Supplementary Online Content


eAppendix. Methods.

This supplementary material has been provided by the authors to give readers additional information about their work.
Clinical trial information

Within the German research network “Panic-Net”, a multicenter randomized controlled clinical trial of cognitive-behavioral therapy (CBT) for patients with panic disorder with agoraphobia (PD/AG) was conducted. Details on the study design, immediate and long-term treatment effects are reported in (1-3). Eight German centers (Aachen, Berlin-Adlershof, Berlin-Charité, Bremen, Dresden, Greifswald, Münster, and Würzburg) participated in the clinical trial, with four of them also participating in the functional magnetic resonance imaging (fMRI) add-on study (Aachen, Berlin-Charité, Dresden, Münster). Three hundred sixty-nine patients who met DSM-IV-TR criteria for PD/AG were treated with a manualized treatment protocol or assigned to a waiting list that consisted of 12 twice-weekly sessions focusing on behavioral exposure in situ. In- and exclusion criteria and sample characteristics of the fMRI subsample are reported in (4;5). In short, only currently (e.g. 4-week washout period) medication-free patients with a primary diagnosis of PD/AG according to DSM-IV-TR criteria (6), a Hamilton Anxiety Scale Score (SIGH-A (7)) ≥ 18, a Clinical Global Impressions Score (CGI (8)) ≥ 4 and aged 18 to 65 years were included. Inability to comply with the study schedule, clinically significant suicidal intent, diagnostic criteria for any psychotic or bipolar disorder, borderline personality disorder, or current alcohol dependence, medical conditions explaining anxiety symptoms and MRI-related contraindications were followed by exclusion. At baseline 49 quality-controlled baseline fMRI data sets available; a detailed description of measures of quality control in this fMRI multicenter study and the patient flowchart for the study have been published elsewhere (5). Response was defined as a reduction in SIGH-A scores exceeding 50% from baseline to post assessment (4). Responders and nonresponders were equally distributed across sites (site 1: 0/1; site 2: 6/4; site 3: 9/13; site 4: 10/6; χ² (df) = 3.108 (3), P = .375).

After a complete description of the study protocol, written informed consent was obtained from every participant and the protocol was approved by the local ethics committees in each fMRI center according to the Declaration of Helsinki.

fMRI data acquisition, preprocessing and preparation for gaussian process classification

Images were acquired using 3-T Philips Achieva (Aachen, Münster, Germany), 3-T Siemens Trio (Dresden, Germany), and 3-T General Electric Healthcare (Berlin, Germany) scanners. A total of 505 axial functional images (matrix = 64 × 64; 30 slices interleaved; field of view = 230; voxel size = 3.6 × 3.6 × 3.8 mm; TE = 30 milliseconds; TR = 2 seconds), covering the whole brain and positioned parallel to the intercommissural line (anterior commissure-posterior commissure), were recorded. MR images were preprocessed using SPM5 (www.fil.ion.ucl.ac.uk) implemented in MATLAB, version 7.1 (MathWorks, Natick, Mass.). The first five volumes were discarded to minimize T1 saturation effects. A high-pass filter (cutoff period, 128 seconds) was applied to remove low-frequency fluctuations in the blood oxygenation level-dependent (BOLD) signal. Following slice time correction, functional images were temporally and spatially aligned and normalized into a standard stereotactic space (Montreal Neurological Institute template: 2 × 2 × 2 mm). An iterative smoothness equalization (9) procedure was performed with a 12-mm full width at half maximum Gaussian isotropic kernel (which is comparable to a kernel of 8 mm full width at half maximum in a normal smoothing procedure).

First level modelling included the BOLD response for each event type (CS+ paired, CS+ unpaired, CS-, and US) and phase (familiarization, acquisition, and extinction) convolved with the canonical hemodynamic response function used in SPM5 within the framework of the general linear model. In a first step, each phase was separated into an early and a late part to account for temporal aspects and habituation, resulting in 16 regressors (familiarization: early CS+, late CS+; early CS-, late CS-; US; acquisition: early CS-, late CS-, early CS presented with the US (CS+paired); early CS+ without US (CS+unpaired), late CS+paired; late CS+unpaired; US; extinction: early CS-, late CS-; early CS+, late CS+; behavioural assessment). Early and late phases were collapsed to one contrast for the current analysis due to power considerations. Realignment parameters were added as regressors of no interest to account for movement artefacts. Parameter estimates (b) and t-statistic images were calculated for each subject. We used the contrast maps reflecting the main contrasts of interest in a differential fear conditioning task yielding two beta maps per subject (acquisition and extinction phase: CS+ > CS-) for gaussian process (GP) classifier analyses.
Classification procedure

Pattern recognition is a field within the area of machine learning which is concerned with automatic discovery of regularities in data through the use of computer algorithms. Using these regularities, it can classify data into different categories. In the context of neuroimaging, brain images are treated as spatial patterns and pattern recognition approaches are used to identify statistical properties of the data that discriminate between 2 groups of subjects (eg, those who respond to treatment and those who do not respond) or 2 cognitive tasks. A classifier based on pattern recognition is trained by providing examples of the form \( <x, c> \) where \( x \) represents a spatial pattern and \( c \) is the class label (e.g. \( c = +1 \) for responders and \( c = -1 \) for nonresponders). Each spatial pattern (eg, whole brain image) corresponds to a point in the input space and each voxel in the brain image represents 1 dimension of this space. During the training phase, the pattern recognition algorithm finds a decision function that separates the examples in the input space according to the class label. Once the decision function is determined from the training data, it can be used to predict the class label of a new test example. There are different approaches to determine the decision function depending on the learning method used. It is important to have a decision function that not only classifies the training data correctly, but also does the same for the test data, ie, one needs to find a classifier, which is able to generalize well for new examples. Gaussian process classifiers (GPCs) are one example of a pattern recognition algorithm, which gives probabilistic outputs. GPCs are based on Bayesian theory, therefore are guaranteed to handle probability distributions correctly. GPCs are most easily understood as a distribution over functions, and GPC inference consists of applying Bayes’ rule to find the (posterior) function distribution that best approximates the training data. GPC is actually an extension of the GP regression model, and data are classified by applying a latent regression model, which is then constrained to the unit interval to produce probabilistic predictions. In this work, GPC was performed using a customized version of the gaussian processes for machine learning (GPML) toolbox for Matlab (http://www.gaussianprocess.org/gpml). We used a linear covariance function and estimated hyper-parameters controlling bias and regularization using an empirical Bayesian approach.

Leave-one-out cross-validation and regional significance testing

The leave-one-out cross-validation (LOO-CV) done in order to predict a subject’s probability to be a responder (\( p_{\text{GP}} \)) and to assess generalizability was conducted as follows: In each leave-one-out run, we used data from all but 1 subject (\( S-1 \) of the \( S \) subjects) to train the classifier. Subsequently, the \( p_{\text{GP}} \) of the remaining subject, which was so far unseen by the algorithm, was calculated. This procedure was repeated \( S \) times, each time leaving out a different subject, yielding each participant’s \( p_{\text{GP}} \) for each condition. In order to determine which single brain areas might contain predictive information regarding therapeutic response, we conducted an atlas-based approach, repeating the procedure outlined above for each of 55 regions drawn from the Harvard-Oxford Brain Atlas as described in Carter et al. (10) separately. For each of the 2 conditions (acquisition and extinction phase) and 55 regions, this yielded a \( p_{\text{GP}} \) as described above. Additionally, we evaluated the performance of the single classifiers (one for each region) by converting the predictive probabilities to categorical predictions. This was achieved by applying a threshold which categorized a subject as a responder (nonresponder), if his/her \( p_{\text{GP}} \) was larger than 0.5 (<0.5). Accuracies were calculated as the ratio of correct predictions to number of cases for each classifier. To establish whether the observed GPC accuracies were statistically significant, we ran each classifier 1000 times with randomly permuted labels and counted the number of permutations which achieved higher accuracy than the one observed with the true labels. The \( P \) value was then calculated by dividing this number by 1000.

Integration of predictive probabilities

In order to obtain a prediction of therapeutic response based on whole-brain information, a second linear classifier was applied which integrates the predictive classification probabilities obtained from the 55 leave-one-out classifiers. Specifications exactly mirror those used for each of the 55 regions before. To calculate the overall prediction accuracy of this approach while attempting to avoid overfitting, a nested leave-one-out procedure was implemented. Specifically, the regional and whole-brain models were trained by cross-validation on the training set only. Then the regional test data were predicted using the weights learned from the training data. Finally, the whole-brain test data were predicted using the weights learned from the training data. This procedure ensured that at each CV fold, the test set was completely independent of the training set. In each nested leave-one-out run, we used the predictive probabilities from all but 1 subject (\( S-1 \) of the \( S \) subjects) to train a classifier model. Subsequently, the class membership of the remaining subject was calculated based on the training model. This procedure was repeated \( S \) times, each time leaving out a different subject, yielding each participant’s predicted overall (ie, whole-brain) class membership. Again, accuracy was calculated as the ratio of correct predictions over number of samples.
To establish whether the observed whole-brain classification accuracies were statistically significant, repeated the procedure described above 1000 times with randomly permuted labels and counted the number of permutations which achieved higher accuracy (based on whole-brain predictions) than the one observed with the true labels. The P value was then calculated by dividing this number by 1000.

**Regional and multivariate spatial mapping**

To visualize the multivariate pattern of brain regions underlying the whole-brain classifier predictions, we computed the mean GPC weight-map from the weight-maps obtained during each cross-validation fold as described in Marquand et al (11). This provides a multivariate estimate of the contribution of each region to classifier performance. One should, however, be aware that the maps describe a multivariate pattern, and one should thus be cautious about interpreting regions in isolation. Generally, weights for each of the 55 regions should be interpreted in the context of the entire multivariate pattern.

Against this background, we additionally present a more readily interpretable, univariate mapping procedure. To this end, we computed classification accuracy for each of the 55 regions separately as described above and display those regions containing predictive information regarding treatment response if the accuracy estimate for this region exceeded chance level ($P < .05$). We corrected for multiple comparisons (for 55 regions) using a false discovery rate (FDR).
References


