Leehr[c], Udo Dannlowski[c], Ulrike Lueken[a,aa]

[a] *Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany*
[b] *Department of Psychology, HMU Health and Medical University Erfurt, Erfurt, Germany*
[c] *Institute for Translational Psychiatry, University of Münster, Germany*
[d] *Université Paris-Saclay, INSERM U1299 "Trajectoires développementales et psychiatrie", CNRS UMR 9010 Centre Borelli, Ecole Normale Supérieure Paris-Saclay, France*
[e] *Mental Health Research and Treatment Center, Faculty of Psychology, Ruhr-Universität Bochum, Bochum, Germany*
[f] *Department of Psychiatry and Psychotherapy, Campus Charité Mitte, Charité - Universitätsmedizin Berlin, Berlin, Germany*
[g] *Department of Clinical Psychology and Psychotherapy, Faculty of Human Sciences, Universität zu Köln, Germany*
[h] *Center for Mental Health, Department of Psychiatry, Psychosomatics, and Psychotherapy, University Hospital of Würzburg, Germany*
[i] *Department of Psychiatry and Psychotherapy, Medical Center - University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany*
[j] *Institute of Medical Psychology and Systems Neuroscience, University of Münster, Germany*
[k] *Department of Biological and Clinical Psychology, University of Greifswald, Greifswald, Germany*
[l] *Institute of Clinical Psychology & Psychotherapy, Technische Universität Dresden, Dresden, Germany*
[m] *Institute for Biomagnetism and Biosignalanalysis, University of Münster, Germany*
[n] *Department of Psychiatry and Psychotherapy, University of Marburg, Marburg, Germany*
[o] *LVR-University-Hospital Essen, Department of Psychiatry and Psychotherapy, University of Duisburg-Essen, Essen, Germany*
[p] *Functional Imaging Unit. Diagnostic Radiology and Neuroradiology, University Medicine Greifswald, Greifswald, Germany*
[q] *Protect-AD Study Site Cologne, Cologne, Germany*
[r] *Institut für Klinische Psychologie und Psychotherapie, TU Chemnitz, Germany*
[s] *Department of Psychology, University of Würzburg, Würzburg, Germany*
[t] *Translational Psychotherapy, Institute of Psychology, University of Göttingen, Germany*
[u] *Department of Psychiatry, Psychotherapy and Psychosomatics, Alexianer Krankenhaus Hedwigshoehe, St. Hedwig Kliniken, Berlin, Germany*
[v] *Department of Experimental Psychopathology, University of Hildesheim, Hildesheim, Germany*
[w] *Department of Clinical Psychology and Psychotherapy, Faculty of Psychology & Center for Mind, Brain and Behavior – CMBB, Philipps-University of Marburg, Marburg, Germany*
[x] *Faculty of Psychology, Clinical Child and Adolescent Psychology, Mental Health Research and Treatment Center, Ruhr-Universität Bochum, Bochum, Germany*
[y] *Institute of Psychology, Unit of Clinical Psychology and Psychotherapy in Childhood and Adolescence, University of Osnabrueck, Osnabruck, Germany*
[z] *Psychiatric University Hospital, Ludwig-Maximilians-University München, Germany*
[aa] *German Center for Mental Health (DZPG), partner site Berlin/Potsdam, Germany*

[*] Corresponding author at: University of Münster, Institute for Translational Psychiatry, Albert-Schweitzer-Campus 1, Building A9a, 48147, Münster.
E-mail address: joscha.boehnlein@uni-muenster.de (J. Böhnlein).

[1] Shared first authorship.

https://doi.org/10.1016/j.neuroimage.2024.120639
Received 8 March 2024; Received in revised form 3 May 2024; Accepted 6 May 2024
Available online 25 May 2024

ARTICLE INFO

Keywords:
Outcome prediction
Precision psychotherapy
Machine learning
Anxiety disorders
Cognitive behavioral therapy
Resting state

ABSTRACT

Data-based predictions of individual Cognitive Behavioral Therapy (CBT) treatment response are a fundamental step towards precision medicine. Past studies demonstrated only moderate prediction accuracy (i.e. ability to discriminate between responders and non-responders of a given treatment) when using clinical routine data such as demographic and questionnaire data, while neuroimaging data achieved superior prediction accuracy. However, these studies may be considerably biased due to very limited sample sizes and bias-prone methodology. Adequately powered and cross-validated samples are a prerequisite to evaluate predictive performance and to identify the most promising predictors. We therefore analyzed resting state functional magnet resonance imaging (rs-fMRI) data from two large clinical trials to test whether functional neuroimaging data continues to provide good prediction accuracy in much larger samples. Data came from two distinct German multicenter studies on exposure-based CBT for anxiety disorders, the Protect-AD and SpiderVR studies. We separately and independently preprocessed baseline rs-fMRI data from $n = 220$ patients (Protect-AD) and $n = 190$ patients (SpiderVR) and extracted a variety of features, including ROI-to-ROI and edge-functional connectivity, sliding-windows, and graph measures. Including these features in sophisticated machine learning pipelines, we found that predictions of individual outcomes never significantly differed from chance level, even when conducting a range of exploratory post-hoc analyses. Moreover, resting state data never provided prediction accuracy beyond the sociodemographic and clinical data. The analyses were independent of each other in terms of selecting methods to process resting state data for prediction input as well as in the used parameters of the machine learning pipelines, corroborating the external validity of the results. These similar findings in two independent studies, analyzed separately, urge caution regarding the interpretation of promising prediction results based on neuroimaging data from small samples and emphasizes that some of the prediction accuracies from previous studies may result from overestimation due to homogeneous data and weak cross-validation schemes. The promise of resting-state neuroimaging data to play an important role in the prediction of CBT treatment outcomes in patients with anxiety disorders remains yet to be delivered.

1. Introduction

Precision medicine represents the idea of tailoring treatment to individual patient characteristics rather than offering a "one size fits all" solution (Ozomaro et al., 2013). It bears high potential for improving treatment outcomes for the cost-intensive group of patients with mental disorders not responding to first-line treatments such as cognitive behavioral therapy (CBT). Data-based predictions of individual treatment outcomes are a fundamental step towards precision medicine, as these predictions are needed to select the most suitable treatment option (Lueken and Hahn, 2020).

With machine learning, a set of tools is available that excels at detecting complex patterns and interactions in multiple predictor variables and translating them into one prediction for the individual patient (Bzdok et al., 2017; Janssen et al., 2018). These approaches compliment those using classical, univariate analyses that have already been able to detect a range of response predictors on a group-level. For example, (Pico-Perez et al., 2022) show that it is possible to associate therapy response using task-based function magnetic resonance imaging (fMRI) data; (Fullana et al., 2017) and (Cyr et al., 2021) found evidence that resting state (rs-)fMRI data can be used to associate response in therapy patients with obsessive-compulsive disorder. Consequently, a variety of studies in recent years has examined a plethora of different potential predictor variables from numerous data processing pipelines and machine learning algorithms in order to identify the most promising approaches, including clinical routine and neuroimaging data (Vieira et al., 2022).

Recent efforts to associate individual-level CBT outcomes with clinical routine data alone have only resulted in moderate prediction accuracies (58–69%) (Hilbert et al., 2021; Hilbert et al., 2020; Hornstein et al., 2021; Taubitz et al., 2022; Wallert et al., 2022; Leehr et al., 2021; Symons et al., 2019; Symons et al., 2020). On the contrary, task-based and resting state neuroimaging data was found informative for predictive models in anxiety disorders with accuracies between 81–92% and might provide incremental accuracy (Hahn et al., 2015; Månsson et al., 2015; Frick et al., 2020; Whitfield-Gabrieli et al., 2016). However, the neuroimaging studies have largely been plagued by very limited sample sizes (Vieira et al., 2022), which are prone to overfitting and only allow for weaker cross-validation strategies with increased risk of biased and overestimated accuracy estimates (Varoquaux, 2018; Varoquaux et al., 2017; Flint et al., 2021). In addition, past reviews emphasized that some clinical predictors could be replicated quite consistently, whereas fMRI-based predictors were reported rather in exploratory studies and replicability is not always given (Deckert and Angelika, 2019). It remains unclear whether neuroimaging data really facilitate superior prediction accuracies compared to clinical routine data, and how much incremental accuracy neuroimaging data provides when both data modalities are included in a common prediction model.

In this study, we established two independent teams that applied sophisticated machine learning analysis pipelines on separate resting state functional MRI data from two multicenter large-scale trials in anxiety disorders. Each team analyzed one dataset, but teams did not directly replicate each other. Our aim was to investigate individual CBT outcome prediction performances (e.g. reduction in relevant questionnaire scores) to examine whether the promise of neuroimaging data to provide good to very good prediction accuracy holds true in these large clinical samples, and to compare the relative contributions of neuroimaging and sociodemographic / clinical features to the final prediction performance. Our approach with two independent analysis teams allowed for separate choices regarding specific predictors (e.g., regarding data extraction, preprocessing, and feature reduction) and machine learning pipelines for both datasets. Hence, we applied distinct powerful methodological approaches mirroring the heterogeneity of analysis pipelines in real-world settings. We hypothesized i) outcome prediction accuracies both immediately after and six months after treatment based on both resting state as well as demographic and clinical data significantly exceeding chance level in both datasets, ii) resting state data providing incremental predictive power beyond the sociodemographic and clinical data alone.

2. Methods

2.1. Datasets

We analyzed data from two German multicenter studies on exposure-based CBT for anxiety disorders, the Protect-AD (NIMH Protocol

Registration System 01EE1402A; ClinicalTrials.gov ID NCT02605668; German Register of Clinical Studies DRKS00008743) and SpiderVR (ClinicalTrials.gov ID NCT03208400) studies. Protect-AD was conducted in Berlin, Bochum, Cologne, Dresden, Greifswald, Marburg, Münster and Würzburg. Patients with a diagnosis of panic disorder, agoraphobia, social anxiety disorder, or multiple specific phobias received exposure-based CBT in vivo across twelve sessions (Heinig et al., 2017). This was a randomized controlled study with two arms, where patients received the exposure sessions in a temporally intensified fashion delivered within 2 weeks or a standard less intensive way delivered within 6 weeks. Overall, participants did improve clinically with large effect sizes (Pittig et al., 2021). As patients in both arms improved comparably in the primary outcome, we consequently did not differentiate for treatment condition for the current outcome prediction analysis and used all patients together. SpiderVR was conducted in Münster and Würzburg. Patients with a diagnosis of spider phobia received exposure-based CBT in virtuo in a single-session (Schwarzmeier et al., 2020). This was a prospective longitudinal study with only one arm, where all patients received an exposure session with a maximum duration of 2.5 h. Overall, participants did improve clinically with large effect sizes (Leehr et al., 2021) that were in line with other studies investigating virtual reality exposure treatment (Opris et al., 2012; Powers and Emmelkamp, 2008).

All participants provided written informed consent before study participation.

Both studies were approved by local Ethics Committees and performed according to the Declaration of Helsinki. Protect-AD: TUD-Ethics Review Committee (EK 234,062,014, 11/14/2014), SpiderVR: Ethics Committees of the Medical Faculties at Wuerzburg University (proposal number 330/15) and Muenster University (proposal number 216–212-b-S).

2.2. Patients and outcomes

Among all patients included in the Protect-AD trial, a subset of $n = 309$ patients completed the fMRI assessment at baseline and was thus available for the current analyses. Of these, we excluded $n = 66$ patients due to data loss and insufficient quality in the rs-fMRI data (Langhammer et al.), and an additional $n = 23$ patients due to missing primary outcome data, which is needed for the predictions. This resulted in a final sample of $n = 220$ patients for Protect-AD. For SpiderVR, all $n = 207$ patients were assessed in the scanner. Due to loss and insufficient quality in the rs-fMRI data, we excluded 17 participants, resulting in a final sample of $n = 190$ patients for SpiderVR. Eleven patients did not take part in the 6-month follow up (FU) assessment and thus were not included in the prediction of FU outcomes, resulting in a sample of $n = 179$ patients for these sub-analyses. Note that this sample is largely the same sample used in (Leehr et al., 2021), but differs slightly: we included patients who were assessed after the analyses for the previous study were completed, but had to exclude patients with bad or missing rs-fMRI images that were included in the previous study. The data used in the analysis of this manuscript have previously been analyzed univariately elsewhere (Leehr et al., 2024).

Treatment response for Protect-AD was defined as a reduction of at least 50% in the Hamilton Anxiety Rating Scale (HAMA-A; administered with the Structured Interview Guide for the Hamilton Anxiety Scale (Shear et al., 2001)) score between baseline and post-therapy assessment. This results in patients having improved their severity by at least one category post-therapy compared to pre-therapy (according to the severity cut-offs of Matza et al., 2010). This response definition was also used as the primary outcome in the clinical trial (Pittig et al., 2021). For SpiderVR, a reduction of at least 30% in the spider phobia questionnaire (SPQ (Hamm, 2006)) score between baseline and post-treatment assessment and between baseline and FU-assessment (primary outcome) was seen as treatment response. We chose this cutoff criterion as such a reduction typically leads to a post- or FU-treatment SPQ score of <20, which is defined as the cutoff for clinically significant symptoms (Hamm, 2006). This is in accordance with all other analyses of the Spider-VR project (Schwarzmeier et al., 2020). As a secondary outcome, a reduction of at least 50% in approach distance in a Behavioral Avoidance Test with a living bird spider (BAT, see (Schwarzmeier et al., 2020) for a thorough description) was seen as treatment response. Patients' demographics and clinical characteristics split between responders and non-responders are presented in Table 1.

2.3. Image acquisition and preprocessing

Protect-AD. MRI scans were acquired on harmonized scanning sequences at seven sites with 3-Tesla MRI scanners (3x Siemens TrioTim, 1x Siemens Verio, 1x Siemens Prisma, 1x Siemens Skyra, 1 x Philips Achieva). Rs- fMRI images were collected using a T2-weighted gradient-echo echoplanar imaging (EPI) sequence (31–33 axial slices of 3.8 mm thickness with 10% gap per volume, TR = 2000 ms, TE = 29–30 ms, flip angle = 90°, resolution = 3.3 × 3.3 × 3.8 mm, matrix size = 64 × 64 voxel, field of view = 210 mm). 237 vol scans were acquired in one run with a total length of approx. 8 min. Patients were instructed to remain still and to close their eyes. The screen was black and the lights inside the MRI scanning room were switched off. Additionally, a T1-weighted magnetization-prepared rapid gradient-echo (MPRAGE) sequence (176 sagittal slices, 1 mm slice thickness, no gap, TR = 1900 ms, TE = 2.26 ms, flip angle = 9°, resolution = 1 × 1 × 1 mm, matrix size = 256 × 256, field of view = 256 mm^3) was used to acquire a structural scan. To minimize carry-over effects, we conducted the RS fMRI paradigm before all other paradigms that were also part of our MRI sessions. Hence, there were only structural MRI/T1-weighted sequences before the RS fMRI.

Rs-fMRI data were preprocessed using the CONN toolbox and the CONN default preprocessing pipeline (Whitfield-Gabrieli, 2012) implemented in MATLAB (R2019b; The MathWorks Inc., MA, USA) and SPM12 (Penny et al., 2006). The pipeline included functional realignment and unwarping, slice timing correction, structural segmentation and normalization, functional normalization, smoothing, outlier identification and denoising. This was done centrally in order to rule out methodological differences between sites. More information on the preprocessing can be found elsewhere (Langhammer et al.). We used a selection of bilateral regions of interest (ROIs) implicated in clinical anxiety in recent meta-analyses (Chavanne and Robinson, 2021; Santos et al., 2019; Lueken and Hahn, 2016): the dorsal, pregenual, and subgenual part of the anterior cingulate cortex (ACC), the dorsomedial prefrontal cortex (DMPFC), the ventromedial prefrontal cortex (VMPFC), the dorsolateral prefrontal cortex (DLPFC), the ventrolateral prefrontal cortex (VLPFC), the orbitofrontal prefrontal cortex (OFPFC), the amygdala, the anterior and posterior insula, the hippocampus, the thalamus, the periacqueductal gray (PAG), and the bed nucleus of the stria terminalis (BNST). ROI definitions were taken from the Brainnetome atlas (Fan et al., 2016) (all ROIs except PAG and BNST; PAG based on masks created by Keuken and colleagues (Keuken et al., 2017), BNST based on the atlas for the hypothalamic region from Neudorfer and colleagues (Neudorfer et al., 2020)). ROI-to-ROI connectivity was extracted as the Fisher's-z transformed correlation between mean time series per ROI.

Additionally, graph metrics were calculated as they can characterize the properties of large-scale brain networks and their parts, which may be more predictive than classical ROI-ROI connectivity. We used ROIs as nodes and z-correlations >0.3 as edges. Four graph metrics were calculated for each node (ROI) and for the whole graph by averaging the results of all nodes: cost, global efficiency, betweenness centrality, and clustering coefficient. An introduction to graph-theory including an explanation of these metrics can be found here (Bullmore and Sporns, 2009).

SpiderVR. Patients were examined in two different 3-Tesla scanners: A Siemens Prisma MRI in Münster and a Siemens Skyra in Würzburg. Rs-fMRI images were collected with the following parameters: 31

Table 1
Core sociodemographic and clinical characteristics of Protect-AD and SpiderVR datasets at pre-treatment, for the final used samples of patients and for responders and non-responders separately. Means (SD), except where noted. The categorization as responders (vs. non-responders) is based on the primary outcome measure, i.e. HAMA-A reductions of 50% (Protect-AD) and SPQ reductions of 30% (SpiderVR) from pre- to post-treatment assessment.

	Protect-AD						SpiderVR					
	Responder ($n = 108$)		Non-responder ($n = 112$)		All ($n = 220$)		Responder ($n = 105$)		Non-responder ($n = 85$)		All ($n = 190$)	
Sociodemographics												
Female [n (%)]	56	(51.90)	60	(53.60)	116	(52.70)	89	(84.80)	72	(85.90)	162	(85.30)
Age (years)	30.77	(10.51)	30.81	(10.21)	30.79	(10.34)	26.77	(7.78)	29.52	(9.82)	28.00	(8.84)
Years of Education[1]	12.02	(1.39)	11.76	(1.39)	11.89	(1.39)	14.57	(3.03)	14.69	(2.96)	14.63	(2.99)
Site: Berlin [n (%)]	23	(21.3)	24	(21.4)	47	(21.4)	–		–		–	
Site: Bochum [n (%)]	16	(14.8)	3	(2.7)	19	(8.6)	–		–		–	
Site: Dresden [n (%)]	22	(20.4)	27	(24.1)	49	(22.3)	–		–		–	
Site: Greifswald [n (%)]	11	(10.2)	10	(8.9)	21	(9.5)	–		–		–	
Site: Marburg [n (%)]	20	(18.5)	25	(22.3)	45	(20.5)	–		–		–	
Site: Münster [n (%)]	9	(8.3)	10	(8.9)	19	(8.6)	50	(47.6)	54	(63.5)	104	(54.7)
Site: Würzburg [n (%)]	7	(6.5)	13	(11.6)	20	(9.1)	55	(52.4)	31	(36.5)	86	(45.3)
Smoking [n (%)][1]	34	(33.00)	22	(19.80)	56	(26.20)	NA		NA		NA	
Clinical characteristics												
Diagnosis: PD + AG [n (%)]	68	(63.00)	72	(64.30)	140	(63.60)						
Diagnosis: PD [n (%)]	61	(56.50)	64	(57.10)	125	(56.80)						
Diagnosis: AG [n (%)]	54	(50.00)	56	(50.00)	110	(50.00)						
Diagnosis: SAD [n (%)]	44	(40.70)	57	(50.90)	101	(45.90)						
Diagnosis: SPH [n (%)]	32	(29.60)	23	(20.50)	55	(25.00)	105	(100.00)	85	(100.00)	190	(100.00)
Diagnosis: Major Depression [n (%)]	29	(26.9)	51	(45.5)	80	(36.4)	2	(1.90)	3	(3.53)	5	(2.63)
Symptom severity												
SIGH-A	24.44	(5.68)	24.61	(5.09)	24.53	(5.37)	NA		NA		NA	
SPQ	NA		NA		NA		23.37	(2.15)	22.11	(1.93)	22.8	(2.15)
BAT final distance	NA		NA		NA		175.89	(63.04)	157.84	(70.13)	167.81	(66.74)
CGI	4.79	(0.70)	5.02	(0.67)	4.90	(0.69)	4.43	(0.71)	4.15	(0.81)	4.31	(0.76)
BDI-II	14.81	(10.20)	18.69	(8.91)	16.79	(9.74)	3.25	(3.25)	3.39	(3.82)	3.31	(3.92)
ASI	26.20	(11.29)	28.55	(10.18)	27.40	(10.78)	13.71	(8.91)	15.74	(10.24)	14.62	(9.56)
PAS	17.45	(9.74)	19.58	(11.05)	18.53	(10.46)	NA		NA		NA	
LSAS	42.20	(29.23)	56.77	(31.25)	49.62	(31.07)	22.91	(16.59)	26.03	(18.39)	24.31	(17.44)
DSM-5 SP	12.93	(10.60)	13.91	(10.18)	13.43	(10.37)	NA		NA		NA	
PROMIS-SPH	NA		NA		NA		1.82	(2.69)	1.99	(2.88)	1.89	(2.77)
Outcome												
SIGH-A posttreatment	6.67	(3.93)	19.21	(6.64)	13.05	(8.33)	NA		NA		NA	
SPQ post	NA		NA		NA		13.30	(2.33)	17.78	(2.02)	15.31	(3.13)
BAT post final distance	NA		NA		NA		70.00	(56.46)	93.35	(61.60)	80.45	(59.79)

PD: panic disorder, AG: agoraphobia, SAD: social anxiety disorder, SPH: specific phobia, SIGH-A: Structured Interview Guide for the Hamilton Anxiety Rating Scale, SPQ: Spider Phobia Questionnaire, CGI: Clinical Global Impression scale, BAT: Behavioral Avoidance Test, BDI-II: Beck Depression Inventory-II, ASI: Anxiety Sensitivity Index, PAS: Panic and Agoraphobia Scale, LSAS: Liebowitz Social Anxiety Scale, DSM-5 SP: Dimensional Specific Phobia Scale for DSM-5, PROMIS-SPH: Patient-Reported Outcomes Measurement Information System.

[1] data from $n = 5$ responder and $n = 1$ non-responder missing in Protect-AD.

(Würzburg) and 33 (Münster) axial slices of 3.8 mm thickness with 10% gap per volume, TR = 2000 ms, TE = 29 (Münster) or 30 ms (Würzburg), flip angle = 90°, resolution = 3.3 × 3.3 × 3.8 mm, matrix size = 64 × 64 voxel, field of view = 210 mm). The slices were positioned transaxially parallel to the intercommissural (AC-PC) plane and tilted 20° to reduce magnetic susceptibility artifacts in prefrontal areas. The T1-weightes MPRAGE sequence had the following parameters: 176 sagittal slices, 1 mm slice thickness, no gap, TR = 2130 ms, TE = 2.28 ms, flip angle = 8°, resolution = 1 × 1 × 1 mm, matrix size = 256 × 256, field of view = 256 mm^3. Again, we conducted the RS fMRI paradigm before all other fMRI paradigms that were also part of our MRI sessions.

Rs-fMRI data were again preprocessed with the CONN toolbox and the CONN default preprocessing pipeline (Whitfield-Gabrieli, 2012), implemented in MATLAB (R2020b) and SPM12 (Penny et al., 2006). The preprocessing pipeline was similar to the one described above. For the first ("static resting state") and second analysis ("combined static resting state and clinical data"), we extracted the ROI-to-ROI connectivity as the Fisher's-z transformed correlation between mean time series per ROI for all the $n = 116$ ROIs in the "automatic anatomic labeling" (AAL) atlas (Rolls et al., 2020) (http://www.gin.cnrs.fr/en/tools/aal/). This resulted in a 116 × 116 matrix. To obtain the correlations between every ROI-ROI pair excluding duplicates and self-correlations, we extracted the upper triangular part (without the diagonal), resulting in $116 \times 115/2 = 6670$ correlation values per participant.

2.4. Clinical and demographic predictor variables

Protect-AD. In addition to the neuroimaging data, we used a minimal sociodemographic and clinical data predictor set that included only age, sex, and baseline severity (HAM-A values). Particularly age and baseline severity have been consistently shown to yield predictive utility in prior CBT outcome prediction studies (Hilbert et al., 2021; Hilbert et al., 2020; Hornstein et al., 2021; Leehr et al., 2021). For a more profound analysis of the clinical data, a separate analysis using the full breadth of clinical data in the Protect-AD dataset is currently under preparation.

Spider-VR. To test whether the combination of clinical and demographic data with Rs-fMRI data would improve predictive accuracy, we repeated the data predictor set described and used in a previous study on SpiderVR data (Leehr et al., 2021), in which we could find a significant, but relatively small predictive value of the clinical and demographic alone.

2.5. Machine learning pipeline

Since the two teams worked separate from each other, the pipelines differ from each other:

Protect-AD. We employed a stacked ensemble learning approach, with three first-level learners providing separate predictions based on i) the clinical and demographic predictors, ii) the ROI-to-ROI resting state connectivity and iii) the graph-derived metrics. These three first-level learners initially used $k = 3$, $k = 435$, and $k = 124$ features as inputs. Then, these separate first-level predictions were in turn used as features for the second-level learner that produced the final predictions. The whole procedure was conducted over 100 iterations to get robust estimates of the performance metrics independent from the current train-test split.

The first-level learners employed a train-test split with 80% of the data going into the train and 20% of the data going into the test sets, i.e., serving as out-of-sample validation samples. Patients of the minority outcome label were oversampled for a balanced outcome frequency. Missing features were imputed with their mode, median and mean as appropriate and rescaled to z-scores. For feature selection, we used an elastic net-regularized logistic regression model with stochastic gradient descent learning; features were selected for the prediction if they had above-average absolute feature weights with the help of scikit-learn's "feature_selection.SelectFromModel" function. For the prediction, we used a random forest (Breiman, 2001) with 1000 estimators and otherwise standard scikit-learn (Pedregosa et al., 2011) hyperparameters. Predictions by first-level learners in the train set were used as features to train the second-level learner and predictions by first-level learners in the test set were used as features for the prediction on the test set (see supplemental figure 1). The primary second-level learner was also a random forest with 1000 estimators and otherwise standard scikit-learn hyperparameters, while we also employed logistic regression, majority voting, softmax voting and weighted softmax voting as exploratory second-level learners (see supplemental methods for details). Significance of the predictive models was assessed by comparing the balanced accuracies of the classifier against a pseudo-classifier with 0.5 balanced accuracy (derived from the balanced outcome frequencies after resampling) over 100 iterations with Nadeau & Bengio's corrected resampled *t*-test (Nadeau and Bengio, 2003; Bouckaert and Frank, 2004). Due to the oversampling procedure the samples for the train and test split varied somewhat across the iterations, for the corrected resampled *t*-test we used the mean sample sizes across all iterations. Given the lack of significant results for this main approach, we conducted additional exploratory post-hoc analyses that varied specific aspects of the machine learning pipeline, particularly the feature selection approach, the examined outcome and the second-level learner. This also included one approach using all available ROIs from the whole brain instead the preselected ROIs based on the previous literature. Additionally, we repeated the main approach with treatment response as a dimensional measurement (change of scores/distance from pre to post treatment in percent) in a regressional machine learning approach and conducted a sanity check with synthetic data. These approaches are described in the supplemental methods.

SpiderVR. We first used only the clinical and demographic data as features, effectively replicating the analysis by Leehr et al. (Leehr et al., 2021). We then used only "static resting state" data, followed by the combination of static resting state and clinical and demographic data. Given the lack of significant results for these main approaches, we repeated the first approach with treatment response this time as a dimensional measurement (change of scores/distance from pre to post treatment in percent) in a regressional machine learning approach. We then conducted additional exploratory post-hoc analyses that used time-dependent features: a "sliding window" and an edge-functional connectivity analysis. Additionally, we repeated the sanity check of the Protect-AD part with our methodology. These exploratory approaches are described in detail in the supplemental methods.

All built pipelines mainly followed a shared identical structure. We used the PHOTONAI toolbox (see https://www.photon-ai.com (Leenings et al., 2021)). Following guidelines by Poldrack and colleagues (Poldrack et al., 2020), we applied a nested cross-validation scheme in which 10 inner validation loops were used to optimize hyperparameters and 10 outer validation loops were used to estimate model performance (with 10% of the samples as test set in every step). Thus, ten percent of the sample were used as out-of-sample validation sample in every outer validation loop. Model performance was optimized for balanced accuracy. The pipeline consisted of an imbalanced data transformer to achieve balanced outcome frequency, a standard scaler to z-score the data and a principal component analysis (PCA) and/ or an F-test based univariate feature selection implemented in PHOTONAI to reduce the dimensionality of the data. Hyperparameter optimization was performed based on grid search and included a test-wise disabling of each pre-processing algorithm (Scaling, PCA, F-test based feature selection). Hence, the grid search for the inner cross-validation included a pipeline variant without the algorithm, or in other words: whether to use the respective algorithm at all was itself a hyperparameter. This approach resulted in different numbers of features available for the classification algorithm, depending on whether PCA or F-test based univariate feature selection was used (see supplemental methods for details). As classifier, we used either a support vector machine (with the parameter c as well as the choice of the kernel, either linear or radial basis function, as hyperparameters and all other parameters at their default value set by scikit learn) or a random forest classifier (with maximum depth of each tree as a hyperparameter and all other parameters again at their default value). To test for statistical significance of the models, we repeated the analyses 1000 times with permuted labels.

There were only small deviations from this approach in the different pipelines. For detailed parameters and differences between the pipelines, see supplemental methods.

3. Results

Predictions across both datasets performed never outperformed chance.

Protect-AD. For Protect-AD, the full prediction approach resulted in a mean balanced accuracy of 0.51 on the second-level learner, with all separate first-level learners likewise performing on chance level only (table 2). Despite being on chance level on average, we found that the corresponding area-under-the-curve for classification varied considerably across iterations. Calibration plots also indicated an inability of the second-level learner to classify responders and non-responders correctly (supplemental figure 2). Results for the exploratory analyses including hyperparameter optimization were similarly non-significant, except for the sanity check that reached a mean balanced accuracy of 99.9% (supplemental Table 1). Given that no learner performed significantly better than chance, we did not examine the importance of individual features for classification.

SpiderVR. For the Spider-VR dataset we were able to replicate the findings of the previous manuscript (Leehr et al., 2021) when using only clinical and demographic data (balanced accuracy of 0.613; see supplemental methods and table 3). However, using "static" resting state data, balanced accuracy never significantly differed from chance (between 0.498 and 0.546; see table 3), with only little difference between outer validation loops and the different approaches concerning pre- to post-treatment or pre-treatment to FU prediction, as well as predicting SPQ or BAT response. Additionally, combining the static resting state data and clinical and demographic data did not yield significant prediction results either(between 0.479 and 0.571; see table 3).This did not change when using sliding-window data (between 0.488 and 0.534) or edge-functional connectivity data (between 0.454 and 0.553; see supplemental table S3), or dimensional targets (R^2 between $-.098$ and $-.15$; see supplemental table S5) as basis for the predictions. This lack of predictive accuracy is not due to the pipeline, as the pipelines with the synthetic data (sanity check) reached a balanced accuracy of 1. Supplemental table S4 shows detailed results of all ten outer folds for all classification approaches. Supplemental table S6 shows this information for the dimensional approaches. Again, we did not examine the importance of individual features for classification given the lack of classifiers performing significantly better than chance.

Table 2
Mean prediction measurements of first and second level learner across 100 iterations in the Protect-AD dataset.

	Classifier	Accuracy (SD)	Balanced Accuracy (SD)	Sensitivity (SD)	Specificity (SD)
First level learners					
Demographics and Clinical data	random forest	0.465 (0.073)	0.465 (0.073)	0.460 (0.121)	0.470 (0.118)
Connectivity	random forest	0.504 (0.073)	0.504 (0.073)	0.535 (0.118)	0.473 (0.120)
Graph Metrics	random forest	0.503 (0.064)	0.503 (0.064)	0.512 (0.114)	0.494 (0.110)
Second level learner					
Random Forest	random forest	0.505 (0.076)	0.505 (0.076)	0.524 (0.118)	0.487 (0.125)

SD: standard deviation.

Table 3
Mean prediction measurements of the different pipelines based on clinical and demographic data, resting-state and combined resting-state and clinical data in the Spider-VR dataset. Classifier is the one chosen in the best performing outer fold.

Label	Classifier	Accuracy (SD)	Balanced Accuracy (SD)	Sensitivity (SD)	Specificity (SD)
Clinical and demographic data					
SPQ pre post	random forest	0.6 (0.11)	0.613 (0.097)	0.687 (0.159)	0.539 (0.162)
Static RS					
SPQ pre post	random forest	0.547 (0.12)	0.546 (0.12)	0.571 (0.12)	0.521 (0.21)
SPQ pre FU	SVC	0.536 (0.11)	0.498 (0.12)	0.566 (0.12)	0.43 (0.20)
BAT pre post	SVC	0.511 (0.08)	0.509 (0.08)	0.512 (0.13)	0.506 (0.18)
BAT pre FU	random forest	0.514 (0.06)	0.512 (0.08)	0.491 (0.17)	0.534 (0.24)
Static RS combined with clinical and demographic data					
SPQ pre post	SVC	0.563 (0.08)	0.571 (0.08)	0.617 (0.17)	0.526 (0.13)
SPQ pre FU	random forest	0.548 (0.12)	0.479 (0.13)	0.583 (0.19)	0.375 (0.33)
BAT pre post	random forest	0.5 (0.11)	0.499 (0.10)	0.501 (0.15)	0.497 (0.19)
BAT pre FU	SVC	0.512 (0.14)	0.500 (0.15)	0.555 (0.16)	0.445 (0.22)

SPQ: Spider Phobia Questionnaire, FU: Follow up, BAT: Behavioral Avoidance Test, SVC: Support Vector Classifier, SD: standard deviation.

4. Discussion

In these independent studies, we applied sophisticated machine learning analysis pipelines on resting state data from two large-scale trials in anxiety disorders in order to investigate whether neuroimaging (resting state) data provide good to very good prediction accuracies in large clinical samples, and to examine the incremental contribution of neuroimaging beyond clinical and demographic data to the final prediction performance. Contrary to our hypotheses, resting state data was not able to predict exposure-based CBT responses above chance level, and did not provide incremental predictive power beyond the sociodemographic and clinical data. This was true for both datasets, for the main analyses as well as for a substantial number of exploratory analyses conducted on these datasets. Of note, prediction accuracy was even reduced when adding RS data to the feature set of clinical and demographic data in the Spider-VR sample instead of improving, suggesting that RS data introduced considerable noise in the feature set. Accounting for the considerable heterogeneity in methodological approaches in the field, we were able to show a robust null-result of predictive performance in two separate powerful analysis pipelines.

Predicting with an accuracy higher than chance is a comparably easy target in machine learning and substantially below clinical utility, which most likely requires accuracies beyond mere statistical significance to inform clinical decisions. This is true particularly for datasets with relatively equal group sizes (responders and non-responders) such as these. It was surpassed by the vast majority of published prediction studies using clinical and sociodemographic data (Vieira et al., 2022; Hilbert et al., 2021; Hilbert et al., 2020; Hornstein et al., 2021; Taubitz et al., 2022; Wallert et al., 2022; Leehr et al., 2021; Symons et al., 2020) and outperformed by previously published studies using neuroimaging data (Hahn et al., 2015; Månsson et al., 2015; Frick et al., 2020). Considering this background literature, the inability of rs-fMRI data to surpass chance level for the prediction of CBT outcomes in two large-scale datasets was surprising. This is even more the case given the high plausibility of resting state data for constructing outcome prediction models: examining resting state functional connectivity is a reliable approach to robust large-scale brain networks (Yang et al., 2020), several of which play major roles in psychopathology (Menon, 2011): For anxiety disorders, this traditionally includes the salience network, central executive network, default-mode network and ventral attention network (Sylvester et al., 2012), while alternative taxonomies add for example a negative affect network (Williams, 2016). The brain areas implicated in these networks are largely in agreement with results from recent and historical meta-analyses (Chavanne and Robinson, 2021; Etkin and Wager, 2007). On a conceptual level, network dysfunction has been related to functional changes in perception, cognition, emotion, and behavior (Sylvester et al., 2012; Williams, 2016). The high plausibility of resting state connectivity data for predicting psychotherapy outcomes has been further underscored by several studies reporting high prediction accuracies ranging between 70–81% for social anxiety disorder, post-traumatic stress disorder and obsessive-compulsive disorder (Whitfield-Gabrieli et al., 2016; Reggente et al., 2018; Zhutovsky et al., 2019; Zhutovsky et al., 2021). However, all of these studies used samples with $n < 50$. This severely limits the ability to adequately evaluate classifiers, as this process is particularly dependent on sufficiently large test sets (Flint et al., 2021) which cannot be provided by these sample sizes. On the contrary, both datasets used in this study are substantially larger and include neuroimaging data collected at more than one site, making the datasets overall more similar to real-world use-cases. The inability to predict outcomes in these datasets therefore calls for caution regarding the interpretation of promising results from small samples and regarding the potential of rs-fMRI for treatment outcome prediction. This result and interpretation are further in line with a recent study aiming to predict treatment response to escitalopram in a moderately sized multi-site sample: in this study, the prediction based on resting state connectivity from baseline alone also was not able to surpass chance level, while changes in connectivity from baseline to week two achieved between 64.0–69.4% accuracy (Harris et al., 2022). The authors related this finding to early change. Including information from the first few sessions of a treatment may be a promising avenue for further research on theranostic biomarkers.

It is important to note that the null-findings in this study were achieved using a powerful methodological approach which mirrored the heterogeneity of analysis pipelines in real-world settings. Random forests and Support Vector Classifiers are particularly well-suited for tabular data such as ours, with random forests even outperforming deep learning approaches under such settings (Grinsztajn et al.). Moreover, in both datasets, we combined different data modalities including fMRI, clinical, and demographic data, as it has been recommended in the literature (Chekroud et al., 2021). On the Protect-AD dataset, we even integrated first-level classifiers in an ensemble learning approach, which often outperforms individual classifiers (Polikar, 2006). The convincing results of the concluding sanity check analysis conducted for this dataset generated further trust in the overall performance of the analytic pipeline. Given the inability to predict CBT outcomes in our main approach, we conducted additional exploratory post-hoc analyses on the Protect-AD dataset. As prediction accuracies did not increase for the approach using ROIs from the whole-brain, we can reasonably conclude that the lack of meaningful prediction was not related to an inappropriate selection of ROIs. For Protect-AD, our selection of ROIs based on the literature (Chavanne and Robinson, 2021; Santos et al., 2019; Lueken and Hahn, 2016) constitutes feature selection by prior knowledge, which has been demonstrated to outperform data-driven feature selection approaches in some cases in the neuroimaging literature (Chu et al., 2012). We initially combined this approach with data driven feature selection methods in the analysis pipeline. As feature weights and selected features showed substantial variability across iterations on the Protect-AD dataset, we employed further feature selection methods, including the extraction of particularly stable features (Nogueira et al., 2018; Meinshausen and Bühlmann, 2010) and a complementary approach using all available ROIs from the whole brain in additional exploratory analyses. As predictive performances did not differ between both hypothesis-based and complementary data-driven feature selection, we concluded that the lack of satisfactory results from feature selection was grounded in the lack of predictive information in the underlying resting state in both datasets. This argument is further corroborated by the lack of significant prediction when combining resting state and clinical and demographic data in the Spider-VR dataset: this can be interpreted as the resting state data decreasing the signal-to-noise ratio in the feature set, thus deteriorating the predictive value of the overall dataset.

Across both datasets, we also implemented sophisticated methodology on the feature level itself by using graph metrics (main analysis), dynamic resting state functional connectivity (sliding windows analysis; exploratory) and edge-centric functional connectivity (exploratory). Graph metrics have received increasing attention as they excel at describing overall characteristics and topology of the large scale networks in the brain (Farahani et al., 2019). Sliding window approaches try to address a commonly raised concern regarding resting state paradigms, namely their lack of temporal resolution: two brain areas may be strongly correlated for part of the paradigm time, but not at all during the rest of the time, making the correlation non-significant for the entire paradigm. Separating different time periods from each other and estimating correlations within these periods might thus show meaningful correlations otherwise overlooked (Yan et al., 2020; Allen et al., 2014). Going even one step further, avoiding temporal blurring altogether, edge-functional connectivity examines co-fluctuations, i.e. shared patterns of activity from one MRI image to the next (Faskowitz et al., 2020; Esfahlani et al., 2020; Novelli and Razi, 2022). Thus, this approach should be even more successful in uncovering dynamic functional connectivity. A growing body of literature corroborates the feasibility and relevance of this approach (Jo et al., 2021; Chumin et al., 2022). However, none of the approaches mentioned above resulted in a meaningful pattern for prediction.

Despite this methodological rigor, the current investigation has some limitations. Although our sample-sizes considerably surpass those of previous studies, they are still inferior to comparable studies on sociodemographic, clinical and routine data that used more than thousand (Hilbert et al., 2020; Hornstein et al., 2021; Symons et al., 2020) to tens of thousands (Wolff et al., 2020) of patients. As model training and evaluation are dependent on the train and test set sizes, larger samples allow for the construction of more robust models and a less biased assessment of model performance, especially when working with high numbers of potential features, as in the case of the sliding-window and edge-functional connectivity approaches. However, sample sizes such as ours may be suitable for an initial assessment of model performance and considerably exceed previous investigations. Furthermore, both included studies were conducted at different sites and MRI data was collected on different scanners, adding additional variance in our data set. One way to tackle this problem would have been to harmonize the data across the sites (Eshaghzadeh Torbati et al., 2021; Yamashita et al., 2019; Yu et al., 2018). However, since harmonization are population-based, they bear the risk of data leakage, so that test set subject information is used during harmonization and thus during classifier training. Alternatively, harmonization could be included in the train-test-split, however this would lead to increased variability between splits. Additionally, in real-life scenarios, a prediction algorithm might be deployed to new sites not included in the training procedure, thus a successful prediction algorithm should not be dependent on the data collection location or hardware. We therefore argue that the heterogeneity adds to our external validity, preventing overfitting. Second, by employing analytic teams that applied different powerful machine learning pipelines on the data, we mirrored the heterogeneity of analysis pipelines in real-world settings. Here, each team analyzed its own dataset, but teams did not directly replicate each other's analysis. Previous research has shown that the specific analysis strategy can have a large impact on the analysis outcome for MRI data (Botvinik-Nezer et al., 2020). But the fact that we found very similar results for different analysis conducted by different teams in different samples, indicates an overall lack of meaningful information for CBT outcome prediction within the broad feature set of resting state functional connectivity data of both datasets. Therefore, an exact replication of analytic strategies across teams and datasets would have made sense only if we had found a significant predictive model in the first place, to ensure the robustness of this analysis. Third, there are several other ways of analyzing resting state activity in addition to the ones applied in our approaches (e.g. decomposition into brain networks using ICA). As the possible combinations between resting-state and machine learning analysis is endless, we cannot definitively rule out that another way of analyzing resting state would have yielded feature sets that would have led to more accurate prediction. But given the amount of applied state-of-the-art approaches and the similarity in results across two independent teams and datasets suggests that the data does not exhibit an apparent predictive signal.

In conclusion, we were not able to predict individual CBT outcomes based on resting state functional connectivity data from two large clinical trials in anxiety disorders. Across a variety of approaches, prediction accuracies were much lower than in comparable yet smaller previous studies. This finding urges caution regarding the interpretation of promising results from small samples and re-iterates that some of the prediction accuracies from these studies may result from overestimation due to homogeneous data and weak cross-validation schemes (Varoquaux, 2018; Varoquaux et al., 2017; Flint et al., 2021). While neuroimaging may still set out to prove its added value for the prediction of treatment outcomes and in precision psychotherapy, adequately powered samples are a necessary prerequisite for an initial evaluation of predictive performance and for a subsequently identification of the most promising candidates.

We used resting-state fMRI data from two large clinical trials to predict individual CBT outcomes. Contrary to previous results, resting-state data never provided incremental predictive power beyond sociodemographic and clinical data. These findings urge caution regarding the interpretation of promising prediction results based on neuroimaging data from small samples.

Data and code availability statement

The data of both studies are available on reasonable request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Code for the main analyses in both datasets is openly available at [https://github.com/wwu-trap/RS_Prediction_in_two_samples_paper].

Funding statement

The German research consortia PROTECT-AD ("Providing Tools for Effective Care and Treatment of Anxiety Disorders: Outcome, Mediators and Moderators of Enhanced Extinction Learning"; PI: HUW), is one out of nine research consortia in the German federal research program "research network on mental disorders", funded by the German Federal Ministry of Education and Research (BMBF; www.fzpe.de; project no. 01EE1402) and comprises six projects. MRI data were collected as part of project P4 (PIs: TK and BS, University of Marburg; project no. 01EE1402E). Clinical recruitment and data collection was conducted in project P1 (grant 01EE1402A). SpiderVR was funded by the Deutsche Forschungsgemeinschaft (DFG)—Projektnummer 44541416-TRR 58 (CRC-TRR58, Project C08 to MJ and TS, Project C09 to UD and UL, Project C07 to TS and MH). YY was supported by the Hessisches Ministerium für Wissenschaft und Kunst (HMWK; project "The Adaptive Mind"; to TK, BS, WR). AW was supported by the German Research Foundation (DFG, STR1146/18-1, 437618337 to BS, and KI588/22-1 to TK as part of the FOR2107). This work was funded by the Deutsche Forschungsgemeinschaft (DFG)— FOR5187 (Projektnummer 442075332).

CRediT authorship contribution statement

Kevin Hilbert: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Data curation, Conceptualization. **Joscha Böhnlein:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Charlotte Meinke:** Writing – review & editing, Investigation, Data curation. **Alice V. Chavanne:** Writing – review & editing, Validation, Software, Formal analysis, Data curation. **Till Langhammer:** Writing – review & editing, Validation, Software, Data curation. **Lara Stumpe:** Investigation, Data curation. **Nils Winter:** Writing – review & editing, Validation, Software, Formal analysis. **Ramona Leenings:** Writing – review & editing, Software, Data curation. **Dirk Adolph:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Volker Arolt:** Funding acquisition, Conceptualization. **Sophie Bischoff:** Resources, Project administration, Funding acquisition, Conceptualization. **Jan C. Cwik:** Resources, Project administration, Funding acquisition, Conceptualization. **Jürgen Deckert:** Supervision, Project administration, Funding acquisition, Conceptualization. **Katharina Domschke:** Resources, Project administration, Funding acquisition, Conceptualization. **Thomas Fydrich:** Resources, Project administration, Funding acquisition, Conceptualization. **Bettina Gathmann:** Writing – review & editing, Resources, Investigation. **Alfons O. Hamm:** Supervision, Investigation, Funding acquisition, Conceptualization. **Ingmar Heinig:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Martin J. Herrmann:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Maike Hollandt:** Project administration, Investigation, Funding acquisition, Conceptualization. **Jürgen Hoyer:** Supervision, Project administration, Funding acquisition, Conceptualization. **Markus Junghöfer:** Writing – review & editing, Supervision, Investigation, Funding acquisition, Conceptualization. **Tilo Kircher:** Supervision, Project administration, Funding acquisition, Conceptualization. **Katja Koelkebeck:** Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization. **Martin Lotze:** Supervision, Resources, Project administration, Conceptualization. **Jürgen Margraf:** Supervision, Project administration, Funding acquisition, Conceptualization. **Jennifer L.M. Mumm:** Project administration, Investigation, Conceptualization. **Peter Neudeck:** Supervision, Project administration, Funding acquisition. **Paul Pauli:** Resources, Project administration, Funding acquisition, Conceptualization. **Andre Pittig:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Data curation, Conceptualization. **Jens Plag:** Supervision, Investigation, Funding acquisition, Conceptualization. **Jan Richter:** Supervision, Project administration, Methodology, Data curation, Conceptualization. **Isabelle C. Ridderbusch:** Project administration, Investigation, Funding acquisition, Conceptualization. **Winfried Rief:** Resources, Project administration, Funding acquisition, Conceptualization. **Silvia Schneider:** Supervision, Project administration, Investigation, Funding acquisition, Conceptualization. **Hanna Schwarzmeier:** Writing – review & editing, Investigation, Supervision, Data curation. **Fabian R. Seeger:** Writing – review & editing, Supervision, Investigation, Data curation. **Niklas Siminski:** Writing – review & editing, Project administration. **Benjamin Straube:** Validation, Project administration, Funding acquisition, Conceptualization. **Thomas Straube:** Resources, Project administration, Funding acquisition, Conceptualization. **Andreas Ströhle:** Resources, Project administration, Funding acquisition, Conceptualization. **Hans-Ulrich Wittchen:** Funding acquisition, Conceptualization. **Adrian Wroblewski:** Supervision, Project administration, Investigation. **Yunbo Yang:** Writing – review & editing, Software, Methodology, Formal analysis, Data curation. **Kati Roesmann:** Writing – review & editing, Supervision, Project administration, Investigation, Data curation. **Elisabeth J. Leehr:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Data curation. **Udo Dannlowski:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Ulrike Lueken:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare there is no conflict of interests.

Data availability

Data will be made available on request.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2024.120639.

References

Allen, E.A., Damaraju, E., Plis, S.M., Erhardt, E.B., Eichele, T., Calhoun, V.D., 2014. Tracking whole-brain connectivity dynamics in the resting state. Cereb. Cortex 24 (3), 663–676.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C.F., Dreber, A., Huber, J., Johannesson, M., et al., 2020. Variability in the analysis of a single neuroimaging dataset by many teams. Nature 582, 84–88.

Bouckaert, R.R., Frank, E., 2004. Evaluating the replicability of significance tests for comparing learning algorithms. In: Dai, H, Srikant, R, Zhang, C (Eds.), Advances in Knowledge Discovery and Data Mining. PAKDD 2004, Lecture Notes in Computer Science, Vol 3056. Springer, Heidelberg.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Bullmore, E., Sporns, O., 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. Nat. Rev. Neurosci. 10 (3), 186–198.

Bzdok, D., Krzywinski, M., Altman, N., 2017. Points of significance: machine learning: a primer. Nat. Methods 14 (12), 1119–1120.

Chavanne, A.V., Robinson, O.J., 2021. The overlapping neurobiology of induced and pathological anxiety: a meta-analysis of functional neural activation. Am. J. Psychiatry 178 (2), 156–164.

Chekroud, A.M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., et al., 2021. The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry 20, 154–170.

Chu, C., Hsu, A.L., Chou, K.H., Bandettini, P., Lin, C.P., Neuroimaging AsD, 2012. Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. Neuroimage 60 (1), 59–70.

Chumin, E.J., Faskowitz, J., Esfahlani, F.Z., Jo, Y., Merritt, H., Tanner, J., et al., 2022. Cortico-subcortical interactions in overlapping communities of edge functional connectivity. Neuroimage 250, 118971.

Cyr, M., Pagliaccio, D., Yanes-Lukin, P., Goldberg, P., Fontaine, M., Rynn, M.A., et al., 2021. Altered fronto-amygdalar functional connectivity predicts response to cognitive behavioral therapy in pediatric obsessive-compulsive disorder. Depress. Anxiety 38 (8), 836–845.

Deckert, J., Angelika, E, 2019. Predicting treatment outcome for anxiety disorders with or without comorbid depression using clinical, imaging and (epi)genetic data. Curr. Opin. Psychiatry 32 (1), 1–6.

Esfahlani, F.Z., Jo, Y., Faskowitz, J., Byrge, L., Kennedy, D.P., Sporns, O., et al., 2020. High-amplitude cofluctuations in cortical activity drive functional connectivity. Proc. Natl. Acad. Sci. USA 117 (45), 28393–28401.

Eshaghzadeh Torbati, M., Minhas, D.S., Ahmad, G., O'Connor, E.E., Muschelli, J., Laymon, C.M., Yang, Z., Cohen, A.D., Aizenstein, H.J., Klunk, W.E., Christian, B.T., Hwang, S.J., Crainiceanu, C.M., Tudorascu, D.L., 2021. A multi-scanner neuroimaging data harmonization using RAVEL and ComBat. Neuroimage 245, 118703.

Etkin, A., Wager, T.D., 2007. Functional neuroimaging of anxiety: a meta-analysis of emotional processing in PTSD, social anxiety disorder, and specific phobia. Am. J. Psychiat. 164 (10), 1476–1488.

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., et al., 2016. The human Brainnetome atlas: a new brain atlas based on connectional architecture. Cereb. Cortex 26 (8), 3508–3526.

Farahani, F.V., Karwowski, W., Lighthall, N.R., 2019. Application of graph theory for identifying connectivity patterns in human brain networks: a systematic review. Front. Neurosci. 13, 585.

Faskowitz, J., Esfahlani, F.Z., Jo, Y., Sporns, O., Betzel, R.F., 2020. Edge-centric functional network representations of human cerebral cortex reveal overlapping system-level architecture. Nat. Neurosci. 23 (12), 1644–1654.

Flint, C., Cearns, M., Opel, N., Redlich, R., Mehler, D.M.A., Emden, D., et al., 2021. Systematic misestimation of machine learning performance in neuroimaging studies of depression. Neuropsychopharmacology 46 (8), 1510–1517.

Frick, A., Engman, J., Alaie, I., Bjorkstrand, J., Gingnell, M., Larsson, E.M., et al., 2020. Neuroimaging, genetic, clinical, and demographic predictors of treatment response in patients with social anxiety disorder. J. Affect. Disord. 261, 230–237.

Fullana, M.A., Zhu, X., Alonso, P., Cardoner, N., Real, E., Lopez-Sola, C., et al., 2017. Basolateral amygdala-ventromedial prefrontal cortex connectivity predicts cognitive behavioural therapy outcome in adults with obsessive-compulsive disorder. J. Psychiatry Neurosci. 42 (6), 378–385.

Grinsztajn L., Oyallon E., Varoquaux G. Why do tree-based models still outperform deep learning on tabular data? *arXiv* preprint: 2207.08815.

Hahn, T., Kircher, T., Straube, B., Wittchen, H.U., Konrad, C., Strohle, A., et al., 2015. Predicting treatment response to cognitive behavioral therapy in panic disorder with agoraphobia by integrating local neural information. JAMA Psychiat. 72 (1), 68–74.

Hamm, A, 2006. Spezifische Phobien. Hogrefe, Göttingen.

Harris, J.K., Hassel, S., Davis, A.D., Zamyadi, M., Arnott, S.R., Milev, R., et al., 2022. Predicting escitalopram treatment response from pre-treatment and early response resting state fMRI in a multi-site sample: a CAN-BIND-1 report. Neuroimage Clin. 35, 103120.

Heinig, I., Pittig, A., Richter, J., Hummel, K., Alt, I., Dickhover, K., et al., 2017. Optimizing exposure-based CBT for anxiety disorders via enhanced extinction: design and methods of a multicentre randomized clinical trial. Int. J. Methods Psychiatr. Res. 26 (2).

Hilbert, K., Jacobi, T., Kunas, S.L., Elsner, B., Reuter, B., Lueken, U., et al., 2021. Identifying CBT non-response among OCD outpatients: a machine-learning approach. Psychother.. Res. 31 (1), 52–62.

Hilbert, K., Kunas, S.L., Lueken, U., Kathmann, N., Fydrich, T., Fehm, L., 2020. Predicting cognitive behavioral therapy outcome in the outpatient sector based on clinical routine data: a machine learning approach. Behav. Res. Ther. 124, 103530.

Hornstein, S., Forman-Hoffman, V., Nazander, A., Ranta, K., Hilbert, K., 2021. Predicting therapy outcome in a digital mental health intervention for depression and anxiety: a machine learning approach. Dig. Health 7, 20552076211060659.

Janssen, R.J., Mourão-Miranda, J., Schnack, H.G., 2018. Making individual prognoses in psychiatry using neuroimaging and machine learning. Biol. Psychiatry 3 (9), 798–808.

Jo, Y., Esfahlani, F.Z., Faskowitz, J., Chumin, E.J., Sporns, O., Betzel, R.F., 2021. The diversity and multiplexity of edge communities within and between brain systems. Cell Rep. 37 (7), 110032.

Keuken, M.C., Bazin, P.L., Backhouse, K., Beekhuizen, S., Himmer, L., Kandola, A., et al., 2017. Effects of aging on T(1), T(2)*, and QSM MRI values in the subcortex. Brain Struct. Funct. 222 (6), 2487–2505.

Langhammer T., Hilbert K., Wroblewski A., Ridderbusch I.C., Yang Y., Richter J. et al. Resting-state functional connectivity in anxiety disorders: a multicenter fMRI study. in preparation.

Leehr, E.J., Roesmann, K., Bohnlein, J., Dannlowski, U., Gathmann, B., Herrmann, M.J., et al., 2021. Clinical predictors of treatment response towards exposure therapy in virtuo in spider phobia: a machine learning and external cross-validation approach. J. Anxiety. Disord. 83, 102448.

Leehr, E.J., Seeger, F.R., Böhnlein, J., Gathmann, B., Straube, T., Roesmann, K., Junghöfer, M., Schwarzmeier, H., Siminski, N., Herrmann, M.J., Langhammer, T., Goltermann, J., Grotegerd, D., Meinert, S., Winter, N.R., Dannlowski, U., Lueken, U., 2024. Association between resting-state connectivity patterns in the defensive system network and treatment response in spider phobia—a replication approach. Transl Psychiatry 14, 137.

Leenings, R., Winter, N.R., Plagwitz, L., Holstein, V., Ernsting, J., Sarink, K., et al., 2021. PHOTONAI-A Python API for rapid machine learning model development. PLoS ONE 16 (7), e0254062.

Lueken, U., Hahn, H., 2020. Personalized mental health: artificial intelligence technologies for treatment response prediction in anxiety disorders. In: Baune, BT (Ed.), Personalized Psychiatry. Academic Press, London, pp. 201–213.

Lueken, U., Hahn, T., 2016. Functional neuroimaging of psychotherapeutic processes in anxiety and depression: from mechanisms to predictions. Curr. Opin. Psychiatry 29 (1), 25–31.

Månsson, K.N.T., Frick, A., Boraxbekk, C.-J., Marquand, A.F., Williams, S.C.R., Carlbring, P., et al., 2015. Predicting long-term outcome of Internet-delivered cognitive behavior therapy for social anxiety disorder using fMRI and support vector machine learning. Transl. Psychiatry 5, e530.

Matza, L.S., Morlock, R., Sexton, C., Malley, K., Feltner, D., 2010. Identifying HAM-A cutoffs for mild, moderate, and severe generalized anxiety disorder. Int. J. Methods Psychiatr. Res. 19, 223–232.

Meinshausen, N., Bühlmann, P., 2010. Stability selection. J. R. Stat. Soc. Ser. B 74 (4), 417–473.

Menon, V., 2011. Large-scale brain networks and psychopathology: a unifying triple network model. Trends. Cogn. Sci. 15 (10), 483–506.

Nadeau, C., Bengio, Y., 2003. Inference for the generalization error. Mach. Learn. 52, 239–281.

Neudorfer, C., Germann, J., Elias, G.J.B., Gramer, R., Boutet, A., Lozano, A.M., 2020. A high-resolution in vivo magnetic resonance imaging atlas of the human hypothalamic region. Sci. Data 7 (1), 305.

Nogueira, S., Sechidis, K., Brown, G., 2018. On the stability of feature selection algorithms. J. Mach. Learn. Res. 18, 1–54.

Novelli, L., Razi, A., 2022. A mathematical perspective on edge-centric brain functional connectivity. Nat. Commun. 13 (1), 2693.

Opris, D., Pintea, S., Garcia-Palacios, A., Botella, C., Szamoskozi, S., David, D., 2012. Virtual reality exposure therapy in anxiety disorders: a quantitative meta-analysis. Depress. Anxiety. 29 (2), 85–93.

Ozomaro, U., Wahlestedt, C., Nemeroff, C.B., 2013. Personalized medicine in psychiatry: problems and promises. BMC. Med. 11, 132.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Penny, W., Friston, K., Ashburner, J., Kiebel, S., Nichols, T.E., 2006. Statistical Parametric Mapping: The Analysis of Functional Brain Images. Academic Press.

Pico-Perez, M., Fullana, M.A., Albajes-Eizagirre, A., Vega, D., Marco-Pallares, J., Vilar, A., et al., 2022. Neural predictors of cognitive-behavior therapy outcome in anxiety-related disorders: a meta-analysis of task-based fMRI studies. Psychol. Med. 1–9.

Pittig, A., Heinig, I., Goerigk, S., Thiel, F., Hummel, K., Scholl, L., et al., 2021. Efficacy of temporally intensified exposure for anxiety disorders: a multicenter randomized clinical trial. Depress. Anxiety. 38 (11), 1169–1181.

Poldrack, R.A., Huckins, G., Varoquaux, G., 2020. Establishment of best practices for evidence for prediction: a review. JAMA Psychiat 77 (5), 534–540.

Polikar, R., 2006. Ensemble based systems in decision making. IEEE Circ. Syst. Mag. 6 (3), 21–45.

Powers, M.B., Emmelkamp, P.M., 2008. Virtual reality exposure therapy for anxiety disorders: a meta-analysis. J. Anxiety. Disord. 22 (3), 561–569.

Reggente, N., Moody, T.D., Morfini, F., Sheen, C., Rissman, J., O'Neill, J., et al., 2018. Multivariate resting-state functional connectivity predicts response to cognitive behavioral therapy in obsessive–compulsive disorder. PNAS 115 (9), 2222–2227.

Rolls, E.T., Huang, C.C., Lin, C.P., Feng, J., Joliot, M., 2020. Automated anatomical labelling atlas 3. Neuroimage 206, 116189.

Santos, V.A., Carvalho, D.D., Van Ameringen, M., Nardi, A.E., Freire, R.C., 2019. Neuroimaging findings as predictors of treatment outcome of psychotherapy in anxiety disorders. Prog. Neuropsychopharmacol. Biol. Psychiatry 91, 60–71.

Schwarzmeier, H., Leehr, E.J., Bohnlein, J., Seeger, F.R., Roesmann, K., Gathmann, B., et al., 2020. Theranostic markers for personalized therapy of spider phobia: methods of a bicentric external cross-validation machine learning approach. Int. J. Methods Psychiatr. Res. 29 (2), e1812.

Shear, M.K., Vander Bilt, J., Rucci, P., Endicott, J., Lydiard, B., Otto, M.W., et al., 2001. Reliability and validity of a Structured Interview Guide for the Hamilton Anxiety Rating Scale (SIGH-A). Depress. Anxiety. 13 (4), 166–178.

Sylvester, C.M., Corbetta, M., Raichle, M.E., Rodebaugh, T.L., Schlaggar, B.L., Sheline, Y. I., et al., 2012. Functional network dysfunction in anxiety and anxiety disorders. Trends. Neurosci. 35 (9), 527–535.

Symons, M., Feeney, G.F.X., Gallagher, M.R., Young, R.M., Connor, J.P., 2019. Machine learning vs addiction therapists: a pilot study predicting alcohol dependence treatment outcome from patient data in behavior therapy with adjunctive medication. J. Subst. Abuse Treat. 99, 156–162.

Symons, M., Feeney, G.F.X., Gallagher, M.R., Young, R.M.D., Connor, J.P., 2020. Predicting alcohol dependence treatment outcomes: a prospective comparative study of clinical psychologists versus 'trained' machine learning models. Addiction 115 (11), 2164–2175.

Taubitz, F.-S., Büdenbender, B., Alpers, G.W., 2022. What the future holds: machine learning to predict success in psychotherapy. Behav. Res. Ther. 156, 104116.

Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B., 2017. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. Neuroimage 145 (Pt B), 166–179.

Varoquaux, G., 2018. Cross-validation failure: small sample sizes lead to large error bars. Neuroimage 180 (Pt A), 68–77.

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. Clin. Psychol. Rev. 97, 102193.

Wallert, J., Boberg, J., Kaldo, V., Mataix-Cols, D., Flygare, O., Crowley, J.J., et al., 2022. Predicting remission after internet-delivered psychotherapy in patients with depression using machine learning and multi-modal data. Transl. Psychiatry 12 (1), 357.

Whitfield-Gabrieli, S., 2012. Nieto-Castanon A. Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. Brain Connect. 2 (3), 125–141.

Whitfield-Gabrieli, S., Ghosh, S.S., Nieto-Castanon, A., Saygin, Z., Doehrmann, O., Chai, X.J., et al., 2016. Brain connectomics predict response to treatment in social anxiety disorder. Mol. Psychiatry 21 (5), 680–685.

Williams, L.M., 2016. Precision psychiatry: a neural circuit taxonomy for depression and anxiety. Lancet Psychiat. 3 (5), 472–480.

Wolff, J., Gary, A., Jung, D., Normann, C., Kaier, K., Binder, H., et al., 2020. Predicting patient outcomes in psychiatric hospitals with routine data: a machine learning approach. BMC. Med. Inform. Decis. Mak. 20 (1), 21.

Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., Okada, N., Yamagata, H., Matsuo, K., Hashimoto, R., Okada, G., Sakai, Y., Morimoto, J., Narumoto, J., Shimada, Y., Kasai, K., Kato, N., Takahashi, H., Okamoto, Y., Tanaka, S.C., Kawato, M., Yamashita, O., Imamizu, H., 2019. Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. PLoS Biol. 17 (4), e3000042.

Yan, B., Xu, X., Liu, M., Zheng, K., Liu, J., Li, J., et al., 2020. Quantitative identification of major depression based on resting-state dynamic functional connectivity: a machine learning approach. Front. Neurosci. 14, 191.

Yang, J., Gohel, S., Vachha, B., 2020. Current methods and new directions in resting state fMRI. Clin. Imaging 65, 47–53.

Yu, M., Linn, K.A., Cook, P.A., Phillips, M.L., McInnis, M., Fava, M., Trivedi, M.H., Weissman, M.M., Shinohara, R.T., Sheline, Y.I., 2018. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. Hum. Brain Mapp. 39 (11), 4213–4227.

Zhutovsky, P., Thomas, R.M., Olff, M., van Rooij, S.J.H., Kennis, M., van Wingen, G.A., et al., 2019. Individual prediction of psychotherapy outcome in posttraumatic stress disorder using neuroimaging data. Transl. Psychiatry 9 (1), 326.

Zhutovsky, P., Zantvoord, J.B., Ensink, J.B.M., Op den Kelder, R., Lindauer, R.J.L., van Wingen, G.A., 2021. Individual prediction of trauma-focused psychotherapy response in youth with posttraumatic stress disorder using resting-state functional connectivity. Neuroimage Clin. 32, 102898.