NeuroImage: Clinical 12 (2016) 320–331

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

Accuracy of automated classification of major depressive disorder as a function of symptom severity

Rajamannar Ramasubbu MD, FRCPC, MSc[a,b,c,d,*,1], Matthew R.G. Brown PhD[e,f,g,1], Filmeno Cortese MSc[d], Ismael Gaxiola MSc[d], Bradley Goodyear PhD[d], Andrew J. Greenshaw PhD[e], Serdar M. Dursun MD, PhD[e], Russell Greiner PhD[f,g]

[a]Department of Psychiatry, University of Calgary, Calgary, AB, Canada
[b]Department of Clinical Neuroscience, University of Calgary, AB, Canada
[c]Mathison Centre for Mental Health Research and Education, University of Calgary, Calgary, AB, Canada
[d]Hotchkiss Brain Institute, University of Calgary, Calgary, AB, Canada
[e]Department of Psychiatry, University of Alberta, Edmonton, AB, Canada
[f]Department of Computing Science, University of Alberta, Edmonton, AB, Canada
[g]Alberta Innovates Centre for Machine Learning, Edmonton, AB, Canada

ARTICLE INFO

Article history:
Received 23 October 2015
Received in revised form 7 July 2016
Accepted 26 July 2016
Available online 27 July 2016

Keywords:
Major depression
Severity of symptoms
Diagnosis
Functional magnetic resonance imaging
Machine learning
Classification
Support vector machine

ABSTRACT

Background: Growing evidence documents the potential of machine learning for developing brain based diagnostic methods for major depressive disorder (MDD). As symptom severity may influence brain activity, we investigated whether the severity of MDD affected the accuracies of machine learned MDD-vs-Control diagnostic classifiers.

Methods: Forty-five medication-free patients with DSM-IV defined MDD and 19 healthy controls participated in the study. Based on depression severity as determined by the Hamilton Rating Scale for Depression (HRSD), MDD patients were sorted into three groups: mild to moderate depression (HRSD 14–19), severe depression (HRSD 20–23), and very severe depression (HRSD ≥24). We collected functional magnetic resonance imaging (fMRI) data during both resting-state and an emotional-face matching task. Patients in each of the three severity groups were compared against controls in separate analyses, using either the resting-state or task-based fMRI data. We use each of these six datasets with linear support vector machine (SVM) binary classifiers for identifying individuals as patients or controls.

Results: The resting-state fMRI data showed statistically significant classification accuracy only for the *very severe depression* group (accuracy 66%, $p = 0.012$ corrected), while *mild to moderate* (accuracy 58%, $p = 1.0$ corrected) and *severe depression* (accuracy 52%, $p = 1.0$ corrected) were only at chance. With task-based fMRI data, the automated classifier performed at chance in all three severity groups.

Conclusions: Binary linear SVM classifiers achieved significant classification of very severe depression with resting-state fMRI, but the contribution of brain measurements may have limited potential in differentiating patients with less severe depression from healthy controls.

1. Introduction

Major depressive disorder (MDD) is a complex brain disorder associated with dysregulation of distributed neuronal networks involving several cortical and limbic regions. This position is based on evidence from the neuroimaging literature that has documented distinct structural and functional alterations in patients with MDD compared to healthy controls (Mayberg, 2003; Drevets et al., 2008; Price and Drevets, 2012). However, these group-level inferences have had minimal impact on clinical translation at the individual patient level – that is, they do not directly lead to a way to determine whether a specific subject has MDD or not. Recently, machine learning techniques have been applied to neuroimaging data to draw inferences for individual subjects, with the potential for improving patient-specific clinical diagnostic and treatment decisions (Orru et al., 2012; Kloppel et al., 2012). Current diagnosis of mental disorders is based on diagnostic criteria drawn from self-reported clinical symptoms without any objective biomarkers. This has led to the search, in recent years, for a diagnostic system that can use objective measurements from a subject's brain to validate and improve the accuracy of psychiatric diagnosis.

* Corresponding author at: Department of Psychiatry and Clinical Neurosciences, University of Calgary, Mathison Centre for Mental Health, Research and Education, TRW building, Room 4D64, 3280 Hospital Drive NW, Calgary, Alberta, T2N4Z6, Canada.
E-mail address: rramasub@ucalgary.ca (R. Ramasubbu).
[1] Equal contributions as first author.

http://dx.doi.org/10.1016/j.nicl.2016.07.012

In the last decade, several neuroimaging studies have examined the classification accuracy of machine learned classifiers in differentiating patients with MDD from healthy controls. One major focus has been the application of machine learning techniques to magnetic resonance imaging (MRI) data, including both structural and function MRI (fMRI) data. Machine learning is a sub-area of artificial intelligence that applies statistical methods to training data, such as high dimensional neuroimaging data, to find patterns that can distinguish patients from healthy controls. Authors reported classification accuracy for MDD ranging from 67 to 90% using structural MRI data (Costafreda et al., 2009; Gong et al., 2011; Mwangi et al., 2012a), 94% using resting-state fMRI data (Zeng et al., 2012; Zeng et al., 2014), 67–86% using task-related fMRI data (Fu et al., 2008; Marquand et al., 2008; Hahn et al., 2011) and 76.3% using combined structural and functional MRI data (Nouretdinov et al., 2011). High accuracy prediction is clinically important, as MDD is heterogeneous in symptom profile and prone to clinician bias with poor inter-rater reliability (Regier et al., 2013). The identification of MDD subtypes based on neural abnormalities or brain imaging methods might improve classification accuracy, facilitate new drug discovery and move toward stratified medicine.

Depression subtypes defined by symptom severity have several clinical implications for the treatment and prognosis. For example, baseline symptom severity is associated with drug-placebo differences in randomized control trials (Kirsch et al., 2008) and antidepressants are recommended as the choice of treatment for severe depression whereas psychosocial interventions as the choice of treatment for mild-moderate subthreshold depression (NICE guidelines CG90, 2009). Additionally, epidemiological studies have shown the association of symptom severity with functional impairment, co-morbidity and increased risk of mortality (Kessler et al., 2003; Kessler et al., 2005; Rutledge et al., 2006). In machine learning approaches, severity-related brain abnormalities have been shown to offer good discriminating potential in the classification of MDD and healthy controls. In emotional task fMRI data, Mourao-Miranda et al. (2011) found significant correlations between the distance of participants' feature vectors from the separating hyperplane of a trained support vector machine, and those participants' severity scores from the Hamilton Rating Scale for Depression (HRSD) (Hamilton, 1960), which suggests a relationship between depression severity and test predictions (Mourao-Miranda et al., 2011). Similarly, another study using structural MRI data reported a strong relationship between the fitted SVM weights and ratings of illness severity (Mwangi et al., 2012b). These findings suggest that fitted machine learned classifiers may capture patterns of brain abnormality in functional and structural neuroimaging data related to MDD severity. A model derived from a machine learned classifier may constitute an objective biomarker for depression severity. To date, no previous study has examined how the performance of machine learning algorithms in differentiating MDD vs. health may differ as a function of MDD symptom severity. This research question has important clinical implications in the context of whether machine learning approaches using fMRI data can yield comparable accuracy in the classification of MDD at various levels of severity.

We examined the accuracy of two-class machine learning classification of three distinct groups of MDD patients, with different levels of symptom severity based on the HRSD Scores, versus healthy controls. The three groups of MDD with severity gradation were: mild to moderate depression (HRSD score 14–19), severe depression (HRSD 20–23), and very severe depression (HRSD ≥24). (While there is no consensus on cutoff scores on the HRSD for identifying MDD severity subtypes, these severity ranges are consistent with several published recommendations (Zimmerman et al., 2013; Rush et al., 2008; DeRubeis et al., 1999)). We expected that the classifiers would achieve higher accuracy for the patient groups with very severe depression compared to those with severe depression or mild-moderate depression. For each range of severity, we also considered two types of fMRI data – from either resting-state or from an emotional-face matching task – hence, we examined classifier performance for 3×2 different situations.

2. Materials & methods

2.1. Participants

Ethics approval was obtained from the local review board. All participants were fluent in English and gave informed, written consent to participate in the study. Forty-five patients meeting DSM-IV criteria for MDD (Association AP, 2000) according to the Structured Clinical Interview for DSM-IV Axis 1 Disorders (First et al., 2002a), were recruited through advertisements. (See Table 1 for participant demographics). Patients included 29 females and 16 males, all right-handed, in the age range of 19–58 years (mean 37 ± 11 SD). The Edinburgh Handedness Inventory was used to assess handedness (Oldfield, 1971). The severity of depressive and anxiety symptoms was assessed using the clinician-administered, 17-item Hamilton Rating Scale for Depression (Hamilton, 1960), the Montgomery Asberg Depression Rating Scale (MADRS) (Montgomery and Asberg, 1979), and the Hamilton Anxiety Rating Scale (HAM-A) (Hamilton, 1959). Patients were also rated for disease severity using the Clinical Global Impression (CGI) scale (Guy, 1976), which allows clinicians to provide a severity rating based on their clinical experience. Patients were included in the study if they met the following inclusion criteria: (1) acute episode of MDD of unipolar subtype and a score of 14 or higher on the HRSD, and (2) free of psychotropic medication for a minimum of three weeks at time of recruitment. Exclusion criteria were: (1) Axis I disorders such as bipolar disorder, anxiety disorder, or psychosis, (2) history of substance abuse within six months of study participation, (3) borderline personality disorder, (4) medical or neurological disorders, (5) severe suicidal symptoms, (6) failure to respond to three trials of antidepressant

Table 1
Characteristics of three MDD patient groups and healthy controls.

Characteristic	All MDD patients	Mild-moderate MDD	Severe MDD	Very severe MDD	Healthy controls	p-Value (patients vs. controls)	p-Value (3 MDD groups omnibus comparison)
n	45	12	18	15	19		
Sex (% female)	64%	42%	67%	80%	58%	0.31	0.09
Age (years)	37 ± 11	33 ± 11	38 ± 10	37 ± 11	33 ± 10	0.18	0.39
Age of onset (years)	24 ± 10	19 ± 5	26 ± 10	27 ± 11	–	–	0.10
Illness duration (years)	12 ± 8	14 ± 11	13 ± 7	10 ± 7	–	–	0.50
Duration of current episode (months)	59 ± 66	42 ± 54	72 ± 73	57 ± 69	–	–	0.48
HRSD score	22 ± 4	17 ± 1	21 ± 1	26 ± 2	3 ± 3	10^{-27}	10^{-15}
HAM-A score	24 ± 5	19 ± 4	24 ± 3	27 ± 5	–	–	10^{-6}
CGI score	4.1 ± 0.9	3.1 ± 0.2	4.1 ± 0.2	5.1 ± 0.4	–	–	0.001
MADRS scores	26 ± 6	20 ± 4	25 ± 4	31 ± 4			0.001

Age, Age of onset, Illness duration, Duration of current episode, HRSD score, HAM-A score, and CGI score rows show mean values ± standard deviations. First p-value column shows p-values for tests comparing all patients vs. controls (*t*-test or proportion test as appropriate). Second p-value column shows p-values from omnibus tests comparing the three patient groups (F-test or chi-squared test as appropriate).