Generalizability of treatment outcome prediction in major depressive disorder using structural MRI
A NeuroPharm study
Beliveau, Vincent; Hedeboe, Ella; Fisher, Patrick M; Dam, Vibeke H; Jørgensen, Martin B; Frokjaer, Vibe G; Knudsen, Gitte M; Ganz, Melanie

Published in:
NeuroImage. Clinical

DOI:
10.1016/j.nicl.2022.103224

Publication date:
2022

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY-NC-ND

Citation for published version (APA):
Generalizability of treatment outcome prediction in major depressive disorder using structural MRI: A NeuroPharm study

Vincent Beliveau\textsuperscript{a,b,*}, Ella Hedebøe\textsuperscript{a}, Patrick M. Fisher\textsuperscript{a}, Vibeke H. Dam\textsuperscript{a}, Martin B. Jørgensen\textsuperscript{a,c,d}, Vibe G. Frokjaer\textsuperscript{a,c,g}, Gitte M. Knudsen\textsuperscript{a,c}, Melanie Ganz\textsuperscript{a,e}

\textsuperscript{a} Neurobiology Research Unit and NeuroPharm, Copenhagen University Hospital Rigshospitalet, Copenhagen, Denmark
\textsuperscript{b} Department of Neurology, Medical University of Innsbruck, Innsbruck, Austria
\textsuperscript{c} Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
\textsuperscript{d} Psychiatric Center Copenhagen, Rigshospitalet, Copenhagen, Denmark
\textsuperscript{e} Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

ARTICLE INFO

Keywords:
MDD
Structural MRI
Prediction
Treatment response
Remission
SSRI

ABSTRACT

Brain morphology has been suggested to be predictive of drug treatment outcome in major depressive disorders (MDD). The current study aims at evaluating the performance of pretreatment structural brain magnetic resonance imaging (MRI) measures in predicting the outcome of a drug treatment of MDD in a large single-site cohort, and, importantly, to assess the generalizability of these findings in an independent cohort. The random forest, boosted trees, support vector machines and elastic net classifiers were evaluated in predicting treatment response and remission following an eight week drug treatment of MDD using structural brain measures derived with FastSurfer (FreeSurfer). Models were trained and tested within a nested cross-validation framework using the NeuroPharm dataset (n = 79, treatment: escitalopram); their generalizability was assessed using an independent clinical dataset, EMBARC (n = 64, treatment: sertraline). Prediction of antidepressant treatment response in the NeuroPharm cohort was statistically significant for the random forest (p = 0.048), whereas none of the models could significantly predict remission. Furthermore, none of the models trained using the entire NeuroPharm dataset could significantly predict treatment outcome in the EMBARC dataset. Although our primary findings in the NeuroPharm cohort support some, but limited value in using pretreatment structural brain MRI to predict drug treatment outcome in MDD, the models did not generalize to an independent cohort suggesting limited clinical applicability. This study emphasizes the importance of assessing model generalizability for establishing clinical utility.

1. Introduction

Major Depressive Disorder (MDD) is one of the most prevalent and severe brain disorders in the world with 6.9% of the European population estimated to suffer from the disease, making it the most burdensome disease in Europe (Wittchen et al., 2011). MDD is a highly heterogeneous disorder where the diagnosis is based on the presence of a set of symptoms leading to 227 unique ways to meet the criteria for the MDD based on the Diagnostic Statistical Manual (DSM-5) (Zimmerman et al., 2015). MDD can be treated in many different ways, including pharmaceuticals, psychotherapy, electroconvulsive therapy, and other somatic therapies, or combinations thereof (Gartlehner et al., 2017). Whereas selective serotonin/noradrenaline reuptake inhibitors (SSRI/SNRI) are first-line pharmaceutical treatment for MDD, only 40–60% of patients respond clinically, and only 30–45% achieve clinical remission (Carvalho et al., 2007; Khin et al., 2011; Thase et al., 2001). Moreover, up to a third of patients will not respond to a second-line medication either (Rush et al., 2006). Identifying biomarkers that enable the clinician to choose the optimal medication for the patient is seen as key for a precision medicine approach and accordingly, many attempts to define such a biomarker have been made (Pu et al., 2012). Such a biomarker would be valuable to guide the treatment of MDD and lower the rate of unsuccessful therapies, which would be of immense benefit to patients and reduce the associated economic burden.

Magnetic resonance imaging (MRI) is a prevalent and non-invasive method for measuring brain structure and function. Structural...
properties of the brain derived from MRI modalities, such as T1-weighted images, have the advantage of being relatively easy to acquire and, therefore, carry a strong potential as biomarkers for clinical application. The structure of the brain has reliably been shown to be altered in MDD patients compared to healthy controls. Although the reported effect sizes are small (Cohen’s d < 0.2) and it seems to be mainly associated with late-onset MDD (Eker and Gonul, 2010), many studies have reported reduced hippocampal volume in MDD (Eker and Gonul, 2010). Furthermore, several studies have identified associations between treatment outcome and regional cortical thickness and volume, as well as whole brain volume (Fonseka et al., 2018; Frodl et al., 2008; Jarvum et al., 2011; Phillips et al., 2015). As reviewed in (Emmeking et al., 2020), multiple studies have thus investigated if structural brain characteristics can be used as biomarkers of response to antidepressant drug treatment in MDD.

A handful of studies have attempted to use morphological features of the brain derived from pre-treatment MRI to predict the outcome (i.e., response or remission) to antidepressant treatment in MDD with varying study protocols (e.g., choice of antidepressant drug) and methodology (Bartlett et al., 2011; Costafreda et al., 2009; Gong et al., 2011; Grzenda et al., 2021, p. 1; Liu et al., 2012; Nouretdinov et al., 2011). Earlier studies in small cohorts (n = 18 to 46) adopted similar approaches and classified treatment outcome using variants of voxel-based morphometry (features) combined with support vector machine classifiers (Costafreda et al., 2009; Gong et al., 2011; Liu et al., 2012; Nouretdinov et al., 2011). In two studies, sparse clusters showed high predictive accuracy for drug treatment response, but these clusters could not readily be associated with neurobiologically meaningful functions (Gong et al., 2011; Liu et al., 2012). Two additional reports by the same group applying a different methodology on the same data reported greater gray matter density in the right rostral anterior cingulate cortex, left posterior cingulate cortex, left middle frontal gyrus, and right occipital cortex as a predictive marker of treatment remission (Costafreda et al., 2009; Nouretdinov et al., 2011). The predictive performances of these studies were evaluated using leave-one-out cross-validation designs, and accuracy ranged from 70% to 89%. However, the use of leave-one-out cross-validation has since been strongly discouraged, particularly for small cohorts, due to an increased risk of performance misestimation (Flint et al., 2021; Varouquaux et al., 2017). In a more recent study including a large (n = 184) multi-center cohort (Establishing moderators and biosignatures of antidepressant response in clinical care; EMBARC), the prediction of remission 8 weeks after initiation of sertraline treatment (n = 87) or placebo (n = 97) was assessed using a random forest classifier applied to estimates of regional cortical thickness and subcortical volumes (Bartlett et al., 2018). Notably, the features were obtained from structural MRIs acquired both before and one week after the intervention was initiated. When the performance of this approach was evaluated using repeated fivefold cross-validation, an accuracy of 64% was achieved. Post hoc analyses also revealed that thickening of the rostral anterior cingulate was associated with better responses to sertraline. In spite of these promising initial findings on the clinical usefulness of brain structural information for predicting the outcome of antidepressant treatment in MDD, these results remain to be replicated in an independent cohort.

Generalizability, the applicability of a model to an unseen dataset, is a core principle of machine learning and is essential to establishing the clinical usefulness of predictive models. Even when all proper procedures for establishing evidence for prediction are followed within a dataset (Pollack et al., 2020), the real limit test for a model is the out-of-sample prediction performance (Cohen et al., 2021). Recent large-scale efforts have been made toward determining the generalizability of models for the classification of MDD patients from healthy controls using functional MRI (Believ et al., Unpublished results). However, to our knowledge, the generalizability of models using pretreatment structural MRI to predict pharmacological treatment outcome in MDD has never been evaluated. Given the wide phenotypic heterogeneity of MDD, it is crucial to determine whether these models can be used to predict the treatment outcome of MDD patients from new sites.

We here assess the ability of structural brain MRI measured prior to treatment to predict the response and remission to 8 weeks of treatment with the SSRI escitalopram among complying patients. To this end, we leveraged data from the NeuroPharm cohort (Köhler-Forsberg et al., 2020), the largest single-site study of drug treatment outcome in MDD. We hypothesized that in MDD, structural brain MRI scans acquired prior to treatment-onset would have predictive value in determining the response to drug treatment. Thereafter, we evaluate the generalizability of the models trained using the complete NeuroPharm cohort in an independent dataset (EMBARC).

2. Patients & methods

2.1. Participants

The current study is based on the NeuroPharm dataset, a large, single-site, non-randomized, single-treatment, naturalistic, open-label clinical trial evaluating neuroimaging, biochemical, EEG, and neuropsychological measures as biomarkers of antidepressant treatment outcome in depressed patients (https://www.clinicaltrials.gov, NCT02869035).

Here we summarize relevant study design elements, which are described elsewhere in greater detail (Köhler-Forsberg et al., 2020). One hundred untreated patients with MDD were initially included in the study. Patients were recruited from a central referral center within the Mental Health Services, Capital Region of Denmark or referred directly from one of five general practitioners. At inclusion, patients had to be between 18 and 65 years old and were required to meet the DSM-5 criteria for single or recurrent unipolar depression. Patients were required to be moderately to severely depressed, i.e., a score greater than 17 on the Hamilton Depression Rating Scale-17 item (HAM-D17) (Hamilton, 1960). Clinical diagnosis was confirmed by an experienced psychiatrist and confirmed with the “Mini-International Neuropsychiatric Interview” (Sheehan et al., 1998).

The study protocol was approved by the Ethics Committee (H-15017713), the Danish Data Protection Agency, and Danish Medicines Agency (protocol number: NeuroPharm-NP1, EudraCT-number 2016-001,626–34). The study complies with the Declaration of Helsinki II, and a Good Clinical Practice unit in the Capital Region of Denmark monitored the project. All participants signed written informed consent after an oral and written description of the study. Patients did not receive compensation for participation.

2.2. Treatment protocol

The full treatment protocol is detailed elsewhere (Köhler-Forsberg et al., 2020). Briefly, after inclusion, MDD individuals underwent, among other things, baseline assessments of depression- and anhedonic severity, neuropsychological testing, and a baseline MRI scan. Patients subsequently entered a treatment protocol with escitalopram to last up to twelve weeks. Escitalopram was the primary pharmacological treatment (flexible dosages: 5–20 mg/day), administered in alignment with current clinical practice. Patients not responding (<25% decrease in HAMD-6) or with unacceptable side effects could switch to duloxetine from week 4 and onward (flexible dosages: 30–120 mg/day), consistent with clinical guidelines. Antidepressant medication was provided free of charge. Non-compliant patients, i.e., reporting to have taken <2/3 of their tablets or with serum concentrations of medicine below the detection limit at week 6, were excluded. Depression severity and side effects were assessed by a study physician or supervised research assistant during clinical follow-up sessions at 1, 2, 4, 8, and 12 weeks after treatment. Patients did not receive any other form of treatment than the described antidepressant regimen, including psychotherapy, for the duration of the trial.
2.3. Clinical assessment

Depression-severity at baseline depression severity and throughout the study was monitored by the Hamilton Depression Rating Scale (HAMD-17) and its subscale of 6 items (HAMD-6) (Timmerby et al., 2017). HAMD-6 was chosen as the primary outcome for the evaluation of depression based on recent evidence of superior sensitivity to change in depression severity and clinimetric properties compared to HAMD-17 (Dunlop et al., 2019). Response to antidepressant treatment was defined as reduction from baseline of at least 50% in HAMD-6 score at week 8, and remission was defined as HAMD-6 \( \leq 4 \) at week 8.

2.4. MRI data acquisition

All MRI scans were acquired using a Siemens (Erlangen, Germany) MAGNETOM Prisma 3T scanner with a 64-channel head coil. High-resolution, structural T1-weighted magnetization-prepared rapid gradient-echo structural scans were acquired (repetition time = 1900 ms, echo time = 2.58 ms, inversion time = 900 ms, flip angle = 9\(^\circ\), number of slices = 224, slice thickness = 0.9 mm, matrix = 256 \times 256, in-plane resolution 0.9 \times 0.9 mm, no gap, and acquisition time = 4 min 26 s).

2.5. Processing of structural MRI

The structural MRI data was processed using FastSurfer (FreeSurfer) (Henschel et al., 2020) (https://github.com/Deep-MI/FastSurfer). An overall quality check of the processing was performed to ensure that no large error in the segmentation process was present, however, no manual editing was performed to provide a fully automated and better clinically applicable evaluation. The average cortical thickness of the 34 regions, per hemisphere, defined by the Desikan-Killiany atlas (29) and the volume of eight subcortical regions (i.e., accumens, amygdala, caudate, cerebellar gray matter, hippocampus, pallidum, putamen, and thalamus), as well the volume of the lateral ventricles and the mean cortical thickness of each hemisphere were extracted to be used for comparison between the different groups and as features in the classification models. Intracranial volume was quantified using SPM12 (v7219, https://www.fil.ion.ucl.ac.uk/spm) in Matlab (R2019a, Math-Works, Natick, MA, USA) by segmenting the gray matter, white matter, and cerebrospinal fluid and summing their combined volume (Malone et al., 2015).

2.6. Assessment of group differences

Group differences in demographic characteristics between responders and non-responders or remitters and non-remitters, within and between datasets, were assessed using Wilcoxon rank sum tests and Pearson’s chi-squared test where appropriate. Similarly, group associations with cortical thickness or subcortical volumes were evaluated independently for each region of interest using linear regression models, with age and sex as covariate for estimates of cortical thickness, and additionally with intracranial volume for subcortical volumes (Malone et al., 2015). P-values were adjusted for multiple comparisons using the false discovery rate (FDR) method (Benjamini and Hochberg, 1995) and q < 0.05 was considered statistically significant. All statistical tests and classification analyses were performed using R v4.1.0 (R Core Team, 2013).

2.7. Prediction of treatment outcome

Prediction of treatment response or remission was evaluated using some of the most commonly used classifiers (caret, v6.0.92) (Kuhn, 2008): the random forest (randomForest, v4.7.1.1) (Breiman, 2001; Liaw and Wiener, 2002), boosted trees (xgboost, xgbTree, v1.6.0.1.0), support vector machines (SVM) (kernlab, svmRadialWeights, v0.9.31), and elastic net (glmnet, v4.1.4). Class weights were set to 1 - p, where p is the proportion of a given class among all samples. These weights were used to ensure equal representation of each class during training. Estimates of regional cortical thickness, subcortical volumes, mean cortical thickness of the whole hemisphere, and intracranial volumes, as well as age, sex, HAMD-6 at baseline (week 0), and recurrence status (i.e., first-episode or recurrent) were used as input features. Numerical features were centered and scaled independently per training set and applied to the respective test set. Model performance was assessed using the area under the ROC Curve (AUC), balanced accuracy, sensitivity, and specificity evaluated within a stratified nested cross-validation with a repeated cross-validation with 5-fold and 25 repeats as outer loop and a 5-fold cross-validation as nested loop for hyperparameters optimization, as recommended in (Varoquaux et al., 2017). Hyperparameters optimization for all classifiers aside from the random forest was performed according to the default implementation in caret. For the random forest, the optimization was performed for both the number of variables randomly sampled as candidates at each split (mtry) and the number of trees to grow (ntree). The statistical significance of the AUCs were empirically derived using a null distribution estimated from 1,000 permutations for the elastic net and random forest and 100 permutations for the boosted trees and SVM. A smaller number of permutations were used for boosted trees and SVM due to the computational load required to train these models.

2.8. Generalizability of the classifiers on the EMBARC dataset

EMBARC is a large, multicenter, double-blind, randomized, placebo-controlled trial evaluating the treatment response to the SSRI sertraline in patients with MDD. The rationale and design of this study have been previously described (Trivedi et al., 2016). Individuals from the EMBARC dataset, meeting the same inclusion criteria as for NeuroPharm, were identified. Notably, only patients having received sertraline treatment, with HAMD-17 greater than 17 at week 0, and with pre-treatment structural MRI at week 0 were included. Pre-treatment structural MRIs were processed using the approach described above and the corresponding brain morphological features were extracted. Models trained on NeuroPharm data were evaluated in the task of predicting treatment response and remission in the EMBARC dataset using the same input features (i.e., pre-treatment MRI measures, age, sex, HAMD-6 at week 0, and recurrence status). The EMBARC data was normalized to match the data used for training using the parametric ComBat (sva, v4.32.0) procedure (Johnson et al., 2007). Model performance was estimated using bootstrap with 1,000 resamples and the statistical significance of the mean AUC was empirically derived using a null distribution estimated from 1,000 permutations.

2.9. Voxel-based morphometry

As previous studies have largely been performed using voxel-based morphometry (VBM) (Costafreda et al., 2009; Gong et al., 2011; Liu et al., 2012; Noureddin et al., 2011), we briefly evaluated this approach in the NeuroPharm dataset. The standard VBM protocol was applied using SPM12: 1) the structural MR images were segmented into gray matter, white matter and cerebrospinal fluid and imported into the NeuroPharm dataset. The standard VBM protocol was applied to the respective test set. Model performance was assessed using measures of volume in subsequent VBM analyses.
Voxel-wise group differences in VBM volumes were assessed using F-tests. P-values were corrected for family-wise error and considered significant at $p < 0.05$. For classification, significant voxels to be included as features were identified using an uncorrected voxel-level threshold of $p < 0.005$ (Costafreda et al., 2009). Significant voxels were identified independently for each training set and used to extract VBM volumes in both the training and test sets. Classification using VBM volumes (in place of cortical thickness or subcortical volumes) was otherwise performed as described in Section 2.7.

2.10. Data and code availability

The NeuroPharm data can be made available upon request through an application to the Cimbi database (https://www.cimbi.dk). The EMBARC dataset can be accessed through the National Institute of Mental Health Data Archive (https://nda.nih.gov). The source code for this manuscript is freely available at https://github.com/vbeliveau/MDD_SSRI_structural_prediction.

3. Results

3.1. Demographics

A flow chart of the participants in the study is shown in Fig. 1. Of the 100 patients enrolled in the NeuroPharm study, three did not complete a pre-treatment structural MRI scan. Eighteen of the 97 remaining patients were lost at follow-up ($n = 18$), resulting in 79 MDD patients having both structural MRI data and clinical assessments at baseline and week 8. Out of 336 participants, 64 patients were included from the EMBARC dataset.

The demographics and clinical variables are presented in Table 1 for the NeuroPharm data and in Table S1 for the EMBARC dataset. No significant differences in demographic characteristics between the treatment outcome groups were found (all $p$-values $> 0.05$), aside from the expected difference in HAMD-6 at week 8 ($p < 0.001$). Based on the relative change in HAMD-6 at week 8 compared to baseline, a total of 46 (58.2 %) patients were labeled as responders and 33 (41.8 %) non-responders, and 34 (43.0 %) patients were labeled as remitters and 45...
derived based on the NeuroPharm data, we tested the models for pre-
treatment remission, none of the models obtained
responders and non-responders (t
EMBARC data, aside from the volume of the right putamen between
dataset in each of the treatment outcome groups.

classification models were trained to predict treatment response and
remission in the NeuroPharm dataset is reported
Tables S3-4 for details.

no significant group difference was identified in the pretreatment
volume pretreatment was found between responders and non-
responders nor between remitters and non-remitters in the Neuro-
Pharm dataset. Detailed results are presented in
Tables S1-2. Similarly, no significant group difference was identified in the pretreatment
EMBARC data, aside from the volume of the right putamen between
responders and non-responders (t = 3.75, q = 0.035, FDR-corrected), see
Tables S3-4 for details.

3.3. Prediction of treatment outcome within the NeuroPharm dataset

Detailed information on the performance of the models in predicting
treatment response and remission in the NeuroPharm dataset is reported in
Table 3. For treatment response, only the random forest classifiers
achieved mean AUCs significantly greater than the null (p = 0.048). For the
prediction of treatment remission, none of the models obtained
mean AUCs significantly greater than the null. Fig. 2 presents the mean
receiver operating characteristic (ROC) curves of the models.

3.4. Assessment of generalizability

To evaluate the generalizability of the identified predictive model
derived based on the NeuroPharm data, we tested the models for pre-
dicting response or remission in the EMBARC dataset. To this end,
classification models were trained to predict treatment response and
remission using all available data (i.e., n = 79) from the NeuroPharm
dataset. Applied to the EMBARC dataset, neither the elastic net nor the
random forest achieved AUCs significantly greater than the null for
treatment response (p = 0.57, and p = 0.53, respectively) and remission
(p = 0.69, and p = 0.83, respectively). Detailed performance metrics are
reported in Table S7. Due to their poor performance in the NeuroPharm
dataset, we did not evaluate the boosted trees and SVM models.

As a post hoc analysis, we evaluated the performance of classification
models trained on only the EMBARC dataset using the cross-validation
framework previously described. These models were not able to pre-
dict treatment response and remission better than chance (all mean
AUCs < 0.5). Fig. S1 presents the corresponding ROC curves and the
associated performance metrics are reported in Table S8.

3.5. Voxel-based morphometry

No significant group differences in VBM volumes were identified
between responders and non-responders nor between remitters and non-
remitters. The predictive performance of the classifiers using VBM vol-
umes were limited at best, with mean AUCs ranging from 0.42 to 0.58.
We note that p-values for the AUCs were not here estimated given their
closeness to 0.5 and the large computations this would entail. ROC
curves are presented in Fig. S2 and the associated performance metrics are
included in Table S9.

4. Discussion

Previous studies have claimed some, but limited success in predicting
the outcome of drug treatments in MDD using structural MRI, but so far,
no study has demonstrated the generalizability of such models in an
independent cohort. Generalizability is a cornerstone to any form of
clinical application and is a prerequisite to claims of usefulness beyond
research interest. This point is critical for assessing the results presented
in our study.

In this work, we evaluated the potential of pre-treatment structural
brain MRI for the prediction of antidepressant treatment outcome in
patients with MDD following an eight-week treatment program starting
with escitalopram. In the NeuroPharm dataset, no significant group
difference in pretreatment regional cortical thickness or subcortical
volume was observed between responders and non-responders, nor be-
tween remitters and non-remitters. Our classification models were first
trained and tested in-sample (NeuroPharm dataset only), and their
generalizability was subsequently assessed using a second independent
clinical test dataset (EMBARC). At first glance, the predictive models
evaluated in-sample using cross-validation suggested that distinct pro-
files of brain structure could be captured by our models enabling us to
predict treatment response above chance. In contrast to earlier, smaller
studies, the classification performances of our models were noticeably
reduced both for response and remission (Bartlett et al., 2018; Costa-
freda et al., 2009; Gong et al., 2011; Liu et al., 2012; Nouretdinov et al.,
2011), but were still comparable to that of (Bartlett et al., 2018) (Bartlett
et al., 2018). However, our models did not generalize to a new dataset

### Table 1
Demographics and clinical variables for remission and response in the NeuroPharm dataset. HAMD-6: Hamilton Depression Rating Scale-6 item.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Response Non-responder, N = 33</th>
<th>Response Responder, N = 46</th>
<th>p-value</th>
<th>Remission Non-remitter, N = 45</th>
<th>Remitter, N = 34</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>24 (22, 27)</td>
<td>25 (23, 29)</td>
<td>0.2</td>
<td>24 (22, 27)</td>
<td>26 (23, 29)</td>
<td>0.14</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td>0.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>26 (79 %)</td>
<td>32 (70 %)</td>
<td></td>
<td>36 (80 %)</td>
<td>22 (65 %)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>7 (21 %)</td>
<td>14 (30 %)</td>
<td></td>
<td>9 (20 %)</td>
<td>12 (35 %)</td>
<td></td>
</tr>
<tr>
<td>HAMD-6 Baseline</td>
<td>12 (11, 13)</td>
<td>12 (12, 13)</td>
<td>0.9</td>
<td>13 (12, 13)</td>
<td>12 (11.25, 13)</td>
<td>0.2</td>
</tr>
<tr>
<td>HAMD-6 Week 8</td>
<td>9 (7, 11)</td>
<td>3 (2, 5)</td>
<td>&lt;0.001</td>
<td>6 (10)</td>
<td>2 (1, 3)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Recurrence status</td>
<td></td>
<td></td>
<td>0.5</td>
<td></td>
<td></td>
<td>0.6</td>
</tr>
<tr>
<td>First-episode</td>
<td>12 (36 %)</td>
<td>20 (42 %)</td>
<td></td>
<td>17 (38 %)</td>
<td>15 (44 %)</td>
<td></td>
</tr>
<tr>
<td>Recurrent</td>
<td>21 (64 %)</td>
<td>26 (57 %)</td>
<td></td>
<td>28 (62 %)</td>
<td>19 (56 %)</td>
<td></td>
</tr>
</tbody>
</table>

1 Median (IQR); n (%)
2 Wilcoxon rank sum exact test; Pearson’s Chi-squared test; Wilcoxon rank sum test

(57.0 %) as non-responders. As expected, responder and remitter groups
overlapped: all remitters were also classified as responders, 12 (15.2 %)
responders were classified as non-responders, and the remaining non-
responders were labeled non-remitters. The number of participants in
each group is presented as a contingency table (Table 2) for the Neu-
roPharm dataset and in Table S2 for the EMBARC dataset.

Significant differences in age (W = 1143, p < 0.0001) and recurrence status ($\chi^2 = 12.409$, df = 1, p < 0.0001) were observed between the
NeuroPharm and EMBARC datasets, where as sex was not different ($\chi^2 = 0.395$, df = 1, p = 0.530).

### Table 2
Contingency table presenting the number of participants from the NeuroPharm
dataset in each of the treatment outcome groups.

<table>
<thead>
<tr>
<th></th>
<th>Non-remitter</th>
<th>Remitter</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-responder</td>
<td>33</td>
<td>0</td>
<td>33</td>
</tr>
<tr>
<td>Responder</td>
<td>12</td>
<td>34</td>
<td>46</td>
</tr>
<tr>
<td>Total</td>
<td>45</td>
<td>34</td>
<td>79</td>
</tr>
</tbody>
</table>
which clearly undermines their clinical value. Subsequent evaluations using VBM, an approach adopted by most earlier studies (Costafreda et al., 2009; Gong et al., 2011; Liu et al., 2012; Nouretdinov et al., 2011), also revealed no group differences in VBM volumes between the treatment outcomes and the models derived from these data achieved only limited predictive performances. As such, these results further perpetuate the current and unfortunate situation where strong predictive performance of small studies using machine learning cannot be replicated by studies using larger datasets (Flint et al., 2021). Indeed, it is worth emphasizing that the training (NeuroPharm) and test (EMBARC) dataset used here represent two of the largest studies with MRI investigating treatment outcome following an SSRI intervention. Even though the primary evaluation of the models was performed within a cross-validation framework, which is generally accepted as unbiased (Varoquaux et al., 2017), estimating the models’ performance using a single dataset (i.e., NeuroPharm dataset only) would have been misleading concerning their generalizability and clinical usefulness.

It is unclear whether the range of predictive performances observed across previous studies and the lack of generalizability from our models on the EMBARC dataset stem from different choices of drugs, intensity of clinical follow-up, study design, methodology, or phenotypic heterogeneity (i.e., MDD subtypes) which may not have been well captured in smaller cohorts. In fact, evidence of disease heterogeneity is readily apparent in both the NeuroPharm and EMBARC cohorts with 27 % and 21 % of the responders, respectively, showing response to the treatment, but not achieving remission, while the remaining responders also achieved remission. Interestingly, our treatment response model achieved numerically better performances compared to that of remission. Although many studies have proposed different definitions for subtypes of MDD (Musil et al., 2017; van Loo et al., 2012), there is currently no overwhelming evidence justifying their application in clinical diagnosis and, as of yet, their usage in conjunction with brain structural measures for the prediction of antidepressant treatment outcome remains to be investigated. Without more information it becomes difficult to identify the specific factors driving these discrepancies across studies and future research should concentrate on establishing model generalizability across different cohorts. One possible approach is to establish models that are invariant and robust to dataset shift (Quiñonero-Candela et al., 2008). However, obtaining the data necessary to train such models can only be accomplished through data sharing efforts by the research community. To this end, we make our models and code publicly available and access to the data can be requested (see the Data and code availability section).

When looking at model interpretability, there is little agreement across previous studies concerning which brain regions provide structural information predictive of treatment outcome following antidepressant intervention in MDD. Across the NeuroPharm and EMBARC datasets, only the volume of the right putamen in the EMBARC dataset was significantly different between responders and non-responders. The orbitofrontal cortex and the hippocampus are two brain regions thought to be primarily implicated in the drug treatment of MDD (Bartlett et al., 2018). Early pathological studies have demonstrated a decrease in cortical thickness, neuronal size, and neuronal and glial densities in the rostral orbitofrontal cortex of depressed individuals (Rajkowska et al.,

Table 3

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Response</th>
<th>Remission</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>Balanced Accuracy</td>
</tr>
<tr>
<td>Elastic Net</td>
<td>0.61 (0.04)</td>
<td>59.4 (2.5)</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.62 (0.03)</td>
<td>58.0 (3.2)</td>
</tr>
<tr>
<td>SVM</td>
<td>0.46 (0.04)</td>
<td>48.0 (3.5)</td>
</tr>
<tr>
<td>Boosted Trees</td>
<td>0.55 (0.04)</td>
<td>53.8 (3.7)</td>
</tr>
</tbody>
</table>

Fig. 2. Mean receiver operating characteristic (ROC) curves of the different classifiers for the classification of (A) response and (B) remission in the NeuroPharm dataset.

Table 3
Performance metrics of the different classifiers for predicting treatment response or remission in the NeuroPharm dataset. AUC: Area under the ROC Curve. Values are given as mean (SD), aside from the p-values.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Response</th>
<th>Remission</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>Balanced Accuracy</td>
</tr>
<tr>
<td>Elastic Net</td>
<td>0.61 (0.04)</td>
<td>59.4 (2.5)</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.62 (0.03)</td>
<td>58.0 (3.2)</td>
</tr>
<tr>
<td>SVM</td>
<td>0.46 (0.04)</td>
<td>48.0 (3.5)</td>
</tr>
<tr>
<td>Boosted Trees</td>
<td>0.55 (0.04)</td>
<td>53.8 (3.7)</td>
</tr>
</tbody>
</table>

Fig. 2. Mean receiver operating characteristic (ROC) curves of the different classifiers for the classification of (A) response and (B) remission in the NeuroPharm dataset.

which clearly undermines their clinical value. Subsequent evaluations using VBM, an approach adopted by most earlier studies (Costafreda et al., 2009; Gong et al., 2011; Liu et al., 2012; Nouretdinov et al., 2011), also revealed no group differences in VBM volumes between the treatment outcomes and the models derived from these data achieved only limited predictive performances. As such, these results further perpetuate the current and unfortunate situation where strong predictive performance of small studies using machine learning cannot be replicated by studies using larger datasets (Flint et al., 2021). Indeed, it is worth emphasizing that the training (NeuroPharm) and test (EMBARC) dataset used here represent two of the largest studies with MRI investigating treatment outcome following an SSRI intervention. Even though the primary evaluation of the models was performed within a cross-validation framework, which is generally accepted as unbiased (Varoquaux et al., 2017), estimating the models’ performance using a single dataset (i.e., NeuroPharm dataset only) would have been misleading concerning their generalizability and clinical usefulness.

It is unclear whether the range of predictive performances observed across previous studies and the lack of generalizability from our models on the EMBARC dataset stem from different choices of drugs, intensity of clinical follow-up, study design, methodology, or phenotypic heterogeneity (i.e., MDD subtypes) which may not have been well captured in smaller cohorts. In fact, evidence of disease heterogeneity is readily apparent in both the NeuroPharm and EMBARC cohorts with 27 % and 21 % of the responders, respectively, showing response to the treatment, but not achieving remission, while the remaining responders also achieved remission. Interestingly, our treatment response model achieved numerically better performances compared to that of remission. Although many studies have proposed different definitions for subtypes of MDD (Musil et al., 2017; van Loo et al., 2012), there is currently no overwhelming evidence justifying their application in clinical diagnosis and, as of yet, their usage in conjunction with brain structural measures for the prediction of antidepressant treatment outcome remains to be investigated. Without more information it becomes difficult to identify the specific factors driving these discrepancies across studies and future research should concentrate on establishing model generalizability across different cohorts. One possible approach is to establish models that are invariant and robust to dataset shift (Quiñonero-Candela et al., 2008). However, obtaining the data necessary to train such models can only be accomplished through data sharing efforts by the research community. To this end, we make our models and code publicly available and access to the data can be requested (see the Data and code availability section).

When looking at model interpretability, there is little agreement across previous studies concerning which brain regions provide structural information predictive of treatment outcome following antidepressant intervention in MDD. Across the NeuroPharm and EMBARC datasets, only the volume of the right putamen in the EMBARC dataset was significantly different between responders and non-responders. The orbitofrontal cortex and the hippocampus are two brain regions thought to be primarily implicated in the drug treatment of MDD (Bartlett et al., 2018). Early pathological studies have demonstrated a decrease in cortical thickness, neuronal size, and neuronal and glial densities in the rostral orbitofrontal cortex of depressed individuals (Rajkowska et al.,
therefore closer to clinical application outside of the research setting.

Secondly, some individuals from the NeuroPharm dataset were excluded to be able to draw conclusions centered on the predictive capabilities of structural MRI and basic demographics. Other studies utilizing the NeuroPharm cohort have investigated prediction of treatment response by reward processing using functional MRI (Brandt et al., 2021) and alpha asymmetry from qEEG (Ip et al., 2021), but with limited success. Future work should concentrate on combining the information of multiple domains to improve the prediction of outcome following antidepressant treatment in MDD.

A few limitations for this work have to be acknowledged. Firstly, some individuals (n = 10) included in the NeuroPharm dataset switched their SSRI treatment to the SNRI duloxetine during the course of the treatment. Although this deviates from a perspective focused solely on the treatment of MDD with a unique drug, it does reflect the naturalistic clinical course of a drug intervention in the treatment of MDD and is therefore closer to clinical application outside of the research setting. Secondly, some individuals from the NeuroPharm dataset were excluded as they did not follow the complete treatment protocol up to week 8 due to reasons which could potentially be linked to the treatment, i.e., acute suicidal, excessive anxiety, adverse side-effect to SNRI, or spontaneous remission. These individuals should not be included in the context of a randomized controlled trial, nonetheless, they were not included here so that the models trained using NeuroPharm data could be applied to the EMBARC dataset where this information is not available, and to make our results relatable to previous studies where this exclusion is also performed. Finally, it is worth noting that there are important intrinsic differences between the NeuroPharm and the EMBARC datasets, with some of the most striking being a difference in mean age of 14 years between the two datasets and the usage of a different SSRI drugs (i.e., some of the most striking being a difference in mean age of 14 years between the two datasets and the usage of a different SSRI drugs (i.e., some the EMBARC dataset where this information is not available, and to make our results relatable to previous studies where this exclusion is also performed. Finally, it is worth noting that there are important intrinsic differences between the NeuroPharm and the EMBARC datasets, with some of the most striking being a difference in mean age of 14 years between the two datasets and the usage of a different SSRI drugs (i.e., escitalopram and sertraline) which may not have identical treatment effects.

5. Conclusion

The usefulness of pre-treatment structural brain MRI in predicting the outcome of antidepressant treatment in MDD is not convincing. Our work utilizing the NeuroPharm cohort indicates that treatment response can be predicted above chance in-sample, but these results did not generalize to an independent test dataset. Future work should concentrate on improving the generalizability of the models across cohorts and combining features from structural brain MRI with information from other domains.

Funding and disclosure

Funding for this study was provided by the Innovation Fund Denmark (grant 4108-00004B, NeuroPharm), and the Lundbeck Foundation (grant R279-2018-1145, BrainDrugs); the Innovation Fund Denmark and the Lundbeck Foundation had no further role in the study design, collection, analysis and interpretation of data, the writing of the report and the decision to submit the paper for publication. GMK has received honoraria as a speaker for Sage Biogen and as a consultant for Sanos. All remaining authors declare that they have no potential conflicts of interest.

CRediT authorship contribution statement

Vincent Beliveau: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Visualization. Elia Hedebøe: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Visualization. Patrick M. Fisher: Conceptualization, Writing – original draft. Vibeke H. Dam: Conceptualization, Writing – original draft. Martin B. Jørgensen: Conceptualization, Writing – original draft. Vibe G. Frokjaer: Conceptualization, Writing – original draft. Gitte M. Knudsen: Conceptualization, Writing – original draft, Funding acquisition. Melanie Ganz: Conceptualization, Writing – original draft, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimagec.2022.103224.

References


