Acta Psychiatrica Scandinavica

Acta Psychiatr Scand 2018: 1–11
All rights reserved
DOI: 10.1111/acps.12848

ACTA PSYCHIATRICA SCANDINAVICA

Decoding diagnosis and lifetime consumption in alcohol dependence from grey-matter pattern information

Guggenmos M, Scheel M, Sekutowicz M, Garbusow M, Sebold M, Sommer C, Charlet K, Beck A, Wittchen H-U, Zimmermann US, Smolka MN, Heinz A, Sterzer P, Schmack K. Decoding diagnosis and lifetime consumption in alcohol dependence from grey-matter pattern information.

Objective: We investigated the potential of computer-based models to decode diagnosis and lifetime consumption in alcohol dependence (AD) from grey-matter pattern information. As machine-learning approaches to psychiatric neuroimaging have recently come under scrutiny due to unclear generalization and the opacity of algorithms, our investigation aimed to address a number of methodological criticisms.
Method: Participants were adult individuals diagnosed with AD ($N = 119$) and substance-naïve controls ($N = 97$) ages 20-65 who underwent structural MRI. Machine-learning models were applied to predict diagnosis and lifetime alcohol consumption.
Results: A classification scheme based on regional grey matter attained 74% diagnostic accuracy and predicted lifetime consumption with high accuracy ($r = 0.56$, $P < 10^{-10}$). A key advantage of the classification scheme was its algorithmic transparency, revealing cingulate, insular and inferior frontal cortices as important brain areas underlying classification. Validation of the classification scheme on data of an independent trial was successful with nearly identical accuracy, addressing the concern of generalization. Finally, compared to a blinded radiologist, computer-based classification showed higher accuracy and sensitivity, reduced age and gender biases, but lower specificity.
Conclusion: Computer-based models applied to whole-brain grey-matter predicted diagnosis and lifetime consumption in AD with good accuracy. Computer-based classification may be particularly suited as a screening tool with high sensitivity.

M. Guggenmos[1], M. Scheel[2], M. Sekutowicz[1], M. Garbusow[1], M. Sebold[1], C. Sommer[3], K. Charlet[1], A. Beck[1], H.-U. Wittchen[4,5], U. S. Zimmermann[3], M. N. Smolka[3,6], A. Heinz[1], P. Sterzer[1,*], K. Schmack[1,*]

[1]Department of Psychiatry and Psychotherapy, Campus Charité Mitte, Charité Universitätsmedizin Berlin, Berlin, Germany, [2]Department of Radiology, Charité Universitätsmedizin Berlin, Berlin, Germany, [3]Department of Psychiatry and Psychotherapy, Technische Universität Dresden, Dresden, Germany, [4]Institute for Clinical Psychology and Psychotherapy, Technische Universität Dresden, Dresden, Germany, [5]Research Group Clinical Psychology and Psychotherapy, Department of Psychiatry and Psychotherapy, Ludwig Maximilans Universität Munich,Munich, Germany and [6]Neuroimaging Center, Technische Universität Dresden, Dresden, Germany

Key words: alcohol drinking; grey matter; machine learning; neuroimaging; radiologists

Matthias Guggenmos, Department of Psychiatry and Psychotherapy, Campus Charité Mitte, Charité Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany. E-mail: matthias.guggenmos@charite.de

*Shared last authorship

Accepted for publication December 7, 2017

Significant outcomes

- Whole-brain grey-matter patterns convey rich information about diagnosis and lifetime consumption in AD.
- Novel transparent and robust classification scheme that shows excellent generalization to novel data.
- Compared to an experienced radiologist, computer-based diagnostic classification of alcohol dependence was more sensitive and less biased by age and gender, but less specific.

Limitations

- The observed grey-matter atrophy may not be specific to alcohol dependence.
- Assessment of brain atrophy in patients with alcohol dependence is a rather rare task for radiologists.

Introduction

In recent years, machine-learning methods have been increasingly applied to neuroimaging data and have been able to separate healthy individuals from patients affected by different psychiatric disorders, such as major depressive disorder (1), autism (2) and schizophrenia (3–6). However, despite these encouraging findings, the real-world diagnostic potential of machine-learning techniques has remained unclear due to a number of substantial methodological limitations and criticisms.

In our current work, we present a thorough and rigorous investigation of machine-learning methods applied to neuroimaging data in the paradigmatic case of alcohol dependence (AD). Two features render AD as a model case. First, post-mortem autopsies and magnetic resonance imaging studies have consistently shown that AD is associated with loss of grey matter in the human brain (7). Hence, AD-associated grey-matter loss seems to be an excellent example of a neuroimaging marker that could potentially inform clinical diagnosis. Second, AD can be diagnosed with high certainty on the basis of clinical anamnesis. Thus, as opposed to many other psychiatric disorders, AD suffers much less from label noise, which otherwise compromises machine-learning methods.

Our investigation was based on structural MRI scans from a recent clinical trial investigating the neurobiology of AD (LeAD study, www.lead-studie.de). Our aim was to predict the diagnosis and alcohol lifetime consumption of AD from structural neuroimaging data, thereby addressing some of the most substantial methodological limitations and criticisms raised in the context of machine-learning approaches to psychiatric neuroimaging:

i ‘Black box’ criticism. Many machine-learning algorithms operate as ‘black boxes’ and offer relatively little insight into affected brain structures or disease mechanisms (8). For instance, widely used kernel-based algorithms (e.g. support vector machine classification) provide only limited information about which parts of the data are crucial for diagnostic classification. Here, we unpacked the black box of machine learning by means of the recently proposed weighted robust distance (WeiRD) classifier, which combines high transparency with competitive performance (9,10).

ii Disregard of graded disease measures in binary classification schemes. A drawback of binary classification between case and control groups is the fact that classification may often capture between-group differences that are unrelated to the disease mechanisms. This concern can be alleviated, when the same data are used to predict a graded marker of disease severity. The present work therefore applied the same data used for the diagnostic classification between AD subjects and control subjects to also predict a continuous measure for the cumulative severity of AD (lifetime alcohol consumption).

iii Overfitting and lack of validation on independent data sets. An increasing concern about the application of machine-learning methods in neuroimaging research is their generalizability (11). The fact that the performance of diagnostic classifiers is typically assessed within single data sets bears the risk that classification is partly based on data set-specific characteristics that interact with the conditions of interest (12). Such overfitting problems are exacerbated when sample sizes are small (13) and when classification schemes are ‘fine-tuned’ outside of a cross-validation procedure. For these reasons, it is of utmost importance to test the performance of classifiers on completely independent data sets, a step that is only rarely undertaken in psychiatric research. Here, we validated our classification results on a large data set from an independent trial.

iv Lack of comparison with human expert judgements. Machine-learning methods for diagnostic classification will only then be of clinical relevance if they perform equal or better than human experts on the same data. Therefore, an important—but almost never undertaken—step to evaluate the diagnostic potential of machine -learning methods applied to neuroimaging data is the direct comparison of the diagnostic performance with a radiologist (14). Here, we consulted an independent and blinded radiologist who likewise judged AD diagnosis on the basis of structural MRI scans.

Aims of the study

The main aim of our study was to investigate the predictive capacity of grey-matter patterns for the diagnosis and severity of alcohol dependence, while setting standards for robust, transparent and generalizable methodology. Secondary aims were the assessment of a generalization to a novel data set and the comparison with a human expert radiologist.

Material and methods

Participants

This study was conducted as part of the Learning and Alcohol Dependence (LeAD) study, a larger

German (Berlin, Dresden) programme investigating the neurobiological basis of alcohol dependence [www.lead-studie.de; clinical trial number: NCT01679145 (15–17)]. We assessed 119 individuals aged 20-65 (18 female) meeting criteria of alcohol dependence according to ICD-10 and DSM-IV-TR (American Psychiatric Association 2000) and 97 healthy controls aged 21-65 (16 female) matched in terms of age, gender and smoking (see Table S1).

We used the computer-assisted interview version Composite International Diagnostic Interview [CAPI-CIDI (18,19)] to verify diagnosis criteria of AD in the patient group and to exclude the possibility of AD in control subjects. For inclusion, individuals with AD had to meet criteria for AD for at least 3 years and had to undergo an in-patient detoxification phase (average duration, ±SEM: 22.8 ± 1 days). Exclusion criteria for all subjects were left-handedness [Edinburgh handedness index below 50 (20)], contraindications for MRI, and a history of or current neurological or mental disorders (excluding nicotine dependence in both groups and alcohol abuse in individuals with AD, but including abuse of other drugs). Mental disorders were assessed according to DSM-IV axis one as verified by the computer-assisted interview version Composite International Diagnostic Interview, CAPI-CIDI (18,19). It was ensured that all subjects were free of psychotropic medication (including detoxification treatment) known to interact with the central nervous system for at least four half-lives. Current non-tobacco/non-alcohol drug abuse was confirmed by means of a dedicated urine test.

The study was conducted in accordance with the declaration of Helsinki and approved by local ethics committees of the Technische Universität Dresden and the Charité Universitätsmedizin Berlin. All participants provided written informed consent after receiving a complete description of the study.

MRI data acquisition and preprocessing

High-resolution T1-weighted structural MRI scans were acquired on a 3-Tesla Siemens Trio scanner using a magnetization-prepared rapid gradient echo sequence (repetition time: 1900 ms; echo time: 5.25 ms; flip angle: 9°; field of view: 256 × 256 mm^2; 192 sagittal slices; voxel size: 1 mm isotropic). Data were preprocessed and analysed using SPM12 (http://www.fil.ion.ucl.ac.uk/spm) and VBM 8 (http://dbm.neuro.uni-jena.de/vbm). Images were spatially normalized to a Montreal Neurological Institute (MNI) template, segmented (grey matter, white matter, cerebrospinal fluid) and resampled to 1.5 mm isotropic. As the process of modulation for volumetric grey-matter estimates (21) has been found to decrease the sensitivity for mesoscopic abnormalities such as cortical thinning (22) and to impair classification accuracy (23), we omitted this step. All analyses are thus based on unmodulated images, which test for regional differences in grey-matter concentration (24). Unmodulated were smoothed with an 8 mm isotropic Gaussian kernel to account for intersubject anatomical variability (24,25).

Univariate whole-brain analysis

To ensure that we could replicate previously reported AD-related grey-matter loss (7), we first assessed univariate group differences. To this end, individual grey-matter concentration images were subjected to a second-level random-effects analysis with the factor group and three covariates of no interest (age, gender, site). After model estimation, contrast images and two-sample t-maps were computed for the hypothesis of reduced grey-matter concentration in individuals with AD.

Classification between individuals with AD and controls

A first methodological objective for the classification between AD and controls subjects was the comparison of four analysis schemes that differed in the way how whole-brain concentration images were prepared for the classification step (Fig. 1):

i Whole-brain voxel patterns (Fig. 1a). Grey-matter concentration of all voxels (N = 712 135) within the grey-matter mask was submitted to a classifier.

ii Whole-brain averages (Fig. 1b). The average grey-matter concentration across all voxels within the grey-matter mask was computed as a single feature and used for classification.

iii Parcelled voxel patterns (Fig. 1c). The brain was parcelled into 110 grey-matter areas according to an independent anatomical atlas (26), which is based on manual delineation of whole-head MRI of human volunteers and includes a comprehensive set of both cortical and subcortical brain areas (henceforth referred to as *JHU atlas*). The resulting 110 multivoxel patterns were each submitted to separate classifiers and overall classification was performed through majority vote.

iv Parcelled averages (Fig. 1d). After parcellation (see 3.), the average grey-matter concentration was computed for each parcel. Whole-brain regional pattern was then submitted to a single classifier.

In all cases, masking ensured that only voxels identified as grey matter in the segmentation step were considered for classification. Prior to classification, grey-matter patterns were linearly corrected for variance of no interest related to demographic variables age, gender and site.

A second methodological objective for the classification between cases and controls was the comparison of two classifiers that differed in complexity of involved data transformations: weighted robust distance (9) and support vector machine (27).

Weighted robust distance (WeiRD) is a distance-to-centroid classifier (code available at https://github.com/m-guggenmos/weird). We have shown previously that WeiRD, despite its simplicity, is competitive with other classifiers (SVM, Random Forest) across a range of real-world and simulated classification tasks similar to the present (9). In more recent work (10), we have compared WeiRD to a larger set of classifiers (including support vector machine, Gaussian Naïve Bayes, linear discriminant analysis) in the MEG modality and could likewise confirm the strong performance of WeiRD.

WeiRD operates in a voting scheme, in which each feature (e.g. voxel or regional averages) casts a vote as to which class a given sample pertains. Votes are continuous values that are based on the statistical importance (t-statistic) of features and their L1 distance to class prototypes that are constructed during training. Where i indexes the feature vector, $\mathrm{a_i}$ and $\mathrm{b_i}$ are the arithmetic averages of feature i for two classes A and B (e.g. patients and controls) during training, t_i is the corresponding t-statistic during training (two-sample t-test between A and B along feature dimension i), and x is a novel test sample, the vote v_i associated with feature i is defined as follows:

$$v_i = t_i^2\left[(x_i - a_i)^2 - (x_i - b_i)^2\right] \tag{1}$$

The final prediction p for a test sample is based on the sign of the sum over these votes:

$$p = \mathrm{sgn}\left(\sum_i v_i\right) \tag{2}$$

where positive values of p predict class A and negative values predict class B.

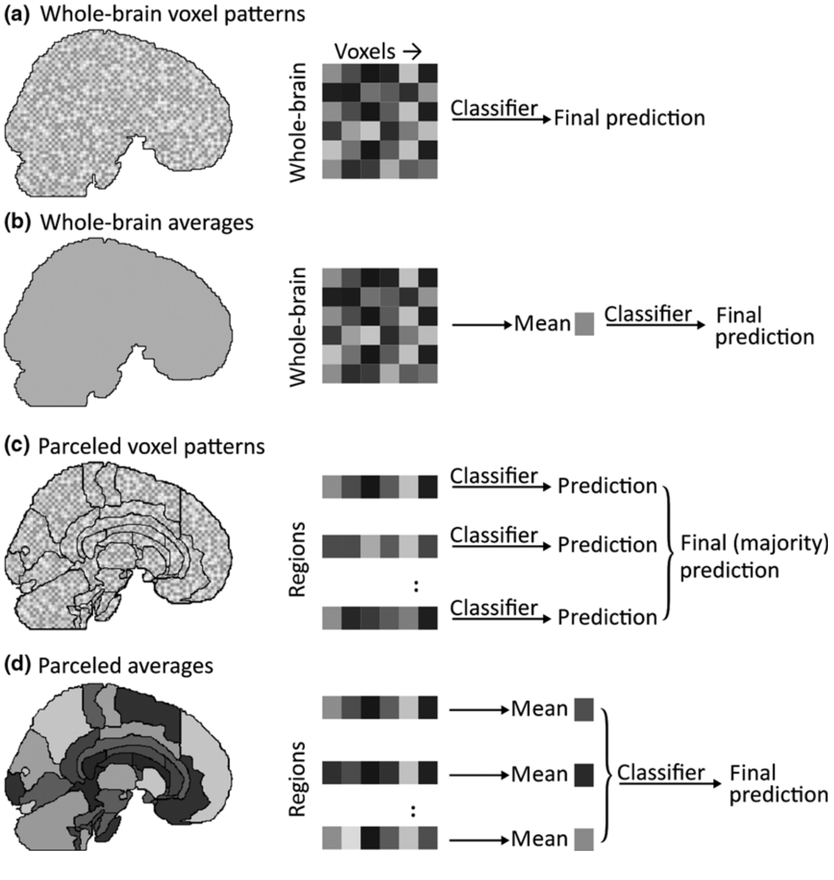

Fig. 1. Analysis schemes. (a) Whole-brain voxel patterns. All voxels within a grey-matter mask are passed to the classifier. (b) Whole-brain averages. Average grey-matter concentration across all voxels is computed and passed as a single feature to classification. (c) Parcelled voxel patterns. Separate classifiers are trained on multivoxel patterns of each region of an atlas-based parcellation. A majority vote is carried out across regional predictions. (d) Parcelled averages. Average values are computed for each region of an atlas-based parcellation and subjected to a single classifier instance.

A key advantage of this vote-based classification scheme is that it makes transparent the contribution of each feature to classification. Moreover, as opposed to nearly all other canonical classifiers, WeiRD requires no parameter tuning. This avoids arbitrary choices of parameters or costly and often unstable optimization procedures. In addition, it increases generalizability, as such optimizations easily lead to overfitting to the particular characteristics of a given data set.

As a reference for the results obtained by the novel WeiRD approach, we chose to additionally perform support vector machine (SVM) classification, for two reasons. First, SVM is the most popular choice for neuroimaging classification problems (28). Second, SVM is still considered to perform best among many current classification techniques when computation time is considered (29) and has been shown to robustly handle data with high dimensionality but few samples per class—two characteristics that are shared with the classification problems in the present study. We used the linear implementation provided by libsvm (30) and used nested cross-validation and grid search to optimize the cost parameter C (range of C: 10^x, x=[-10; 10]).

All classification schemes operated in a leave-one-sample-out cross-validation procedure. To assess the accuracy of classification in a cross-validated setting, we used balanced accuracy (31), which is defined as the average of sensitivity and specificity. Balanced accuracy has two key advantages. First, it avoids the problem of overestimated classification accuracies in unbalanced samples. Second, it enables computing confidence intervals and P-values to reject the null hypothesis of chance-level performance by replacing the conventional point estimate of accuracy (average accuracy across cross-validation folds) by an estimate of the posterior distribution of the balanced accuracy (32). In addition to the balanced accuracy, we provide sensitivity and specificity for our key classification analyses.

Predicting lifetime consumption

To assess whether grey-matter information captured the severity of the disease within the AD group, we predicted the amount of lifetime ethanol consumption based on grey-matter data. Life time alcohol consumption was estimated using the measures provided by CAPI-CIDI (18,19) regarding drinking amounts during the last 12 months and periods of maximal alcohol consumption in combination with representative data of alcohol-dependent patients in Germany ((33); see references (16) and (34) for previous uses of this measure). Grey-matter data consisted of the 110 regional atlas-based grey-matter concentration averages of each participant. In an initial preprocessing step, we corrected the regional averages for age, gender and site using linear regression. Next, a linear ridge regression model was used to regress grey-matter pattern information on the kg lifetime consumption variable (R-syntax: lifetime consumption ~ $\text{region}_1 + \ldots + \text{region}_{110}$).

The regression model was trained in a leave-one-subject-out cross-validation scheme, such that in each fold, the model was trained on $N-1$ subjects and predicted the amount of consumed ethanol (in kg) for the left-out subject. To optimize the regularization parameter λ of Ridge regression, in each fold a range of parameters values (10^x, x=[-5; 5]) was tested in a nested leave-one-subject cross-validation procedure and the best parameter was used to train the model in the current (outer) fold. This procedure yielded an optimal value of $\lambda = 0.1$ for each outer cross-validation fold. Finally, to evaluate the accuracy of ensuing predictions of lifetime consumption, we correlated the predicted amounts of all subjects with measured amounts.

Validation on an independent data set

Crucial to our present work, we aimed to assess the generalizability of the applied machine-learning methods. We therefore validated our machine-learning approach on data of the National Genome Research Network Plus (henceforth NGFN+; http://www.ngfn.de/en/alkoholabh__ngigkeit.html; 19). This independent sample included structural MRI data from 83 substance-naive controls and 94 individuals with AD (see Table S1 for sample characteristics), which were acquired on a 3-Tesla Siemens MAGNETOM Verio scanner using a magnetization-prepared rapid gradient echo sequence (repetition time: 2300 ms; echo time: 3.03 ms; flip angle: 9°; field of view: 256×256 mm^2; 192 sagittal slices; voxel size: 1 mm isotropic). Importantly, the data set was acquired in a different scanning facility and by a different research group, and differed in terms of gender balance (36% female) from our original data set (16% female), thereby providing an excellent opportunity to test the generalizability of the machine-learning algorithms. To assess the generalization performance of a classifier, it was trained on the original LeAD data set and tested on an identically preprocessed NGFN+ data set.

Blinded evaluation by an independent radiologist

Another central goal of our present work was to directly compare the diagnostic classification performance of our machine-learning algorithms to human expert performance. To this end, the structural MRI scans of the LeAD study were rated by an independent radiologist (M. Sch.) based at the Charité Universitätsmedizin Berlin with 6 years of experience in neuroradiology. The radiologist was blind with respect to the original diagnostic labels and the computer-based labels of the machine-learning algorithm. Like the algorithm, the radiologist was provided information about the age and gender of subject. In addition, the radiologist was informed that age and gender were matched between groups and could thus only be used as a corrective for the judgement of brain scans, but not as discriminatory information by themselves. To probe the judgements of the radiologist in a way that the performance would be as representative as possible of typical radiologists in a clinical setting, the radiologist did not receive any labelled training scans and did not receive feedback during the judgement process. The task of the radiologist was to provide binary diagnostic labels for each scan ('In your opinion, is it a subject with or without a diagnosis of alcohol dependence?'). To avoid strategic judgements, the radiologist was not informed about the base rate of alcohol dependence in the sample.

To assess demographic biases in the judgments of the radiologist, as well as our classification scheme, a logistic regression model was used with judgement (control=0, case=1) as the regressand as well as age (in years) and gender (0 = female, 1 = male) as regressors.

Results

Whole-brain grey-matter loss in alcohol dependence

To ensure that we could replicate the previously reported grey-matter loss in AD (7), we first characterized grey-matter concentration differences between cases and controls in a univariate analysis. A voxel-wise whole-brain t-test revealed significantly higher grey-matter concentration in a number of brain regions in controls as compared to alcohol-dependent patients (see Fig. 2 and Table S2). The statistically strongest reduction in concentration was observed in a continuous set of regions stretching from precuneus via posterior, middle and anterior cingulate cortex to orbitofrontal cortex. Other affected brain structures included the middle frontal gyrus, pre- and postcentral gyrus, insula, cerebellum and thalamus. These results are in line with previous studies investigating grey-matter reduction in abstinent individuals with AD (35–37).

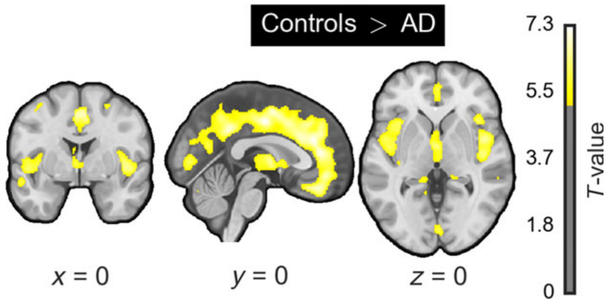

Fig. 2. Whole-brain univariate analysis. Whole-brain t-map for AD-related grey-matter concentration loss based on the contrast control >AD. No significant effects were observed for the reverse contrast control <AD. The t-map is thresholded at $P < 0.001$, corrected for family-wise errors.

Parcellation-based regional grey-matter averages yield the highest diagnostic classification accuracy

We considered and compared four different general analysis schemes for the classification between cases and controls based on grey-matter concentration (Fig. 1). In addition, all analysis schemes were submitted to two different classifiers: weighted robust distance (WeiRD) (9,10) and support vector machine (SVM). Figure 3a shows that parcellation-based schemes (parcelled voxel patterns and averages) were generally superior to non-parcelled schemes (whole-brain averages and voxel patterns). Moreover, comparing schemes based on voxel patterns (whole-brain and parcelled voxel patterns) to their averaged counterparts (whole-brain and parcelled averages respectively) indicated that fine-grained voxel patterns did not improve classification performance. Classification based on parcelled averages and WeiRD yielded the overall best performance (balanced accuracy: 74%; 95% posterior probability interval: [67; 79]; infraliminal probability: $P < 10^{-12}$; see Table 1 for sensitivity and specificity).

We reasoned that the benefit of parcellation was based on the anatomical prior information incorporated into the classification scheme. To test this hypothesis, for each anatomical region, we computed (i) the average across-subject correlation between (unsmoothed) voxels located *within* the region, and (ii) the average across-subject correlation between each voxel of the region and the voxels of all other regions. We found that the average within-region correlation ($r_{\text{z-transform}} = 0.077 \pm 0.023$ [95% CI]) was clearly higher than the average between-region correlation ($r_{\text{z-transform}} = 0.014 \pm 0.001$ [95% CI]). This difference was significant ($t_{218} = 5.9$, $P < 10^{-7}$, two-sample t-test). Thus, voxels within regions

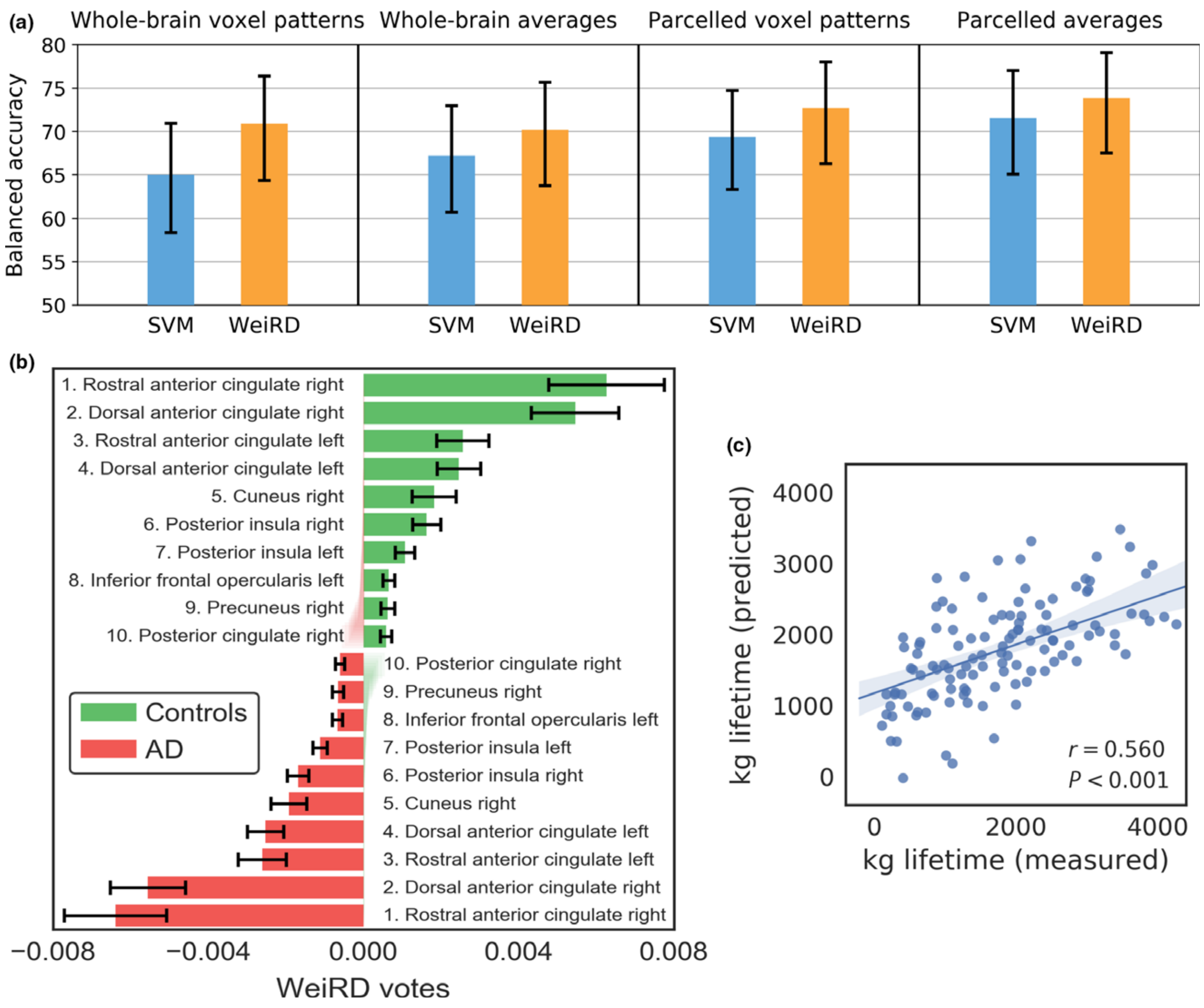

Fig. 3. Predicting the diagnosis and lifetime consumption in alcohol dependence. (a) Classification performance of analysis schemes. Balanced accuracy for classification based on whole-brain voxel patterns, whole-brain averages, parcelled voxel patterns and parcelled voxel averages. Error bars represent the 95% posterior probability interval of the balanced accuracy (31). (b) Votes of the WeiRD classifier for predictions of control (green) and AD subjects (red). Positive values indicate that the feature/region voted, on average, for ‘control’, negative values for ‘AD’. Error bars indicate SEM across subjects. Labelled are the 10 regions with the highest absolute votes, separately for control and AD subjects. Average votes of lower-ranking regions are displayed in a semi-transparent condensed manner. (c) Predicting AD severity based on grey-matter concentration. Scatter plot of measured lifetime consumption of ethanol (in kg) vs. the predicted amount (in kg). Translucent bands around the regression line depict the 95% confidence interval derived by bootstrapping.

Table 1. Comparison with radiologist

	Balanced Accuracy	Specificity	Sensitivity	Odds Ratio age [CI]	Odds Ratio Gender [CI]
Radiologist	66%	81%	51%	1.05 [1.02; 1.08]**	2.94 [1.18; 7.35]*
Classifier (corrected)	74%	76%	71%	1.02 [0.99; 1.04]	1.03 [0.49; 2.15]
Classifier (uncorrected)	69%	71%	67%	1.15 [1.11; 1.20]***	1.63 [0.64; 4.20]

Characterizing the judgements of the radiologist and the classifier without and with demographic correction. Asterisks indicate statistical significance of the logistic odds ratios (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$).

showed a more similar pattern across subjects than voxels between different regions.

To test whether there is any benefit in using our specific anatomical atlas as opposed to a random parcellation scheme, we repeatedly and randomly (10 000 iterations) chose grey-matter voxels and defined spheres around these voxels as artificial ‘regions’. To assimilate the region sizes (number of voxels) of the atlas, in each iteration, we randomly draw a designated region size from the atlas-based region size distribution and matched the radius of the artificial sphere to the chosen region size. We found that the within-region correlation of the random parcellation scheme ($r_{z\text{-transform}} = 0.052 \pm 0.004$ [95% CI]) was indeed significantly lower than the correlation yielded by the JHU atlas ($t_{10108} = 2.0$, $P < 0.05$, two-sample t-test). Overall, this exploratory analysis indicates that the advantage of parcellation-based schemes could be due to incorporated atlas-based anatomical knowledge that combines more similarly behaving voxels.

An additional advantage of the WeiRD classifier was that it allowed for a direct introspection of the most relevant regions underlying classification. Figure 3b shows the average classification votes casted by different regions, separately for predictions of control and AD subjects. Higher absolute votes reflect greater contribution of a region to a particular prediction, where positive and negative values are votes in favour of control and AD subjects respectively. The most important features for classification were regions of the cingulate and insular cortex as well as a region of inferior frontal gyrus. This pattern was consistent between control and AD predictions. Given its superior performance, parsimony and transparency, we focused on the parcelled averages classification scheme in combination with WeiRD in all following analyses.

Grey-matter patterns predict lifetime alcohol consumption

While the above results demonstrate the potential of grey-matter patterns for binary classification, they leave open the question whether grey-matter information also captures the severity of the disease within the AD group. To clarify this question, we regressed grey-matter patterns on the kg lifetime alcohol consumption variable. As shown in Fig. 3c, the cross-validated regression analysis revealed a strong correlation between the measured and the predicted amount of lifetime consumption ($r = 0.56$, $P < 10^{-10}$). Thus, grey-matter concentration patterns were a meaningful marker above and beyond the binary distinction between cases and controls.

Successful validation on an independent data set

A key step to assessing the validity of our approach was to apply our main classification scheme (parcelled averages + WeiRD) to the independent NGFN+ data set (38). Importantly, the model was exclusively trained on the full original data set and tested on the unseen NGFN+ data set. The balanced accuracy on the validation data set was 73% ([66; 79]; $P < 10^{-9}$; sensitivity: 71%, specificity: 76%), which was only 1% below the original accuracy (74%) and thus demonstrating excellent generalization of the classification scheme.

Classification-based approach outperforms an independent radiologist

A second key benchmark for our classification scheme was the comparison with an independent and blinded radiologist. We found that the radiologist classified 140 of the 216 individuals correctly, yielding a balanced accuracy of 66% ([60; 72]; $P < 10^{-6}$). While the accuracy was clearly above chance, it was significantly below our classification-based approach ($P = 0.004$, Wald test). Specificity (81%) was much higher than sensitivity (51%) for the radiologist, indicating a bias towards keeping the false-positive rate low.

As the individual age and gender were available to the radiologist, we were interested to what degree this information was related to the judgements of the radiologist. As shown in Table 1, an age increase of 1 year was associated with a 5% increase in the odds for a case judgement (over a control judgement, i.e. odds ratio). In addition, being male increased the odds for a case judgement by a factor of around 3, although this estimate was associated with a relatively large degree of uncertainty (possibly due to the small number of females in the sample). Interestingly, we found that computer-based classification showed an analogous bias, when the data were not corrected for variance in age and gender, but no bias, when the corrected data were used as in the original analysis (Table 1, last two rows). Overall, these results show that the radiologist was less accurate in the diagnosis of AD based on structural data and prone to bias with respect to age and gender.

Discussion

In the present study, we examined the predictive capacity of whole-brain grey-matter patterns both for the diagnosis of AD and for lifetime alcohol consumption. We found that a parsimonious classification scheme predicted diagnostic class labels with a balanced accuracy of 74% (76% specificity, 71% sensitivity), while a regression-based approach accurately predicted individual lifetime consumption in the AD group. High generalizability to an independent data set and superiority to human expert-level classification performance attested to the quality of the machine-learning approach.

Four different computer-based classification schemes were compared in this investigation (whole-brain voxel patterns and averages, parcelled voxel patterns and averages), yielding a number of insights. First, the fact that a classification scheme based on regional or whole-brain averages outperformed costlier voxel-based schemes shows that fine-grained voxel patterns did not provide meaningful information over and above regional averages. This suggests that if fine-grained voxel patterns contained additional information, it was obliterated by the increased noisiness of voxel-level signals.

Second, it is notable that univariate classification based on whole-brain averages of grey-matter concentration performed en par with whole-brain voxel patterns. This finding shows that even a single feature, total average concentration, already provided a simple and relatively accurate marker of alcohol-related brain damage.

Third, classification on whole-brain averages was nevertheless outperformed by parcellation-based schemes, demonstrating that additional discriminative information was indeed contained in a multivariate characterization of grey matter. The superiority of parcellation-based classification thus justifies a key aspect of machine-learning-based approaches, that is multivariate pattern analysis, and indicates that regional averages are the appropriate spatial scale for these patterns.

An exploratory analysis of the applied atlas-based parcellation scheme showed that voxels within regions were more similar across subjects than (i) voxels between different regions and (ii) voxels within regions of a random parcellation scheme. The latter aspect is particularly interesting, as it suggests a specific benefit of structuring grey-matter data based on an anatomically informed parcellation scheme. In the case of parcelled averages, such a structuring of the data can be regarded as a form of anatomically informed dimensionality reduction, as averages are taken across more similarly behaving voxels. Importantly, effective feature reduction reduces noise and often leads to a better generalization on novel data. In the case of parcelled *voxel patterns*, grouping voxels based on anatomically informed similarity may yield a set of relatively non-redundant classifiers with likewise positive effects on ensemble classification.

Using the parcelled averages scheme and exploiting the transparent voting-based classification scheme of WeiRD, we identified a number of brain areas particularly relevant to the distinction between AD and control subjects. The most important areas comprised regions of the cingulate, insular and inferior frontal cortex, which are largely in line with our and previous univariate results (35–37). Although feature importances can be computed for other classifiers, their interpretation is often more opaque. For instance, in the case of the SVM classifier used in the present study, the hyperplane-defining weight vector is often taken as a measure for the importance of features. However, recently, it has been shown that even entirely uninformative features can produce significant nonzero SVM weights (39), calling into question the interpretability of SVM weight vectors. On the other hand, while simple classification strategies such as decision trees, which implement a hierarchical chain of decision rules, can expose the contribution and role of each feature in a human-readable format, they suffer from overfitting and thus low generalization performance (40). By contrast, the voting-based classification scheme used by WeiRD combines high performance, as shown in this and previous studies (9,10), with high transparency, by quantifying the contribution of each feature to classification in the most direct way possible (i.e. votes). And as opposed to unsupervised analyses of data structure, such as principal or independent component analysis, WeiRD votes describe the importance of features not in terms of variance explained, but directly in terms of their discriminatory power for classification.

An important aspect of the present work is that the key classification result was validated on an independent data set of the National Genome Research Network Plus [NGFN+ (38)] with nearly identical accuracy. Likely, reasons for the high generalizability of our approach are (i) the fact that WeiRD is a parameter-free classifier and requires no fine-tuning or prior choices of parameters that are prone to overfitting, and (ii) the step of regional averaging within anatomical regions instead of analysing multivoxel patterns, thereby avoiding the challenge of fragile cross-individual correspondences between voxels. The fact that a number of critical variables differed between the two data sets, including the participating subjects, site of measurement and involved operators, provides reason to be optimistic that the classification model will generalize also to other independent data sets.

The comparison with an independent radiologist served to assess whether computer-based classification presented any advantage over human expert judgements. We found that the radiologist's accuracy at 66% was significantly lower than the computer-based accuracy. This result demonstrates the benefit of computer-based classification, but also attests to the general intricacy of the classification task. By comparison, in one of the few previous studies comparing computer-based and human performance (14), radiologists correctly classified sporadic Alzheimer's disease and controls with a median of 83% of 89% (on two different data sets), and even sporadic Alzheimer's disease from fronto-temporal lobar degeneration with a median of 71%. It is thus no surprise that also the computer-based accuracies were higher for this task (95% of 93% and 89% respectively). However, as a limitation to the present results, it should be noted that assessment of brain atrophy in patients with alcohol dependence is a rather rare task for radiologist. Moreover, as these results are based on the judgements of a single radiologist, no inference about interindividual variation in radiological judgements can be made. By contrast, as several different classification schemes were assessed in computer-based classification, it should be pointed out that compared to voxel-based SVM classification, the radiologist's performance was equal (parcelled voxel patterns) or slightly better (whole-brain voxel patterns).

It is interesting to note that the judgements of both the radiologist and the computer-based classification without demographic correction showed biases with respect to age and gender: being male and being older increased the probability of being classified as an individual with AD. As the computer-based classification is not influenced by sociodemographic stereotypes, this result suggests that the bias was data-driven, that is, there were features in the data that misled both the classifier and the radiologist. Crucially, this classification bias was abolished in computer-based classification when the data were corrected for age and gender, pointing to an advantage of computer-based systems in systematically accounting for variables of no interest. Finally, it is noteworthy that the radiologist's judgements showed much higher specificity than sensitivity, whereas this imbalance was greatly reduced (and in part reversed) in the case of computer-based classification. Thus, computer-aided classification may be particularly suited as a screening tool with high sensitivity, delivering candidate cases to a radiologist with high specificity to reduce the number of false positives.

Finally, given the fact that AD diagnosis is associated with much less uncertainty than anamnestic reports on lifetime alcohol consumption, it is remarkable that grey-matter concentration patterns predicted lifetime alcohol consumption within the AD group with high accuracy. Although no conclusive evidence, these results suggest that grey-matter concentration may be a particularly sensitive marker within the group of AD subjects.

In sum, the present work provides the first classification-based approach to diagnostic classification and severity prediction of current adult AD on the basis of structural neuroimaging data. By applying a transparent and robust classification algorithm, by validating our key algorithm on an independent data set and by providing direct comparison to human expert-level performance, we aimed to set an example for machine-learning approaches to psychiatric neuroimaging. While our results attest to the rich diagnostic information contained in whole-brain grey-matter patterns, they also highlight the current limits of computer-based classification in terms of accuracy when rigorous methodology is applied.

Acknowledgements

This work was supported by the following institutions: German Research Foundation (Deutsche Forschungsgemeinschaft, DFG, FOR 1617: grants STE 1430/6-1, STE 1430/6-2, SCHM 3209/1-2, ZI 1119/3-1, ZI 1119/3-2, HE 2597/14-1, HE 2597/14-2, WI 709/10-1, WI 709/10-2 and SFB 940/2 of the Excellence Cluster Exc 257); Federal Ministry of Education and Research (BMBF grants 01GS08159; 01ZX1311H; 01ZX1311D/1611D and 01ZX1311E/1611E; and in part by 01EE1406A and 01EE1406B). KS is participant in the Charité Clinical Scientist Program funded by the Charité — Universitätsmedizin Berlin and the Berlin Institute of Health. We express our gratitude to the staff of the Berlin Center for Advanced Neuroimaging (BCAN).

References

1. Mwangi B, Ebmeier KP, Matthews K, Douglas Steele J. Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder. Brain 2012;**135**:1508–1521.
2. Ecker C, Marquand A, Mourão-Miranda J et al. Describing the Brain in Autism in Five Dimensions—Magnetic Resonance Imaging-Assisted Diagnosis of Autism Spectrum Disorder Using a Multiparameter Classification Approach. J Neurosci 2010;**30**:10612–10623.
3. Douaud G, Smith S, Jenkinson M et al. Anatomically related grey and white matter abnormalities in adolescent-onset schizophrenia. Brain 2007;**130**:2375–2386.
4. Zanetti MV, Schaufelberger MS, Doshi J et al. Neuroanatomical pattern classification in a population-based sample of first-episode schizophrenia. Prog Neuropsychopharmacol Biol Psychiatry 2013;**43**:116–125.
5. Gould IC, Shepherd AM, Laurens KR, Cairns MJ, Carr VJ, Green MJ. Multivariate neuroanatomical classification of cognitive subtypes in schizophrenia: a support vector machine learning approach. Neuroimage Clin 2014;**6**:229–236.
6. Koutsouleris N, Meisenzahl EM, Borgwardt S et al. Individualized differential diagnosis of schizophrenia and mood disorders using neuroanatomical biomarkers. Brain 2015;**138**:2059–2073.
7. Harper C, Matsumoto I. Ethanol and brain damage. Curr Opin Pharmacol 2005;**5**:73–78.
8. Brodersen KH, Deserno L, Schlagenhauf F et al. Dissecting psychiatric spectrum disorders by generative embedding. Neuroimage Clin 2014;**4**:98–111.

9. Guggenmos M, Schmack K, Sterzer P. WeiRD - a fast and performant multivoxel pattern classifier. In: 2016 International Workshop on Pattern Recognition in Neuroimaging (PRNI), Trento, 2016, pp. 1–4.
10. Guggenmos M, Sterzer P, Cichy RM. Multivariate pattern analysis for MEG: a comprehensive comparison of dissimilarity measures. BioRxiv Published Online First: 2017. https://doi.org/10.1101/172619
11. Demirci O, Clark VP, Magnotta VA et al. A review of challenges in the use of fMRI for disease classification/characterization and A projection pursuit application from A multi-site fMRI schizophrenia study. Brain Imaging Behav 2008;**2**:207–226.
12. Whelan R, Garavan H. When Optimism Hurts: Inflated Predictions in Psychiatric Neuroimaging. Biol. Psychiatry 2014;**75**:746–748. http://doi.org/10.1016/j.biopsych.2013.05.014
13. Schnack HG, Kahn RS. Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters. Front Psychiatry 2016;**7**:50. https://doi.org/10.3389/fpsyt.2016.00050
14. Klöppel S, Stonnington CM, Barnes J et al. Accuracy of dementia diagnosis - A direct comparison between radiologists and a computerized method. Brain 2008;**131**:2969–2974.
15. Sebold M, Deserno L, Nebe S et al. Model-Based and Model-Free Decisions in Alcohol Dependence. Neuropsychobiology 2014;**70**:122–131.
16. Garbusow M, Schad DJ, Sebold M et al. Pavlovian-to-instrumental transfer effects in the nucleus accumbens relate to relapse in alcohol dependence. Addict Biol 2016;**21**:719–731.
17. Garbusow M, Sebold M, Beck A, Heinz A. Too Difficult to Stop: Mechanisms Facilitating Relapse in Alcohol Dependence. Neuropsychobiology 2014;**70**:103–110.
18. Wittchen H-U, Pfister H. DIA-X-Interviews: Manual fur Screening-Verfahren und Interview; Interviewheft. Frankfurt: Swets & Zeitlinger, 1997.
19. Jacobi F, Mack S, Gerschler A et al. The design and methods of the mental health module in the German Health Interview and Examination Survey for Adults (DEGS1-MH). Int J Methods Psychiatr Res 2013;**22**:83–99.
20. Oldfield RC. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 1971;**9**:97–113.
21. Good CD, Johnsrude IS, Ashburner J, Henson RN, Friston KJ, Frackowiak RS. A voxel-based morphometric study of ageing in 465 normal adult human brains. NeuroImage 2001;**14**:21–36.
22. Radua J, Canales-Rodríguez EJ, Pomarol-Clotet E, Salvador R. Validity of modulation and optimal settings for advanced voxel-based morphometry. NeuroImage 2014;**86**:81–90.
23. Dashjamts T, Yoshiura T, Hiwatashi A et al. Alzheimer's disease: diagnosis by different methods of voxel-based morphometry. Fukuoka Igaku Zasshi 2012;**103**:59–69.
24. Ashburner J, Friston KJ. Voxel-Based Morphometry—The Methods. NeuroImage 2000;**11**:805–821.
25. Salmond CH, Ashburner J, Vargha-Khadem F, Connelly A, Gadian DG, Friston KJ. Distributional assumptions in voxel-based morphometry. NeuroImage 2002;**17**:1027–1030.
26. Faria AV, Joel SE, Zhang Y et al. Atlas-based analysis of resting-state functional connectivity: evaluation for reproducibility and multi-modal anatomy–function correlation studies. NeuroImage 2012;**61**:613–621.
27. Cortes C, Vapnik V. Support-Vector Networks. Mach Learn 1995;**20**:273–297.
28. Zhou L, Wang L, Liu L, Ogunbona PO, Shen D. Support vector machines for neuroimage analysis: interpretation from discrimination. In: MA Y, GUO G. eds. Support vector machines applications. Cham: Springer, 2014:191–220.
29. Cherkassky V, Mulier F. Learning from data. Hoboken, NJ: John Wiley & Sons, Inc., 1998.
30. Chang C, Lin C. LIBSVM: A Library for Support Vector Machines. ACM Trans Intell Syst Technol 2011;**2**:27:1–27:27.
31. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. Proc - Int Conf Pattern Recognit 2010;3121–3124.
32. Bishop CM. Pattern recognition and machine learning. New York, NY: Springer Science + Business Media, LLC; 2006.
33. Jacobi F, Höfler M, Strehle J et al. Psychische störungen in der allgemeinbevölkerung. Studie zur gesundheit erwachsener in Deutschland und ihr zusatzmodul psychische gesundheit (DEGS1-MH). Nervenarzt 2014;**85**:77–87.
34. Sommer C, Garbusow M, Jünger E et al. Strong seduction: impulsivity and the impact of contextual cues on instrumental behavior in alcohol dependence. Transl Psychiatry 2017;**7**:e1183.
35. Chanraud S, Martelli C, Delain F et al. Brain morphometry and cognitive performance in detoxified alcohol-dependents with preserved psychosocial functioning. Neuropsychopharmacology 2007;**32**:429–438.
36. Tanabe J, Tregellas JR, Dalwani M et al. Medial Orbitofrontal Cortex Gray Matter Is Reduced in Abstinent Substance-Dependent Individuals. Biol Psychiatry 2009;**65**:160–164.
37. Demirakca T, Ende G, Kämmerer N et al. Effects of Alcoholism and Continued Abstinence on Brain Volumes in Both Genders. Alcohol Clin Exp Res 2011;**35**:1678–1685.
38. Spanagel R. Alcoholism : a systems approach from molecular physiology to addictive ehavior. Physiol Rev 2009;**89**:649–705.
39. Haufe S, Meinecke F, Görgen K et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage 2013;**87**:96–110.
40. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. New York, NY: Springer-Verlag; 2008.
41. Schmidt L, Gastpar M, Falkai P, Gäbel W. Evidenzbasierte Suchtmedizin. In: Substanzbezogene Störungen. Köln: Deutscher Ärzteverlag, 2006.
42. Skinner H, Horn J. Alcohol dependence scale (ADS): user' guide. Toronto, ON: Addiction Research Foundation, 1984.
43. Mann K, Ackermann K. Die OCDS-G: Psychometrische Kennwerte der deutschen Version der Obsessive Compulsive Drinking Scale [The OCDS-G: Psychometric Characteristics of the German version of the Obsessive Compulsive Drinking Scale]. Sucht 2000;**46**:90–100.
44. Meule A, Vögele C, Kübler A. Psychometrische evaluation der Deutschen Barratt Impulsiveness scale - Kurzversion (BIS-15). Diagnostica 2011;**57**:126–133.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Sample characteristics (NGFN+ dataset) for alcohol-dependent (AD) and healthy control (HC) subjects.
Table S2. Whole–brain group contrast.