Development and validation of a neuroimaging biomarker for electroconvulsive therapy outcome in depression: a multicenter machine learning analysis

*Willem B. Bruin[1], MSc, Leif Oltedal[2,3], MD, PhD, Hauke Bartsch[2], PhD, Christopher C. Abbott[4], MD, MS, Miklos Argyelan[5,6], MD, PhD, Tracy Barbour[7], MD, Joan A. Camprodon[7], MD, PhD, Samadrita Chowdhury[7], PhD, Randall Espinoza[8], MD, Peter C. R. Mulders[9], MD, PhD, Katherine L. Narr[10], PhD, Mardien L. Oudega[11], MD, PhD, Didi Rhebergen[12], MD, PhD, Freek ten Doesschate[13,1], MSc, Indira Tendolkar[9], MD, PhD, Philip van Eijndhoven[9], MD, PhD, Eric van Exel[11], MD, PhD, Mike van Verseveld[13], MD, Benjamin Wade[14], PhD, Jeroen van Waarde[13], MD, PhD, **Annemiek Dols[11], MD, PhD & **Guido A. van Wingen[15,16], MD, PhD

*= full author
**=shared senior author

[1]:Amsterdam UMC, University of Amsterdam, Department of Psychiatry, Amsterdam Neuroscience, Amsterdam, The Netherlands, [2]:Mohn Medical Imaging and Visualization Centre, Department of Radiology, Haukeland University Hospital, Bergen, Norway, [3]:Department of Clinical Medicine, University of Bergen, Bergen, Norway, [4]:Department of Psychiatry, University of New Mexico, Albuquerque, NM, [5]:The Feinstein Institutes for Medical Research, Manhasset, NY, [6]:The Zucker Hillside Hospital, Glen Oaks, NY, [7]:Division of Neuropsychiatry and Neuromodulation, Massachusetts General Hospital, Harvard Medical School. Boston, MA (USA), [8]:Department of Psychiatry and Biobehavioral Sciences, UCLA, Los Angeles, USA, [9]:Donders Institute for Brain, Cognition and Behavior, Department of Psychiatry, Nijmegen, [10]:Ahmanson-Lovelace Brain Mapping Center, Departments of Neurology, and Psychiatry and Biobehavioral Sciences, UCLA, Los Angeles, USA, [11]:Departement of Old Age Psychiatry, GGZinGeest, Department of Psychiatry, Amsterdam UMC, location VUmc, Amsterdam Neuroscience, Amsterdam, The Netherlands, [12]:Mental Health Institute GGZ Centraal, Amersfoort; Department of Psychiatry, Amsterdam UMC, location VUmc, Amsterdam Neuroscience, Amsterdam, The Netherlands, [13]:Rijnstate, Department of Psychiatry, Arnhem, The Netherlands, [14]:Ahmanson-Lovelace Brain Mapping Center, Department of Neurology, UCLA, Los Angeles, USA, [15]:Amsterdam UMC, University of Amsterdam, Department of Psychiatry, Amsterdam Neuroscience, Amsterdam, the Netherlands, [16]:Amsterdam Brain and Cognition, University of Amsterdam, the Netherlands.

Corresponding authors:

Willem B. Bruin (w.b.bruin@amsterdamumc.nl)

Guido A. van Wingen (g.a.vanwingen@amsterdamumc.nl)

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

Abstract

Electroconvulsive therapy (ECT) is the most effective intervention for patients with treatment resistant depression. A clinical decision support tool could guide patient selection to improve the overall response rate and avoid ineffective treatments with adverse effects. Initial small-scale, mono-center studies indicate that both structural magnetic resonance imaging (MRI) and functional MRI biomarkers may predict ECT outcome, but it is not known whether those results can generalize to data from other centers. Here, we used MRI data of 189 depressed patients from seven participating centers of the Global ECT-MRI Research Collaboration (GEMRIC) to develop and validate neuroimaging biomarkers for ECT outcome in a multi-center setting. We used multimodal data (i.e., clinical, structural MRI and resting-state functional MRI) and evaluated which data modalities or combinations thereof could provide the best predictions for treatment response (≥50% symptom reduction) or remission (minimal symptoms after treatment) using a support vector machine (SVM) classifier. Remission classification using a combination of gray matter volume with functional connectivity led to good performing models with 0.82-0.84 area under the curve (AUC) when trained and tested on samples coming from all centers, and remained acceptable when validated on other centers with 0.71-0.73 AUC. These results show that multimodal neuroimaging data is able to provide good prediction of remission with ECT for individual patients across different treatment centers, despite significant variability in clinical characteristics across centers. This suggests that these biomarkers are robust,

indicating that future development of a clinical decision support tool applying these biomarkers may be feasible.

Introduction

Electroconvulsive therapy (ECT) is currently the most effective intervention for patients with treatment resistant depression[1]. Despite its high efficacy, ECT remains underutilized, as only 1-2% of patients with severe or persistent depression receive ECT[2, 3]. Although approximately 48% of treatment resistant patients recover with ECT, it is also associated with adverse cognitive effects and may be regarded as more invasive than other treatment options because the use of anesthesia is essential[4]. Furthermore, ECT is relatively expensive and non-responsiveness can only be determined after multiple sessions. Information that better predicts treatment outcome would enable patient selection thereby further improving the overall response rate and avoiding ineffective treatment with adverse effects. A personalized recommendation about the expected benefit of ECT would be a valuable addition to the treating physician's clinical judgement, and may increase its use in clinical practice.

Attempts to develop instruments that may predict ECT outcome date back to the 1950s[5]. Meta-analyses have associated several clinical characteristics with beneficial ECT outcome, in particular no history of treatment resistance, older age and psychotic symptoms[6, 7]. However, their predictive power is insufficient to guide individual patient selection[4, 8-11]. Recent studies have started using neuroimaging data to predict ECT outcome at the individual level using machine learning analysis, which can construct multivariate prediction models using all the available data. Initial small-scale studies have shown that both structural magnetic resonance imaging (MRI) and functional MRI

findings can be used to predict ECT outcome with approximately 80% accuracy, which is considered sufficiently good for clinical use[11-20]. These initial results have been confirmed by subsequent studies, and a recent meta-analysis showed an average prediction accuracy of 82%[21].

Despite these promising results, the existing studies have been limited by using small samples and mono-center settings. This reduces the possibility for models to generalize to new samples across centers. Although machine learning models typically perform better when trained on larger samples from the same center, classification accuracy of larger multicenter studies tends to decrease, presumably due to increased clinical (e.g., adults vs. elderly) and technological (e.g., different MRI hardware and protocols) variability across centers[22-24]. In order to develop robust and generalizable neuroimaging biomarkers for ECT outcome, we used data from the Global ECT-MRI Research Collaboration (GEMRIC) and validated classification performance in a multi-center setting[25]. We used multimodal data (i.e., clinical, structural MRI (sMRI), and resting-state functional MRI (rs-fMRI)) and evaluated which data modalities or combinations thereof might provide the best predictions. As previous studies and clinical trials have used either treatment response (at least 50% symptom reduction) or remission (minimal symptoms after treatment) as outcome criterion, we assessed prediction accuracy for both criteria. Additionally, we evaluated whether model performance would increase when only data from centers with reasonable sample sizes are used. Finally, we visualized the brain regions that were most informative to the classifications, in order to gain insight into the brain regions predictive of ECT

outcome. To adhere to guidelines on transparent reporting of multivariable prediction models for individual prognosis or diagnosis (TRIPOD), the checklist is included in the **Supplementary Files**[26].

Methods

Participants

We used data from GEMRIC, an international consortium that contains the largest multi-center database of neuroimaging on ECT[25, 27]. All contributing sites received ethics approval from their local ethics committee or institutional review board. In addition, the centralized mega-analysis was approved by the Regional Ethics Committee South-East in Norway (No. 2018/769). Analyses contained a selection of sMRI and rs-fMRI data from seven centers across Europe and North America, accounting for a total of 189 clinically depressed patients according to ICD-10 (167 unipolar, 22 bipolar) who had received right unilateral or bilateral ECT (or both; **Supplementary Table 1**). Treatment outcome was measured using the 17-item Hamilton Depression Rating Scale (HAM-D) or Montgomery-Åsberg Depression Rating Scale (MADRS) that was converted to HAM-D (**Supplementary Methods**)[28]. Treatment response was defined as ≥50% HAM-D decrease compared to baseline and remission as post-ECT HAM-D score ≤7. ECT stimulus parameters varied between different centers, including electrode placement. As GEMRIC consists of samples ranging from very small (<20 patients) to relatively large (>40 patients), we performed all analyses on the entire cohort and for centers with ≥20 patients (three centers, N=109) in order to ensure classifiers were provided with sufficient examples per center. A description of centers-specific ECT procedures and image acquisition is provided elsewhere[25, 27].

MRI data and preprocessing

MRI acquisition parameters are listed in **Supplementary Tables 2-3**. Structural T1-weighted scans were acquired using 1.5T and 3T scanners with a minimum resolution of 1.33 mm^3 and preprocessed using the CAT12 toolbox for voxel-based morphometry (VBM). Images were segmented into gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF), normalized to MNI space using DARTEL registration, resampled to 1.5 mm^3 isotropic and spatially smoothed with an 8mm isotropic Gaussian kernel[29]. GM data were masked at 0.2 to exclude WM.

150-266 rs-fMRI volumes were acquired with a TR of 1.7-3.0 seconds, in-plane resolution of 2.4-3.75 mm, and slice thickness of 3-5 mm. Preprocessing was performed using ANTs (https://github.com/ANTsX/ANTs) and FSL (http://fsl.fmrib.ox.ac.uk/), including brain extraction, boundary-based co-registration, motion correction, spatial smoothing with a 5mm isotropic Gaussian kernel, and normalization to a 2mm MNI template. Denoising was performed using ICA-AROMA, and depending on the type of analysis, high-pass ($f>0.01$) or bandpass filtering ($0.009<f<0.08$) was applied together with WM and CSF nuisance regression[30]. Denoised rs-fMRI data were resampled to 4mm isotropic. Subjects showing excessive motion were excluded[31, 32].

Only subjects that passed quality control for both rs-fMRI and sMRI were included for analysis, leading to a final sample of 189 patients (**Supplementary Figure**

1 for a flowchart). Details on MRI preprocessing, quality control and machine learning are provided in the **Supplementary Methods.**

Feature extraction

We extracted commonly used MRI features from the preprocessed data. For sMRI, we used voxel-wise modulated GM maps (VBM) and 142 cortical and subcortical parcellations using the Neuromorphometrics atlas (NMM; provided by Neuromorphometrics, Inc). For rs-fMRI, we used group independent component analysis (ICA) to extract physiologically meaningful resting-state networks and reduce data dimensionality to 70 independent components[33]. Components reflecting non-neural signals were discarded, resulting in 53 spatial components for analysis. Group-information guided ICA was used to derive subject-specific time-series and spatial maps for the 53 signal components[34]. Time-series were used to calculate individual functional connectivity (FC) matrices that described pairwise connectivity between signal components with Pearson correlations (ICA-DR FC). Additionally, we used an atlas-based approach from Power *et al.,* and extracted time-series from 264 functional areas to compute FC matrices (Power FC)[35]. Correlations were converted to z-scores with Fisher r-to-z transformation before entering classification.

Machine learning

Machine learning classifications were performed using linear support vector machine (SVM; LIBSVM[36]) implemented in scikit-learn with stratified shuffle-split cross-

validation (CV) with 100 iterations. At each iteration, stratified-splits were made by preserving the proportion of responders/remitters and non-responders/remitters from each center to obtain maximally homogeneous train-test splits in which 80% data was used for classifier training and 20% for testing. This CV procedure is further referred to as 'internal validation'. In addition, we addressed leave-one-site-out (LOSO) CV, in which all but one center was used to train the SVM while the remaining center was used to assess model performance (further referred to as 'external validation'). This procedure was repeated so that each center is used once for testing. LOSO reduces the risk of overfitting data from a single center but may result in large between-sample heterogeneity of training and test sets, resulting in lower classification performance compared to internal validation[37]. Hyper-parameters for SVM were optimized with gridsearch using nested cross-validation. We assessed classification performance using different sets of MRI features (VBM, NMM, ICA-DR FC, Power FC, and ICA spatial components), as well as using clinical data only (i.e., age, sex and pre-ECT HAMD scores) for baseline classification. Clinical data were always included for each classification. The primary performance metric was area under the receiver operator characteristic curve (AUC) and reported metrics were averaged across CV iterations[38, 39]. Balanced accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) are reported in **Supplementary Tables 6-13**.

Statistical significance of classification performance was assessed using a label permutation-testing framework with 1000 iterations[40]. Obtained p-values were corrected for multiple comparisons using False Discovery Rate (FDR; two-stage (non-

negative); alpha=0.05). 95% confidence intervals (CI) for AUC were computed using the modified Wald-method[41]. To reduce computational burden, only spatial ICA classifications that resulted in AUC>0.75 were tested for significance. Finally, we assessed classification performance for multi-modal classifications combining anatomical and functional features: regional neuromorphometrics GM volumes with either ICA or Power-atlas based FC, and voxel-wise GM with either ICA or Power-atlas based FC.

Anatomical localization

To investigate which regions contributed most to the voxel-wise classification, we employed a method to estimate p-values for the weights of the SVM[42]. A statistic was computed incorporating the weight component value and the size of the margin, and an analytical approximation to the null-distribution obtained through permutation testing was used to calculate p-values.

Results

Demographic data

Demographic data is presented in Table 1. Of the 189 included patients, 113 patients were ECT responders and 76 non-responders, and 76 were remitters and 113 non-remitters. As expected, patients with a favorable outcome were older, and higher symptom severity at baseline was associated with ECT response but not remission. No significant differences in sex, initial electrode placement and total number of ECT-sessions were observed.

We assessed differences in sample demographics and clinical characteristics between the different centers regardless of ECT outcome using one-way analysis of variance (ANOVA) and χ^2. Age ($F(7,181)=14.08$, $p<0.001$), pre-treatment HAMD scores ($F(7,181)=7.40$, $p<0.001$), post-treatment HAMD scores ($F(7,181)=5.24$, $p<0.001$), HAM-D change ($F(7,181)=8.65$, $p<0.001$), number of ECT sessions ($F(7,178)=10.78$, p= $p<0.001$), depression type ($X^2(7, N=189)=19.10$, $p=0.008$) and initial electrode placement laterality ($X^2(7, N=189)=109.8$, $p<0.001$) differed significantly between centers. In contrast, sex did not differ between centers ($X^2(7, N=189)=3.84$, $p=0.80$). Demographic data for the three largest centers (with N≥20) used for additional analyses are described in **Supplementary Tables 4-5.** Differences in sample demographics and clinical characteristics between the three largest centers were similar to those seen in the entire sample. These findings highlight that there is considerable clinical heterogeneity between centers.

	Total sample (n=189)		Responders (n=113)		Non-Responders (n=76)		Resp vs Non-Resp	Remitters (n=76)		Non-Remitters (n=113)		Rem vs Non-Rem
	mean	std	mean	std	mean	std	p	mean	std	mean	std	p
Age	51.7	14.5	54.400	13.7	47.6	16.9	0.0045*	56.3	14.2	48.6	15.5	0.0006*
Sex (m/f)	83/106	n.a.	52/61	n.a.	31/45	n.a.	0.5705	32/44	n.a.	51/62	n.a.	0.7935
Laterality (RUL/BL; n=188)	148/40	n.a.	87/25	n.a.	61/15	n.a.	0.8	60/15	n.a.	88/25	n.a.	0.86
HAM-D pre-treatment	25	7.7	26.300	7.3	23	7.7	0.0031*	25.8	8.2	24.5	7.2	0.26
HAM-D post-treatment	11	8.3	5.700	4.2	18.9	6.3	5.82E-31*	3.3	2.3	16.2	6.6	3.73E-42*
HAM-D change	14	10.7	20.600	7.7	4.1	5.8	6.31E-39*	22.5	8.3	8.2	7.8	3.77E-23*
Diagnosis (UP/BP)	167/22	n.a.	99/14	n.a.	68/8	n.a.	0.8726	67/9	n.a.	100/13	n.a.	0.8726
Total ECT sessions (n=186)	13.4	6.2	13.000	6.4	14	5.8	0.2665	12.9	6.7	13.8	5.8	0.3474

Table 1. Demographics of patients included in data analysis, with subject demographics and comparisons between ECT responders and non-responders, and between remitters and non-remitters. Abbreviations: m: male; f: female; RUL: right unilateral ECT initially, BL: bilateral ECT initially; HAM-D: Hamilton Rating scale for depression; UP: unipolar depression; BP: bipolar depression; n.a.: not available. Asterisks depict significance using independent t-test or χ^2 test.

Response prediction

The majority of the classification models for response prediction performed poorly with AUC<0.7 for internal validation, and none of the models remained significant with external validation after permutation testing with FDR correction. The results for response classification are presented in Figures 1 and 3 and **Supplementary Tables 6, 8, 10-11**. Although in clinical patient care response to ECT may be beneficial, reaching remission after treatment is most preferable. Therefore, we focus on the results obtained for remission prediction below.

Remission prediction

Unimodal analysis

All centers

We first evaluated prediction performance across centers using all data (N=189) with internal validation. Sample size per center ranged from 14 to 42. Prediction performance with internal validation was poor with AUC ranging between 0.58-0.67 across different MRI modalities (Figure 1A). Classification using clinical variables resulted in a comparable AUC of 0.62. All these AUCs were statistically significant. Classification using external validation hardly exceeded chance-level, with AUCs ranging between 0.51-0.58 and none were statistically significant. Classification using ICA networks did not exceed AUC>0.75 for either internal or external validation.

Balanced accuracy, sensitivity, specificity, PPV and NPV, p-values for AUC statistical significance and 95% CIs are provided in **Supplementary Table 7.**

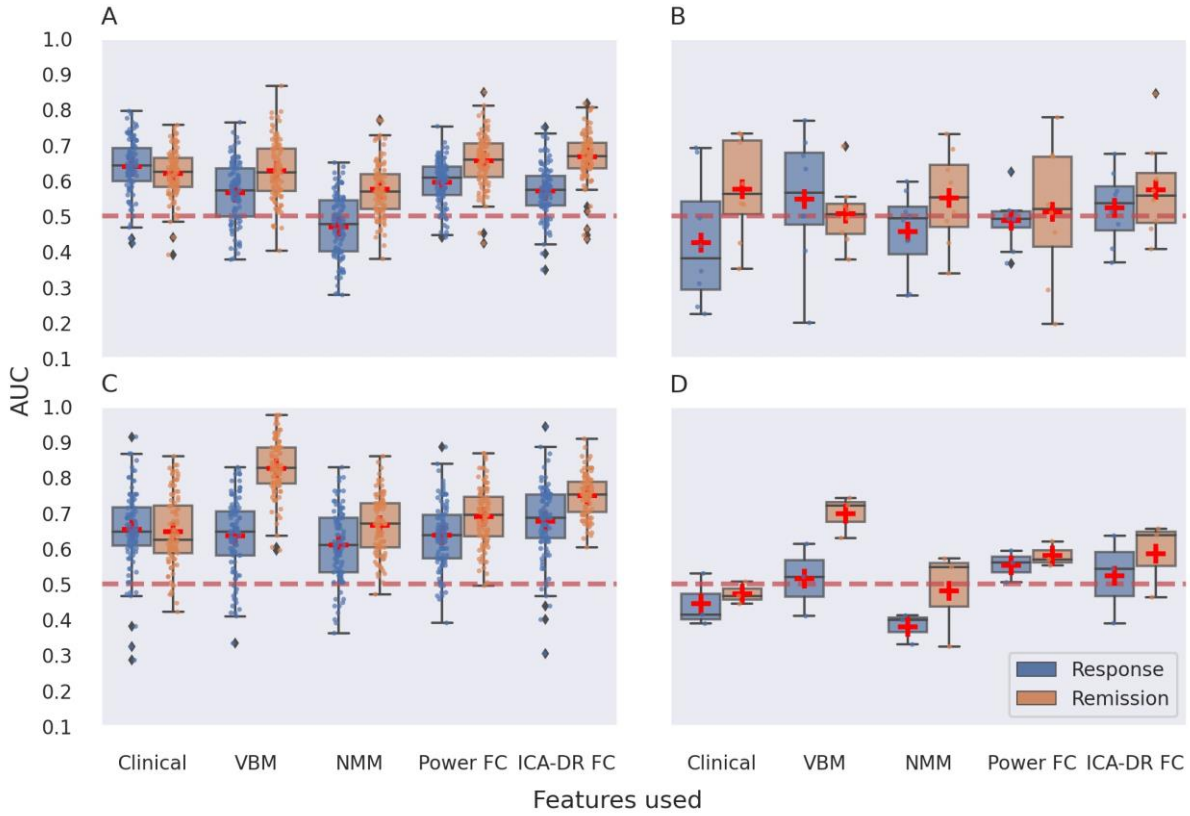

Figure 1. Multi-center predictions for ECT treatment response and remission using unimodal MR data modalities. Panel A depicts classification performance using data from all centers and different MR modalities with internal validation (AUC is averaged over 100 stratified cross-validation splits). Panel B shows classification performance using data from all centers with external validation (leave-one-site-out cross-validation, scores are averaged across different center left out for model testing). Panel C depicts classification performance using data from the three largest centers with internal validation. Panel D shows classification performance using data from the three largest centers with external validation. VBM = voxel-based morphometry; NMM = Neuromorphometrics atlas; FC = functional connectivity; ICA = group information guided independent component analysis. Red dashed line depicts chance level performance (0.5 AUC). Asterisks indicate significant difference from chance level after permutation testing with false discovery rate correction for multiple comparisons ($p<0.05$, corrected).

Three largest centers

We next assessed prediction performance using a subsample of data containing three centers with N≥20 (N=109) to provide the machine learning classifier with sufficient samples per center. Classification performance with internal validation ranged between 0.52-0.83 AUC across different features used, and 0.65 AUC was obtained for classifications using clinical variables only (Figure 1C). All AUCs obtained with internal validation showed statistical significance. Notably, the highest performance was achieved using voxel-wise GM data with 0.83 AUC. Four out of 53 ICA networks resulted in AUC>0.75 (Figure 2). A network incorporating right posterior parietal cortex and part of central executive network (CEN) resulted in 0.78 AUC, a network centered on the right pre- and postcentral gyrus resulted in 0.76 AUC, one located in posterior cingulate gyrus resulted in 0.77 AUC, and a thalamic network resulted in 0.80 AUC. All aforementioned AUCs were found to be statistically significant. Classifications with external validation ranged between 0.47-0.72 AUC (Figure 1D). The highest performance obtained was reduced from 0.83 AUC with internal validation to 0.70 AUC with external validation, and failed to obtain statistical significance following permutation testing with multiple comparison correction ($p_{uncorrected}$=0.018). None of the ICA networks resulted in AUC>0.75 with external validation (**Supplementary Table 9**).

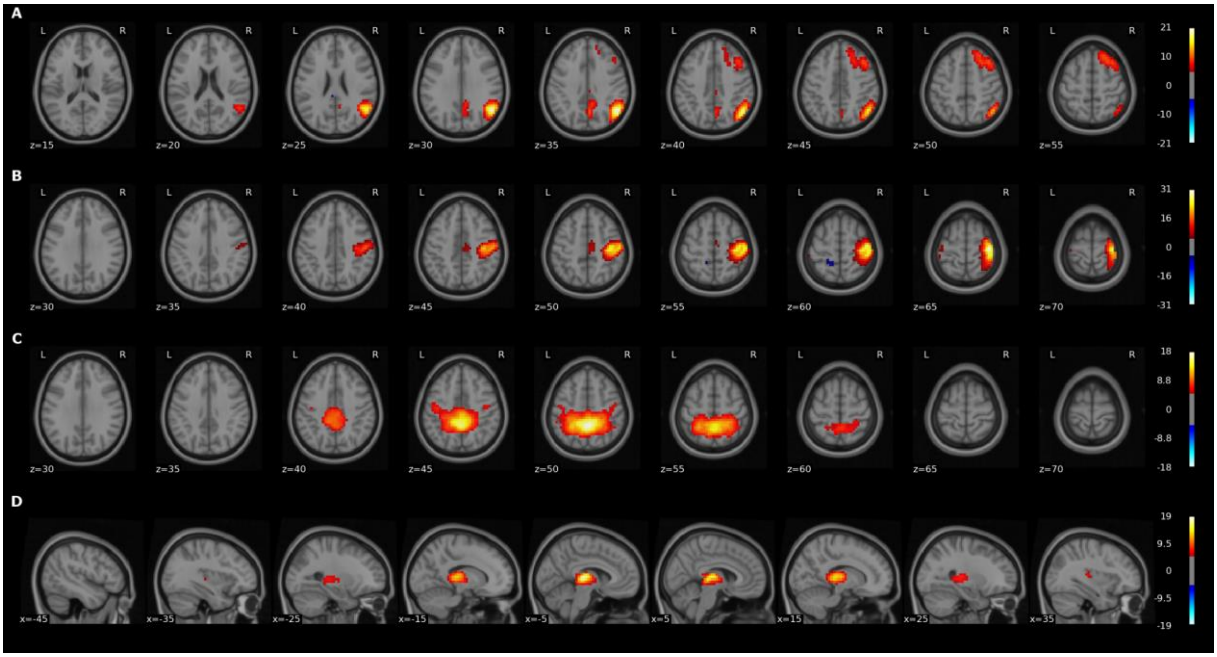

Figure 2. Visual representation of the four spatial components obtained from group ICA that led to AUC>0.75 for either response or remission classification. Top panel A depicts a network located in right posterior parietal cortex, part of the right central executive network. The second panel B shows a network located in right pre- and postcentral gyrus. The third panel C shows a network located in posterior cingulate gyrus. Finally, panel D illustrates a thalamic network. Images are thresholded at Z≥5 and overlaid on a standard 2mm MNI template.

Multimodal analysis

Classification using a combination of anatomical and functional MRI measures with samples from all centers led to a maximum of 0.68 AUC using internal validation which was statistically significant, whereas 0.64 AUC for external validation did not obtain significance. We then assessed multimodal classification performance using the three largest centers only. Classification of voxel-wise GM with ICA-based FC led to the best performing model, 0.84 AUC using internal validation, which remained acceptable using external validation with 0.71 AUC. Classifications for voxel-wise GM with Power-atlas FC led to similar performances with 0.82 AUC for internal validation and 0.73 AUC for external validation. All of the aforementioned AUCs were statistically significant for

both internal and external validation. Classification performance for regional neuromorphometrics volumes with ICA-based FC resulted in 0.72 AUC with internal validation and 0.52 AUC for external validation. Classifications for regional neuromorphometrics volumes with Power-atlas FC led to 0.67 AUC using internal validation and 0.55 AUC for external validation. AUCs obtained for classifications using regional neuromorphometrics were statistically significant for internal validation but not for external validation (**Supplementary Tables 12-13**).

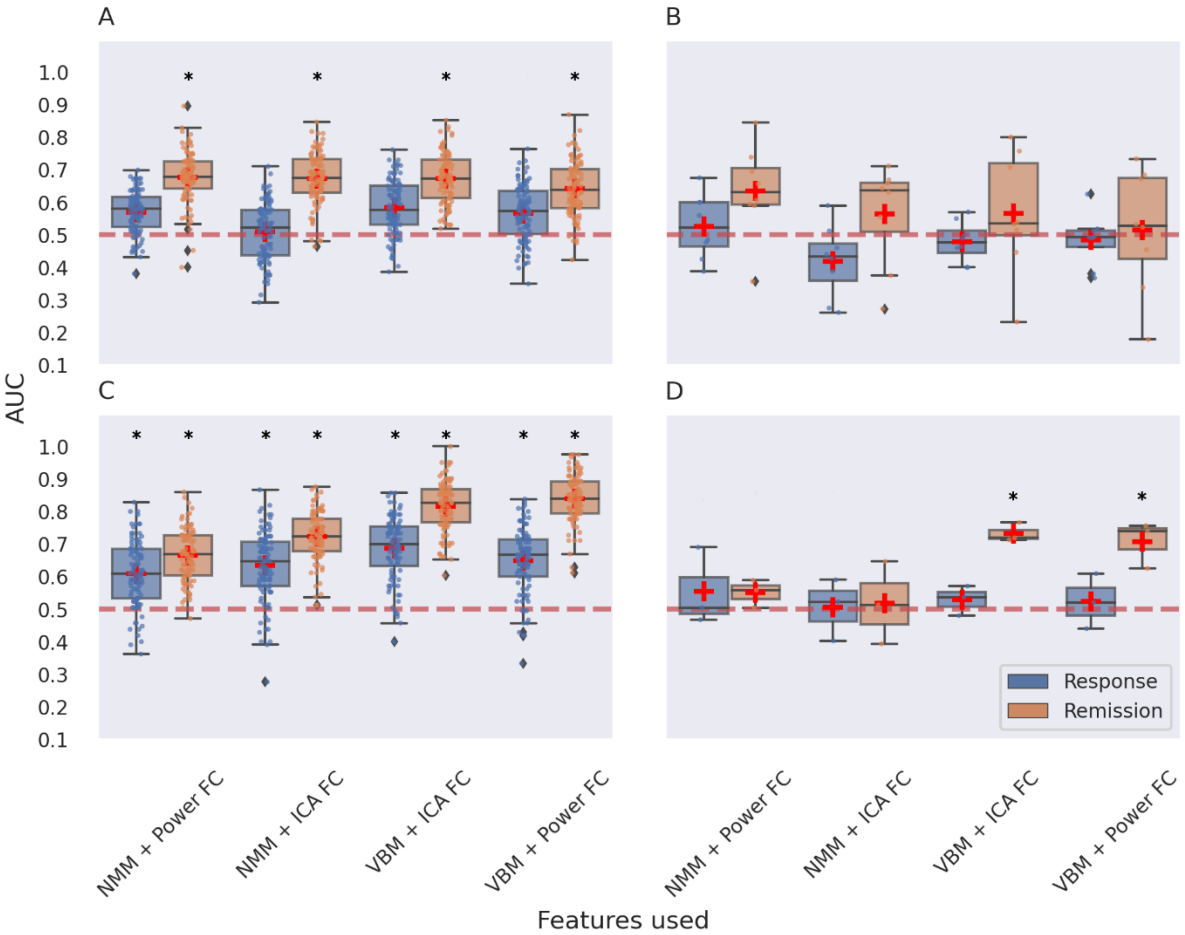

Figure 3. Multimodal multi-center predictions for ECT response and remission. Panel A depicts classification performance using data from all centers and different combinations of features with internal validation (AUC is averaged over 100 stratified cross-validation splits). Panel B shows classification performance using data from all centers and different combinations of features with external validation. Panel C depicts classification performance using data from the three largest centers with internal validation. Panel D shows classification performance using data from the three largest centers with external validation. VBM = voxel-based morphometry; NMM = Neuromorphometrics atlas; FC = functional connectivity; ICA = group information guided independent component analysis. Red dashed line depicts chance level performance (0.5 AUC). Asterisks indicate significant difference from chance level after permutation testing with false discovery rate correction for multiple comparisons ($p < 0.05$, corrected).

Learning curves

To evaluate the relation between sample size and classification performance, we examined learning curves for the best performing models (i.e. remission classification using data from the three largest centers) by subsampling the data using different proportions. Classification accuracy reached 0.83-0.84 AUC for unimodal (voxel-wise GM) and multimodal (voxel-wise GM and ICA-based FC) classifiers, with averaged AUC>0.75 for resamplings at 50% of the data (N=55) and AUC>0.8 for resamplings at 85% of the data (N=88). See **Supplementary Figure 2** for full learning curves.

Anatomical localization

We investigated which brain regions contributed most to treatment classification for voxel-wise GM data. We only focus on our best performing unimodal model, which for remission classification resulted in 0.83 AUC using data from the three largest samples. P-values were plotted for GM weights only as we were interested in brain regions rather than the influence of covariates. As shown in Figure 5, regions located in dorsomedial prefrontal (dmPFC), precuneus and thalamus exhibited high contribution to the classification task. The sign of weights within thalamus was mostly negative, implying a high chance for non-remission classification**,** whereas signs of weights within dmPFC and precuneus were mostly positive, implying a high chance for remission classification. Note that these results reflected the contribution of these brain regions to the multivariate pattern used by the SVM classifier.

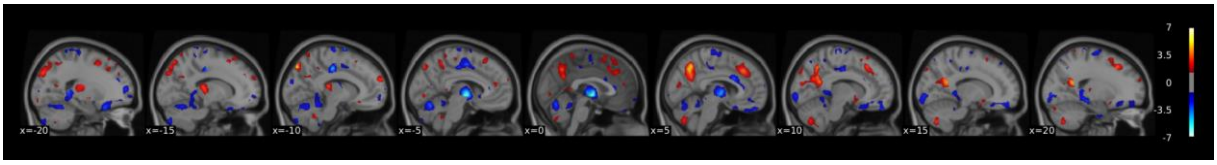

Figure 4. Thresholded -log(p) value maps characterizing the regions important for the treatment remission classification using voxel-wise GM data of the three largest centers (thresholded at p<0.05 uncorrected). Hot colors indicate positive weights and cold colors indicate negative weights of the SVM. The figure was made with the nilearn package (http://nilearn.github.io).

Discussion

These results show that neuroimaging data can provide a good prediction of ECT remission for individual patients across different centers. In line with recent meta-analyses, older age and higher depression severity at baseline were associated with better ECT outcome[7, 43, 44]. However, our classification results show that this information is not sufficient for making individual predictions, highlighting the relevance of obtaining neuroimaging data for accurate predictions. Remission classification using a combination of voxel-wise GM with either ICA-based FC or Power-atlas based FC led to good performing models when trained and tested on samples coming from each center (internal validation AUC>0.8), and remained acceptable when validated on completely new data from other centers (external validation AUC>0.7). These results indicate that multimodal neuroimaging data may provide a robust biomarker that could be used to guide clinical decision-making. By providing patients and clinicians a patient-specific prognosis, this could ultimately increase the success rate of ECT, avoid ineffective treatments and accompanying adverse effects, and increase the use of the most effective antidepressive treatment available.

Previous monocenter studies using neuroimaging data to predict ECT outcome with either structural or functional MRI were able to obtain up to 0.84 AUC[21]. Here we achieved similar classification performance in a multicenter setting. Using data from different samples involves many additional sources of technological (e.g., different MR hardware and scanner protocols) and clinical (e.g. different ECT protocols, patient

cohort and recruitment procedures) variability[23]. These additional sources of variability may decrease prediction accuracy of MRI measurements for ECT outcome[23, 45]. Conversely, a multicenter study avoids cohort-specific solutions and so helps test generalizability of the results across different samples, increasing the likelihood that features identified as discriminatory between remitters and non-remitters reflect generic properties related to treatment outcome across datasets. Our results showed that generalizability to new samples came at the cost of lower accuracy, as classifications performed with internal validation (AUC≈0.83) outperformed those using external validation (AUC≈0.72). Additionally, we found that using a subsample of the data containing three centers with N≥20 each (N=109) led to better model performance compared to using all eight centers (N=189). This improvement could not be attributed solely to reduced clinical heterogeneity, as differences in sample demographics and clinical characteristics between the three largest centers were found to be similar to those seen in the entire sample (**Supplementary Tables 1-2**). We therefore hypothesize that the exclusion of smaller centers ensured that the model had sufficient examples per center for training.

Brain regions that contributed most to remission classification using structural MRI data included dmPFC, precuneus and (hypo)thalamus. Our results also indicated a role for thalamus FC, as classification of the thalamus ICA resulted in the best performing functional MRI classifier. The thalamus is a hub connecting all cortico-cortical circuits with links to hippocampus and medial PFC. It is a central hub in the affective network and plays an important role in emotion dysregulation[12, 46-49]. There

is evidence of decreased thalamic volume in depression[50-53] and hyperactivity during rest and cognitive and emotion processing[54, 55]. Additionally, it has been suggested that seizure propagation between distant brain regions through cortical–thalamocortical and direct cortical–cortical connections is pivotal for ECT effectiveness[56-60]. There is also evidence that links reduced thalamic volume and altered rs-fMRI connectivity with clinical improvement[12, 19]. The precuneus is the core of the posterior default mode network and is associated with self-related processing and episodic memory retrieval, and has shown altered FC in depression[61-63]. Preliminary evidence links changes in precuneus network connectivity and structure with ECT treatment outcome[64, 65]. Altogether, these results provide evidence for the importance of thalamic and precuneus structure and their functional connectivity with other brain regions for both depression and ECT-related clinical response. Several ECT outcome prediction studies using structural MRI have also implicated the precuneus[15, 19, 20], and studies using rs-fMRI have reported functional connectivity with the thalamus as important regions[18, 60]. Notably, the identification of brain regions contributing most to the classification resulted from a multivariate analysis, and the localization of these regions should therefore be interpreted with caution as these regions may not only be related to treatment outcome but also contribute to denoising during the classification process[66]. Several limitations have to be taken into account when interpreting our findings. We used a retrospectively pooled sample from existing data across the world, without harmonized protocols for scanning, inclusion criteria or demographic and clinical characteristics. Not surprisingly, we found significant differences in sample

demographics and clinical characteristics between the different data collection centers. These sources of heterogeneity may limit classification performance but also provide an opportunity for model development using independent data sets and the discovery of generalizable biomarkers that are reproducible across centers. However, classification performance might be improved by using standardized acquisition parameters for possible future clinical utility. Additionally, our findings show that the prediction of treatment response was poor, while prediction of remission was good. This indicates that ECT outcome prediction is limited to remission, which may also provide a better outcome criterion compared to response. Remission has become the gold standard for depression treatment, because patients who do not remit have a poorer prognosis and greater chance of relapse and recurrence than those who do. Remission is also associated with a lower full symptomatic recurrence rate compared with achieving treatment response[7, 67, 68]. Furthermore, while unimodal and multimodal models performed comparable for remission classification using data from the largest centers with internal validation, only the multimodal classifications remained acceptable with external validation on different centers. We speculate that multimodal data may increase the probability that either the structural or functional MRI data overlaps across centers.

Taken together, this study suggests that ECT remission can be accurately predicted using MRI data in a large, ecologically valid, multi-center sample of patients receiving ECT, indicating that future development of a clinical decision support tool might be feasible. MRI could easily be incorporated during decision making, as ECT is

typically provided in a hospital setting. And as MRI is inexpensive compared to ECT, the additional costs are expected to outweigh the costs of unsuccessful treatments.

Acknowledgements

We would like to thank the logistic and academic support of the entire GEMRIC consortium. The full overview of the GEMRIC board members can be found here: https://mmiv.no/gemric/. This work was supported by the Netherlands Organization for Scientific Research (NWO/ZonMW Vidi 917.15.318, Dr. van Wingen), Western Norway Regional Health Authority (Grant No. 91223, Dr. Oltedal), NARSAD Young Investigator Grant (No. 27786 to BW) and a K99 Pathway to Independence Award (Grant No. MH119314) and the National Institute of Mental Health (Grant No. MH092301 and MH110008 for Dr. Narr and Dr. Randall; MH111826 and MH125126 for Dr. Abbott; MH119616 for Dr. Argyelan and R01MH112737 for Dr. Camprodon).

Conflict of interest

Dr. van Wingen has received research grant support from Philips. Dr. Camprodon serves in the Scientific Advisory Board of Hyka Therapeutics and Feelmore Labs, and has been a consultant for Neuronetics. All other individually-named co-authors in the GEMRIC working group reported no biomedical financial interests or potential conflicts of interest.

References

1. Kellner CH, Greenberg RM, Murrough JW, Bryson EO, Briggs MC, Pasculli RM. ECT in treatment-resistant depression. *Am J Psychiatry* 2012; **169**(12)**:** 1238-1244.

2. Scheepens D, Van Waarde J, Lok A, Zantvoord J, Pont BD, Ruhé H *et al.* Elektroconvulsietherapie bij persisterende depressie in Nederland; zeer lage toepassingsgraad. *Tijdschrift voor Psychiatrie* 2019**:** 16-21.

3. Slade EP, Jahn DR, Regenold WT, Case BG. Association of Electroconvulsive Therapy With Psychiatric Readmissions in US Hospitals. *JAMA Psychiatry* 2017; **74**(8)**:** 798-804.

4. Heijnen WT, Birkenhager TK, Wierdsma AI, van den Broek WW. Antidepressant pharmacotherapy failure and response to subsequent electroconvulsive therapy: a meta-analysis. *J Clin Psychopharmacol* 2010; **30**(5)**:** 616-619.

5. Hobson RF. Prognostic factors in electric convulsive therapy. *J Neurol Neurosurg Psychiatry* 1953; **16**(4)**:** 275-281.

6. Haq AU, Sitzmann AF, Goldman ML, Maixner DF, Mickey BJ. Response of depression to electroconvulsive therapy: A meta-analysis of clinical predictors. *Journal of Clinical Psychiatry* 2015; **76**(10)**:** 1374-1384.

7. van Diermen L, van den Ameele S, Kamperman AM, Sabbe BCG, Vermeulen T, Schrijvers D *et al.* Prediction of electroconvulsive therapy response and remission in major depression: meta-analysis. *British Journal of Psychiatry* 2018; **212**(2)**:** 71-80.

8. Loo CK, Mahon M, Katalinic N, Lyndon B, Hadzi-Pavlovic D. Predictors of response to ultrabrief right unilateral electroconvulsive therapy. *J Affect Disord* 2011; **130**(1-2)**:** 192-197.

9. Lopez AD, Murray CC. The global burden of disease, 1990-2020. *Nature Medicine* 1998; **4**(11)**:** 1241-1243.

10. Petrides G, Fink M, Husain MM, Knapp RG, Rush AJ, Mueller M *et al.* ECT remission rates in psychotic versus nonpsychotic depressed patients: a report from CORE. *The Journal of ECT* 2001; **17**(4)**:** 244-253.

11. van Waarde JA, Scholte HS, van Oudheusden LJ, Verwey B, Denys D, van Wingen GA. A functional MRI marker may predict the outcome of electroconvulsive therapy in severe and treatment-resistant depression. *Molecular psychiatry* 2015; **20**(5)**:** 609-614.

12. Leaver AM, Espinoza R, Pirnia T, Joshi SH, Woods RP, Narr KL. Modulation of Intrinsic Brain Activity by Electroconvulsive Therapy in Major Depression. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 2016; **1:** 77-86.

13. Cao B, Luo Q, Fu Y, Du L, Qiu T, Yang X *et al.* Predicting individual responses to the electroconvulsive therapy with hippocampal subfield volumes in major depression disorder. *Scientific Reports* 2018; **8**(1)**:** 5434.

14. First M, Botteron K, Carter C, Castellanos FX, Dickstein DP, Drevets W *et al.* Consensus Report of the APA Work Group on Neuroimaging Markers of Psychiatric Disorders. *APA Ofiicial Actions* 2012**:** 1-38.

15. Jiang R, Abbott CC, Jiang T, Du Y, Espinoza R, Narr KL *et al.* SMRI Biomarkers Predict Electroconvulsive Treatment Outcomes: Accuracy with Independent Data Sets. *Neuropsychopharmacology* 2018; **43**(5)**:** 1078-1087.

16. Leaver AM, Wade B, Vasavada M, Hellemann G, Joshi SH, Espinoza R *et al.* Fronto-Temporal Connectivity Predicts ECT Outcome in Major Depression. *Front Psychiatry* 2018; **9:** 92.

17. Redlich R, Opel N, Grotegerd D, Dohm K, Zaremba D, Burger C *et al.* Prediction of Individual Response to Electroconvulsive Therapy via Machine Learning on Structural Magnetic Resonance Imaging Data. *JAMA Psychiatry* 2016; **73**(6)**:** 557-564.

18. Sun H, Jiang R, Qi S, Narr KL, Wade BS, Upston J *et al.* Preliminary prediction of individual response to electroconvulsive therapy using whole-brain functional magnetic resonance imaging data. *NeuroImage: Clinical* 2020; **26:** 102080.

19. Takamiya A, Liang KC, Nishikata S, Tarumi R, Sawada K, Kurokawa S *et al.* Predicting Individual Remission After Electroconvulsive Therapy Based on Structural Magnetic Resonance Imaging: A Machine Learning Approach. *The Journal of ECT* 2020; **36**(3)**:** 205-210.

20. Wade BS, Joshi SH, Njau S, Leaver AM, Vasavada M, Woods RP *et al.* Effect of Electroconvulsive Therapy on Striatal Morphometry in Major Depressive Disorder. *Neuropsychopharmacology* 2016; **41**(10)**:** 2481-2491.

21. Cohen SE, Zantvoord JB, Wezenberg BN, Bockting CLH, van Wingen GA. Magnetic resonance imaging for individual prediction of treatment response in major depressive disorder: a systematic review and meta-analysis. *Translational Psychiatry* 2021; **11**(1)**:** 168.

22. Nieuwenhuis M, van Haren NE, Hulshoff Pol HE, Cahn W, Kahn RS, Schnack HG. Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. *Neuroimage* 2012; **61**(3)**:** 606-612.

23. Schnack HG, Kahn RS. Detecting Neuroimaging Biomarkers for Psychiatric Disorders: Sample Size Matters. *Frontiers in Psychiatry* 2016; **7:** 50.

24. Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *Neuroimage* 2017; **145**(Pt B)**:** 166-179.

25. Oltedal L, Bartsch H, Sørhaug OJE, Kessler U, Abbott C, Dols A *et al.* The Global ECT-MRI Research Collaboration (GEMRIC): Establishing a multi-site investigation of the neural mechanisms underlying response to electroconvulsive therapy. *NeuroImage: Clinical* 2017; **14:** 422-432.

26. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *Eur J Clin Invest* 2015; **45**(2)**:** 204-214.

27. Oltedal L, Narr KL, Abbott C, Anand A, Argyelan M, Bartsch H *et al.* Volume of the Human Hippocampus and Clinical Response Following Electroconvulsive Therapy. *Biological Psychiatry* 2018; **84**(8)**:** 574-581.

28. Heo M, Murphy CF, Meyers BS. Relationship between the hamilton depression rating scale and the montgomery-...sberg depression rating scale in depressed elderly: A meta-manalysis. *American Journal of Geriatric Psychiatry* 2007; **15**(10)**:** 899-905.

29. Ashburner J. A fast diffeomorphic image registration algorithm. *Neuroimage* 2007; **38**(1)**:** 95-113.

30. Pruim RHR, Mennes M, van Rooij D, Llera A, Buitelaar JK, Beckmann CF. ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. *Neuroimage* 2015; **112:** 267-277.

31. Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* 2002; **17**(2)**:** 825-841.

32. Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. FSL. *Neuroimage* 2012; **62**(2)**:** 782-790.

33. Salman MS, Du Y, Lin D, Fu Z, Fedorov A, Damaraju E *et al.* Group ICA for identifying biomarkers in schizophrenia: 'Adaptive' networks via spatially constrained ICA show more sensitivity to group differences than spatio-temporal regression. *NeuroImage: Clinical* 2019; **22:** 101747.

34. Du Y, Fan Y. Group information guided ICA for fMRI data analysis. *Neuroimage* 2013; **69:** 157-197.

35. Power JD, Cohen AL, Nelson SM, Wig GS, Barnes KA, Church JA *et al.* Functional network organization of the human brain. *Neuron* 2011; **72**(4)**:** 665-678.

36. Chang C-c, Lin C-j. LIBSVM : A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2013; **2:** 1-39.

37. Abraham A, Milham MP, Di Martino A, Craddock RC, Samaras D, Thirion B *et al.* Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. *Neuroimage* 2017; **147:** 736-745.

38. Bradley AP. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition* 1997; **30**(7)**:** 1145-1159.

39. Hosmer Jr DW, Lemeshow S, Sturdivant RX. *Applied logistic regression*, vol. 398. John Wiley & Sons2013.

40. Ojala M, Garriga GC. Permutation Tests for Studying Classi er Performance. *Journal ofMachine Learning Research* 2010; **11:** 1833-1863.

41. Kottas M, Kuss O, Zapf A. A modified Wald interval for the area under the ROC curve (AUC) in diagnostic case-control studies. *BMC Medical Research Methodology* 2014; **14:** 26.

42. Gaonkara B, Shinohara RT, Davatzikos C. Interpreting support vector machine models for multivariate group wise analysis in neuroimaging. *Medical Image Analysis* 2016; **1848:** 3047-3054.

43. Nordenskjold A, von Knorring L, Engstrom I. Predictors of the short-term responder rate of Electroconvulsive therapy in depressive disorders--a population based study. *BMC Psychiatry* 2012; **12:** 115.

44. Yao Z, McCall WV, Essali N, Wohl E, Parker C, Rosenquist PB *et al.* Precision ECT for major depressive disorder: A review of clinical factors, laboratory, and physiologic biomarkers as predictors of response and remission. *Personalized Medicine in Psychiatry* 2019; **17-18:** 23-31.

45. Woo CW, Chang LJ, Lindquist MA, Wager TD. Building better biomarkers: brain models in translational neuroimaging. *Nature Neuroscience Reviews* 2017; **20**(3)**:** 365-377.

46. Mayberg HS. Limbic-cortical dysregulation: a proposed model of depression. *J Neuropsychiatry Clin Neurosci* 1997; **9**(3)**:** 471-481.

47. Phillips ML, Drevets WC, Rauch SL, Lane R. Neurobiology of emotion perception II: Implications for major psychiatric disorders. *Biological Psychiatry* 2003; **54**(5)**:** 515-528.

48. Price JL, Drevets WC. Neural circuits underlying the pathophysiology of mood disorders. *Trends Cogn Sci* 2012; **16**(1)**:** 61-71.

49. Vertes RP. Interactions among the medial prefrontal cortex, hippocampus and midline thalamus in emotional and cognitive processing in the rat. *Neuroscience* 2006; **142**(1)**:** 1-20.

50. Bora E, Harrison BJ, Davey CG, Yucel M, Pantelis C. Meta-analysis of volumetric abnormalities in cortico-striatal-pallidal-thalamic circuits in major depressive disorder. *Psychol Med* 2012; **42**(4)**:** 671-681.

51. Kempton MJ, Salvador Z, Munafo MR, Geddes JR, Simmons A, Frangou S *et al.* Structural neuroimaging studies in major depressive disorder. Meta-analysis and comparison with bipolar disorder. *Archives of General Psychiatry* 2011; **68**(7)**:** 675-690.

52. Sartorius A, Demirakca T, Bohringer A, Clemm von Hohenberg C, Aksay SS, Bumb JM *et al.* Electroconvulsive therapy increases temporal gray matter volume and cortical thickness. *Eur Neuropsychopharmacol* 2016; **26**(3)**:** 506-517.

53. Soriano-Mas C, Hernandez-Ribas R, Pujol J, Urretavizcaya M, Deus J, Harrison BJ *et al.* Cross-sectional and longitudinal assessment of structural brain alterations in melancholic depression. *Biol Psychiatry* 2011; **69**(4)**:** 318-325.

54. Hamilton JP, Farmer M, Fogelman P, Gotlib IH. Depressive Rumination, the Default-Mode Network, and the Dark Matter of Clinical Neuroscience. *Biol Psychiatry* 2015; **78**(4)**:** 224-230.

55. Palmer SM, Crewther SG, Carey LM, Team SP. A meta-analysis of changes in brain activity in clinical depression. *Frontiers in Human Neuroscience* 2014; **8:** 1045.

56. Fink M, Ottosson JO. A theory of convulsive therapy in endogenous depression: significance of hypothalamic functions. *Psychiatry Res* 1980; **2**(1)**:** 49-61.

57. Leaver AM, Vasavada M, Kubicki A, Wade B, Loureiro J, Hellemann G *et al.* Hippocampal subregions and networks linked with antidepressant response to electroconvulsive therapy. *Molecular Psychiatry* 2020.

58. McNally KA, Blumenfeld H. Focal network involvement in generalized seizures: new insights from electroconvulsive therapy. *Epilepsy Behav* 2004; **5**(1)**:** 3-12.

59. Singh A, Kar SK. How Electroconvulsive Therapy Works?: Understanding the Neurobiological Mechanisms. *Clinical Psychopharmacology and Neuroscience* 2017; **15**(3)**:** 210-221.

60. Takamiya A, Kishimoto T, Liang KC, Terasawa Y, Nishikata S, Tarumi R *et al.* Thalamic volume, resting-state activity, and their association with the efficacy of electroconvulsive therapy. *Journal of Psychiatric Research* 2019; **117:** 135-141.

61. Cavanna AE, Trimble MR. The precuneus: a review of its functional anatomy and behavioural correlates. *Brain* 2006; **129**(Pt 3)**:** 564-583.

62. Mulders PC, van Eijndhoven PF, Schene AH, Beckmann CF, Tendolkar I. Resting-state functional connectivity in major depressive disorder: A review. *Neurosci Biobehav Rev* 2015; **56:** 330-344.

63. Utevsky AV, Smith DV, Huettel SA. Precuneus is a functional core of the default-mode network. *J Neurosci* 2014; **34**(3)**:** 932-940.

64. Mulders PC, van Eijndhoven PF, Pluijmen J, Schene AH, Tendolkar I, Beckmann CF. Default mode network coherence in treatment-resistant major depressive disorder during electroconvulsive therapy. *Journal of Affective Disorders* 2016; **205:** 130-137.

65. Mulders PCR, Llera A, Beckmann CF, Vandenbulcke M, Stek M, Sienaert P *et al.* Structural changes induced by electroconvulsive therapy are associated with clinical outcome. *Brain Stimulation* 2020; **13**(3)**:** 696-704.

66. Haufe S, Meinecke F, Gorgen K, Dahne S, Haynes JD, Blankertz B *et al.* On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage* 2014; **87:** 96-110.

67. McIntyre RS, O'Donovan C. The human cost of not achieving full remission in depression. *Canadian Journal of Psychiatry* 2004; **49**(1)**:** 10-16.

68. Trivedi HM, Daly EJ. Treatment strategies to improve and sustain remission in major depressive disorder. *Dialogues in Clinical Neuroscience* 2009; **10**(4)**:** 377-384.