NeuroImage: Clinical 35 (2022) 103120

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

Predicting escitalopram treatment response from pre-treatment and early response resting state fMRI in a multi-site sample: A CAN-BIND-1 report

Jacqueline K. Harris [a,b,*], Stefanie Hassel [c,d], Andrew D. Davis [e], Mojdeh Zamyadi [e,f], Stephen R. Arnott [e], Roumen Milev [g], Raymond W. Lam [h], Benicio N. Frey [i,j], Geoffrey B. Hall [k], Daniel J. Müller [l,m,n,o], Susan Rotzinger [l,p], Sidney H. Kennedy [l,p], Stephen C. Strother [e,f], Glenda M. MacQueen [c,d], Russell Greiner [a,b,q]

[a] Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada
[b] Alberta Machine Intelligence Institute, Edmonton, Alberta, Canada
[c] Department of Psychiatry, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada
[d] Hotchkiss Brain Institute, Mathison Centre for Mental Health Research and Education, University of Calgary, Calgary, Alberta, Canada
[e] Rotman Research Institute, Baycrest Health Sciences, Toronto, Ontario, Canada
[f] Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada
[g] Departments of Psychiatry and Psychology, Queen's University, and Providence Care, Kingston, Ontario, Canada
[h] Department of Psychiatry, University of British Columbia, Vancouver, British Columbia, Canada
[i] Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, Ontario, Canada
[j] Mood Disorders Program and Women's Health Concerns Clinic, St. Joseph's Healthcare, Hamilton, Ontario, Canada
[k] Department of Psychology, Neuroscience & Behaviour, McMaster University, and St. Joseph's Healthcare Hamilton, Hamilton, Ontario, Canada
[l] Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada
[m] Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Ontario, Canada
[n] Department of Pharmacology & Toxicology, University of Toronto, Toronto, Ontario, Canada
[o] Institute of Medical Sciences, University of Toronto, Toronto, Ontario, Canada
[p] Centre for Depression and Suicide Studies, St. Michael's Hospital, Toronto, Ontario, Canada
[q] Department of Psychiatry, University of Alberta, Edmonton, Alberta, Canada

ARTICLE INFO

Keywords:
Resting State
fMRI
Machine learning
Treatment response
Depression
Functional connectivity

ABSTRACT

Many previous intervention studies have used functional magnetic resonance imaging (fMRI) data to predict the antidepressant response of patients with major depressive disorder (MDD); however, practical constraints have limited many of those attempts to small, single centre studies which may not adequately reflect how these models will generalize when used in clinical practice. Not only does the act of collecting data at multiple sites generally increase sample sizes (a critical point in machine learning development) it also generates a more heterogeneous dataset due to systematic differences in scanners at different sites, and geographical differences in patient populations. As part of the Canadian Biomarker Integration Network in Depression (CAN-BIND-1) study, 144 MDD patients from six sites underwent resting state fMRI prior to starting escitalopram treatment, and again two weeks after the start. Here, we consider ways to use machine learning techniques to produce models that can predict response (measured at eight weeks after initiation), based on various parcellations, functional connectivity (FC) metrics, dimensionality reduction algorithms, and base learners, and also whether to use scans from one or both time points. Models that use only baseline (pre-treatment) or only week 2 (early-response) whole-brain FC features consistently failed to perform significantly better than default models. Utilizing the change in FC between these two time points, however, yielded significant results, with the best performing analytical pipeline achieving 69.6% (SD 10.8) accuracy. These results appear contrary to findings from many smaller single-site studies, which report substantially higher predictive accuracies from models trained on only baseline resting state FC features, suggesting these models may not generalize well beyond data used for development. Further, these results indicate the potential value of collecting data both before and shortly after treatment initiation.

* Corresponding author at: University of Alberta, Department of Computing Science, Edmonton, Alberta, Canada.
E-mail address: jkh@ualberta.ca (J.K. Harris).

https://doi.org/10.1016/j.nicl.2022.103120
Received 2 March 2022; Received in revised form 17 May 2022; Accepted 14 July 2022
Available online 16 July 2022

1. Introduction

Major depressive disorder (MDD) is a prevalent, and potentially chronic disorder with a highly variable symptom profile (Fried and Nesse, 2015). While there are many effective treatments, including multiple classes of pharmacological interventions, response to any given treatment at the individual level is quite variable. Approximately one third of patients remit from depressive symptoms following their first antidepressant treatment, and many will require multiple trials of different medications or other treatments to reach remission (Rush et al., 2006). While remission after multiple treatments may be possible, many patients drop out before an adequate response is found (Thornicroft et al., 2017), or experience prolonged distress and frustration while undergoing various (often unsuccessful) treatments. This has motivated the development of tools that can guide clinicians to quickly and accurately inform the treatment choice that is best for each individual MDD patient.

The heterogeneity of MDD suggests that response to treatment may be complex (Fried and Nesse, 2015). Certain clinical factors appear to correlate with response, such as duration, and severity of the disorder (Kraus et al., 2019), suggesting the existence of several depressive subtypes. In general, however, attempts to find stable clinical subgroups with associations that are robust enough to guide clinical decision making have not been successful (Arnow et al., 2015). Machine learning provides a promising alternative to these approaches, as it can produce models that can map complex interactions among input features to outcome measures, enabling accurate, individual predictions of response.

Resting state (RS) functional connectivity (FC) has frequently been used to study neurophysiological alterations in MDD. This type of analysis has been used to reproduce positron emission tomography (PET) findings implicating corticolimbic dysregulation in MDD etiology (Anand et al., 2005; Mayberg, 1997); with subsequent studies consistently reporting abnormal connectivity between the prefrontal cortex (PFC), anterior cingulate cortex (ACC), and limbic regions (Cullen et al., 2009; Veer et al., 2010). Findings have also pointed to alterations in whole RS networks such as hyperconnectivity in the default mode network (DMN) (Greicius et al., 2007; Sheline et al., 2009). Changes in FC following antidepressant treatment have also been reported including normalization of DMN hyperconnectivity (Posner et al., 2013; Wang et al., 2015), modulation of striatal connectivity (Wang et al., 2019), and increased corticolimbic connectivity (Anand et al., 2005).

The sensitivity of RS FC measures to detect differences between MDD and healthy control groups as well as changes with antidepressant treatment, make them an attractive measurement for models predicting individual response to treatment, with several studies reporting accuracies exceeding 80% (Cohen et al., 2021; Gao et al., 2018). Most of these studies, however, use relatively small, single-site datasets, which tend to be more homogeneous, making it difficult to know if these results would translate to clinical practice (Schnack and Kahn, 2016). The magnitude of such site-induced biases in FC measures has also been shown to exceed the effect size of many psychiatric disorders (including MDD), making multi-site data collection essential for modelling the overall population (Yamashita et al., 2019).

Empirical results from predictive studies further support these claims. Sundermann et al. (2017), for example, found that models derived from a larger, more heterogeneous sample – presumably more representative of the true target patient population – predicted MDD diagnosis with lower accuracy (45.0 to 56.1% accuracy) than models that had typically been reported in the literature (based on smaller sample sizes). Ramasubbu et al. (2016) similarly showed that MDD diagnostic models only outperformed default accuracy after limiting their patient population to those with severe depressive symptoms. All of these observations have been recently reinforced by the *meta*-analysis of Sajjadian et al. (2021). They report a mean accuracy of 63% [95% confidence interval (CI) 56–71] from 8 of 54 eligible literature reports before November 2020 considered to be of high-enough quality to enter into the *meta*-analysis. The remaining 46 reported studies provided a significantly higher mean accuracy of 75% (95% CI 72–78).

Variability in data-processing pipelines can also make it challenging to compare results from different studies. In addition to the pre-processing required in all fMRI analysis, studies utilizing FC must also choose definitions for regions of interest and the metric used to calculate FC, both of which have been shown to have substantial impact on overall model performance (Abraham et al., 2017; Dadi et al., 2019; Kalmady et al., 2019). We must also consider other modelling choices, such as methods to reduce the number of features (if at all), what type of base learner will be used, and with what parameters. Published results may also be artificially inflated if various pipeline choices are tested and chosen only after reviewing performance on the test set (Hosseini et al., 2020).

Here we investigate the utility of machine learning models that use whole-brain RS-fMRI features to predict escitalopram treatment response in a relatively large, multi-site MDD sample. We assess both the utility of pre-treatment FC measures and consider the utility of also using data collected after two weeks of treatment, which may serve as an early indicator of response. Further, we consider the impact of altering various steps in the analysis pipeline to assess what effect they may have on prediction accuracy.

2. Methods

2.1. Participants

We used data from CAN-BIND-1, a multi-centre study developed to examine potential biomarkers of response to antidepressant therapies in MDD; see Lam et al. (2016) for a detailed overview of the full study protocol and Kennedy et al. (2019) for a full report on clinical outcomes. Of relevance to this analysis, patients were recruited at six sites across Canada: University of British Columbia (UBC), University of Calgary (UCA), Queen's University at Kingston (QNS), McMaster University (MCU), Toronto General Hospital (TGH), and The Centre for Addiction and Mental Health, Toronto (CAMH). Informed consent was obtained from all participants, and the protocol was approved by the Research Ethics Boards at each institution, with additional approval from the University of Alberta Research Ethics Board for secondary data analysis.

Eligible participants were outpatients between the ages of 18 and 60, fluent English speakers, meeting DSM-IV-TR criteria for MDD in a current Major Depressive Episode, as assessed by the Mini International Neuropsychiatric Interview (MINI; Sheehan et al., 1998), with episode duration of at least three months, and a score of at least 24 on the Montgomery-Asberg Depression Rating Scale (MADRS; Montgomery and Åsberg, 1979). Participants were also required to be free of psychotropic medications for at least five half-lives prior to initial visit and to have not started psychological treatment within the previous three months.

Individuals were ineligible if they had a previous unsuccessful, or not tolerated, trial of either of the study medications (escitalopram or aripiprazole), had failed four or more previous pharmacological treatments, or were at high risk for hypomanic switch (history of antidepressant-induced hypomania). Participants were also screened for other psychiatric disorders and excluded if they had a diagnosis of Bipolar I or II, or any primary psychiatric diagnosis other than MDD. Individuals with high suicidal risk, substance abuse/dependence in the past six months, psychosis in the current depressive episode, significant neurological disorders, head trauma, unstable medical conditions, pregnant or breastfeeding, or contraindications to MRI were also excluded.

Participants were included in current analyses if they were part of the treatment group and had complete pre-treatment (baseline) and week 2 RS-fMRI data, as well as treatment response data at week 8. Of the 157 participants with complete data, 13 were excluded during

manual quality control for severe imaging artefacts or incidental MRI findings. Scans were also screened for excessive motion based on the 'lenient' criteria outlined in Parkes et al. (2018), and none exceeded the threshold of 0.55 mm for the temporal mean of the frame displacement time series. Hence, data from 144 participants (age 34.9 ± 12.43, 90 females; see Supplementary Material Table S1 for detailed patient clinical and demographic information) were included in model building and assessment.

2.2. Treatment protocol and outcomes

Once enrolled, patients underwent extensive clinical data collection, after which, they began treatment with escitalopram for 8 weeks. MRI data, including RS-fMRI, were collected at baseline, two weeks, and eight weeks after treatment initiation. After eight weeks, patients were assessed for response to escitalopram, as defined as ≥ 50% reduction in total MADRS score from baseline. Those who achieved ≥ 50% decrease in MADRS were classified as 'responders', and those who did not, as 'non-responders'.

2.3. MR image acquisition

Imaging data were acquired using 3 Tesla MRI scanners at six centers across Canada, varying by model and manufacturer (GE Healthcare Signa HDxT (TGH), GE Healthcare Discovery MR750 (CAM, MCU, UCA), Phillips Intera (UBC), Siemens Trio Tim (QNS)). Imaging protocols

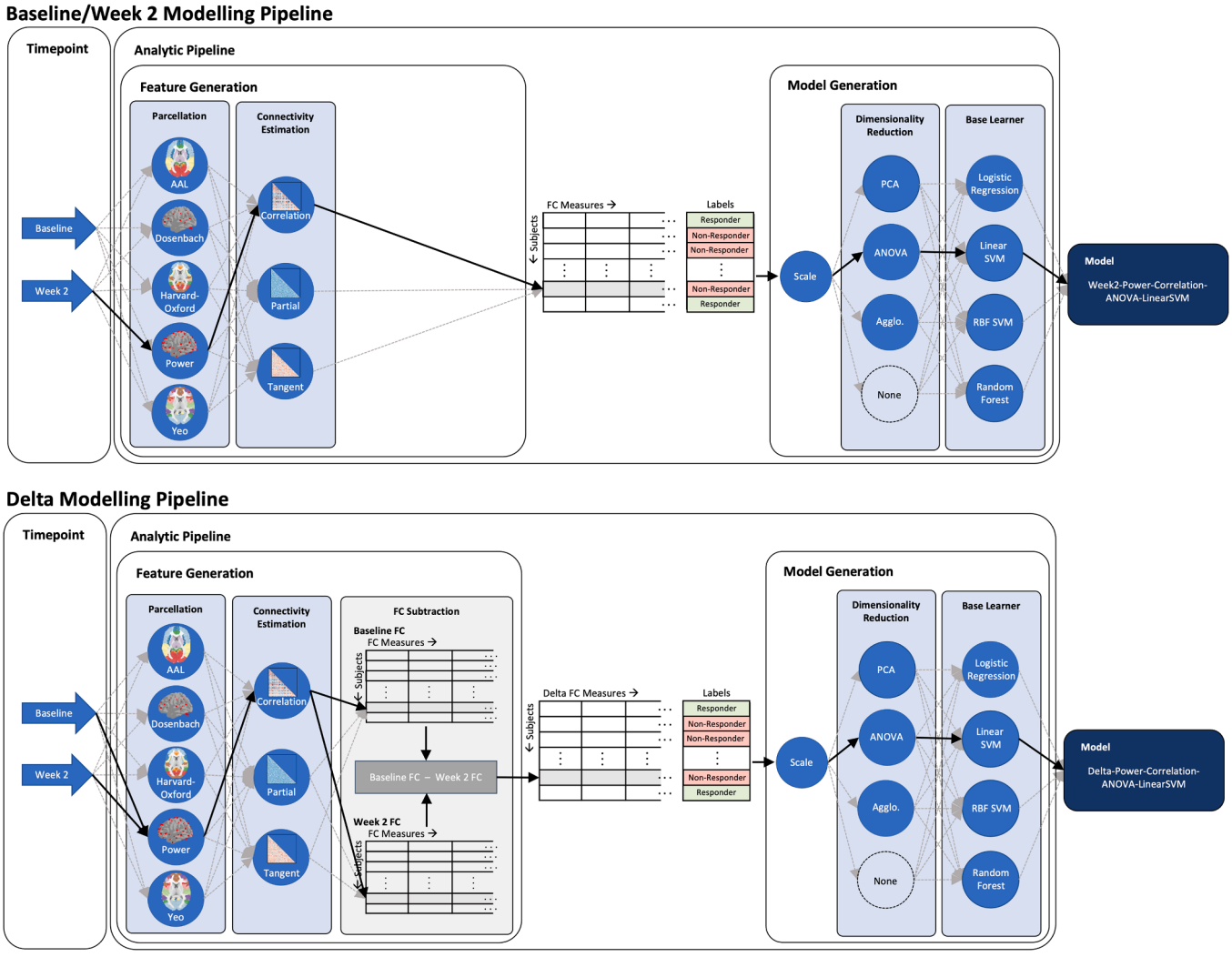


Fig. 1. (Top) Diagram of data flow through analytic pipelines. Pre-processed resting state fMRI data from either baseline or week 2 is fed through a series of operations that first generate functional connectivity (FC) features and then generate a predictive model based on these features. Alternate approaches are used for parcellation, connectivity estimation, dimensionality reduction, and base learner, every combination of which is tested for each set of input data, resulting in a total of 240 models for each of the baseline and week 2 datasets (5 parcellations × 3 connectivity metrics × 4 dimensionality reduction techniques × 4 classifiers) – leading to a total of 480 models. Along the highlighted pathway, for example, pre-processed data collected at week 2 is first parcellated using the Power coordinates, resulting in a 259x295 matrix for each participant's data, where 259 is the number of regions of interest (ROI) included in the power coordinates after SNR masking, and 295 is the length of the temporal dimension of the original fMRI dataset. Between every pair of ROIs, the correlation between time-courses is then calculated, resulting in a single value for every ROI pair. The full set of correlations corresponds to the lower triangle of the full correlation matrix, which is then vectorized with a length of 33,411 for each participant. This process of feature generation is repeated for each participant, resulting in a 144x33,411 matrix of FC features. Since feature generation is independent of response label, this procedure is completed prior to model generation. Features then fed into model generation are scaled, and passed to ANOVA feature selection, which chooses the k (value obtained based on internal cross-validation) most relevant features to be used in the linear SVM classifier. **(Bottom)** Delta models are processed through the same analytic pathway, with the addition of a subtraction step at the end of feature generation. Here, both baseline and week 2 FC features are generated, and the difference of the two matrices (the delta FC feature matrix) is used in subsequent predictive modelling. An additional 240 models are generated using these delta features, considering all possible combinations of processing steps.

varied slightly among sites, to accommodate for scanner and manufacturer differences (see MacQueen et al., 2019 for detailed protocols). RS-fMRI was collected over a 10-minute scan (2000 ms repetition time) using a whole-brain T_2*-sensitive blood-oxygen-level-dependent echo planar imaging sequence (30 ms echo time, 4 mm × 4 mm × 4 mm resolution). Accompanying whole-brain structural 3D T_1-weighted images were acquired with a 1 mm isotropic resolution. In addition to protocol harmonization, substantial quality control and assurance efforts were implemented to ensure consistent high-quality data was collected from all sites. These methods included, but were not limited to, automated file name and imaging protocol adherence checks, manual image quality control rating, and longitudinal monitoring of scanner stability using monthly phantom scans from all sites (see MacQueen et al. 2019). Details of image pre-processing are included in the Supplementary Materials.

2.4. Data analysis

FC was extracted as a connectome, measuring the pair-wise synchronicity between the time-series of each pair of spatial ROIs; where the time-series for a single ROI is the average temporally changing signal intensity of all voxels included in the ROI. Many studies have noted that in both the calculation and the use of these connectomes in machine learning models, various steps in the pipeline can have a substantial impact on overall model performance (Abraham et al., 2017). To ensure that our reported results are not specific to a single pipeline, we tested multiple pipelines, altering decisions made at various steps, as depicted in Fig. 1. In relation to the computation of the connectome, we considered different combinations of parcellations and connectivity estimation metrics. We also considered various methods for feature dimensionality reduction, and various types of base learners.

Demonstration of the effect of acquiring imaging data at different sites on these FC measures can be found in the Supplementary Materials.

2.5. Parcellation

To reduce dimensionality, and obtain neurobiologically relevant features, pre-processed RS images were parcellated based on either an *a priori* atlas or set of pre-defined ROI coordinates. A time-series for each region from either atlas or coordinate set was generated by taking an element-wise average of all voxels included in the region. Methods for defining these regions, and also the number of regions in each parcellation, vary greatly depending on the parcellation scheme. We considered the five schemes listed below. Parcellations were all implemented using *Nilearn* (Abraham et al., 2014); with default parameters unless otherwise stated.

Automated Anatomical Labelling (AAL) (Tzourio-Mazoyer et al., 2002) – 116 region anatomical atlas based on manual segmentation.

Dosenbach (Dosenbach et al., 2010) – 160 ROIs based on *meta*-analyses of task-based fMRI (radius = 4.5).

Harvard-Oxford (Desikan et al., 2006) – 48 region cortical atlas based semi-automated segmentation (atlas_name = 'cort-maxprob-thr25-2 mm').

Power (Power et al., 2011) – 264 ROIs generated based on functional homogeneity (radius = 5.0).

Yeo (Thomas Yeo et al., 2011) – 17 cortical networks based on clustering of functional connectivity in 1000 participants (data = 'thick_17').

To ensure sufficient signal quality, we generated a binary mask that set a voxel position to zero, for exclusion, if the signal-to-noise ratio (SNR, time-series mean divided by standard deviation) for that position was less than 100 in greater than 5% of participants (Drysdale et al., 2017). Voxels with insufficient SNR were excluded during mean time series calculation for parcellations. Based on this low SNR criteria, we excluded (all voxels in) five regions from the Power parcellation and one from the Dosenbach parcellation, generally along the inferior frontal and temporal lobes (Supplementary Fig. S1).

2.6. Connectivity estimation

For each parcellation, we then considered the following three measures to estimate pairwise FC between each pair of brain regions; again, calculated using *Nilearn* (Abraham et al., 2014).

Correlation

Partial Correlation – A variant of correlation that infers only direct connectivity between regions, as opposed to including indirect connectivity through other regions.

Tangent – optimized covariance metric for statistical learning (Abraham et al., 2017; Dadi et al., 2019).

As part of the training procedure, FC features were scaled to zero mean, and unit variance with respect to the entire training set. This, and subsequent steps, including dimensionality reduction techniques, base learners, cross-validation, and hyperparameter optimization, were implemented using *Scikit-learn* (Pedregosa et al., 2011).

2.7. Dimensionality reduction

Various techniques were applied to further reduce the number of input features. The number of features selected, k, was optimized during internal cross-validation to be between 5 and 50; the upper limit selected in consideration of the sample size. In addition to the strategies listed below, models were also tested with this step omitted, to allow the base learners to use the full set of FC features.

Principal components analysis (PCA) – Project data into a lower dimensional space; here keeping the first k components.

ANOVA – Univariate Feature selection based on ANOVA f-value between the input features and labels, keeping features with k highest f-values.

Agglomeration – Recursively merge features, stopping when k clusters are reached.

None – Omit dimensionality reduction step and allow base learner to use all features.

2.8. Base learner

We then ran base learners on the data described above, to produce models that could discriminate between treatment responders and non-responders. We considered four different base learners, both linear and nonlinear, to assess their discriminative capacity. Here, we considered support vector machine (SVM) learners as they have been shown to be highly effective for classification and are one of the most commonly used. We also considered logistic regression for its simplicity and fast training time, and random forest as a more complex, and often effective, model.

Logistic Regression – Basic linear classification model.

Support vector Machine (SVM) with linear kernel.

SVM with Radial Basis Function (RBF) Kernel – Non-linear SVM.

Random Forest – Ensemble bagging algorithm based on decision trees.

2.9. Timepoints

We trained each combination of parcellation, connectivity estimation, dimensionality reduction, and base learner described above with each of the three sets of input data based on the time(s) when the data was collected: (1) just baseline data, which was collected prior to treatment, (2) data collected 2 weeks into treatment, and (3) a combination of the two time points. For (3), we computed the features for a given parcellation and connectivity metric for both baseline and week 2, then subtracted the week 2 features from the baseline features; going forward, we refer to such models as "delta" models. We compared the performance of models trained using data from these three different

timepoints using Wilcoxon matched-pairs signed rank test, matching results from the same analytical pipelines trained using different data, and corrected for multiple comparisons using Bonferroni correction.

Altogether, for each of the three timepoints, we consider 240 = 5x3x4x4 different models, as we consider five different parcellations (AAL, Dosenbach, Harvard-Oxford, Power, and Yeo), three different connectivity measures (correlation, partial correlation, and tangent embedding) four different dimensionality reduction strategies (PCA, ANOVA, agglomeration, and none), and four base learners (logistic regression, linear SVM, SVM with RBF kernel, and random forest). For each timepoint, and each setting (particular parcellation, connectivity measure, dimensionality reduction strategy, and base learner), we used internal 5-fold cross-validation to identify the best hyperparameter values, using a random search of 100 iterations over the prescribed parameter space for both dimensionality reduction and base learner (details of hyperparameter optimization included in Supplementary Materials).

2.10. Cross-validation and assessment

The quality of the learned models was assessed using an external 10-fold cross-validation strategy. Here, we partitioned the data into 10 disjoint equal-sized folds, each with roughly the same number of instances from each scan site and each classification label. Since all the parcellations used in this analysis were derived *a priori,* the parcellation and connectome calculation were performed outside of cross-validation.

This approach produces a model for each fold (9/10 of the data), whose accuracy is assessed on the held-out remaining 1/10 of the data. The overall performance reported for each model is the mean accuracy across all folds. A default 'dummy' classifier was also generated to assign each patient to the majority class label from the training set. Each model for a given timepoint was compared to the default model to determine if it could significantly outperform chance level predictions. Models from each timepoint were first compared using an ANOVA analysis, with subsequent paired t-tests if significance was reached. During external cross-validation, the resubstitution accuracy was also assessed by making predictions on the training data to be used as an indicator of overfitting.

The impact of changes to the analytical pipeline was statistically tested using the Wilcoxon matched-pairs rank test corrected with Bonferroni correction, matching pipelines differing in only a single step.

3. Results

The default model, predicting the majority class, had a prediction accuracy of 53.5% (SD 3.7). Fig. 2 shows the cross-validation accuracy results for the top 10 models from each timepoint, along with details of corresponding analytic pipelines.

3.1. Baseline models

Models trained on baseline data had accuracies ranging from 39.0 (SD 11.7) to 61.2% (SD 10.5), none of which performed significantly differently from the default classifier (p-value $>$ 0.05). Only three models exceeded 60% mean accuracy.

3.2. Week 2 models

Models trained using week 2 data showed performance similar to baseline models, with mean accuracies between 37.5 (SD 8.8) to 66.5% (SD 12.0), only the highest of which performed significantly better than the default classifier ($p = 0.014$). The best performing model was also the only model to exceed 60% mean accuracy, having over 6% higher accuracy than the next best performing model. Corrected Wilcoxon rank test showed no significant difference between models from baseline and week 2 ($p = 1.0$).

3.3. Delta models

40 models of the 240 modelling pipelines tested exceeded a mean accuracy of 60%. Thirty of these 40 models came from pipelines utilizing correlation as the metric to calculate connectome matrices. Mean test accuracies (over all 240 models) ranged from 41.6 (10.1) to 69.6% (SD 10.8), with 17 models performing significantly better than the default model at a p-value of < 0.05. Statistical comparison with models trained with baseline, or week 2 data, show highly significant differences ($p = 1.6e-11$, $p = 2.4e-13$, respectively) although only an overall modest mean performance improvement of 3.1 and 3.4% respectively.

3.4. Pipeline choices

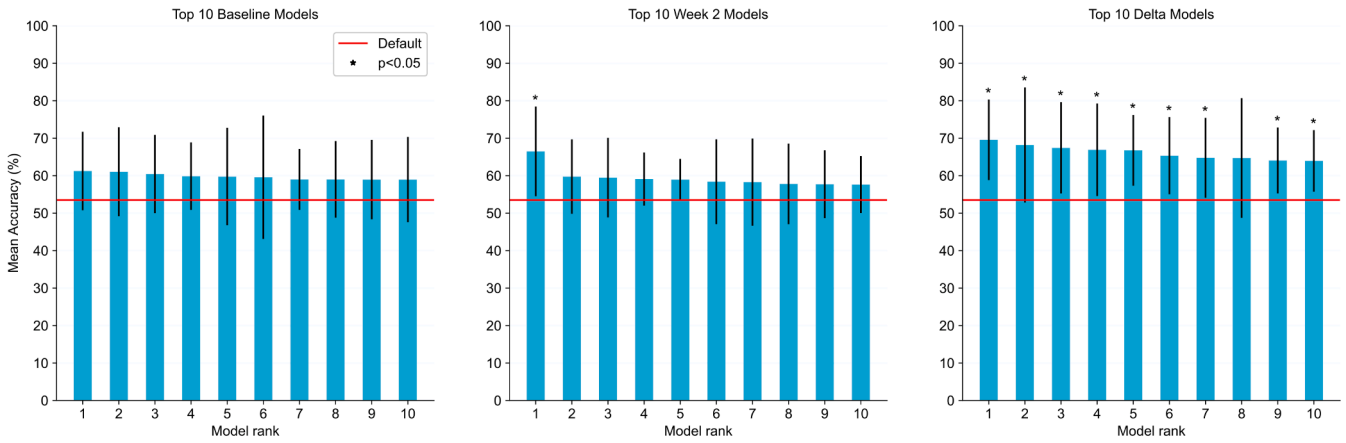
We assessed the impact of different choices in the analytic pipeline only in the delta models (Fig. 3), as neither baseline nor week 2 models indicated predictive capacity beyond chance level. Connectivity metric showed the greatest impact on predictive accuracy, with partial correlation models performing significantly worse than both correlation and tangent ($p = 3.6e-12$, $p = 3.0e-5$, respectively). Models trained with correlation data also significantly outperformed tangent models ($p = 3.0e-7$).

Parcellation choice had a small impact on accuracy but was significant between AAL and Yeo ($p = 0.008$), Dosenbach and Yeo ($p = 0.003$), and Power and Yeo ($p = 0.016$) parcellations. In general, parcellations with fewer regions tended to outperform those with more. Neither base learner nor dimensionality reduction technique had a significant impact on prediction accuracy, although the random forest base learner and ANOVA-based dimensionality reduction showed moderate improvement over other techniques.

4. Discussion

Early symptom response to antidepressant medication has been well established to be one of the strongest predictors of treatment response to date (Jakubovski and Bloch, 2014; Szegedi et al., 2009). However, it is not known if neurobiological measures carry this same predictive power. Our analysis, utilizing a relatively large heterogeneous sample, shows that FC features collected from a single time-point, either prior to treatment initiation, or 2 weeks into treatment, were not capable of predicting response significantly above chance level, except for a single anomalous case amongst the week 2 models. In the same sample, however, the change in FC matrices between baseline and week 2 predicting response after eight weeks of treatment, was significantly above chance, with the highest performing model achieving 69.6% (SD 10.8) accuracy.

Previous authors utilizing RS-fMRI data to predict individual response to pharmacological treatment have reported predictive accuracies in excess of 80% (Cohen et al., 2021; Gao et al., 2018); although these rarely included imaging data from more than one site, or samples with more than 50 participants. One notable result from the iSPOT trials, utilizing 80 participants, reported prediction accuracy for remission exceeding 80%, based on models trained from intrinsic FC measures between the posterior cingulate cortex (PCC) and anterior cingulate cortex (ACC)/medial prefrontal cortex (mPFC) (Goldstein-Piekarski et al., 2018); regions that have been implicated in depressive symptomology (Cullen et al., 2009; Veer et al., 2010). Problematically, the mPFC, ACC, and other limbic regions lie in the inferior frontal and temporal lobes, which are known to experience substantial magnetic field inhomogeneities due to numerous air/tissue interfaces in the region (Truong et al., 2002), resulting in susceptibility artefacts and signal loss. SNR masks produced as part of this analysis (Supplementary Fig. S1) to ensure adequate signal coverage in regions included in model development excluded significant portions of regions of interest including subgenual ACC and raphe nuclei, which are important components of emotional regulation and serotonergic pathways. Poor signal



Top 10 Baseline Models

	Parcellation	Connectivity Metric	Dimensionality Reduction	Classifier	Mean Accuracy (%)	Standard Deviation
1	Power	Tangent	ANOVA	Linear SVM	61.2	10.5
2	Harvard-Oxford	Correlation	Agglomeration	Linear SVM	61.0	11.9
3	Harvard-Oxford	Correlation	PCA	Logistic Regression	60.4	10.5
4	AAL	Correlation	PCA	Logistic Regression	59.9	9.0
5	Harvard-Oxford	Partial Correlation	PCA	Linear SVM	59.8	13.0
6	Harvard-Oxford	Correlation	None	Linear SVM	59.6	16.5
7	AAL	Partial Correlation	PCA	Random Forest	59.0	8.1
8	Yeo	Correlation	None	Linear SVM	59.0	10.2
9	Harvard-Oxford	Partial Correlation	PCA	Random Forest	59.0	10.6
10	Harvard-Oxford	Correlation	PCA	Linear SVM	59.0	11.4

Top 10 Week 2 Models

	Parcellation	Connectivity Metric	Dimensionality Reduction	Classifier	Mean Accuracy (%)	Standard Deviation
1	AAL	Tangent	None	Random Forest	66.5	12.0
2	Dosenbach	Tangent	ANOVA	Random Forest	59.8	9.9
3	AAL	Partial Correlation	None	Random Forest	59.5	10.6
4	Dosenbach	Tangent	ANOVA	Linear SVM	59.1	7.1
5	Harvard-Oxford	Tangent	Agglomeration	SVM with RBF Kernel	59.0	5.5
6	Yeo	Correlation	Agglomeration	SVM with RBF Kernel	58.4	11.3
7	Doesnbach	Tangent	None	Random Forest	58.3	11.6
8	Dosenbach	Partial Correlation	None	Random Forest	57.8	10.8
9	Dosenbach	Partial Correlation	ANOVA	SVM with RBF Kernel	57.7	9.0
10	Harvard-Oxford	Tangent	Agglomeration	Linear SVM	57.6	7.6

Top 10 Delta Models

	Parcellation	Connectivity Metric	Dimensionality Reduction	Classifier	Mean Accuracy	Standard Deviation
1	Yeo	Correlation	ANOVA	Random Forest	69.6	10.8
2	Yeo	Correlation	ANOVA	SVM with RBF Kernel	68.2	15.4
3	Yeo	Correlation	None	Random Forest	67.4	12.2
4	Harvard-Oxford	Tangent	ANOVA	Random Forest	66.9	12.3
5	Harvard-Oxford	Tangent	None	Random Forest	66.8	9.4
6	AAL	Correlation	None	Random Forest	65.3	10.3
7	Yeo	Correlation	ANOVA	Logistic Regression	64.8	10.7
8	AAL	Correlation	None	Logistic Regression	64.7	16.0
9	Dosenbach	Correlation	ANOVA	Logistic Regression	64.0	8.8
10	Harvard-Oxford	Partial Correlation	ANOVA	Linear SVM	64.0	8.2

Fig. 2. Ranked models with highest mean cross-validation accuracy for data from baseline (left), week 2 (center), and delta (right) timepoints. Bar charts depict the mean test accuracy for each pipeline with error bars representing standard deviation across folds. Pipelines that performed significantly better than default accuracy (p-value ≤ 0.05) are indicated with an asterisk, which occurs only in the delta models and a single week 2 model. Tables below further detail the pipeline settings of top performing models along with mean cross-validation accuracy, and standard deviation.

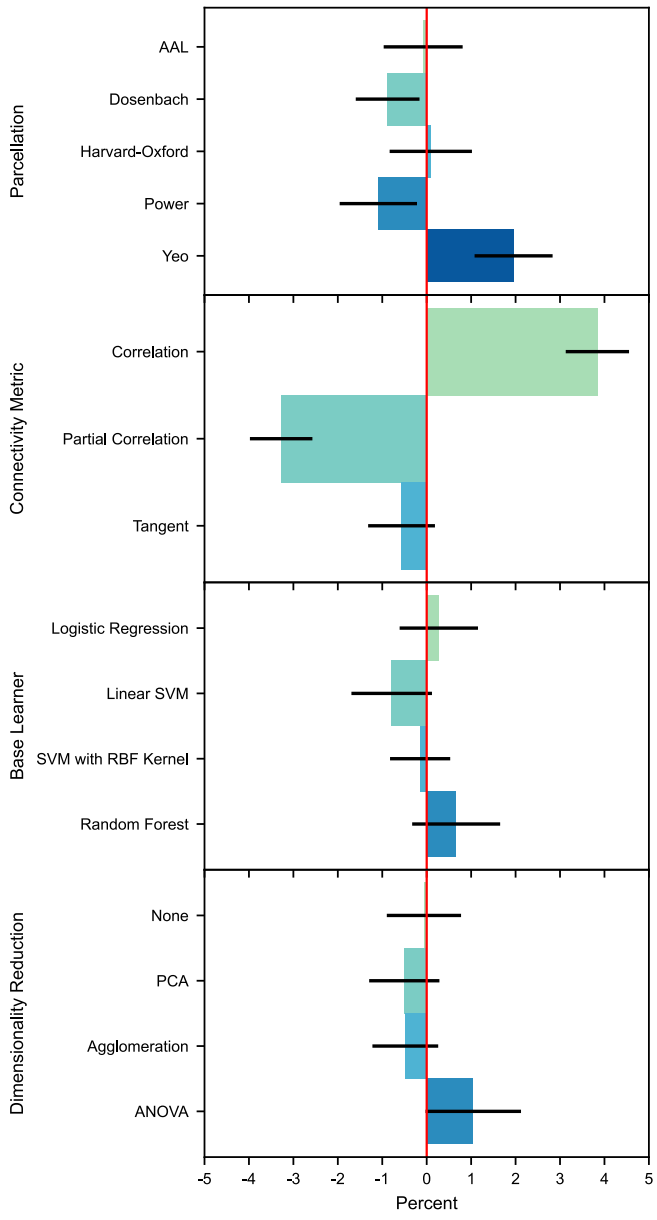


Fig. 3. Impact of analytic pipeline choices on prediction accuracy. For each evaluated step of the analytic pipeline, the figure shows the mean difference between pipelines with indicated alternatives and mean overall predictive accuracy. Error bars are scaled to one-sixth standard deviation. For each step in the pipeline, alternative operations were compared using a Wilcoxon matched-pairs rank test with Bonferroni correction. Steps involved in feature generation had the greatest impact on model performance, specifically the choice of connectivity metric, where correlation showed the highest mean performance. Neither step in model generation (dimensionality reduction or base learner) had a substantial impact on performance.

coverage in these areas may have failed to capture neuronal interactions pivotal to predicting treatment response.

Lower accuracies observed in this study are also consistent with a previously observed, paradoxical relationship between sample size and accuracy in psychiatric predictive models (Kalmady et al., 2019; Varoquaux, 2018; Wolfers et al., 2015); wherein *larger* samples tend to report *lower* overall accuracy. An overall increase in dataset heterogeneity is likely a large contributing factor to this decline, as it becomes more challenging to control heterogeneity as sample sizes increase. Although this heterogeneity may appear to decrease overall accuracy, results of less tightly controlled studies are likely more generalizable to clinical practice, as they tend to encompass more realistic patient populations and multiple collection sites.

Instability in model performance across cross-validation folds, resulting in large error bars, is also a concern in small samples, and leads to uncertainty around the capacity of these models to generalize to new samples. All models tested exhibited a large degree of variance in performance across cross-validation folds; with standard deviations reaching up to 18.7%, which is consistent with empirical results from Varoquaux (2018) for samples of similar size. This result, however, draws into question whether the proposed models are indeed able to classify treatment response above a chance level. Moreover, the number of models run may raise concerns of 'overhyping' (Hosseini et al., 2020), where analysis choices are made after observing test set accuracies. This effect becomes clearly evident in the week 2 model results, where the only significant model has an accuracy over six percent higher than the next best performing model; an effect that is more likely to be spurious, rather than a reflection of superior modelling choices. Even the more consistent performance of the delta models may be questionable in light of the number of models run, although they were found to statistically outperform baseline and week 2 models with a high degree of significance.

Model stability and performance are also negatively impacted by small sample sizes relative to the number of features; a cited concern of neuroimaging features in machine learning (Du et al., 2018; Varoquaux, 2018). Under these conditions, the training samples inadequately cover the feature space leading to overfitting. Fig. 4 demonstrates that across time point and across model type, resubstitution accuracy consistently exceeds test accuracy, an indication of overfitting. Given the inherently high-dimensional nature of FC data, considerably larger sample sizes may be required to elucidate subtle patterns necessary for prediction, although resubstitution accuracy would still be expected to exceed test accuracy in larger real-world data sets but by smaller amounts. Similar concerns have been suggested in utilizing genetic features (Wray et al., 2013), which was cited as a major concern in a recent report by Shumake et al. (2021). Here, authors failed to show improvement in accuracy when including genetic data in models that predict SSRI treatment response. These results were somewhat disappointing given the considerable support in the literature for a heritable component to SSRI response and the interest in the predictive capacity of genetic data in treatment response (Amare et al., 2017).

Amassing data sets of adequate size is extremely challenging, due, in part, to the high costs and scarce resources for image acquisition. Prognostic studies, such as ours, where additional follow-up and patient monitoring is required, bring additional challenges. Improved predictive accuracy observed in the delta models, which include imaging data for each patient from two separate timepoints, requires yet additional resources. Results from Klöbl et al. (2020), however, suggest medication-induced FC alterations may be achieved in a much shorter timeframe by an acute intravenous SSRI challenge. This may allow the week 2 timepoint used in delta models to be acquired at the baseline visit following an intravenous challenge, eliminating the need for a second scanning session and 2-week medication trial before response prediction can be made.

Whether we consider one- or two-timepoints, our current approach requires a large number of participants to find meaningful patterns, leading to good generalization. An alternative approach to increasing sample size may be to reduce the number of features used to describe each patient; either through parcellation choice or *a priori* manual feature selection. When comparing performance of models utilizing different parcellations in Fig. 3 (top), a trend emerges favouring those with fewer regions (and therefore fewer features), which may suggest that larger parcellations produce too many features to be adequately modelled within the given sample. Ultimately, however, smaller parcellations require more voxels to be averaged together in a single region, implicitly imposing regional homogeneity, which may lead to information loss (Dornas and Braun, 2018) important in the predictive task.

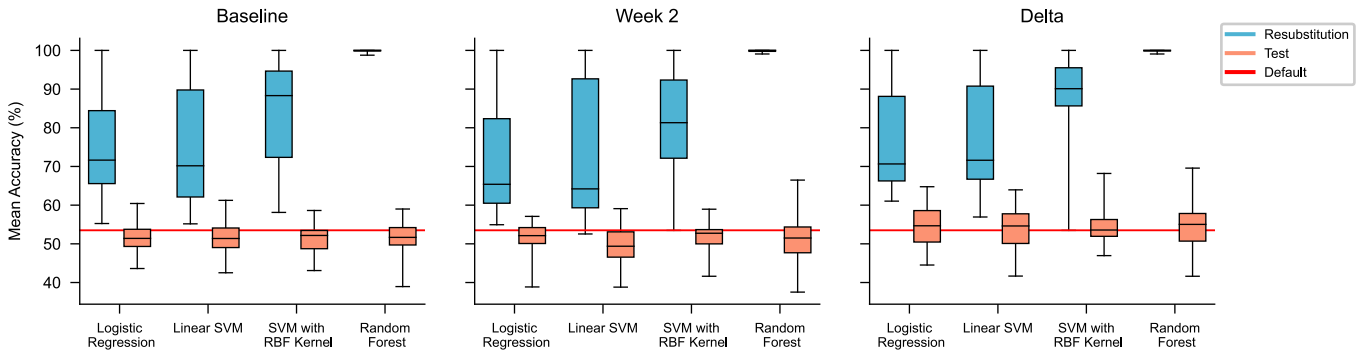


Fig. 4. Mean resubstitution (blue) and testing (orange) accuracy across timepoints and base learners. Mean accuracy results for all models separated by time point and base learner are plotted, where each box extends from the first quartile to the third quartile with median line; whiskers extend to show full data range with mean default model accuracy indicated by the red horizontal line through plots.

Manual feature selection similarly has the potential to overlook connections that may meaningfully contribute to predictions.

Results from delta models in our analysis show a modest improvement over default accuracy, similar to reports from larger, multisite studies such as (1) Chekroud et al. (2016), utilizing clinical measures from the STAR*D study to predict treatment response after 2 weeks reporting a mean accuracy of 67.9% (SD 3.8), which improved over baseline prediction accuracy of 64.65% (SD 3.2), (2) Shumake et al. (2021) using pre-treatment clinical and small nucleotide polymorphism (SNP) features to achieve 62.8% accuracy, and (3) Bartlett et al. (2018) who achieved 63.9% accuracy predicting SSRI remission using both pre-treatment, and also early-treatment cortical thickness measurements.

A plateau in accuracy approaching 70% across these larger studies may be an indication of a theoretical upper limit on the capacity of models to predict response to pharmacological treatment in MDD, likely impacted to some degree by uncertainty associated with the response labels (Cortes et al., 1995). This limit, however, may be different for alternative prediction targets of treatment response such as remission, and treatment resistance (Sajjadian et al., 2021). MADRS scores, the basis of the binary labels of response in this analysis, are inherently noisy due to variability in rater assessment (Davidson et al., 1986), and may also be biased due to factors independent of biological response that impact mood such as changes in external stressors, spontaneous recovery, and placebo effects.

5. Limitations

This study is not without limitations. As an open label trial, there was no placebo group to control for known placebo effects in treatment response. While this dataset is from one of the larger fMRI studies to examine treatment response in MDD, it may still be substantially smaller than what might be necessary for modelling inherently high-dimensional FC data. As discussed earlier, the inability of single-timepoint models (i.e., baseline, or week 2) to exceed chance level performance may be mitigated by including more patients in the training data.

In this analysis we chose to use cross-validation splits with equal proportions of data from each acquisition site. This approach, however, does not assess how the model would perform on data with uncontrolled site-specific variability. Alternatively, defining folds by acquisition site, where each fold consists of data from a single site, might better assess a model's capacity to generalize to new sites. We did not take this approach in our analysis, given the uneven distribution of patients across imaging sites (see Supplementary Materials Table S1) and limited sample size. Ideally, future analyses may evaluate model performance on a completely independent and external testing data set.

Further, while we attempted to explore the impact of different modelling choices on overall performance in predicting treatment response in MDD, this was not an exhaustive search of the space. For example, we limited parcellations to pre-defined atlases and co-ordinate sets but did not consider the potential of data-driven strategies. Novel approaches to each step of the analytic pipeline continue to be developed and may improve classification accuracy. In addition, we note there are state-of-the-art approaches for learning from resting-state fMRI (such as Chen et al., 2022; Santana et al., 2019; Zhao et al., 2022). However, as the main point of this paper is showing that including week 2 data can significantly improve the accuracy over just using the baseline data, we decided it was sufficient to show this on these standard machine learning approaches.

6. Conclusion

This study explored ways to learn a model that can predict the effectiveness of escitalopram for treating MDD, based on RS-fMRI data, taken at baseline and/or at 2 weeks. Our extensive empirical analysis (over 720 possible models) suggests that models that use only the data at baseline, or that use only the data at 2 weeks, are not sufficient for this task. However, we were able to produce several effective learned models that combine the data from both timepoints. While we acknowledge that baseline data might be sufficient, given a larger sample, additional features, and perhaps other learning techniques, we anticipate, even then, including early biological changes following treatment initiation may lead to yet more accurate predictions.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: **RM** has received consulting and speaking honoraria from AbbVie, Allergan, Eisai, Janssen, KYE, Lallemand, Lundbeck, Neonmind, Otsuka, and Sunovion, and research grants from CAN-BIND, Canadian Institutes for Health Research (CIHR), Janssen, Lallemand, Lundbeck, Nubiyota, Ontario Brain Institute (OBI) and Ontario Mental Health Foundation (OMHF). **RWL** has received honoraria for ad hoc speaking or advising/consulting, or received research funds, from: Asia-Pacific Economic Cooperation, BC Leading Edge Foundation, CIHR, Canadian Network for Mood and Anxiety Treatments, Healthy Minds Canada, Janssen, Lundbeck, Lundbeck Institute, Medscape, Michael Smith Foundation for Health Research, MITACS, Myriad Neuroscience, Ontario Brain Institute, Otsuka, Pfizer, Sanofi, Unity Health, Vancouver Coastal Health Research Institute, and VGH/UBCH Foundation. **SHK** has received research funding or honoraria from the following sources: Abbott, AbbVie, Alkermes, Allergan, Boehringer-Ingelheim, Brain Canada, CIHR, Field Trip (stock), Janssen, Lundbeck, Lundbeck Institute, Merck, OBI, Ontario Research Fund (ORF), Otsuka, Pfizer, Servier, SPOR, Sun, and Sunovion. **SCS** is a senior scientific advisor and shareholder of

ADMdx, Inc., which receives NIH funding, and he currently has research grants from the Ontario Brain Institute in Canada. The remaining authors have no conflicts to disclose.

Acknowledgments

This work was completed on behalf of the Canadian Biomarker Integration Network in Depression (CAN-BIND). CAN-BIND is an Integrated Discovery Program carried out in partnership with, and financial support from, the Ontario Brain Institute, an independent non-profit corporation, funded partially by the Ontario government. The opinions, results and conclusions are those of the authors and no endorsement by the Ontario Brain Institute is intended or should be inferred. Additional funding was provided by the Canadian Institutes of Health Research (CIHR), Lundbeck, and Servier. Funding and/or in-kind support is also provided by the investigators' universities and academic institutions. All study medications are independently purchased at wholesale market values. The neuroimaging platform was supported in part by a CIHR grant (Co-PIs: Drs. Kennedy and MacQueen, MOP 125880). This work was supported in part by funding provided by Brain Canada, in partnership with Health Canada, for the Canadian Open Neuroscience Platform (CONP) initiative. JKH is partially supported by grants from the Alberta Machine Intelligence Institute (AMII), NSERC, and CONP.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.nicl.2022.103120.

References

Abraham, A., Milham, M.P., Di Martino, A., Craddock, R.C., Samaras, D., Thirion, B., Varoquaux, G., 2017. Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. Neuroimage 147, 736–745.

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., Varoquaux, G., 2014. Machine learning for neuroimaging with scikit-learn. Front Neuroinform 8, 14.

Amare, A.T., Schubert, K.O., Baune, B.T., 2017. Pharmacogenomics in the treatment of mood disorders: Strategies and Opportunities for personalized psychiatry. EPMA J 8 (3), 211–227.

Anand, A., Li, Y.u., Wang, Y., Wu, J., Gao, S., Bukhari, L., Mathews, V.P., Kalnin, A., Lowe, M.J., 2005. Antidepressant effect on connectivity of the mood-regulating circuit: an FMRI study. Neuropsychopharmacology 30 (7), 1334–1344.

Arnow, B.A., Blasey, C., Williams, L.M., Palmer, D.M., Rekshan, W., Schatzberg, A.F., Etkin, A., Kulkarni, J., Luther, J.F., Rush, A.J., 2015. Depression Subtypes in Predicting Antidepressant Response: A Report From the iSPOT-D Trial. Am J Psychiatry 172 (8), 743–750.

Bartlett, E.A., DeLorenzo, C., Sharma, P., Yang, J., Zhang, M., Petkova, E., Weissman, M., McGrath, P.J., Fava, M., Ogden, R.T., Kurian, B.T., Malchow, A., Cooper, C.M., Trombello, J.M., McInnis, M., Adams, P., Oquendo, M.A., Pizzagalli, D.A., Trivedi, M., Parsey, R.V., 2018. Pretreatment and early-treatment cortical thickness is associated with SSRI treatment response in major depressive disorder. Neuropsychopharmacology 43 (11), 2221–2230.

Chekroud, A.M., Zotti, R.J., Shehzad, Z., Gueorguieva, R., Johnson, M.K., Trivedi, M.H., Cannon, T.D., Krystal, J.H., Corlett, P.R., 2016. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry 3 (3), 243–250.

Chen, Y., Yan, J., Jiang, M., Zhang, T., Zhao, Z., Zhao, W., Zheng, J., Yao, D., Zhang, R., Kendrick, K.M., Jiang, X. 2022. Adversarial Learning Based Node-Edge Graph Attention Networks for Autism Spectrum Disorder Identification. IEEE Trans Neural Netw Learn Syst PP.

Cohen, S.E., Zantvoord, J.B., Wezenberg, B.N., Bockting, C.L.H., van Wingen, G.A., 2021. Magnetic resonance imaging for individual prediction of treatment response in major depressive disorder: a systematic review and meta-analysis. Transl Psychiatry 11, 168.

Cortes, C., Jackel, L.D., Chiang, W.-P., 1995. Limits on learning machine accuracy imposed by data quality. In: Proceedings of the First International Conference on Knowledge Discovery and Data Mining. AAAI Press, Montréal, Québec, Canada, pp. 57–62.

Cullen, K.R., Gee, D.G., Klimes-Dougan, B., Gabbay, V., Hulvershorn, L., Mueller, B.A., Camchong, J., Bell, C.J., Houri, A., Kumra, S., Lim, K.O., Castellanos, F.X., Milham, M.P., 2009. A preliminary study of functional connectivity in comorbid adolescent depression. Neurosci Lett 460 (3), 227–231.

Dadi, K., Rahim, M., Abraham, A., Chyzhyk, D., Milham, M., Thirion, B., Varoquaux, G., Initiative, A.s.D.N., 2019. Benchmarking functional connectome-based predictive models for resting-state fMRI. Neuroimage 192, 115–134.

Davidson, J., Turnbull, C.D., Strickland, R., Miller, R., Graves, K., 1986. The Montgomery-Asberg Depression Scale: reliability and validity. Acta Psychiatr Scand 73, 544–548.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31 (3), 968–980.

Dornas, J.V., Braun, J., 2018. Finer parcellation reveals detailed correlational structure of resting-state fMRI signals. J Neurosci Methods 294, 15–33.

Dosenbach, N.U.F., Nardos, B., Cohen, A.L., Fair, D.A., Power, J.D., Church, J.A., Nelson, S.M., Wig, G.S., Vogel, A.C., Lessov-Schlaggar, C.N., Barnes, K.A., Dubis, J. W., Feczko, E., Coalson, R.S., Pruett, J.R., Barch, D.M., Petersen, S.E., Schlaggar, B. L., 2010. Prediction of individual brain maturity using fMRI. Science 329 (5997), 1358–1361.

Drysdale, A.T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R. N., Zebley, B., Oathes, D.J., Etkin, A., Schatzberg, A.F., Sudheimer, K., Keller, J., Mayberg, H.S., Gunning, F.M., Alexopoulos, G.S., Fox, M.D., Pascual-Leone, A., Voss, H.U., Casey, B.J., Dubin, M.J., Liston, C., 2017. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat Med 23 (1), 28–38.

Du, Y., Fu, Z., Calhoun, V.D., 2018. Classification and Prediction of Brain Disorders Using Functional Connectivity: Promising but Challenging. Front Neurosci 12, 525.

Fried, E.I., Nesse, R.M., 2015. Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. J Affect Disord 172, 96–102.

Gao, S., Calhoun, V.D., Sui, J., 2018. Machine learning in major depression: From classification to treatment outcome prediction. CNS Neurosci Ther 24, 1037–1052.

Goldstein-Piekarski, A.N., Staveland, B.R., Ball, T.M., Yesavage, J., Korgaonkar, M.S., Williams, L.M., 2018. Intrinsic functional connectivity predicts remission on antidepressants: a randomized controlled trial to identify clinically applicable imaging biomarkers. Transl Psychiatry 8, 57.

Greicius, M.D., Flores, B.H., Menon, V., Glover, G.H., Solvason, H.B., Kenna, H., Reiss, A. L., Schatzberg, A.F., 2007. Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus. Biol Psychiatry 62 (5), 429–437.

Hosseini, M., Powell, M., Collins, J., Callahan-Flintoft, C., Jones, W., Bowman, H., Wyble, B., 2020. I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. Neurosci Biobehav Rev 119, 456–467.

Jakubovski, E., Bloch, M.H., 2014. Prognostic subgroups for citalopram response in the STAR*D trial. J Clin Psychiatry 75 (07), 738–747.

Kalmady, S.V., Greiner, R., Agrawal, R., Shivakumar, V., Narayanaswamy, J.C., Brown, M.R.G., Greenshaw, A.J., Dursun, S.M., Venkatasubramanian, G., 2019. Towards artificial intelligence in mental health by improving schizophrenia prediction with multiple brain parcellation ensemble-learning. NPJ Schizophr 5, 2.

Kennedy, S.H., Lam, R.W., Rotzinger, S., Milev, R.V., Blier, P., Downar, J., Evans, K.R., Farzan, F., Foster, J.A., Frey, B.N., Giacobbe, P., Hall, G.B., Harkness, K.L., Hassel, S., Ismail, Z., Leri, F., McInerney, S., MacQueen, G.M., Minuzzi, L., Müller, D.J., Parikh, S.V., Placenza, F.M., Quilty, L.C., Ravindran, A.V., Sassi, R.B., Soares, C.N., Strother, S.C., Turecki, G., Vaccarino, A.L., Vila-Rodriguez, F., Yu, J., Uher, R., 2019. Symptomatic and Functional Outcomes and Early Prediction of Response to Escitalopram Monotherapy and Sequential Adjunctive Aripiprazole Therapy in Patients With Major Depressive Disorder: A CAN-BIND-1 Report. J Clin Psychiatry 80 (2).

Klöbl, M., Gryglewski, G., Rischka, L., Godbersen, G.M., Unterholzner, J., Reed, M.B., Michenthaler, P., Vanicek, T., Winkler-Pjrek, E., Hahn, A., Kasper, S., Lanzenberger, R., 2020. Predicting Antidepressant Citalopram Treatment Response via Changes in Brain Functional Connectivity After Acute Intravenous Challenge. Front Comput Neurosci 14, 554186.

Kraus, C., Kadriu, B., Lanzenberger, R., Zarate, C.A., Kasper, S., 2019. Prognosis and improved outcomes in major depression: a review. Transl Psychiatry 9, 127.

Lam, R.W., Milev, R., Rotzinger, S., Andreazza, A.C., Blier, P., Brenner, C., Daskalakis, Z. J., Dharsee, M., Downar, J., Evans, K.R., Farzan, F., Foster, J.A., Frey, B.N., Geraci, J., Giacobbe, P., Feilotter, H.E., Hall, G.B., Harkness, K.L., Hassel, S., Ismail, Z., Leri, F., Liotti, M., MacQueen, G.M., McAndrews, M.P., Minuzzi, L., Müller, D.J., Parikh, S.V., Placenza, F.M., Quilty, L.C., Ravindran, A.V., Salomons, T. V., Soares, C.N., Strother, S.C., Turecki, G., Vaccarino, A.L., Vila-Rodriguez, F., Kennedy, S.H., 2016. Discovering biomarkers for antidepressant response: protocol from the Canadian biomarker integration network in depression (CAN-BIND) and clinical characteristics of the first patient cohort. BMC Psychiatry 16 (1).

MacQueen, G.M., Hassel, S., Arnott, S.R., Addington, J., Bowie, C.R., Bray, S.L., Davis, A. D., Downar, J., Foster, J.A., Frey, B.N., Goldstein, B.I., Hall, G.B., Harkness, K.L., Harris, J., Lam, R.W., Lebel, C., Milev, R., Müller, D.J., Parikh, S.V., Rizvi, S., Rotzinger, S., Sharma, G.B., Soares, C.N., Turecki, G., 'Vila-Rodriguez, F., Yu, J., Zamyadi, M., Strother, S.C., Kennedy, S.H., 2019. The Canadian Biomarker Integration Network in Depression (CAN-BIND): magnetic resonance imaging protocols. J Psychiatry Neurosci 44 (4), 223–236.

Mayberg, H.S., 1997. Limbic-cortical dysregulation: a proposed model of depression. J Neuropsychiatry Clin Neurosci 9, 471–481.

Montgomery, S.A., Åsberg, M., 1979. A new depression scale designed to be sensitive to change. Br J Psychiatry 134 (4), 382–389.

Parkes, L., Fulcher, B., Yücel, M., Fornito, A., 2018. An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI. Neuroimage 171, 415–436.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2825–2830.

Posner, J., Hellerstein, D.J., Gat, I., Mechling, A., Klahr, K., Wang, Z., McGrath, P.J., Stewart, J.W., Peterson, B.S., 2013. Antidepressants normalize the default mode network in patients with dysthymia. JAMA Psychiatry 70, 373–382.

Power, J., Cohen, A., Nelson, S., Wig, G., Barnes, K., Church, J., Vogel, A., Laumann, T., Miezin, F., Schlaggar, B., Petersen, S., 2011. Functional network organization of the human brain. Neuron 72 (4), 665–678.

Ramasubbu, R., Brown, M.R., Cortese, F., Gaxiola, I., Goodyear, B., Greenshaw, A.J., Dursun, S.M., Greiner, R., 2016. Accuracy of automated classification of major depressive disorder as a function of symptom severity. Neuroimage Clin 12, 320–331.

Rush, A.J., Trivedi, M.H., Wisniewski, S.R., Nierenberg, A.A., Stewart, J.W., Warden, D., Niederehe, G., Thase, M.E., Lavori, P.W., Lebowitz, B.D., McGrath, P.J., Rosenbaum, J.F., Sackeim, H.A., Kupfer, D.J., Luther, J., Fava, M., 2006. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. Am J Psychiatry 163 (11), 1905–1917.

Sajjadian, M., Lam, R.W., Milev, R., Rotzinger, S., Frey, B.N., Soares, C.N., Parikh, S.V., Foster, J.A., Turecki, G., Müller, D.J., Strother, S.C., Farzan, F., Kennedy, S.H., Uher, R., 2021. Machine learning in the prediction of depression treatment outcomes: a systematic review and meta-analysis. Psychological Medicine 51 (16), 2742–2751.

Santana, A.N., Cifre, I., de Santana, C.N., Montoya, P., 2019. Using Deep Learning and Resting-State fMRI to Classify Chronic Pain Conditions. Front Neurosci 13, 1313.

Schnack, H.G., Kahn, R.S., 2016. Detecting Neuroimaging Biomarkers for Psychiatric Disorders: Sample Size Matters. Front Psychiatry 7, 50.

Sheehan, D.V., Lecrubier, Y., Sheehan, K.H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., Dunbar, G.C. 1998. The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry. 59 Suppl 20. 22-33. quiz 34-57.

Sheline, Y.I., Barch, D.M., Price, J.L., Rundle, M.M., Vaishnavi, S.N., Snyder, A.Z., Mintun, M.A., Wang, S., Coalson, R.S., Raichle, M.E., 2009. The default mode network and self-referential processes in depression. Proc Natl Acad Sci U S A 106 (6), 1942–1947.

Shumake, J., Mallard, T.T., McGeary, J.E., Beevers, C.G., 2021. Inclusion of genetic variants in an ensemble of gradient boosting decision trees does not improve the prediction of citalopram treatment response. Sci Rep 11, 3780.

Sundermann, B., Feder, S., Wersching, H., Teuber, A., Schwindt, W., Kugel, H., Heindel, W., Arolt, V., Berger, K., Pfleiderer, B., 2017. Diagnostic classification of unipolar depression based on resting-state functional connectivity MRI: effects of generalization to a diverse sample. J Neural Transm (Vienna) 124 (5), 589–605.

Szegedi, A., Jansen, W.T., van Willigenburg, A.P.P., van der Meulen, E., Stassen, H.H., Thase, M.E., 2009. Early improvement in the first 2 weeks as a predictor of treatment outcome in patients with major depressive disorder: a meta-analysis including 6562 patients. J Clin Psychiatry 70 (3), 344–353.

Thornicroft, G., Chatterji, S., Evans-Lacko, S., Gruber, M., Sampson, N., Aguilar-Gaxiola, S., Al-Hamzawi, A., Alonso, J., Andrade, L., Borges, G., Bruffaerts, R., Bunting, B., de Almeida, J.M.C., Florescu, S., de Girolamo, G., Gureje, O., Haro, J.M., He, Y., Hinkov, H., Karam, E., Kawakami, N., Lee, S., Navarro-Mateu, F., Piazza, M., Posada-Villa, J., de Galvis, Y.T., Kessler, R.C., 2017. Undertreatment of people with major depressive disorder in 21 countries. Br J Psychiatry 210 (2), 119–124.

Truong, T.-K., Clymer, B.D., Chakeres, D.W., Schmalbrock, P., 2002. Three-dimensional numerical simulations of susceptibility-induced magnetic field inhomogeneities in the human head. Magn Reson Imaging 20 (10), 759–770.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15 (1), 273–289.

Varoquaux, G., 2018. Cross-validation failure: Small sample sizes lead to large error bars. Neuroimage 180, 68–77.

Veer, I.M., Beckmann, C.F., van Tol, M.J., Ferrarini, L., Milles, J., Veltman, D.J., Aleman, A., van Buchem, M.A., van der Wee, N.J., Rombouts, S.A. 2010. Whole brain resting-state analysis reveals decreased functional connectivity in major depression. Front Syst Neurosci. 4.

Wang, L.i., An, J., Gao, H.-M., Zhang, P., Chen, C., Li, K.e., Mitchell, P.B., Si, T.-M., 2019. Duloxetine effects on striatal resting-state functional connectivity in patients with major depressive disorder. Hum Brain Mapp 40 (11), 3338–3346.

Wang, L.i., Xia, M., Li, K.e., Zeng, Y., Su, Y., Dai, W., Zhang, Q., Jin, Z., Mitchell, P.B., Yu, X., He, Y., Si, T., 2015. The effects of antidepressant treatment on resting-state functional brain networks in patients with major depressive disorder. Hum Brain Mapp 36 (2), 768–778.

Wolfers, T., Buitelaar, J.K., Beckmann, C.F., Franke, B., Marquand, A.F., 2015. From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. Neurosci Biobehav Rev 57, 328–349.

Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E., Visscher, P.M., 2013. Pitfalls of predicting complex traits from SNPs. Nat Rev Genet 14 (7), 507–515.

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

Thomas Yeo, B.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zöllei, L., Polimeni, J.R., Fischl, B., Liu, H., Buckner, R.L., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J Neurophysiol 106 (3), 1125–1165.

Zhao, M., Yan, W., Luo, N.a., Zhi, D., Fu, Z., Du, Y., Yu, S., Jiang, T., Calhoun, V.D., Sui, J., 2022. An attention-based hybrid deep learning framework integrating brain connectivity and activity of resting-state functional MRI data. Med Image Anal 78, 102413.